Università
degli Studi
di Padova

Head Office: Università degli Studi di Padova
Department of Industrial Engineering

Ph.D. COURSE IN: Industrial Engineering
CURRICULUM: Chemical and Environmental Engineering
SERIES: XXXV

DIGITAL MODELS TO SUPPORT MONOCLONAL ANTIBODIES DEVELOPMENT IN THE BIOPHARMACEUTICAL INDUSTRY 4.0

Coordinator: Prof. Giulio Rosati
Supervisor: Prof. Pierantonio Facco

Ph.D. student: Gianmarco Barberi

Foreword

The fulfillment of the research results included in this Dissertation involved the intellectual and financial support of many people and institutions, to whom the author is very grateful.

Most of the research activity that led to the results reported in this Dissertation has been carried out at CAPE-Lab, Computer-Aided Process Engineering Laboratory, at the Department of Industrial Engineering of the University of Padova (Italy), under the supervision of Prof. Pierantonio Facco, with the help of Prof. Massimialino Barolo and Prof. Fabrizio Bezzo. Part of the work has been carried out at Imperial College London (UK) during a 6-month stay under the supervision of Prof. Cleo Kontoravdi. Part of the work resulted from a collaboration with Mr. Antonio Benedetti, Mrs. Paloma Diaz-Fernandez and Mr. Gary Finka form GlaxoSmithKline Process Engineering & Analytics and Biopharm Process Research – Stevenage (UK).

Financial support has been provided by the University of Padova and by *Fondazione Ing. Aldo Gini*, Padova (Italy).

DISCLOSURE STATEMENT

All the material reported in this Dissertation is original, unless an explicit indication is given in this statement or explicit references to studies carried out by other peoples are indicated. The work contained in Chapter 7 "Next-FLUX: Neural-net extracellular trained flux" is the results of the collaboration with the research group of Prof. Cleo Kontoravdi at Imperial College London (U.K.). Specifically, Mr. Gianmarco Barberi developed and deployed the machine learning model presented in Chapter 7. Mr. Benjamin Strain and Mr. James Morrisey developed and applied the strategy to constrain metabolic models with the outcomes of the machine learning model, which is briefly presented in Chapter 7.

In the following, a list of publications stemmed from this project is reported.

CONTRIBUTIONS IN INTERNATIONAL JOURNALS

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

Botton, A., Barberi, G., Facco, P. (2022). Data augmentation to support biopharmaceutical process development through digital models – A proof of concept. *Processes*, **10**(9), 1796.

CONTRIBUTIONS SUBMITTED TO INTERNATIONAL JOURNALS

Barberi, G., Benedetti, A., Diaz-Fernandez, P., Sévin, D. C., Vappiani, J., Finka, G., Bezzo, F., Facco, P. (2022). Identification of CHO platform metabolic traits for the selection of productive cell lines in biopharmaceutical process development through metabolomic dynamic data-driven modeling. Submitted to *Metabolic Engineering*

CONTRIBUTIONS IN INTERNATIONAL JOURNALS (in preparation)

Barberi, G., Giacopuzzi, C., Facco, P. Bioprocess feeding optimization through in silico dynamic experiments and hybrid digital models - A proof of concept.

CONTRIBUTIONS IN PEER-REVIEWD CONFERENCE PROCEEDINGS

Barberi, G., Benedetti, A., Diaz-Fernandez, P., Finka, G., Bezzo, F., Barolo, M., Facco, P. (2021). Anticipated cell lines selection in bioprocess scale-up through machine learning on metabolomics dynamics. *IFAC-PapersOnLine*, **54**(3), 85-90.

CONFERENCE PRESENTATIONS

Barberi, G., Benedetti, A., Diaz-Fernandez, P., Finka, G., Bezzo, F., Barolo, M., Facco, P. (2021). Anticipated cell lines selection in bioprocess scale-up through machine learning on metabolomics dynamics. Online presentation at: *16th IFAC Symposium on Advanced Control of Chemical Processes - ADCHEM 2021, June 13-16 2021* (Venice, Italy).

Barberi, G., Benedetti, A., Diaz-Fernandez, P., Sévin, D. C., Vappiani, J., Finka, G., Bezzo, F., Facco, P. (2022). Accelerating cell lines selection in biopharmaceutical process development through machine learning on process and metabolomic dynamics information. Oral presentation at: *36th International Forum Process Analytical Technology, IFPAC, June 12-15 2022* (North Bethesda, MA – USA).

Barberi, G., Benedetti, A., Diaz-Fernandez, P., Finka, G., Bezzo, F., Barolo, M., Facco, P. (2022). Accelerating cell lines selection in biopharmaceutical process development through machine learning on process and metabolomic dynamics information. Oral presentation at: *Convegno nazionale GRICU 2022, July 3-6 2022* (Ischia, Italy).

Abstract

In the last decade, the biopharmaceutical industry has been expanding in an impressive fashion, and more and more biopharmaceuticals have reached approval every year. More than one half of the total yearly approvals of biopharmaceuticals is represented by monoclonal antibodies (mAbs), which are an important class of therapeutics used for the treatment of autoimmune, oncological, and infectious diseases. Monoclonal antibodies are typically produced by genetically modified mammalian cells (specifically Chinese Hamster Ovary cells), which are cultivated in large bioreactors. The development of new mAbs is a resource-intensive and time-consuming procedure, typically lasting several years and costing more than 2 billion dollars each. The main steps of mAb development are cell line generation and engineering, cell line selection and scale-up, process characterization, and process optimization. Due to the long timelines and investments required for the development of new mAbs, biopharmaceutical companies are looking for innovative solutions to support and accelerate each of those steps of the drug development.

