Empirical, Mechanistic and Hybrid Models for Mammalian Cell Cultures

by

Mariana Carvalho

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Chemical Engineering

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Prof. Kim McAuley Department of Chemical Engineering, Queen's University
Supervisor(s):	Prof. Hector Budman Department of Chemical Engineering, University of Waterloo
Internal Member:	Prof. William A. Anderson Department of Chemical Engineering, University of Waterloo
Internal Member:	Prof. Valerie Ward Department of Chemical Engineering, University of Waterloo
Internal-External Member:	Prof. Brendan McConkey Department of Biology, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapter 3 has been published as the paper below, author contributed as the first author in developing the idea of the paper, implementation of the method and writing the manuscript:

Mariana Carvalho, Ali Nikdel, Jeremiah Riesberg, Delia Lyons, Hector Budman. Identification of a Dynamic Metabolic Flux Model for a Mammalian Cell Culture. Part of special issue: *12th IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems*, Volume 52, Issue 1, Pages 88-93, 2019. ISSN 2405-8963.

Chapter 4 has been published as the paper below, author contributed as the first author in developing the idea, implementation of the method and writing the manuscript:

Mariana Carvalho, Jeremiah Riesberg, Hector Budman. Development of new media formulations for cell culture operations based on regression models. *Bioprocess Biosyst Eng.*, 44(3):453-472, 2021. doi: 10.1007/s00449-020-02456-9. Epub 2020 Oct 28. PMID: 33111178.

Chapter 5 has been submitted to the Biochemical Engineering Journal.

Abstract

To reach the increasing demand for monoclonal antibodies the pharmaceutical industry has been looking into ways to optimize productivity. Monoclonal antibodies (mAb) are commonly synthesized in mammalian cell cultures. Chinese Hamster Ovary (CHO) cells are among the most used cells by the industry for mAb production. Changes in medium formulation used in cell culturing is a crucial factor among others for maximizing mAb productivity. Mathematical modelling is a useful tool for better understanding the behavior of each medium component on cellular metabolism and for developing new medium formulations by model based optimization.

Mechanistic models, such as metabolic flux models and nonlinear differential kinetic models, have been used in the literature to study cellular performance. These mechanistic models are developed based on fundamental laws of mass balance and stoichiometric information. However, several assumptions and simplifications are done during model development due to the complexity of mammalian cells metabolism. For example, media formulations contain over one hundred chemical components and the effect of each component on growth and productivity is sometimes not well understood. On the other hand, empirical models have the potential ability to describe the effect of many metabolites using experimental data but they are expected to lack prediction accuracy beyond the data used for model calibration. Part of the lack of prediction accuracy of empirical models is related to the presence of many media components which may require many model parameters and thus tend to overfit the data. Following the above, this work presents novel extensions and applications of empirical, mechanistic and hybrid models that combine mechanistic and empirical terms.

Dynamic metabolic flux analysis (DMFA) is a modelling approach that has gained attention in recent years due to its potential to describe dynamic cell behaviour while correctly accounting for metabolic network information. This approach assumes that the cell optimally allocates resources to maximize/minimize a biological objective function of interest. Although DMFA has been applied extensively to bacteria it has not been much investigated for mammalian cell cultures due to the relatively complex behaviour of the latter. Chapter 3 presents an extension of a Dynamic Metabolic Flux Model for mammalian cells with novel extensions to an earlier model that include i- prediction of biomass, ii- inclusion of cell death in the model, and iii- application for both batch and perfusion systems. The final DMFM model was proved very satisfactory, being able to predict the data with an average error of 15%.

Empirical approaches are used by the industry to develop media formulation. However, the reported empirical models are linear and they have not accounted for possible interactions

among the components present in the cellular medium. Chapter 4 presents a new empirical approach that includes linear, two-way interactions and squared terms representing possible components interactions, including interactions with minor components present in the medium, such as vitamins, metals, etc. Due to the large number of components present in the media formulation, the resulting problem is defined as a high dimensional problem where the number of predictor variables p is much larger than the number of observations n. The goal is to find parsimonious models to avoid over-fitting of the training data. A number of techniques were investigated to reduce the parameters such as Lasso and multivariate statistical methods. PCR and PLS model regressions were found suitable to address the high dimensional problem and their predictions have shown good fit with the data. The resulting PCR and PLS models were used for development of new medium formulation by robust optimization. Also, an approach that combines PCR and PLS regression with D-optimal design is presented to select a subset of most informative media to be used for media optimization thus reducing the required number of experiments.

An alternative type of modelling approach involves the use of mechanistic models based on mass balances and kinetic rate expressions. When applying these models with different media formulations they often exhibit discrepancies between predictions and experimental data. A key source of error is due to lack of knowledge about the metabolic effect of minor elements. To correct for these errors hybrid models have been investigated in this thesis which aim to incorporate effect of minor elements by empirical terms into the mechanistic model of the major components. Following this approach Chapter 5 proposes the use of a hybrid model that combines a mechanistic model of amino acids, glucose, by-products and biomass with an empirical PLS based regression model that captures the effect of the minor media elements. The original mechanistic model is described by a set of nonlinear differential equations that uses kinetic terms of Michaelis–Menten form to predict the profile of major medium metabolites. The empirical model is used to capture information regarding medium components interactions, including the effect of minor components, to the final hybrid model. Results have shown that better predictions were obtained by the hybrid model in comparison with the original mechanistic model. Reduction of the dynamic autocorrelation error between model and data was also observed by the hybrid model prediction.

The model approaches presented in this thesis are shown to be efficient for describing the performance of mammalian cell cultures. A significant part of this work has been developed in collaboration with an industrial partner (MilliporeSigma - A business of Merck KGaA, Darmstadt, Germany) on the development of growth media formulations for mammalian cell cultures in batch and perfusion operations.

Acknowledgements

I would like to sincerely thank my supervisor Professor Hector Budman for his continuous support, guidance, and for being a mentor to me, encouraging me throughout the course of my research.

I also would like to thank the other members of my committee, Professors William Anderson, Raymond Legge, Brendan McConkey, Valerie Ward and Kim McAuley for agreeing to be part of my PhD advising committee.

Thank you to Jeremiah Riesberg and Delia Lyons for their support and great discussions. Thank you to Judy Caron and Ralph Dickhout for their help in the Chemical Engineering Department. Thank you to Rubin Hille and Ali Nikdel for sharing their experience and knowledge with their past work. Thank you to my colleagues Piyush Agarwal, Xin Shen, Meghana Chepuru, Sima Lashkari, Yue Yuan and Mina Rafiei for sharing their experience and a great time together.

I would like to express my gratitude to my family, my partner Vicente, my parents Margarete and Marinaldo, my brother Murilo, and my cats Muna and Fred for their support and encouragement throughout my study.

Finally, I would like to thank Natural Science and Engineering Research Council (NSERC), Fields – CQAM and MilliporeSigma - A business of Merck KGaA, Darmstadt, Germany for funding my research.

Dedication

This thesis is dedicated to my family.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADP	Adenosine di-phosphate
AIC	Akaike Information Criteria
AQC	6-aminoquinolyl-N-hydroxysuccinimidyl carbomate
ATP	Adenosine tri-phosphate
CHO	Chinese Hamster Ovary
CD	Cell Density
DFBA	Dynamic Flux Balance Analysis
DMFM	Dynamic Metabolic Flux Model
DoE	Design of Experiments
FADH	Flavin Adenine Dinucleotide reduced
FBA	Flux Balance Analysis
GLC	Glucose
HCA	Hierarchical Cluster Analysis
HPLC	High Performance Liquid Chromatography
JCR	Joint Confidence Region
LASSO	Least Absolute Shrinkage Selector Operator
LP	Linear Programming
mAb,MAB	Monoclonal Antibody
MBDoE	Model-Based Design of Experiments

MFA	Metabolic Flux Analysis
MPLS	Multi-way Partial Least Squares
NAD+	Nicotinamide Adenine Dinucleotide oxidized
NADH	Nicotinamide Adenine Dinucleotide reduced
NAD(P)H	Nicotinamide Adenine Dinucleotide Phosphate reduced
OFAT	One-factor-at-a-time
PC	Principal Components
PCA	Principal Components Analysis
PCR	Principal Components Regression
PLS	Partial Least Squares
Qp	Specific Productivity
RIP	Restricted Isometry Property
SSE	Sum of Squared Error
TCA	Tricarboxylic Acid
VCD	Viable Cells Density
VIP	Variable Importance in Projection

List of Symbols

α	mixing parameter between ridge
Amm	Ammonia concentration
β	regression coefficients
B	bleeding rate
$\mathbf{c}$	vector of constant values
c_1, c_2	cell death function constants
ε	noise
F	feeding rate
Glc	Glucose concentration
Glu	Glutamate concentration
H	harvest rate
Δk	sampling time interval
k	sampling time
K	kinetic parameters
Lac	Lactate concentration
$\boldsymbol{\lambda}$	vector of Lagrange multipliers
λ'	regularization penalty parameter
L	lower bound
M	number of major medium components

n	number of samples (observations)
N	number of minor medium components
n_{PC}	number of principal components
n_t	total number of time intervals
n_y	total number of response variables
p	number of variables (regressors)
ϕ	minor component concentration
ψ	major component concentration
$\boldsymbol{\Psi}$	vector of metabolites concentration
q	number of fluxes in the stoichiometric matrix
Qp_t	specific productivity at time t
r	number of metabolites in the stoichiometric matrix
R	uptake/production rate
$\mathbf{S}$	stoichiometric matrix
S_ϑ	sensitivity
t	time
T	final experimental data time
t_f	final sampling time
θ	robust optimization variance constant
ϑ	hybrid model parameter
U	upper bound
v	specific metabolic flux
V	volume of reactor
$\mathbf{v}$	vector of metabolic fluxes
var	variance
$\mathbf{X}$	input variable matrix

X0	mean-centered input matrix
XL	input loading matrix
$\mathbf{X}_{regressor}$	regressor matrix
XS	input scores matrix
X_v	biomass concentration
W	matrix of PLS weights
y	output response
Y	output response vector
Y0	mean-centered output vector
YL	output loading matrix

Chapter 1

Introduction

The demand for therapeutic proteins, such as monoclonal antibodies, have been exponentially increasing through the last decade due to their successful use in medical treatments. Monoclonal antibodies (mAbs) are recombinant proteins that are currently used in the treatment of rheumatoid arthritis, multiple sclerosis, HIV infection, Alzheimer disease, breast cancer and colorectal cancer [127, 128]. By 2025, it is estimated that the antibody global market will reach over 179 billion dollars, an increase of over 50% if compared to the market in 2020 [35], as seen in Figure 1.1. Therefore, in order to meet this growing demand for mAbs, pharmaceutical industries have been looking for more efficient manufacturing techniques to increase the productivity of mAbs while maintaining their therapeutic efficacy.

Therapeutic proteins are commonly synthesized in mammalian cell cultures. Chinese Hamster Ovary (CHO) cells are among the most used organisms for monoclonal antibody production [84]. CHO cells are widely used due to their ability to adapt to different culture media, for achieving higher titer than other mammalian cell lines and for being a bio-safe host cell, e.g. less prone to virus infection [77, 83, 84, 81]. Furthermore, CHO cells achieve high protein productivity due to their gene amplification qualities [114], and are able to provide efficient human-like post translational modifications such as glycosylation [179]. Aiming for cost-effective process production, technological solutions in cell culture systems have been developed to improve cell growth and specific productivity such as feed medium strategies and optimization of medium formulation [65]. The optimization results investigated in this thesis focused on media formulation.

The medium used in cell culture process has a significant effect on cell proliferation and cellular functions thus impacting antibody titer and growth [175, 59, 67]. Furthermore,

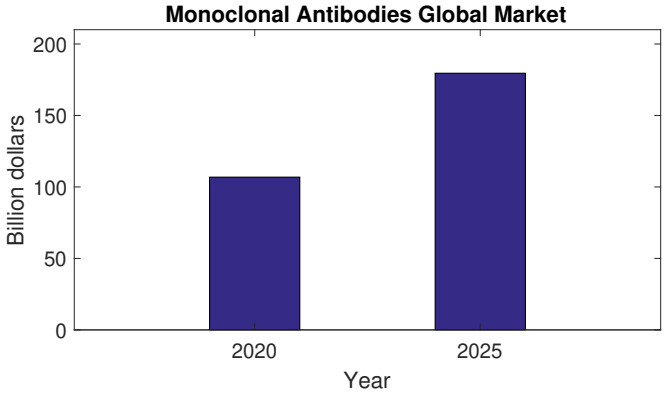


Figure 1.1: The global market for monoclonal antibodies [35].

cell glycosylation and protein aggregation can also be influenced by medium composition [70, 135]. Specifically, a cell culture medium should contain all the nutrients that cells need to grow and reproduce, such as amino acids, inorganic salts, metals, vitamins, lipids and hormones. Because small variations in some of these components, or, a interaction between two or more components can affect culture performance [67], finding the appropriate medium is a challenging, time-consuming and costly process [59].

A cell culture system under different modes of operation, e.g. batch and perfusion, is a dynamic process in which different nutrients are required at different phases of the cellular growth [123]. Moreover, it is known that the lack or excess of certain components in the cell environment can cause accumulation of harmful compounds or even cell death. Thus, to meet the cells nutrients demand at all times during the culturing process, medium optimization serves as an important tool to increase cellular productivity [171] [67].

Experimental analysis can be used to quantify the effect of some medium components. Although several analytical techniques are available to measure amino acids and other metabolites, e.g. chromatography and spectrometry, these procedures are expensive and time-consuming [123]. Thus, it is crucial to limit the number of experiments that are required to do modelling and optimization. Furthermore, due to the diverse interactions that exist between the components in the medium, experimental identification of the effects of components on cellular performance is a arduous task. Alternatively, metabolic modelling has become a useful approach to gain theoretical understanding of cell nutrients' demand and their effect on growth and productivity [171, 49]. Although metabolic models have not been thoroughly investigated for mammalian cells they have been successfully used for bacterial cultures. Hence, this modelling approach has potential for improving understanding

of mammalian cell cultures and for guiding the optimization of cell culture medium [131].

The traditional mathematical modelling approach for cell cultures consisted of formulating a set of dynamic mass balances for the most important metabolites where these balances may involve different degree of detail on kinetic reactions. In the more detailed approaches each possible reaction among metabolites is described by a kinetic expression of Michaelis-Menten type or other. However, due to the complexity of the mammalian cell metabolic network, this approach results in models with large number of equations and many parameters. These over-parameterized models are generally difficult to calibrate when data is relatively scarce and noisy and are prone to over-fitting. Constraints' based models, in which quasi-steady state conditions are assumed for intracellular fluxes, have been proposed in the literature to describe the evolution of metabolites concentration in a culture system. Generally, the key potential advantage of these constraint-based models is that they do not require extensive kinetic information as compared to the more traditional models mentioned above [49]. For example, Metabolic Flux Analysis (MFA) and Flux Balance Analysis (FBA) are constraint based models that have been used to understand cell behavior at steady state conditions including the estimation of nutrients demands, and prediction of cellular growth and productivity [171].

To account for the dynamic behavior of cellular systems, dynamic metabolic flux approaches, such as Dynamic Flux Balance Analysis (DFBA) [97, 76, 112, 113], have been developed in the last years. These models follow the idea of an general optimization problem based on FBA or MFA formulation, as given by equation 1.1. Therefore, for each time interval, the following optimization problem is solved:

$$\begin{aligned} &\max_{v} \text{ or } \min \quad \text{biological function} \\ &\text{subject to} \quad \text{set of constraints} \end{aligned} \tag{1.1}$$

where v is the specific metabolic flux. The biological function and the set of constraints are both function of the specific metabolic flux. This optimization problem considers that through natural evolution cells have the ability to optimally allocate resources so as to maximize or minimize a biological objective that is most beneficial for their cellular activities. This optimization is performed subject to stoichiometric and kinetic (reaction rate) constraints. Generally, constraints in few amino acids are sufficient to constrain their behaviour whereas metabolites that are not rate limited follow the rate limited metabolites according to their relative stoichiometric relations. Consequently, this type of model requires a smaller number of model parameters as compared to other modelling approaches and thus are less prone to over-fitting.

On the other hand, only a subset of medium components are usually described in

metabolic models, such as macro nutrients and amino acids. Trace elements, such as inorganic salts, vitamins, fatty acids, and metals, are frequently not present in metabolic models due to a lack of complete information about their influence in cell metabolism combined with the scarcity of data available for model training. Although these minor components are found in very low concentration, it has been reported that they can have a significant influence in cell behavior [23, 140]. For instance, vitamins function as cofactors for some enzymes, and the lack of them can cause cell death, reduction of cellular growth or a decrease in productivity [23]. A minimal variation in the concentration of other trace elements can even affect glycosylation patterns [140]. Therefore, the investigation on how these minor components and how the interaction between them can affect yield and protein quality have been the focus of some current studies.

Empirical modelling techniques have also been proposed for media composition optimization. For instance, optimal design of experiments (DoE) in combination with regression analysis can serve to identify important components and components interactions in the medium. Statistical tools may be particularly practical for understanding the influence of trace element components for which their exact metabolic roles is not clear. Nevertheless, media optimization based only on statistical analysis may still require a large number of experiments and outcome values, which is not economically viable. Hence, it was hypothesized in this thesis that a hybrid model that combines mechanistic and empirical expressions can be suitable for capturing the effect of cell culture medium components on cell culture performance.

Hybrid models started to be used in the biopharmaceutical field in recent years to address the limitations of either purely mechanistic or purely empirical modelling approaches [150, 159, 118]. While mechanistic models describes the bioprocesses with fundamental laws of mass balance, stoichiometric and kinetic rate constraints, data-driven models extract information given by experimental data that cannot be explained by the mechanistic equations. Hybrid models have been used to incorporate information about metabolic fluxes [150] or effects of bioreactor operation conditions [118]. To our knowledge, hybrid models have not been used yet with the aim to incorporate information regarding the variability of media formulations on cellular performance, including the variations of minor components (such as vitamins, hormones, metals, etc) and effects from possible interactions among all medium formulation components.

Following the above, this work presents novel contributions for empirical, mechanistic and hybrid models for mammalian cell cultures. The thesis is organized as follows. The second chapter presents a literature review about the topics discussed in this thesis. Chapter 3 presents an expansion of the Dynamic Metabolic Flux Model (DMFM) for mammalian cells developed by Nikdel [114] by i- incorporating the prediction of biomass into

the model; ii- incorporating kinetic expressions as function of metabolites concentration for limiting components identified by the model; and iii- expanding the model for perfusion operation. Chapter 4 presents the development of new media formulations based on empirical PCR and PLS regression models that accounts for medium components interactions and minor components effects. This chapter also presents the use of D-optimal Design of Experiments (DoE) method to identify the most informative experiments about culture productivity thus reducing the number of experiments needed to develop of new media formulations. Chapter 5 presents a hybrid dynamic model that combines a mechanistic model given by a set of dynamic balances of main metabolites and reaction rate expressions with an empirical model given by PLS regression where the latter accounts for all medium components interactions. Finally, Chapter 6 presents the conclusions and future work.

Chapter 2

Background and Literature Review

2.1 Cell Culture and Media Optimization

Mammalian cells are cultivated under specific operating conditions and in a medium solution in which all components required for their metabolic activities are present [161]. A medium formulation can directly influence cell culture productivity [112]. In order to ensure that cells remain healthy and exhibit high growth, an optimal range of medium nutrients concentration need to be reached [135].

Several types of media have been developed since the inception of cell culture technologies. The first media generation used to be derived from natural animal substances, such as serum. Animal serum is a great source of proteins, growth factors and hormones [175]. However, due to variations in composition for different supplied serum, and due to the risk of microbiological contamination (such as by bacteria, yeast or virus) that may promptly cause the death of cells, there is a general drive in the industry to replace serum containing media by synthetic serum-free medium. Such medium is composed of a basal medium, which contains all compounds important for cells survival, and it is also supplemented with specific components according to the needs of the cell line being cultivated [161]. In general chemically defined media, has among its advantages a known specific composition that results in reproducible cell culture outcomes with smaller risk of contamination [175]. With the continuously acquired knowledge of cell growth requirements it became clear that media optimization is crucial for cell culture based manufacturing processes.

It is widely recognized that to ensure cell culture health and growth elements such as hormones, growth factors, protease inhibitors, amino acids, vitamins, lipids, proteins,

trace elements, glutamine, glucose and antibiotics are required [161]. Each one of these components has a direct crucial role in the cells metabolism but also they can indirectly affect the evolution of the culture by changing the cell culture environment. For instance, amino acids levels may affect pH and osmolality of culture solutions while other nutrients can affect mAb quality, glycosylation and protein aggregation [136].

Reinhart *et al.* [130] compared eight different chemically defined media for CHO (Chinese Hamster Ovary) cell culture and confirmed that media composition and cellular adaptation might be the main reasons for cell culture performance results. Furthermore, the authors shown that higher amino acids levels in culture media are not necessarily associated with higher cellular growth and antibody production, but the presence of particular amino acids is essential for cell metabolism. Reinhart *et al.* [130] also confirmed that high glucose level could affect glycolysis rate and lactate formation.

Kishishita *et al.* [79] analyzed amino acids supplementation and their influence on monoclonal antibody production by comparing a medium before and after the CHO cell culture. The authors concluded that amino acids supplements such as serine, cysteine and tyrosine increase cell culture performance by raising monoclonal antibody production. Besides, Kishishita *et al.* [79] found that by-products formation can be reduced by controlling amino acids supplementation. Lactate and ammonia are examples of by-products that if present in large concentrations they significantly affect productivity.

The concentration levels of key media nutrients, such as glucose and glutamine, that affect cells' glycosylation process was also investigated by Aghamohseni *et al.* [5] and Liu *et al.* [90]. Because glycoproteins are able to influence mAb activities, these structures are important features to evaluate the quality of monoclonal antibodies [87]. Liu *et al.* [90] confirmed that glycan structures production are directly correlated with glucose concentration present in the medium. Aghamohseni *et al.* [5] used a combination of data and mathematical modelling based on metabolic flux analysis to study the effect of glucose, glutamine and ammonia on glycosylation of mAb production and cell growth. These authors found that higher initial glutamine concentration, consequently increases ammonia level reducing galactosylation and sialylation indexes.

Sun *et al.* [149] investigated feeding strategies using glucose and galactose focused on lowering the production of lactate in order to raise cell growth and mAb production in a fed-batch reactor. Experimental results showed that the culture performance is related to the occurrence of a metabolic shift of lactate which at some point during the culture stops to be produced and starts to be consumed. The authors also associated an energy metabolism change to the metabolic shift in lactate production/consumption. The optimized fed medium with glucose and galactose did not present accumulation or insuffi-

ciency of amino acids in their cell culture; and, although lactose was being consumed, the ammonia concentration increased.

Through an analytical experimental approach involving experiments in 96-well plates, shake flasks and bioreactors, Bai *et al.* [11] explored the effect of iron and sodium citrate on protein-free media for CHO cell culture. Iron is known as an essential component to keep cells healthy and proliferate; yet, a non-optimal iron concentration is also toxic to cells. Citrate, on the other hand, is involved in the TCA cycle, and consequently with protein biosynthesis [11]. The authors reported that a combination of iron and sodium citrate are essential components for cellular growth and antibody yield where their optimal concentrations vary within ranges of 0.1–0.5 mM and 0.125–1 mM, respectively. Furthermore, the authors concluded that citrate may be responsible for a mAb productivity increase of 30 - 40%.

Aiming the optimization of media nutrients levels for a CHO cell culture, Rouiller *et al.* [135] developed a novel approach for media blending design using a high-throughput method. The approach included the use of empirical models, multivariate analysis and design of experiments techniques. The authors concluded that the new blended media obtained with their approach was able to improve by 20% cell growth and by 40% titer. Furthermore, they identified the medium components ferric ammonium citrate, panthothenic acid, valine, methionine, arginine, biotin and serine as the ones that most affect the mAb titer concentration in CHO-S cell line culture.

The importance of several other supplements, minor components such as non-amino acids, in the cell culture media is discussed in Yao & Asayama [175]. In general, metabolites that are present in the medium in small amounts are optimized using experimental analysis. Often, the minor components are not optimized simultaneously but they are rather optimized in smaller subsets, e.g. subset of vitamins, subset of metals, etc [11, 52, 169, 137]. Mathematical modelling and statistical tools have been frequently used in industrial practice for medium optimization [135, 114, 143, 105, 158, 49, 68]. However most of the mathematical models reported in literature only describe amino acids and other major byproducts such as lactate and ammonia. To the author's knowledge, mathematical modelling accounting simultaneously for the influence of all medium components, including amino acids, major and minor components, has not been reported as yet.

In summary, different approaches have been reported for media optimization such as analytical experiments, design of experiments, statistic tools and metabolic mathematical models. To satisfy the constant need for reducing process costs and culturing times while improving quality production and performance of cell culture, further research is still required to identify the effect of different media components on cell behavior.

2.2 Statistical Tools for Metabolic Data

Several statistical techniques are available for interpreting and understanding correlations among data. Multivariate statistical tools, such as Principal Components Analysis (PCA) and Partial Least Squares (PLS) regression have been used to describe correlations among the input data (PCA) or between input and output data (PLS). On the other hand optimal Design of Experiments (DoE) based on statistical premises is an additional tool that has been used to identify particular experimental conditions that are most informative about the effect of medium components on cell culture products. A brief description of the statistical methods used to interpret cell culture data is presented bellow:

i. **Principal Component Analysis (PCA)**: Compresses a larger set of variables into a smaller set of new variables, referred to as 'principal components' (PC), which are linear combinations of the original variables. The PC also account for most of the variance of the original large set of variables.

ii. **Principal Component Regression (PCR)**: This method is a linear regression model based on Principal Component Analysis (PCA). It models a response variable using a new set of predictor variables, the 'principal components' (PC), which explain the observed variability in the predictor variables, without considering the response variable.

iii. **Partial Least Squared (PLS)**: Finds the linear relationship between a large set of independent predictor variables and a set of dependent response variables. The PLS approach identifies multidimensional directions in the input space that explains most of the variance of the output space.

iv. **Least Absolute Shrinkage Selector Operator (LASSO)**: Selects regression parameters through a shrinkage process based on an optimization cost that involves a regularization term. LASSO identifies the most important independent predictor variables related to the dependent response variables that minimize the overall prediction error. LASSO can help to interpret the resulting regression model because insignificant variables have their parameters reduced to zero. A tuning parameter (λ) is used to control the strength of the penalty weight on the regularization term. As λ increases the bias increases and the variance decreases.

v. **Elastic Net**: A regularized regression method that can be described as an extension of LASSO regression. The elastic net approach presents the same LASSO optimization problem with an additional quadratic term into the penalty weight.

vi. **Hierarchical Cluster Analysis (HCA)**: Identifies similarities among the available data and groups them into clusters. Then, a hierarchy is created among the identified clusters based on an "agglomerative" and "divisive" strategies. The HCA results can be visualized in heat maps, dendrograms or cluster trees.

Since the uses of the above algorithms are pervasive in process systems the following review focuses only on application related to cell culture. Selvarasu *et al.* [143] used PCA and PLS to analyze results of a CHO fed-batch cell culture conducted with protein free chemically defined medium. The authors were interested in finding the nutrient's components that were most correlated to lactate and ammonia secretion since the latter are responsible for inhibiting cell growth. As expected, the cell culture results indicated that glucose and glutamine depletion are highly correlated with lactate and ammonia, respectively. Yet, a high correlation between asparagine and ammonia responses were also found. PCA was also used by Hong *et al.* [66] to understand the influence of amino acids concentrations from different cell culture media on the lactate production/consumption shift.

Recently, Moris *et al.* [107] used PLS regression for identification of amino acids that most affect titer production in CHO cells. The resulting PLS model was used to propose new feeding strategies for fed-batch reactors. Eyster *et al.* [165] also used a PLS model for predicting the evolution of glucose and lactate with the aim to develop a nutrient control strategy for a fed-batch reactor used for mAb production in CHO cell culture.

Templeton *et al.* [153] used hierarchical clustering analysis (HCA) and PLS in order to evaluate important metabolic fluxes in a C-MFA model of CHO cells. They identified that the IgG specific productivity was associated to the lactate dehydrogenase (LDH) flux, concluding that higher productivity is achieved when lactate consumption increases. Using PLS, the authors also identified that glutamine has a strong negative correlation with recombinant monoclonal antibody production.

The Least Absolute Shrinkage and Selection Operator (LASSO) have been used in the literature for modelling processes and for identifying variables that affect the predictions of product quality. LASSO is an original variable selection method that outcomes with a smaller regression model when its selection operator turn irrelevant coefficients to zero [174]. In this way, LASSO regression provide a better model accuracy by only selecting meaningful input variables [10]. For example, Yan *et al.* [174] shown the effectiveness of their proposed LASSO approach for an injection molding process case study, showing that LASSO method could predict with more accuracy, e.g. with a smaller root mean squared error, the batch operation results if compared to multi-way partial least squares (MPLS) and PLS approaches.

Badsha *et al.*[10] used Elastic net regression, an extension of LASSO method, to elucidate significant metabolites that influence, for instance, cell growth, glucose and lactate rates in CHO-K1 cells operating in a batch culture system. The analysis identify correlations between cell growth and the metabolites ribose-5-phosphate, lactic acid, arginine, glucose and ammonium ion. Factors UDP-glucose, S-adenosylmethionine, arginine, and glutathione were found important for glucose and lactate uptake rate.

It is common in metabolic engineering that the number of factors being analyzed in a system is much larger than the number of samples/observations available. Hence, the least squares approach is not viable due to the under-determinacy of the problem. Instead, other statistical techniques that uses regularization (such as LASSO) or compressed (such as PCA and PLS) process are an acceptable approach to describe the system with a parsimonious linear model where over fitting of noise is avoided.

The concept of sparse and compressed linear regression models has also been used to perform an optimal DoE that is limited to specific data regions of interest. The main point of this strategy is to select a minimal number of input variables, from all the variables of the system, that helps to estimate model parameters and helps to design experiments without affecting the predicted statistical variability of the parameters of the model [126]. This coupling of DoE and statistical regression techniques is presented in the current work.

2.2.1 Design of Experiments

Biological systems are dynamic processes that must be modeled by a combination of differential and algebraic equations. Once the dynamic model is formulated, experiments need to be carried out in order to estimate the parameters of the model. Cell culture and media analysis experiments are expensive and time consuming. For these reasons, when modelling biological systems, academics and industrial practitioners seek for a minimum number of experiments that will be able generate the most informative data for model calibration [50, 151].

In the DoE approach, variables input are exploited with the aim to select experiment samples that will provide significant information about the model and its intended application [93]. Since one of the key practical endeavors goals is model based optimization, the metabolic models are focused on understanding which inputs give the best cellular performance. However, in many cases it is desired to calibrate models that describe the system in a wide region of operating conditions corresponding to either good or bad cellular performance. In this case, experimental samples should generate relevant information about conditions and metabolites concentration that most affect cells behavior.

A Model-Based Design of Experiments (MBDoE) is an useful technique to choose experiments that will generate helpful information for the development of dynamic deterministic models [17, 50]. In this method, an optimal design problem is solved by optimizing a scalar measure of the Fisher information matrix or variance-covariance matrix that reduces the uncertainty in the model parameters, e.g. as given by their confidence intervals [50, 151, 152]. Since MBDoE is formulated as a nonlinear optimization problem it generally require initial model parameters' values to start the optimization search. Data that can be used for initial guesses about the parameters can be found in literature or acquired from previous DoE results [93]. After the hypothesized model is fitted with the initial parameters' values, optimal experiments can be designed by a sequential iterative process of model fitting and parameterization to find the final parameter values that result in good fitting between predictions and data [17, 50].

The final goal of an optimal design of experiments is to utilize a minimum number of experiments to obtain the maximum experimental data information (or variability) [151, 93]. Specific design criteria that shapes the confidence region of the model parameters, are used to solve optimal design problems [93, 108]. The current optimal DoE methods are referred by the corresponding design criteria used: A-design, D-design, E-design, Modified-E-design and V-design [108]. By using these design criteria, optimal experiments can be chosen that will generate the highest information data for model parameter estimation [111]. If the calibrated model by one of these methods is subsequently used for optimization, improving the model fitting with fewer experiments will lead to the improvement of productivity in biological systems with a low cost of experimentation [93].

Design of Experiments can be run sequentially, in parallel or even in a combination of both modes. Usually, dynamic models are calibrated with sequential DoE based approaches [93]. In a sequential design of experiments data information is collected from a wide range of conditions throughout the design space. Then, a region of interest can be identified based on the experiments and further application of the optimization design criteria problem is used to plan additional following experiments [148]. Figure 2.1 presents a flowchart in which the sequential design of experiment steps are followed for a maximum of N experiments.

Cockshott & Sulivan [32] used a combination of both sequential statistical design and analysis methods to optimize fermentation media for *Aspergillus nidulans*. The sequential statistical experiment design techniques, Plackett–Burman, factorial, response surface and ridge analysis, were used to screen the effect of fifteen media nutrients. The sequential experimental procedure identified a group of five significant media components which were used to obtain a final optimal media formulation that increased title by 46%.

Several other works in literature combine design of experiments with modelling ap-

proaches in order to optimize the process conditions of interest [93, 152, 148]. However, to the knowledge of the author, the use of regression models and design of experiments to understand the effect of interactions among all medium components and for optimizing complex media have not been reported before.

Media used for mammalian cell culture are composed of a large number of components (∼100). These components can also present interaction effects on the productivity results. The use of novel mathematical modelling in bioprocesses that involve many parameters, e.g. dynamic metabolic flux analysis, has also generated the need for more informative design of experiments [80]. Efficient design of experiments for model calibration can promote the use of more sophisticated dynamic modelling technique and provide a better understanding on the influence of specific media components on cell performance [17].

Chapter 4 section 4.3 presents complementary literature review on medium design and mathematical approaches to the "$n < p$ regression problem", e.g., cases where the number of samples are much smaller than the number of variables.

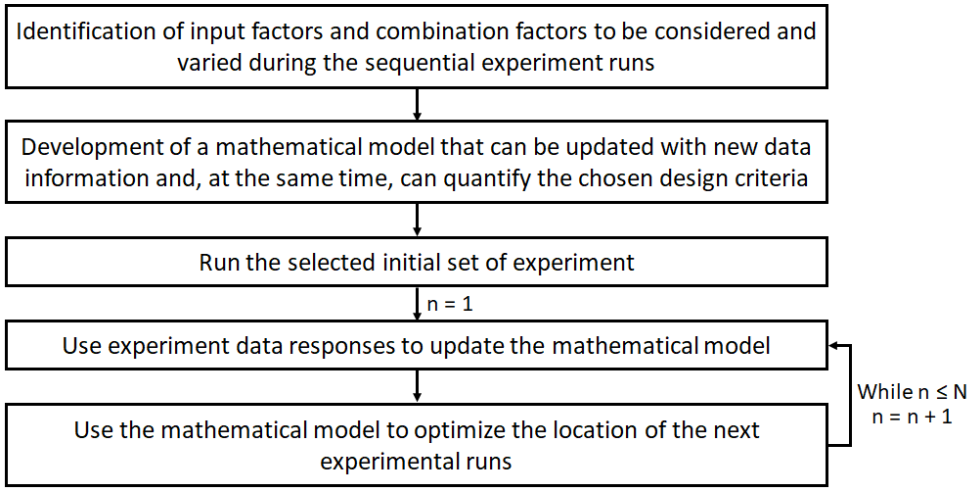


Figure 2.1: Sequential design of experiments steps. Based on Soepyan *et al.* [148]

2.3 Mathematical Metabolic Modelling

Mathematical modelling has among its advantages the ability to test hypotheses, theories and concepts and to simulate and predict different process conditions [14]. Also, models can

be used for optimizing operating conditions thus requesting a lower number of experiments as compared to empirical optimization approaches. Furthermore, mathematical models based on first principles and knowledge about the cellular metabolism can serve to identify biological mechanisms that may not be obvious from either experimental data or statistical models.

The complex cellular metabolism can be described by a large network system in which nutrients are transformed into essential molecules necessary for cells' growth. The metabolic network can also serve to describe reactions that generate or consume energy needed for cell proliferation and maintenance [96, 158]. Based on the metabolic network, mass balances can be formulated for both intracellular and extracellular metabolites using stoichiometric proportions to describe steady state conditions or suitable kinetic expressions to describe dynamic situations [14].

Metabolic Flux Analysis (MFA) is an important mathematical approach used to quantitatively estimate metabolic fluxes [158]. Since metabolic networks may involve many reactions for intracellular and extracellular metabolites some simplifications are generally done. For example, assuming a steady state the mass balances can be expressed as a function of the metabolic fluxes as per the following equation 2.1:

$$\mathbf{Sv} = \mathbf{0} \tag{2.1}$$

where $\mathbf{S} \in \mathbb{R}^{r \times q}$ is the stoichiometric matrix, $\mathbf{v} \in \mathbb{R}^{q \times 1}$ is the vector of fluxes for each metabolic flux $\mathbf{v} = (v_1, ..., v_q)$, r and q are the number of metabolites and the number of fluxes, respectively, involved in the metabolic pathway.

MFA has been used to identify main metabolic fluxes of CHO cells in order to reduce the number of mass balances necessary to describe dynamic scenarios [105]. For example, Naderi *et al.* [110] and Hille [64] employed MFA to identify significant fluxes. Assuming fluxes that were less than 1% of the total flux sum could neglected, the authors identified significant macro-reactions and used them as a basis for the formulation of dynamic balances of metabolites using kinetic terms such as Michaelis–Menten.

Due to the large number of reactions and the fact that each metabolite may participate in more than one reaction, the MFA may result in underdetermined or overdetermined system of algebraic equations thus requiring the addition of constraints to obtain unique solutions. Recently, Kastelic *et al.*[75] and Erklavec Zajec *et al.* [46] used a constrained metabolic flux approach to describe a metabolic network of over 100 reactions. The approach was used to identify elementary modes that were used for developing dynamic kinetic equations. These dynamic equations were then used to study the effect of different operating conditions for CHO cell culture operated in fed-batch mode.

Thereby, a constrained version of MFA, often referred as Flux Balance Analysis (FBA), is used where a cellular biological function is optimized subject to constraints. Limits on metabolic rates, stoichiometric relations as well as thermodynamic information may be used as constraints [49]. Lee *et al.* [85] used FBA to address a problem that was underdetermined since the number of metabolites was smaller than the number of reaction fluxes. An FBA formulation can be written according to equation 2.2:

$$\begin{aligned} \min_{\mathbf{v}} \quad & \mathbf{c}^T\mathbf{v} \\ \text{subject to} \quad & \mathbf{S}\mathbf{v} = \mathbf{0} \\ & \mathbf{v}^L \leq \mathbf{v} \leq \mathbf{v}^U. \end{aligned} \tag{2.2}$$

where $\mathbf{c}^T\mathbf{v} \in \mathbb{R}^{1\times 1}$ is a biological objective function to be optimized (such as growth rate), $\mathbf{v}^L$ and $\mathbf{v}^U \in \mathbb{R}^{q\times 1}$ are, respectively, the vectors for lower and upper bounds of the specific metabolic flux $\mathbf{v}$. The rationale for this optimization is that cells have acquired through natural evolution the ability to optimally allocate resources so as to maximize or minimize a biological objective that is most beneficial for their proliferation.

FBA solves the metabolic flux systems in steady state conditions. On the other hand, cell cultures are dynamic in nature. The extension of FBA to describe dynamic cell culture processes was first proposed by of Savinell & Palsson [141]. The key idea by these authors is that the FBA static model shown in equation 2.2 for steady state, can be instead solved at each time interval of a dynamic operation, e.g. batch or fed-batch, and the fluxes calculated at each time can be used for calculating the changes in metabolites' concentrations using a numerical integration approach. Mahadevan *et al.* [97] proposed a Dynamic Metabolic Flux Analysis (DFBA) to analyze the diauxic growth in *E. coli* cells. Assuming the cells are able to maximize the resources that are provided to them, the DFBA incorporates the rate-of-change constraints on metabolic fluxes.

Following what was presented above, Nikdel and Budman [112] and Nikdel *et al.* [113] developed a systematic 3-step approach (described in Chapter 3) to identify a Dynamic Metabolic Flux Model (DMFM) that was successfully applied to bacteria and other microorganisms. The DMFM typically requires a smaller number of parameters as compared to models that require modelling of each reaction in the metabolic network. A DMFM can

be formulated as in equation 2.3, for each sampling interval k ($k = 1, ..., t_f$).

$$\begin{aligned} \max_{\mathbf{v}_k} \quad & \mathbf{c}^{\mathrm{T}}\mathbf{v}_k \\ \text{s.t.} \quad & f(\boldsymbol{\psi}_k) \leq \mathbf{S}\mathbf{v}_k \leq g(\boldsymbol{\psi}_k) \\ & \mathbf{v}_k \geq \mathbf{0} \\ & \boldsymbol{\psi}_{k+1} = \boldsymbol{\psi}_k + \mathbf{S}\mathbf{v}_k X_{v,k} \Delta k \geq \mathbf{0} \end{aligned} \tag{2.3}$$

where k is the sampling time, t_f is the final sampling time, $\mathbf{v}_k = (v_1, ..., v_q)_k \in \mathbb{R}^{q\times 1}$ is the vector of fluxes for each q metabolic flux at time interval k, $\boldsymbol{\psi} \in \mathbb{R}^{r\times 1}$ is the vector of concentration for each metabolite r, $X_v \in \mathbb{R}^{1\times 1}$ is the biomass concentration, $f(\boldsymbol{\psi}) \in \mathbb{R}^{r\times 1}$ and $g(\boldsymbol{\psi}) \in \mathbb{R}^{r\times 1}$ are lower and upper kinetic reaction rates vector defined as a function of the metabolite's concentration participating in the corresponding reaction. These rate constraints can be represented by Michaelis-Menten expressions or from interpolation of look-up tables as done in the current study [112]. The last equality in equation 2.3 represents an Euler integration that serves to update the metabolites' concentrations with time as a function of the fluxes $\mathbf{v}$ calculated at each time interval.

The systematic 3-step approach to identify DMFM was recently studied in another work of Nikdel [114] for mammalian CHO cells in a batch system. However, in this model biomass was not being predicted by the DMFM. Also, the DMFM model by Nikdel [114] is applied only to batch reactor process. To the knowledge of the authors, no such systematic identification of a DMFM, including biomass prediction, has been conducted for mammalian CHO cells.

One reason for this lack of mammalian cell descriptions is that their general regulatory behavior is highly complex. The regulation of metabolism in mammalian cells is closely related to several processes such as different forms of death [105], and varying energy requirements. For example, while bacteria growth is mostly limited by availability of nutrients, mammalian cells' growth is also limited by apoptosis (programmed cell death) or by exposure to high concentrations of by-products (ammonia, lactate). In the literature, few metabolic analysis incorporate these process into the models. For mammalian cells, there is not yet a dynamic metabolic flux model which incorporates apoptosis and/or describes a perfusion system as intended in the current work.

2.4 Hybrid Models

Metabolic dynamic models that are used to describe cellular systems requires mechanistic knowledge about the metabolic network system and extensive experimental information, especially about kinetic parameters involved in bioreactions' rate terms. One of the limitations of such models are the lack of information about kinetic metabolic parameters involved in (complex) cellular systems and the large amount of data required to calibrate these type of models [36]. To overcome these challenges, hybrid models have been proposed as an alternative description of the metabolic behavior of cells [38].

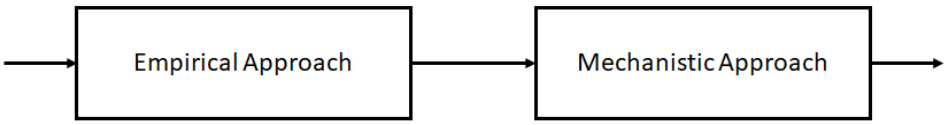
Hybrid models combine mechanistic and empirical information into an overall model. Mechanistic approaches, or first principles' models, take into account the prior knowledge about the process through fundamental conservation laws. The key advantage of mechanistic models is their superior extrapolation capabilities beyond the operating conditions used for model calibration. Still, first principles' models may not be able to predict accurately due to incomplete knowledge about all the physical/biological phenomena involved in the process and thus it may not be able to fit the experimental data [43].

On the other side, empirical models, or data-driven models, can be adjusted to fit the experimental data by minimizing some norm of the error between model and data [43]. Empirical approaches can be based on regression analysis, artificial neural networks, Fourier series, smoothing splines, and other similar correlations. However, despite their ability to fit the data, empirical models have limited extrapolation capability and it may provide non-physical results outside the region of calibration since it does not account for physical/biological constraints [43]. Also, empirical models can bring information about interactions and patterns observed by the data that are not described in mechanistic models [147].

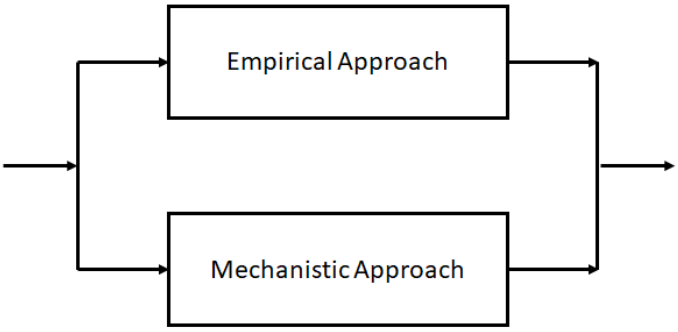
In a hybrid model, mechanistic and empirical models are combined together in either serial or parallel structures. Hybrid models that follow a serial arrangement use mechanistic and empirical models one after the other [54]. Then, the final prediction is obtained from the sum of the predicted output of the first principles model plus the error where the latter is calculated with the empirical model. Figure 2.2a presents the flowchart of a hybrid model in a serial mode [43]. It should be noticed that in this arrangement the input variables is only used in the empirical model and is not used directly in the mechanistic model.

On the other hand for hybrid models of parallel structure the input variables is fed simultaneously to both the mechanistic and empirical models where the mechanistic part is based on the available physical/biological knowledge about the process while the empirical component is used to capture the behavior that is not captured by the mechanistic

component. Figure 2.2b shows the hybrid model with parallel structure [43].



(a) Serial arrangement.



(b) Parallel arrangement.

Figure 2.2: Hybrid model dispositions. Adapted from Duarte *et al.* [43]

Azevedo *et al.* [38] developed a hybrid formulation that combines the mass balance equations (mechanistic component) and an artificial neural network (empirical component) to predict biomass formation in the baker's yeast. In this case the artificial neural network used the available data measurements to complete the information about the kinetic reaction and the latter it was further fed into the set of mass balance equations. Following a dynamic serial hybrid model structure, the output from the mechanistic equations was used to train the neural network by minimizing the sum of square errors between the mechanistic model predictions and the data. The authors concluded that the hybrid model was able to predict and explain experimental data with good accuracy. Furthermore, the authors emphasize the importance of hybrid approaches in the biochemical engineering field, in which a lack of metabolic information and data are key challenges for developing accurate models.

Teixeira *et al.* [150] include information from an Elementary Mode analysis into a hybrid model for the identification of prevalent pathways in recombinant Baby Hamster Kidney (BHK-21A) cell cultures to improve production of recombinant fusion of glycoprotein.

The authors incorporated into the Elementary Mode kinetics the product of two functions, in which the first one is defined by a mechanistic/empirical function and the second is a nonparametric (purely empirical) function that was calibrated with data. Thus, this parallel hybrid model integrated the knowledge from metabolic reactions with intracellular kinetics. It was shown by Texeira *et al.*. [150] that optimal feeding strategies based on the hybrid model resulted in higher final product concentration.

The work of Costa *et al.* [36] investigated a hybrid formulation that could reduce the complexity of models for large metabolic network. The authors studied four hybrid models, in which the mechanistic approach was described by the Michaelis-Menten kinetics for one specific substrate, and four different empirical kinetic structures (generalized mass action, convenience kinetics, power-law, and lin-log) were applied for the other reactions. The formulated hybrid models results were compared with a model that was developed solely based on mechanistic equations. The main idea of the authors was to use simplified models to describe large metabolic networks for which the precise kinetic rate for part of the metabolites are not known. Costa *et al.* [36] concluded that the hybrid model coupled to the lin-log based empirical component successfully predicted the results of *E. coli* metabolism.

O'Brien *et al.* [118] presented a hybrid model to study the bioprocess systems of an in silico mammalian cell culture. Their model combined mechanistic models for central metabolism, cell signaling, cell growth, and the reactor environment and used a large manufacturing data set which presented a great variability to simultaneously calibrate all model parameters. The final model was used to optimize the reactor process conditions and to predict cellular performance.

Recently, Ghosh *et al.* [54] presented a parallel hybrid model that combined a mechanistic model with PLS regression to describe a batch poly(methyl methacrylate) (PMMA) polymerization reactor. The results of their work have shown improved modeling predictions with the hybrid model compared with mechanistic or data-driven model alone. However,it is important to notice that, in contrast with the current work, the model was simpler in terms of the number of variables (only 2 input variables were considered) and the hybrid model parameters were not re-calibrated.

Hybrid formulations can be developed in order to make use of the benefits of both approaches, using in this way all available information and knowledge from the system being studied [38]. In the current study we are proposing a parallel hybrid model that combine a mechanistic model given by dynamic balances of metabolites using kinetic terms such as Michaelis–Menten with an empirical regression based on Partial Least Squared (PLS) regression. The proposed model, which is presented in Chapter 5 is considered a

parallel structure since a subset of the input variables is used simultaneously by both the dynamic mechanistic equations and by the PLS regression. Also, parameter re-calibration was performed for the hybrid model. To the knowledge of the author, the proposed hybrid approach was never presented in the literature to study media optimization with respect to major and minor components and their nonlinear interactions.

Chapter 3

Identification of a Dynamic Metabolic Flux Model for a Mammalian Cell Culture

This chapter is based on the author's work presented at the XXIX Interamerican Congress of Chemical Engineering held jointly with the 68th Canadian Chemical Engineering Conference (CsChe 2018) [27] and in the author's paper presented at the 12th IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems (DYCOPS 2019) [28]. This work follows a 3-step approach to identify DMFM for mammalian CHO cells recently studied in thesis of Nikdel [114]. In the work of Nikdel [114] CHO cells were studied in a batch system and biomass was not being predicted by the DMFM. The contribution of this current work focuses on expanding the DMFM initially developed by Nikdel [114] to account for biomass prediction, apoptotic (cell death) behavior and perfusion operation.

3.1 Introduction

Therapeutic proteins, such as monoclonal antibodies (mAbs), are commonly synthesized in mammalian cells, in particular, Chinese Hamster Ovary (CHO) cells [84]. To optimize cell growth and specific productivity it is imperative to understand the metabolic behavior of mammalian cells cultivated in cell cultures and to predict the evolution of nutrients and by-products' concentrations with time. Metabolic modelling has become a useful approach to build predictive mass balance models that conform to the metabolic network and potentially identify targets for genetic engineered modifications [131, 112].

Biochemical models that involve mass balances where each possible metabolic reaction must be described by a kinetic rate concentration dependent expression involve many calibration parameters. Furthermore, lack of data for many of these metabolites often makes it difficult to calibrate these over-parameterized kinetic metabolic models. Constraint based models, in which quasi-steady-state conditions are assumed for intracellular fluxes, have been proposed in the literature to describe the evolution of metabolites' concentrations for cell culture systems. These constraint-based models typically require less model parameters as compared to models that require kinetic expressions to represent each possible reaction in the network [49]. Metabolic Flux Analysis (MFA) and Flux Balance Analysis (FBA) are examples of constraint based models that have been used to understand cell behavior and estimate nutrients demands at steady state [171].