The objective of this Dissertation is to develop digital models to support and accelerate the monoclonal antibody development favoring the transition to the Biopharmaceutical Industry 4.0. This Dissertation concerns descriptive and diagnostic models, which provide a better comprehension of the biopharmaceutical processes and their behavior, but also predictive and prescriptive models, which allow to forecast and even improve the performance of the biopharmaceutical process. Specifically, in this Dissertation: *i*) cell line selection is accelerated by integrating process and biological information and examining their dynamics; *ii*) the identification of high performing cell lines in scenarios with limited available data is improved through *in silico* data generation; *iii*) the optimization of the culture feeding schedule is accelerated by means of hybrid models; *iv*) a new constraining method based on deep learning is proposed to improve of the metabolic description of genome-scale metabolic models (GSMMs); and *v*) an novel method to identify genetic engineering targets exploiting GSMMs is developed by means of latent-variables regression model inversion.

Furthermore, the models proposed in this Dissertation fulfill some of the regulatory requirements for the development of new drugs, such as providing enhanced process understanding, managing process variability, reducing the risk of poor-quality product, and predicting critical quality attributes (CQAs).

Two works deal with the **acceleration of cell lines selection by integrating process and biological information and examining their dynamics**. Since a workflow on the fusion of metabolomics and process data for a broad understanding of the relationship between cell metabolism and process CQAs, and the exploitation of metabolomic data dynamics to accelerate cell line selection are still missing, in the first work, an innovative machine learning approach is proposed to integrate time-varying process and biological information (i.e., metabolomics), explicitly accounting for their dynamics. The proposed framework is aimed at understanding the metabolic state changes occurring along the cultivation process, and how they are associated with process performance. Furthermore, product titer is estimated with good accuracy ($Q^2 > 40\%$), providing insights into its relationship with underlying metabolic mechanisms and enabling the identification of biomarkers to be further investigated, such as *propinol adenylate* and *L-lactic acid*. The biological insight obtained through the proposed approach provides an in-depth metabolic understanding of the process and allows the early identification of high performing cell lines.

In the second work, a machine learning methodology based on multivariate linear classification, explicitly exploiting dynamic metabolomic data, is proposed to accelerate the selection of high productive cell lines. Specifically, the information contained in the dynamic biological information allows to identify the cell lines with high productivity with 100% accuracy, already from the early stages of the culture. Moreover, this allows identifying the biomarkers that are most related to high cell productivity, such as *Citric acid*, *Thiamine*, and *UDP-glucose*, and to study how the relevant metabolic pathways for the discrimination of cell productivity vary along the cultivation. In the exponential growth and stationary phases, the metabolic pathways connected to energy production and DNA replication are found to be important for cell productivity, while in the decline phase the cell physiological state is totally connected to the metabolism of nucleotide and other sugars. Such biological understanding provides at the same time insight for the improvement of the host cells. The methodologies developed in these works were implemented in a software named ADAM, which is internally used by GlaxoSmithKline for the analysis of metabolomic data.

Concerning the **improved identification of high performing cell lines in scenarios with limited available data**, biopharmaceutical process development is typically characterized by the availability of few experiments, especially at large process scales, such as the pilot one, because of their high cost and long duration. This limits the use of science-based methods, such as multivariate statistical techniques, which demonstrated to be extremely beneficial to support various stages of process development. Data augmentation strategies are a viable solution to artificially increase the quantity of available data from experiments. However, they are underexplored in the biopharmaceutical sector. In this work, an innovative data augmentation methodology for *in silico* data generation is proposed to augment the amount of data available from real (i.e., *in vitro*) experiments. *In silico* data generated by two digital models (one based on a first principles model, the other on a hybrid model) are used to improve the identification of high performing cell lines by means of multivariate models in a simulated biopharmaceutical process for the production of mAbs. The simulated process allows better control of both the process behavior and the biological diversity in the experiments. The generation of *in silico* data through digital models effectively support the identification of high-productive cell lines (i.e., high mAb titer) even when a very low number of real experimental batches (< 6) is available, by predicting the mAb titer with errors that are comparable with the experimental one (180-220 mg/L). This allows to reduce expenses and timelines of mAbs development, and a more effective identification of the process variables with the largest influence on mAb titer.

Regarding the **accelerated optimization of the feeding schedule by means of hybrid models**, hybrid models proved to be effective for the optimization of the feeding schedule, but a proof of their advantages over the state-of-the-art experimental strategies is missing. In this work, a novel methodology for *in silico* experimental campaign through hybrid models is proposed. In particular, the *in silico* experimental campaign on hybrid digital models, trained with experiments planned through a Design of Dynamic Experiments is compared with two *in vitro* experimental campaigns with different numbers of planned experiments. The *in silico* experimental campaign identifies better process optimum (*in silico*: 3222.8 mg/L vs. *in vitro*: 3136.3 mg/L), and reduces the number of experiments required to identify the best feeding schedule. The proposed methodology is tested on a simulated biopharmaceutical process for the production of mAbs; the simulated process allows to know if the *in silico* experimentation captures in a correct way the real relationship between nutrients and antibody titer, and if it is able to identify the real optimal feeding strategy.

With respect to **the new constraining method based on deep learning to improve of the metabolic description of GSMMs**, the complexity of mammalian cell metabolic networks limits the accuracy in representing the metabolic state and phenotype of cells. The introduction of better methods for constraining the intracellular flux values in GSMMs could provide accurate modeling of cell metabolism, but methodologies are either too expensive (i.e., ^{13}C labeling experiments) or still not enough accurate (i.e., FBA, pFBA, ccFBA). Furthermore, a reliable, accurate, and cheap method based on experimental data to properly define the intracellular constrains of GSMMs is missing. In this work, a deep learning method, named Next-FLUX, is proposed to estimate the constraints of GSMMs from cheap and easily available measurements. Next-FLUX accurately predicts most of the intracellular fluxes ($Q^2 > 65\%$). Furthermore, the constraints to apply in GSMMs are estimated in an innovative way by means of the neural network prediction intervals. The estimated constraints allow a GSMM to improve the calculation of the intracellular fluxes (Person correlation between GSMM calculated and experimental intracellular fluxes: 0.765 vs. 0.343) and the biomass prediction ($Q^2 = 61.6\%$ vs.

$Q^2 = 13.3\%$) with respect to the state-of-the-art methodologies. Accordingly, the more accurate predictions of both metabolic state and phenotypes of cells lead to a better description of cell metabolism.

Regarding the **novel method to identify genetic engineering targets by means of GSMMs and latent variable regression model inversion**, targets for genetic engineering are typically identified by mean of GSMMs and optimization methods, which are complex to implement, computationally demanding and time-consuming, especially for large metabolic networks. The application these algorithms to mammalian cells, such as CHO, is limited by the large size of the mammalian metabolic network. Furthermore, alternative methods based on simple data-based mathematical methodologies are missing. In this work, a method based on latent variable regression model inversion is proposed to identify genetic modifications that improve the mAb productivity of mammalian cells. The proposed methodology suggested genetic modifications concerning the metabolism of specific amino acids, such as *L-valine* and *L-tryptophan*, and the recirculation of *mannose* in the early stages of mAb glycosylation to improve cell productivity. The mAb productivity improvements are verified in synthetic experiments run on the GSMMs, which is the first step towards the *in vitro* testing. Furthermore, the proposed methodology is faster and simpler than traditional methods to identify targets for genetic engineering through GSMMs which are based on optimization algorithms.

Riassunto esteso

Negli ultimi dieci anni, l'industria biofarmaceutica è cresciuta notevolmente e sempre più biofarmaci vengono approvati ogni anno. Più di metà delle approvazioni annuali è costituita da anticorpi monoclonali (mAbs), una classe di medicinali utilizzati per il trattamento di malattie autoimmuni, oncologiche e infettive. A livello industriale, gli anticorpi monoclonali sono tipicamente prodotti da cellule di mammifero geneticamente modificate (in particolare cellule ovariche di criceto cinese, CHO) coltivate in grandi bioreattori. La fase di sviluppo di nuovi anticorpi monoclonali è una procedura dalla lunga durata ed elevato costo, che in generale può durare diversi anni e costare più di 2 miliardi di dollari per ogni nuovo farmaco. Le fasi principali dello sviluppo di nuovi anticorpi monoclonali sono la generazione e l'ingegnerizzazione delle linee cellulari, la selezione delle linee cellulari più promettenti, e la caratterizzazione e l'ottimizzazione del processo produttivo. A causa delle lunghe tempistiche per lo sviluppo di nuovi anticorpi monoclonali e degli ingenti investimenti richiesti, le aziende biofarmaceutiche sono alla ricerca di soluzioni innovative per adiuvare ed accelerare lo sviluppo di nuovi farmaci.

L'obiettivo di questa Dissertazione è lo sviluppo di modelli digitali per adiuvare and accelerare le varie fasi dello sviluppo di nuovi anticorpi monoclonali, in modo da favorire la transizione verso l'Industria Biofarmaceutica 4.0. Questa Dissertazione tratta di modelli descrittivi e diagnostici, che consentono di raggiungere una migliore comprensione del processo produttivo e delle sue prestazioni, ma anche di modelli predittivi e prescrittivi, che permettono di prevedere e perfino di migliorare le prestazioni dei processi biofarmaceutici. In particolare, in questa Dissertazione: *i*) la selezione di linee cellulari promettenti è stata accelerata attraverso la fusione di informazioni biologiche e di processo, anche sfruttando la loro dinamica; *ii*) l'identificazione di linee cellulari ad elevate prestazioni in scenari con disponibilità limitata di dati è stata migliorata attraverso la generazione di dati *in silico*; *iii*) l'ottimizzazione della strategia di alimentazione delle colture cellulari è stata accelerata attraverso modelli ibridi; *iv*) un nuovo metodo per vincolare i flussi metabolici basato su tecniche di *deep learning* è stato proposto per migliorare la capacità descrittiva del metabolismo cellulare da parte di modelli metabolici a scala genomica (GSMM); e *v*) un nuovo metodo per identificare modifiche genetiche sfruttando GSMM è stato sviluppato attraverso l'inversione di modelli di regressione a variabili latenti.

Inoltre, i modelli proposti in questa Dissertazione soddisfano alcune richieste degli enti regolatori riguardo lo sviluppo di nuovi farmaci, come una profonda conoscenza del processo,

la gestione della sua variabilità, la riduzione del rischio di ottenere prodotti di scarsa qualità e la predizione di attributi di qualità critici (CQA).