Dynamic Metabolic Flux Modelling (DMFM), an extension of FBA, is a dynamic modelling approach [97, 98] where a specific biological function is optimized at each time interval with respect to a vector of fluxes subject to some constraints. The DMFM typically requires a smaller number of parameters as compared to models that require modelling of each reaction in the metabolic network. The inherent assumption in DMFM is that the cell is acting as an optimizing agent where an biological objective, e.g. growth rate, is minimized or maximized subject to few limiting kinetic constraints while most species are assumed to be correlated with each other through stoichiometric relations. A DMFM can be formulated for each sampling interval k ($k = 1, ..., t_f$) as follows:

$$\begin{aligned} \max_{\mathbf{v}_k} \quad & \mathbf{c}^{\mathrm{T}}\mathbf{v}_k \\ \text{s.t.} \quad & f(\boldsymbol{\psi}_k) \leq \mathbf{S}\mathbf{v}_k \leq g(\boldsymbol{\psi}_k) \\ & \mathbf{v}_k \geq \mathbf{0} \\ & \boldsymbol{\psi}_{k+1} = \boldsymbol{\psi}_k + \mathbf{S}\mathbf{v}_k X_{v,k}\Delta k \geq \mathbf{0} \end{aligned} \tag{3.1}$$

where k is the sampling time, t_f is the final sampling time, $\mathbf{c}^{\mathrm{T}}\mathbf{v}_k \in \mathbb{R}^{1\times 1}$ is a biological objective function to be optimized (such as growth rate, metabolic burden, etc), $\mathbf{v}_k = (v_1, ..., v_q)_k \in \mathbb{R}^{q\times 1}$ is the vector of fluxes for each q metabolic flux at time interval k, $\mathbf{c} \in \mathbb{R}^{q\times 1}$ is a vector of coefficients, $\boldsymbol{\psi} \in \mathbb{R}^{r\times 1}$ is the vector of concentration for each metabolite n, $\mathbf{S} \in \mathbb{R}^{r\times q}$ is the stoichiometric matrix, $X_v \in \mathbb{R}^{1\times 1}$ is the biomass concentration, $f(\boldsymbol{\psi}) \in \mathbb{R}^{r\times 1}$ and $g(\boldsymbol{\psi}) \in \mathbb{R}^{r\times 1}$ are lower and upper kinetic reaction rates vector defined as a function of the metabolite's concentration participating in the corresponding reaction. These rate constraints can be represented by standard kinetic expressions such as Michaelis-Menten or can be obtained from interpolation within look-up tables as done in past works [114] and [27]. The last equality in equation 3.1 represents an Euler numerical integration

operation that serves to update the metabolites' concentrations with time as a function of the fluxes $\mathbf{v}$ calculated at each time interval.

Following this modelling approach, the DMFM (equation 3.1) does not require the calibration of kinetic parameters for each possible reaction involved in the metabolic network, but only for a subset of them. Thus, only parameters of metabolites considered as relevant constraints in the functions $f(\psi)$ and $g(\psi)$ need to be estimated. The evolution with time of many metabolites that are not rate limited are calculated by mass balance stoichiometric relations. However, while the DMFM can potentially result in compact models, a key challenge for this approach is the identification of limiting constraints that best describe the data.

A systematic approach to identify DMFM has been recently proposed and successfully applied to bacteria and other microorganisms [112, 113]. This systematic approach was also used to identify a preliminary DMFM for mammalian CHO cells [114]. However, in this last work, biomass concentration which is crucial since it affects each of the mass balance equations was not being predicted by the model. Also, to the knowledge of the author, no such other systematic identification of a DMFM has been conducted for mammalian CHO (Chinese Hamster Ovary) cells. In general the lack of reported DMFM models for mammalian cells, such as CHO cells, is due to the fact that they exhibit a more complex behavior as compared to bacteria. For example, while bacteria growth is mostly limited by availability of nutrients, mammalian cells' growth is also limited by apoptosis (programmed cell death) or by exposure to high concentrations of toxic by-products (ammonia, lactate). In this current work, a DMFM for mammalian cells is systematically identified using methods proposed earlier by our group [112, 113, 114], and a new step is proposed for the extension of model to describe perfusion operation. Comparisons between model predictions and experiments are shown during the growing phase in a batch and perfusion operation. Furthermore, this work presents biomass prediction and incorporation of cell death into the optimization problem.

3.2 Steps of the Algorithm to Identify DMFM Constraints

In this section, we briefly review the approach presented by Nikdel and Budman [112] for identifying the limiting constraints of a DMFM model from data. The approach is based on the use of set based bounds of the experimental data. It is assumed that, due to measurement noise or other unmeasured disturbances, the metabolites' concentration

trajectories with time are bounded by convex sets defined by upper and lower values, at each sampling time. For instance, Fig. 3.1 shows set based bounds for glucose concentration consumption during a batch culture [114]. The use of these bounds combined with the fact that the problem defined in the following equations is an LP considerably simplifies the identification of limiting constraints as further explained below. The approach involves 3 steps as follows.

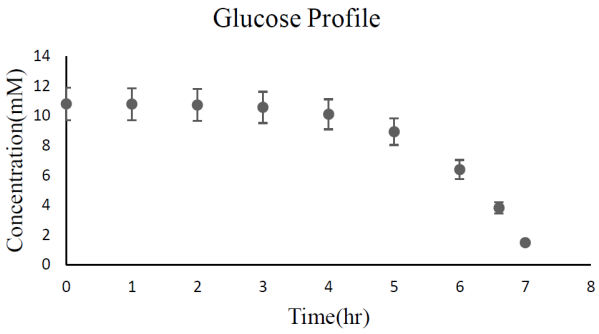


Figure 3.1: Bounded set for glucose concentration during a batch culture. From [114].

3.2.1 Step 1

Solve, for each time interval, the optimization problem described by equations 3.2a to 3.2d in which a flux distribution (v) that maximizes the biological objective function subject to a set of constraints is optimized. In this first step, the functions $f(\boldsymbol{\psi})$ and $g(\boldsymbol{\psi})$, which are the upper and lower uptake/consumption rates constraints, are replaced by the set based constraints obtained from data, as presented in equation 3.2b. The problem is an LP where the objective function is maximized subject to set based constraints and positivity constraints. At each sampling interval k ($k = 1, ..., t_f$):

$$\max_{\mathbf{v}_k} \quad \mathbf{c}^T \mathbf{v}_k \tag{3.2a}$$

$$\text{s.t.} \quad \frac{1}{\Delta k X_{v,k}}(\boldsymbol{\psi}^L_{k+1} - \boldsymbol{\psi}_k) \leq \mathbf{S}\mathbf{v}_k \leq \frac{1}{\Delta k X_{v,k}}(\boldsymbol{\psi}^U_{k+1} - \boldsymbol{\psi}_k) \tag{3.2b}$$

$$\mathbf{v}_k \geq \mathbf{0} \tag{3.2c}$$

$$\boldsymbol{\psi}_{k+1} = \boldsymbol{\psi}_k + \mathbf{S}\mathbf{v}_k X_{v,k} \Delta k \geq \mathbf{0} \tag{3.2d}$$

where $\boldsymbol{\psi}^U \in \mathbb{R}^{r\times 1}$ and $\boldsymbol{\psi}^L \in \mathbb{R}^{r\times 1}$, respectively, are the upper and lower bounds of the metabolites concentration (identified directly from data). The inequality constraints in equation 3.2b are obtained from the metabolites mass balance equation ($\frac{1}{X_v}\frac{d\boldsymbol{\psi}}{dt} = \mathbf{S}\mathbf{v}$). It is important to observe that the same objective function is being considered in equations (3.1) and (3.2a), but, because the metabolite kinetics are not known a priori, the equation 3.2b uses the set based bounds on uptake/consumption rates observed from data, instead of the concentrations' dependent constraints $f(\boldsymbol{\psi})$ and $g(\boldsymbol{\psi})$.

In addition, soft constraints are added for all metabolites consumption/production rates. Soft constraints are constant with respect to concentrations. It was found that soft constraints are needed to limit the solution space since the DMFM optimization is often under-determined (smaller number of constraints than unknowns) thus resulting in multiple solutions. The initial values of the soft constraints can be set as the maximum or minimum consumption/production rates obtained through the data. These soft constraints are added to the set of equations 3.2a to 3.2d.

3.2.2 Step 2

Once the flux for all metabolites (v) are identified for each time sampling k in Step 1, a minimal number of limiting metabolites are determined in Step 2 based on the magnitudes of the corresponding Lagrange multipliers calculated from the LP solution. To this end, consider the equations 3.2a to 3.2d can be re-written as presented in equation 3.3. At each sampling interval k ($k = 1, ..., t_f$):

$$\begin{aligned} \max_{\mathbf{v}_k} \quad & \mathbf{c}^T\mathbf{v}_k \\ \text{s.t.} \quad & h(\mathbf{v}_k) \leq \mathbf{d} \end{aligned} \tag{3.3}$$

where $h(\mathbf{v}_k) \in \mathbb{R}^{(3r+q)\times 1}$ and $\mathbf{d} \in \mathbb{R}^{(3r+q)\times 1}$ are vectors given by equations (3.4) and (3.5).

$$h(\mathbf{v}_k) = \begin{bmatrix} \mathbf{S}\mathbf{v}_k \\ -\mathbf{S}\mathbf{v}_k \\ -\mathbf{v}_k \\ -\mathbf{S}\mathbf{v}_k X_{v,k}\Delta k \end{bmatrix} \tag{3.4}$$

$$\mathbf{d} = \begin{bmatrix} \frac{1}{\Delta k X_k}(\boldsymbol{\psi}^U_{k+1} - \boldsymbol{\psi}_k) \\ -\frac{1}{\Delta k X_k}(\boldsymbol{\psi}^L_{k+1} - \boldsymbol{\psi}_k) \\ \mathbf{0} \\ \boldsymbol{\psi}_k \end{bmatrix} \tag{3.5}$$

In this second step, $f(\boldsymbol{\psi})$ and $g(\boldsymbol{\psi})$ as functions of the corresponding metabolite concentration were calculated based on the sets of bounded data. The Lagrangian for equation (3.3) is given by equation 3.6.

$$L(\mathbf{v}_k, \boldsymbol{\lambda}) = \mathbf{c}^T \mathbf{v}_k + \boldsymbol{\lambda}(\mathbf{d} - h(\mathbf{v}_k)) \tag{3.6}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{1\times(3r+q)}$ is the vector of Lagrange multipliers for a total of $(3r + q)$ inequality constraints described in equation 3.3. Then, solving the Lagrange formulation described by equations 3.7a to 3.7d, optimal fluxes $(\mathbf{v}_k)$ and the corresponding Lagrange multipliers values $(\boldsymbol{\lambda})$ can be found for each constraint at each time interval.

$$\nabla_{\mathbf{v}_k}(\mathbf{c}^T \mathbf{v}_k) + \boldsymbol{\lambda}\nabla_{\mathbf{v}_k}(h(\mathbf{v}_k)) = \mathbf{0} \tag{3.7a}$$
$$\boldsymbol{\lambda}(h(\mathbf{v}_k) - \mathbf{d}) = \mathbf{0} \tag{3.7b}$$
$$\boldsymbol{\lambda} \geq \mathbf{0} \tag{3.7c}$$
$$h(\mathbf{v}_k) \leq \mathbf{d} \tag{3.7d}$$

Following the inequality equation 3.7c, either the Lagrange multiplier is zero or non-zero. A constraint is not active if the corresponding multiplier is zero or very close to zero, whereas a non-zero Lagrange multiplier indicates that the corresponding constraint associated to a particular metabolite is active. The corresponding constraint is included further in the model even if the constraint was found active for only few time intervals during the entire batch duration.

In case of multiplicity

If multiplicity is present in the LP problem, it is possible that the active constraints cannot be properly identified from the values of the Lagrange multipliers. This was observed in cases where among the original set of constraints, there is one of them that is parallel to the objective function. In these cases, only the constraint parallel to the objective function is found active by the Lagrange multipliers approach. A simple toy example that illustrates this case is shown in Appendix G.

Instead, the residuals of each constraint should be monitored to identify which constraints are active in the solution. For instance, if the constraint follows the structure $Ax \leq b$ the residuals $(Ax - b)$ should be equal or smaller than zero to be considered active, e.g. $Ax - b \leq 0$.

In this current work, multiplicity is present. It can be noticed that the biological objective function is defined as the maximization of growth, e.g. maximization of the biomass

flux, while the biomass flux is also constrained by an upper bound. Thus, the constraint is geometrically parallel to the objective function. This situation was not considered in the original algorithm of Nikdel and Budman [112] because the biomass data was used in the solution instead of the biomass prediction. Thus, the corresponding biomass flux was not explicitly considered as done in the current work.

It will be shown in the results' section that the biomass flux is dominant in the solution. The reason that the biomass flux has a particular impact on the overall solution is that since the fluxes are given per unit biomass, the concentration of biomass multiplies each one of the fluxes of the mass balances of the amino acids. Thus all the concentrations are significantly impacted by the biomass concentration that results from the integration of the biomass flux over time. Also, because the consumption/production rates of metabolites are multiplied by the biomass it is possible that some of the rate constraints are redundant. Based on this special role of biomass concentration in the model we hypothesized that if biomass flux constraint is found active this will allow removing constraints in the fluxes related to other metabolites. It should be emphasized that it is important to end up with the smallest number of constraints to avoid over-parameterization of the model to avoid over-fitting.

3.2.3 Step 3

The last step of this identification approach consists in the estimation of the functions $f(\boldsymbol{\psi})$ and $g(\boldsymbol{\psi})$ for the metabolites that were found limiting (active constraints) according to Step 2. Accordingly, at each sampling interval, the consumption/production rates of active constraints are described as kinetic expressions as functions of the corresponding metabolites' concentrations using typical kinetic expressions, such as Michaelis-Menten (Eq. 3.8) or Hill equation (Eq. 3.9) depending of which function results on better fitting of data.

$$\frac{d\psi}{dt} = \frac{K_1 \psi}{K_2 + \psi} \tag{3.8}$$

$$\frac{d\psi}{dt} = \frac{\psi^{K_3}}{K_4 + \psi^{K_3}} \tag{3.9}$$

where $\frac{d\boldsymbol{\psi}}{dk}$ is the uptake/production rates of the relevant metabolite for DMFM (active constraints identified in Step 2), K is the Michaelis-Menten or Hill kinetic parameter to be estimated also for each relevant metabolite involved in the active constraints.

The parameters are estimated by fitting the values of uptake/consumption rates at each time interval with the corresponding data values of metabolites concentration through the

duration of a batch culture. Once these parameters are identified, the predictive DMFM model presented in equation 3.1 can be formulated, with functions $f(\boldsymbol{\psi})$ or $g(\boldsymbol{\psi})$ being set as expressions of the form of equation 3.8 or 3.9. Alternatively, it is possible to use interpolation within look-up tables of the uptake/consumption rates as a function of the corresponding metabolite concentration if closed kinetic expressions such as 3.8 does not result in good fit.

In summary, the numerical mathematical advantages of this 3-step approach are that only linear optimization problems have to be solved in each step and that the functions $f(\boldsymbol{\psi})$ and $g(\boldsymbol{\psi})$ can be identified in Step 3 separately for each of the metabolites that were found limiting in Step 2. This is in clear contrast with other modelling approaches where the nonlinear kinetic expressions related to all metabolites must be simultaneously identified to fit the data resulting in a more difficult nonlinear optimization problem.

3.3 Materials and Methods

3.3.1 Experimental data and cell culture process

The experimental data used to calibrate and validate the DMFM model were provided by MilliporeSigma [114] [64] The experiments were conducted in two bioreactors operated in batch for an initial period of time and then switched to perfusion operation. Both culture processes were performed with CHOZN GS® cell line, at 37°C, 40% DO, 150 rpm in bioreactors of 5 L with working volume of 2 L. In each bioreactor, a different media from MilliporeSigma was used. The initial cell density was about 1×10^6 cells/ml for either bioreactor, and the experiments run from 0 to 62 hours in batch mode, and from 62 to 254 hours in perfusion operation.

Cell density [114] was measured by trypan blue exclusion method; a Nova Bioprofile analyzer was used to quantify glucose, lactate, ammonia and glutamine concentrations and HPLC was used to measure amino acids concentration. Biomass (viable cells), glucose, lactate and ammonia concentration were measured at times 0 h, 22.08 h, 62.40 h, 86.64 h, 110.64 h, 134.64 h, 158.64 h, 188.64 h, 209.28 h, 230.40 h and 254.64 h. Feeding, harvest and bleeding rates were also measured in these same sample times. The amino acids concentration, were measured at times 0 hs, 86.58 hs, 110.75 hs, 134.58 hs, 158.58 hs and 254.58 hs. These data was interpolated hourly using the function "makima" from Matlab. The "makima" function performs a modified Akima piecewise cubic Hermite interpolation with continuous first-order derivatives. The modified Akima algorithm was preferred over

the "spline" algorithm because the former interpolation method resulted in less oscillations as compared to the latter. Using this smoothing algorithm profiles of different amino acids' concentrations were generated from the original data for the time period being studied, e.g. from 0 to 254 hours.

3.3.2 Metabolic network and stoichiometric matrix

The metabolic network used for this three-step DMFM modelling approach is based in the most relevant reactions that take place in CHO cells [114] [110] [178], such as TCA cycle, glycolysis, amino acids' synthesis reactions, glutamine synthesis reaction, biomass formation (as a function of amino acids). For simplicity, balances of co-metabolites, for instance, ATP/ADP and NADH/NAD+, were not considered. The reactions that were considered accounted for the main contributions of carbon and nitrogen molecules resulting in a final network that involves a total of 47 reactions (including reversible reactions, see Appendix A). Once the metabolic network reactions are identified, the stoichiometric matrix ($\mathbf{S}$) can be created based on the stoichiometric coefficients of the reactions involved. The metabolic network used in this work can be found at the Appendix A. All units are in mM and the conversion factors used for biomass assumed a dry cell weight of 350 pg/cell and a cellular molecular weight of 150 g/mol [53].

3.3.3 Solver

The solver "cplexlp" (Dual-simplex Optimizer algorithm), from IBM ILOG CPLEX for Matlab Toolbox, was used to solve the linear programming (LPs) problems presented in section 3.2.

3.4 Results

To develop the DMFM model based on the approach presented above, two data set of experiments performed with the same cell line were conducted with the same bioreactor operating conditions but with different growth media (Media 1 and Media 2). The stoichiometric matrix was built according to the metabolic network reactions shown in the Appendix A. The biomass amino acids' composition is based on the one presented at the work of [51] for hybridoma cells since it is a common cell composition used in previous studies [47] [59] [119] for different mammalian cells' studies. The coefficients of four amino

acids (alanine, serine, proline and methionine) in the biomass composition were slightly adjusted for the cell line used in this work.

This modification of the cell composition coefficients was done by solving the linear optimization program in step 1 above as described by equations 3.2a to 3.2d, after imposing very narrow bounds (equation 3.2b) corresponding to $\pm 0.5\%$ of the original data around the amino acids experimental profiles. The four biomass coefficients referred above were adjusted in order to satisfy the given biomass data. It was found that without these adjustments to the biomass coefficients it was impossible to obtain feasible solutions. The final DMFM model involves a total of 36 species components of which only 24 concentrations were experimentally measured. The components that are not measured are internal metabolites which are produced and consumed inside the cells at a very high rate but are not secreted to the medium in significant quantities [162] [117].

The objective function used in the current DMFM model aims to maximize the growth rate at each time interval. It should be pointed out that growth has been the objective generally proposed for bacterial cells that have been conditioned by natural evolution to maximize their growth. However, mammalian cells are significantly different from bacteria since they involve programmed cell death processes (apoptosis) to avoid unbounded growth in living organisms. In a previous work, Carvalho *et al.* [28] hypothesized and tested other two objective functions for comparison with the growth maximization objective. The purpose of this comparison was to test whether alternative objectives can result in better fit of data based on the sum of squared error (SSE). For this comparison we applied the systematic three-step approach for DMFM modelling presented in section 3.2 for each one of the 3 candidate objectives: Obj-1 - Maximization of growth rate at each time, Obj-2 - Minimization of NADH production in the cytosol plus minimization of NAD(P)H consumption in mitochondria at each time interval, and Obj-3 Combination of Obj-1 and Obj-2. Obj-1 is the objective function typically used for bacteria as explained above. Obj-2 is based on the hypothesis that cells attempt to preserve themselves by mitigation of apoptosis (programmed cell death) and oxidative stress mechanisms. Apoptosis is highly correlated to the levels of cytochrome-C where the latter is the main protein in the respiratory electron transport chain which is highly coupled to the production of NADH/FADH in the TCA cycle. Hence, the minimization of NADH production is expected to result in a reduction of cytochrome-C mediated apoptosis. Similar arguments can be given regarding the coupling of NADH generation with oxidative stress that is damaging to the cell [15]. The production of NADH/FADH can be calculated through the sum of the fluxes of the reactions producing NADH in cytosol and reaction consuming NAD(P)H in mitochondria as per the following sum $v_5 + v_8 + v_{11} + v_{13} + v_{34} + v_{40} + v_{43}$, found in the Appendix A. Obj-3 is an objective function that combines both Obj-1 and Obj-2 into one, thus implying that

the cell allocates resources so as to maximize growth while minimizing apoptosis/oxidative stress.

Carvalho *et al.* [28] tested the DMFM model with the three different objective functions (Obj-1, Obj-2, Obj-3), and concluded that Obj-1 and Obj-3 resulted in the best fit, with the model using Obj-1 presenting a SSE only slightly higher than Model 3, 0.7728 and 0.7529, respectively. In view of the small difference for the sake of simplicity, in the current DMFM model presented in this thesis, the maximization of growth rate at each time interval (Obj-1) was chosen as the objective function.

The two sets of data, obtained from the two different bioreactors' runs described before, were used for calibration and validation of the model. The Bioreactor 1 data was used for model calibration and Bioreactor 2 data was used for model validation. The cell culture process in these bioreactors start operating in batch system for the first 63 hours, and after this period, the bioreactor is operated in perfusion mode, with bleeding, feeding, and harvesting rates set so as to maintain approximately constant volume in the bioreactor. Hourly interpolated data within the interval of 0 to 62.4 hours was used for the identification of the DMFM in batch system. The model describing the perfusion mode was developed considering the hourly interpolated data from 86 to 254 hours. As described in the section 3.3.1 measurements of amino acids concentrations were not available at time 62.4 hour. Hence within the period of transition from batch to perfusion operation between 62.4 to 86 hours could not be calibrated accurately.

Following preliminary trials to calibrate the model for the combined batch and perfusion operations it was found very difficult to find one model that can fit with equivalent accuracy both the batch and perfusion operation periods. Therefore, it was decided to focus the calibration effort on the batch period and then separately assess how the batch model applies to the perfusion period. Accordingly, the calibration for batch and perfusion are presented below separately. In the Conclusions of the chapter a rationale will be provided for possible reasons that explain why the batch and perfusion operations may require slightky different constraints in order to fit the data.

3.4.1 Batch Operation

For the identification of the DMFM during batch system, the Step 1 was solved assuming values for $\boldsymbol{\psi}^L$ and $\boldsymbol{\psi}^U$ that were 8%, respectively, lower and bigger than the original data at the time interval k. The bounds were set at $\pm 8\%$ to ensure that the errors between the predicted metabolites concentrations and the data be below the expected average error given by HPLC measurements [114]. In Step 2 of the 3-step procedure described above,

the Lagrange multipliers approached the identified biomass flux as an active constraint. Presence of multiplicity was observed in the problem, since biomass flux is also part of the set of original constraints. Therefore, based on the constraints residuals, as explained in 3.2.2, the fluxes of the metabolites glutamine, glutamate, lactate and ammonia were then also identified as active constraints while the rest of the rate constraints could be removed. The soft constraints values were adjusted and can be found in the Table B.1 in the Appendix B.

To formulate the predictive model in Step 3, several kinetic expressions (Michaelis-Menten, Hill, etc) were tested to fit the uptake/production rates for the 4 metabolites that were found to be limiting (active). The Hill equation was the kinetic expression that described the data with most accuracy. Using the average metabolites concentration data from the two bioreactors, the production/consumption rates of the limiting components were fit as function of metabolites concentration. The kinetic expressions are function of a metabolites that are reactants in the corresponding metabolic flux reactions. Correspondingly, functions $f(\boldsymbol{\psi})$ and $g(\boldsymbol{\psi})$ were described by the equations 3.10 to 3.13 as follows:

$$R_{Gln} = \frac{\left(\frac{[Asn]}{[Gln]}\right)^{1.7186}}{4.4432 \times 10^4 + \left(\frac{[Asn]}{[Gln]}\right)^{1.7186}} - 0.0037 \tag{3.10}$$

$$R_{Glu} = \frac{\left(\frac{[Glu]}{[Ala]}\right)^{9.4615}}{6.211 \times 10^5 + \left(\frac{[Glu]}{[Ala]}\right)^{9.4615}} - 0.0169 \tag{3.11}$$

$$R_{Lac} = \frac{[Glc]^{20.4181}}{1.3069 \times 10^{30} + [Glc]^{20.4181}} + 0.0756 \tag{3.12}$$

$$R_{Amm} = \frac{\left(\frac{[Glu]}{[Amm]}\right)^{6.0909}}{258.8157 + \left(\frac{[Glu]}{[Amm]}\right)^{6.0909}} + 0.0045 \tag{3.13}$$

where $[Gln]$ is the concentration of glutamine (mM), $[Glu]$ is the concentration of glutamate (mM), $[Lac]$ is the concentration of lactate (mM), $[Amm]$ is the concentration of ammonia (mM), $[Asn]$ is the concentration of asparagine (mM), $[Ala]$ is the concentration of alanine

(mM), $[Glc]$ is the concentration of glucose (mM), and R is production/consumption rate per biomass (mM/h/$g_{biomass}$). It can be observed that the equations 3.10 to 3.13 also have an intercept constant that accounts for the reversing of the corresponding metabolic reactions involving the metabolite under consideration.

It should be noticed that these expressions constrain the rate of consumption or production of a particular metabolite and therefore are obtained from a combination of one or more individual reactions involving that metabolite. Thus, the rate of consumption/production may be also dependent on the concentrations of different metabolites involved in the set of reactions determining this rate. For example, according to the metabolic network shown in Appendix A the consumption/production rate of glutamate is determined by the combination of reactions v_{16}, v_{21}, v_{22}, v_{23}, v_{28}, v_{32}, v_{33}, v_{34}, v_{35} and v_{36}. Accordingly the rate of consumption/production of glutamate may depend on glutamate concentration but also on the concentrations of other metabolites involved in reactions related to glutamate, e.g. ammonia, alanine, aspartate etc. To determine the dependency of the rate with respect to concentrations of metabolites the following procedure was applied. First, plot the production/consumption rate data of the metabolite of interest as function of its own concentration or as function of other metabolite concentration that is involved in one its reactions. If the plot shows that for each metabolite concentration exactly one production/consumption rate can be obtained, this metabolite is used to fit the kinetic function. If for a metabolite concentration two or more production/consumption rate values can be obtained, analyse the plots of the rates as function of metabolites fraction (metabolites should be involved in the reactions). If the plot shows that the metabolites fraction is able to provide only one value for the consumption/production rate, use this metabolites fraction to fit the kinetic function. If more than one metabolite, or metabolite fraction, is able to provide exactly one output value for the production/consumption rate, choose the one that better fits the kinetic function, e.g. the metabolite or fraction of metabolites that provides smaller sum of squared errors for the predicted production/consumption rate.

It was also observed in equations 3.10 to 3.13 that some of the coefficients in the denominator present high values hinting at some metabolic switch between two states determined by genetic regulation. However, such regulation was not explicitly modelled in the current work. These switches could also be represented by sigmoid functions but it was decided to leave them in a Hill kinetic rate form since it is a common biological description of reaction kinetics. Figure 3.2 shows the fitting of equations 3.10 to 3.13 to data. The fits were able to predict with the data error.

It is important to note that among the metabolites that were identified as limiting there were by-products such as ammonia and lactate that are known to be toxic to the cells as they accumulate. Thus, the fact that they were found limiting may indicate that the growth

media should be further optimized to reduce the accumulation of these compounds.

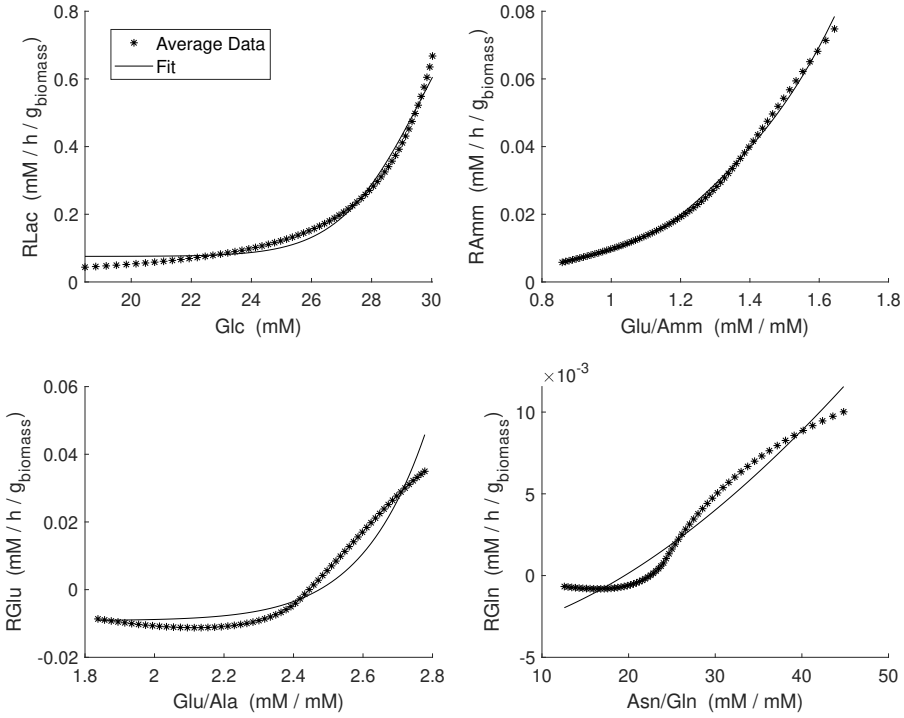


Figure 3.2: Fitting of equations 3.10 to 3.13 compared to data

In theory, the four constraints found in Step 3 (Equation 3.8) should have been sufficient to formulate a DMFM predictive model. However, with only these constraints, multiple solutions of the optimization problem were found due to under-determinacy of the LP problem. Therefore, additional soft constraints were needed to further limit the solution space. These additional constraints consist of constant upper or lower bounds, i.e. independent of concentration, for the metabolites' concentrations. The need for additional soft constraints to limit the solution space of dynamic models has been recognized and reported in previous studies [176].

The model predictions resulting from the DMFM are compared to the hourly interpolated data for Bioreactor 1 that was used for calibration (Fig. 3.3) and model predictions are also compared for Bioreactor 2 that was used for validation (Fig. 3.4) of the model.

Overall, the results presented in Fig. 3.3 and 3.4 indicate good agreement between data and model predictions. Most of the components presented model results within ±15% of error. Observe that in both calibration and validation the Glucose and Biomass predictions exhibit accurate fitting results with the data. Cysteine and Glutamine exhibited over 15% error but those are measured with a high level of uncertainty due to their low concentrations. The lack of accurate fit in some metabolites can be explained by two main reasons: (i) - the use of a biomass composition that was formulated for hybridoma cells and not specifically CHO cells, (ii) - the lack of metabolic regulation related constraints and (iii) - lack of thermodynamic constraints that could inform about reversal of certain reactions.

The DMFM model predictions could be further improved with more data. Data from several bioreactors could be used to improve parameter calibration and to test with more accuracy the validation predictions. Also, the availability of more frequent measurements over time could help to avoid interpolation error resulting from the use of splines with the infrequent data.

3.4.2 Perfusion Operation

The DMFM approach applied to batch operations was assessed and applied for the perfusion operation that followed the batch system. During perfusion operation the bioreactor is continuously fed with fresh media while part of the culture volume is perfused out. The perfusion rate is also referred to as harvest rate since the product is harvested from the perfused volume. The cells from the outflow are retained, typically in a filter, and recirculated back into the bioreactor. Therefore, a perfusion system is able to reach high cell density and operate for a long period of time since toxicants such as ammonia are partially eliminated.

In order to incorporate the perfusion behavior into the model, the last equation of the DMFM approach (equation 3.1) was modified by adding the feeding rate (F) and harvest rate (H) on each of the metabolite concentrations, as shown in equation 3.14. The feeding and harvest rates are assumed constant within time intervals. They were adjusted only at the times samples was collected. The volume of the reactor (V) is assumed constant although some minor deviations were observed in the data. The bleeding rate (B) was also added into the biomass (ψ_{Bio}) Euler integration, as shown in equation 3.15. Bleeding is often used to keep a high viability of the cell and to avoid accumulation of dead cells [40]. Figure 3.5 shows the perfusion rates of bioreactors 1 and 2. Observe that bleed rate was not present in Bioreactor 2 during the operation time being studied.

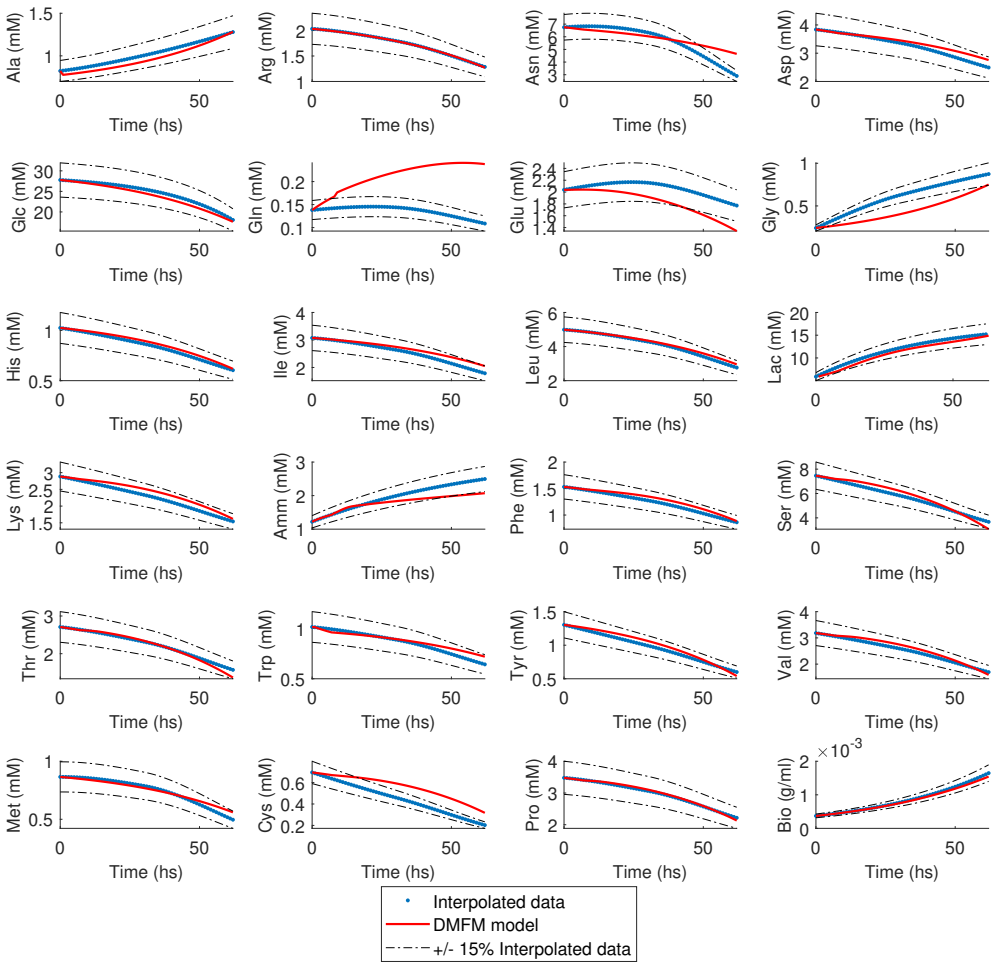


Figure 3.3: DMFM prediction calibration and the hourly interpolated data for the batch system. Abbreviations used in the figure: Ala - Alanine, Arg - Arginine, Asn - Asparagine, Asp - Aspartate, Glc - Glucose, Gln - Glutamine, Glu - Glutamate, Gly - Glycine, His - Histidine, Ile - Isoleucine, Leu - Leucine, Lac - Lactate, Lys - Lysine, Amm - Ammonia, Phe - Phenylalanine, Ser - Serine, Thr - Threonine, Trp - Tryptophan, Tyr - Tyrosine, Val - Valine, Met - Methionine, Cys - Cysteine, Pro - Proline, Bio - Biomass.

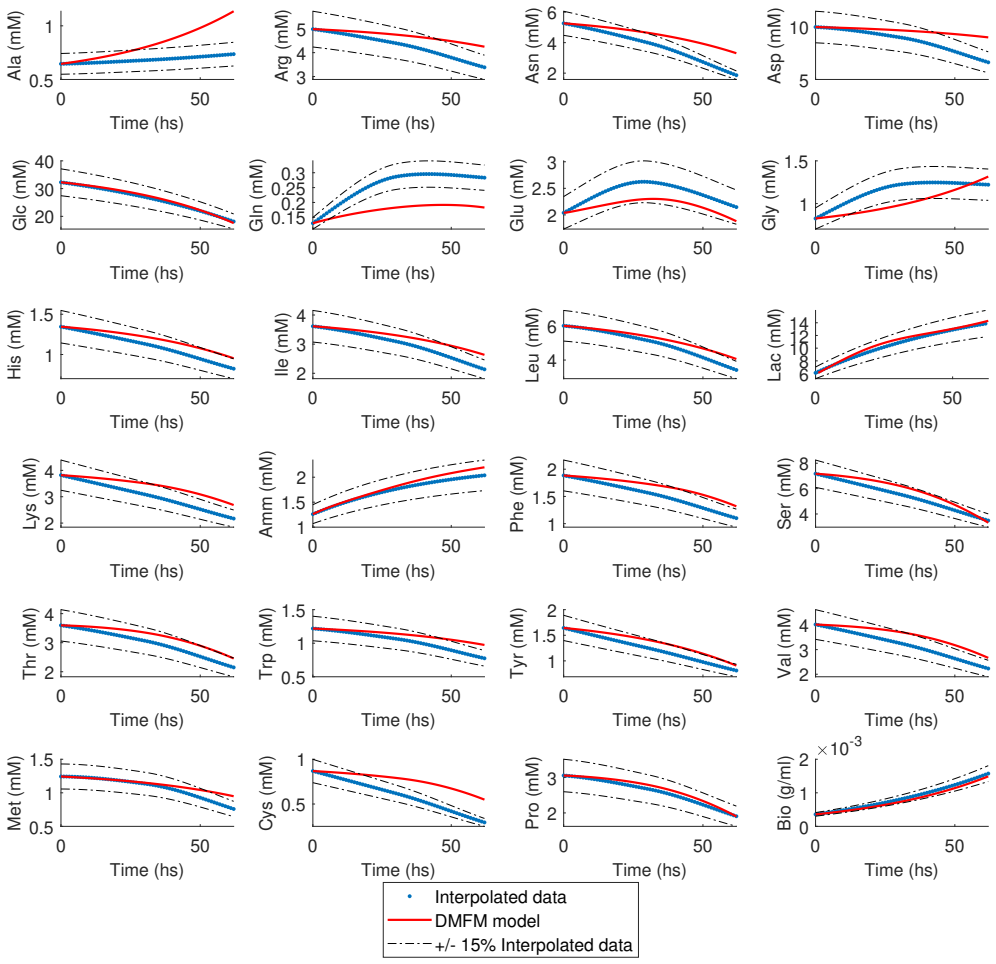


Figure 3.4: DMFM prediction validation and the hourly interpolated data for the batch system. Abbreviations used in the figure: Ala - Alanine, Arg - Arginine, Asn - Asparagine, Asp - Aspartate, Glc - Glucose, Gln - Glutamine, Glu - Glutamate, Gly - Glycine, His - Histidine, Ile - Isoleucine, Leu - Leucine, Lac - Lactate, Lys - Lysine, Amm - Ammonia, Phe - Phenylalanine, Ser - Serine, Thr - Threonine, Trp - Tryptophan, Tyr - Tyrosine, Val - Valine, Met - Methionine, Cys - Cysteine, Pro - Proline, Bio - Biomass.

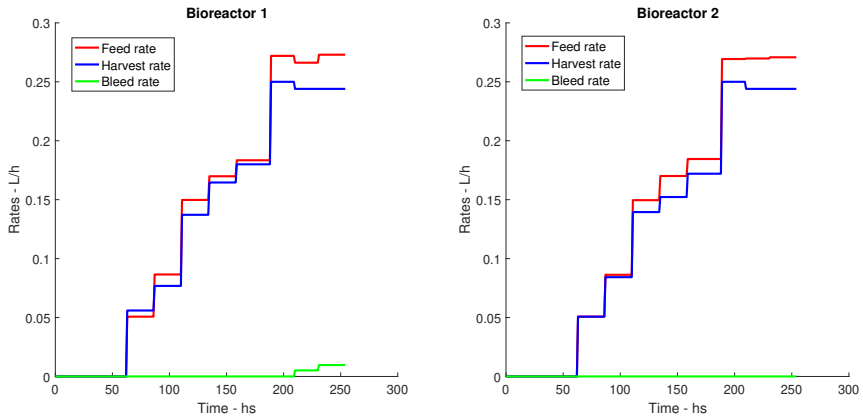


Figure 3.5: Perfusion operation rates for Bioreactor 1 and Bioreactor 2.

$$\boldsymbol{\psi}_{k+1} = \boldsymbol{\psi}_k + \mathbf{S}\mathbf{v}_k X_{v,k}\Delta k + \frac{F}{V}\boldsymbol{\psi}_0 - \frac{H}{V}\boldsymbol{\psi}_k \geq \mathbf{0} \tag{3.14}$$

$$\psi_{Bio,k+1} = \psi_{Bio,k} + \mathbf{sv}_k X_{v,k}\Delta k - \frac{B}{V}\psi_{Bio,k} \geq 0 \tag{3.15}$$

Initially, the DMFM that was identified for the batch operation was expanded by only including the feeding, harvest and bleeding rates into the model. However, this preliminary model was found inadequate to describe the evolution of species especially the evolution of biomass concentration. It was hypothesized that for increasing cell density there may be an increasing death rate due to the occurrence of apoptosis. The effect of cell density on apoptosis (programmed cell death) has been reported in the literature [105, 110]. It is possible to observe in Figure 3.6 that the biomass prediction was consistently higher than the interpolated biomass data thus hinting at the occurrence of cell death. Different expressions were considered for describing the cell death. The best prediction results of biomass concentration were obtained considering the cell death as a function of $c_1\psi_{Bio,k-1}^{c_2}\Delta k$, where c_1 and c_2 are constants. The cell death (apoptosis) term was inserted into the biomass prediction as shown in Equation 3.16. The cell death terms were calibrated aiming at minimizing the sum of squared errors between model fitting and biomass data defined as $c_1 = 0.9$ and $c_2 = 1.9$.

$$\psi_{Bio,k+1} = \psi_{Bio,k} + \mathbf{sv}_k X_{v,k}\Delta k - \frac{B}{V}\psi_{Bio,k} - c_1\psi_{Bio,k}^{c_2}\Delta k \geq 0 \tag{3.16}$$

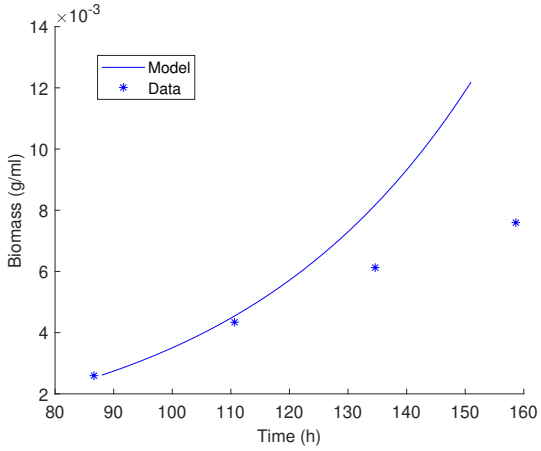


Figure 3.6: DMFM prediction and data for biomass the perfusion system without considering apoptosis.

It should be noticed that the perfusion operation was modeled starting at 86 hours, using the data collected at this time as initial concentration for metabolites. Due to the lack of data during the transition from batch to perfusion system occurring at 62.4 hours, the predictions at the initial hours of perfusion operation could not be accurately assessed. Thus, by setting the initial conditions for the model at 86 hour to the corresponding data we compensated for the lack of information between the beginning of perfusion (62.4 hours) and the first point where data for all amino acids were available (86 hours).

Following the re-application of the 3-step procedure to the perfusion operation it was found that the constraints related to biomass, glutamine, glutamate, lactate and ammmonia were still active during perfusion as they were during batch operation. However, it was found that the kinetic rate expressions that if the kinetic rate expressions used for batch were applied for perfusion the resulting fitting was not satisfactory. Instead, good fitting could only be obtained if the active constraints for glutamine, glutamate, lactate and ammonia were left as a function of time as per equation 3.2b. It is hypothesized that the assumption that many intermediates' concentration remained at quasi-steady state may be inaccurate and may explained the observed differences between batch and perfusion operation. For example, it is known that citrate and pyruvate are present in the media and thus their consumption during batch and feeding during perfusion operation may result in different dynamics and non-steady state behaviour for these metabolites. Since data for these intermediates were not available this argument could not be explicitly checked. How-

ever, future work currently conducted in our group will account for this dynamic behaviour by adjusting the stoichiometric coefficients between measured metabolites.

Also, the soft constraints values (constant with time) have to be slightly adjusted with respect to their values during batch to improve the fitting. Soft constraints values used for batch and perfusion operations can be found at Tables B.1 and B.2 in Appendix B. However, a key finding of this work was that the constraint on biomass flux remained the same value for batch and perfusion thus hinting at the fact that the production of biomass is not directly determined by the consumption of amino acids but is due to another limiting intermediate, e.g. ribose that contributes to the production of DNA and RNA [74, 72, 73].

Figure 3.7 shows the comparisons between model predictions and data for the metabolites and biomass prediction (the interpolated data is not shown in this figure) during perfusion operation. Good fitting is obtained both in the calibration (Bioreactor 1) and validation (Bioreactor 2) of the model. Although the production of Alanine and Glycine were over predicted, the media components predicted the calibration and validation data with an average error of 13.14% and 15.38%. The Cysteine and Asparagine concentrations predictions also were inaccurate for both bioreactors; cysteine concentration data was expected to be inaccurate due to equipment measurement limitations reported to us by MilliporeSigma, while Asparagine concentration may be inaccurate due to the closeness of Asparagine and Aspartate peaks in the HPLC chromatogram. The glucose, lactate, ammonia and biomass profiles were predicted with an error smaller than 8% for the Bioreactor 1 (calibration) and less than 15% error for the Bioreactor 2 (validation). The model was also able to describe the significant consumption occurring around 180 hours, observed in the data. It was also observed that the significant decrease of glucose at that time propagated to the other metabolites present in the network since glucose is the major nutrient in terms of carbon content. Although the minima in amino acids concentrations were not measured at this time, the model predicted such minima due to the behaviour of glucose.

It should be emphasized that the model obtained for perfusion operation is not predictive since the constraints are given as a function of time. For the model to be predictive, these constraints should be expressed as a function of concentrations. However, it was not possible to find time independent kinetic expressions for these constraints as was done for batch operation. The dynamics of intermediates that was ignored in the present work could possibly explain the need for time varying constraints to describe the data. An additional possible source for the discrepancy observed between batch and perfusion operations could be due to the objective function that was maximized in the model. Although the maximization of growth rate was also used as objective function in the DMFM during perfusion, this may not be the best possible choice of objective for mammalian systems, specially during the perfusion operation. Although during the initial exponential growth

phase cells invest most of their energy in growth, during perfusion growth may not be the main focus while minimizing cell death may become more important. In this way, different objective functions, or possibly a weighted combination of different objective functions, could be considered in future work during perfusion.

3.5 Conclusion

A systematic approach was applied to identify the limiting constraints of a DMFM model for mammalian cells. The novel contributions of the work with respect to a previous study are: i- the biomass was predicted by the model instead of using data , ii- the consideration of cell death in the model and iii- the model was applied to both batch and perfusion operation.

A key finding of the modelling study is that the biomass flux is an active constraint during the entire duration of the batch and subsequent perfusion operation. Also, the flux of biomass was almost constant during the entire operation. However, cell death had to be considered in order to fit the biomass measurements especially towards the end of the perfusion operation. The fact that a constraint on biomass flux is needed indicates that the dynamic of the measured amino acids are not sufficient to describe the growth rate. Instead it can be argued that other species such as ribose or NADPH which are known contributors to biomass growth maybe limiting.

Also, active constraints related to four metabolites: glutamine, glutamate, lactate, and ammonia were required to describe the data. During the batch operation the consumption/production rates of the active constraints were properly described by kinetic functions that follow the Hill equation. The model was able to satisfactorily predict the experimental data from batch systems. It should be noticed that in the current work only 4 kinetic constraints and a constant constraint on biomass flux were necessary to explain the data as compared to a larger number of kinetic constraints (6) needed in a previous study [28]. The ability to describe the model with a smaller number of constraints is advantageous since it leads to a smaller number of parameters that need to be calibrated. The fact that the limiting constraints correspond to metabolites that are produced in the culture, where some of those are particularly toxic (ammonia, lactate), indicates that the growth media should be further optimized to reduce such accumulation.

The DMFM approach, with same objective function and same active constraints for batch systems, was extended to perfusion operation. However, the Hill expressions that were used for batch operation did not result in good fitting during perfusion. Instead,

the consumption and/or production rates of the limiting components had to be described as functions of time. The use of such time varying constraints results in a model that is not predictive. We hypothesize that the dynamic behaviour of intermediate species, which were assumed to be at quasi-steady state in the current model, may explain the need for time varying constraints. Thus, we expect that the dynamic of these intermediates is significantly different for batch and perfusion thus explaining why the Hill kinetic expressions that applied to batch did not apply for the perfusion operation. Moreover, we argue that the dynamic accumulation/consumption of intermediate species may depend on the perfusion and harvesting rates which change during perfusion operation thus explaining the need for time varying constraints. For example intermediates such as citrate and pyruvate are present in the media and their dynamics may be significant. On the other hand these intermediates were not measured. We are currently investigating the possibility to introduce a time varying stoichiometric matrix that is a function of measured metabolites and perfusion rates as a way to account for the varying dynamics of intermediate species.

The effect of cell death was found to be very significant in order to explain the evolution of cell mass especially during perfusion operation.

In addition to kinetic constraints, soft constraints that are constant during the entire operation were necessary due to the multiplicity of solutions of the LP problem. Thermodynamic or genetic regulation related constraints could be added in the future to address the multiplicity of solutions and to reduce the number of soft constraints.

An additional source of model error may be related to the choice of the objective function used in the LP problem. For instance, the perfusion operation may require a different objective from batch. While in batch operation cell growth is the dominant feature during perfusion energy may have to be directed to cell maintenance. Although for the objective functions that were tried in this work the maximization of growth rates resulted in the best fitting, other objective functions related to cell maintenance or thermodynamic objectives (Gibbs energy) could be tried to improve the fitting.

In conclusion, the DMFM approach was proved very satisfactory, being able to predict the batch and perfusion data with an average error of 15%. One of the main advantages of the DMFM modelling approach is the ability to predict the metabolites concentration with a smaller number of parameters. For example, while the presented DMFM used a total of 40 parameters, the conventional kinetic models [64] usually require over 80 parameters to obtain the same level of prediction error.

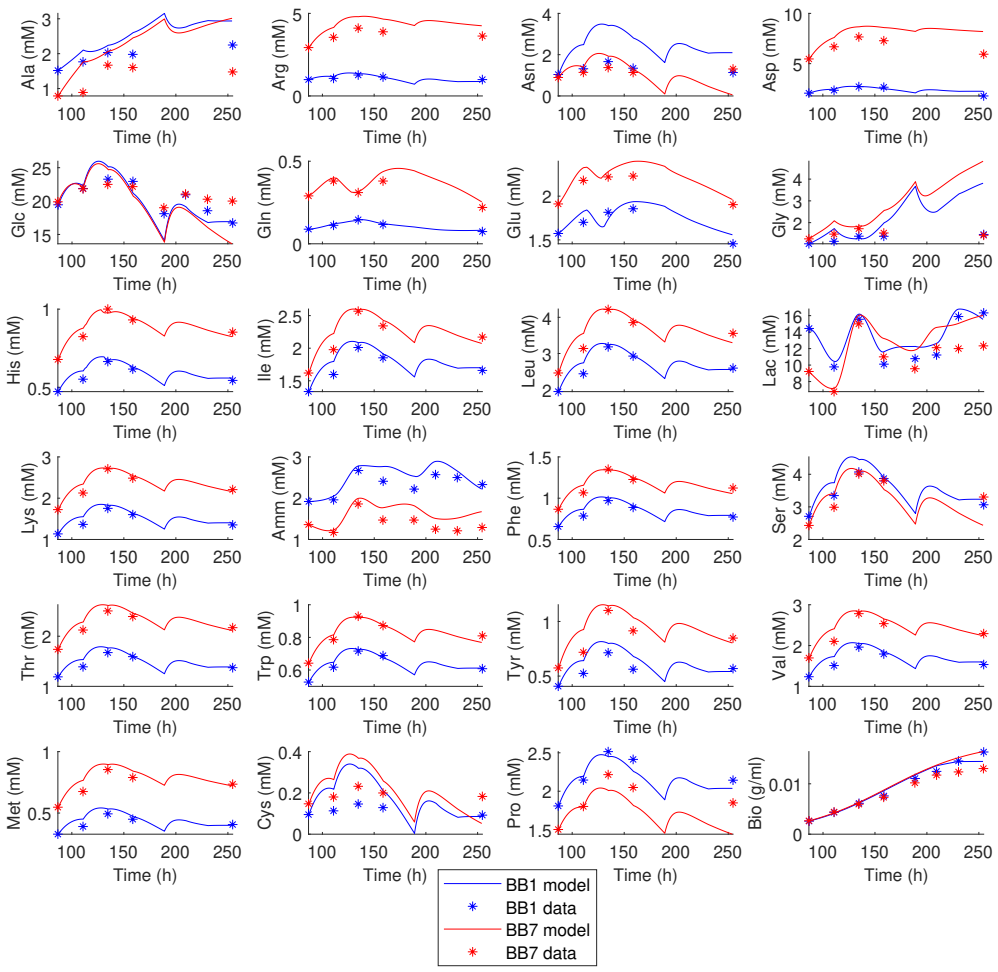


Figure 3.7: DMFM prediction calibration (BB1, Bioreactor 1) and validation (BB7, Bioreactor 2) and data for the perfusion system. Abbreviations used in the figure: Ala - Alanine, Arg - Arginine, Asn - Asparagine, Asp - Aspartate, Glc - Glucose, Gln - Glutamine, Glu - Glutamate, Gly - Glycine, His - Histidine, Ile - Isoleucine, Leu - Leucine, Lac - Lactate, Lys - Lysine, Amm - Ammonia, Phe - Phenylalanine, Ser - Serine, Thr - Threonine, Trp - Tryptophan, Tyr - Tyrosine, Val - Valine, Met - Methionine, Cys - Cysteine, Pro - Proline, Bio - Biomass.

Chapter 4

Development of new media formulations for cell culture operations based on regression models

This chapter is adapted from the published paper [29] Mariana Carvalho, Jeremiah Riesberg, Hector Budman. Development of new media formulations for cell culture operations based on regression models. Bioprocess Biosyst Eng., 44(3):453-472, 2021.

4.1 Overview

This chapter discusses modelling and optimization of a multi-component cell culture medium. The approach in this chapter, in contrast with the other chapters of this thesis, is to model the process by empirical equations. The specific productivity (Qp) was considered to be a function of the medium components and possible interactions described by linear factors, two-way interactions and squared terms that results in a high dimensional problem where the number of variables p (represented by the medium components and their interactions) is much larger than the number of observations n. This problem has been identified as particularly challenging for regression and it has been referred in the literature as $n < p$ problem. Principal Components Regression (PCR), Partial Least Squared (PLS), LASSO and Elastic Net regressions were compared as modelling tools to deal with the $n < p$ problem. PCR and PLS regression models resulted in better prediction results and were used for robust optimization of the medium composition by a nonlinear optimization. The

experimental results of the case studies show that it is possible to formulate new media that result in equivalent Qp values to the highest Qp values obtained by the industrial partner within the margin of measurement error. On the other hand the new formulations resulted in consistently higher cell density and mAb values throughout the cell culture as compared to the existent formulations. Also, the multivariate statistical approach permitted us to select a subset of media that is most informative about the optimum thus permitting modelling and optimization with a reduced set of initial experiments.

4.2 Introduction

Cell culture medium is a complex mixture composed of major components, such as amino acids, and other minor components that are the source of nutrients for cell growth and protein production. Interactions among chemical components present in the medium are known to have a major impact on cell culture performance [175]. Due to the large number of different components utilized in media the optimization and development of new formulations is a challenging task. The typical approach for medium design used by the pharmaceutical industry has been based on the blending of a particular set of pre-mixed formulations ("master" media) in different proportions as calculated by a Design of Experiments (DoE) approach, e.g one-factor-at-a-time (OFAT) method, Plackett-Burman design or Response Surface based methodologies [170, 124, 133, 71]. However, a key difficulty with the use of input data that is based on the mixing of "master" media is the introduction of very high correlation in the data which introduces numerical challenges for identification of empirical models based on these data.

Design of experiments is a common technique used specifically to reduce the number of experiments needed for a design without losing significant information about the process variables. The occurrence of a large number of components in cellular medium that can be manipulated for optimizing the cell culture performance poses a major challenge for the application of standard design of experiments approaches. To tackle this high dimensionality multivariate statistical tools, e.g. Principal Component Analysis (PCA), can be combined together with design of experiments approaches. However, the design and the interpretation of the design results remain highly complex when the number of variables (p), as given by the candidate components in medium and the potentially important interactions among them, is much larger than the number of observations (n) which is a very common scenario in design of cell culture media. Then, if a model is sought for prediction and optimization, the number of model parameters will be equal to the number of variables (p). The problem of parameter estimation with a reduced number of experiments has recently

received increased attention in the context of gene regulation modelling problems based on limited amount of data [24, 132]. In this case, because the number of observations is much smaller than the number of variables $n < p$, the least square approach is not viable due to the under-determinancy of the problem. Instead, other statistical techniques must be used to evaluate the model parameters such as Ridge regression, LASSO regression, Elastic net or Dantzig Selector. The dimensions of the problem can also be reduced (compressed) based on multivariate methods such as Principal Component Analysis (PCA) or Partial Least-Squares (PLS) regression.

In the current study we discuss the optimization of the medium utilized in a mammalian CHO cell culture producing a monoclonal antibody. The main objective for this optimization problem is to design medium that results in maximal specific productivity over the duration of a simulated perfusion operation. The experiments to be used for this design must provide a set of data that is informative enough about the level of productivity but the number of experiments involved in this design must be kept small due to cost and duration of the experiments. The approach adopted in the current study is to develop regression models based on the optimally designed experiments between the concentrations of media components and the quantity of interest, e.g. specific productivity, and to use this model to optimize the productivity with respect to the media composition. The regression model assumed for this task is of the general following form:

$$\begin{aligned} Qp_t = & \sum_{i=1}^{n_\psi} \left(\beta_{\psi_i}\right) \psi_i + \sum_{j=1}^{n_\phi} \left(\beta_{\phi_j}\right) \phi_j + \sum_{i=1}^{n_\psi - 1} \sum_{k=1}^{n_\psi} \left(\beta_{\psi_i \psi_k}\right) \psi_i \psi_k \left\{i \neq k\right\} + \\ & \sum_{j=1}^{n_\phi - 1} \sum_{l=1}^{n_\phi} \left(\beta_{\phi_j \phi_l}\right) \phi_j \phi_l \left\{j \neq l\right\} + \sum_{i=1}^{n_\psi} \sum_{j=1}^{n_\phi} \left(\beta_{\psi_i \phi_j}\right) \psi_i \phi_j + \\ & \sum_{i=1}^{n_\psi} \left(\beta_{{\psi_i}^2}\right) {\psi_i}^2 + \sum_{j=1}^{n_\phi} \left(\beta_{{\phi_j}^2}\right) {\phi_j}^2 \end{aligned} \tag{4.1}$$

where Qp_t is the specific productivity at the end of the simulated perfusion operation, ψ_i is the major component i concentration in the medium formulation (e.g. amino acids, major by-products, such as ammonia and lactate, and main nutrients such as glucose and glutamate), ϕ_j is the minor component j concentration in the medium formulation (e.g. vitamins, hormones, metals, etc), n_ψ and n_ϕ are, respectively, the number of major and minor components that are being considered in the medium formulation, and $\beta_{\psi_i}, \beta_{\phi_j}, \beta_{\psi_i \psi_k}, \beta_{\phi_j \phi_l}, \beta_{\psi_i \phi_j}, \beta_{{\psi_i}^2}, \beta_{{\phi_j}^2}$ the regression coefficients. This regression model assumes that the predicted response, the specific productivity (Qp_t) at the end of the

simulated perfusion operation, is a function of the regressors composed by linear factors ($\sum_{i=1}^{n_\psi} \psi_i$ and $\sum_{j=1}^{n_\phi} \phi_j$), two-way interactions ($\sum_{i=1,k=1}^{n_\psi} \psi_i\psi_k \{i \neq k\}$, $\sum_{j=1,l=1}^{n_\phi} \phi_j\phi_l \{j \neq l\}$, and $\sum_{i=1}^{n_\psi} \sum_{j=1}^{n_\phi} \psi_j\phi_l$) and squared terms ($\sum_{i=1}^{n_\psi} {\psi_i}^2$ and $\sum_{j=1}^{n_\phi} {\phi_j}^2$) of chemical components, i.e. major (ψ) and minor component (ϕ) concentrations in the medium formulation. The interactions assumed in this model are common due to the correlated nature of the metabolic reactions occurring in a mammalian cell.

While the experimental design is mainly directed towards identifying the composition of the medium that will result in maximal productivity, it is often the case that after the initial set of experiments is performed, follow-up experiments are needed to measure additional quantities of interest. For example, if it is desired to conduct HPLC studies on the dynamic evolution of amino acids along the culture or to measure the level of glycosylation of the antibody, it is often necessary to limit these additional measurements to a smaller subset of experiments to avoid costly experimentation.

Following these arguments this work presents three case studies: 1- A small scale toy example is used to compare the performance of different regression modelling tools with particular focus on the case where the number of samples is much smaller than the number of model regression predictors; 2- Optimization of the cell culture medium using all observations from the experimental design that is typically used in an industrial setting which is based on the blending of a small set of "master media" in different ratios; 3- Optimization of the cell culture medium using only a reduced subset of experiments (subset of observations) used in Case Study 2. For all cases the number of regression predictors, that include linear factors, two-way interaction and squared terms, is much larger than the number of available experiments. This case is referred in the literature to as the "$n < p$ regression problem" where n is the number of samples and p is the number of input predictors considered for regression [168].

To tackle this problem, first we compared different regression methods including PCR, PLS, LASSO or Elastic Net regressions to better understand the limitations regarding each regression method for the "$n < p$ regression problem". Then, in Case Studies 2 and 3 nonlinear optimization algorithms were applied to the PCR and PLS models to find new optimal media formulations. The optimization was made robust to uncertainty by considering the variance in the predictions. Specifically, in Case Study 3 we proposed the use of Principal Component Analysis (PCA) combined with the D-optimal criteria to design, through a genetic algorithm, the set of experiments that are most informative, among all experiments available, for finding new optimal media formulations.

4.3 Model regression and Design of Experiments for high dimensional n < p problem

Formulations of cell culture media include several chemical components to support cellular growth and protein yield, such as amino acids, peptides, vitamins, growth factors, fatty acids, trace elements, and salts. Either in large or small concentrations, each of these components has a critical role in cellular metabolism. Furthermore, these components can interact with each other affecting the performance of cells and protein titer [170]. For instance, trace metal elements are found in very low concentration in cellular medium but are essential for metabolic pathways regulation and enzymes activity [133]. Also, amino acids and metal ions can interact affecting their availability and stability in the cell culture medium [138].

The optimization of cell culture media for mammalian cells is a challenging process due to the large number of variables (chemical components) and the possible interactions among these variables combined with the need to limit the number of experiments to be performed due their cost and time constraints. Furthermore, the optimal medium components' concentrations can vary according to the cell line and to a desired performance indicators of the culture such as specific productivity, viability and cell density. Although deterministic stoichiometric models have been used in recent years to study optimal media composition for mammalian cell lines, these models often do not account for the effects of all media nutrients, especially minor components, such as non-amino acids and trace elements, on the cellular metabolism [133]. Therefore, in order to determine the most favorable medium formulation the pharmaceutical industry still relies on empirical (black-box) models based on blending of specific formulations, often referred to as "master" media, in combination with optimal design of experiments approaches [133].

One-factor-at-a-time (OFAT), Plackett-Burman and Response surface methodologies have been proposed for media optimization. The OFAT is very time and resource consuming [124], and fails to distinguish interactions among the medium components [133]. Plackett-Burman designs are usually used to evaluate significant components [133, 170]. However, this method requires at least one more experimental point than the number of variables (medium components) being studied [170], which means it requires a relatively large number of experiments if the number of components is large. The response surface methodology is a design of experiments approach that uses stochastic search to obtain an optimal set of experiments [124, 120], but the training of the response surface based model is computationally challenging when there is high correlation among the variables in the input data used for model training which is often the case when the input data is based

on different combinations of "master" media [170].

Currently, statistical design of experiments have been largely used for media optimization due to its ability to assess the simultaneous influence of several components and components interactions while reducing the total number of experiments required for modelling the data [133, 71]. Sequential statistical experimental design for media optimization involves [170, 124] several steps as follows: i-the screening of factors (identification of significant media nutrients); ii- the establishment of optimal ranges for the nutrients that were identified as important, iii- search for an optimum medium composition and iv- experimental check of the new formulation.

Cockshott and Sullivan [32] used a combination of sequential Plackett–Burman, factorial, response surface and ridge analysis to optimize the medium used for Echinocandin B production by *Aspergillus nidulans*. The author analysed 15 medium nutrients using 108 experimental samples in which 5 factors were found important and titer was increased by 46%. Rajendran and Thangavelu [124] used Plackett–Burman experimental design, response surface methodology and an artificial neural network to optimize the medium used for lipase production by *Bacillus sphaericus*. The authors studied the effect of a moderate number of medium components (12 components) using 16 experimental samples, finding that only 5 medium components were important for improving lipase production.

The works of Ramesh and Murty [125], Parthasarathy and Gnanadoss [120], Ju *et al.* [71] and Zhang *et al.* [180] also used a sequential experimental design to optimize media formulation aiming to enhance the production of a substrate of interest.

It is widely recognized that mathematical models can be used to improve guided design of experiments [80, 17]. The combination of mathematical models and optimal design of experiments can generate new experimental possibilities where maximum statistical information can be obtained and analyzed [80, 122]. Following this idea, the current work proposes the use of multivariate regression models combined with the D-optimal design approach to analyse experimental data to optimize the media components to improve the specific productivity results. The focus is on cases where there are many components and thus the number of components (p) and interactions among them is much larger than the number of observations (n). In view of the large number of chemical components and interactions that will be each assigned a parameter in the regression model, a key challenge is to avoid the model to over-fitting the noise in the data by reducing the number of model parameters [89]. Thus, it is crucial to mathematically compress the data into a lower dimensional space without losing important information for the optimization task. Data compression tools such as Principal Components Regression (PCR) and Least Absolute Shrinkage and Selection Operator (LASSO) [155] are good candidates to reduce the

complexity of the resulting models to be used for optimization of medium composition and thus they are investigated in this study.

PCA uses orthogonal transformations to find a subset of principal components (corresponded eigenvectors), the "new variables" (scores), that are able to explain a desired level of variability present in the original data [19, 132]. The PCA transformation of the inputs is then used in a Principal Component Regression model (PCR), in which an output interest, e.g. specific productivity, is regressed with respect to the scores of a subset of principal components of the input data. Because the scores fitting the principal components are orthogonal, PCR is able to reduce the sensitivity to noise arising from the collinearity present in the data [58]. It should be noticed that regardless of the number of principal components (PC) that are ultimately considered, the PCR model retains information regarding all original variables since each principal component involves a weighted linear combination of all input variables.

While PCR leads to models that consider all variables within the principal components, it is often of interest to identify variables in a media formulation that affect or do not affect an output of interest such as cell growth or cell specific productivity. The Least Absolute Shrinkage and Selection Operator (LASSO) [155] is an analysis technique that performs such variable selection while also providing a regression model with respect to the significant variables. LASSO have been applied in the "$n < p$ regression problem", leading to the selection of a subset of variables that that results in the smaller prediction error [82] but it has some known limitations. For instance, in case of collinearity in the input data LASSO is deficient and other methods such as Ridge regression are found more accurate. Also, for the "$n < p$ regression problem", LASSO can only select a maximum of significant regressors that is equal to the number of samples [166]. LASSO also ignores any meaningful variable order, and may select an incorrect two-way interaction variable which has a component that is actually important when it is present in another two-way interaction variable that was not identified as significant [82]. Moreover, in the "$n < p$ regression problem" LASSO may not necessarily select the real important variables (false positive or false negative significant coefficients) [92, 166].

Several studies have been recently conducted to address these LASSO limitations. To deal with inconsistent variable selection by the regular LASSO, Zou [182] proposed an Adaptive LASSO method, in which adaptive weights were assigned to specific coefficients to penalize the resulting regression model but the choice of these weights is often arbitrary. Luo and Chen [95] proposed a Sequential LASSO approach but it also requires particular conditions with respect to the number of samples versus the number of variables.

Tibshirani [156] discussed the LASSO solutions when the number of predictor variables

is larger than the number of observations. Tibshirani [156] identified special cases with unique solutions occurring when the input data matrix satisfies a restricted eigenvalue condition known as Restricted Isometry Property (RIP). However, as has been discussed in the literature [26, 25, 2, 45, 13, 121, 4], there are currently no systematic ways to design a regressor matrix that satisfies this condition. Also, testing the RIP condition of a given large regressor matrix involves a difficult numerical problem (NP-hard) [13]. Some simpler tests have been proposed to test RIP condition, such as mutual coherence [45, 2, 4], but these are sufficient but not necessary conditions and thus may be conservative.

There are other extensions of LASSO such as Ridge regression and Elastic Net that have been proposed for developing regressions between input to an output of interest while reducing the number of model parameters. However, it has been recognized by Waldmann *et al.* [166] that when there is high collinearity in the data and the number of variables and interactions is much larger than the number of samples they may provide inaccurate results.

Candes and Tao [24] proposed the Dantzig selector to find the significant input variables with respect to an output variable when the number of samples were much smaller than the number of factors. However, the Dantzig selector only identifies the significant variables provided that the input data satisfies a specific sparcity condition [24]. Such condition is not trivial to satisfy when designing media based on a set of master solutions as commonly done in the pharmaceutical industry.

In view of the challenges presented by the different regression methods for the $n < p$ case, in this study we are comparing these techniques using numerical examples that were tailored to elucidate the efficiency of the methods in the specific context of culture medium design and optimization.

All the scenarios considered in this work deal with input data where the number of predictors (p), including linear terms, two-way interactions and squared terms as per the model assumed in equation 4.1, is larger than the number of observations (n).

Three case studies are presented in this work as follows:

- Case 1- Comparison of PCR, PLS, LASSO and Elastic Net regression models calibrated with simulated data in terms of accuracy and their ability to identify significant predictors and prediction of new observations;
- Case 2- Development of optimal media formulation to maximize specific productivity, referred henceforth as Qp, using observations from a set of data based on blending

of certain number of master media (used by our industrial partner) in different proportions;

- Case 3- Development of optimal media formulation using only a reduced subset of observations from the set assumed in Case 2, chosen according to the D-optimal criteria, to reduce experimental costs.

The approach used for calibration and validation of all models considered in this study are based on the use of 3 sets to be referred henceforth as training, testing and prediction sets. The general idea is that the data in the training and testing sets are both used to calibrate different parameters of the regression model. In order to add robustness to the model, data is exchanged between the training and testing sets as per the "leave m out" procedure proposed in literature [145] through several loops. The partitioning between training and testing set is based on a 70/30 ratio. Finally the model that has been calibrated by using the training and testing sets is subsequently validated with an additional prediction set that was not used in the development of the model. The validation of the models in Case Study 2 and 3 is done for both simulated data and experimental data. For all the regression models presented in this study confidence intervals are calculated for the model parameters. Then, in Case studies 2 and 3 we conducted robust optimizations to calculate solutions that are tolerant/robust to the uncertainty in the model parameters. The validation of the robust optimization results is done only with simulated data. The following sections describes the methods and results for each of these case studies.

4.4 Case Study 1: Comparison of regression modelling tools used in this study

4.4.1 Methods

Simplifying the media formulation or identifying which components have effects on outcomes of interest, e.g. productivity or cell growth, are crucial in the design of culture media. Regularization based regression methods such as LASSO or Elastic Net have the potential to identify components or interactions among components that have no effects on outcomes by penalizing the parameters of the regression model. However, although these methods are known to be effective when a large number of experiments is available, these methods are often less effective when applied to the $n < p$ case which is typical in media design problems. To illustrate this point a toy example is formulated where a

process describing the relation between the media components' concentrations and the Qp is simulated as follows:

- The assumed model involves four medium components: two major components (ψ_1 and ψ_2) and two minor components (ϕ_1 and ϕ_2);
- As described in Eq. 4.1, the Qp model response was assumed to be function of linear terms, two-way interactions and squared terms as per equation (Eq. 4.2);

$$\begin{aligned} Qp_t = &\beta_1\psi_1 + \beta_2\psi_2 + \beta_3\phi_1 + \beta_4\phi_2 + \beta_5\psi_1\psi_2 + \beta_6\psi_1\phi_1 + \\ &\beta_7\psi_1\phi_2 + \beta_8\psi_2\phi_1 + \beta_9\psi_2\phi_2 + \beta_{10}\phi_1\phi_2 + \\ &\beta_{11}{\psi_1}^2 + \beta_{12}{\psi_2}^2 + \beta_{13}{\phi_1}^2 + \beta_{14}{\phi_2}^2 \end{aligned} \tag{4.2}$$

- The concentrations are generated according to uniformly distributed random data, varying from 0 to 1. The regressor matrix is built using the corresponding values of these linear factors, two-way interaction and squared terms. The regressor matrix is represented by $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$;
- Six media formulations are considered in this study case with $n = 6$ observations and $p = 14$ regressors;
- The regressor matrix values are mean-centered and normalized by its regressor standard deviation;
- The simulated Qp response is assumed as follows: $Qp_t = 5\psi_1 + \phi_2 - 3\psi_1\psi_2$. The Qp vector has dimensions $\mathbf{Qp}_t \in \mathbb{R}^{n \times 1}$. Mean centered and normalized regressors values are used to generate Qp data;
- Random noise ε with a magnitude of 8, 10 and 15% percent of the full scale variation of the output data is incorporated into the simulated response data;
- Lastly, a set with 60 media is used to evaluate the model prediction.

PCR based model

A regression model of Qp as a function of media components concentrations based on Principal Components Regression (PCR) is developed as per the following steps:

i. Apply PCA to the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$;

ii. Select a number of principal components (n_{PC}) so as the scores of the PCA ($\mathbf{XS}_{n_{PC}} \in \mathbb{R}^{n \times n_{PC}}$) are able to describe a specific level of input variability.

iii. Use the simulated specific productivity $\mathbf{Qp}_t \in \mathbb{R}^{n \times 1}$ and the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$ from the training set in combination with the loadings of PCA ($\mathbf{XL}_{n_{PC}} \in \mathbb{R}^{p \times n_{PC}}$) to estimate the PCR regression coefficient vector $\boldsymbol{\beta}_{PCR} \in \mathbb{R}^{n_{PC} \times 1}$ according to the equation given by: $\hat{\mathbf{Qp}}_t = (\mathbf{X}_{regressor})(\mathbf{XL}_{n_{PC}})\boldsymbol{\beta}_{PCR}$. Evaluate the predicted Qp and regression error given by the PCR model regression for the training set;

iv. Use the testing set to predict the Qp and calculate the prediction error for the testing set;

v. The number of principal components n_{PC} used in the PCR regression model is adjusted such that both the errors for both the training and testing sets are smaller than the noise ε assumed in the simulated data.

PLS based model

A regression model of Qp as function of the initial media components concentration based on Partial Least-Squares (PLS) regression is developed as follows:

i. Apply PLS to the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$ and the simulated specific productivity vector $\mathbf{Qp}_t \in \mathbb{R}^{n \times 1}$ from the training set;

ii. Select a number of latent variables (n_{PC}) such as the predictor and response scores of PLS (respectively $\mathbf{XS}_{n_{PC}} \in \mathbb{R}^{n \times n_{PC}}$ and $\mathbf{YS}_{n_{PC}} \in \mathbb{R}^{n \times n_{PC}}$), satisfy a specific level of covariance.

iii. Estimate the PLS regression coefficient vector $\boldsymbol{\beta}_{PLS} \in \mathbb{R}^{p \times 1}$ based on the training data as follows: $\boldsymbol{\beta}_{PLS} = \mathbf{W}_{n_{PC}}(\mathbf{YL}_{n_{PC}})^T$. The predicted Qp according to the PLS based model is given by: $\hat{\mathbf{Qp}}_t = (\mathbf{X}_{regressor})\boldsymbol{\beta}_{PLS}$;

iv. Evaluate the predicted Qp and regression error given by the PLS based model regression for the training set;

v. Use the PLS regression coefficient vector $\boldsymbol{\beta}_{PLS} \in \mathbb{R}^{p \times 1}$ to predict the Qp and the prediction error of the testing set;

vi. The number of latent variables is adjusted such that the prediction error for the training and testing sets in the model calibration are smaller than the noise ε assumed in the simulated response data.

In both PCR and PLS approaches, there is a trade-off between the variability explained by the scores and the fitting accuracy. As the number of principal components selected increases, the training error decreases and the fitting accuracy improves. However, this improvement is generally a result of over-fitting of the noise training data thus often resulting in a higher prediction error.

LASSO and Elastic Net Based Models

With the goal of reducing the number of model parameters and identifying media components that do not contribute significantly to the output of interest LASSO and the Elastic Net regression models were developed. In these methods the number and values of active coefficients (non-zero) $\beta \in \mathbb{R}^{p\times 1}$ strongly depend on the choice of parameters $\lambda' \in \mathbb{R}^{1\times 1}$ and $\alpha \in \mathbb{R}^{1\times 1}$. λ' is non-negative and α can assume values from 0 to 1 (observe that for LASSO based model $\alpha = 1$). In the current study the parameter λ' was developed in a particular way so as to ensure the testing error is smaller than the magnitude of noise that is assumed to be known a priori. Accordingly, these models are developed as per the following steps (for LASSO $\alpha = 1$):

i. Using the training set, the values of λ' and α are obtained through an optimization Eq. 4.3 whose objective function is the maximization of λ', subject to the constraint that the prediction error for the testing set data is smaller than the magnitude of the noise ε assumed in the simulated data;

$$\begin{aligned} \underset{\lambda',\alpha}{\text{maximize}} \quad & \lambda' \\ \text{subject to} \quad & \boldsymbol{\beta} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \quad \left\{ \frac{1}{n} \left\| \mathbf{Q}\mathbf{p}_t - \mathbf{X}_{regressor}\boldsymbol{\beta} \right\|_2^2 + \lambda' \left((1-\alpha) \left\| \boldsymbol{\beta} \right\|_2^2 + \alpha \left\| \boldsymbol{\beta} \right\|_1 \right) \right\} \\ & \sqrt{\sum \left(\mathbf{Q}\mathbf{p}_t - \mathbf{X}_{regressor}\boldsymbol{\beta} \right)^2} \leq \varepsilon \\ & 0 < \alpha < 1 \end{aligned} \tag{4.3}$$

As λ increases, the number of non-zero regression coefficients decreases, thus resulting in an increasingly sparse vector β.

ii. If the testing and predictor error are bigger than the noise ε assumed in the simulated response data, the problem is infeasible. The magnitude of the noise can be generally obtained from replicates. If the noise related constraint is not included in the optimization problem then λ will be unbounded.

The second constraint from Eq. 4.3 ensures that the regression coefficient vector β will be such that the prediction error for the training set in the model calibration is smaller than the noise ε assumed in the simulated response data.

4.4.2 Results

A toy simulation example where $n < p$ is used here to compare the performance of PCR, PLS, LASSO and Elastic Net.

- As described in the Section 4.4, the toy example used in this case study presents 6 media formulations and four media components, e.g two major components (ψ_1 and ψ_2) and two minor components (ϕ_1 and ϕ_2). The values of ψ_i and ϕ_j concentrations are assumed from uniformly distributed random data, varying from 0 to 1, as given in equation 4.4;

$$\mathbf{X} = \begin{bmatrix} 0.8147 & 0.2784 & 0.9571 & 0.7922 \\ 0.9057 & 0.5468 & 0.4853 & 0.9594 \\ 0.1269 & 0.9575 & 0.8002 & 0.6557 \\ 0.9133 & 0.9648 & 0.1418 & 0.0357 \\ 0.6323 & 0.1576 & 0.4217 & 0.8491 \\ 0.0975 & 0.9705 & 0.9157 & 0.9339 \end{bmatrix} \tag{4.4}$$

- The regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$ is generated assuming linear factors, two-way interactions and squared terms of major components (ψ_i) and minor components (ϕ_j). Thus the number of observations is $n = 6$ and the number of regressors is $p = 14$ regressors;

- The simulated Qp response is obtained using the regressor matrix generated from the matrix in equation 4.4, following the function $Qp_t = 5\psi_1 + \phi_2 - 3\psi_1\psi_2$. The specific productivity response vector is given by $\mathbf{Qp}_t \in \mathbb{R}^{n \times 1}$. Observe that as described in the Methods section, the Qp response is obtained using regressors values mean-centered and normalized by its regressor standard deviation;

- Random noise ε, uniformly distributed, was added to the response vector. Different levels of noise with magnitudes of 8%, 10% and 15% of the full range of variation of the simulated Qp, respectively 0.6796, 0.8495 and 1.2742, were considered in this case study.

- The 6 media formulation described in equation 4.4 and its correspondent simulated Qp were used in the cross-validation to estimate the regression coefficients of the PCR, PLS, LASSO and Elastic Net based models;
- Other 60 media formulations (not shown), also composed by uniformly distributed random data, varying from 0 to 1 , and their corresponded simulated Qp are used to compare the sum of squared error (SSE) of the models for the purpose of model prediction.

Table 4.1 presents the results of the calibration errors when PCR, PLS, LASSO and Elastic Net based models are used. PCR and PLS based models consistently resulted in smaller training and testing errors than LASSO and Elastic Net models. Also, for LASSO model as the noise in the data decreases, the training error was larger in magnitude than the noise in the data. From Table 4.1 we concluded that the PLS based model is the best in terms of training error and the PCR model results in slightly better results in testing the media. Table 4.2 shows the sum of squared errors (SSE) for the 60 different media used for model prediction. Thus, for the particular case of $n < p$ the PLS and PCR models perform better than LASSO and Elastic Net. It is very important to emphasize that this result is specific to a case where $n < p$ and for input data that has been sampled from a uniformly random distribution.

To understand the superior results of PLS and PCR as compared to LASSO and Elastic Net models for this particular toy example, we compared the values of the regression coefficients obtained by these regression methods (β_{LASSO} and $\beta_{ElasticNet}$) with the regression coefficients used to generate the simulated Qp data (β). As can be seen in Table 4.3, the non-zero model regression coefficients found by the LASSO and Elastic Net approaches are not the ones used to generate the simulated data. In fact this type behavior for regularization based models, i.e. LASSO and Elastic Net, has been reported in the literature [82, 166, 92] for cases where the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$ presents less observations than the regressor variables, i.e. $n < p$ scenario. Although not shown for brevity, we have corroborated that if the number of observations is equal or larger than the number of regressor variables, e.g. n≥p where the total number of media observations is 14 or more, LASSO and Elastic Net would be able to find exactly the same important coefficient regressors (β_{LASSO} and $\beta_{ElasticNet}$) as the ones used to define the simulated data (β). Thus, the results of this toy example are very specific to the case of $n < p$ which happen to be the common situation for cellular medium design where the number of experiments is limited. For instance, it has been found that for cases where $n < p$, the models based on LASSO and Elastic Net do not present a unique solution. For those cases it has been shown that the exact model coefficients can only be recovered for particular regressor

matrix designs that satisfy a mathematical condition referred to as Restricted Isometry Property (RIP) [24, 156]. However, it is numerically challenging to satisfy this condition. Moreover, when the experiments are based on blending of master media as done in the industry, this condition could not be satisfied due to high correlation resulting from the blending of formulations.

Table 4.1: Training and testing error for the cross-validation when a random noise ε proportional to 8%, 10% and 15% to the range of the simulated Qp, respectively 0.6796, 0.8495 and 1.2742, is considered.

	CROSS-VALIDATION ERROR					
Noise in the data	0.6796		0.8495		1.2742	
Error	Train	Test	Train	Test	Train	Test
PCR	0.1571	0.2324	0.1928	0.2792	0.2828	0.3975
PLS	0.0949	0.2793	0.1142	0.3177	0.1717	0.4488
LASSO	0.7433	0.6796	0.9124	0.8236	0.9618	0.8062
Elastic Net	0.4264	0.3963	0.4548	0.4216	0.5352	0.4670

Table 4.2: SSE presented by the sixty media formulation when a random noise ε proportional to 8%, 10% and 15% to the range of the simulated Qp, respectively 0.6796, 0.8495 and 1.2742, is considered.

	SSE prediction		
Noise in the data	0.6796	0.8495	1.2742
PCR	34.4806	36.8596	43.7585
PLS	48.1961	53.4652	66.2500
LASSO	114.7288	154.9956	172.6365
Elastic Net	66.3108	73.3360	90.3790

Furthermore, the relative accuracy of the models is of particular significance for model based optimization. For example, considering the possible minimum level concentration equal to 0 and the possible maximum level concentration equal to 1, the maximum simulated Qp according to the defined function $Qp_t = 5\psi_1 + \phi_2 - 3\psi_1\psi_2$ would be obtained when the mean-centered and normalized values of the decision variables at the optimum are $\psi_1 = 1.1074$, $\psi_2 = -1.7421$ and $\phi_2 = 0.8563$ and maximum productivity it is equal to 2.5743. However, using these concentration levels, LASSO and Elastic Net models are not able to predict Qp values that are larger than the ones predicted by PCR and PLS models. While the predicted Qp for LASSO and Elastic Net models are 0.0420 and 1.0029

respectively, the Qp values predicted by PCR and PLS are 2.1896 and 1.8138, i.e. much closer to the true optimum. For this case study, the reason for these results is that the term containing the interaction $\psi_1\psi_2$ vanishes in both the LASSO and Elastic Net models while this term is very important about the optimum.

Table 4.3: Comparison among the coefficient regression used to generate the simulated Qp data and the coefficients regression given by LASSO and Elastic Net based model, assuming a noise proportional to 10% of the range of the simulated response values, e.g. noise equal to 0.8495.

Coefficient	Regressor	$\boldsymbol{\beta}$	$\boldsymbol{\beta}_{LASSO}$	$\boldsymbol{\beta}_{ElasticNet}$
β_1	ψ_1	5	0	0.0137
β_2	ψ_2	0	-0.0033	-0.1563
β_3	ϕ_1	0	0	0
β_4	ϕ_2	1	0	0.0026
β_5	$\psi_1\psi_2$	-3	0	0
β_6	$\psi_1\phi_1$	0	0	0.0836
β_7	$\psi_1\phi_2$	0	6.9391e-15	0.1408
β_8	$\psi_2\phi_1$	0	0	-0.0245
β_9	$\psi_2\phi_2$	0	0	-4.6432e-05
β_{10}	$\phi_1\phi_2$	0	0	0
β_{11}	ψ_1^2	0	0	0
β_{12}	ψ_2^2	0	-0.0303	-0.1697
β_{13}	ϕ_1^2	0	0	0
β_{14}	ϕ_2^2	0	0	0.0149

4.5 Case Study 2: Development of new media formulations using all observations from the blending of the master media

4.5.1 Methods

This case study uses the full data set of media formulations provided by the industrial collaborator, given by matrix $\mathbf{X} \in \mathbb{R}^{n \times p_o}$, with $n = 80$ media and $p_o = 90$ chemical elements (among major and minor components). To generate the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$

linear factors, two-way interactions and squared terms are considered, leading to a high-dimensional $n < p$ data with 80 rows ($n = 80$) and 4185 columns ($p = 4185$). The 80 media formulations are developed based on current industrial practice of blending pre-defined formulations (master formulations) that have resulted in good outcomes, e.g. high productivity, in earlier experiments. These formulations are blended with each other in specific proportions for further experimentation and optimization.

Data of monoclonal antibody Qp and the concentration of 4 key metabolites, i.e. glucose, glutamate, ammonia and lactate, at end of the culture experiments are also available. These 4 metabolites are of key importance in mammalian cell cultures and therefore they are commonly monitored during culture. Glucose and glutamate are key nutrients and ammonia and lactate are by-products that are often inhibitory when accumulating to high levels [22]. We assume that the output of interest for the current study is the mAb specific productivity at the end of the simulated perfusion operation (Qp_t), which is a function of the concentration of major components (ψ) and minor components (ϕ) in the initial medium formulation, as per the following equation:

$$Qp_t = f\left(\psi_i, \phi_j\right) \tag{4.5}$$

where ψ_i is the medium formulation concentration of the major component i and ϕ_j is the medium formulation concentration of the minor component j. Assuming the Qp is a function of linear factors, two-way interaction and squared terms of major components (ψ) and minor components (ϕ) medium formulation concentrations, the Eq. 4.5 can be re-written as Eq. 4.1. All concentrations' values used in the model are mean centered and thus no intercept term is present in the regression models. Also, the concentrations are normalized by the standard deviation of each component based on the different media formulations that are considered for model calibration.

A key challenge for calibrating a regression model based on samples obtained from blending of pre-defined media is that there exist large correlations among the inputs. The level of correlation in the input data is further exacerbated by considering interactions among the media components. In the presence of such correlations conventional least squares approaches are ineffective and statistical regression approaches that can deal with such correlations are needed [99, 144].

Since Case Study 1 indicated that LASSO and Elastic Net are less accurate than PCR and PLS for prediction in $n < p$ cases, only PCR and PLS models were considered for model based optimization. In both PCR and PLS based model regression, the same regressor matrix is used but two different sets of predicted responses are used separately in each regression modelling: i- simulated Qp data, and ii- industrial Qp data. When the

simulated Qp data is used, simulated data for glucose, glutamate, ammonia and lactate is also generated. The use of simulated output data served to check our results in terms of prediction ability and in terms of accuracy of the resulting optima. Then, the industrial data was used to check the ability of the regression models to predict the output measurements.

The objective of the robust optimization is to develop a new medium formulation capable of predicting monoclonal antibody Qp that is higher than the ones already given by the 80 original media formulation. In order to generate solutions that are robust to model error a robust optimization problem was solved. The variance of the new Qp to be used for robust optimization is given as follows:

For PCR based model:

$$\mathrm{var}\left(Qp_t\right) = \sqrt{\left(\left(\mathbf{X}_{regressor}\right)\left(\mathbf{XL}_{n_{PC}}\right)\right)^2 \mathrm{var}\left(\boldsymbol{\beta}_{PCR}\right)} \tag{4.6}$$

For PLS based model:

$$\mathrm{var}\left(Qp_t\right) = \sqrt{\left(\mathbf{X}_{regressor}\right)^2 \mathrm{var}\left(\boldsymbol{\beta}_{PLS}\right)} \tag{4.7}$$

where $Qp_t \in \mathbb{R}^{1\times 1}$ is the predicted specific productivity for the medium formulation, $\mathrm{var}\left(Qp_t\right) \in \mathbb{R}^{1\times 1}$ is the variance of the predicted specific productivity for the medium formulation, $\mathbf{X}_{regressor} \in \mathbb{R}^{1\times p}$ is the regressor matrix given by the medium formulation considering linear factors, two-way interactions and squared terms, $\mathbf{XL}_{n_{PC}} \in \mathbb{R}^{p\times n_{PC}}$ is the loadings from PCA considering n_{PC}, $\mathrm{var}\left(\boldsymbol{\beta}_{PCR}\right) \in \mathbb{R}^{n_{PC}\times 1}$ is the variance given by the coefficients for PCR regression, and $\mathrm{var}\left(\boldsymbol{\beta}_{PLS}\right) \in \mathbb{R}^{p\times 1}$ is the variance given by the coefficients for PLS regression.

Constraints are added to avoid that the decision variables, i.e. the concentrations of chemical components in the new medium formulation, result in too high or negative concentrations. The constraints on key metabolites, such as glutamate, glucose, lactate and ammonia, are chosen according to values provided by the industrial partner. These constraints aim to guarantee that the new medium formulation do not result in cultures where all glutamate and glucose are consumed or where lactate and ammonia are overproduced thus leading to inhibition of cell growth and/or productivity.

This robust optimization problem can generate 3 possible new media formulations $\mathbf{X}_{new} \in \mathbb{R}^{1\times p_o}$, according to the value of constant θ: if is equal to zero, the average prediction of $Qp_{t,new}$ is obtained, if is equal to -1 or $+1$, the lower or upper bound average prediction of $Qp_{t,new}$ is obtained, respectively.

Development of new media formulations using a PCR based model

i. Build the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$, with $n = 80$ and $p = 4185$ with the medium formulation data;

ii. Using the monoclonal antibody Qp (simulated data or industrial data), follow the steps described in Section 4.4.1 to generate the PCR based model.

iii. Additional PCR based models relating the concentration of the 4 key metabolites, i.e. glucose, glutamate, ammonia and lactate, to the initial media components are also obtained following the steps described in Section 4.4.1 from PCR regressions. By using these additional models it is possible to impose constraints on the concentrations of these 4 metabolites within the robust optimization problem as follows;

iv. Perform robust optimization as described by equations 4.8 to find the new medium composition.

$$
\begin{aligned}
\underset{\mathbf{X}_{new}}{\text{maximize}} \quad & Qp_{t,new} + \theta \left(\text{var}\left(Qp_{t,new}\right)\right) \\
\text{subject to} \quad & \mathbf{X}_{new,regressor} = \text{regressor matrix}(\mathbf{X}_{new}) \\
& Qp_{t,new} = \left(\mathbf{X}_{new,regressor}\right)\left(\mathbf{XL}_{n_{PC}}\right)\boldsymbol{\beta}_{PCR} \\
& \text{var}\left(Qp_{t,new}\right) = \sqrt{\left(\left(\mathbf{X}_{new,regressor}\right)\left(\mathbf{XL}_{n_{PC}}\right)\right)^2 \text{var}\left(\boldsymbol{\beta}_{PCR}\right)} \\
& Glu = \left(\mathbf{X}_{new,regressor}\right)\left(\mathbf{XL}_{n_{PC,Glu}}\right)\boldsymbol{\beta}_{PCR,Glu} \\
& Glc = \left(\mathbf{X}_{new,regressor}\right)\left(\mathbf{XL}_{n_{PC,Glc}}\right)\boldsymbol{\beta}_{PCR,Glc} \\
& Lac = \left(\mathbf{X}_{new,regressor}\right)\left(\mathbf{XL}_{n_{PC,Glc},Glc}\right)\boldsymbol{\beta}_{PCR,Lac} \\
& Amm = \left(\mathbf{X}_{new,regressor}\right)\left(\mathbf{XL}_{n_{PC,Amm}}\right)\boldsymbol{\beta}_{PCR,Amm} \\
& Qp_{t,new} \leq Qp_{t,max} \\
& Glu > 0 \\
& Glc > 0 \\
& Lac \leq Lac_{max} \\
& Amm \leq Amm_{max} \\
& \mathbf{X}_{min} \leq \mathbf{X}_{new} \leq \mathbf{X}_{max}
\end{aligned}
\tag{4.8}
$$

where $\mathbf{X}_{new} \in \mathbb{R}^{1 \times p_o}$ is new medium formulation, $\mathbf{X}_{new,regressor} \in \mathbb{R}^{1 \times p}$ is the regressor matrix of $\mathbf{X}_{new}$ considering linear factors, two-way interactions and squared terms, $Qp_{t,new} \in \mathbb{R}^{1 \times 1}$ is the predicted output for the new formulation, $\theta \in \mathbb{R}^{1 \times 1}$ is a constant

that can be -1, 0 or +1, var $(Qp_{t,new}) \in \mathbb{R}^{1\times1}$ is the variance of the predicted output for the new formulation, $\mathbf{XL}_{n_{PC}} \in \mathbb{R}^{p\times n_{PC}}$ is the loadings from PCA considering n_{PC}; n_{PC}, $n_{PC,Glu}$, $n_{PC,Glc}$, $n_{PC,Lac}$, and $n_{PC,Amm}$ are the number of principal components selected for, respectively, the response of interest, glutamate, glucose, lactate and ammonia, $\boldsymbol{\beta}_{PCR} \in \mathbb{R}^{n_{PC}\times1}$ is the vector of coefficients for PCR regression, $\boldsymbol{\beta}_{PCR,Glu} \in \mathbb{R}^{n_{PC}\times1}$ is the vector of coefficients for PCR regression of the glutamate concentration, $\boldsymbol{\beta}_{PCR,Glc} \in \mathbb{R}^{n_{PC}\times1}$ is the vector of coefficients for PCR regression of the glucose concentration, $\boldsymbol{\beta}_{PCR,Lac} \in \mathbb{R}^{n_{PC}\times1}$ is the vector of coefficients for PCR regression of the lactate concentration, $\boldsymbol{\beta}_{PCR,Amm} \in \mathbb{R}^{n_{PC}\times1}$ is the vector of coefficients for PCR regression of the ammonia concentration, $Qp_{t,max} \in \mathbb{R}^{1\times1}$ is the upper bound for the output of interest prediction, $Lac_{max} \in \mathbb{R}^{1\times1}$ is the upper bound for the lactate concentration, $Amm_{max} \in \mathbb{R}^{1\times1}$ is the upper bound for the ammonia concentration, $\mathbf{X}_{min} \in \mathbb{R}^{1\times p_o}$ and $\mathbf{X}_{max} \in \mathbb{R}^{1\times p_o}$ are the lower and upper bounds for the new medium composition values, respectively.

Development of new media formulation using a PLS based model

i. Build the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n\times p}$, with $n = 80$ and $p = 4185$ with the data provided by the industrial collaborator;

ii. Using the monoclonal antibody Qp (simulated data or the industrial data), follow the steps described in Section 4.4.1 to generate the PLS based model.

iii. Additional PLS based models relating the concentration of the 4 key metabolites, i.e. glucose, glutamate, ammonia and lactate, to the initial media components are also obtained following the steps described in Section 4.4.1 from PLS regressions.

iv. Perform robust optimization as described by equations 4.9 to find the new medium composition.

$$
\begin{aligned}
\underset{\mathbf{X}_{new}}{\text{maximize}} \quad & Qp_{t,new} + \theta \left(\text{var}\left(Qp_{t,new}\right)\right) \\
\text{subject to} \quad & \mathbf{X}_{new,regressor} = \text{regressor matrix}(\mathbf{X}_{new}) \\
& Qp_{t,new} = \left(\mathbf{X}_{new,regressor}\right)\boldsymbol{\beta}_{PLS} \\
& \text{var}\left(Qp_{t,new}\right) = \sqrt{\left(\mathbf{X}_{new,regressor}\right)^2 \text{var}\left(\boldsymbol{\beta}_{PLS}\right)} \\
& Glu = \left(\mathbf{X}_{new,regressor}\right)\boldsymbol{\beta}_{PLS,Glu} \\
& Glc = \left(\mathbf{X}_{new,regressor}\right)\boldsymbol{\beta}_{PLS,Glc} \\
& Lac = \left(\mathbf{X}_{new,regressor}\right)\boldsymbol{\beta}_{PLS,Lac} \\
& Amm = \left(\mathbf{X}_{new,regressor}\right)\boldsymbol{\beta}_{PLS,Amm} \\
& Qp_{t,new} \leq Qp_{t,max} \\
& Glu > 0 \\
& Glc > 0 \\
& Lac \leq Lac_{max} \\
& Amm \leq Amm_{max} \\
& \mathbf{X}_{min} \leq \mathbf{X}_{new} \leq \mathbf{X}_{max}
\end{aligned}
\tag{4.9}
$$

where $\mathbf{X}_{new} \in \mathbb{R}^{1\times p_o}$ is new medium formulation, $\mathbf{X}_{new,regressor} \in \mathbb{R}^{1\times p}$ is the regressor matrix of $\mathbf{X}_{new}$ considering linear factors, two-way interactions and squared terms, $Qp_{t,new} \in \mathbb{R}^{1\times 1}$ is the predicted output for the new formulation, $\theta \in \mathbb{R}^{1\times 1}$ is a constant that can be -1, 0 or +1, $\text{var}\left(Qp_{t,new}\right) \in \mathbb{R}^{1\times 1}$ is the variance of the predicted output for the new formulation, $\boldsymbol{\beta}_{PLS} \in \mathbb{R}^{p\times 1}$ is the vector of coefficients for PLS regression, $\boldsymbol{\beta}_{PLS,Glu} \in \mathbb{R}^{p\times 1}$ is the vector of coefficients for PLS regression of the glutamate concentration, $\boldsymbol{\beta}_{PLS,Glc} \in \mathbb{R}^{p\times 1}$ is the vector of coefficients for PLS regression of the glucose concentration, $\boldsymbol{\beta}_{PLS,Lac} \in \mathbb{R}^{p\times 1}$ is the vector of coefficients for PLS regression of the lactate concentration, $\boldsymbol{\beta}_{PLS,Amm} \in \mathbb{R}^{p\times 1}$ is the vector of coefficients for PLS regression of the ammonia concentration, $Qp_{t,max} \in \mathbb{R}^{1\times 1}$ is the upper bound for the output of interest prediction, $Lac_{max} \in \mathbb{R}^{1\times 1}$ is the upper bound for the lactate concentration, $Amm_{max} \in \mathbb{R}^{1\times 1}$ is the upper bound for the ammonia concentration, $\mathbf{X}_{min} \in \mathbb{R}^{1\times p_o}$ and $\mathbf{X}_{max} \in \mathbb{R}^{1\times p_o}$ are the lower and upper bounds for the new medium composition values, respectively.

4.5.2 Results

Following the results in Case Study 1 and the limitations identified for regularization methods in the $n < p$ case, the models used in Case Study 2 were based on either PCR or PLS regressions. First a simulated model is created in order to test the methodology. The input data corresponds to the one provided by our industrial partner and it is based on the blending of master media as explained in the Methods' section while the outputs corresponding to Qp, glutamate, glucose, lactate and ammonia data are simulated.

Modelling and optimization of simulated response data

The simulated model for Qp, glutamate, glucose, lactate and ammonia was obtained as follows:

- The regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$, used was the one built with the blending media formulations (composed by 20 major components ψ, and 70 minor components ϕ) provided by the industrial collaborator, assuming linear factors, two-way interactions and squared terms, with $n = 80$ and $p = 4185$;
- A sparse regression coefficient vector with few random regressors defined, as presented by Table 4.4, is used to simulate the output of interest Qp_t, e.g. mAb specific productivity at the end of the simulated perfusion operation. These coefficients are related to the importance of certain regressors from $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$. In this case study, 9 coefficients were non-zero (see Table 4.4);
- Four additional sparse vectors were defined to generate data for glutamate, glucose, lactate and ammonia. Other significant regressors than the ones defined for Y generation were used to define the coefficient sparse vectors $\boldsymbol{\beta}_{Glu}$, $\boldsymbol{\beta}_{Glc}$, $\boldsymbol{\beta}_{Lac}$ and $\boldsymbol{\beta}_{Amm}$, all $\in \mathbb{R}^{p \times 1}$ (see Tables 4.5, 4.6, 4.7, and 4.8);
- With $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$ mean-centered and normalized and the coefficients regression $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ defined, the output of interest $\mathbf{Qp}_t \in \mathbb{R}^{n \times 1}$ was obtained by $\mathbf{Qp}_t = \mathbf{X}_{regressor}\boldsymbol{\beta}_o$. Random noise proportional to 15% of the range of simulated Qp were added to $\mathbf{Qp}_t$;
- Output data for metabolites **Glu**, **Glc**, **Lac**, and **Amm** all $\in \mathbb{R}^{n \times 1}$ were obtained by $\mathbf{Glu} = \mathbf{X}_{regressor}\boldsymbol{\beta}_{Glu}$, $\mathbf{Glc} = \mathbf{X}_{regressor}\boldsymbol{\beta}_{Glc}$, $\mathbf{Lac} = \mathbf{X}_{regressor}\boldsymbol{\beta}_{Lac}$, and $\mathbf{Amm} = \mathbf{X}_{regressor}\boldsymbol{\beta}_{Amm}$. Random noise proportional to 15% of the range of metabolite output were also added to the output metabolites' concentrations.