L'**accelerazione della fase di selezione delle linee cellulari più promettenti attraverso l'integrazione di informazioni di processo e biologiche sfruttando le loro variazioni temporali** è stata considerata in due lavori, contenuti nei Capitoli 3 e 4. In questo caso, un approccio che combini informazioni biologiche e di processo per una migliore comprensione della relazione tra il metabolismo cellulare e i CQA, e l'utilizzo di dati di metabolomica variabili nel tempo per velocizzare la selezione di linee cellulari promettenti sono ancora mancanti. Nel primo lavoro, viene proposto un approccio innovativo basato sul *machine learning* che combini informazioni biologiche (come dati di metabolomica) e di processo variabili nel tempo e sfrutti in modo esplicito la loro dinamica. La metodologia proposta viene utilizzata per studiare come il metabolismo cellulare cambia durante il processo di coltivazione e come questi cambiamenti siano associati alle prestazioni del processo produttivo. Inoltre, il titolo di anticorpi viene stimato con buona accuratezza ($Q^2 > 40\%$), permettendo di ottenere preziose informazioni sui meccanismi metabolici associati alla produzione di anticorpi e biomarcatori, come *propinol adenylate* e *L-lactic acid*, che dovranno essere analizzati approfonditamente in ulteriori studi. Tutte le conoscenze sui fenomeni biologici ottenute attraverso la metodologia proposta consentendo l'identificazione precoce delle linee cellulari con elevate prestazioni.

Nel secondo lavoro, viene proposta una metodologia basata sul *machine learning* che utilizza in modo esplicito informazioni biologiche variabili nel tempo per accelerare l'identificazione di linee cellulari che presentano una elevata produttività. In particolare, le informazioni biologiche variabili nel tempo permettono di identificare già dagli stadi iniziali del processo le linee cellulari che presentano una elevata produttività con un'accuratezza del 100%. Inoltre, la metodologia proposta permette di identificare i biomarcatori associati ad un'elevata produttività cellulare, come *Citric acid*, *Thiamine* e *UDP-glucose*, e di studiare come i percorsi metabolici legati alla produttività cellulare cambiano nel tempo lungo la coltura cellulare. Nella fase di crescita esponenziale e in quella stazionaria, i percorsi metabolici connessi alla produttività riguardano la produzione di energia e la replicazione del DNA, mentre nella fase di declino, i percorsi metabolici connessi alla produttività sono principalmente legati al metabolismo degli zuccheri e degli zuccheri nucleici. La conoscenza della relazione tra il metabolismo cellulare ed un'elevata produttività cellulare fornisce importanti informazioni per migliorare le caratteristiche delle cellule utilizzate per la produzione di anticorpi monoclonali. Le metodologie sviluppate in questi lavori sono state implementate in un software chiamato ADAM, il quale viene utilizzato internamente da GlaxoSmithKline per l'analisi di dati metabolomici.

Per quanto riguarda la **migliore identificazione di linee cellulari più performanti in situazioni con limitata disponibilità di dati**, lo sviluppo di processi biofarmaceutici è tipicamente caratterizzato dalla limitata disponibilità di esperimenti, specialmente alle scale di processo più grandi come quella pilota, a causa dell'elevato costo e lunga durata di ogni esperimento. Questo limita notevolmente l'uso di tecniche scientifiche, come quelle statistiche multivariate, le quali hanno dimostrato di essere particolarmente utili per lo sviluppo di processi biofarmaceutici. Strategie di *data augmentation* sono una soluzione per aumentare il numero di dati disponibili dagli esperimenti, ma purtroppo sono poco considerate nel settore biofarmaceutico. In questo lavoro viene proposta una metodologia innovativa per la generazione di dati *in silico* in modo da aumentare la quantità dei dati disponibili da esperimenti reali, che viene riportata nel Capitolo 5. I dati sono generati *in silico* attraverso due diversi modelli digitali (uno basato su un modello a principi primi e l'altro basato su un modello ibrido) e vengono utilizzati per migliorare l'identificazione delle linee cellulari con le migliori prestazioni da parte di modelli multivariati in un processo simulato di produzione di anticorpi monoclonali. Questo processo simulato permette un migliore controllo del comportamento del processo e della diversità biologica tra i vari esperimenti. Utilizzo di dati generati *in silico* attraverso modelli digitali permette di identificare efficacemente linee cellulari molto produttive e che raggiungono un titolo di anticorpi elevato, anche quando un numero molto ridotto di esperimenti reali è disponibile per l'analisi (< 6), in quanto permettono di predire il titolo di anticorpi con errori comparabili a quelli sperimentali (180-220 mg/L). Tutto ciò permette di ridurre i costi e le tempistiche di sviluppo degli anticorpi monoclonali assieme ad una identificazione più efficace delle variabili di processo che influenzano maggiormente il titolo di anticorpi.

Per quanto riguarda **l'accelerazione della strategia di alimentazione delle colture cellulari attraverso modelli ibridi**, i modelli ibridi hanno dimostrato la loro efficacia nell'ottimizzare la strategia di alimentazione, ma i vantaggi del loro utilizzo rispetto a strategie sperimentali non sono ancora stati provati. In questo lavoro viene proposta una metodologia innovativa per l'esecuzione di una campagna sperimentale *in silico* utilizzando dei modelli digitali ibridi, che viene riportata nel Capitolo 6. In particolare, la campagna sperimentale *in silico* basata su modelli ibridi addestrati a partire da esperimenti pianificati in modo dinamico è stata confrontata con due campagne sperimentali *in vitro* che comprendono un diverso numero di esperimenti. La campagna sperimentale *in silico* identifica un punto di ottimo del processo migliore delle campagne puramente sperimentali (*in silico*: 3222.8 mg/L contro *in vitro*: 3136.3 mg/L), e permette di ridurre il numero di esperimenti richiesti per identificare la migliore strategia di alimentazione per le colture cellulari. La metodologia proposta è stata testata su un processo biofarmaceutico simulato per la produzione di anticorpi monoclonali. Questo processo simulato permette di conoscere se la campagna sperimentale *in silico* cattura in modo corretto la relazione tra nutrienti e titolo di anticorpi e se è anche in grado di identificare il punto di ottimo reale del processo.

Per quanto riguarda il **nuovo metodo per vincolare i flussi metabolici basato su tecniche di *deep learning* che migliora la capacità descrittiva del metabolismo cellulare da parte di GSMM**, la complessità del network metabolico delle cellule di mammifero limita l'accuratezza con cui lo stato metabolico e il fenotipo delle cellule viene descritto. L'introduzione di ulteriori metodi per vincolare il valore accettabile dei flussi metabolici in GSMM può migliorare sensibilmente l'accuratezza con cui il metabolismo cellulare viene descritto. Purtroppo, le metodologie disponibili sono troppo costose (come gli esperimenti basati sugli isotopi ^{13}C) o non sufficientemente accurate (come FBA, pFBA, ccFBA). Inoltre, un metodo affidabile, accurato ed economico basato su dati sperimentali per definire i vincoli sui valori accettabili dei flussi metabolici in GSMM non è al momento disponibile. In questo lavoro è stato sviluppato un modello di *deep learning* basato su reti neurali denominato Next-FLUX, riportato nel Capitolo 7. Questo modello stima i valori ammissibili dei flussi metabolici (ovvero vincoli necessari ai modelli metabolici per essere risolti) a partire da dati economici e facilmente ottenibili. Next-FLUX è in grado di predire con grande accuratezza i flussi metabolici intracellulari ($Q^2 > 65\%$). Inoltre, i vincoli sui valori accettabili dei flussi metabolici da applicare nei GSMM sono stimati in modo innovativo utilizzando gli intervalli di predizione delle reti neurali. I vincoli stimati in questo modo permettono ad un GSMM di calcolare in modo più accurato i flussi metabolici (con una correlazione Pearson tra i flussi calcolati dal GSMM e quelli sperimentali di 0.765 contro 0.343) e la biomassa prodotta ($Q^2 = 61.6\%$ contro $Q^2 = 13.3\%$) rispetto alle metodologie più all'avanguardia. Tutto ciò ha permesso di migliorare la capacità dei modelli metabolici di descrivere il metabolismo cellulare.

Per quanto riguarda il **nuovo metodo per indentificare possibili modifiche genetiche attraverso modelli metabolici e l'inversione di modelli regressivi a variabili latenti**, l'identificazione di possibili modifiche genetiche viene tipicamente eseguita attraverso GSMM e tecniche di ottimizzazione, le quali risultano complesse da implementare e richiedono elevate risorse computazionali e molto tempo, specialmente nel caso di grandi *network* metabolici. L'applicazione di questi algoritmi di ottimizzazione a cellule di mammifero, come ad esempio le CHO, è fortemente limitata dalle elevate dimensioni dei *network* metabolici delle cellule di mammifero. Inoltre, metodi alternativi per identificare modifiche genetiche che si basano su metodologie basate su dati non sono al momento disponibili. In questo lavoro è stato sviluppato un metodo basato sull'inversione di modelli regressivi a variabili latenti per identificare modifiche genetiche che migliorino la produttività di anticorpi monoclonali da parte di cellule di mammifero. La metodologia proposta ha suggerito delle modifiche genetiche riguardanti il metabolismo di specifici aminoacidi, come *L-valine* e *L-tryptophan*, e il ricircolo di *mannose* negli stadi iniziali della glicosilazione degli anticorpi. Gli aumenti di produttività sono stati

verificati attraverso esperimenti sintetici su GSMM, i quali sono il primo passo prima dei test *in vitro*. Inoltre, la metodologia proposta risulta più semplice e veloce rispetto alle metodiche tipicamente utilizzate che si basano sull'utilizzo di modelli metabolici e algoritmi di ottimizzazione.