Table 4.4: Regression coefficients used for simulated response of interest (**Qp**) data - Case Study 2. Note that the other coefficients not described in the table have zero value.

Coefficient Vector β_o	Value	Corresponding Variable
β_{80}	0.3333	ϕ_{60}
β_{111}	-0.3333	$\psi_1\phi_2$
β_{272}	0.2222	$\psi_3\psi_8$
β_{520}	0.6667	$\psi_5\phi_{65}$
β_{685}	0.6667	$\psi_7\phi_{63}$
β_{1039}	-0.1111	$\psi_{12}\phi_{17}$
β_{2112}	0.1111	$\phi_7\phi_{40}$
β_{3555}	0.8889	$\phi_{37}\phi_{58}$
β_{3895}	-0.2222	$\phi_{50}\phi_{60}$

Table 4.5: Regression coefficients used for simulated response of glutamate (**Glu**) data - Case Study 2. Note that the other coefficients not described in the table have zero value.

Coefficient Vector β_{Glu}	Value	Corresponding Variable
β_1	1	ψ_1
β_{15}	-1	ψ_{15}
β_{3000}	2	$\phi_{23}\phi_{56}$
β_{4022}	1	$\phi_{58}\phi_{63}$

Table 4.6: Regression coefficients used for simulated response of glucose (**Glc**) data - Case Study 2. Note that the other coefficients not described in the table have zero value.

Coefficient Vector β_{Glc}	Value	Corresponding Variable
β_{12}	-3	ψ_{12}
β_{20}	2	ψ_{20}
β_{2856}	8	$\phi_{20}\phi_{56}$

Table 4.7: Regression coefficients used for simulated response of lactate (**Lac**) data - Case Study 2. Note that the other coefficients not described in the table have zero value.

Coefficient Vector β_{Lac}	Value	Corresponding Variable
β_{46}	5	ϕ_{26}
β_{272}	10	$\psi_3\psi_8$
β_{1685}	-1	$\phi_1\phi_6$
β_{2039}	-5	$\phi_6\phi_{30}$

Table 4.8: Regression coefficients used for simulated response of ammonia (**Amm**) data - Case Study 2. Note that the other coefficients not described in the table have zero value.

Coefficient Vector β_{Amm}	Value	Corresponding Variable
β_{30}	-5	ϕ_{10}
β_{32}	6	ϕ_{32}
β_{852}	1	$\psi_9\phi_{67}$
β_{1022}	2	$\psi_1\psi_{20}$
β_{2489}	-2	$\phi_{13}\phi_{60}$

Table 4.9: Cross-validation error for PCR and PLS based models using simulated Qp data - Case Study 2.

	Error Cross-Validation	
	Training	Testing
PCR	0.4931	0.6433
PLS	0.2614	0.7279

The magnitude of the noise is 15% of the full range of variation of the simulated Qp, i.e. 0.89, which is at the same level as the noise present in the industrial response data. Aiming to achieve a training and testing error of the order of the noise, 14 principal components were considered for the PCR based regression model. Similarly, the number of principal components selected for the individual models of glutamate, glucose, lactate and ammonia was also equal to 14. For the PLS based regression model, the number of latent variables considered in all the models was equal to 12. These number of principal components or latent variables were chosen based on the training and testing methods explained in the Methods' section. Table 4.9 presents the training and testing errors for Qp for both PCR and PLS based approaches. Both PCR and PLS based models resulted in cross-validation errors smaller than the magnitude of the noise.

A set of 15 different media was used for testing the prediction ability of the PCR and PLS models. Clearly, the prediction accuracy largely depends on the projections of the data to be predicted onto the principal components for PCR or the principal latent variables for PLS of the calibrated models.

To test this statement two data sets were tried for the PCR model: set 1- data set that has large scores for the principal components identified during calibration and set 2- data set for which the scores for the first principal components are small and large along the less significant components. Accordingly for set 1 the media formulations for the prediction set were obtained using the PCA scores and the PCA loadings of the original 80 media formulation. However, the first ten scores principal components (which explain approximately 87% of the input variability) were used with higher weight than the other principal components. The aim was to create a media prediction set which the input was around the region where the model is valid. For the described prediction set, the SSE and prediction error for the PCR based model were 13.10 and 0.9346. The worst prediction error of the PCR model is just 4% larger than the noise magnitude. On the other hand, set 2 was generated by reducing the weight of the scores of the first principal components and increasing the weight on the last principal components scores (the ones which account for less variability of the input data). For set 2 the SSE and prediction error are 21.83 and 1.2066. Thus, as expected, as the data is farther apart from the data used for model training, the prediction is less accurate and this will have a clear impact on optimization results.

Similar testing of the prediction accuracy can be done for a PLS model. Also, two data sets were tried for PLS model prediction, set 1 with a data set that has large scores for the first latent variables and set 2 with a data set for which the smaller scores for the first latent variables and large scores for the latest latent variables. Set 1 presented SSE and prediction error for the PLS model equal to 11.56 and 0.8777, observe that the prediction error fits the noise in the data. For set 2 the SSE and prediction error were equal to 91.24 and 2.4664. Observe that for set 2 the prediction error is much larger than the noise present in the data. Again, as expected and as happened for PCR model, as the data is farther apart from the data used for model training, the prediction is less accurate.

Based on the observation that the prediction accuracy largely depends on the proximity of the data to the original data, as given by the projection of the new data onto the principal components or principal latent variables, it was decided to impose constraints on the outputs in the optimization problem. The objective was to ensure that the calculated optimum will not be located too far from the data used for model calibration. For the robust optimization based on either the PCR or PLS models, $Qp_{t,max}$, Lac_{max} and Amm_{max} were considered as 30% of, respectively, the maximum Qp, lactate concentration, and ammonia

concentration obtained by the simulated data given by the 80 media observations. For each robust optimization problem three new media formulations were found: one for $\theta = -1$ (lower bound prediction), one for $\theta = 0$ (average prediction), and the last one for $\theta = 1$ (upper bound prediction).

Figure 4.1 shows the maximum Qp obtained with the original media formulation (in red) compared with the predicted and true Qp for the new media obtained through the optimization when θ is equal 0, -1 and 1 for the PCR and PLS regression models. The predicted Qp results, indicated in light blue and yellow, were obtained by substituting the values of each optimal medium formulation into the respective PCR and PLS models. The corresponding true Qp results (in dark blue and orange) were obtained by using the optimal media formulations found by PCR and PLS models into the true simulated model (given by the coefficients described in Table 4.4). It can be seen that with either the PCR and PLS regression models it is possible to find new media compositions that can predict an Qp about 30% larger than the maximum Qp found among the 80 original media formulation, except for the case of the lower bound predicted with the PLS model $\theta = -1$, see Figure 4.1 (yellow bar for PLS $\theta = -1$). The maximum Qp found among the 80 original media formulation is 4.0861 (red bar). For the PCR based model, the Qp predictions (light blue bars) found were 5.3187 ($\theta = 0$), 4.9965 ($\theta = -1$) and 7.3182 ($\theta = 1$). For the PLS based model, the Qp predictions (yellow bars) found were 5.3187 ($\theta = 0$), 3.8144 ($\theta = -1$) and 7.0619 ($\theta = 1$).

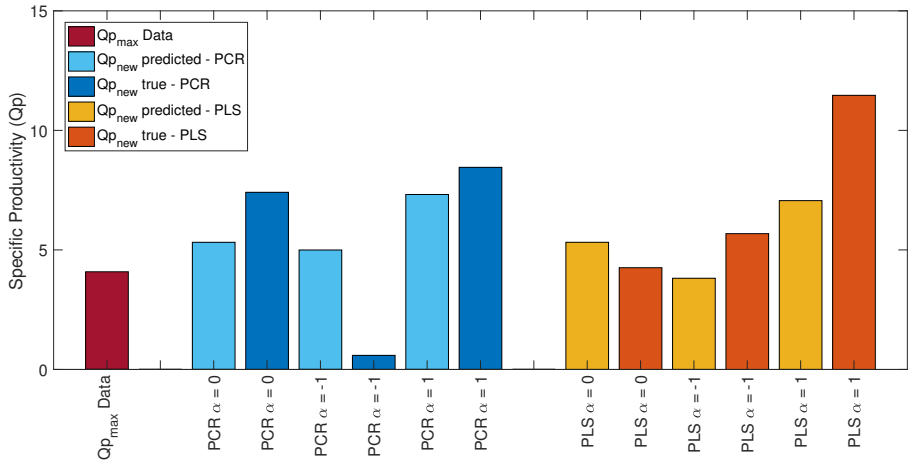


Figure 4.1: Maximum specific productivity (Qp) given by the original media formulation compared with the predicted and simulated specific productivity for the new media obtained when θ is equal 0, -1 and 1 for PCR and PLS regression models.

The true Qp results for the new media found by the robust optimization for PCR and PLS models confirms that all these new media resulted in Qp values that are larger than the maximum value obtained with the 80 initial media, except for the medium found by the robust optimization considering the PCR model and $\theta = -1$ which corresponds to a lower bound, see Figure 4.1 (dark blue bar for PCR $\theta = -1$).

The predicted and true Qp values found for the new media, for PCR and PLS based approaches, are presented in Table 4.10. The differences between the true and predicted values are smaller with the PLS based model thus indicating that this model can better predict the Qp than the PCR based model. On the other hand, although the PCR model has slightly lower prediction accuracy, it leads to a higher Qp optimum as compared to the optimum obtained with the PLS model. The explanation is that PCR can better explain the variability of the input variables space as compared to PLS that focuses on the region of the input space that better explains the variability of the output which limits the ability of PLS based optimization to find the global optimum in the input space.

Table 4.10: Predicted and true Qp values found for the new media, for PCR and PLS based approaches - Case Study 2

	$\theta = 0$		$\theta = -1$		$\theta = 1$	
	Predicted	True	Predicted	True	Predicted	True
PCR	5.3187	7.4101	4.9965	0.5890	7.3182	8.4544
PLS	5.3187	4.2563	3.8144	5.6805	7.0619	11.4664

Note that the differences of predicted and simulated Qp given by the robust optimization using PCR regression with $\theta = 0$ are only slightly larger than the noise (0.894) which is the desired outcome of the robust optimization [86].

It was also observed (not shown in this work) that when $\theta = -1$ or $\theta = 1$,corresponding to the lower and upper bounds of the optimum respectively, for both PCR and PLS based models the robust optimizations found the components in the new media formulations to be at their extreme levels for most of the medium components. However, this is not the case for the average value $\theta = 0$.

Figure 4.2 shows the optimum medium composition found from the robust optimization for PCR and PLS when $\theta = 0$. The composition for most of the medium components are generally larger from the PCR based optimization except for some components, such as ψ_2, ψ_4, ϕ_4 ϕ_6, ϕ_{15}, ϕ_{24} and ϕ_{41}. We also observed that as the difference in the new formulation of media components obtained through PCR and PLS increases, that component tends to be not significant in the true model. For example, medium components ϕ_{15}, ϕ_{18}, ϕ_{34} and ϕ_{70} are not present in the true model that generated the Qp data (see Table 4.4).

For comparison, we searched for the true optimum i.e. when the true model was used to predict the Qp along with its constraint functions for glutamate, glucose, ammonia and lactate. The optimum Qp was equal to the one calculated by the PCR model (5.3187) but this is not surprising since this value is at the constraint on Qp used in the optimization, i.e. 30% higher than the highest value of Qp obtained from the original 80 media. If such constraint is not imposed the optimal productivity with the true model is much higher. However, as we shown above, the constraint is necessary so as to ensure validity of the PCR and PLS models. On the other hand it was found that for the same Qp value the input formulation was somewhat different from the optimal formulations predicted by the PCR and PLS models. These differences can be explained by the true optimum having high scores on the later principal components or principal latent variables of the PCR and PLS model respectively, which accounts for very low variability of the data and are not in the region of validity of PCR and PLS models. It was observed that the PCR scores projection of the optimum formulation obtained when the true model was used to predict the Qp presented scores values out of the PCR model projection space (defined in the calibration) for most of the principal components. For example, while the projection values of first 12 principal components are found into the PCR based model scores range, the last 12 principal components present scores values totally out of the calibration region.The same behavior was observed with the latent variables for the PLS based model.

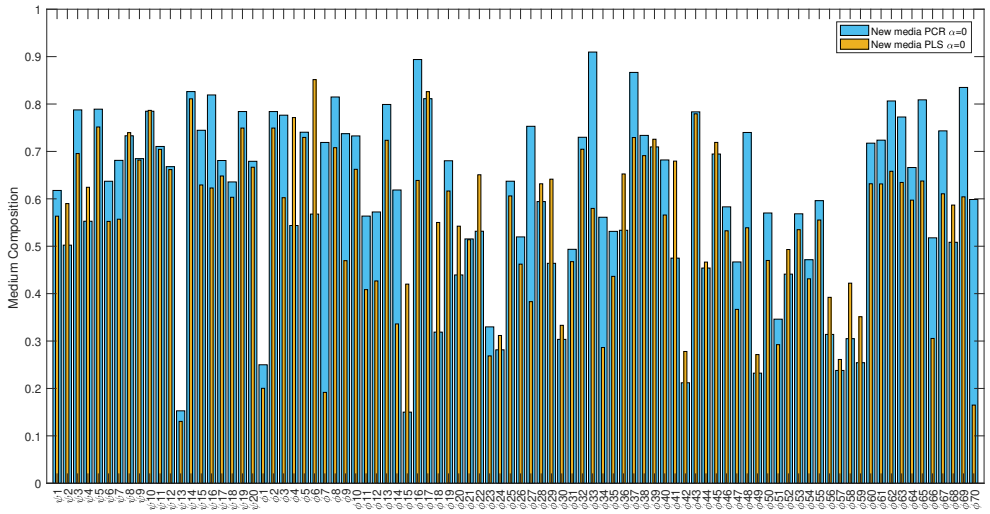


Figure 4.2: Optimal medium components concentration (X_{new}) found by robust optimization when $\theta = 0$ for PCR and PLS based models.

Modelling results for industrial output data

As described in the methods section 4.5, the blending of the master media in different proportions generated 80 media that were used in this case study for modelling. The resulting model was then used for finding an optimal media. Three out of the 80 media are replicates and, therefore, they serve to estimate the level of the noise that was found to be equal to 0.80 pg/cell. Eight randomly chosen media formulations, were used for testing and were not used for model training or validation. The other 72 media were used for model calibration (testing and training). In this part of the work only the formulations that were found by robust optimization with the PCR model were experimentally tested. Optimal medium formulations found by PLS model approach will be experimentally tested in future work. The reason for testing the PCR based formulations was based on our earlier results with synthetic models that indicated better results with PCR.

Initially the model was developed without including constraints on glutamate, glucose, lactate and ammonia. Thus, the goal was the maximization of Qp subject to constraints of Qp maximum value and medium components composition bounds, as shown in equation 4.10.

$$\begin{aligned} \underset{\mathbf{X}_{new}}{\text{maximize}} \quad & Qp_{t,new} + \theta \left(\text{var}\left(Qp_{t,new}\right)\right) \\ \text{subject to} \quad & \mathbf{X}_{new,regressor} = \text{regressor matrix}(\mathbf{X}_{new}) \\ & Qp_{t,new} = \left(\mathbf{X}_{new,regressor}\right)\left(\mathbf{X}\mathbf{L}_{n_{PC}}\right)\boldsymbol{\beta}_{PCR} \\ & \text{var}\left(Qp_{t,new}\right) = \sqrt{\left(\left(\mathbf{X}_{new,regressor}\right)\left(\mathbf{X}\mathbf{L}_{n_{PC}}\right)\right)^2 \text{var}\left(\boldsymbol{\beta}_{PCR}\right)} \\ & Qp_{t,new} \leq Qp_{t,max} \\ & \mathbf{X}_{min} \leq \mathbf{X}_{new} \leq \mathbf{X}_{max} \end{aligned} \tag{4.10}$$

The new medium composition obtained to be referred as "New A", assumed $Qp_{t,max}$ equal 30% of the maximum Qp obtained by the 80 media experiments. The reason for imposing an upper bound constraint on productivity was to avoid large extrapolations of the model into a region of the variable space that was not used for model calibration due to lack of data in that region. In this initial approach, 25 principal components were used in the PCR model for Qp, which explain 96% of the total input variance. The Qp predicted error for the training and testing sets were 0.46 pg/cell and 0.97 pg/cell, respectively. The prediction error for the 8 media used for model prediction were equal to 0.63 pg/cell.

A second new optimum medium to be referred as "New B", was obtained using the full robust optimization as described in equation 4.8. The $Qp_{t,max}$ was also set as 30%

of the maximum Qp obtained by the 80 media experiments. In this second approach, in order to calibrate the model such as the prediction error will be equal or smaller than the magnitude of the noise (+-0.80 pg/cell), 14 principal components were used in the PCR based regression models for the Qp, glutamate, glucose, lactate and ammonia. The 14 principal components were able to explain 91.2% of the input variability. The Qp predicted error for the training and testing sets were 0.54 pg/cell and 0.78 pg/cell, respectively. The prediction error for the 8 media used for model prediction were equal to 0.69 pg/cell. Observe that for both media "New A" and "New B" , the prediction errors were smaller than the estimated noise (+-0.80 pg/cell) thus confirming the ability of the PCR model to predict data not used for model calibration.

Figure 4.3 shows a comparison of the cell density, mAb and Qp obtained with "New A" and "New B", and with four different media used by the industrial partner, "Run 1", "Run 9", "Run 53" and "Run 81". "Run 9" and "Run 53" are two out of eleven master media and "Run 1" is a weighted combination of the 11 master media; "Run 9", "Run 53" and "Run 1" were present in the 80 media experiments. The medium "Run 81" is a new proportion combination of the master media that is not one of the 80 media considered in our modelling exercise.

Figure 4.3 shows that in the last day of the cell culture (day 4) the proposed optimum media presented a Qp value equal to 10.30 pg/cell for "New A" and 10.21 pg/cell for "New B", approximately 20% smaller than the Qp value predicted by the PCR model approach, which was 12.24. However, this new run of experiments presented a experimental data noise of +-1.55 pg/cell. Therefore, the measured values are not significantly different from the ones predicted by PCR model. In general, the performance for cell density, mAb and Qp over the days was very similar for "New A" and "New B".

Although the measured Qp values obtained by Run 1 (11.79 pg/cell), Run 9 (11.39 pg/cell), Run 53 (10.69 pg/cell) and Run 81 (10.68 pg/cell) were relatively high as compared to the "New A" and "New B" media, the latter two optimized media presented consistently higher values of cell density and mAb through the days. This higher growth was of great interest for the company due its potential to shorten perfusion operations. Larger growth has the potential to shorten the batch period needed for reaching a certain density before perfusion is started.

Figure 4.4 compares the new media compositions, "New A" and "New B", with medium "Run 9", which is one of the initial 80 media experiment. "Run 9" medium corresponds to one of the 11 master media and it is considered a good performing media by our industrial partner, presenting a high Qp value in the experiments. It can be noticed that the two media proposed by the PCR model exhibits component concentrations that are significantly

different from the ones used by "Run 9" but they are also resulting in cell cultures with high Qp values. For instance, medium components that were not present in the "Run 9" medium presents a significant concentration in the "New 1", such as the minor components ϕ_4, ϕ_6, ϕ_{18}, ϕ_{20}, ϕ_{36}, ϕ_{45} and ϕ_{55}. The presence of these medium components may explain the consistent results for high growth and mab presented by the "New A" and "New B" medium. Figure 4.4 also shows that the new proposed media "New A" and "New B" are very similar, with few components varying over +-40% in composition, such as ϕ_6, ϕ_{10}, ϕ_{12}, ϕ_{19}, ϕ_{24}, ϕ_{29} and ϕ_{70}.

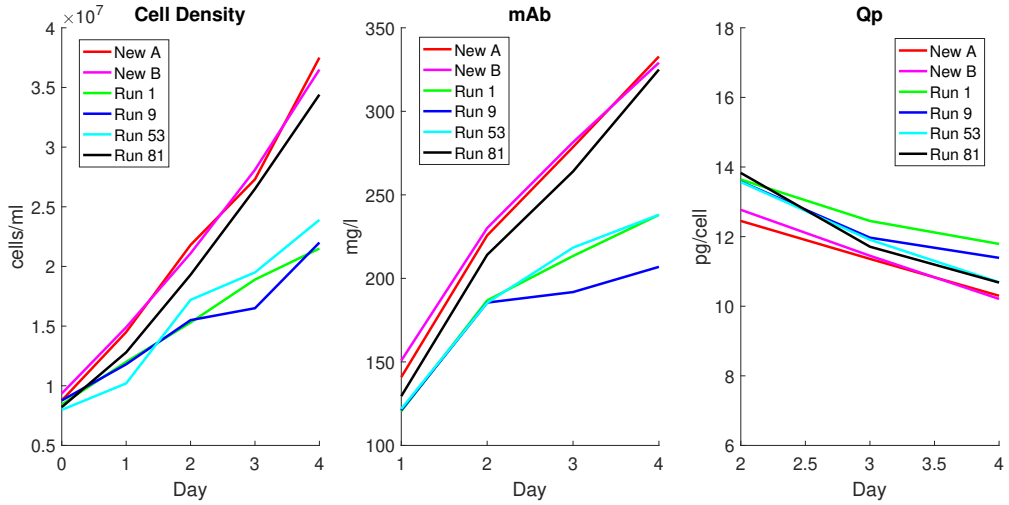


Figure 4.3: Daily experimental results for cell density, mAb and Qp for media "New A", "New B", "Run 1", "Run 9", "Run 53", and "Run 81".

It is also important to notice that the new media experiments were run with most of the operating conditions of the original 80 media experiments but with a different RPM value. While the initial 80 experiments were done with 200 RPM, the new proposed optimum media ("New A" and "New B") and the new batches using media from "Run1", "Run 9", "Run 53" and "Run 81" were all run at 230 RPM. Higher RPM generally favours mass transfer with resulting higher cellular growth, and therefore, affects mAb and Qp results.

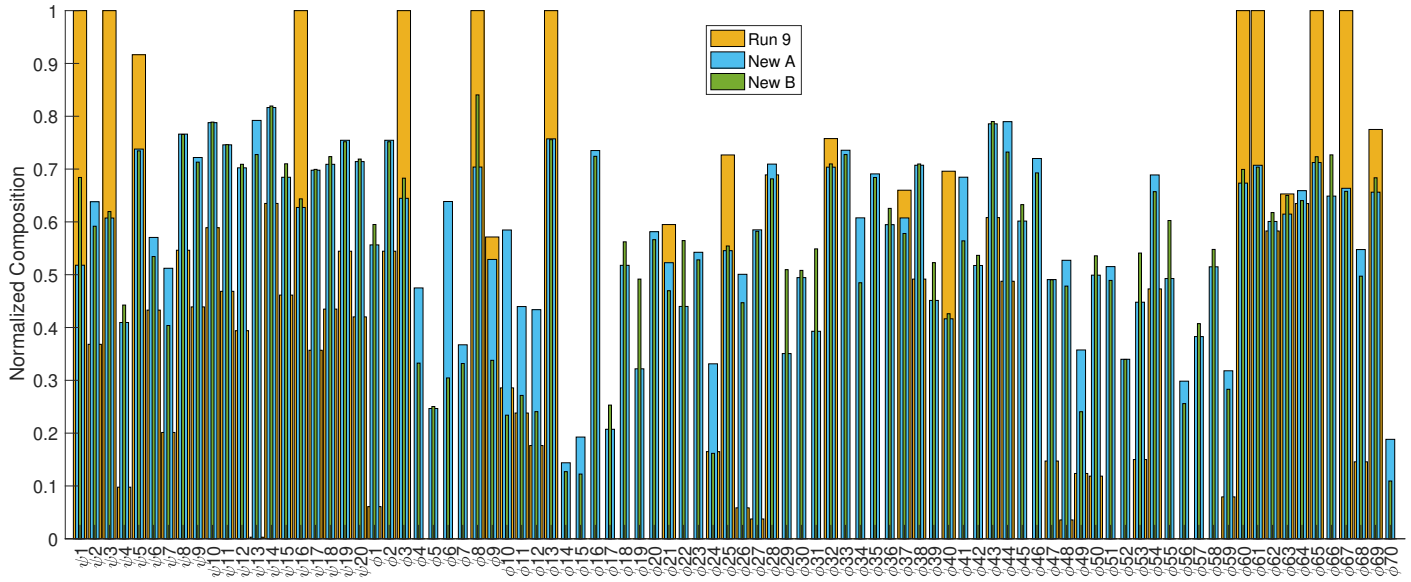


Figure 4.4: Comparison of optimal media formulations found by robust optimization in Case Study 2, "New A" and "New B", and industrial master media "Run 9" when $\theta = 0$ for PCR based models.

To further improve the productivity results we conducted a new round of modelling with old and new data followed by optimization based on the new models. Accordingly,the coefficients of the PCR model were re-calibrated and other two new media formulation was proposed. However, in this second round, new constraints were added to the robust optimization problem described in equation 4.8. First, a cell density (CD) constraint was incorporated in the optimization model in addition to the constraints for Qp, glucose, lactate, glutamate and ammonia. 14 principal components were used to describe the cell density data based on model training and validation. The aim was to understand how cell density constraints affect the final results of Qp for the new proposed media.

In addition, to avoid large deviations from the region of the input space where training data was available, we incorporated constraints on principal scores of the input data. Correspondingly, we incorporated constraints on the first 5 scores given by the principal components, XS1, XS2, XS3, XS4 and XS5. The first 5 scores represents 65.3% of the input variability. The constraints were set in a way that the new proposed media would be within the region of the input space where the media "Run 1", "Run 9", "Run 53" and "Run 81" resided.

Following optimization two new media were obtained to be referred as "New C" and "New D". Both media were obtained with same constraints Qp and metabolites constraints but differed only in terms of the CD constraint value. In terms of mean centered and normalized CD values, the media "New C" used CD > 0 while media "New D" used CD > 1. Thus, in the former the objective was to have CD slighter than the average whereas in the latter the goal was to obtain cell density larger than the maximum measured in past experiments. Figure 4.5 shows a comparison of the cell density, mAb and Qp obtained with the new media, "New C" and "New D", and with the three media chosen by the industrial partner, "Run 1", "Run 9" and "Run 81".

At the last day of the cell culture (day 4) the proposed optimum media presented a Qp value equal to 10.66 pg/cell for "New C" and 10.51 pg/cell for "New D", approximately 15% smaller than the Qp predicted by the PCR model approach, which was 12.24 pg/cell. On the other hand based on the estimate of experimental noise of ± 1.55 pg/cell, the results with "New C" and "New D" were similar to the ones obtained with Run 1 (12.60 pg/cell), Run 9 (12.13 pg/cell) and Run 81 (10.97 pg/cell). In general, the performance for cell density, mAb and Qp over the days for "Run 81" was similar as the ones presented by "New C" and "New D".

Although the scores region was restricted by the optimization model, again it is possible to observe that the new proposed media presents a higher consistent growth and mab results that the ones presented by Run 1 and Run 9.

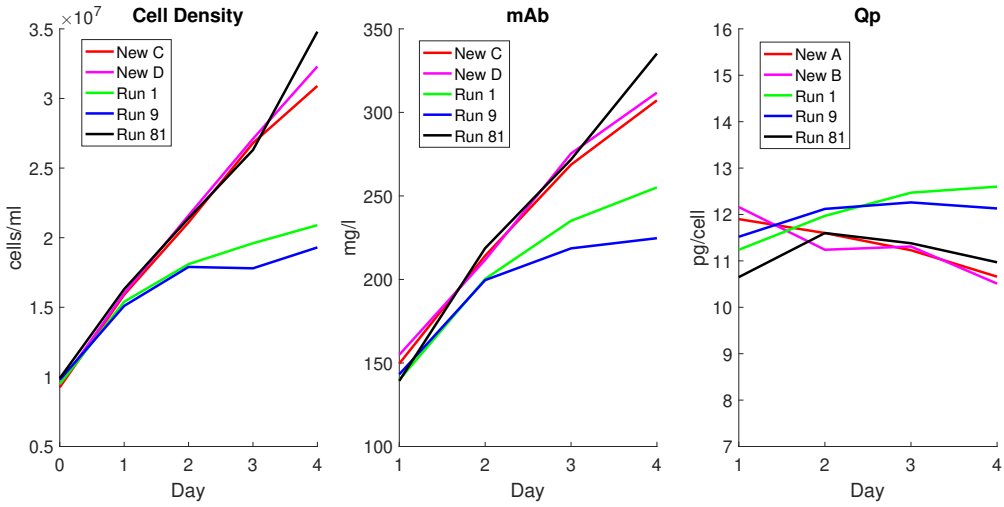


Figure 4.5: Daily experimental results for cell density, mAb and Qp for media "New A", "New B", "Run 1", "Run 9" and "Run 81".

It is also possible to observe that the medium "New D" (CD>1) presents a higher cell density but smaller Qp compared to "New C" (CD>0), confirming that an apparent tradeoff between cell density and Qp is captured by the PCR model. The trade off between cell density and Qp is also shown in Figure 4.6. Such trade off between growth and productivity has been reported [7] [42] by the relative consumption of resources (nutrients) towards cell mass versus product. To verify whether the PCR model was able to predict the trade off between cell density and Qp, twenty five random media were generated from the proposed medium "New C" with random medium composition variation up to +- 30%. As Figure 4.6 shows, as cell density increases, the Qp values decreases.

In summary, the main advantage of the proposed new media ("New A", "New B", "New C" and "New D") over the media used by our industrial partner is the consistently higher cell growth while maintaining similar productivity values within the experimental noise. In view of the correlation between growth and productivity it was hypothesized that a PLS regression that will explicitly account for this correlation may lead to more accurate predictions. However, due to time limitations, this was left for future study.

Figure 4.7 shows a comparison between the new media compositions, "New C" and "New D", with medium "Run 9". Since in this case the input score space was constrained within the region spanned by the industrial media "Run 1", "Run 9", "Run 53" and "Run 81" the figure shows that the concentration of the new media "New C" and "New D" is more similar to "Run 9" as compared to "New A" and "New B", where the latent variables

(scores of principal components) were not constrained (see Figure 4.4). However, despite of the similarities, some components present a large difference of concentration between the new media and "Run 9", such as ψ_3, ψ_4, ψ_5, ψ_6, ϕ_9, ϕ_{37} and ϕ_{53}. Components that in the medium "Run 9"were zero, are present in the new media, such as, ϕ_4, ϕ_6, ϕ_{18} and ϕ_{36}. This results confirm that the PCR model is able to find new medium composition different from the ones presented by the design of the 80 media experiments even the input space was restricted.

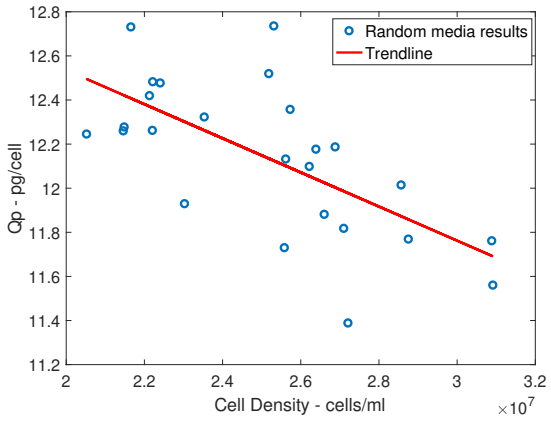


Figure 4.6: Qp and cell density relation predicted by PCR model approach.

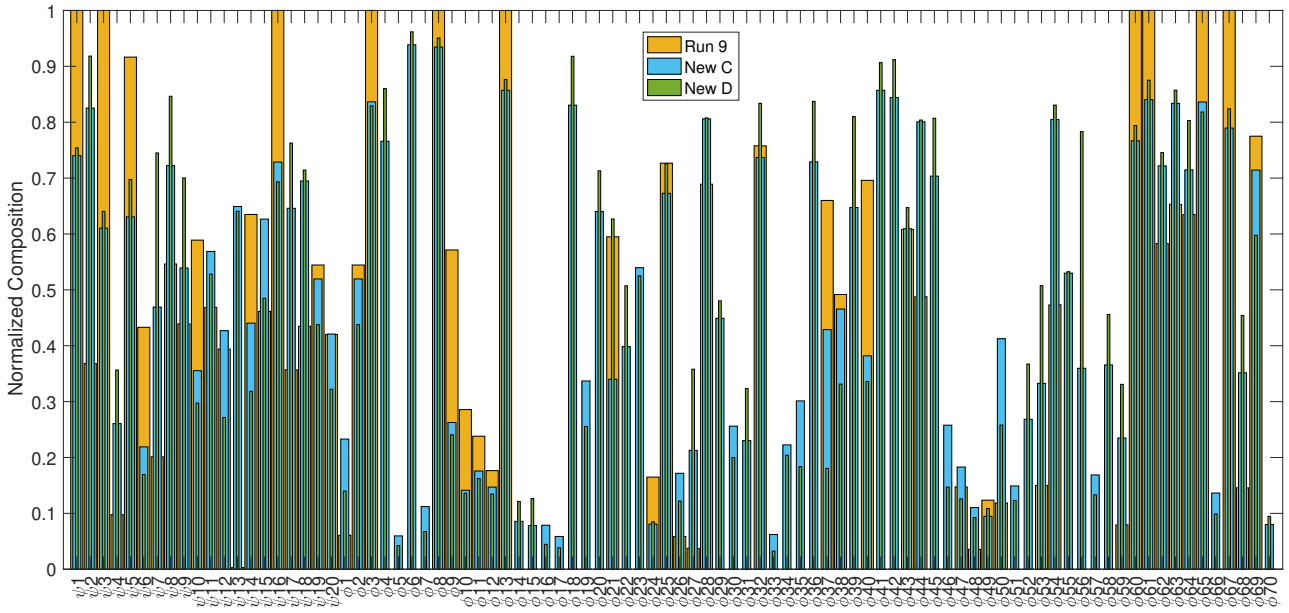


Figure 4.7: Comparison of optimal media formulations found by robust optimization in Case Study 2, "New C" and "New D", and industrial master media "Run 9" when $\theta = 0$ for PCR based models.

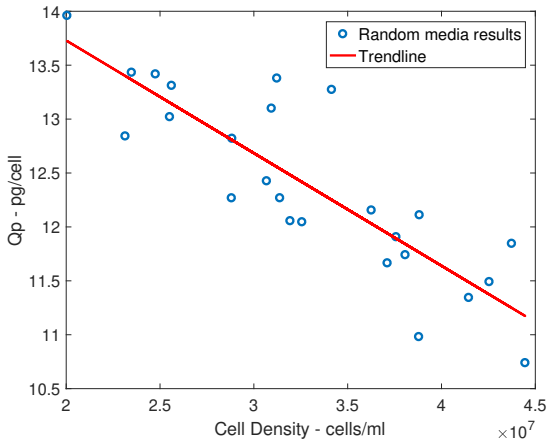


Figure 4.8: Qp and cell density relation predicted by PLS model approach.

It is important to notice, that the media "Run 1", "Run 9", "Run 53" and "Run 81" were prepared by combining the master media set used by our industrial partner. These master media was developed based on years of experimental studies and using run to run improvements. Although we were not able to find media that resulted in significantly higher productivity than existing media, the new media provided similar results in terms of productivity with higher cell growth with few experiments and model based optimization.

As mentioned above, PLS modelling could help to model the media while accounting for correlation between growth and productivity. A PLS model with 12 latent variables resulted in cross-validation training and testing errors of 0.25 pg/cell and 0.79 pg/cell, respectively, which are smaller than the original noise magnitude in the Qp data ($\pm$0.80 pg/cell). The prediction error of the PLS model for the 8 media used for testing was 0.6674 pg/cell, which is also smaller than the data noise ($\pm$0.80 pg/cell) thus confirming the ability of the PLS model to predict data not used for model calibration. The PLS model was also able to predict the trade off between cell density and Qp. As shown in Figure 4.8, as cell density increases, the Qp values decreases. Again, twenty five random media were generated from the proposed medium given by PLS regression with random medium composition variation up to $\pm$ 30%.

Experiments based on the PLS model have not been conducted as yet due to time limitations.

4.6 Case Study 3: Development of new media formulation using only a reduced subset of observations of the set used in Case Study 2

4.6.1 Methods

In this case study we investigate the possibility of optimizing the media formulation based on a subset of the original input regressor considered in Case Study 2. The goal is finding new media formulations according to the PCR based model described in Case Study 2 where instead of using all media observations (n) to define the PCR regression coefficients only the information of a reduced media subset is used. Note that in Case Study 3 it is assumed that initially only the media formulation is provided, but no outputs, e.g. productivity or metabolites concentrations, are available a priori. The idea is to define inputs that capture large variability in the input space and then select a small set of experiments for which outputs will be subsequently measured. This problem is very relevant in industrial practice since the measurements of outputs involve time consuming and expensive analysis. Thus there is great incentive to reduce the number of samples for measuring these outputs.

To find the reduced subset of media the D-optimal criteria was used. D-optimal designs are frequently used for non classical designs (e.g. designs that are not factorials or fractional factorials), and are a good option for training any type of model that is linear with respect to the model parameters. [115, 106].

Consider a matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ with dimensions satisfying $n > p$, the D-optimal criteria seeks to minimize the determinant of $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$, which leads to the minimization of the volume of the Joint Confidence Region (JCR) of the regression parameters. The smaller the volume of the JCR the more precise the coefficients estimates [106]. Note that minimizing the $|(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}|$ corresponds to maximizing the determinant of the information matrix $(\mathbf{A}^{\mathrm{T}}\mathbf{A})$. The information matrix carries information on both variances and covariances (correlations) of model regression parameters. Therefore, the D-optimality will select the media subset of observations such that the variance of the estimated parameters is minimized.

It should be noticed that the use of the D-optimal criterion requires the matrix $\mathbf{A}$ to satisfy $n > p$, since the information matrix given by $\mathbf{A}^{\mathrm{T}}\mathbf{A} \in \mathbb{R}^{p \times p}$ requires full rank. Because in our studies, the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$ has dimensions such as $n < p$, the dimension of inputs p must be reduced. For this purpose PCA is used to select a set of principal components such that a certain percentage of the original input data

variability is explained. It is evident that there is a cost for using a smaller data set in the form of a larger variance in parameters' values and corresponding prediction. Thus, the robust optimization based on a smaller data set is expected to result in an optimal value with a larger corresponding variance as shown in the Results section.

Assuming the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$ provides information of all media observations obtained through the blending of the master media, considering the linear factors, two-way interaction and squared terms, therefore, $n = 80$ and $p = 4185$ ($n < p$). Aiming to select a reduced number of observations (n_{subset}) as compared to Case Study 2, the D-optimal criteria was applied as follows:

i. Define the number of media subset to be selected, n_{subset};

ii. Apply Principal Components Analysis (PCA) to the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$;

iii. Select a number of principal components nPC such that such that a certain percentage of the original input data variability is explained. It should be noticed that the number of principal components n_{PC} must be smaller than the number of media in the chosen subset n_{subset} to satisfy the minimal rank condition;

iv. The regressor matrix is now re-written in terms of the scores from the PCA $\mathbf{XS}_{n_{PC}} \in \mathbb{R}^{n \times n_{PC}}$. Observe that the scores of PCA are given by $\mathbf{XS}_{n_{PC}} = \mathbf{X}_{regressor}\mathbf{XL}_{n_{PC}}$, with $\mathbf{XL}_{n_{PC}} \in \mathbb{R}^{p \times n_{PC}}$ being the loadings from PCA considering n_{PC} principal components;

v. The optimization problem described in equations 4.11 is solved by a genetic algorithm to find the combination of media subset observations, here described by vector $\mathbf{s}$, that maximize the D-optimal criteria.

$$\begin{aligned} \underset{\mathbf{subset}}{\text{maximize}} \quad & \left| \left(\mathbf{XS}_{n_{PC}}^{subset}\right)^{\mathrm{T}} \left(\mathbf{XS}_{n_{PC}}^{subset}\right) \right| \\ \text{subject to} \quad & \mathbf{XS}_{n_{PC}}^{subset} = \mathbf{XS}_{n_{PC}}\left(\mathbf{s}, 1 : n_{PC}\right) \end{aligned} \tag{4.11}$$

where $\mathbf{XS}_{n_{PC}}^{subset} \in \mathbb{R}^{n_{subset} \times n_{PC}}$ is the score matrix in which the rows are composed by a subset of the original n media considered in Case Study 2, $\mathbf{s} \in \mathbb{R}^{1 \times n_{subset}}$ is a vector with a combination of unique subset of media, e.g. the vector s contain a set of integer numbers that can be from 1 to n. A genetic algorithm is used because the problem involves integer decision variables that are used to describe the subset of media selected from the set used in Case Study 2. The implementation was done using Matlab genetic algorithm function *ga* for global optimization in parallel mode.

Only the subset of media found by the D-optimality approach and its correspondent Qp subset data is used to develop new media. The steps for the development of new media using PCR model based robust optimization are the same as described in section 4.4.1.

It should be noticed that the method described in this case study can be also applied for designing experiments in which the goal is to find information other than productivity or growth. Although these quantities are the primary variables to be optimized by media design, it is often of necessary to measure quantities other than growth or productivity for further modelling or for improving the quality of the product, e.g. amino acids, glycosylation, etc. On the other hand, it is often time consuming and expensive to measure these variables for a large number of different media and therefore it is of practical interest to reduce the number of samples that will require these measurements. The D-optimality method proposed here can serve to reduce the number of experiments for this purpose.

4.6.2 Results

Modelling and optimization for simulated output data

The method in this case study is similar to Case Study 2, but here only a subset of the 80 media considered in case 2 is used to calibrate a model for prediction of new media formulation and for robust optimization based on this model. The key goal of the current case, in contrast with Case Study 2, is to test whether it is possible to formulate a model that will result in similar prediction accuracy and similar optimization results as the ones obtained with the larger set of data used in Case Study 2. This has significant importance in an industrial setting since it can save costly and time consuming experiments for design of new media.

The subset of media out of all the 80 media used in Case Study 2 was selected based on a D-optimal criteria. In principle we could search for a minimal number of media that will result in acceptable prediction and optimization accuracy. This accuracy could be decided on the basis of a desired confidence interval for prediction and optimal solution. However, this results in a challenging combinatorial optimization problem which is left for future study. Instead, the number was chosen arbitrarily to be equal to 50.

The subset of media was chosen based on the scores of PCA analysis of the inputs according to the steps described in section 4.6. For the D-optimal optimization, 14 principal components were selected which explained 90% of the variability of the input data of the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$, assuming linear factors, two-way interactions and squared terms, with $n = 80$ and $p = 4185$. Since the input was selected based on PCA

analysis only PCR regression was conducted since PCA is specifically targeted to minimize the variance of the coefficients of the PCR regression model.

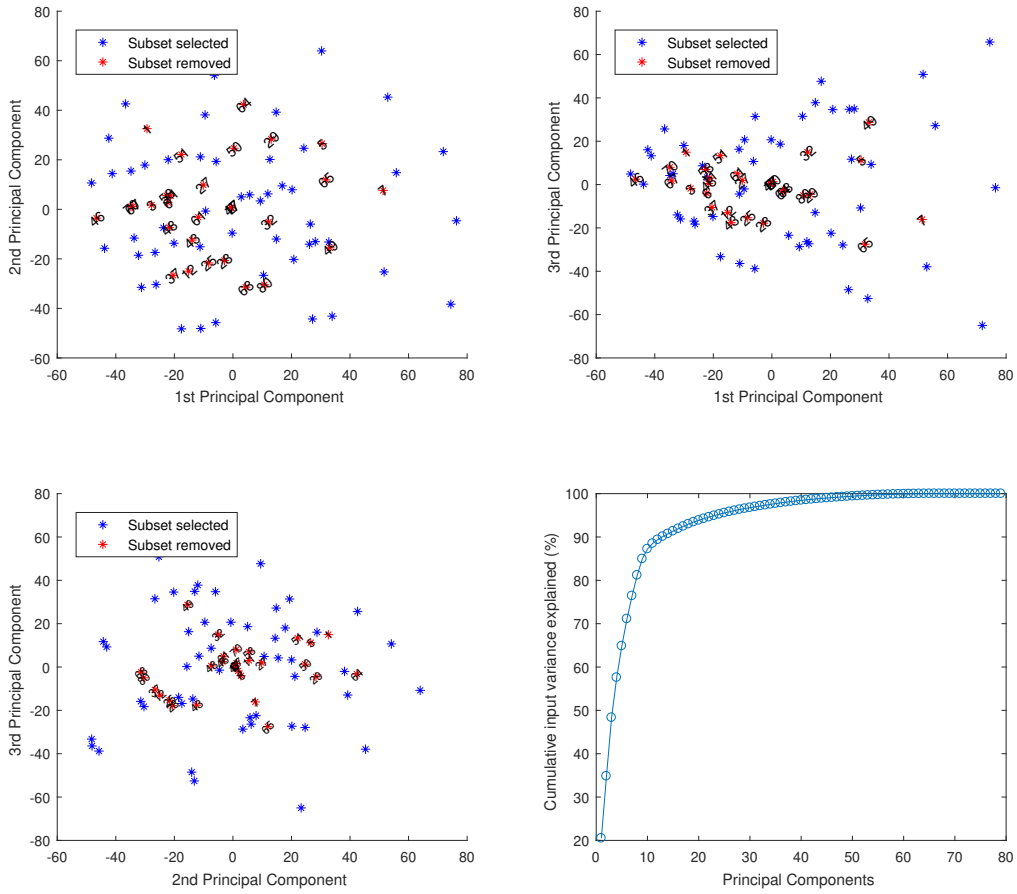


Figure 4.9: Subset of media selected and removed identified in the scores principal components plots and cumulative input variance explained by the principal components of the 80 media formulation.

Figure 4.9 shows the subset of 50 media selected (in blue) and the subset of media removed (in red) identified in the principal components plots. The number on top of the red symbols of the subset removed represents the number of original corresponded media that was removed. Comparing the selected and the removed subsets, it can be observed

that the selected subset corresponds to media that is present in the region that better explain the variability of input space. The media that was removed by the D-optimal optimization are media that are located closer to other media in the principal components space. Figure 4.9 also shows the percentage of the cumulative input variance explained by the principal components and at least 14 principal components are needed to explain approximately 90% of the total variance in the input data. This observation cannot be fully justified in terms of the first two principal components since they capture only 35% of the variance. For instance, media 1 and 46 that were removed appear in the plot of the first two principal components far from other media. However, these two media formulations are located very close in the space spanned by the 1st and 3rd and by the 2nd and 3rd principal components.

Using only the 50 media selected by PCA, 14 principal components were needed to obtain training and testing errors that are of the same order of magnitude and that are smaller than the noise assumed in the simulated data (0.8946). The cross-validation error for training was 0.4931, while the testing 0.6433. Two sets of 15 media, not used before, were used to evaluate the prediction error. In the prediction set 1, the media formulations were created using higher weight on the first principal components of the PCA scores of the subset regressor matrix, and the prediction set 2, using higher weight on the last scores of PCA scores of the subset regressor matrix. While the SSE and the prediction error for the set 1 was 3.20 and 0.4620, respectively, the set 2 presented SSE and prediction error equal to 33.29 and 1.490. Similar to Case Study 2, these results confirm that the PCR based model that is calibrated with a subset of 50 media is able to provide acceptable prediction accuracy for media formulations values located in the region of validity of the model as determined by the scores along the principal components.

Figure 4.10 shows the maximum Qp given by the selected media subset formulation (in red) compared with the predicted (light blue) and simulated (dark blue) Qp for the new media when θ is equal 0, -1 and 1 for PCR based models corresponding to the average, lower and upper bounds values respectively. The Qp predictions (in light blue) were obtained using PCR model regression, and the corresponding true Qp values (in dark blue and orange) were obtained by multiplying the regressor matrix of each new medium formulation by the regression coefficients as defined in Table 4.4. For the PCR based model, the Qp predictions were 5.282 ($\theta = 0$), 4.801 ($\theta = -1$) and 7.932 ($\theta = 1$). Thus, the results for Qp prediction using the subset of media are very close to the ones found in Case Study 2 for PCR model. It is important to notice that the bounds are slightly larger when 50 media is used (5.282 and 7.932) versus the case when 80 media is used (4.9965 and 7.3182) thus corroborating that there is a trade-off between the number of experiments and the variance of the optimal solution. The Qp true results for the new media formulations found by PCR

model in the optimization was 8.823 that is relatively similar to the value obtained in Case Study 2.

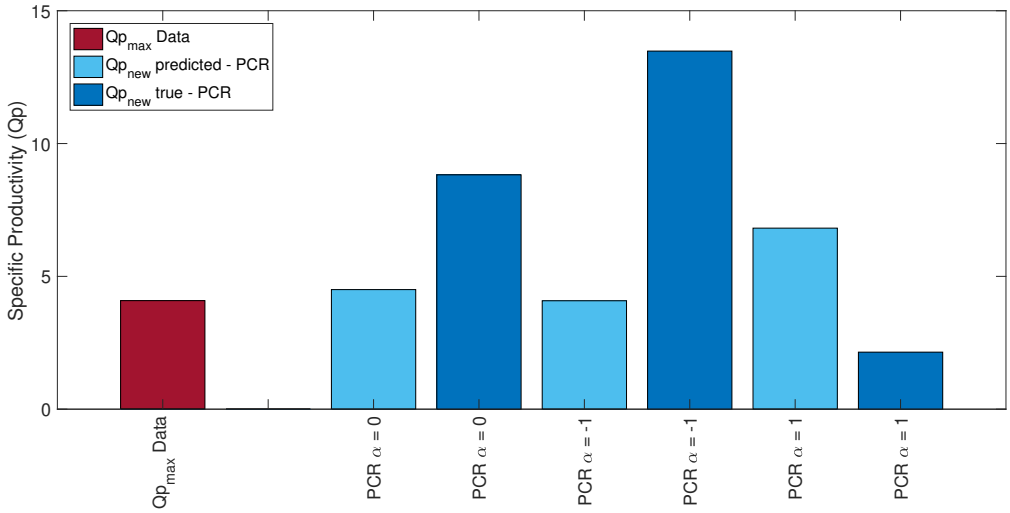


Figure 4.10: Maximum specific productivity (Qp) given by the selected subset of media formulation compared with the predicted and simulated specific productivity for the new media obtained when θ is equal 0, -1 and 1 for PCR regression models.

Figure 4.11 compares the new media formulation found by Case Study 2 and Case Study 3 for PCR model and $\theta = 0$, corresponding to the average Qp value. The optimum medium found in Case Study 2 and 3 were similar for most of the medium components, thus corroborating the option of reducing the number of experiments without significantly affecting the optimization results. The majority of the components vary no more than 20% of the composition between the results obtained with the initial 80 media and the results obtained with 50 media. The medium components that presented the highest differences among the optimum media obtained in Case Study 2 and 3, such as ψ_4, ϕ_1, ϕ_{10}, ϕ_{48} and ϕ_{68}, are not present in the true model, indicating that their differences in concentrations do not have large impact on the Qp results.

We also investigated the effect of the choice of n_{subset} media selected for two extreme cases of n=30 and n=75. We corroborated that the number of media chosen for calibration affects the model validation results for the media not used for calibration. The validation errors are 0.96 and 0.70 for n=30 and n=75 respectively. Also we could verify that as more media is used for calibration the resulting optimal media composition converges to the optimal media composition obtained with all the 80 samples. For example, the average

differences in media components with respect to the optimal composition obtained with all the 80 media are 14.3% and -0.1 % for n=30 and 75 respectively.