Table of Contents

FOREWORD i

ABSTRACT iii

RIASSUNTO ESTESO vii

TABLE OF CONTENT xiii

LIST OF SYMBOLS xxi

CHAPTER 1 – STATE OF THE ART AND MOTIVATION 1

1.1 BIOPHARMACEUTICAL INDUSTRY: AN OVERVIEW 1

1.2 MONOCLONAL ANTIBODIES 3

1.3 CELL CULTURES 4

1.3.1 UPSTREAM PROCESS 4

1.3.2 DOWNSTREAM PROCESS 5

1.3.3 DATA IN CELL CULTURES 6

1.3.3.1 Process data 6

1.3.3.2 Biological data 7

1.4 BIOPRODUCT DEVELOPMENT 8

1.4.1 PRODUCT DEVELOPMENT 9

1.4.1 PROCESS DEVELOPMENT 10

1.4.1.1 Cell generation and engineering 11

1.4.1.2 Cell lines selection and scale-up 12

1.4.1.3 Process characterization 13

1.4.1.4 Media and feed optimization 14

1.4.1.5 Process optimization 15

1.4.2 REGULATORY ASPECTS OF BIOPROCESS DEVELOPMENT 15

1.5 MATHEMATICAL MODELING IN BIOPHARMA INDUSTRY 4.0 17

1.5.1 DATA-DRIVEN MODELLING 19

1.5.1.1 Descriptive activities 19

1.5.1.2 Diagnostic activities 23

1.5.1.3 Predictive activities 25
1.5.2 HYBRID MODELING 27
1.5.2.1 Descriptive activities 30
1.5.2.2 Diagnostic activities 31
1.5.2.3 Predictive activities 31
1.5.2.4 Prescriptive activities 32
1.5.3 FIRST PRINCIPLES MODELLING OF CELL METABOLISM 34
1.5.3.1 Descriptive activities 37
1.5.3.2 Predictive activities 40
1.5.3.3 Prescriptive activities 41
1.6 OBJECTIVE OF THE RESEARCH 43
1.7 DISSERTATION ROADMAP 46
CHAPTER 2 – MATHEMATICAL METHODOLOGIES 51
2.1 MULTIVARIATE MODELLING 51
2.1.1 PRINCIPAL COMPONENT ANALYSIS 51
2.1.1.2 Multiblock PCA 53
2.1.2 PARTIAL LEAST-SQUARES REGRESSION 53
2.1.2.1 PLS Discriminant Analysis 55
2.1.3 MULTIWAY MODELING 55
2.1.4 EVOLVING MODELING 56
2.2.5 INVERSION OF LATENT VARIABLE MODELS 56
2.2 NEURAL NETWORKS 58
2.2.1 TRAINING OF THE NETWORK 59
2.3 GENOME-SCALE METABOLIC MODELS 61
2.3.1 FLUX BALANCE ANALYSIS 62
2.3.2 FLUX SAMPLING 63
CHAPTER 3 – INTEGRATING METABOLOME DYNAMICS AND PROCESS DATA TO GUIDE CELL LINE SELECTION 65
3.1 INTRODUCTION 65
3.2 MATERIALS AND METHODS 67
3.2.1 AVAILABLE DATA 67
3.2.2 METHODOLOGY TO INTEGRATE PROCESS AND BIOLOGY 68

3.2.3 Multivariate statistical analysis 69

3.3 Results and Discussion 72

3.3.1 Process understanding 72

3.3.1.1 Metabolic mapping of cell lines according to process performance 72

3.3.1.2 Time course similarity in biological phenomena 75

3.3.2 Time course changes in process and biology 77

3.3.3 QAs and cell performance estimation and biological phenomena identification 80

3.4 Conclusions 84

CHAPTER 4 – METABOLIC TRAITS FOR THE SELECTION OF PRODUCTIVE CELL LINES THROUGH METABOLOMIC DYNAMIC DATA-DRIVEN MODELING 87

4.1 Introduction 87

4.2 Materials and methods 90

4.2.1 Cell culture data 90

4.2.2 Multiway and multivariate data analysis 91

4.2.3 Important biomarker and biological function identification 93

4.2.3.1 Biomarker identification 93

4.2.3.2 Metabolic pathway identification 94

4.3 Results and Discussion 95

4.3.1 Discrimination analysis to study the correlation between cell productivity and metabolomic profiles dynamics 95

4.3.2 Anticipated discrimination of cell productivity 97

4.3.3 Identification of cell productivity biomarkers 99

4.3.4 Identification of important biological functions in the discrimination of cell productivity 103

4.5 ADAM: Application for the Digital Analysis of Metabolites 107

4.4 Conclusions 107

CHAPTER 5 – DATA AUGMENTATION TO SUPPORT BIOPHARMACEUTICAL PROCESS DEVELOPMENT THROUGH DIGITAL MODELS 109

5.1 Introduction 109

5.2 Materials and Methods 111

5.2.1 Methodological procedure 111

5.2.2 PROCESS FOR THE PRODUCTION OF MONOCLONAL ANTIBODIES 113
5.2.3 MODELING STRATEGY 1: FIRST PRINCIPLES DIGITAL MODEL 114
5.2.3.1 *In silico* batch generation through First Principles Digital Model 115
5.2.4 MODELING STRATEGY 2: HYBRID DIGITAL MODEL 115
5.2.4.1 *In silico* batch generation through Hybrid Digital Model 118
5.2.5 MULTIVARIATE PREDICTIVE MODELING 119
5.3 RESULTS AND DISCUSSION 121
5.3.1 MONOCLONAL ANTIBODIES TITER ESTIMATION 121
5.3.1.1 Titer estimation performance and sensitivity to the available number of process calibration batches 121
5.3.1.2 Effect of data augmentation on the estimation performance 122
5.3.2 PROCESS UNDERSTANDING FOR MABS TITER ESTIMATION 124
5.3.2.1 Process understanding with process batches only 124
5.3.2.2 Process understanding supported by FPDM *in silico* data augmentation 126
5.3.2.3 Process understanding supported by HDM *in silico* data augmentation 127
5.4 CONCLUSIONS 128

CHAPTER 6 – BIOPROCESS FEEDING OPTIMIZATION THROUGH IN SILICO DYNAMIC EXPERIMENTS AND HYBRID DIGITAL MODELS 131
6.1 INTRODUCTION 131
6.2 MATERIALS AND METHODS 134
6.2.1 PROPOSED METHODOLOGY 134
6.2.2 DESIGN OF DYNAMIC EXPERIMENTS 135
6.2.3 PROCESS FOR THE PRODUCTION OF MONOCLONAL ANTIBODIES AT 1-L SHAKE FLASKS SCALE 139
6.2.4 HYBRID MODEL 140
6.2.5 FEEDING OPTIMIZATION 142
6.3 RESULTS 143
6.3.1 EXPERIMENTAL CAMPAIGN A FOR FEEDING SCHEDULE OPTIMIZATION 144
6.3.2 EXPERIMENTAL CAMPAIGN B FOR FEEDING SCHEDULE OPTIMIZATION 146
6.3.3 IN SILICO CELL CULTURE OPTIMIZATION THOUGH HYBRID MODEL 148
6.3.4 REAL OPTIMAL FEEDING SCHEDULE 149
6.4 DISCUSSION 150
6.4.1 THE OPTIMAL FEEDING SCHEDULE 150