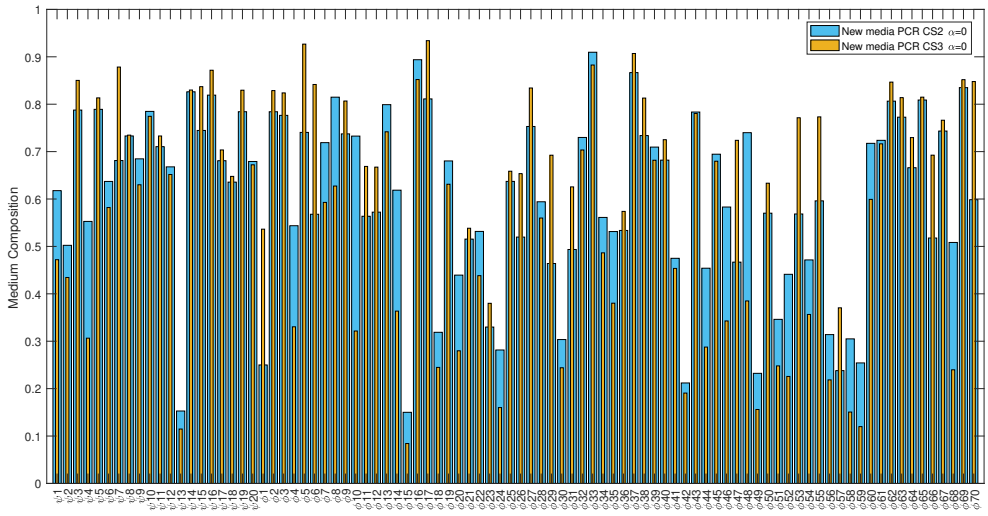


Figure 4.11: Comparison of optimal medium components concentration (X_{new}) found by robust optimization in Case Study 2 and 3 when $\theta = 0$ for PCR based models.

Modelling accuracy for industrial data based on reduced set of experiments

A PCA model of the input data was found with 35 principal components that explained about 97.7% of the variability of the input data of the regressor matrix $\mathbf{X}_{regressor} \in \mathbb{R}^{n \times p}$. The model assume linear factors, two-way interactions and squared terms, with $n = 80$ and $p = 4185$. A subset of media was selected using a D optimal criteria that was calculated from the scores of the PCA model.

An additional PCA model was found based on the data of the 50 most informative subset of media as calculated from D optimality. With 20 principal components the training cross-validation error was 0.42 pg/cell, while the testing cross-validation error was 0.93 pg/cell, about 15% larger than the noise. The 30 media that was not selected by the D-optimality based selection were used to validate the PCR model prediction. The prediction error of the removed 30 media set was 0.70 pg/cell, i.e. smaller than the data noise (0.80 pg/cell). This confirmed that a good predictive model could also be obtained with a subset of only 50 media formulations.

In this work, only the robust optimization formulations found with the PCR model were experimentally tested. Optimal medium formulation found by PLS model approach will be experimentally tested in a future work. The robust optimization was developed without including the constraints on glutamate, glucose, lactate and ammonia. The model was focused in the maximization of Qp and constrained only by the Qp maximum value and components composition bounds, as shown in equation 4.10.

The new optimum medium formulation obtained, "New E", assumed $Qp_{t,max}$ equal 30% of the maximum Qp obtained by the 80 media experiments. The objective is to show that the optimum subset of media selected out of the 80 media experiments using D-optimal criteria is able to predict new medium formulation with as good or similar Qp as the new medium obtained when using information of all 80 media experiments. Figure 4.12 compares the medium composition for "New E" and "New A" (which was obtain using same constraints in Case Study 2). The medium composition of "New E" varies an average of $\pm 15\%$ with respect to medium "New A". The components ϕ_{14} and ϕ_{70} are the ones with the largest difference in concentration, presenting almost double concentrations in "New E" as compared to "New A".

The performance of the medium "New E" is also similar to "New A", as can be seen in Figure 4.13. Cell growth and mAb were consistently high through the days of culture. The Qp at the end of the operation, day 4, is 10.77 pg/cell. The predicted Qp for the PCR based model was 12.24 pg/cell, therefore the experiment value is within the error of the new run of experiments (+-1.55 pg/cell).

The results confirm the PCR based model is able to predict a new media formulation using an optimal subset of media obtained by D-optimal criteria. This has important implications for the industrial partner since the preparation of media formulation is time consuming and costly.

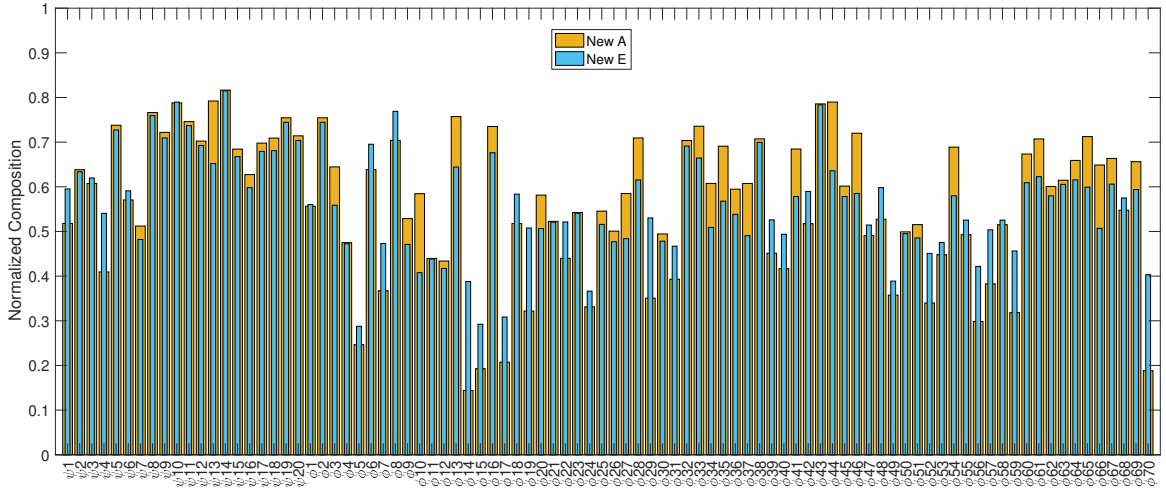


Figure 4.12: Experimental comparison of optimal medium components concentration "New A" and "New E" found by robust optimization in Case Study 2 and 3 when $\theta = 0$ for PCR based models.

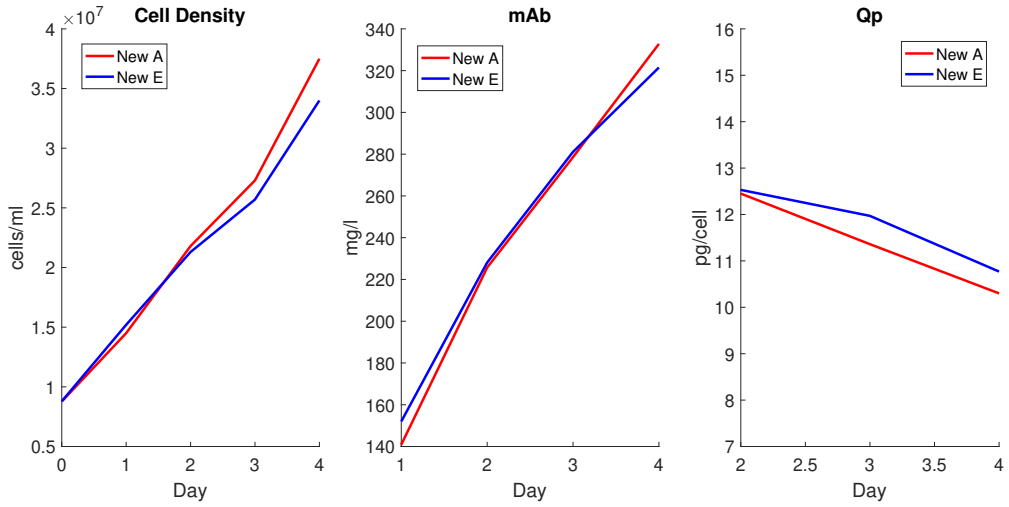


Figure 4.13: Daily experimental results for cell density, mAb and Qp for media "New A", "New E".

4.7 Conclusions

The problem of culture medium optimization was approached by a combination of empirical modelling and robust optimization based on these models. The problem presents the particular challenge that the number of samples is smaller than the number of variables ($n < p$) thus increasing the sensitivity to noise in the model parameters. By means of a toy example it was shown that regularization methods such as LASSO and Elastic net may result in important coefficients being neglected during training while these coefficients may be important for subsequent optimization. PCR and PLS were found to mitigate this problem by considering all components.

Through simulated studies it was shown that PCR and PLS result in models with good prediction accuracy within the magnitude of the noise. The results were corroborated with output measurements from industrial data. It was shown that the prediction accuracy largely depends on the proximity of new media formulations being located within the input space spanned by the principal components or principal latent variables of the trained PCR and PLS models. Then, to ensure the applicability of the model, constraints were introduced to limit the search space of the robust optimization problems.

It was shown that by applying a D-optimal approach it is possible to reduce the number of experiments required for further media optimization. The optima resulting from the

model calibrated with the smaller subset of media formulation led to very similar optima to the ones obtained with the larger data set. The main drawback of optimizing the media with a model based on a smaller number of formulations is a slightly larger variance of the optimum value but the savings in the number of experiments can be potentially significant.

Experimental results confirmed that the PCR based model is able to generate in both Case Study 2 and Case Study 3, new media formulation that presents equivalent Qp values to the highest Qp values obtained by the industrial partner. The differences between the Qp values of the new media and best existent formulations are within the margin of error in the experiments. Furthermore, The proposed new media presented consistently high cell density and mAb values throughout the cell culture as compared to the existent formulations. The results of Case Study 3 proved that a new media formulation, similar to the one obtained in Case Study 2, can be obtained by using a subset of 50 media selected based on a D-optimality criteria using the initial 80 media experiments. This method can be used by the industry to reduce the number of experiments to be performed and, consequently, their costs during experiments of media formulation.

Chapter 5

Hybrid Modelling Approach for Mammalian Cells

The work presented in this chapter has been submitted to the Biochemical Engineering Journal.

This chapter presents a hybrid modeling approach to describe mammalian cell cultures. The model was calibrated with both data provided by an industrial collaborator and data collected at Waterloo. The model is initially developed for a batch operation and its main goal is to include the effect of minor components usually not included in most dynamic metabolic models. Minor components will generally refer in this proposal to species that are present in very low levels in the media, e.g. metals, vitamins etc., as compared to nutrients that occur in higher concentrations such as amino acids and glucose or key by-products such as lactate and ammonia. One of the key challenges for modelling the effect of the minor components is that for confidentiality reasons manufacturers of media, including our industrial collaborator, are only able to provide their relative amounts but cannot identified them by name. Since the nature of the minor element is not known its role in the metabolism cannot be specifically accounted for. Instead, this work presents a way to incorporate the contribution of the minor elements through an empirical correction that is incorporated into a mechanistic model, that describe mass balances of amino acids. The combination of the empirical correction and the mechanistic model results in a hybrid model that was calibrated and validated with experimental data. This part of the study was pursued in close collaboration with MilliporeSigma - A business of Merck KGaA, Darmstadt, Germany. The goal was to develop a comprehensive model that can be used in the future to optimize batch, fed-batch and perfusion operations with respect to the media formulation.

5.1 Introduction

In addition to the nutrients present in cell culture medium, minor components such as vitamins, hormones and metals, play a significant role in cells metabolism [140]. Often, the concentration of those minor components over time are hard to measure due to the lack of specific equipment and methods. The lack of experimental data for minor elements poses a challenge for modelling their effect. Moreover, dynamic experimental measurements of all medium components are an expensive and time consuming task and thus the models must be generally calibrated with scarce data.

When developing new media formulations for cell culture, knowledge about how each medium component affects cell behavior is crucial. Commonly, a media development procedure involves the manipulation of a large number of input variables while generally only a small number of samples is available for analysis. Regression models have generally been used to correlate response variables, such as cell growth, with the input variables (for instance, medium components and their interactions). However, standard least square regression approach are not suitable in cases where there is a small number of samples available and a large number of predictor variables that are being considered [94]. The particular challenges regarding the regression modelling of high dimensional $n < p$ problems have been discussed in Chapters 2 and 4. In order to obtain a concise regression model, Partial Least-Squares (PLS) has been reported as an convenient empirical modelling approach to deal with a large number of input variables. The PLS model can also be used to find the most significant predictors of output variables such as cell growth or productivity through regression of input and output data [134]. However, the key disadvantage of the empirical regression models that were presented in an earlier chapter is that they cannot properly capture the dynamic behaviour of the culture. Hence, those static models cannot be effectively used for optimization of fed-batch and perfusion operations where process dynamics is important. Although dynamics could be approximately incorporated into empirical models using techniques such as dynamic PLS, it is challenging to account in these techniques for the correct effect of perfusion and feeding rates and for a priori known interactions between amino acids, byproducts (ammonia, lactate), biomass and product. On the other hand, while these effects and interactions can be correctly modelled by mechanistic mass balances, the inclusion of the effects of minor elements into a mechanistic model is challenging due to lack of knowledge about their effects. Thus, the aim of the hybrid model proposed in this chapter is to combine mechanistic models based on prior knowledge of the metabolic network with an empirical model that describes the impact of minor elements which metabolic impact is not a priori known.

This chapter presents a hybrid approach that combines the mechanistic information

from the kinetic dynamic model developed by Hille [64] and the empirical regression model based on a PLS regression approach. The PLS regression aims to capture the influence of all medium components (inclusively minor nutrients) and possible interaction between them and the amino acids. This model was formulated based in the fact that the concentration of all cell culture medium nutrients is known at the initial time (including the minor components), but only few major metabolites and amino acids are measured over time.

5.2 Experimental Materials and Methods

The batch experiments presented in this work were conducted by MilliporeSigma and the data collected from these experiments were provided to us for modelling and analysis purposes. Analysis of amino acids for different media formulations were conducted at the University of Waterloo.

5.2.1 Cell Culture

A Chinese Hamster Ovarian (CHO) cell line producing monoclonal antibodies was used in the experiments. The cell culture experiments were done by MilliporeSigma and performed for 11 different media formulations. The cells were cultivated in TPP tubes (50 mL centrifuge tube with air exchange in the cap) in high batch mode with a initial working volume of 32 ml and a seeding density of 2×10^6 cells/ml. Cells were cultivated in a CO2 incubator set at 200 rpm, 37°C, 5% CO_2 and 80% humidity. 300 μl of glucose (with concentration of 450 g/L) was fed at day 3. Samples of 1.8 ml were collected at days 0, 3, 4, 5 and 6 (MilliporeSigma) for analysis.

Cell Count, Analysis of Main Metabolites and Monoclonal Antibody

Cell count, analysis of main metabolites and quantification of mAb were conducted at MilliporeSigma. Cell count were performed using a Vi-CELL (Beckman Coulter Life Sciences, IN, USA). Lactate, ammonia, glucose and glutamine were analyzed using a BioProfile Flex2 (Nova Biomedical, MA, USA). Quantification of monoclonal antibody was conducted using a ForteBio (Pall ForteBio LLC, CA, USA).

5.2.2 Amino Acid Analysis

Samples from cell culture experiments collected at days 0, 3, 4, 5 and 6 were sent by MilliporeSigma to Waterloo for amino acid analysis. For amino acids quantification, the AccQ.Tag^{TM} Method with a pre-column derivatization made of 6-aminoquinolyl-N-hydroxysuccinimidyl carbamate (AQC) (Waters, MA.USA) was used [33]. The Waters AccQ.Fluor^{TM} Reagent Kit (Waters, MA.USA) was used to quantify the amino acids concentration in the CHO cell culture medium samples. The set of derivatizing reagents includes a AccQ.Fluor Borate Buffer, a AccQ.Fluor Reagent Powder (6-aminoquinolyl-N-hydroxysuccinimidyl carbomate - AQC) and a AccQ.Fluor Reagent Diluent (acetonitrile) (Waters, MA.USA). This set of reagents result in highly stable fluorescent derivatives. The AccQ.Tag^{TM} Method uses the AQC pre-column to separate the amino acids present in the derivatized samples which were subsequently quantified through fluorescence detection.

Reconstituting Accq.Fluor Reagent

Following the Waters AccQ.Fluor^{TM} Reagent Kit instructions sheet (Waters, MA, USA), 1 ml of reagent diluent (Vial 2B) was transfered to reagent powder (Vial 2A) container and vortex for 10 seconds. Then the reconstituted reagent was placed in a water bath at 55°C, vortexed at intervals, until the powder was completely dissolved.

Calibration Standard

The calibration standard solution was prepared by mixing 80 μl of Sigma-Aldrich Amino Acid Standard (Sigma Aldrich, MI, USA) with 920 μl of 18.2 MΩ water in a micro centrifuge tube. In this standard, the amino acids are in a concentration of 2.5 μmoles/ml, except L-cystine at 1.25 μmoles/ml. Accordingly, the calibration standard solution contains 200 pmol/μl, except L-cystine at 100 pmol/μl. In the micro centrifuge tubes, the calibration standard was successively diluted in a 1:2 dilution factor. The derivatized calibration standard results, at different dilution rates, were used to built a standard calibration curve for amino acids analysis.

Derivatizing the Calibration Standard and Samples

Each calibration standard and each cell culture sample were derivatized using the same procedure. To derivatize, 10 μl of the calibration solution/sample was mixed with 70 μl of

Waters AccQ.Fluor^{TM} Reagent Kit (Vial 1) in a micro centrifuge tube and was vortexed for 10 seconds. Then, 20 μl of reconstituted reagent (subsection 5.2.2) was added and vortexed for 10 seconds. After being incubated for 1 minute at room temperature, the derivatized calibration standard/sample was transfered to an autosampler vial with a limited volume insert and closed with a silicone-lined septum cap. This vial was then placed at the water bath at 55° for 10 minutes.

HPLC Amino Acid Analysis

The amino acid analysis was performed in a Waters 1525 chromatography system (Waters Corporation, MA, USA) coupled with a fluorescence detector (W2475; Waters Corporation, MA, USA). The AQC pre-column (Waters AccQ.Tag^{TM} Amino Acid Analysis Column C18, dimensions 4 μm, 3.9 mm $\times$ 150 mm) used for this method was set at 38°C. A sample volume of 5μl was injected for each analysis. The fluorescence detector was set with a excitation wavelength of 248 nm and an emission wavelength of 395 nm. The eluent A used was an acetate-phosphate buffer solution prepared by mixing 100 mL AccQ.Tag^{TM} Eluent A concentrate (Waters, MA.USA) with 1-liter Millipore water. The eluent B was HPLC grade acetonitrile, and eluent C was Millipore water. The flow rate of each eluent at different times for HPLC run, were adapted from the work of Cohen [33] and are presented in Table 5.1. The HPLC analysis for each sample runs for 40 min. Before each run starts, and between samples run, the column was equilibrated for 10 min using 99% of eluent A and 1% of eluent B for 10 minutes at 1 ml/min.

Table 5.1: Gradient table for the HPLC runs for ternary eluent system.

Time (min)	Flow rate (ml/min)	Eluent A Buffer (%)	Eluent B ACN - HPLC grading (%)	Eluent C Water (%)	Gradient Curve
0	1	99	1	0	-
0.5	1	98	2	0	6
18	1	95	5	0	6
19	1	91	9	0	6
29.5	1	80	20	0	6
33	1	0	60	40	11
40	1	99	1	0	11

5.3 Model Development

5.3.1 PLS Regression based Models

Regression models are commonly used for estimation of results obtained through experimental designs. However, when the number of input variables (predictors) are larger than the number of sample observations (response), the traditional least squared regression approach are likely to provide a model that over-fits the sampled data [157]. In high dimensional n < p problems, although many input variables (p) are present, there may be only a few underlying or latent variables that account for most of the variability of the output response [157, 134].

The Partial Least Squares (PLS) regression approach compresses the original input data by projecting it onto a new set of predictors orthogonal variables referred to as latent variables [1, 134]. Differently from Principal Components Analysis (PCA) or Principal Components Regression (PCR), the PLS regression is focused on the prediction of the output responses and not necessarily on trying to understand the underlying relationship amongst the input variables [157]. Specifically, the PLS approach calculates a set of latent variables that maximizes the covariance between the input and the output response variables [134]. This maximization is achieved by combining information from the variance and correlations among both input predictors and output observations [101].

This work uses the *plsregress* function from MatLab [101], which follows the SIMPLS algorithm [39]. In this approach, the centered input variables ($X0 \in \mathbb{R}^{n \times p}$) and centered output responses ($Y0 \in \mathbb{R}^{n \times m}$) can be reconstructed using the following relations (equations 5.1 and 5.2):

$$\mathbf{X0} = (\mathbf{XS})(\mathbf{XL})' \tag{5.1}$$

$$\mathbf{Y0} = (\mathbf{XS})(\mathbf{YL})' \tag{5.2}$$

where $\mathbf{XS} \in \mathbb{R}^{n \times nPC}$ are the input scores (magnitudes of latent variables), $\mathbf{XL} \in \mathbb{R}^{p \times nPC}$ are the input loadings, $\mathbf{YL} \in \mathbb{R}^{m \times nPC}$ are the response loadings, n is the number of samples, p the number of original predictors, m the number of response variables and nPC is the number of PLS components that are considered statistically significant.

The input scores **XS** is obtained by multiplying the centered input matrix by the matrix of PLS weights ($\mathbf{W} \in \mathbb{R}^{p \times nPC}$), obtained by the SIMPLS algorithm, as shown in equation 5.3.

$$\mathbf{XS} = (\mathbf{X0})(\mathbf{W}) \tag{5.3}$$

The equation resulting from substitution of equation 5.3 into equation 5.2 can be used

to reconstruct the output centered response as shown in equation 5.4.

$$\mathbf{Y0} = (\mathbf{X0})(\mathbf{W})(\mathbf{YL})' \tag{5.4}$$

In this work, the PLS regression is used to describe how the different minor elements and their possible interactions among themselves and with amino acids can affect the cell performance (biomass growth, productivity and/or main metabolites consumption/production rate).

5.3.2 Dynamic Kinetic Model

The dynamic kinetic model used in this work is adapted from Hille [64]. The model developed by Hille [64] describes the main metabolites and will be referred as mechanistic since it is based on the use of Metabolic Flux Analysis (MFA) for identification of significant intracellular fluxes and on dynamic mass balances of the modelled species. After conducting flux balance analysis and neglecting non-significant fluxes, i.e. fluxes that are smaller than 1% of the total fluxes, a set of macro reactions was developed where each of these reactions was described by a kinetic expression of Michaelis-Menten or other form. Using the kinetic expressions representing the macro reactions dynamic mass balance equations for each modeled metabolite are formulated [64] . The resulting set of non-linear differential equations for a batch operation can be found in Appendix C.

The model involves 24 equations and 83 parameters. The set of equations used in the model is described in Appendix C. The coupling among the nonlinear equations and the relatively large number of parameters pose challenges for parameter estimation and requires large computational times. Also the fact that only sparse measurements, i.e. samples are collected only at day 1, 3, 4, 5 and 6, are available for training makes the model calibration more difficult. To reduce the computations, Hille [64] proposed a sequential parameter optimization approach, also used in this current work, in which successive optimization problems involving a smaller number of parameters are solved. The ability of splitting the problem into smaller ones is due to the lack or weak coupling between some of the equations in the model. According to Hille [64]:

> The main motivation behind this sequential model calibration procedure is that the dynamics of most metabolites mainly depend on the change in biomass. This allows us to divide the problem into the following steps: 1. Approximation of the measured viable cell density profile by a simple piecewise linear interpolation to capture the

> dynamic behaviour of biomass. 2. Successive estimation of all the minor amino acids which do not contribute to the evolution of other metabolites in the sense that their fluxes are insignificant as per the MFA. These smaller parameter estimation steps can be executed independently from each other. 3. Estimation of major metabolites which are involved in the macro-reactions, where the dynamics of coupled metabolites are estimated simultaneously. 4. Finally, separate estimation of the parameters related to the dead cell and viable cell density equations. (Hille, 2018, p. 150)

More details about the sequential parameter optimization proposed by Hille can be found at [64]. The table with the estimated parameters for the dynamic kinetic model is presented in Appendix C section D.1.

5.3.3 Proposed Hybrid Model

The dynamic metabolic model developed by Hille [64] for CHO cells does not include the effect of minor components, such as hormones, vitamins, lipids, etc, that have been found to have significant effect on the evolution of the cell culture[161]. This work proposes the incorporation of a PLS regression model into the mechanistic metabolic model developed by Hille [64] to account for the effect of minor components. The rationale is that the original mechanistic dynamic model will account for the evolution of major metabolites and amino acid concentrations through time (ψ) and the PLS regression approach will correct the mechanistic model for the influence of minor nutrients (ϕ) and possible interactions among all medium components that can affect the cell culture behavior.

The formulation of the hybrid model is motivated by the concept of Taylor expansion of nonlinear functions as follows. The system of nonlinear differential equations to describe the evolution of amino acids in the cell culture is given as shown in equation 5.5:

$$\frac{1}{X_v}\frac{d(\psi(t))}{dt} = f\left(\psi(t), \phi\right) \tag{5.5}$$

where ψ denote the major components that refer, as stated above, to amino acids, major by-products, i.e. ammonia and lactate and main nutrients such as glucose and glutamate, ϕ denote the minor components such as vitamins, hormones and metals at time zero (media formulation concentration), X_v is biomass, and f denotes a nonlinear function of the concentration of all chemical species considered in the model. It is important to notice that only the initial concentrations of minor elements are known but not by their name due to confidentiality. Thus, we were not able to account for the interactions of the

minor elements in the metabolic network with the amino acids, main nutrients and main by-products.

The idea to include the effect of the minor elements is to add their effect with respect to concentration trajectories defined in terms of the time dependent concentration of elements accounted for in the mechanistic model and average values of the initial concentrations of the minor elements. Accordingly, the nonlinear function f is approximated by a first order Taylor expansion around the time trajectories of the major metabolites and the average initial compositions of the minor components as follows:

$$\frac{1}{X_v}\frac{d(\psi(t))}{dt} = f\left(\psi(t), \bar{\phi}\right) + \left(\frac{df}{d\psi}\right)_{\psi(t),\bar{\phi}} (\psi(t) - \psi(t)) + \left(\frac{df}{d\phi}\right)_{\psi(t),\bar{\phi}} \left(\phi - \bar{\phi}\right) \tag{5.6}$$

$$\frac{1}{X_v}\frac{d(\psi(t))}{dt} = f\left(\psi(t), \bar{\phi}\right) + \left(\frac{df}{d\phi}\right)_{\psi(t),\bar{\phi}} (\delta\phi) \tag{5.7}$$

$$\frac{1}{X_v}\frac{d(\psi(t))}{dt} = f\left(\psi(t), \bar{\phi}\right) + g\left(\psi(t), \delta\phi\right) \tag{5.8}$$

where $\delta\phi$ for a medium i is the difference between ϕ of medium i and the average of ϕ for all media formulation.

Therefore, using the Taylor expansion's based argument regarding the superposition of the effects of major components and minor components, the final parallel hybrid model is described as shown in equation 5.8. The function $f\left(\psi(t), \bar{\phi}\right)$, represents the consumption/production of the major components with time by biomass based on the dynamic mechanistic model of Hille [64]. Simultaneously with function $f\left(\psi(t), \bar{\phi}\right)$, the function $g\left(\psi(t), \delta\phi\right)$ accounts for the influence of minor components concentrations and possible interaction among all components, which is described by the PLS regression model, e.g., $\left(\frac{df}{d\phi}\right)_{\psi(t),\bar{\phi}} (\delta\phi) = g\left(\psi(t), \delta\phi\right) = Y_{PLS}$. Observe that the general regression model given by function $g\left(\psi(t), \delta\phi\right)$ to describe the influence of medium components on an given output h follows the form:

$$\begin{aligned} h\left(\psi(t), \delta\phi\right) = & \sum_{i=1}^{n_\psi} \left(\beta_{\psi_i}\right) \psi_i(t) + \sum_{j=1}^{n_\phi} \left(\beta_{\phi_j}\right) \phi_j + \sum_{i=1}^{n_\psi - 1} \sum_{k=1}^{n_\psi} \left(\beta_{\psi_i \psi_k}\right) \psi_i(t) \psi_k(t) \left\{i \neq k\right\} + \\ & \sum_{j=1}^{n_\phi - 1} \sum_{l=1}^{n_\phi} \left(\beta_{\phi_j \phi_l}\right) \phi_j \phi_l \left\{j \neq l\right\} + \sum_{i=1}^{n_\psi} \sum_{j=1}^{n_\phi} \left(\beta_{\psi_i \phi_j}\right) \psi_i(t) \phi_j \end{aligned} \tag{5.9}$$

where $h\left(\psi(t), \delta\phi\right)$ is the output response of interest, $\psi_i(t)$ is the major metabolite i concentration at time t, ϕ_j is the minor component j concentration in the medium formulation, n_ψ and n_ϕ are, respectively, the total number of major and minor components, and $\beta_{\psi_i}, \beta_{\phi_j}, \beta_{\psi_i\psi_k}, \beta_{\phi_j\phi_l}, \beta_{\psi_i\phi_j}, \beta_{{\psi_i}^2}, \beta_{{\phi_j}^2}$ are the regression coefficients. This regression model assumes that the predicted response, is a function of the regressors composed by linear factors ($\sum_{i=1}^{n_\psi} \psi_i(t)$ and $\sum_{j=1}^{n_\phi} \phi_j$) and two-way interactions ($\sum_{i=1,k=1}^{n_\psi} \psi_i(t)\psi_k(t)\,\{i \neq k\}$, $\sum_{j=1,l=1}^{n_\phi} \phi_j\phi_l\,\{j \neq l\}$, and $\sum_{i=1}^{n_\psi}\sum_{j=1}^{n_\phi} \psi_j(t)\phi_l$) given by medium components concentrations. The interactions assumed in this model are common due to the correlated nature of the metabolic reactions occurring in a mammalian cell.

It should be noticed that only first order terms were included in the Taylor expansion for the sole purpose of justifying the approximation of the original f function in equation 5.5 by a superposition of the functions shown in equation 5.8. However, the actual hybrid model does not use the first order derivative information and instead the first order term in equation 5.7 is described by a nonlinear empirical function of ψ and the deviations of ϕ from their corresponding average values. Furthermore, $f\left(\psi(t), \bar{\phi}\right)$ is trained, as further explained below, with different media formulations and thus the values of $\bar{\phi}$ are assumed to be embedded in the model parameters of this function.

In the proposed hybrid approach, the $g\left(\psi(t), \delta\phi\right)$ function, based on the PLS regression, was obtained according to the following steps:

i. Based on the available infrequent experimental data (samples were collected only at days 0, 3, 4, 5 and 6), interpolated hourly values were generated (e.g. each 0.04167 day) using the Matlab function *makima*, which performs the Modified Akima piecewise cubic Hermite interpolation.

ii. Using the interpolated experimental data, the terms $\frac{1}{X_v}\frac{d(\psi(t))}{dt}$ and $f\left(\psi(t), \bar{\phi}\right)$ were calculated at each discrete time k, for each medium, as follows:

$$\frac{1}{X_v}\frac{d(\psi(t))}{dt} = \frac{1}{X_{v,k}}\frac{(\psi_{k+1} - \psi_k)}{t_{k+1} - t_k} \tag{5.10}$$

$$f\left(\psi(t), \bar{\phi}\right) = f\left(\psi_k, \bar{\phi}\right) \tag{5.11}$$

The output response $y_{i,k}$, for a specific major metabolite i, at each time interval k was obtained from equation 5.8 as follows:

$$y_{i,k} = \frac{1}{X_{v,k}}\frac{(\psi_{i,k+1} - \psi_{i,k})}{t_{k+1} - t_k} - f\left(\psi_k, \bar{\phi}\right)\Big|_{\psi_{i,k}} \tag{5.12}$$

where $\psi_{i,k}$ denotes the concentration of major metabolite i at time k.

iii. The output response vector Y_i^j for a specific major metabolite i and medium j was generated as per equation 5.13, with $k = 1, .., T$, where T indicates the final experimental time step.

$$\mathbf{Y}_i^j = \begin{bmatrix} \frac{1}{X_{v,1}} \frac{(\psi_{i,2} - \psi_{i,1})}{t_2 - t_1} - f\left(\psi_1, \bar{\phi}\right)\big|_{\psi_{i,1}} \\ \frac{1}{X_{v,2}} \frac{(\psi_{i,3} - \psi_{i,2})}{t_3 - t_2} - f\left(\psi_2, \bar{\phi}\right)\big|_{\psi_{i,2}} \\ \vdots \\ \frac{1}{X_{v,T-1}} \frac{(\psi_{i,T} - \psi_{i,T-1})}{t_T - t_{T-1}} - f\left(\psi_{T-1}, \bar{\phi}\right)\big|_{\psi_{i,T-1}} \end{bmatrix}^j \tag{5.13}$$

iv. The input variables matrix $\mathbf{X}^j$ for each medium j was generated to include the major medium components concentration over time ($\psi_{i,k}$) in the first M columns, the minor components concentration minus the average concentration of its respective minor component at time zero ($\delta\phi_j$) in the subsequent N columns, and two-terms interactions among those major and centered minor components ($\psi_{i,k}\psi_{i+1,k}$, $\psi_{i,k}\delta\phi_j$ and $\delta\phi_j\delta\phi_{j+1}$), as shown in equation 5.14 in the last columns of the matrix. $\psi_{i,k}$ denotes the concentration of major metabolite i at time k and $\delta\phi_j$ denotes the deviation of the concentration of minor metabolite j from the average concentration of its minor component at time zero, with $i = 1, .., M$, $j = 1, .., N$ and $k = 1, .., T$, which M is number of major metabolites, N is number of minor components and T the final experimental time.

$$\mathbf{X}^j = \begin{bmatrix} \psi_{1,1} & \cdots & \psi_{M,1} & \delta\phi_1 & \cdots & \delta\phi_N & \psi_{1,1}\psi_{2,1} & \cdots & \delta\phi_{N-1}\delta\phi_N \\ \psi_{1,2} & \cdots & \psi_{M,2} & \delta\phi_1 & \cdots & \delta\phi_N & \psi_{1,2}\psi_{2,2} & \cdots & \delta\phi_{N-1}\delta\phi_N \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \ddots & \vdots \\ \psi_{1,T} & \cdots & \psi_{M,T} & \delta\phi_1 & \cdots & \delta\phi_N & \psi_{1,T}\psi_{2,T} & \cdots & \delta\phi_{N-1}\delta\phi_N \end{bmatrix}^j \tag{5.14}$$

v. The experimental data obtained for all tested media is divided into two sets: a data set of different media formulations that is used for model calibration (Set 1) and a second set obtained with different media formulations that is used for model validation (Set 2). For each one of these sets, the input variables' matrix $\mathbf{X}_{PLS}$ and the output response vector $\mathbf{Y}_{PLS,i}$ are arrange into matrices, with j being the media being used in each set:

$$\mathbf{X}_{PLS} = \begin{bmatrix} \mathbf{X}^{Medium_1} \\ \mathbf{X}^{Medium_2} \\ \vdots \\ \mathbf{X}^{Medium_j} \end{bmatrix} \tag{5.15}$$

$$\mathbf{Y}_{PLS,i} = \begin{bmatrix} \mathbf{Y}_i^{Medium_1} \\ \mathbf{Y}_i^{Medium_2} \\ \vdots \\ \mathbf{Y}_i^{Medium_j} \end{bmatrix} \tag{5.16}$$

vi. The data for set 1 are mean centered and normalized. The $\mathbf{Y}_{PLS,i}^{Set1}$ and $\mathbf{X}_{PLS}^{Set1}$ were subtracted from their respectively mean ($\bar{Y}_{PLS,i}^{Set1}$ and $\bar{\mathbf{X}}_{PLS}^{Set1}$) and divided by their respectively standard deviation ($Y_{PLS,i,\sigma}^{Set1}$ and $\mathbf{X}_{PLS,\sigma}^{Set1}$), through each column. The mean centered and normalized output response vector and input variable matrix for the calibration set are represented by $\hat{\mathbf{Y}}_{PLS,i}^{Set1}$ and $\hat{\mathbf{X}}_{PLS}^{Set1}$, respectively.

vii. The data for set 2 are also mean centered and normalized. The $\mathbf{Y}_{PLS,i}^{Set2}$ and $\mathbf{X}_{PLS}^{Set2}$ were subtracted from the mean and divided by the standard deviation, through each column, using $\bar{Y}_{PLS,i}^{Set1}$, $\bar{\mathbf{X}}_{PLS}^{Set1}$, $Y_{PLS,i,\sigma}^{Set1}$ and $\mathbf{X}_{PLS,\sigma}^{Set1}$ from the calibration set. The mean centered and normalized output response vector and input variable matrix for the validation set are represented by $\hat{\mathbf{Y}}_{PLS,i}^{Set2}$ and $\hat{\mathbf{X}}_{PLS}^{Set2}$, respectively.

viii. $\hat{\mathbf{Y}}_{PLS,i}^{Set1}$ and $\hat{\mathbf{X}}_{PLS}^{Set1}$ were then used to perform PLS regression, using the MatLab function *plsregress*. The matrix of PLS weight $\mathbf{W}$ and the vector of response loadings $\mathbf{YL}$ obtained with the PLS regression of the calibration set were used to generate predictions for both calibration and validation sets.

ix. The number of principal components nPC (latent variables) was selected such as the maximum output variability of the calibration set could be explained and, at the same time, smaller predictive error could be obtained for both calibration and validation output results. The predicted normalized output response was described by the PLS regression model as follows:

$$\hat{\mathbf{Y}}_{PLS,i}^{Set1}\Big|_{Prediction} = \left(\hat{\mathbf{X}}_{PLS}^{Set1}\right)(\mathbf{W})(\mathbf{YL}) \tag{5.17}$$

$$\hat{\mathbf{Y}}_{PLS,i}^{Set2}\Big|_{Prediction} = \left(\hat{\mathbf{X}}_{PLS}^{Set2}\right)(\mathbf{W})(\mathbf{YL}) \tag{5.18}$$

And, the original output response vector for the calibration and validation sets could be reconstructed as follows:

$$\mathbf{Y}_{PLS,i}^{Set1}\big|_{Prediction} = \left(\hat{\mathbf{Y}}_{PLS,i}^{Set1}\right)\left(Y_{PLS,i,\sigma}^{Set1}\right) + \bar{Y}_{PLS,i}^{Set1} \tag{5.19}$$

$$\mathbf{Y}_{PLS,i}^{Set2}\big|_{Prediction} = \left(\hat{\mathbf{Y}}_{PLS,i}^{Set2}\right)\left(Y_{PLS,i,\sigma}^{Set1}\right) + \bar{Y}_{PLS,i}^{Set1} \tag{5.20}$$

Once nPC, $\mathbf{W}$ and $\mathbf{YL}$ are determined as described by the steps above, the function g can be generated for each major metabolite i, at each time t as follows:

$$g\left(\psi(t), \delta\phi\right)\big|_{i,t} = \left(\hat{\mathbf{x}}_t\right)\left(\mathbf{W}\right)\left(\mathbf{YL}\right)\left(Y_{PLS,i,\sigma}^{Set1}\right) + \bar{Y}_{PLS,i}^{Set1} \tag{5.21}$$

where $\hat{\mathbf{x}}_t$ is the vector containing concentrations of major metabolites over time ($\psi(t)_i$), minor components concentration minus the average concentration of its respective minor component at time zero ($\delta\phi_j$), and two-terms interactions among those major components' concentrations and the deviations in minor components' concentrations with respect to their mean values ($\psi(t)_i\psi(t)_{i+1}$, $\psi(t)_i\delta\phi_j$ and $\delta\phi_j\delta\phi_{j+1}$), as shown in equation 5.14. Note that $\hat{\mathbf{x}}_t$ is mean-centered using $\bar{\mathbf{X}}_{PLS}^{Set1}$ and normalized by $\mathbf{X}_{PLS,\sigma}^{Set1}$.

It should also be noticed that the empirical and mechanistic parts of the hybrid model are coupled because common information, i.e. the current concentration of amino acids at each time t, is used by both f and g functions. Therefore, the hybrid model parameters including the kinetic parameters from the mechanistic part and the PLS loadings parameters, should be simultaneously re-calibrated to best fit the model predictions with the experimental data. It is important to observe that the calibration of too many parameters in a non-linear differential model is challenging task. Furthermore, the calibration of too many parameters may lead to overfitting of noise with resulting poor predictions. To address this issue only a subset of informative parameters should be selected to perform re-calibration. The selected parameters are chosen based on a sensitivity analysis, which consider the variation of the output response with respect to perturbations in each model parameter [64]. The sensitivity with respect to each parameter averaged over all responses, $S_{\vartheta_i}^{y}$, can be obtained as follows:

$$S_{\vartheta_i}^{y_j}(t_k) = \frac{\partial y_j}{\partial \vartheta_i}(\vartheta, t_k)\frac{\vartheta_i}{y_j(t_k)} \tag{5.22}$$

$$S^{y}_{\vartheta_i} = \frac{1}{n_y}\sum_{j=1}^{n_y}\left(\frac{1}{n_t}\sum_{k=1}^{n_t}\left|S^{y_j}_{\vartheta_i}(t_k)\right|\right) \tag{5.23}$$

where y is the predicted response of a metabolite j by the hybrid model, ϑ is the hybrid model parameter i, t_k is each time interval, n_t is the total number of time intervals k, n_y is the total number of response variables. Following the sensitivity analysis, only parameters with high sensitivity values were re-calibrated.

The hybrid modelling structure presented above (equation 5.8) is referred in the literature as parallel since information is simultaneously used in the mechanistic and empirical parts. Also, within this hybrid modeling approach, the function $g\left(\psi(t), \delta\phi\right)$ that are based on the PLS regression model can be incorporated into one or more equations of the system of nonlinear differential equations that describe the evolution of amino acids in the cell culture. However, it should be remembered that a key objective in any modelling exercise is to minimize the number of model parameters required to fit the data so as to reduce the over-fitting of noise. Since the fine tuning of the model involved a re-calibration of the PLS response loading vector it was very important to limit the addition of empirical corrections corresponding to the function g into the mechanistic equations.

Thus, in this study, to limit the number of model parameters, the empirical correction by the additional function g (Equation 5.8) was implemented only in the biomass equation. The rationale is that the biomass X_v concentration enters in all the differential equations of the model since the kinetic expressions in the function f are defined per unit biomass. Consequently, it was hypothesized that an improvement of the biomass prediction by the empirical correction term g will also improve the prediction of the other major metabolites. Therefore, less parameters would be needed into the new hybrid model as compared to the case where empirical corrections will be added in more differential equations of the mechanistic model.

5.4 Results

Eleven cell culture experiments in high density batch mode were used in this work for model calibration and validation. Each batch used a different media formulation. From the 11 media formulations, 9 were used for model calibration (Media 10, 17, 19, 22, 32, 51, 56, 64, 65) and 2 for model validation (Media 11 and 16). A proportion of 80%/20% was used to split the data for the different media into the calibration and validation sets, respectively. As described in Section 5.3.3 all components were normalized by their standard deviation.

Hence, Media 11 and 16 were the ones selected for the model validation because was the only combination set which the normalization of minor components did not result in division by zero at time=0.

Samples for each cell culture experiment were collected at days 0, 3, 4, 5 and 6. The concentrations for the major metabolites including viable cells, glucose, lactate, ammonia and amino acids over time, for each medium, can be found in Tables E.1 to E.11 in Appendix E. The concentrations are reported in normalized values from 0 to 1 for confidentiality reasons. The normalized concentrations for the minor components (hormones, vitamins, lipids, etc.) for each medium formulation is shown in Tables E.12 to E.14 in Appendix E. It is also important to emphasize that minor components concentrations were only available at the initial time. Based on communications with our industrial collaborator it is impractical to attempt to measure the concentrations of minor metabolites over time due to their presence in small amounts in the culture supernatant. Also, for confidentiality reasons, our industrial collaborator was only able to provide the amounts but could not identify the minor components by name. Each of the concentration values is normalized between 0 and 1.

5.4.1 Amino Acids Concentration measured by HPLC analysis

Figure 5.1 shows the amino acids chromatogram for the calibration standard at the highest concentration level (20 pmol/ul) used. The chromatogram shows good separation between the peaks except for the Asparagine peak that it is not identified in the Figure 5.1. Asparagine cannot be clearly identified due to its relatively low concentration and to its closeness to the Aspartate peak in the HPLC chromatogram. A calibration curve was built using four levels of amino acids standard for each of the days 0, 3, 4, 5 and 6. The tables showing the values of the amino acids concentration for different times as measured by HPLC analysis can be found in Appendix E.

5.4.2 Mechanistic Model

Following the procedure presented in the previous section the cell culture experiments chosen for model calibration (calibration set) were used to estimate the parameters of the mechanistic model described by the system of non-linear differential equations shown in Appendix C. The mechanistic model has a total of 83 parameters. Table D.1 in Appendix D shows the estimated parameters values for the mechanistic model. The concentration profiles as a function of time that are predicted by the mechanistic model exhibit significant

errors with respect to the data, especially for viable cells, mAb and glucose concentrations (see Figures F.1 to F.11 in Appendix F).

To assess whether the mechanistic model has structural error due to the absence of the effect of minor components in the model the concept of autocorrelation and the correlogram were used. The autocorrelation function presents a serial correlation (autocorrelation) among the errors of the mechanistic model. If the errors are random and consequently uncorrelated, the autocorrelation should be 1 at zero lag and within the noise limits for any other lag. Figure 5.5a presents the correlogram plot for lag-1, lag-2, and lag-3 for the errors in viable cells (vcd), glucose (glc), and monoclonal antibody (mAb) based on the mechanistic model predictions. The correlogram was obtained using the prediction errors at day 3, 4, 5 and 6 for all media. Assuming 95% confidence bounds, the sample autocorrelation function analysis for lag-1 and lag-2 were considered significant for vcd, glc, and mAb error prediction thus implying that further improvement is possible. The addition of the empirical model correction is expected to provide such improvement.

The incorporation of the empirical model in the current model aims to describe the influence of minor components on major metabolites concentration that is not captured by the mechanistic approach. As mentioned in the previous section the empirical model correction was added into the biomass (viable cells - vcd) differential equation only.

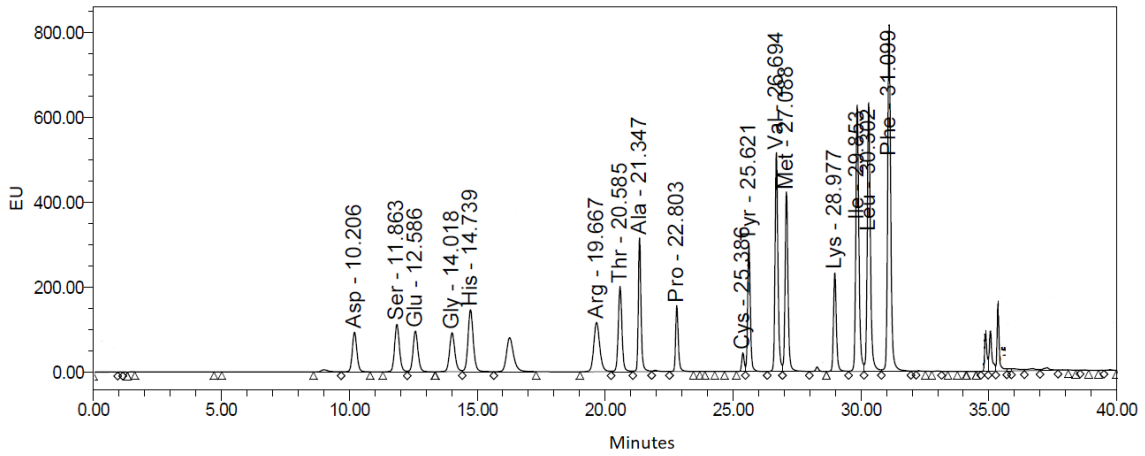


Figure 5.1: Amino acids separation peaks in plot of EU HPLC unit per minute time.

5.4.3 Empirical Model

The empirical model correction to be added in the biomass differential equation was generated based on the PLS regression, as described in section 5.3.3. The correction described the difference observed in the derivative $\frac{1}{X_v}\frac{d(\psi(t))}{dt}$ and $f(\psi(t), \phi)$ for biomass (viable cells - vcd) concentration results.

To train the empirical correction term the data was interpolated hourly, e.g. each 0.04167 day, using the Matlab function *makima*. The input PLS matrix $\mathbf{X}_{PLS}$ considered linear terms of major and minor metabolites concentration, and possible two-way interactions among them, resulting a total of 6328 input variables (as described in equations 5.9 and 5.14). Eleven principal components (latent variables) were chosen which explained approximately 81% of the output response variability of the calibration set. With this number of latent variables the resulting calibration and validation errors were of similar magnitude, 0.43 and 0.58 respectively. This indicated that the calibrated model did not significantly over-fit the calibration data.

The variable importance in projection (VIP) scores were calculated for the PLS regression. The VIP scores can be used to infer the importance of each latent variable in the PLS regression model and is generally used to identify important predictors, specially when large multicollinearity exists among input variables [101, 31]. The equation used to calculate the VIP scores can be found in Chong and Jun [31]; input variables presenting a VIP score greater than 1 are considered important. The VIP scores identified 1126 important input variables out of the total 6328 original input variables. Several interactions among glucose and amino acids and also between amino acids were found significant for biomass profile. Also, interactions between glucose and minor components and interactions between serine and minor components were the most common predictors affecting biomass performance. Components 7, 23, 26, 66, and 70 are the most common minor components found significant interacting with amino acids. Based on recent communications with our industrial partner all components that were found to be relevant for our model they were known to have significant impact on the cell culture.

The 6328-by-11 weight matrix $\mathbf{W}$ and the 1-by-11 vector of response loadings $\mathbf{YL}$ obtained with the PLS regression were used to define the empirical correction of the biomass balance in the hybrid model. The weight matrix $\mathbf{W}$ projects the original 6328 input variables into a space of 11 latent variables. As mentioned in the methodology section, the values of the vector of response loadings $\mathbf{YL}$, which corresponds to the parameters from the PLS regression, can be further re-calibrated in the final hybrid model. In this regard, the significant advantage of using PLS is that due to the projection onto latent variables the number of parameters that can be further re-tuned is drastically lower than

the number that would be required if a multiple linear regression with 6328 would be used. The current coefficient estimates values for the PLS regression, before re-calibration, and their corresponded lower and upper confidence bounds for each coefficient, assuming 95% confidence, are given in Table 5.2.

Table 5.2: PLS regression coefficient values (**YL**) estimate and confidence bounds, assuming 95% confidence.

	Coefficient Estimate	Lower Confidence Bound	Upper Confidence Bound
YL(1)	16.77	15.92	17.62
YL(2)	8.66	7.81	9.51
YL(3)	16.22	15.37	17.07
YL(4)	10.19	9.34	11.04
YL(5)	10.82	9.97	11.67
YL(6)	11.14	10.29	11.99
YL(7)	5.05	4.20	5.90
YL(8)	4.46	3.61	5.31
YL(9)	2.79	1.94	3.63
YL(10)	4.28	3.44	5.13
YL(11)	4.23	3.39	5.08

5.4.4 Hybrid Model

The empirical $g(\psi(t), \delta\phi)$ function based on the PLS regression was incorporated into the biomass non-linear differential equation of the mechanistic model, as described by equations 5.8 and 5.21. It should be remembered that the empirical PLS model was calibrated, as described in the previous section, with the mechanistic model parameters set at their original calibrated values. It was hypothesized that due to the coupling between the empirical and mechanistic parts, further improvements could be obtained by simultaneous re-calibration of the mechanistic and empirical parts of the hybrid model. Thus, the comparisons in this section will involve 3 different models as follows: 1- mechanistic, 2-hybrid before re-calibration and 3-hybrid after re-calibration.

Since the hybrid model involves a large number of parameters it was computationally prohibitive to re-calibrate all the parameters. In addition, due to the limited amount of data, re-calibrating all the parameters may have resulted in over-fitting. Therefore, a sensitivity analysis was conducted to find a subset of parameters that have most impact on

the results and which re-calibration may provide the largest improvements. The sensitivity analysis of the hybrid model was conducted with respect to parameter perturbations varying from ±10 to 20%. A parameter that resulted in an average sensitivity larger than 50 was deemed significant. The choice of this threshold was mostly driven by the number of parameters that we could re-calibrate within an acceptable window of time and that improved the results in terms of the AIC (Akaike Information Criterion). However, the threshold has been found by trial and error but may be further optimized. The sensitivity analysis identified a total of 33 parameters to be sensitive that are directly related to the metabolites biomass, glucose, lactate, ammonia, aspartate, alanine, glutamate and serine. The first 10 PLS regression coefficients were also found among the parameters identified as significant by the sensitivity analysis. Table D.2 in Appendix D shows the parameters values for the hybrid model.

Notice that 11 new parameters were added to the hybrid approach. Therefore, a fair comparison of the mechanistic model versus the hybrid models, before and after re-calibration, should take into account the larger number of parameters that are present in the models. For this purpose the estimation of prediction error relative to the number of model parameters was evaluated for the mechanistic and hybrid models by using the AIC. The AIC can be calculated as function of the residuals sum of squares [20] according to the following equation:

$$AIC = n\ln\left(\frac{\Sigma\epsilon_i^2}{n}\right) + 2K \tag{5.24}$$

where n is the number of samples, ϵ is the residuals, and K the number of parameters. Smaller AIC values indicate more accurate model predictions regardless of the larger number of model parameters.

Figures F.1 to F.11 in Appendix F compares the concentration data with the predictions obtained with the mechanistic model, the hybrid model prior to the final re-calibration, and the hybrid model after re-calibration for all major metabolites for each medium. Figures 5.2, 5.3, and 5.4 compares data with model predictions for biomass (viable cells - vcd), glucose (glc) and monoclonal antibodies (mAb) concentration. The plots show that the hybrid approaches, specially the re-calibrated hybrid model, provided a most accurate fit to the data than the mechanistic model for most of the media and major metabolites.

Although the amino acids predictions of the re-calibrated hybrid model is only slightly better than the predictions of the mechanistic model, the predictions for biomass, glucose and mAb given by the re-calibrated hybrid model are significantly better than the ones given by the mechanistic approach. It should be emphasized that biomass, mab and glucose (main nutrient) are probably the most important variables for future process optimization.

Furthermore, re-calibration of the hybrid model improve the predictions as compared to the hybrid model before re-calibration as shown below.

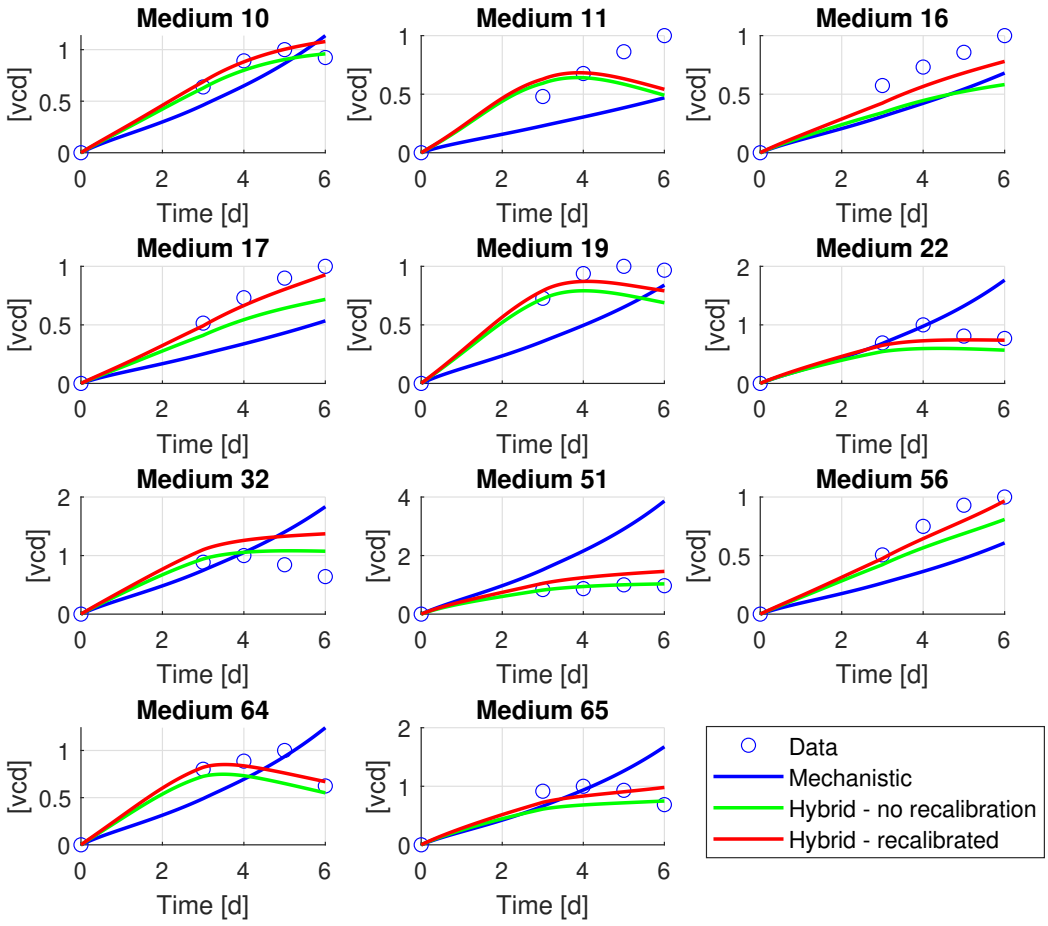


Figure 5.2: Viable cells concentration ([vcd]) profile given by data, mechanistic model and hybrid model. The units of [vcd] are given in 10^6 cells/ml/mM of Glc.

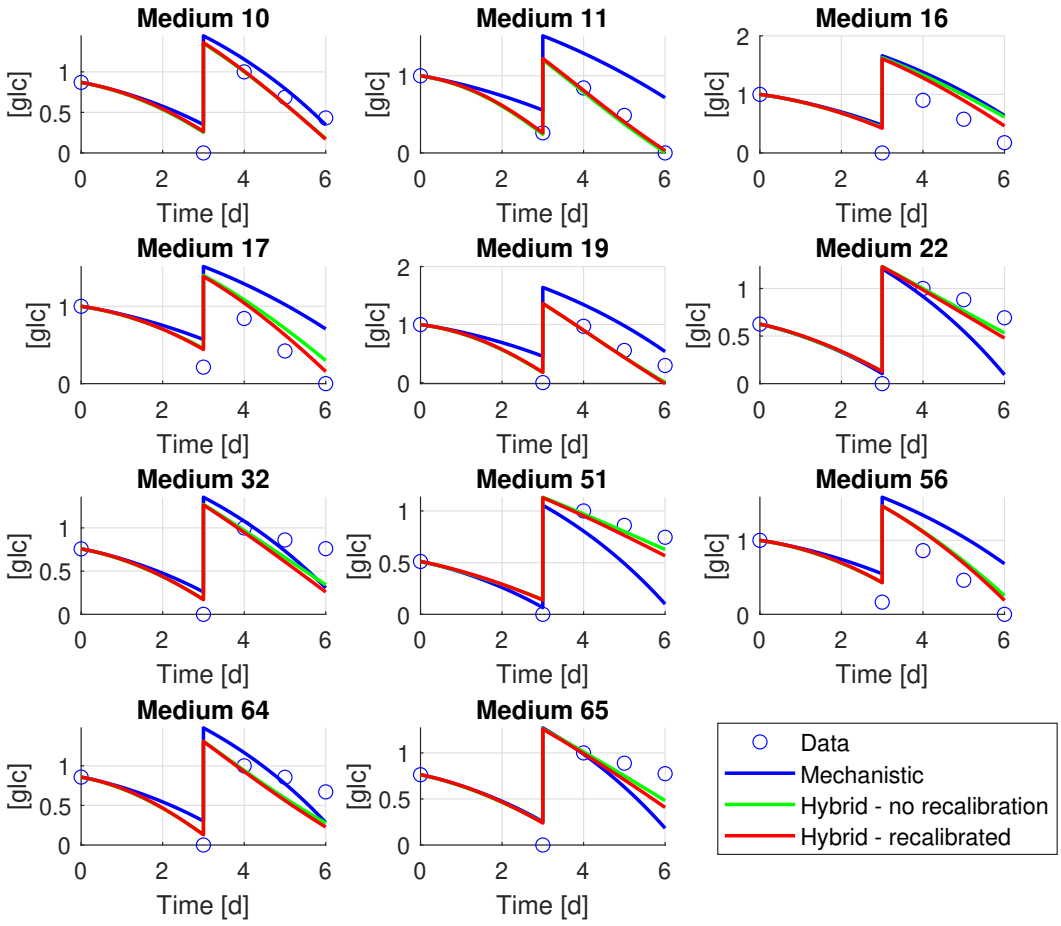


Figure 5.3: Glucose concentration ([glc]) profile given by data, mechanistic model and hybrid model. The units of [glc] are given in mM/mM of Glc.

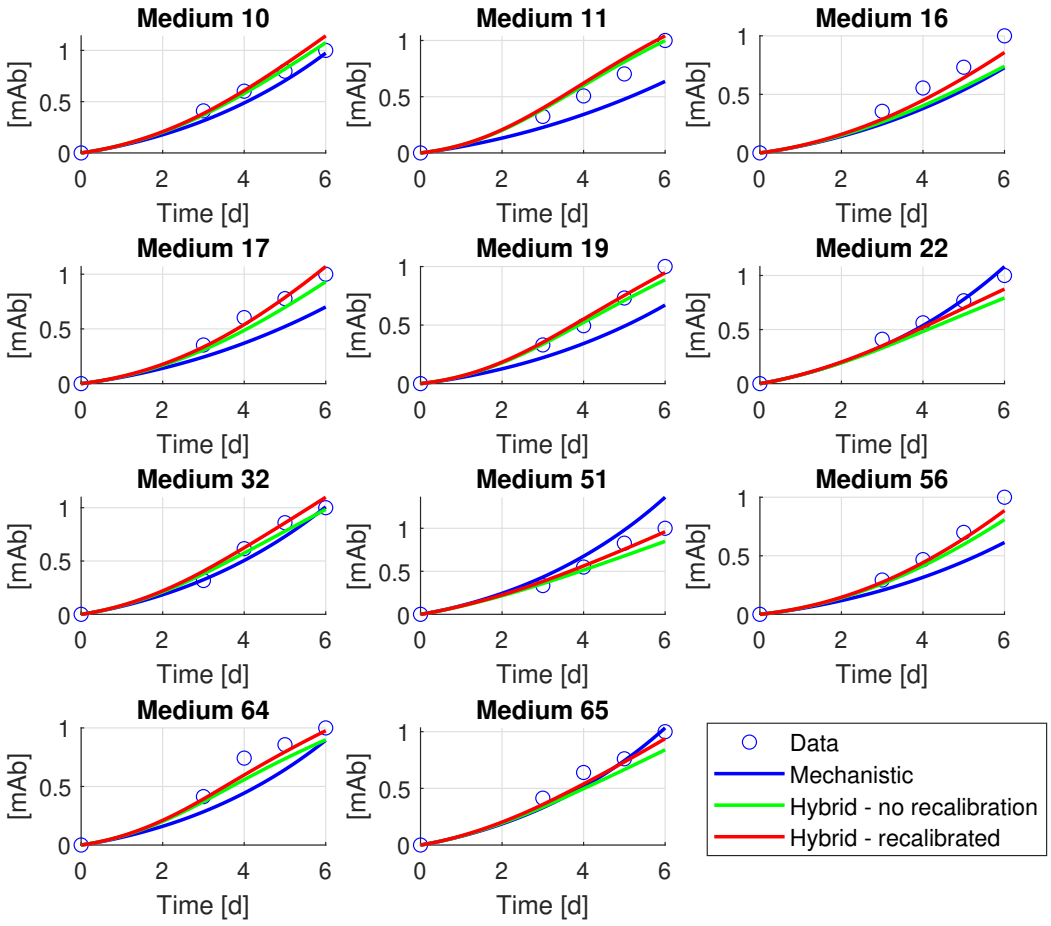


Figure 5.4: Monoclonal antibody concentration ([mAb]) profile given by data, mechanistic model and hybrid model. The units of [mAb] are given in mg/L/mM of Glc.

Table 5.3 compares the AIC, given by each media, for the mechanistic model, hybrid model prior to the re-calibration and hybrid model after re-calibration. Significantly smaller AIC values are obtained by most of the media when the hybrid approach is used. AIC values for the hybrid model after parameter re-calibration are even smaller, for both calibration (Media 10,17, 19, 22, 32, 51, 56, 64 and 65) and validation (Media 11 and 16) sets. The average AIC for all media is about 23% smaller for the re-calibrated hybrid model compared to the original mechanistic model. The average AIC values for the calibration and validation sets for the hybrid approach after re-calibration, 420.27 and 560.72 respectively, are also smaller than the AIC average values for the mechanistic model, 562.67 for calibration set and 671.54 for validation set, i.e. about 25% and 17% smaller respectively.

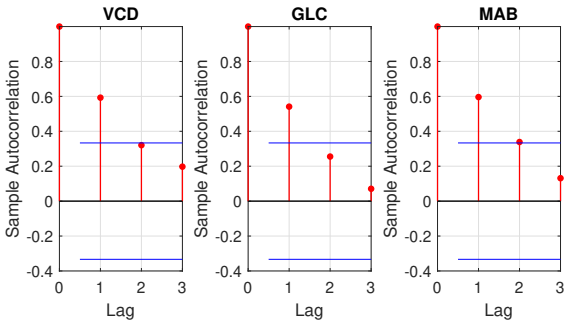
Table 5.3: AIC values given by the mechanistic and hybrid models for each media.

	Mechanistic	Hybrid no recalibration	Hybrid recalibrated
Medium 10	424.99	325.08	402.23
Medium 11	702.66	561.34	586.72
Medium 16	640.42	622.06	534.72
Medium 17	678.05	480.81	448.06
Medium 19	675.87	428.51	398.92
Medium 22	422.35	487.55	385.64
Medium 32	469.42	405.40	434.76
Medium 51	595.82	427.72	336.80
Medium 56	718.16	568.24	517.03
Medium 64	606.86	506.31	438.95
Medium 65	472.50	506.35	420.07
Sum	6407.11	5319.37	4903.92
Average Calibration	562.67	459.55	420.27
Average Validation	671.54	591.70	560.72
Average All	582.46	483.58	445.81

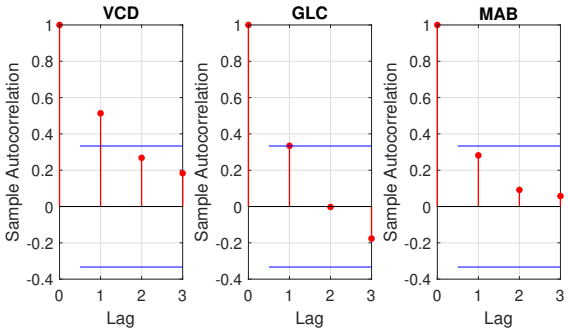
To further test whether the hybrid model corrects for the model structure error Figure 5.5 presents the autocorrelation error for lag-1, lag-2, and lag-3 for viable cells (VCD), glucose (GLC), and monoclonal antibody (MAB) for the mechanistic and hybrid models results. Notice that both hybrid models, before and after re-calibration, present better autocorrelation results than the ones given by the mechanistic model. Comparing the autocorrelation results for the hybrid model before and after re-calibration, it is observed

a slightly improvement in the VCD autocorrelation after re-calibration of hybrid model parameters, reduced from 0.5135 to 0.5012. Also, while autocorrelation for GLC become insignificant for the re-calibrated hybrid model in comparison to the results before re-calibration, the MAB lag-1 autocorrelation is still siginifcant for the re-calibrated hybrid model, but observe that parameters related to MAB were not re-calibrated (were not selected in the sensibility analysis). Although the sample autocorrelation is still present for lag-1 in VCD and MAB results in the hibrid re-calibated model, the autocorrelation for GLC is found insignificant and the overall autorrelation for VCD, GLC and MAB given by the re-calibrated hybrid prediction model are about 15% to 30% smaller than the one given by the mechanistic model. Furthermore, the autocorrelation for lag-2 and lag-3 become insignificant for VCD, GLC and MAB when the re-calibrated hybrid approach is used.

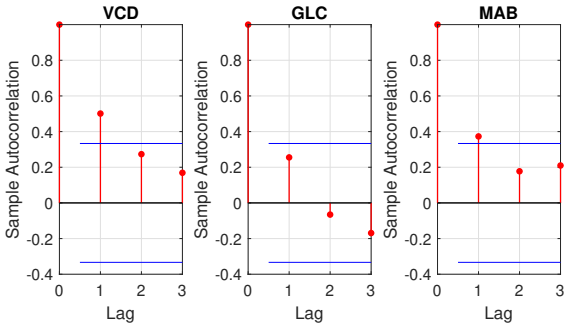
The presence of model structure error that remains according to the lag-1 autocorrelation error that it is still present can be explained by different reasons such as: i-the minor components concentrations were accounted for only through their initial concentrations while in reality these components may be consumed or produced along the batch; ii- empirical corrections were only introduced in the biomass equation while correcting for the remaining error may require adding empirical corrections in balances of other metabolites, e.g. glucose; and iii- the PLS model was calibrated using derivatives obtained from interpolated data.



(a) Autocorrelation error for mechanistic model.



(b) Autocorrelation error for hybrid model before recalibration.



(c) Autocorrelation error for hybrid model afer recalibration.

Figure 5.5: Correlogram for mechanistic and hybrid models errors for the calibration set. The red line/dots represents the sample autocorrelation for a specific lag and the blue lines represents the upper and lower autocorrelation confidence bounds, assuming 95% confidence bounds.

5.5 Conclusion

A parallel hybrid approach was proposed in order to incorporate the effect of minor components present in cell culture media into a mechanistic model of major metabolites. Samples at day 0, 3, 4, 5 and 6 were collected from batch experiments that used 11 different media formulations. Amino acids concentration were measured using HPLC analysis. Minor elements concentration were only available at time zero. Experiments using 9 different media were used for model calibration, while 2 media were used for model validation.

Following a Taylor expansion rationale the hybrid model consists of a superposition of the mechanistic model terms and an empirical correction term that describes the effects of minor components and possible interaction among all medium components. The empirical correction is introduced through a polynomial function containing empirical information based on a PLS regression. A methodology to obtain the PLS coefficients was presented. At the moment, the PLS regression function was incorporated only at the biomass nonlinear differential equation from the mechanistic model. The biomass concentration is present in all other differential equations of the mechanistic model, therefore, a better fit in biomass would also improve the prediction of all other major metabolites. Although additional correction terms could be incorporated in mass balances of metabolites other than biomass the aim was to keep a small number of new parameters in the model. A subset of parameters from the proposed hybrid model, selected by sensitivity analysis, was re-calibrated to improve model predictions.

The results for the re-calibrated hybrid parallel model prediction have shown good improvement of fitting between data and predictions compared to the original mechanistic model, specially for the main major metabolites, biomass (viable cells), glucose and mAb. The AIC values proved the fidelity of the hybrid model over the mechanistic model despite the additional model parameters that were considered. Although the prediction error of the re-calibrated hybrid model indicates the presence of autocorrelation for lag-1 in VCD and MAB, these values were smaller than the ones presented for the mechanistic model. On the other hand the autocorrelation for lag-1 in GLC is found insignificant when the re-calibrated hybrid model is used, as well as the autocorrelation for lag-2 and lag-3 for VCD, GLC and MAB, indicating that the re-calibrated hybrid model resulted in net reduction of model structure error.

In conclusion, the re-calibrated hybrid model performed better predictions than the original mechanistic model. The autocorrelation error for lag-1 presented by the hybrid model may be related to two main factors: i- the potential depletion of minor medium components over time are not accounted for and ii- the empirical correction was only added

in the biomass equation . Thus, further reduction of the auto-correlation are expected for the hybrid approach with the addition of other PLS regression functions into the set of non-linear differential equations of the mechanistic model.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Since 2000 the therapeutic market for monoclonal antibodies (mAbs) has increased exponentially. The pharmaceutical industry is continuously searching for more efficient methods to supply the increasing demand for monoclonal antibodies. Several factors influence the cellular production performance and among them the medium used for cell culture is a crucial determinant of cell growth and mAb productivity. The cost of cellular medium is high and the experimental procedures to develop new formulations are also costly and time consuming. Mathematical modelling is an useful tool to understand the production/consumption of medium components and to optimize the culture performance. Also, empirical mathematical modelling is used this thesis, for reducing costs during optimal medium development.

As discussed in the thesis the mathematical modelling of bioprocess systems is a challenging task. The main challenges addressed in this thesis are related to: i- the complexity of the cell culture medium and ii- dynamic perfusion operation. These challenges are further exacerbated by the fact that the metabolism of mammalian cells is more complex as compared to bacteria. Hence, modelling approaches such as dynamic metabolic flux models (DMFM) that work well to describe bacterial cultures require further modifications for their application to mammalian cultures. In view that pharmaceutical companies experiment and implement highly variable media they require robust models that can correctly describe the effect of changes in concentrations of media components. An added difficulty is that for many of the minor components in media, i.e. elements that are present in low concentrations, their effect on cellular metabolism is unclear. Furthermore, because of

confidentiality, part of the media components could not be identified by name and thus their effect on the culture could be solely inferred from data. To address these challenges this work investigates different modelling approaches including mechanistic empirical and hybrid models. The different models are then used in different applications as follows: i- dynamic metabolic flux models (DMFM) are used with one single media to describe batch and perfusion operations (Chapter 3), ii- empirical models are used for media optimization such as productivity is maximized and for reducing the experiments needed for media development (Chapter 4) and iii- hybrid models are used to describe dynamic responses of metabolites for different media compositions (Chapter 5). More detailed conclusions are provided below for each one of these three parts of the thesis.

Chapter 3 presents the DMFM model for mammalian cells in which a systematic approach was applied to identify the limiting constraints of the model. The goal was to minimize the number of constraints needed to describe the batch and perfusion dynamic data. The work was based on a preliminary model reported by Nikdel [114]. This model was extended to include dynamic biomass predictions by the model, consideration of cell death behavior in the modelling and application to both batch and perfusion system operations. Active constraints in glutamine, glutamate, lactate, and ammonia were required to describe the data during the batch operation while in the earlier model constraints for 6 amino acids were required. The constraints in this work were expressed by kinetic functions that follow the Hill equation. At this point, time-varying constraints in the 4 limiting amino acids were necessary during perfusion since the Hill expressions that were found for batch did not provided good fitting for perfusion. It may be necessary to use different constraint related expressions during perfusion from the ones during batch. A possible reason for why different kinetic expressions may be needed during perfusion is that the dynamic of intermediates was ignored. In view that the dynamic of intermediates will depend on the mode of operation, batch or perfusion, different kinetic expressions may be needed for different mode of operations to compensate for the ignored dynamics of intermediate metabolites. A key finding in this work was that biomass flux was also an active constraint maintaining a constant value during batch and perfusion operation. This was an important observation that helped to improve the fitting of data. It also implies that the biomass growth is not limited by the consumption of the measured amino acids as assumed in Nikdel's approach. The biomass flux constraint indicates that the dynamic of the measured amino acids are not sufficient to describe the growth rate, which may indicates that other metabolites such as ribose or NADPH can be limiting to biomass growth. It was also observed that the cell death had to be considered in order to fit the evolution of cell mass during perfusion operation. Cell death/apoptosis are known to be significant in mammalian cells and thus it is not surprising that its inclusion improves the

model considerably. The fact that biomass flux was an active constraint during the entire operation required the biomass balance to be accurate in order to improve the overall fitting of the model. Thus, in view that the inclusion of the cell death improved the biomass fitting, it also indirectly improve the fitting of the other measured metabolites. In addition to the identified kinetic constraints, soft constraints were also required to limit the flux solution space and to address the multiplicity of solutions of the LP problem. These soft constraints consisted of upper or lower limits that are kept constant over time. In conclusion, the DMFM approach was proved to be very satisfactory, being able to predict the batch and perfusion data with an average error of 15%. Furthermore, the final DMFM approach was able to provide equivalent accuracy during batch operation using a smaller number of parameters when compared to conventional kinetic models, such as the one used in Chapter 5. While the DMFM requires a total of 40 parameters, the differential kinetic model requires over 80 parameters to obtain the same level of prediction error.

The focus of the study in Chapter 4 was to model the effect of changes in concentrations of minor components in media formulations and to use the resulting models for optimizing productivity with respect to media composition. Since part of the components were not explicitly identified due to confidentiality the problem was approached by empirical modelling followed by robust optimization based on the identified model. The particular challenge presented by the large number of components in the media is that the number of samples is smaller than the number of variables ($n < p$ problem). To avoid over-parameterization of the model regression approaches such as Lasso and Elastic Net were investigated. However, some of these regression approaches were found inaccurate for the $n < p$ problem. To develop a suitable empirical modelling approach a toy example was presented to compare different regression models. The results shown that regularization methods, such as Lasso and Elastic net, may neglect important coefficients during model training while PCR and PLS were found to be a better approach by considering all components. In Case Study 2 of Chapter 4, both PCR and PLS model results showed good prediction accuracy with data. The robust optimization proposed new optimum media for both simulated and industrial scenarios. It was observed that the prediction accuracy largely depends on the proximity of new media formulations within the input space defined by the principal components or principal latent variables of the trained PCR and PLS models. Case Study 3 showed that a proposed D-optimal based DOE approach was able to reduce the number of experiments required for further media optimization by selecting a smaller subset of media that was capable to describe most of the variance in the data set. This result has industrial importance because the experimentation required for media optimization is very time consuming and costly. Experimental results showed that the optimization based on a PCR model resulted in new media formulations that present equivalent Qp values to the

highest Qp values obtained by the industrial partner within the margin of measurement error. On the other hand the new formulations resulted in consistently higher cell density and mAb values throughout the cell culture as compared to the existent formulations. The inaccurate predictions of the empirical models observed in the current work are a direct result of the inability of empirical models to extrapolate beyond the region of data used for model training. The use of upper bounds on productivity were instrumental for avoiding too large extrapolations beyond the range of training data.

Chapter 5 presented a parallel hybrid approach that was developed with the aim of incorporating the effect of minor components present in cell culture media into a mechanistic model of major metabolites. 11 different media formulations were used for model training and validation. Amino acids concentrations through time were measured using HPLC analysis while minor elements concentration were only made available at time zero. The structure of the hybrid model was justified by a Taylor expansion argument. The effects of the minor elements were incorporated into a preliminary mechanistic model of major metabolites through an empirical correction based on PLS regression. A crucial advantage of the hybrid structure is that by using the PLS compression the numbers of tuning parameters of the empirical part is reduced to the number of principal components that explain most of the variability that cannot be explained by the mechanistic part of the model. The PLS based correction was used to describe the effects of minor components and possible interaction among all medium components. The PLS regression function was incorporated only in the biomass non-linear differential equation from the mechanistic model. The aim was to add as small number of new parameters as possible and since biomass concentration is present in all other differential equations of the mechanistic model, a better fit in biomass would also expected to improve the prediction of all other major metabolites. The re-calibrated hybrid model predictions showed significant improvement in predictions as compared to the original mechanistic model, especially for biomass, mAb and glucose. The Aikaike Information Criteria (AIC) calculations confirmed the better prediction accuracy of the re-calibrated hybrid approach despite the larger number of parameters that were used as compared to the original mechanistic model. It was also observed that the re-calibrated hybrid model resulted in reduction of the autocorrelation lag-1 error thus indicating a reduction in the dynamic model structure error. Moreover, the lag-2 and lag-3 autocorrelation errors became insignificant for the re-calibrated hybrid model as compared to the original mechanistic model. Overall, the hybrid model provided more accurate predictions than the original mechanistic model.

In conclusion, this thesis provided a comprehensive study of empirical, mechanistic and hybrid modelling approaches to describe and understand the performance of CHO cells in different cellular medium. The influence of different media formulations on cellular growth

and productivity was studied and an approach to reduce the number of experiments during media optimization procedure was provided.

6.2 Future Work

6.2.1 Dynamic metabolic flux model approaches

i. One of the main challenges with the dynamic metabolic flux modelling approach is the multiplicity of solutions present with this approach. Multiplicity was explicitly observed in the current work and was a major obstacle for fitting the data. To address multiplicity it is important to consider the addition of constraints other than kinetic constraints that were already considered in the current work. Artificial constraints such as the soft constraints used in the current work are helpful but are not directly motivated by the biochemical phenomena underlying the process. Instead, biochemically motivated bounds such as thermodynamic Gibbs energy related constraints and genetic regulatory related constraints should be investigated.

ii. If measurements for intracellular intermediates become available, e.g by isotope based LC-MS measurements, they could also be incorporated into the model.

iii. Other objective functions can be considered for constrained optimization, such as maximization of the entropy production rate or a weighted combination of different single objective functions since mammalian cells may pursue different goals during different phases of the cell culture.

iv. Hybridization of dynamic metabolic flux models should be investigated (see further discussion below under the item of future work for hybrid models).

6.2.2 Robust medium optimization using empirical models

i. Media optimization should be further explored by the empirical approaches proposed in this work. A key challenge found in this thesis is that the empirical model is inaccurate far away from the region of training data. To address this problem PCR or PLS empirical models should be developed favoring a particular region of interest, e.g. around a pre-defined medium, by weighting more the information given by this particular medium. This can be done by repeating this particular medium composition several times in the input variables matrix. Then, based on this enhanced local

fitting approach, the results could be progressively improved through a run to run optimization procedure.

ii. If the dynamic profiles of key minor components, such as citrate or pyruvate, could be measured their transient behavior can be incorporated into the empirical model to better describe their interactions with other medium components over time.

iii. To further improve the extrapolation accuracy the negative correlations found between cell growth and productivity could be exploited in the future to improve model predictions with the available data.

6.2.3 Hybrid models

i. The current parallel hybrid model can be further extended by incorporating the influence of minor components in other non-linear differential equations used to described the mechanistic behavior of major metabolites. Hybrid approach can better explore the influence of minor elements on other medium components if dynamic information of key minor components could be included in the model. Of course the incorporation of additional empirical corrections will require the calibration of a larger number of parameters. to mitigate this problem either a larger amount of data, if available, should be used for model calibration or, alternatively, parametric sensitivity analysis should be done to identify the parameters that have most effect on the results.

ii. Incorporation of empirical information into DMFM models presented in Chapter 3, can be very advantageous since these type of models requires less number of parameters as compared to the mechanistic models used in Chapter 5. For instance, empirical models can be incorporated in a serial approach into the DMFM by considering time varying stoichiometric matrix coefficients. These coefficients could be expressed as empirical functions of measured metabolites and perfusion rates. Allowing the stoichiometric coefficients to change with time can be used as a mean to compensate for transients in unmeasured concentrations of intermediates that were assumed in DMFM to be at quasi-steady state. This could also justify the use of different kinetic constraints during batch and perfusion operations as discussed above in the conclusions of Chapter 3.

References

[1] Hervé Abdi. Partial least squares regression and projection on latent structure regression (pls regression). *Wiley interdisciplinary reviews: computational statistics*, 2(1):97–106, 2010.

[2] Vahid Abolghasemi, Saideh Ferdowsi, Bahador Makkiabadi, and Saeid Sanei. On optimization of the measurement matrix for compressive sensing. In *2010 18th European Signal Processing Conference*, pages 427–431. IEEE, 2010.

[3] Gary K. Ackers and D. Wayne Bolen. The Gibbs Conference on Biothermodynamics: Origins and evolution. *Biophysical Chemistry*, 64(1-3):3–5, 1997.

[4] Piyush Agarwal and Arun K. Tangirala. Reconstruction of missing data in multivariate processes with applications to causality analysis. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, 9(4):196–213, 2017.

[5] Hengameh Aghamohseni, Kaveh Ohadi, Maureen Spearman, Natalie Krahn, Murray Moo-Young, Jeno M. Scharer, Mike Butler, and Hector M. Budman. Effects of nutrient levels and average culture pH on the glycosylation pattern of camelid-humanized monoclonal antibody. *Journal of Biotechnology*, 186:98–109, 2014.

[6] Woo Suk Ahn and Maciek R. Antoniewicz. Towards dynamic metabolic flux analysis in cho cell cultures. *Biotechnology Journal*, 7(1):61–74, 2012.

[7] Amr S. Ali, Ravali Raju, Somak Ray, Rashmi Kshirsagar, Alan Gilbert, Li Zang, and Barry L Karger. Lipidomics of cho cell bioprocessing: relation to cell growth and specific productivity of a monoclonal antibody. *Biotechnology journal*, 13(10):1700745, 2018.

[8] Nur Afny C. Andryani, Kadek Dwi Pradnyana, and Dadang Gunawan. The critical study of mutual coherence properties on compressive sensing framework for sparse

reconstruction performance: Compression vs measurement system. In *Journal of Physics: Conference Series*, volume 1196, page 012074. IOP Publishing, 2019.

[9] Veronica Avello, Bethzabeth Tapia, Mauricio Vergara, Cristian Acevedo, Julio Berrios, Juan G. Reyes, and Claudia Altamirano. Impact of sodium butyrate and mild hypothermia on metabolic and physiological behaviour of CHO TF 70R cells. *Electronic Journal of Biotechnology*, 27:55–62, 2017.

[10] Md Bahadur Badsha, Hiroyuki Kurata, Masayoshi Onitsuka, Takushi Oga, and Takeshi Omasa. Metabolic analysis of antibody producing Chinese hamster ovary cell culture under different stresses conditions. *Journal of Bioscience and Bioengineering*, 122(1):117–124, 2016.

[11] Yunling Bai, Changjian Wu, Jia Zhao, Yan-Hui Liu, Wei Ding, and Wai Lam W. Ling. Role of iron and sodium citrate in animal protein-free CHO cell culture medium on cell growth and monoclonal antibody production. *Biotechnology Progress*, 27(1):209–219, 2011.

[12] Samuel Bandara, Johannes P. Schlöder, Roland Eils, Hans Georg Bock, and Tobias Meyer. Optimal experimental design for parameter estimation of a cell signaling model. *PLoS computational biology*, 5(11):e1000558, 2009.

[13] Afonso S. Bandeira, Edgar Dobriban, Dustin G. Mixon, and William F. Sawin. Certifying the restricted isometry property is hard. *IEEE transactions on information theory*, 59(6):3448–3450, 2013.

[14] Bassem Ben Yahia, Laetitia Malphettes, and Elmar Heinzle. Macroscopic modeling of mammalian cell growth and metabolism. *Applied Microbiology and Biotechnology*, 99(17):7009–7024, 2015.

[15] Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer. *Biochemistry.* W H Freeman, 5th edition, 2002.

[16] Dimitris Bertsimas and John Tsitsiklis. *Introduction to Linear Optimization.* Athena Scientific, 1st edition, 1997.

[17] Nathan Braniff and Brian Ingalls. New opportunities for optimal design of dynamic experiments in systems and synthetic biology. *Current Opinion in Systems Biology*, 9:42–48, 2018.

[18] Matthias Brunner, Klara Kolb, Alena Keitel, Fabian Stiefel, Thomas Wucherpfennig, Jan Bechmann, Andreas Unsoeld, and Jochen Schaub. Application of metabolic modeling for targeted optimization of high seeding density processes. *Biotechnology and Bioengineering*, 118(5):1793–1804, 2021.

[19] Fred B. Bryant and Paul R. Yarnold. Principal-components analysis and exploratory and confirmatory factor analysis. *Reading and understanding multivariate statistics*, page 99–136, 1995.

[20] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach 2nd Edition*. Springer, New York, New York, 2002.

[21] Michael Butler. Animal cell cultures: recent achievements and perspectives in the production of biopharmaceuticals. *Applied Microbiology and Biotechnology*, 68(3):283–291, 2005.

[22] Michael Butler, Thomas Hassell, Christopher Doyle, Susan Gleave, and Philip Jennings. The effect of metabolic by-products on animal cells in culture. In *Production of biologicals from animal cells in culture*, pages 226–228. Elsevier, 1991.

[23] Heino Büntemeyer and Jürgen Lehmann. *The Role of Vitamins in Cell Culture Media*, pages 204–206. 2001.

[24] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[25] Emmanuel J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[26] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[27] Mariana Carvalho, Ali Nikdel, Rubin Hille, and Hector Budman. A comparison between dynamic modelling approaches for mammalian cells. Unpublished, work presented at the XXIX Interamerican Congress of Chemical Engineering Incorporating the 68th Canadian Chemical Engineering Conference (CsChe 2018), 2018.

[28] Mariana Carvalho, Ali Nikdel, Jeremiah Riesberg, Delia Lyons, and Hector Budman. Identification of a dynamic metabolic flux model for a mammalian cell culture. Unpublished, work submitted to the 12th IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems (DYCOPS 2019).

[29] Mariana Carvalho, Jeremiah Riesberg, and Hector Budman. Development of new media formulations for cell culture operations based on regression models. *Bioprocess and Biosystems Engineering*, 44(3):453 – 472, 2021.

[30] Zhao Chen, Jianqing Fan, and Runze Li. Error variance estimation in ultrahigh-dimensional additive models. *Journal of the American Statistical Association*, 113(521):315–327, 2018.

[31] Il-Gyo Chong and Chi-Hyuck Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems*, 78(1-2):103–112, 2005.

[32] A.R. Cockshott and Gary R. Sullivan. Improving the fermentation medium for echinocandin b production. part i: sequential statistical experimental design. *Process Biochemistry*, 36(7):647–660, 2001.

[33] Steven A. Cohen. *Amino Acid Analysis Using Precolumn Derivatization with 6-Aminoquinolyl-N-Hydroxysuccinimidyl Carbamate*, pages 39–47. Humana Press, Totowa, NJ, 2000.

[34] Paul Collier and Anke Hoeffler. Greed and grievance in civil war. *Oxford Economic Papers*, 56(4):563–595, 2004.

[35] The Business Research Company. *Monoclonal Antibodies (MAbS) Global Market Report 2021: COVID 19 Impact and Recovery to 2030*. Report Linker, 2021.

[36] Rafael S. Costa, Daniel Machado, Isabel Rocha, and Eugénio C. Ferreira. Hybrid dynamic modeling of escherichia coli central metabolic network combining michaelis–menten and approximate kinetic equations. *Biosystems*, 100(2):150 – 157, 2010.

[37] Paula Fernandes de Aguiar, B. Bourguignon, M.S. Khots, Desire Massart, and R. Phan-Than-Luu. D-optimal designs. *Chemometrics and intelligent laboratory systems*, 30(2):199–210, 1995.

[38] Sebastião Feyo de Azevedo, B. Dahm, and F.R. Oliveira. Hybrid modelling of biochemical processes: A comparison with the conventional approach. *Computers and Chemical Engineering*, 21:S751 – S756, 1997. Supplement to Computers and Chemical Engineering.

[39] Sijmen De Jong. Simpls: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*, 18(3):251–263, 1993.

[40] Jean-Sébastien Deschênes, Andre Desbiens, Michel Perrier, and Amine Kamen. Use of cell bleed in a high cell density perfusion culture and multivariable control of biomass and metabolite concentrations. *Asia-Pacific Journal of Chemical Engineering*, 1(1-2):82–91, 2006.

[41] Dimiter S. Dimitrov. Therapeutic proteins. In Vladimir Voynov and Justin A. Caravella, editors, *Therapeutic Proteins: Methods and Protocols, Methods in Molecular Biology, vol. 899*, chapter 1, pages 1–26. Humana Press, second edition, 2012.

[42] Zhimei Du, David Treiber, John D. McCarter, Dina Fomina-Yadlin, Ramsey A. Saleem, Rebecca E. McCoy, Yuling Zhang, Tharmala Tharmalingam, Matthew Leith, Brian D. Follstad, Brad Dell, Brent Grisim, Craig Zupke, Carole Heath, Arvia E. Morris, and Pranhitha Reddy. Use of a small molecule cell cycle inhibitor to control cell growth and improve specific productivity and product quality of recombinant proteins in cho cell cultures. *Biotechnology and bioengineering*, 112(1):141–155, 2015.

[43] Belmiro Duarte, Pedro M. Saraiva, and Constantinos Pantelides. Combined mechanistic and empirical modelling. *International Journal of Chemical Reactor Engineering - INT J CHEM REACT ENG*, 2, 2004.

[44] Dawn M. Ecker, Susan Dana Jones, and Howard L. Levine. The therapeutic monoclonal antibody market. *mAbs*, 7(1):9–14, 2015.

[45] Michael Elad. Optimized projections for compressed sensing. *IEEE Transactions on Signal Processing*, 55(12):5695–5702, 2007.

[46] Vivian Erklavec Zajec, Uroš Novak, Miha Kastelic, Boštjan Japelj, Ljerka Lah, Andrej Pohar, and Blaž Likozar. Dynamic multiscale metabolic network modeling of chinese hamster ovary cell metabolism integrating n-linked glycosylation in industrial biopharmaceutical manufacturing. *Biotechnology and Bioengineering*, 118(1):397–411, 2021.

[47] Osmán Fernández, Julio C. Dustet, and Ernesto Chico. Mathematical model for the application of metabolic flux analysis to cho cells producing recombinant human erythropoietin. *Biotecnología Aplicada*, 29:246 – 252, 2012.

[48] Kevin D. Foust and Brian K. Kaspar. NIH Public Access. 8(24):4017–4018, 2010.

[49] Sarah N. Galleguillos, David Ruckerbauer, Matthias P. Gerstl, Nicole Borth, Michael Hanscho, and Jürgen Zanghellini. What can mathematical modelling say about CHO metabolism and protein glycosylation? *Computational and Structural Biotechnology Journal*, 15:212–221, 2017.

[50] Federico Galvanin, Massimiliano Barolo, and Fabrizio Bezzo. A framework for model-based design of experiments in the presence of continuous measurement systems. *IFAC Proceedings Volumes*, 43(5):571 – 576, 2010. 9th IFAC Symposium on Dynamics and Control of Process Systems.

[51] Anshu Gambhir, Rashmi Korke, Jongchan Lee, Peng-Cheng Fu, Anna Europa, and Wei-Shou Hu. Analysis of cellular metabolism of hybridoma cells at distinct physiological states. *Journal of Bioscience and Bioengineering*, 95(4):317 – 327, 2003.

[52] Amin Ghafuri-Esfahani, Rahman Shokri, Athar Sharifi, Lida Shafiee, Roya Khosravi, Hooman Kaghazian, and Marouf Khalili. Optimization of parameters affecting on cho cell culture producing recombinant erythropoietin. *Preparative biochemistry & biotechnology*, 50(8):834–841, 2020.

[53] Atefeh Ghorbaniaghdam, Jingkui Chen, Olivier Henry, and Mario Jolicoeur. Analyzing clonal variation of monoclonal antibody-producing cho cell lines using an in silico metabolomic platform. *PLOS ONE*, 9(3):1–18, 2014.

[54] Debanjan Ghosh, Emma Hermonat, Prashant Mhaskar, Spencer Snowling, and Rajeev Goel. Hybrid modeling approach integrating first-principles models with subspace identification. *Industrial & Engineering Chemistry Research*, 58(30):13533–13543, 2019.

[55] Erwin P. Gianchandani, Arvind K. Chavali, and Jason A. Papin. The application of flux balance analysis in systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(3):372–382, 2010.

[56] Robert N. Goldberg. Standards in biothermodynamics. *Perspectives in Science*, 1(1-6):7–14, 2014.

[57] Oliver Götz, Kerstin Liehr-Gobbers, and Manfred Krafft. *Evaluation of Structural Equation Models Using the Partial Least Squares (PLS) Approach*, pages 691–711. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[58] Ali S. Hadi and Robert F. Ling. Some cautionary notes on the use of principal components regression. *The American Statistician*, 52(1):15–19, 1998.

[59] Erika Hagrot, Hildur Æsa Oddsdóttir, Joan Gonzalez Hosta, Elling W. Jacobsen, and Véronique Chotteau. Poly-pathway model, a novel approach to simulate multiple metabolic states by reaction network-based model – Application to amino acid depletion in CHO cell culture. *Journal of Biotechnology*, 265(July):127, 2017.

[60] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[61] B. Heinritz. Biothermodynamic studies for optimization of cell-mass production. *Thermochimica Acta*, (172):241–250, 1990.

[62] Christopher S. Henry, Linda J. Broadbelt, and Vassily Hatzimanikatis. Thermodynamics-based metabolic flux analysis. *Biophysical Journal*, 92(5):1792–1805, 2007.

[63] Christopher S. Henry, Matthew D. Jankowski, Linda J. Broadbelt, and Vassily Hatzimanikatis. Genome-scale thermodynamic analysis of Escherichia coli metabolism. *Biophysical Journal*, 90(4):1453–1461, 2006.

[64] Hille, Rubin. *Run-to-Run Optimization of Biochemical Batch Processes in the Presence of Model-Plant Mismatch*. PhD thesis, University of Waterloo, 2018.

[65] Tomoharu Hogiri, Hiroshi Tamashima, Akitoshi Nishizawa, and Masahiro Okamoto. Optimization of a pH-shift control strategy for producing monoclonal antibodies in Chinese hamster ovary cell cultures using a pH-dependent dynamic model. *Journal of Bioscience and Bioengineering*, 125(2):245–250, 2018.

[66] Jong Kwang Hong, Shilpa Nargund, Meiyappan Lakshmanan, Sarantos Kyriakopoulos, Do Yun Kim, Kok Siong Ang, Dawn Leong, Yuansheng Yang, and Dong-Yup Lee. Comparative phenotypic analysis of cho clones and culture media for lactate shift. *Journal of biotechnology*, 283:97–104, 2018.

[67] Yao-Ming Huang, WeiWei Hu, Eddie Rustandi, Kevin Chang, Helena Yusuf-Makagiansar, and Thomas Ryll. Maximizing productivity of cho cell-based fed-batch culture using chemically defined media conditions and typical manufacturing equipment. *Biotechnology Progress*, 26(5):1400–1410, 2010.

[68] Zhuangrong Huang, Jianlin Xu, Andrew Yongky, Caitlin S Morris, Ashli L. Polanco, Michael Reily, Michael C. Borys, Zheng Jian Li, and Seongkyu Yoon. Cho cell productivity improvement by genome-scale modeling and pathway analysis: Application to feed supplements. *Biochemical Engineering Journal*, 160:107638, 2020.

[69] Qianru Jiang, Sheng Li, Huang Bai, Rodrigo C. de Lamare, and Xiongxiong He. Gradient-based algorithm for designing sensing matrix considering real mutual coherence for compressed sensing systems. *IET Signal Processing*, 11(4):356–363, 2017.

[70] Ying Jing, Michael Borys, Samiksha Nayak, Susan Egan, Yueming Qian, Shih Hsie Pan, and Zheng Jian Li. Identification of cell culture conditions to control protein aggregation of IgG fusion proteins expressed in Chinese hamster ovary cells. *Process Biochemistry*, 47(1):69–75, 2012.

[71] Yoonjung Ju, Kwang-Hee Son, Chunzhi Jin, Byung Soon Hwang, Dong-Jin Park, and Chang-Jin Kim. Statistical optimization of culture medium for improved production of antimicrobial compound by streptomyces rimosus ag-p1441. *Food science and biotechnology*, 27(2):581–590, 2018.

[72] Minoru Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947–1951, 2019.

[73] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe, and Mao Tanabe. Kegg: integrating viruses and cellular organisms. *Nucleic acids research*, 49(D1):D545–D551, 2021.

[74] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[75] Miha Kastelic, Drejc Kopač, Uroš Novak, and Blaž Likozar. Dynamic metabolic network modeling of mammalian chinese hamster ovary (cho) cell cultures with continuous phase kinetics transitions. *Biochemical Engineering Journal*, 142:124–134, 2019.

[76] William Kelly, Sorelle Veigne, Xianhua Li, Shyam Sundar Subramanian, Zuyi Huang, and Eugene Schaefer. Optimizing performance of semi-continuous cell culture in an ambr15™ microbioreactor using dynamic flux balance modeling. *Biotechnology Progress*, 34(2):420–431, 2018.

[77] Jee Yon Kim, Yeon Gu Kim, and Gyun Min Lee. CHO cells in biotechnology for production of recombinant proteins: Current state and further potential. *Applied Microbiology and Biotechnology*, 93(3):917–930, 2012.

[78] Jungyeon Kim and Kyoung Heon Kim. Effects of minimal media vs. complex media on the metabolite profiles of Escherichia coli and Saccharomyces cerevisiae. *Process Biochemistry*, 57(December 2016):64–71, 2017.

[79] Shohei Kishishita, Satoshi Katayama, Kunihiko Kodaira, Yoshinori Takagi, Hiroki Matsuda, Hiroshi Okamoto, Shinya Takuma, Chikashi Hirashima, and Hideki Aoyagi. Optimization of chemically defined feed media for monoclonal antibody production in Chinese hamster ovary cells. *Journal of Bioscience and Bioengineering*, 120(1):78–84, 2015.

[80] Clemens Kreutz and Jens Timmer. Systems biology: experimental design. *The FEBS Journal*, 276(4):923–942, 2009.

[81] Chih-Chung Kuo, Austin W.T. Chiang, Isaac Shamie, Mojtaba Samoudi, Jahir M. Gutierrez, and Nathan E. Lewis. The emerging role of systems biology for engineering protein production in cho cells. *Current Opinion in Biotechnology*, 51:64 – 69, 2018. Systems Biology Nanobiotechnology.

[82] Minjung Kyung, Jeff Gill, Malay Ghosh, George Casella, et al. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

[83] Tingfeng Lai, Yuansheng Yang, and Say Kong Ng. Advances in mammalian cell line development technologies for recombinant protein production. *Pharmaceuticals*, 6(5):579–603, 2013.

[84] Marie Eve Lalonde and Yves Durocher. Therapeutic glycoprotein production in mammalian cells. *Journal of Biotechnology*, 251(December 2016):128–140, 2017.

[85] Hae Woo Lee, Andrew Christie, Jason A. Starkey, Erik K. Read, and Seongkyu Yoon. Intracellular metabolic flux analysis of cho cells supplemented with wheat hydrolysates for improved mab production and cell-growth. *Journal of Chemical Technology & Biotechnology*, 90(2):291–302, 2015.

[86] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: Principal component analysis. *Nature Methods*, 17(7):641–642, 2017.

[87] Feng Li, Natarajan Vijayasankaran, Amy Shen, Robert Kiss, and Ashraf Amanullah. Cell culture processes for monoclonal antibody production. *mAbs*, 2(5):466–479, 2010.

[88] Yujie Li, Gaorong Li, and Tiejun Tong. Sequential profile lasso for ultra-high-dimensional partially linear models. *Statistical Theory and Related Fields*, 1(2):234–245, 2017.

[89] Juliane Liepe, Sarah Filippi, Michał Komorowski, and Michael PH Stumpf. Maximizing the information content of experiments in systems biology. *PLoS computational biology*, 9(1):e1002888, 2013.

[90] Bo Liu, Maureen Spearman, John Doering, Erica Lattová, Hélène Perreault, and Michael Butler. The availability of glucose to CHO cells affects the intracellular lipid-linked oligosaccharide distribution, site occupancy and the N-glycosylation profile of a monoclonal antibody. *Journal of Biotechnology*, 170(1):17–27, 2014.

[91] Hanzhong Liu, Xin Xu, and Jingyi Jessica Li. A bootstrap lasso+ partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models. *arXiv preprint arXiv:1706.02150*, 2017.

[92] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[93] Martin F. Luna and Ernesto C. Martínez. Optimal design of dynamic experiments in the development of cybernetic models for bioreactors. *Chemical Engineering Research and Design*, 136:334 – 346, 2018.

[94] Shan Luo and Zehua Chen. Sequential Lasso for feature selection with ultra-high dimensional feature space. *arXiv e-prints*, page arXiv:1107.2734, 2011.

[95] Shan Luo and Zehua Chen. Sequential lasso cum ebic for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*, 109(507):1229–1240, 2014.

[96] Timo R. Maarleveld, Ruchir A. Khandelwal, Brett G. Olivier, Bas Teusink, and Frank J. Bruggeman. Basic concepts and principles of stoichiometric modeling of metabolic networks. *Biotechnology Journal*, 8(9):997–1008, 2013.