6.4.2 COMPARISON OF THE OPTIMAL FEEDING SCHEDULE 151
6.4.3 THE BEST OPTIMIZATION STRATEGY 151
6.5 CONCLUSIONS 156

CHAPTER 7 – NEXT-FLUX: NEURAL-NET EXTRACELLULAR TRAINED FLUX 157

7.1. MATERIALS AND METHODS 157
7.1.1 ^{13}C ISOTOPE LABELING DATASET 157
7.1.2 PREDICTION OF INTRACELLULAR FLUXES FROM EXTRACELLULAR METABOLITE UPTAKE RATES THROUGH ARTIFICIAL NEURAL NETWORKS 158
7.1.2.1 Data management 158
7.1.2.2 Data-driven modeling strategy 159
7.1.2.3 Ensemble prediction interval calculation method 161
7.1.2.4 Gradient prediction interval calculation method 162
7.1.3 GENOME-SCALE METABOLIC MODEL 162
7.1.3.1 Reaction mapping 163
7.1.3.2 Maximization of the feasible constraints 163
7.1.3.3 GSMM solution 163
7.1.3.4 Performance evaluation 163
7.4 PREDICTION OF INTRACELLULAR FLUXES 164
7.2.1 PERFORMANCE OF DIFFERENT PREDICTION INTERVAL CALCULATION METHODS 164
7.2.2 SENSITIVITY TO DIFFERENT NUMBER OF NETWORK NEURONS 170
7.3 GENOME-SCALE METABOLIC MODEL PREDICTIONS 172
7.4 DEPLOYMENT OF NEXT-FLUX 174
7.5 CONCLUDING REMARKS AND FUTURE WORK 175

CHAPTER 8 – DATA-DRIVEN GENETIC ENGINEERING 177

8.1. MATERIAL AND METHODS 177
8.1.1 AVAILABLE CULTURE DATA 177
8.1.2 PROPOSED STRATEGY 178
8.1.3 STRAIN SPECIFIC GENOME-SCALE METABOLIC MODEL AND METABOLIC DATA GENERATION 179
8.1.4 LATENT VARIABLES REGRESSION MODEL INVERSION 180
8.1.5 GENETIC MODIFICATIONS IDENTIFICATION 181
8.2. PREDICTION OF BIOMASS AND PRODUCTIVITY 186