[97] Radhakrishnan Mahadevan, Jeremy S. Edwards, and Francis J. Doyle. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophysical journal*, 83(3):1331 – 13340, 2002.

[98] Radhakrishnan Mahadevan and Christopher H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5(4):264 – 276, 2003.

[99] Saikat Maitra and Jun Yan. Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Applying Multivariate Statistical Models*, 79:79–90, 2008.

[100] MathWorks. *Statistics and Machine Learning Toolbox™ User's Guide MATLAB*, volume R2019b. The MathWorks, Inc., 2019.

[101] MATLAB. *Statistics and Machine Learning Toolbox User's Guide (R2021a)*. The MathWorks Inc., Natick, Massachusetts, 2021.

[102] Michael L. Mavrovouniotis. Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnology and Bioengineering*, 36(10):1070–1082, 1990.

[103] Michael L. Mavrovouniotis. Estimation of standard Gibbs energy changes of biotransformations. *J Biol Chem*, 266(22):14440–14445, 1991.

[104] Trudy McKee and James R. McKee. *Biochemistry: The Molecular Basis of Life*. Oxford University Press, Oxford, New York, 2009.

[105] Mukesh Meshram, Saeideh Naderi, Brendan McConkey, Brian Ingalls, Jeno Scharer, and Hector Budman. Modeling the coupled extracellular and intracellular environments in mammalian cell culture. *Metabolic Engineering*, 19:57–68, 2013.

[106] Douglas C. Montgomery. *Design and Analysis of Experiments (8th Edition)*. John Wiley & Sons, 2013.

[107] Caitlin Morris, Ashli Polanco, Andrew Yongky, Jianlin Xu, Zhuangrong Huang, Jia Zhao, Kevin S McFarland, Seoyoung Park, Bethanne Warrack, Michael Reily, et al. Bigdata analytics identifies metabolic inhibitors and promoters for productivity improvement and optimization of monoclonal antibody (mab) production process. *Bioresources and Bioprocessing*, 7:1–13, 2020.

[108] Anwesh Reddy Gottu Mukkula and Radoslav Paulen. Model-based optimal experiment design for nonlinear parameter estimation using exact confidence regions**this research was funded by the european commission under grant agreement number 291458 (erc advanced investigator grant mobocon). *IFAC-PapersOnLine*, 50(1):13760 – 13765, 2017. 20th IFAC World Congress.

[109] Isabelle Nadeau, Danielle Jacob, Michel Perrier, and Amine Kamen. 293sf metabolic flux analysis during cell growth and infection with an adenoviral vector. *Biotechnology Progress*, 16(5):872–884, 2000.

[110] Saeideh Naderi, Mukesh Meshram, Catherine Wei, Brendan Mcconkey, Brian Ingalls, Hector Budman, and Jeno Scharer. Development of a mathematical model

for evaluating the dynamics of normal and apoptotic Chinese hamster ovary cells. *Biotechnology Progress*, 27(5):1197–1205, 2011.

[111] Flavia Neddermeyer, Volker Marhold, Christoph Menzel, Dominik Krämer, and Rudibert King. Modelling the production of soluble hydrogenase in ralstonia eutropha by on-line optimal experimental design**this work was supported by the dfg in the framework of the cluster of excellence unicat. *IFAC-PapersOnLine*, 49(7):627 – 632, 2016. 11th IFAC Symposium on Dynamics and Control of Process SystemsIncluding Biosystems DYCOPS-CAB 2016.

[112] Ali Nikdel and Hector Budman. Identification of active constraints in dynamic flux balance analysis. *Biotechnology Progress*, 33(1):26–36, 2017.

[113] Ali Nikdel, Richard D. Braatz, and Hector M. Budman. A systematic approach for finding the objective function and active constraints for dynamic flux balance analysis. *Bioprocess and Biosystems Engineering*, 41:641–655, 2018.

[114] Nikdel, Ali. *Systematic Approaches to Identification of Dynamic Flux Balance Models*. PhD thesis, University of Waterloo, 2018.

[115] NIST/SEMATECH. *NIST/SEMATECH e-Handbook of Statistical Methods*. 2012.

[116] Hongxing Niu, Zakaria Amribt, Patrick Fickers, Wensong Tan, and Philippe Bogaerts. Metabolic pathway analysis and reduction for mammalian cell cultures—towards macroscopic modeling. *Chemical Engineering Science*, 102:461 – 473, 2013.

[117] Ryan P. Nolan and Kyongbum Lee. Dynamic model of cho cell metabolism. *Metabolic engineering*, 13(1):108–124, 2011.

[118] Conor M. O'Brien, Qi Zhang, Prodromos Daoutidis, and Wei-Shou Hu. A hybrid mechanistic-empirical model for in silico mammalian cell bioprocess simulation. *Metabolic Engineering*, 66:31–40, 2021.

[119] Xiao Pan, Ciska Dalm, René H. Wijffels, and Dirk E. Martens. Metabolic characterization of a cho cell size increase phase in fed-batch cultures. *Applied Microbiology and Biotechnology*, 101(22):8101–8113, 2017.

[120] Muthukumarasamy Parthasarathy and Joel Gnanadoss. Medium formulation and its optimization to enhance protease production by streptomyces sp. isolated from mangroves. *Biosciences Biotechnology Research Asia*, 15(3):719–728, 2018.

[121] Satheesh K. Perepu and Arun K. Tangirala. Reconstruction of missing data using compressed sensing techniques with adaptive dictionary. *Journal of Process Control*, 47:175–190, 2016.

[122] Elena Pesce and Eva Riccomagno. Large datasets, bias and model oriented optimal design of experiments. *arXiv preprint arXiv:1811.12682*, 2018.

[123] Jinshu Qiu, Pik Kay Chan, and Pavel V. Bondarenko. Monitoring utilizations of amino acids and vitamins in culture media and chinese hamster ovary cells by liquid chromatography tandem mass spectrometry. *Journal of Pharmaceutical and Biomedical Analysis*, 117:163 – 172, 2016.

[124] Aravindan Rajendran and Viruthagiri Thangavelu. Sequential optimization of culture medium composition for extracellular lipase production by bacillus sphaericus using statistical methods. *Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental & Clean Technology*, 82(5):460–470, 2007.

[125] Vinayagam Ramesh and Vytla Ramachandra Murty. Sequential statistical optimization of media components for the production of glucoamylase by thermophilic fungus humicola grisea mtcc 352. *Enzyme research*, 2014, 2014.

[126] Sathya N. Ravi, Vamsi K. Ithapu, Sterling C. Johnson, and Vikas Singh. Experimental design on a budget for sparse linear models and applications. *JMLR Workshop Conf Proc*, 48:583 – 592, 2016. JMLR workshop and conference proceedings.

[127] Janice M. Reichert. Antibodies to watch in 2016. *mAbs*, 8(2):197–204, 2016.

[128] Janice M. Reichert. Antibodies to watch in 2017. *mAbs*, 9(2):167–181, 2017.

[129] Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016.

[130] David Reinhart, Lukas Damjanovic, Christian Kaisermayer, and Renate Kunert. Benchmarking of commercially available CHO cell culture media for antibody production. *Applied Microbiology and Biotechnology*, 99(11):4645–4657, 2015.

[131] Živa Rejc, Lidija Magdevska, Tilen Tršelič, Timotej Osolin, Rok Vodopivec, Jakob Mraz, Eva Pavliha, Nikolaj Zimic, Tanja Cvitanović, Damjana Rozman, Miha Moškon, and Miha Mraz. Computational modelling of genome-scale metabolic networks and its application to CHO cell cultures. *Computers in Biology and Medicine*, 88:150–160, 2017.

[132] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303, 2008.

[133] Frank V. Ritacco, Yongqi Wu, and Anurag Khetan. Cell culture media for recombinant protein expression in chinese hamster ovary (cho) cells: History, key components, and optimization strategies. *Biotechnology progress*, 34(6):1407–1426, 2018.

[134] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In Craig Saunders, Marko Grobelnik, Steve Gunn, and John Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection*, pages 34–51, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[135] Yolande Rouiller, Arnaud Périlleux, Natacha Collet, Martin Jordan, Matthieu Stettler, and Hervé Broly. A high-throughput media design approach for high performance mammalian fed-batch cultures. *mAbs*, 5(3):501–511, 2013.

[136] Brian Russell, Guillermo Miro-Quesada, Qu Limin, and Sanjeev Ahuja. Characterizing the preparation of a concentrated nutrient feed solution for a large-scale cell culture process. *Biochemical Engineering Journal*, 134:120–128, 2018.

[137] Maliheh Safavi, Mahroo Seyed Jafari Olia, Mohammad Haji Abolhasani, Mohsen Amini, and Mehran Kianirad. Optimization of the culture medium and characterization of antioxidant compounds of a marine isolated microalga as a promising source in aquaculture feed. *Biocatalysis and Agricultural Biotechnology*, page 102098, 2021.

[138] Andrew Salazar, Michael Keusgen, and Jörg von Hagen. Amino acids in the cultivation of mammalian cells. *Amino acids*, 48(5):1161–1171, 2016.

[139] Taha Salim, Gaurav Chauhan, Neil Templeton, and Wai Lam Ling. Using mvda with stoichiometric balances to optimize amino acid concentrations in chemically defined cho cell culture medium for improved culture performance. *Authorea Preprints*, 2021.

[140] Brandy Sargent. The impact of trace elements on cell culture media and upstream processing. `https://cellculturedish.com/impact-trace-elements-cell-culture-media-and-upstream-processing/`, Aug 2018. Accessed: 2020-09-30.

[141] Joanne M. Savinell and Bernhard O. Palsson. Network analysis of intermediary metabolism using linear optimization. i. development of mathematical formalism. *Journal of Theoretical Biology*, 154(4):421 – 454, 1992.

[142] Mark Schmidt. Least squares optimization with l1-norm regularization. *CS542B Project Report*, 504:195–221, 2005.

[143] Suresh Selvarasu, Ying Swan Ho, William P.K. Chong, Niki S.C. Wong, Faraaz N.K. Yusufi, Yih Yean Lee, Miranda G.S. Yap, and Dong Yup Lee. Combined in silico modeling and metabolomics analysis to characterize fed-batch CHO cell culture. *Biotechnology and Bioengineering*, 109(6):1415–1429, 2012.

[144] Kristen A. Severson, Jeremy G. VanAntwerp, Venkatesh Natarajan, Chris Antoniou, Jörg Thömmes, and Richard D. Braatz. A systematic approach to process data analytics in pharmaceutical manufacturing: The data analytics triangle and its application to the manufacturing of a monoclonal antibody. In *Multivariate Analysis in the Pharmaceutical Industry*, pages 295–312. Elsevier, 2018.

[145] Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

[146] Madeline A. Shea, John J. Correia, and Michael D. Brenowitz. Introduction: Twenty five years of the Gibbs Conference on Biothermodynamics. *Biophysical Chemistry*, 159(1):1–5, 2011.

[147] Coral Fung Shek, Pavlos Kotidis, and Michael Betenbaugh. Mechanistic and data-driven modeling of protein glycosylation. *Current Opinion in Chemical Engineering*, 32:100690, 2021.

[148] Frits Byron Soepyan, Christine M. Anderson-Cook, Joshua C. Morgan, Charles H. Tong, Debangsu Bhattacharyya, Benjamin P. Omell, Michael S. Matuszewski, K. Sham Bhat, Miguel A. Zamarripa, John C. Eslick, Joel D. Kress, James R. Gattiker, Christopher S. Russell, Brenda Ng, Jeremy C. Ou, and David C. Miller. Sequential design of experiments to maximize learning from carbon capture pilot plant testing. In Mario R. Eden, Marianthi G. Ierapetritou, and Gavin P. Towler, editors, *13th International Symposium on Process Systems Engineering (PSE 2018)*, volume 44 of *Computer Aided Chemical Engineering*, pages 283 – 288. Elsevier, 2018.

[149] Ya-ting Sun, Liang Zhao, Zhaoyang Ye, Li Fan, Xu ping Liu, and Wen Song Tan. Development of a fed-batch cultivation for antibody-producing cells based on combined feeding strategy of glucose and galactose. *Biochemical Engineering Journal*, 81:126–135, 2013.

[150] Ana P. Teixeira, Carlos Alves, Paula M. Alves, Manuel J.T. Carrondo, and Rui Oliveira. Hybrid elementary flux analysis/nonparametric modeling: application for bioprocess control. *BMC Bioinformatics*, 8(1):30, 2007.

[151] Dries Telen, Filip Logist, Eva Van Derlinden, Ignace Tack, and Jan Van Impe. Optimal experiment design for dynamic bioprocesses: A multi-objective approach. *Chemical Engineering Science*, 78:82 – 97, 2012.

[152] Dries Telen, Filip Logist, Rien Quirynen, Boris Houska, Moritz Diehl, and Jan Van Impe. Optimal experiment design for nonlinear dynamic (bio)chemical systems using sequential semidefinite programming. *AIChE Journal*, 60(5):1728–1739, 2014.

[153] Neil Templeton, Kevin D. Smith, Allison G. McAtee-Pereira, Haimanti Dorai, Michael J. Betenbaugh, Steven E. Lang, and Jamey D. Young. Application of 13c flux analysis to identify high-productivity cho metabolic phenotypes. *Metabolic Engineering*, 43:218 – 225, 2017. Engineering approaches to study cancer metabolism.

[154] Neil Templeton and Jamey D. Young. Biochemical and metabolic engineering approaches to enhance production of therapeutic proteins in animal cell cultures. *Biochemical Engineering Journal*, 136:40 – 50, 2018.

[155] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[156] Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[157] Randall D. Tobias. An introduction to partial least squares regression. In *Proceedings of the twentieth annual SAS users group international conference*, volume 20. Citeseer, 1995.

[158] Yoshihiro Toya, Nobuaki Kono, Kazuharu Arakawa, and Masaru Tomita. Metabolic Flux Analysis and Visualization. *Journal of Proteome Research*, 10(8):3313–3323, 2011.

[159] Apostolos Tsopanoglou and Ioscani Jiménez del Val. Moving towards an era of hybrid modelling: advantages and challenges of coupling mechanistic and data-driven models for upstream pharmaceutical bioprocesses. *Current Opinion in Chemical Engineering*, 32:100691, 2021.

[160] Pornkamol Unrean and Friedrich Srienc. Metabolic networks evolve towards states of maximum entropy production. *Metabolic Engineering*, 13(6):666 – 673, 2011.

[161] Jan van der Valk, Daniel Brunner, Karen De Smet, Asa Fex Svenningsen, Paul Honegger, Lisbeth E. Knudsen, Toni Lindl, Jens Noraberg, Anna Price, Maria Laura Scarino, and Gerhard Gstraunthaler. Optimization of chemically defined cell culture media – replacing fetal bovine serum in mammalian in vitro methods. *Toxicology in Vitro*, 24(4):1053 – 1063, 2010.

[162] Amit Varma and Bernhard O. Palsson. Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/technology*, 12(10):994–998, 1994.

[163] Urs Von Stockar. Biothermodynamics of live cells: A tool for biotechnology and biochemical engineering. *Journal of Non-Equilibrium Thermodynamics*, 35(4):415–475, 2010.

[164] Urs Von Stockar. The hole of thermodynamics in biochemical engineering. *Journal of Non-Equilibrium Thermodynamics*, (38):225–240, 2013.

[165] Thomas W. Eyster, Sameer Talwar, Janice Fernandez, Shelby Foster, James Hayes, Randal Allen, Scot Reidinger, Boyong Wan, Xiaodan Ji, Juan Aon, et al. Tuning monoclonal antibody galactosylation using raman spectroscopy-controlled lactic acid feeding. *Biotechnology Progress*, 37(1):e3085, 2021.

[166] Patrik Waldmann, Gábor Mészáros, Birgit Gredler, Christian Fuerst, and Johann Sölkner. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics*, 4:270, 2013.

[167] John M. Walker. *Therapeutic proteins: Methods and Protocols*, volume 531. 2012.

[168] HaiYing Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405, 2018.

[169] Zhibing Weng, Jian Jin, ChunHua Shao, and Huazhong Li. Reduction of charge variants by cho cell culture process optimization. *Cytotechnology*, 72(2):259–269, 2020.

[170] Dirk Weuster-Botz. Experimental design for fermentation media development: statistical design or global random search? *Journal of bioscience and bioengineering*, 90(5):473–483, 2000.

[171] Zizhuo Xing, Brian Kenty, Inna Koyrakh, Michael Borys, Shih-Hsie Pan, and Zheng Jian Li. Optimizing amino acid composition of cho cell culture media for a fusion protein production. *Process Biochemistry*, 46(7):1423 – 1429, 2011.

[172] Jianping Xu, Yiming Pi, and Zongjie Cao. Optimized projection matrix for compressive sensing. *EURASIP Journal on Advances in Signal Processing*, 2010(1):560349, 2010.

[173] Ningning Xu, Chao Ma, Jianfa Ou, Wanqi Wendy Sun, Lufang Zhou, Hui Hu, and Xiaoguang Margaret Liu. Comparative proteomic analysis of three Chinese hamster ovary (CHO) host cells. *Biochemical Engineering Journal*, 124:122–129, 2017.

[174] Zhengbing Yan, Chih-Chiun Chiu, Weiwei Dong, and Yuan Yao. A lasso-based batch process modeling and end-product quality prediction method. *IFAC Proceedings Volumes*, 47(3):6704–6709, 2014.

[175] Tatsuma Yao and Yuta Asayama. Animal-cell culture media: history, characteristics, and current issues. *Reproductive medicine and biology*, 16(2):99–117, 2017.

[176] Piotr Zakrzewski, Marnix H. Medema, Albert Gevorgyan, Andrzej M. Kierzek, Rainer Breitling, and Eriko Takano. Multimeteval: Comparative and multi-objective analysis of genome-scale metabolic models. *PLOS ONE*, 7(12):1–9, 2012.

[177] Dénes Zalai, Krisztina Koczka, László Párta, Patrick Wechselberger, Tobias Klein, and Christoph Herwig. Combining mechanistic and data-driven approaches to gain process knowledge on the control of the metabolic shift to lactate uptake in a fed-batch cho process. *Biotechnology progress*, 31(6):1657–1668, 2015.

[178] Francisca Zamorano, Alain Vande Wouwer, and Georges Bastin. A detailed metabolic flux analysis of an underdetermined network of cho cells. *Journal of Biotechnology*, 150(4):497 – 508, 2010.

[179] Jinyou Zhang. *Mammalian Cell Culture for Biopharmaceutical Production*, pages 157 – 178. 2010.

[180] Yun Zhang, Yang Zhang, Jie Gao, Qiuxuan Shen, Zhihui Bai, Xuliang Zhuang, and Guoqiang Zhuang. Optimization of the medium for the growth of nitrobacter winogradskyi by statistical method. *Letters in applied microbiology*, 67(3):306–313, 2018.

[181] Quan Zhou, Shiji Song, Gao Huang, and Cheng Wu. Efficient lasso training from a geometrical perspective. *Neurocomputing*, 168:234–239, 2015.

[182] Hui Zou. The adaptive lasso ad its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

APPENDICES

Appendix A

Metabolic Network of CHO Cells

Metabolic network according to Nikdel [114]. The symbol v represents each metabolic flux. Biomass flux (v_{37}) coefficients were adapted according to the cell line used in this work.

v_1: Glc → G6P
v_2: G6P → 2·3phosphaglycerate
v_3: 3phosphoglycerate → Pyr
v_4: Pyr → Lac
v_5: Pyr → AcCoA + CO2
v_6: AcCoA + Oxal → Cit
v_7: Cit → KG + CO2
v_8: KG → SucCoA + CO2
v_9: SucCoA → Suc
v_{10}: Suc → Mal
v_{11}: Mal → Oxal
v_{12}: Mal → Pyr + CO2
v_{13}: Thr → Gly + AcCoA
v_{14}: Trp → Ala + NH4 +2·AcCoA
v_{15}: Lys → NH4 + KG
v_{16}: Ile → Glu + AcCoA + SucCoA
v_{17}: Leu → 2·AcCoA + 2·CO2
v_{18}: Tyr → Mal + Oxal + CO2
v_{19}: Ser + Met → Cys + NH4
v_{20}: Val → SucCoA + KG
v_{21}: Glu + Oxal → Asp + KG
v_{22}: Glu → KG + NH4

v_{23}: Glu + Pyr → Ala + KG
v_{24}: Cys → Pyr
v_{25}: Ser → NH4 + Pyr
v_{26}: Gly → NH4 + CO2
v_{27}: Ser + Thr → SucCoA
v_{28}: Glu + 3phosphoglycerate → Ser + KG
v_{29}: Ser → Gly
v_{30}: Phe → Tyr
v_{31}: Asn → Asp + NH4
v_{32}: Gln → Glu + NH4
v_{33}: Arg → Glu
v_{34}: Glu → Pro
v_{35}: His → Glu + NH4
v_{36}: Gln → Asp → Glu + Asn
v_{37}: 0.0208·Glc + 0.0377·Gln + 0.0006·Glu + 0.007·Arg + 0.003·Hist + 0.0084·Ile + 0.0133·Leu + 0.0101·Lys + 0.005·Met + 0.0055·Phe + 0.008·Thr + 0.004·Trp + 0.0096·Val + 0.01·Ala + 0.026·Asp + 0.0004·Cys + 0.0165·Gly + 0.02·Pro + 0.05·Ser + 0.0077·Tyr → Bio
v_{38}: 0.0104·Gln + 0.0107·Glu + 0.0050·Arg + 0.0035·Hist + 0.0050·Ile + 0.0142·Leu + 0.0145·Lys + 0.0028·Met + 0.0072·Phe + 0.0160·Thr + 0.0189·Val + 0.0110·Ala + 0.0082·Asp + 0.0050·Cys + 0.0145·Gly + 0.0148·Pro + 0.0267·Ser + 0.0085·Tyr + 0.0072·Asn → IgG
v_{39}, v_{40}, v_{41}, v_{42}, v_{43}, v_{44}, v_{45}, v_{46} and v_{47} are, respectively, the reversible reactions v_4, v_{21}, v_{22}, v_{23}, v_{29}, v_{30}, v_{31}, v_{32} and v_{36}.

Appendix B

Soft constraint in the DMFM model batch system

Table B.1: Soft constraint values used in the DMFM model for the batch system. Soft constraints bound the metabolite consumption/production rate by a constant value b^L in the lower bound (LB) or b^U in the upper bound (UB).

Metabolite	Bound	Soft constraint value b^L or b^U
Alanine	UB	0.01
Arginine	UB	-0.015
Asparagine	UB	-0.04
Aspartate	LB & UB	-0.04 & -0.02
Glucose	UB	-0.2
Glutamine	UB	0.02
Glutamate	LB	-0.03
Glycine	LB	0.01
Histidine	UB	-0.008
Isoleucine	UB	-0.02
Leucine	UB	-0.04
Lactate	UB	1.00
Lysine	UB	-0.02
Ammonia	UB	0.15
Phenylalanine	UB	-0.01
Serine	LB	-0.25
Threonine	LB	-0.03
Tryptophan	UB	-0.005
Tyrosine	UB	-0.015
Valine	UB	-0.015
Methionine	UB	-0.006
Cysteine	UB	-0.005
Proline	UB	-0.01
Biomass	UB	0.165

Table B.2: Soft constraint values used in the DMFM model for the perfusion system. Soft constraints bound the metabolite consumption/production rate by a constant value b^L in the lower bound (LB) or b^U in the upper bound (UB).

Metabolite	Bound	Soft constraint value b^L or b^U
Alanine	LB & UB	0.013 & 0.02
Arginine	UB	-0.015
Asparagine	UB	-0.044
Aspartate	UB	-0.015
Glucose	UB	-0.26
Glutamine	LB	-0.00062
Glutamate	LB	-0.02
Glycine	LB	0.013
Histidine	UB	-0.0044
Isoleucine	UB	-0.013
Leucine	UB	-0.023
Lactate	UB	1.00
Lysine	UB	-0.014
Ammonia	UB	0.05
Phenylalanine	UB	-0.007
Serine	LB	-0.04
Threonine	LB	-0.0125
Tryptophan	UB	-0.0038
Tyrosine	UB	-0.0085
Valine	UB	-0.015
Methionine	UB	-0.0044
Cysteine	UB	-0.0075
Proline	UB	-0.012
Biomass	UB	0.165

Appendix C

Dynamic Kinetic Metabolic Model adapted from Hille (2018)

$$\frac{dX_v}{dt} = \mu X_v \left(\frac{[Glc]}{(K_{21} + [Glc])} \frac{1}{\left(1 + \frac{[Amm]}{K_{23}}\right)} \frac{1}{\left(1 + \frac{[Lac]}{K_{25}}\right)} \right) ... \\ - k_d X_v^2 \left(\frac{1}{1 + \left(\frac{K_{24}}{[Amm]}\right)^n} + \frac{K_{26}}{[Glc] + K_{22}} \right) \tag{C.1}$$

$$\frac{dX_d}{dt} = k_d X_v^2 \left(\frac{1}{1 + \left(\frac{K_{24}}{[Amm]}\right)^n} + \frac{K_{26}}{[Glc] + K_{22}} \right) - K_{lys} X_d \tag{C.2}$$

$$\frac{d[Glc]}{dt} = -K_{41} X_v - X_v \left(\frac{K_{42}[Glc]}{K_{43} + [Glc]} \right) \tag{C.3}$$

$$\frac{d[Lac]}{dt} = K_{51} X_v + X_v \left(\frac{K_{52}[Glc]}{K_{43} + [Glc]} - \frac{K_{53}[Lac]}{K_{54} + [Lac]} \right) \tag{C.4}$$

$$\frac{d[Amm]}{dt} = X_v \left(\frac{K_{64}[Glu]}{K_{1003} + [Glu]} + \frac{K_{601}[Gln]}{K_{602} + [Gln]} \right) ... \\ - X_v \left(\frac{K_{62}[Amm][Glu]}{(K_{61} + [Amm])(K_{63} + [Glu])} + \frac{K_{66}[Amm][Asp]}{(K_{67} + [Asp])(K_{68} + [Amm])} \right) \tag{C.5}$$

$$\frac{d[Asp]}{dt} = X_v \left(K_{81} Asp - \frac{K_{82}[Asp]}{K_{83} + [Asp]} \right) \tag{C.6}$$

$$\frac{d[Ala]}{dt} = X_v \left(\frac{K_{91}[Glu]}{K_{1003} + [Glu]} - \frac{K_{92}[Ala]}{K_{93} + [Ala]} \right) \tag{C.7}$$

$$\frac{d[Glu]}{dt} = X_v \left(\frac{K_{106}[Gln]}{K_{144} + [Gln]} + \frac{K_{104}[Ala]}{K_{93} + [Ala]} + \frac{K_{102}[Asp]}{K_{83} + [Asp]} \right) ... \\ + X_v \left(\frac{K_{103}[Amm]}{K_{101} + [Amm]} + \frac{K_{1001}[Glc]}{K_{1002} + [Glc]} \right) ... \\ - X_v \left(\frac{K_{105}[Glu]}{K_{1003} + [Glu]} + \frac{K_{1007}[Amm][Glu]}{(K_{61} + [Amm])(K_{63} + [Glu])} \right) \tag{C.8}$$

$$\frac{d[Pro]}{dt} = -X_v \left(\frac{K_{110}[Pro]}{K_{111} + [Pro]} \right) \tag{C.9}$$

$$\frac{d[Ser]}{dt} = -X_v \left(\frac{K_{120}[Ser]}{K_{121} + [Ser]} + \frac{K_{122}[Amm][Ser]}{(K_{133} + [Amm])(K_{134} + [Ser])} \right) \tag{C.10}$$

$$\frac{d[Gly]}{dt} = X_v \left(\frac{K_{130}[Ser]}{(K_{133} + [Ser])} - \frac{K_{131}[Gly]}{K_{132} + [Gly]} \right) \tag{C.11}$$

$$\frac{d[Gln]}{dt} = -X_v \left(\frac{K_{140}[Amm][Glu]}{(K_{61} + [Amm])(K_{63} + [Glu])} - \frac{K_{143}[Gln]}{K_{144} + [Gln]} \right) \tag{C.12}$$

$$\frac{d[Thr]}{dt} = -X_v \left(\frac{K_{150}[Thr]}{K_{151} + [Thr]} \right) \tag{C.13}$$

$$\frac{d[His]}{dt} = -X_v \left(\frac{K_{160}[His]}{K_{161} + [His]} \right) \tag{C.14}$$

$$\frac{d[Arg]}{dt} = -X_v \left(\frac{K_{170}[Arg]}{K_{171} + [Arg]} \right) \tag{C.15}$$

$$\frac{d[Cys]}{dt} = -X_v \left(\frac{K_{180}[Cys]}{K_{181} + [Cys]} \right) \tag{C.16}$$

$$\frac{d[Lys]}{dt} = -X_v \left(\frac{K_{190}[Lys]}{K_{191} + [Lys]} \right) \tag{C.17}$$

$$\frac{d[Tyr]}{dt} = -X_v \left(\frac{K_{200}[Tyr]}{K_{201} + [Tyr]} \right) \tag{C.18}$$

$$\frac{d[Met]}{dt} = -X_v \left(\frac{K_{210}[Met]}{K_{211} + [Met]} \right) \tag{C.19}$$

$$\frac{d[Val]}{dt} = -X_v \left(\frac{K_{220}[Val]}{K_{221} + [Val]} \right) \tag{C.20}$$

$$\frac{d[Ile]}{dt} = -X_v \left(\frac{K_{230}[Ile]}{K_{231} + [Ile]} \right) \tag{C.21}$$

$$\frac{d[Leu]}{dt} = -X_v \left(\frac{K_{240}[Leu]}{K_{241} + [Leu]} \right) \tag{C.22}$$

$$\frac{d[Phe]}{dt} = -X_v \left(\frac{K_{250}[Phe]}{K_{251} + [Phe]} \right) \tag{C.23}$$

$$\frac{d[mAb]}{dt} = -X_v \left(K_{270} + K_{271}[Glc] \right) \tag{C.24}$$

Appendix D

Parameters used in the Mechanistic and Hybrid Models

D.1 Parameters used in the Mechanistic Dynamic Kinetic Metabolic Model adapted from Hille (2018)

Table D.1: Parameter values used in the mechanistic dynamic kinetic model.

Parameter	Value	Parameter	Value	Parameter	Value
μ	1.421	K_{91}	49.123	K_{144}	19.673
K_{21}	1.233	K_{92}	138.192	K_{150}	0.036
K_{23}	12.501	K_{93}	14.836	K_{151}	0.228
K_{25}	0.803	K_{101}	0.358	K_{160}	0.754
K_d	0.005	K_{102}	41.598	K_{161}	39.940
K_{22}	3.622	K_{103}	2.386	K_{170}	0.101
K_{24}	65.685	K_{104}	0.802	K_{171}	3.487
K_{lys}	0.430	K_{105}	66.373	K_{180}	0.021
K_{41}	1.047	K_{106}	25.556	K_{181}	0.074
K_{42}	1.20E-05	K_{1001}	29.186	K_{190}	0.190
K_{43}	112.239	K_{1002}	19.816	K_{191}	4.759
K_{51}	0.850	K_{1003}	1.257	K_{100}	3.063
K_{52}	2.125	K_{1004}	5.397	K_{201}	100.00
K_{53}	46.769	K_{110}	0.137	K_{210}	1.799
K_{54}	106.698	K_{111}	3.381	K_{211}	44.371
K_{61}	0.374	K_{120}	8.555	K_{220}	4.216
K_{62}	0.113	K_{121}	60.028	K_{221}	81.768
K_{63}	0.054	K_{122}	0.061	K_{230}	0.119
K_{64}	0.330	K_{123}	39.450	K_{231}	1.490
K_{66}	0.247	K_{124}	4.167	K_{240}	0.091
K_{67}	0.156	K_{130}	0.350	K_{241}	0.001
K_{68}	18.491	K_{131}	0.604	K_{250}	1.990
K_{601}	0.176	K_{132}	1.465	K_{251}	87.052
K_{602}	0.310	K_{133}	0.123	K_{270}	5.046
K_{81}	0.026	K_{134}	0.031	K_{271}	9.697
K_{82}	0.386	K_{140}	0.002	n	0.105
K_{83}	4.293	K_{143}	0.447		

D.2 Parameters used in the re-calibrated Hybrid Model

Table D.2: Parameter values used in the hybrid model (re-calibrated).

Parameter	Value	Parameter	Value	Parameter	Value
μ	1.544	K_{91}	49.123	K_{144}	19.673
K_{21}	1.153	K_{92}	130.908	K_{150}	0.036
K_{23}	11.707	K_{93}	15.207	K_{151}	0.228
K_{25}	0.786	K_{101}	0.358	K_{160}	0.754
K_d	0.005	K_{102}	38.916	K_{161}	39.940
K_{22}	3.622	K_{103}	2.258	K_{170}	0.101
K_{24}	65.685	K_{104}	0.802	K_{171}	3.487
K_{lys}	0.430	K_{105}	66.280	K_{180}	0.021
K_{41}	0.983	K_{106}	25.556	K_{181}	0.074
K_{42}	1.20E-05	K_{1001}	28.653	K_{190}	0.190
K_{43}	118.215	K_{1002}	19.438	K_{191}	4.759
K_{51}	0.795	K_{1003}	1.272	K_{100}	3.063
K_{52}	2.025	K_{1004}	5.397	K_{201}	100.000
K_{53}	48.804	K_{110}	0.137	K_{210}	1.799
K_{54}	109.557	K_{111}	3.381	K_{211}	44.371
K_{61}	0.374	K_{120}	8.819	K_{220}	4.216
K_{62}	0.113	K_{121}	60.028	K_{221}	81.768
K_{63}	0.054	K_{122}	0.061	K_{230}	0.119
K_{64}	0.331	K_{123}	39.450	K_{231}	1.490
K_{66}	0.247	K_{124}	4.167	K_{240}	0.091
K_{67}	0.156	K_{130}	0.350	K_{241}	0.001
K_{68}	18.491	K_{131}	0.604	K_{250}	1.990
K_{601}	0.176	K_{132}	1.465	K_{251}	87.052
K_{602}	0.310	K_{133}	0.123	K_{270}	5.046
K_{81}	0.026	K_{134}	0.031	K_{271}	9.697
K_{82}	0.370	K_{140}	0.002	n	0.105
K_{83}	4.293	K_{143}	0.447	YL_{10}	4.529
YL_1	16.419	YL_4	10.040	YL_7	4.917
YL_2	9.166	YL_5	10.146	YL_8	4.550
YL_3	16.101	YL_6	11.770	YL_9	2.879

Appendix E

Metabolites Concentration

Notice that each major metabolite concentration is normalized between 0 and 1 through day 0 to day 6, for each media. Tables E.1 to E.11 presents the major metabolites concentration, unitless, over time.

Tables E.12 to E.14 presents the concentration of minor metabolites normalized between 0 and 1 in each media formulation. Minor metabolites concentration are also unitless.

Table E.1: Major metabolites concentration for Medium 10.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.6381	0.8912	1.0000	0.9234
Dead Cells	0.0000	0.1399	0.0940	0.2913	1.0000
mAb	0.0000	0.4101	0.6022	0.7968	1.0000
Glutamine	0.5455	1.0000	0.2727	0.0000	0.8182
Glucose	0.8701	0.0000	1.0000	0.6863	0.4314
Lactate	0.0000	1.0000	0.9417	0.8641	0.7767
Ammonia	0.0000	0.7196	0.7232	0.7528	1.0000
Aspartate	1.0000	0.7016	0.2882	0.0025	0.0000
Serine	1.0000	0.1461	0.0442	0.0201	0.0000
Glutamate	1.0000	0.6542	0.4179	0.0022	0.0000
Glycine	0.0000	1.0000	0.8363	0.7634	0.5683
Histidine	1.0000	0.5289	0.2589	0.1593	0.0000
Arginine	1.0000	0.4825	0.2394	0.1228	0.0000
Threonine	1.0000	0.6232	0.2811	0.1053	0.0000
Alanine	0.0000	0.8018	0.9122	1.0000	0.6929
Proline	1.0000	0.4943	0.3037	0.1757	0.0000
Cysteine	1.0000	0.2954	0.1164	0.0000	0.0064
Tyrosine	1.0000	0.2972	0.2154	0.1034	0.0000
Valine	1.0000	0.5127	0.2945	0.1326	0.0000
Methionine	1.0000	0.4227	0.2318	0.1229	0.0000
Lysine	1.0000	0.3804	0.1747	0.0354	0.0000
Isoleucine	1.0000	0.4853	0.2708	0.1091	0.0000
Leucine	1.0000	0.4813	0.2499	0.0894	0.0000
Phenylalanine	1.0000	0.4184	0.2183	0.1064	0.0000

Table E.2: Major metabolites concentration for Medium 11.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.4802	0.6766	0.8626	1.0000
Dead Cells	0.0000	0.0350	0.1329	0.3322	1.0000
mAb	0.0000	0.3249	0.5069	0.7026	1.0000
Glutamine	0.4773	1.0000	0.9318	0.7273	0.0000
Glucose	1.0000	0.2608	0.8427	0.4871	0.0000
Lactate	0.0000	1.0000	0.9810	0.9905	0.9238
Ammonia	0.0000	0.5636	0.8315	1.0000	0.9770
Aspartate	0.4752	0.6877	1.0000	0.6918	0.0000
Serine	1.0000	0.4353	0.2672	0.1251	0.0000
Glutamate	0.0000	0.2865	0.6971	1.0000	0.7596
Glycine	0.0000	0.4188	0.5936	0.9111	1.0000
Histidine	1.0000	0.8022	0.7587	0.7271	0.0000
Arginine	1.0000	0.2705	0.2788	0.4315	0.0000
Threonine	1.0000	0.5214	0.4333	0.3225	0.0000
Alanine	0.0000	0.2415	0.5718	0.9863	1.0000
Proline	1.0000	0.3936	0.3508	0.4102	0.0000
Cysteine	1.0000	0.3745	0.3068	0.2647	0.0000
Tyrosine	1.0000	0.0000	0.7500	0.2032	0.0339
Valine	1.0000	0.4469	0.4106	0.3365	0.0000
Methionine	1.0000	0.4041	0.3460	0.2869	0.0000
Lysine	1.0000	0.4609	0.5084	0.4170	0.0000
Isoleucine	1.0000	0.4216	0.4042	0.3531	0.0000
Leucine	1.0000	0.4547	0.4020	0.3284	0.0000
Phenylalanine	1.0000	0.4137	0.2894	0.2422	0.0000

Table E.3: Major metabolites concentration for Medium 16.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.5739	0.7317	0.8575	1.0000
Dead Cells	0.0000	0.1010	0.2146	0.4899	1.0000
mAb	0.0000	0.3556	0.5546	0.7319	1.0000
Glutamine	1.0000	0.3750	0.7500	0.3125	0.0000
Glucose	1.0000	0.0000	0.8992	0.5756	0.1751
Lactate	0.0000	1.0000	0.9752	0.9917	0.9008
Ammonia	0.0000	0.9050	1.0000	0.9441	0.8212
Aspartate	0.7797	1.0000	0.5232	0.1392	0.0000
Serine	1.0000	0.2087	0.0553	0.0000	0.0076
Glutamate	0.0000	1.0000	0.6027	0.0682	0.0659
Glycine	0.0000	1.0000	0.9138	0.5065	0.6120
Histidine	1.0000	0.6967	0.3990	0.0000	0.0554
Arginine	1.0000	0.7901	0.5898	0.0123	0.0000
Threonine	1.0000	0.7427	0.4639	0.1022	0.0000
Alanine	0.0000	0.3648	0.4963	0.6135	1.0000
Proline	1.0000	0.6704	0.4200	0.0960	0.0000
Cysteine	1.0000	0.3628	0.1664	0.0101	0.0000
Tyrosine	1.0000	0.1007	0.3360	0.0129	0.0000
Valine	1.0000	0.6460	0.3508	0.0815	0.0000
Methionine	1.0000	0.4893	0.2813	0.0397	0.0000
Lysine	0.9947	1.0000	0.4925	0.0000	0.0089
Isoleucine	1.0000	0.6094	0.3386	0.0828	0.0000
Leucine	1.0000	0.5800	0.2915	0.0659	0.0000
Phenylalanine	1.0000	0.4275	0.2339	0.0254	0.0000

Table E.4: Major metabolites concentration for Medium 17.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.5149	0.7319	0.8992	1.0000
Dead Cells	0.0000	0.1368	0.2393	0.5684	1.0000
mAb	0.0000	0.3530	0.6034	0.7767	1.0000
Glutamine	0.6429	1.0000	0.6071	0.5357	0.0000
Glucose	1.0000	0.2131	0.8418	0.4219	0.0000
Lactate	0.0000	1.0000	1.0000	0.9010	0.8218
Ammonia	0.0000	0.7465	0.9452	0.9615	1.0000
Aspartate	0.4077	1.0000	0.4904	0.0899	0.0000
Serine	1.0000	0.4652	0.2215	0.0350	0.0000
Glutamate	0.0000	1.0000	0.8543	0.6022	0.9842
Glycine	0.0000	0.6342	0.8422	0.7206	1.0000
Histidine	1.0000	0.8571	0.5447	0.0041	0.0000
Arginine	1.0000	0.6849	0.4879	0.0000	0.0196
Threonine	1.0000	0.7624	0.4721	0.0000	0.0881
Alanine	0.0000	0.4691	0.7599	0.7673	1.0000
Proline	1.0000	0.7768	0.5278	0.0000	0.0196
Cysteine	1.0000	0.5306	0.3651	0.0124	0.0000
Tyrosine	1.0000	0.2831	0.2871	0.0000	0.0374
Valine	1.0000	0.7728	0.4917	0.0735	0.0000
Methionine	1.0000	0.6198	0.4127	0.0150	0.0000
Lysine	0.8682	1.0000	0.4768	0.0342	0.0000
Isoleucine	1.0000	0.7482	0.4878	0.0592	0.0000
Leucine	1.0000	0.7532	0.4968	0.0840	0.0000
Phenylalanine	1.0000	0.5411	0.3966	0.0000	0.0207

Table E.5: Major metabolites concentration for Medium 19.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.7270	0.9380	1.0000	0.9670
Dead Cells	0.0000	0.0830	0.2380	0.3384	1.0000
mAb	0.0000	0.3305	0.4945	0.7312	1.0000
Glutamine	0.1607	0.3393	0.0000	0.3214	1.0000
Glucose	1.0000	0.0000	0.9708	0.5544	0.2944
Lactate	0.0990	1.0000	0.8614	0.3861	0.0000
Ammonia	0.0000	0.9502	0.9042	0.7778	1.0000
Aspartate	1.0000	0.8464	0.4905	0.0990	0.0000
Serine	1.0000	0.1719	0.0560	0.0278	0.0000
Glutamate	1.0000	0.8879	0.7148	0.3165	0.0000
Glycine	0.0000	1.0000	0.5474	0.0940	0.5106
Histidine	1.0000	0.6675	0.3484	0.0000	0.6259
Arginine	1.0000	0.4942	0.2795	0.0688	0.0000
Threonine	1.0000	0.8162	0.4852	0.1112	0.0000
Alanine	0.0000	0.9256	0.9206	1.0000	0.3785
Proline	1.0000	0.5673	0.3527	0.1200	0.0000
Cysteine	1.0000	0.5252	0.2889	0.0830	0.0000
Tyrosine	1.0000	0.8065	0.3040	0.0698	0.0000
Valine	1.0000	0.6061	0.3448	0.1006	0.0000
Methionine	1.0000	0.4755	0.2237	0.0411	0.0000
Lysine	1.0000	0.5313	0.2655	0.0580	0.0000
Isoleucine	1.0000	0.6082	0.3292	0.0723	0.0000
Leucine	1.0000	0.5733	0.2898	0.0561	0.0000
Phenylalanine	1.0000	0.5376	0.2674	0.0660	0.0000

Table E.6: Major metabolites concentration for Medium 22.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.6920	1.0000	0.8084	0.7680
Dead Cells	0.0000	0.0926	0.6728	0.9198	1.0000
mAb	0.0000	0.4105	0.5625	0.7662	1.0000
Glutamine	0.5000	0.0000	0.1667	0.1667	1.0000
Glucose	0.6256	0.0000	1.0000	0.8818	0.6921
Lactate	0.0000	1.0000	0.8548	0.7097	0.4194
Ammonia	0.0000	1.0000	0.8614	0.8119	0.6832
Aspartate	1.0000	0.7689	0.0851	0.0000	0.0011
Serine	1.0000	0.0843	0.0100	0.0211	0.0000
Glutamate	1.0000	0.9829	0.3525	0.0070	0.0000
Glycine	0.0000	1.0000	0.7546	0.6447	0.4152
Histidine	1.0000	0.8815	0.4647	0.3171	0.0000
Arginine	1.0000	0.7517	0.3985	0.2962	0.0000
Threonine	1.0000	0.9460	0.4932	0.3045	0.0000
Alanine	0.0000	1.0000	0.6580	0.8859	0.2894
Proline	1.0000	0.7050	0.3612	0.2622	0.0000
Cysteine	1.0000	0.4948	0.2739	0.0691	0.0000
Tyrosine	0.5460	1.0000	0.8990	0.5163	0.0000
Valine	1.0000	0.6579	0.3470	0.1841	0.0000
Methionine	1.0000	0.6035	0.3408	0.2156	0.0000
Lysine	1.0000	0.7224	0.2270	0.2485	0.0000
Isoleucine	1.0000	0.5601	0.2719	0.1239	0.0000
Leucine	1.0000	0.4502	0.1641	0.0480	0.0000
Phenylalanine	1.0000	0.5524	0.2824	0.1473	0.0000

Table E.7: Major metabolites concentration for Medium 32.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.8861	1.0000	0.8433	0.6408
Dead Cells	0.0000	0.1134	0.0988	0.3256	1.0000
mAb	0.0000	0.3169	0.6157	0.8594	1.0000
Glutamine	1.0000	0.1250	0.0000	0.2500	0.2500
Glucose	0.7574	0.0000	1.0000	0.8603	0.7598
Lactate	0.0000	1.0000	0.8590	0.7051	0.5641
Ammonia	0.0000	1.0000	0.9525	0.9119	0.9153
Aspartate	0.8075	1.0000	0.1913	0.0000	0.0321
Serine	1.0000	0.0902	0.0109	0.0011	0.0000
Glutamate	1.0000	0.9295	0.5180	0.0096	0.0000
Glycine	0.0000	0.8612	0.8615	0.8805	1.0000
Histidine	1.0000	0.4860	0.1034	0.0000	0.0324
Arginine	1.0000	0.3721	0.1142	0.0000	0.0047
Threonine	1.0000	0.3887	0.0725	0.0074	0.0000
Alanine	0.0000	0.8388	0.8133	1.0000	0.9529
Proline	1.0000	0.4308	0.1298	0.0282	0.0000
Cysteine	1.0000	0.2123	0.0904	0.0000	0.0007
Tyrosine	1.0000	0.3508	0.1478	0.0085	0.0000
Valine	1.0000	0.3960	0.1314	0.0311	0.0000
Methionine	1.0000	0.3043	0.0622	0.0000	0.0205
Lysine	1.0000	0.3180	0.0316	0.0133	0.0000
Isoleucine	1.0000	0.4291	0.1424	0.0344	0.0000
Leucine	1.0000	0.3056	0.0462	0.0000	0.0017
Phenylalanine	1.0000	0.3473	0.0811	0.0000	0.0080

Table E.8: Major metabolites concentration for Medium 51.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.8393	0.8705	1.0000	0.9707
Dead Cells	0.0000	0.2179	0.3205	0.5705	1.0000
mAb	0.0000	0.3341	0.5495	0.8270	1.0000
Glutamine	0.0000	1.0000	0.6667	0.5833	0.5000
Glucose	0.5111	0.0000	1.0000	0.8606	0.7456
Lactate	0.0000	1.0000	0.8667	0.7467	0.6267
Ammonia	0.0000	1.0000	1.0000	0.9091	0.9818
Aspartate	1.0000	0.6221	0.3921	0.2031	0.0000
Serine	1.0000	0.0761	0.0062	0.0074	0.0000
Glutamate	1.0000	0.9583	0.7648	0.5372	0.0000
Glycine	0.2798	1.0000	0.4187	0.1667	0.0000
Histidine	1.0000	0.7598	0.2139	0.1798	0.0000
Arginine	1.0000	0.5442	0.2359	0.2244	0.0000
Threonine	1.0000	0.7557	0.3007	0.1716	0.0000
Alanine	0.0000	0.3412	0.5569	0.8888	1.0000
Proline	1.0000	0.5216	0.3253	0.2350	0.0000
Cysteine	1.0000	0.5018	0.2630	0.1335	0.0000
Tyrosine	0.4427	0.0400	1.0000	0.0884	0.0000
Valine	1.0000	0.5937	0.3403	0.1927	0.0000
Methionine	1.0000	0.4552	0.2384	0.1342	0.0000
Lysine	1.0000	0.5327	0.2929	0.2509	0.0000
Isoleucine	1.0000	0.5993	0.3484	0.1768	0.0000
Leucine	1.0000	0.5504	0.3079	0.1368	0.0000
Phenylalanine	1.0000	0.4656	0.2875	0.1225	0.0000

Table E.9: Major metabolites concentration for Medium 56.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.5057	0.7495	0.9300	1.0000
Dead Cells	0.0000	0.1532	0.2339	0.4879	1.0000
mAb	0.0000	0.2929	0.4673	0.6997	1.0000
Glutamine	0.0000	1.0000	0.7838	0.6486	0.8378
Glucose	1.0000	0.1644	0.8611	0.4606	0.0000
Lactate	0.0000	1.0000	0.8992	0.6891	0.3025
Ammonia	0.0000	0.9488	1.0000	0.8841	0.7951
Aspartate	0.5781	1.0000	0.6622	0.1229	0.0000
Serine	1.0000	0.4124	0.2111	0.0363	0.0000
Glutamate	0.6415	1.0000	0.8135	0.1450	0.0000
Glycine	0.0000	0.8429	1.0000	0.5881	0.4047
Histidine	0.9671	1.0000	0.7050	0.1736	0.0000
Arginine	0.9701	1.0000	0.7441	0.1998	0.0000
Threonine	0.9468	1.0000	0.7438	0.1985	0.0000
Alanine	0.0000	0.7482	1.0000	0.4281	0.3738
Proline	1.0000	0.8365	0.6686	0.1582	0.0000
Cysteine	1.0000	0.8907	0.6583	0.1469	0.0000
Tyrosine	1.0000	0.4124	0.3723	0.2270	0.0000
Valine	1.0000	0.9578	0.7260	0.1694	0.0000
Methionine	1.0000	0.7850	0.5985	0.1617	0.0000
Lysine	0.8498	1.0000	0.7143	0.0786	0.0000
Isoleucine	0.9955	1.0000	0.7730	0.1833	0.0000
Leucine	1.0000	0.8639	0.6262	0.1535	0.0000
Phenylalanine	1.0000	0.7024	0.5550	0.2372	0.0000

Table E.10: Major metabolites concentration for Medium 64.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.8009	0.8871	1.0000	0.6239
Dead Cells	0.0000	0.0468	0.1047	0.5799	1.0000
mAb	0.0000	0.4135	0.7422	0.8567	1.0000
Glutamine	0.2000	0.0000	0.0000	0.4667	1.0000
Glucose	0.8579	0.0000	1.0000	0.8553	0.6711
Lactate	0.0135	1.0000	0.5676	0.1622	0.0000
Ammonia	0.0000	1.0000	0.7519	0.7218	0.8271
Aspartate	1.0000	0.7286	0.0276	0.0000	0.0064
Serine	1.0000	0.0641	0.0128	0.0000	0.0101
Glutamate	1.0000	0.7680	0.1581	0.0000	0.0034
Glycine	0.0127	1.0000	0.0962	0.0000	0.2429
Histidine	1.0000	0.5573	0.2610	0.0000	0.1010
Arginine	1.0000	0.4458	0.1119	0.0000	0.0158
Threonine	1.0000	0.4828	0.1006	0.0000	0.0108
Alanine	0.6553	0.9637	1.0000	0.4992	0.0000
Proline	1.0000	0.4826	0.1835	0.0262	0.0000
Cysteine	1.0000	0.2421	0.0550	0.0000	0.0013
Tyrosine	0.7783	1.0000	0.3377	0.0034	0.0000
Valine	1.0000	0.5391	0.2249	0.0666	0.0000
Methionine	1.0000	0.3785	0.0916	0.0000	0.0057
Lysine	1.0000	0.3477	0.1582	0.0000	0.0149
Isoleucine	1.0000	0.4118	0.1071	0.0246	0.0000
Leucine	1.0000	0.3478	0.0420	0.0000	0.0009
Phenylalanine	1.0000	0.3965	0.0697	0.0000	0.0109

Table E.11: Major metabolites concentration for Medium 65.

Components	Day 0	Day 3	Day 4	Day 5	Day 6
Viable Cells	0.0000	0.9156	1.0000	0.9307	0.6860
Dead Cells	0.0000	0.0432	0.1131	0.3408	1.0000
mAb	0.0000	0.4135	0.6399	0.7607	1.0000
Glutamine	0.4000	0.0000	0.6000	0.8000	1.0000
Glucose	0.7631	0.0000	1.0000	0.8884	0.7745
Lactate	0.0000	1.0000	0.6462	0.5385	0.3385
Ammonia	0.0000	1.0000	0.7739	0.7826	0.8174
Aspartate	1.0000	0.7186	0.0872	0.0000	0.0062
Serine	1.0000	0.0205	0.0000	0.0682	0.0623
Glutamate	1.0000	0.8640	0.3032	0.0000	0.0039
Glycine	0.4014	1.0000	0.0000	0.8267	0.9149
Histidine	1.0000	0.3884	0.0080	0.3232	0.0000
Arginine	1.0000	0.3520	0.0000	0.3494	0.2866
Threonine	1.0000	0.3835	0.0000	0.0722	0.0324
Alanine	0.0000	0.0649	0.2445	1.0000	0.8225
Proline	1.0000	0.3429	0.0000	0.1837	0.0616
Cysteine	1.0000	0.3179	0.0575	0.0131	0.0000
Tyrosine	1.0000	0.0000	0.0790	0.1741	0.1142
Valine	1.0000	0.3566	0.0703	0.0157	0.0000
Methionine	1.0000	0.2166	0.0000	0.0442	0.0685
Lysine	1.0000	0.3490	0.0000	0.2580	0.2623
Isoleucine	1.0000	0.2575	0.0221	0.0036	0.0000
Leucine	1.0000	0.1780	0.0009	0.0000	0.0015
Phenylalanine	1.0000	0.2316	0.0000	0.0613	0.0476

Table E.12: Minor metabolites concentration in the media formulation for Media 10, 11, 16 and 17.

Minor Metabolites	Concentration			
	Medium 10	Medium 11	Medium 16	Medium 17
Component 01	0.0195	0.0082	8.00E-04	0.0054
Component 02	2.02E-04	1.05E-04	1.50E-04	1.31E-04
Component 03	3.27E-04	1.70E-04	2.43E-04	2.13E-04
Component 04	2.49E-04	1.71E-04	3.03E-04	2.55E-04
Component 05	2.69E-04	1.40E-04	3.64E-04	1.75E-04
Component 06	0.0015	6.26E-04	7.14E-04	0.0010
Component 07	9.12E-04	0	4.57E-05	2.00E-05
Component 08	1.02E-04	1.80E-06	5.26E-05	2.25E-06
Component 09	0.3417	0.1565	0.2714	0.2366
Component 11	0.2761	0.1197	0.1857	0.1936
Component 12	0.0043	0.0030	0.0024	0.0023
Component 13	0.0015	9.36E-04	0.0011	9.14E-04
Component 14	0.0019	2.84E-04	0.0010	2.15E-04
Component 16	2.22E-04	7.28E-05	0	4.55E-05
Component 17	2.22E-04	5.67E-05	0	3.55E-05
Component 19	1.34E-06	7.00E-07	1.00E-06	8.75E-07
Component 21	5.87E-10	2.46E-10	2.86E-06	7.81E-06
Component 22	3.36E-06	1.30E-06	2.43E-06	1.75E-06
Component 23	3.06E-10	2.43E-10	2.56E-10	2.89E-10
Component 24	1.92E-06	1.00E-06	2.57E-06	1.25E-06
Component 25	5.36E-06	2.00E-06	2.86E-06	2.50E-06
Component 26	7.55E-06	2.00E-06	4.71E-06	2.50E-06
Component 27	5.76E-05	3.00E-05	4.29E-05	3.75E-05
Component 29	0	1.20E-06	0	7.50E-07
Component 31	0.0054	0.0013	0.0036	0.0020
Component 32	0.0869	0.0364	0	0.0228
Component 33	0.0983	0.0404	0.0086	0.0290
Component 34	0.0089	0.0125	0.0107	0.0078
Component 35	0.0236	0.0061	0.0118	0.0039
Component 36	0.0438	0.0198	0.0229	0.0219
Component 37	0.1060	0.0489	0.0530	0.0569
Component 38	0.0224	0.0109	0.0636	0.0509

Component 39	2.60E-04	1.19E-04	2.59E-04	2.09E-04
Component 40	0.0012	3.00E-06	8.57E-04	7.52E-04
Component 41	0.0012	6.76E-04	0.0010	0.0009
Component 43	0.0029	0.0024	0.0036	0.0027
Component 44	0.0308	0.0191	0.0229	0.0191
Component 45	0.0018	0.0012	0.0020	0.0017
Component 46	0	0	0	0
Component 47	0.0055	0	0.0041	4.38E-04
Component 48	0	0	0	0
Component 49	0.0122	0.0030	0.0063	0.0022
Component 50	1.08E-04	1.02E-04	5.86E-05	7.06E-05
Component 51	0.0002	1.05E-04	1.50E-04	1.31E-04
Component 52	0.2503	0.1687	0.2214	0.2398
Component 53	0.0204	0.0096	0.0150	0.0125
Component 54	0.0087	0	0	0.0019
Component 56	3.39E-06	1.08E-06	2.36E-06	1.84E-06
Component 57	1.56E-06	0	1.86E-06	1.44E-06
Component 58	8.40E-05	2.00E-05	1.66E-04	1.10E-04
Component 59	5.22E-07	2.19E-07	1.08E-09	3.25E-07
Component 61	1.0000	1.0000	1.0000	1.0000
Component 62	3.11E-05	0	3.31E-05	2.90E-05
Component 63	2.49E-04	1.03E-04	1.50E-04	1.61E-04
Component 64	2.02E-04	1.05E-04	1.50E-04	1.31E-04
Component 65	6.72E-04	4.50E-04	0.0013	9.28E-04
Component 66	1.50E-04	5.70E-04	0.0012	5.51E-04
Component 67	0.0066	0.0049	0.0052	0.0049
Component 68	0.0025	0	0.0019	8.13E-04
Component 69	0.0031	4.90E-04	0.0031	0.0020
Component 70	0.0127	0.0031	0.0071	0.0025
Component 71	0.0019	0.0010	0.0157	0.0013
Component 72	0.0147	0.0094	0.0109	0.0111
Component 74	0.0023	0.0013	0.0033	0.0026
Component 75	5.72E-10	2.40E-10	5.67E-10	5.22E-10
Component 76	0	0	0	0
Component 77	1.25E-09	5.23E-10	1.23E-09	1.14E-09
Component 78	8.96E-10	3.75E-10	8.79E-10	8.17E-10
Component 79	0	0	0	0

Component 80	2.64E-10	1.10E-10	2.59E-10	2.41E-10
Component 81	1.16E-09	4.88E-10	1.14E-09	1.06E-09
Component 82	0	0	0	0
Component 83	0	0	0	0
Component 84	2.33E-10	9.78E-11	0	6.11E-11
Component 85	0.0085	0	0.0043	0.0019
Component 86	2.32E-07	0	1.07E-07	4.69E-08
Component 87	1.50E-10	6.26E-11	4.34E-10	1.36E-10
Component 88	0.00E+00	0	0	0
Component 89	5.86E-11	2.45E-11	5.75E-11	5.35E-11
Component 90	8.31E-11	3.48E-11	8.16E-11	7.58E-11
Component 91	0	0	0	0
Component 92	0	0	0	0
Component 94	0	0	0	0
Component 95	5.92E-10	2.48E-10	5.81E-10	5.40E-10
Component 96	0	0	0	0
Component 97	7.48E-11	3.13E-11	7.41E-11	6.82E-11
Component 98	2.06E-09	8.61E-10	2.02E-09	1.88E-09
Component 99	0	0	4.29E-10	0
Component 100	1.17E-10	4.91E-11	1.15E-10	1.07E-10
Component 101	1.58E-09	6.60E-10	1.55E-09	1.44E-09
Component 102	2.40E-04	0	0	0

Table E.13: Minor metabolites concentration in the media formulation for Media 19, 22, 32 and 51.

Minor Metabolites	Concentration			
	Medium 19	Medium 22	Medium 32	Medium 51
Component 01	0.0004	0.0039	0.0020	0.0024
Component 02	7.55E-05	2.58E-04	1.82E-04	2.41E-04
Component 03	1.22E-04	4.18E-04	2.95E-04	3.90E-04
Component 04	3.69E-04	2.52E-04	3.67E-04	1.06E-04
Component 05	4.93E-04	3.44E-04	2.43E-04	3.24E-04
Component 06	2.48E-04	0	0	0
Component 07	3.92E-04	0.0012	6.70E-04	7.33E-05
Component 08	3.82E-05	3.85E-04	5.30E-06	2.40E-04
Component 09	0.0899	0.3272	0.2262	0.3352
Component 11	0.2032	0.2550	0.1571	0.2521
Component 12	0.0026	0.0053	0.0077	0.0016
Component 13	8.13E-04	0.0018	0.0021	6.66E-04
Component 14	0.0011	0.0062	7.22E-04	0.0036
Component 16	0	4.91E-04	2.52E-05	4.58E-04
Component 17	0	3.32E-04	1.19E-04	6.19E-04
Component 19	2.32E-05	1.72E-06	1.22E-06	1.60E-06
Component 21	0	0	2.34E-10	0
Component 22	1.08E-06	5.53E-06	2.78E-06	3.55E-05
Component 23	4.39E-09	2.46E-10	2.25E-10	2.29E-10
Component 24	1.91E-06	2.46E-06	1.74E-06	8.02E-06
Component 25	2.44E-06	1.07E-05	8.41E-06	8.44E-06
Component 26	4.35E-05	1.92E-05	6.02E-06	1.40E-05
Component 27	5.31E-05	7.37E-05	5.21E-05	6.88E-05
Component 29	7.61E-06	0	2.48E-06	0
Component 31	0.0019	0.0073	0.0027	0.0067
Component 32	0	0	0	0
Component 33	0.0187	0.0680	0.0397	0.0928
Component 34	0.0173	0.0211	0.0142	0.0215
Component 35	0.0065	0.0302	0.0183	5.27E-04
Component 36	0.0576	0.0468	0.0260	0.0444
Component 37	0.0495	0.0856	0.0654	0.0819
Component 38	0.0361	0.0337	0.0209	0.0296

Component 39	7.49E-05	1.22E-04	9.70E-05	4.54E-04
Component 40	0	0	0	0
Component 41	1.08E-04	6.67E-04	6.95E-04	3.44E-04
Component 43	0.0010	0.0025	0.0017	0.0023
Component 44	0.0140	0.0310	0.0405	0.0205
Component 45	0.0012	0.0021	0.0027	8.54E-04
Component 46	0	0	4.34E-05	0
Component 47	5.04E-04	0.0071	0.0213	0.0016
Component 48	0	0	0.0017	0
Component 49	0.0067	0.0615	0.0079	0.0636
Component 50	7.91E-06	1.38E-04	2.70E-04	2.56E-05
Component 51	7.55E-05	2.58E-04	1.82E-04	2.41E-04
Component 52	0.2518	0.5399	0.4949	0.4097
Component 53	0.0068	0.0254	0.0161	0.0214
Component 54	0	0.0031	0	0.0086
Component 56	9.35E-07	5.35E-06	1.62E-06	5.42E-06
Component 57	1.45E-06	2.42E-06	5.78E-07	1.60E-06
Component 58	6.91E-05	7.52E-05	6.54E-05	1.54E-04
Component 59	4.17E-08	0	4.93E-10	0
Component 61	1.0000	1.0000	1.0000	1.0000
Component 62	4.45E-06	2.95E-06	4.43E-06	2.75E-06
Component 63	2.16E-05	1.43E-04	6.86E-05	1.69E-04
Component 64	7.55E-05	2.58E-04	1.82E-04	2.41E-04
Component 65	6.04E-04	8.97E-04	0.0010	5.44E-04
Component 66	0	5.70E-04	1.88E-05	6.23E-04
Component 67	0.0032	0.0083	0.0126	0.0014
Component 68	9.35E-04	0.0092	0.0044	0.0142
Component 69	4.60E-04	0.0016	0.0011	0.0015
Component 70	0.0062	0.0157	0.0150	8.72E-04
Component 71	7.19E-04	0.0025	0.0628	0.0023
Component 72	0.0120	0.0120	0.0155	0.0140
Component 74	0.0014	0.0040	0.0030	0.0024
Component 75	0	0	2.53E-10	0
Component 76	0	0	0	5.73E-04
Component 77	0	0	4.97E-10	0
Component 78	0	0	3.56E-10	0
Component 79	0	0	1.62E-10	0

Component 80	0	0	1.05E-10	0
Component 81	0	0	4.64E-10	0
Component 82	2.09E-04	0	2.52E-04	5.73E-04
Component 83	0	1.54E-06	0	4.30E-06
Component 84	0	0	1.03E-10	0
Component 85	0.0022	0.0109	0.0052	0.0069
Component 86	5.96E-08	2.97E-07	1.30E-07	1.72E-07
Component 87	0	0	6.61E-11	0
Component 88	0	0	0	4.87E-04
Component 89	0	0	2.33E-11	0
Component 90	0	0	3.31E-11	0
Component 91	0	2.33E-06	0	2.18E-06
Component 92	0	5.53E-06	0	1.20E-05
Component 94	0	0	1.74E-05	1.15E-04
Component 95	0	0	2.36E-10	0
Component 96	0	0	3.02E-11	0
Component 97	0	0	3.31E-11	0
Component 98	0	0	8.18E-10	0
Component 99	3.26E-09	0	0	0
Component 100	0	0	4.74E-11	0
Component 101	0	0	6.27E-10	0
Component 102	0	9.22E-04	1.74E-04	0.0017

Table E.14: Minor metabolites concentration in the media formulation for Media 56, 64 and 65.

Minor Metabolites	Concentration		
	Medium 56	Medium 64	Medium 65
Component 01	0	0.0022	0.0010
Component 02	0	2.12E-04	1.91E-04
Component 03	0	3.43E-04	3.09E-04
Component 04	3.66E-04	3.63E-04	3.74E-04
Component 05	4.40E-04	4.59E-04	5.72E-04
Component 06	4.11E-04	0	0
Component 07	2.54E-04	0.0013	6.02E-04
Component 08	0	1.59E-04	9.65E-05
Component 09	0	0.2610	0.2273
Component 11	0.2221	0.1951	0.1744
Component 12	0.0021	0.0057	0.0034
Component 13	6.36E-04	0.0017	0.0011
Component 14	0	0.0035	0.0028
Component 16	0	5.04E-05	0
Component 17	0	9.08E-05	0
Component 19	0	3.32E-05	5.85E-05
Component 21	0	0	0
Component 22	0	3.53E-06	2.73E-06
Component 23	0	6.25E-09	1.11E-08
Component 24	0	3.68E-06	4.82E-06
Component 25	1.65E-06	5.63E-06	3.64E-06
Component 26	8.25E-06	5.99E-05	9.73E-05
Component 27	5.22E-05	6.05E-05	5.45E-05
Component 29	1.26E-05	0	0
Component 31	0.0011	0.0052	0.0030
Component 32	0	0	0
Component 33	0	0.0505	0.0473
Component 34	0.0120	0.0235	0.0254
Component 35	0	0.0337	0.0165
Component 36	0.0794	0.0317	0.0243
Component 37	0.0346	0.0762	0.0721
Component 38	0.0484	0.0236	0.0173

Component 39	5.95E-05	1.11E-04	9.84E-05
Component 40	0	0	0
Component 41	0	5.48E-04	2.73E-04
Component 43	3.93E-04	0.0020	0.0018
Component 44	0.0108	0.0311	0.0189
Component 45	0.0012	0.0021	0.0013
Component 46	0	0	0
Component 47	0	0.0058	0.0013
Component 48	0	0	0
Component 49	9.40E-04	0.0209	0.0156
Component 50	0	1.14E-04	2.00E-05
Component 51	0	2.12E-04	1.91E-04
Component 52	0.1404	0.4970	0.4218
Component 53	0.0039	0.0173	0.0114
Component 54	0	0	0
Component 56	0	2.98E-06	2.36E-06
Component 57	0	2.92E-06	3.67E-06
Component 58	6.88E-05	7.37E-05	6.96E-05
Component 59	0	5.84E-08	1.05E-07
Component 61	1.0000	1.0000	1.0000
Component 62	2.37E-06	5.45E-06	7.64E-06
Component 63	0	7.94E-05	5.45E-05
Component 64	0	2.12E-04	1.91E-04
Component 65	2.14E-04	0.0011	0.0012
Component 66	0	0	0
Component 67	0.0014	0.0094	0.0059
Component 68	0	0.0026	0.0024
Component 69	0	0.0013	0.0012
Component 70	0	0.0212	0.0157
Component 71	0	0.0020	0.0018
Component 72	0.0102	0.0167	0.0149
Component 74	0.0012	0.0027	0.0016
Component 75	0	0	0
Component 76	0	0	0
Component 77	0	0	0
Component 78	0	0	0
Component 79	0	0	0

Component 80	0	0	0
Component 81	0	0	0
Component 82	0	2.93E-04	5.27E-04
Component 83	0	0	0
Component 84	0	0	0
Component 85	0	0.0089	0.0055
Component 86	0	2.52E-07	1.51E-07
Component 87	0	0	0
Component 88	0	0	0
Component 89	0	0	0
Component 90	0	0	0
Component 91	0	0	0
Component 92	0	0	0
Component 94	0	0	0
Component 95	0	0	0
Component 96	0	0	0
Component 97	0	0	0
Component 98	0	0	0
Component 99	5.40E-09	0	0
Component 100	0	0	0
Component 101	0	0	0
Component 102	0	2.52E-04	0

Appendix F

Mechanistic and Hybrid Prediction Plots

The plots presented in this appendix are discussed in Chapter 5.

The nomenclature used in the plots are as follows: [vcd]: Biomass concentration; [mAb]: Monoclonal antibodies concentration; [glc]: Glucose concentration; [lac]: Lactate concentration; [amm]: Ammonia concentration; [gln]: Glutamine concentration; [asp]: Aspartate concentration; [ser]: Serine concentration; [glu]: Glutamate concentration; [gly]: Glycine concentration; [his]: Histidine concentration; [arg]: Arginine concentration; [thr]: Threonine concentration; [ala]: Alanine concentration; [pro]: Proline concentration; [cys]: Cysteine concentration; [tyr]: Tyrosine concentration; [val]: Valine concentration; [met]: Methionine concentration; [lys]: Lysine concentration; [ile]: Isoleucine concentration; [leu]: Leucine concentration; [phe]: Phenylalanine concentration. All concentrations are unitless and normalized.

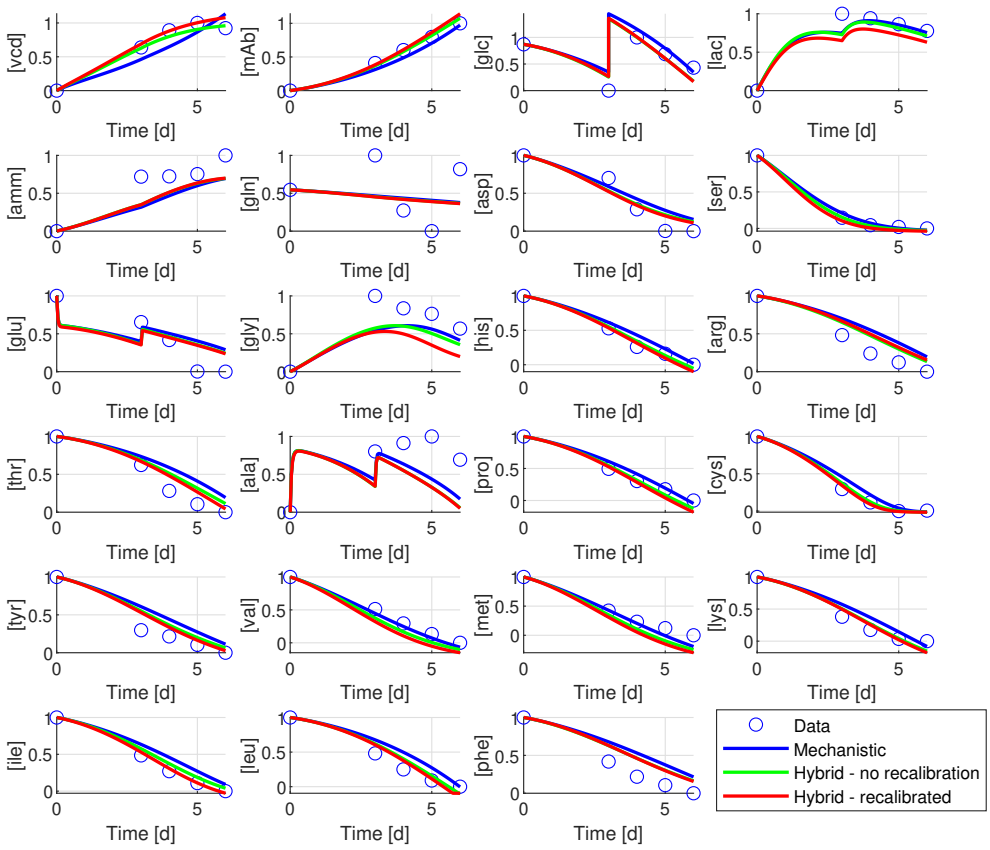


Figure F.1: Data, Mechanistic Model and Hybrid Model profiles for Medium 10.

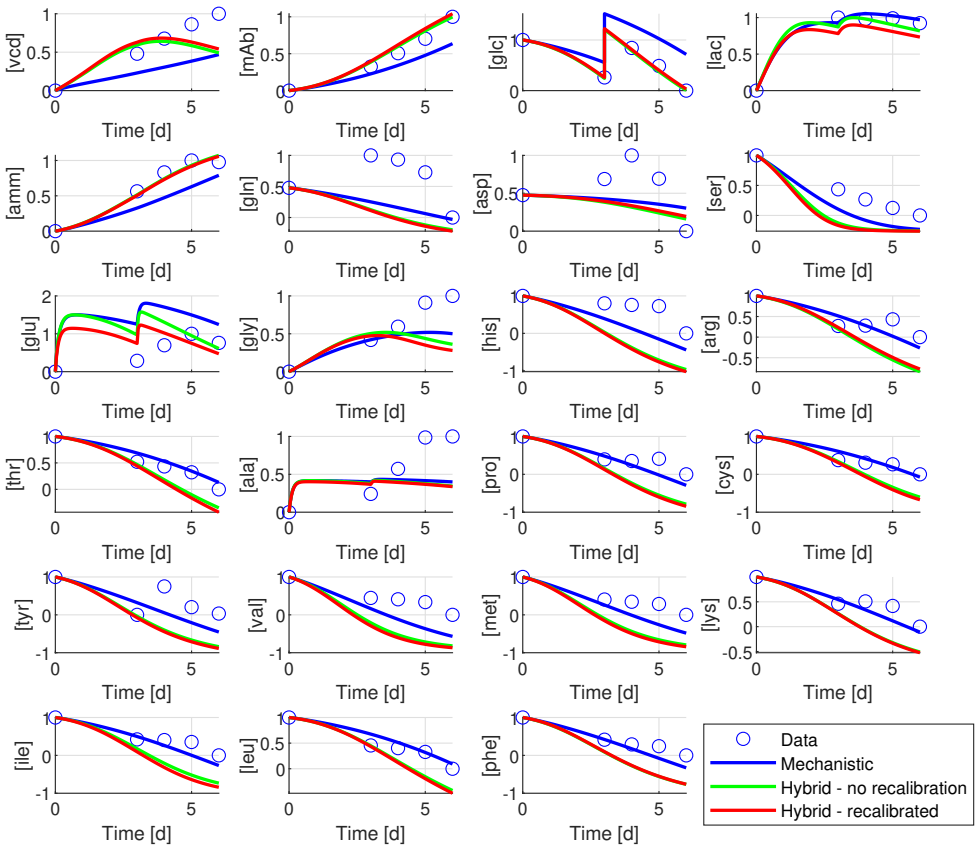


Figure F.2: Data, Mechanistic Model and Hybrid Model profiles for Medium 11.

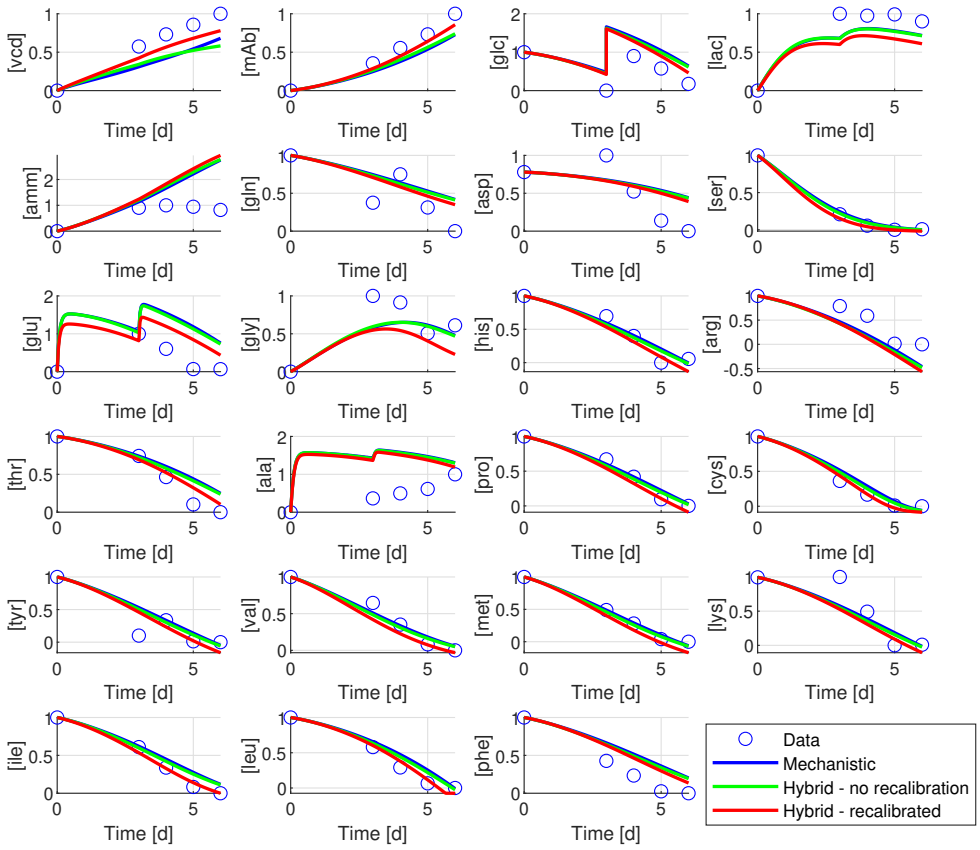


Figure F.3: Data, Mechanistic Model and Hybrid Model profiles for Medium 16.

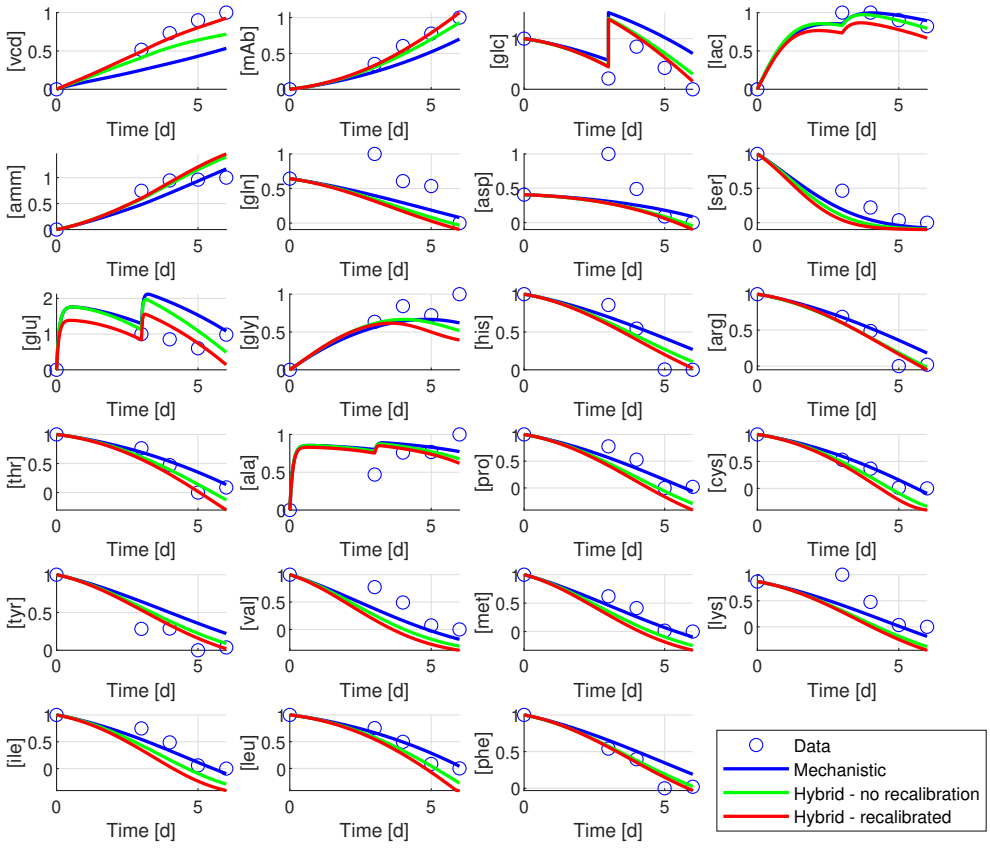


Figure F.4: Data, Mechanistic Model and Hybrid Model profiles for Medium 17.

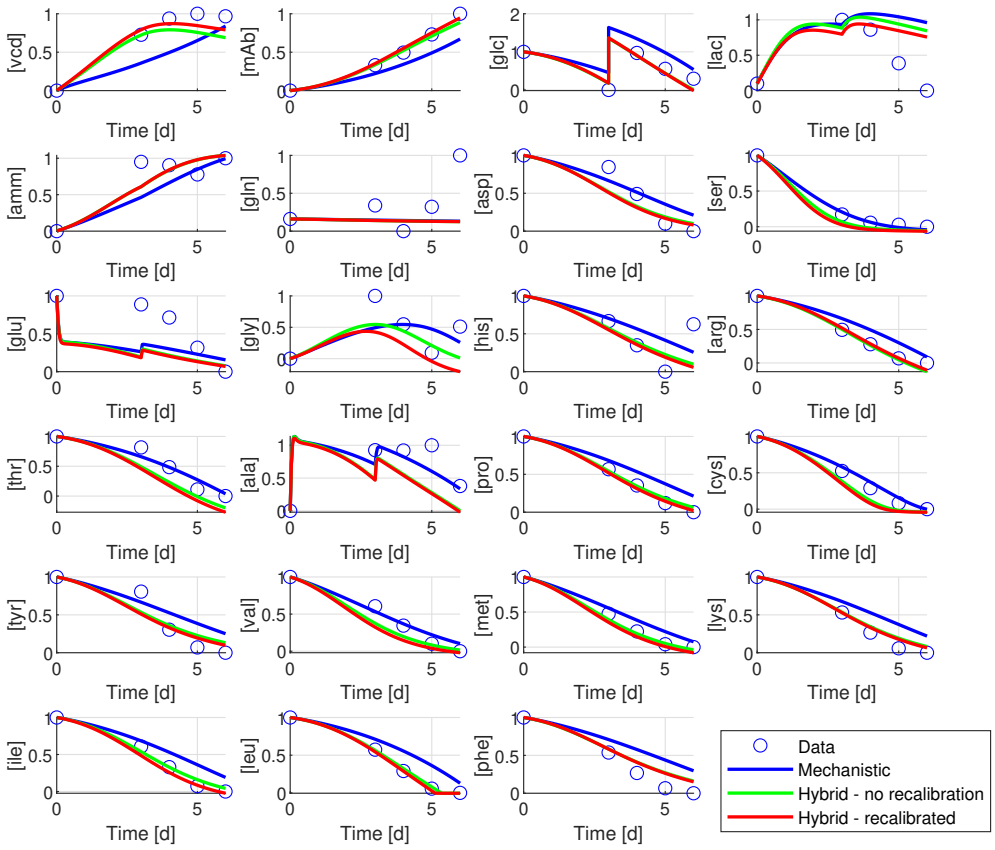


Figure F.5: Data, Mechanistic Model and Hybrid Model profiles for Medium 19.

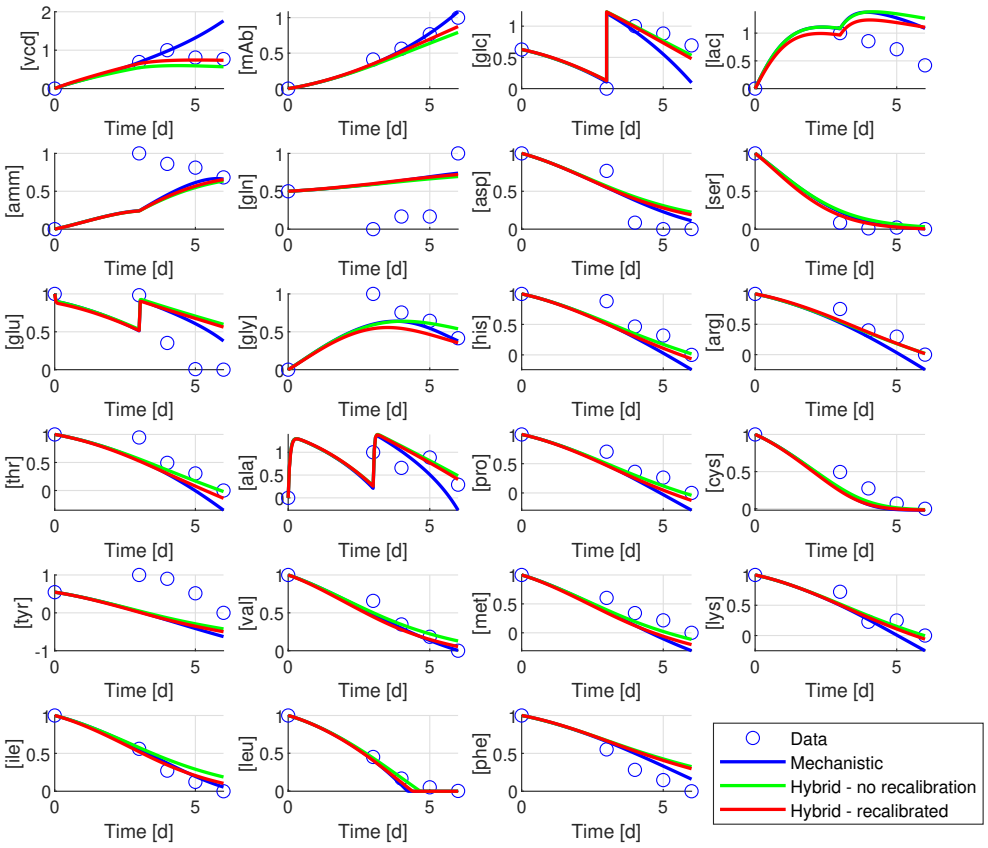


Figure F.6: Data, Mechanistic Model and Hybrid Model profiles for Medium 22.

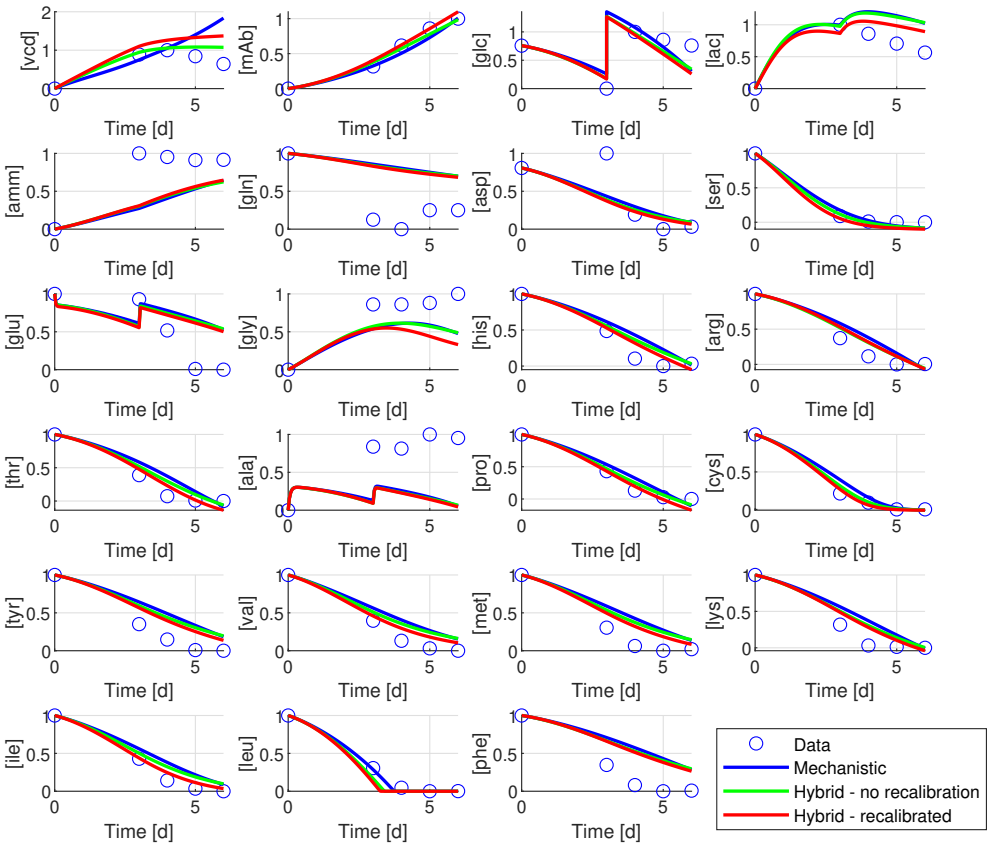


Figure F.7: Data, Mechanistic Model and Hybrid Model profiles for Medium 32.

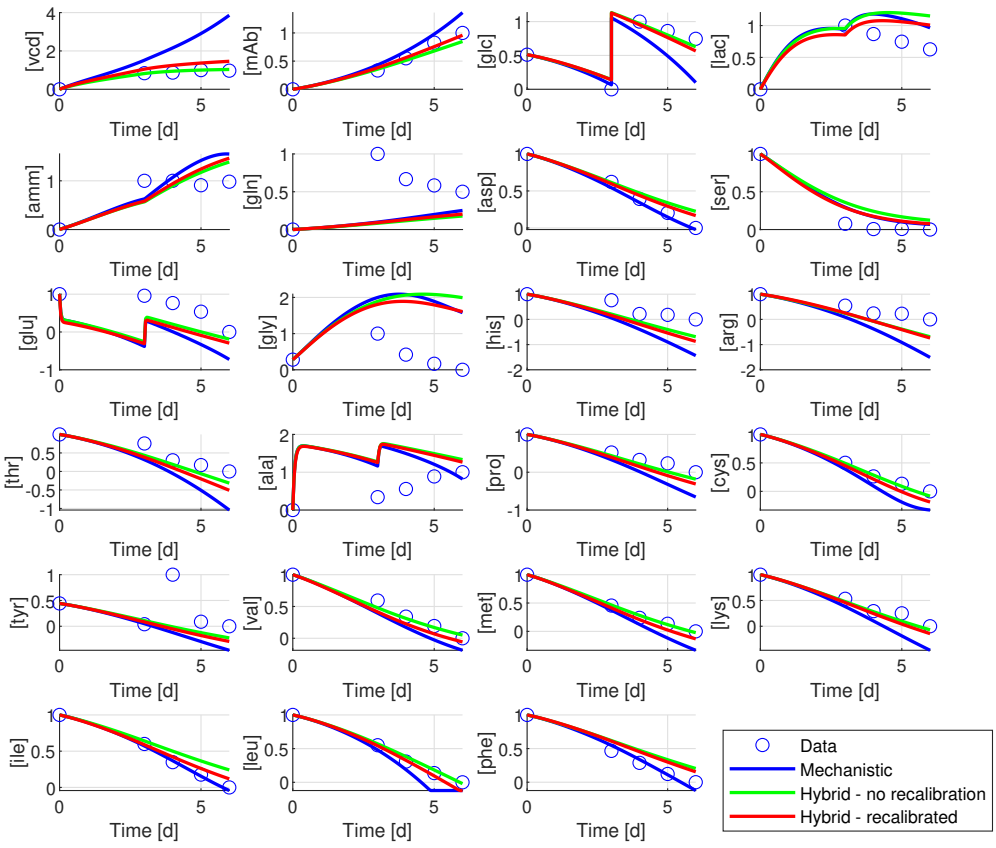


Figure F.8: Data, Mechanistic Model and Hybrid Model profiles for Medium 51.

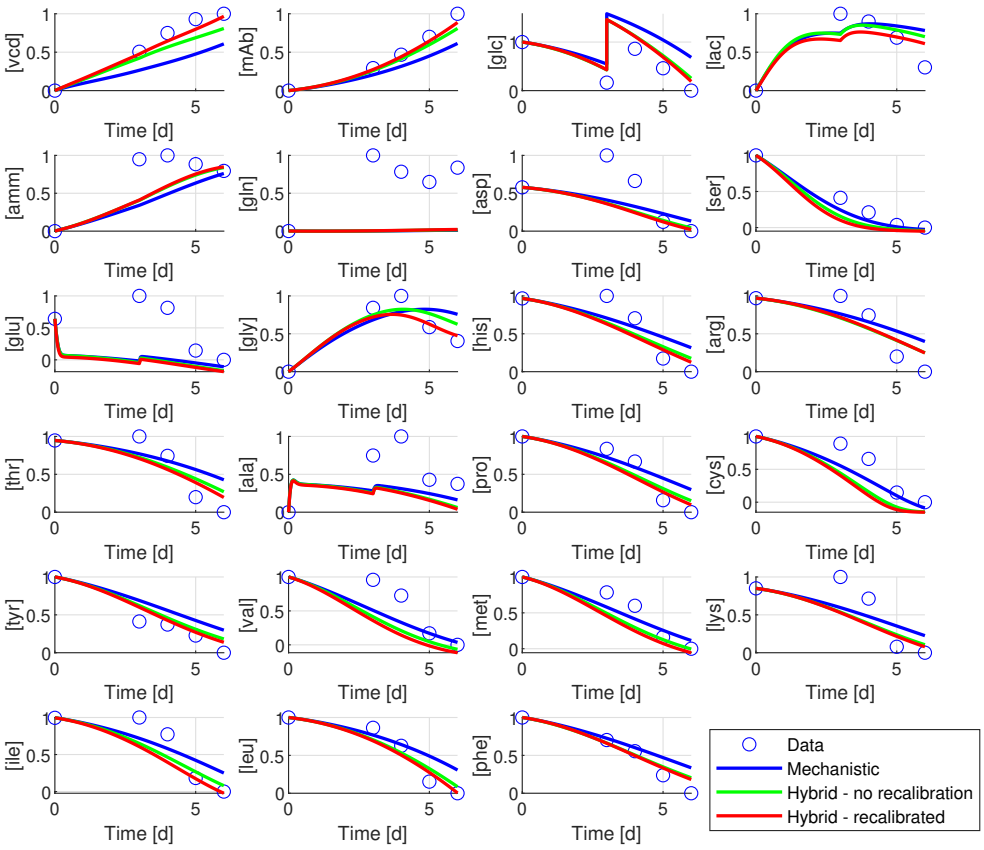


Figure F.9: Data, Mechanistic Model and Hybrid Model profiles for Medium 56.

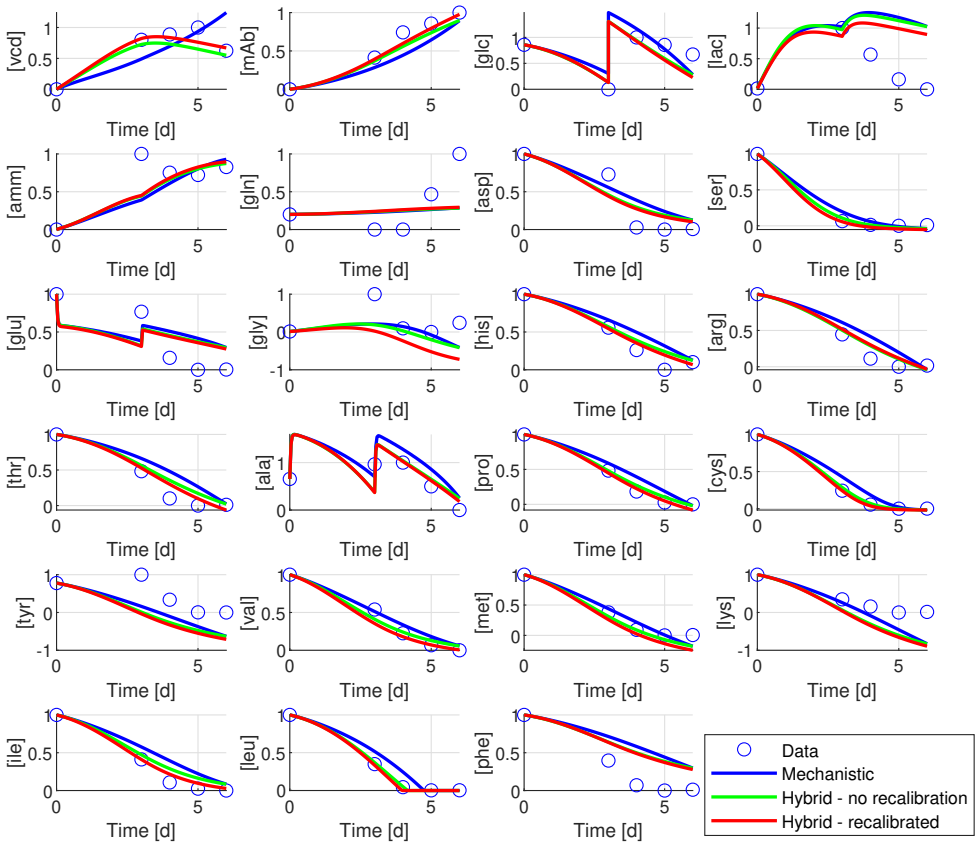


Figure F.10: Data, Mechanistic Model and Hybrid Model profiles for Medium 64.

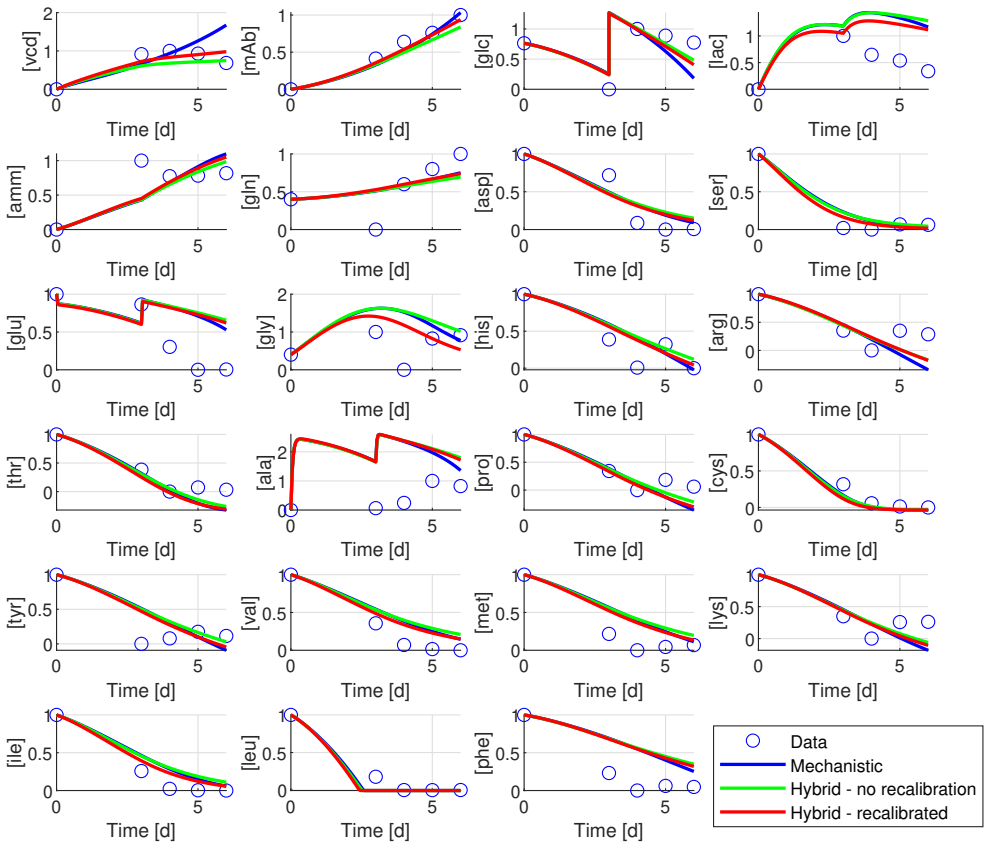


Figure F.11: Data, Mechanistic Model and Hybrid Model profiles for Medium 65.

Appendix G

Example of Multiplicity in LP Problem

Suppose you have the following LP problem:

$$\begin{aligned} \min_{j} \quad & -j_3 \\ \text{s.t.} \quad & j_3 \leq 1 \\ & j_1 + \frac{1}{3} j_2 - j_3 \leq 1 \\ & 2j_1 + j_2 - 2j_3 \leq 3 \end{aligned} \tag{G.1}$$

Multiple solutions can be find for this LP problem, since the first constraint is parallel to the objective function.

For instance, one possible solution is $j_1 = 1$, $j_2 = 3$ and $j_3 = 1$. Notice that for this solution all constraints are active, and the residuals presented by these constraints are equal to zero.

Now, observe the Lagrange multipliers obtained at this solution. The Lagrangian is obtained as follows:

$$\mathcal{L} = -j_3 + \lambda_1 (j_3 - 1) + \lambda_2 \left(j_1 + \frac{1}{3} j_2 - j_3 - 1 \right) + \lambda_3 (2j_1 + j_2 - 2j_3 - 3) \tag{G.2}$$

At stationary conditions, the partial Lagrangian derivatives for each constraint are as follows:

$$\frac{\partial \mathcal{L}}{\partial j_1} = \lambda_2 + 2\lambda_3 = 0 \tag{G.3}$$

$$\frac{\partial \mathcal{L}}{\partial j_2} = \frac{1}{3}\lambda_2 + \lambda_3 = 0 \tag{G.4}$$

$$\frac{\partial \mathcal{L}}{\partial j_3} = -1 + \lambda_1 - \lambda_2 - 2\lambda_3 = 0 \tag{G.5}$$

Observe that when Lagrange multipliers are equal to zero they are considered inactive, while different from zero are considered active. Notice that the solution of equations G.3, and give $\lambda_1 = 1$, $\lambda_2 = 0$, and $\lambda_3 = 0$, e.g. only the first constraint is found active by Lagrange multipliers approach. Remember that it was proved before that at this solution, all constraints are active.

The reason for $\lambda_2 = 0$ and $\lambda_3 = 0$ when even their respective constraints are in reality active is that the partial Lagrange derivatives $\frac{\partial \mathcal{L}}{\partial j_1}$ and $\frac{\partial \mathcal{L}}{\partial j_2}$ are not dependent of λ_1, which is directly associated to j_3. In conclusion, $\lambda_2 = 0$ and $\lambda_3 = 0$ are not find active by Lagrangian approach because j_1 and j_2 are not involved in the objective function.