8.3 IDENTIFICATION OF THE OPTIMAL FLUX DISTRIBUTION 188

8.4 GENETIC MODIFICATIONS 190

8.4.1 METABOLIC GENETIC MODIFICATIONS 190

8.4.1.1 Valine metabolism........ 190

8.4.1.2 Tryptophan metabolism........ 193

8.4.2 SECRETORY PATHWAY GENETIC MODIFICATIONS 197

8.5 CONCLUSIONS 205

CONCLUSIONS AND FUTURE PERSPECTIVES 207

FUTURE PERSPECTIVES 212

APPENDIX A – MONOCLONAL ANTIBODIES AND CELL CULTURES 215

A.1 MONOCLONAL ANTIBODIES 215

A.1.1 MONOCLONAL ANTIBODY STRUCTURE AND FUNCTION........ 215

A.1.2 PRODUCTION OF MONOCLONAL ANTIBODIES........ 217

A.2 CELL CULTURES........ 217

A.2.1 UPSTREAM PROCESS........ 217

A.2.1.1 Operating modes........ 217

A.2.1.2 Bioreactor types and operating parameters 218

A.2.1.3 Elements for cell survival........ 220

A.2.1.4 Upstream phases........ 221

A.2.2 DOWNSTREAM PROCESS........ 221

A.2.3 FUTURE MANUFACTURING TRENDS: CONTINUOUS PRODUCTION 222

APPENDIX B – ADDITIONAL DETAILS ON THE INTEGRATION OF METABOLOME DYNAMICS AND PROCESS DATA 223

B.1 MEASUREMENT REPLICATE UNFOLDING 223

B.2 MULTIWAY PRINCIPAL COMPONENT ANALYSIS 223

B.3 SIMILARITY ANALYSIS 224

B.4 MULTI-BLOCK PRINCIPAL COMPONENT ANALYSIS 225

B.5 MULTIWAY PARTIAL LEAST-SQUARES REGRESSION 225

B.6 VARIABLE SELECTION........ 225

B.7 ADDITIONAL INFORMATION ON CULTURE VARIABLE CORRELATION 227

APPENDIX C – MATHEMATICAL DETAILS OF FOR THE SELECTION OF HIGH PRODUCTIVE CELL LINES 229

C.1 DATA UNFOLDING 229

C.2 PLS-DA 230

C.3 E-MPLS-DA 230

C.4 VARIABLE SELECTION 231

C.5 VIP BOOTSTRAP 231

C.6 DISCRIMINATION OF HIGH PRODUCTIVE CELL LINES THROUGH EXTRACELLULAR METABOLOMICS DATA 232

C.7 ANTICIPATED DISCRIMINATION OF HIGH-PRODUCTIVE CELL LINES THROUGH EXTRACELLULAR METABOLOMICS DATA 233

APPENDIX D – ADDITIONAL DETAIL ON DATA AUGMENTATION 235

D.1 MODEL OF MAMMALIAN CELL CULTURED PRODUCING MABS 235

D.2 FIRST PRINCIPLES DIGITAL MODEL 238

D.3 PARAMETERS FOR *IN SILICO* DATA GENERATION BY HYBRID DIGITAL MODEL 239

APPENDIX E – EXPERIMENTAL CAMPAIGNS FOR FEEDING SCHEDULE OPTIMIZATION 241

E.1 EXPERIMENTAL CAMPAIGNS 241

APPENDIX F – ^{13}C INTRACELLULAR REACTIONS AND METABOLITES 243

F.1 INTRACELLULAR REACTIONS 243

F.2 METABOLITES 245

REFERENCES 247

ACKNOWLEDGMENTS 275

List of Symbols

Acronyms

ANN	=	artificial neural networks
ccFBA	=	carbon constraining flux balance analysis
CHO	=	Chinese Hamster Ovary cell
CMAs	=	critical material attributes
CPPs	=	critical process parameters
CQAs	=	critical quality attributes
DO	=	dissolved oxygen
DoDE	=	design of dynamic experiments
DoE	=	design of experiments
EMA	=	European Medicine Agency
E-PLS-DA	=	evolving partial least-squares discriminant analysis
FBA	=	flux balance analysis
FDA	=	U.S. Food and Drug Administration
GC-MS	=	gas chromatography - mass spectrometry
GlcNAc	=	N-acetylglucosamine
GPU	=	graphics processing unit
GSMM	=	genome-scale metabolic model
HEK	=	human embryonic kidney cell
ICH	=	International Council on Harmonization
Ig	=	immunoglobulin (i.e., monoclonal antibody)
LCL	=	lower confidence limits
LC-MS	=	liquid chromatography - mass spectrometry
LVs	=	latent variables
mAbs	=	monoclonal antibodies
MAPE	=	mean absolute prediction error
MB-MPCA	=	multi-block multiway principal component analysis
MB-PCA	=	multi-block principal component analysis
MFA	=	metabolic flux analysis
MPCA	=	multiway principal component analysis
MPLS	=	multiway partial least-squares
MSE	=	mean squared error

MV	=	multivariate methods
NIPALS	=	nonlinear iterative partial least-square algorithm
NOC	=	normal operating conditions
OPLS-DA	=	orthogonal partial least-squares discriminant analysis
PAT	=	process analytical technology
PCA	=	principal component analysis
PCs	=	principal components
pFBA	=	parsimonious enzyme usage flux balance analysis
PI	=	prediction interval
PLS	=	partial least-squares
PLS-DA	=	partial least-squares discriminant analysis
QAs	=	quality attributes
QbD	=	quality by design
QTTP	=	quality target product profile
RMSE	=	root mean squared error
RSM	=	response surface methodology
SPE	=	squared prediction error
SR	=	selectivity ratio
SVM	=	support vector machine
TCA	=	tricarboxylic acid
VCC	=	viable cell concentration
VIP	=	variable importance in projection index

Symbols

A	=	number of orthogonal principal components of latent variables
a_h	=	activation of the h-th neuron
a'_h	=	neuron's output prior the application of the activation function
B	=	number of classes
$\mathbf{B}$	=	matrix of PLS regression coefficients
$\mathbf{c}$	=	vector of concentrations for the culture variables
$\mathbf{c}^*$	=	reduced concentration vector
c_v	=	concentration of the variable v
$\hat{c}_v$	=	predicted concentration of the variable v
D	=	number of metabolites
D_V	=	culture dilution factor
E	=	number of effects in the RSM
$\mathbf{E}$	=	residual regressor matrix

$\mathbf{e}_n$	=	residual vector of the n-th observation
F_{in}	=	inlet flow rate of the bioreactor
F_{out}	=	outlet flow rate of the bioreactor
$\mathbf{F}$	=	residual response matrix
$\mathbf{f}_n$	=	response residual vector of the n-th observation
$f()$	=	activation function or genetic function
f_{lim}	=	limiting factor
$\mathbf{g}_{\mathrm{NEW}}$	=	gradient vector of the loss function calculated in the new datapoint
g_1, g_2, g_3	=	constants in PLS model inversion
g_{DCW}	=	dry cell weight in grams
H	=	number of neurons in a layer
$\mathbf{H}$	=	matrix of known kinetic expressions in hybrid models
K	=	total number of subfactors
K_{gln}	=	Monod constant for glutamine
K_{glc}	=	Monod constant for the growth on glucose
K_{miss}	=	number of metabolites used for missing data imputation
k_{reg}	=	scaling coefficient for reaction regulation
$\mathbf{J}$	=	matrix of the gradient vectors of the loss function calculated for all training experiments
j	=	index for nutrients in cell cultures
I	=	number of dynamic subfactors
$\mathbf{I}$	=	identity matrix
it	=	iteration
it_{max}	=	maximum number of iterations
L	=	loss function of the ANN
$\mathcal{L}$	=	ANN cost function
$\mathbf{L}$	=	diagonal matrix of the square root of the eigenvalues of PCA or PLS
M	=	number of response variables
m_{glc}	=	glucose maintenance constant
N	=	number of observations (i.e, samples or experiments)
N_{models}	=	number of models trained
O	=	number of new observations (i.e., samples, experiments, batches)
PI	=	half width prediction interval
P_{i-1}	=	shifted Legendre polynomial of degree $i-1$
$\mathbf{P}$	=	PCA or PLS loading matrix
p^*	=	factor for prediction interval calculation
Q^2	=	coefficient of determination for new samples
Q_{P}	=	specific mAb productivity

Q_{glc}	=	specific glucose consumption rate
$\mathbf{Q}$	=	response loading matrix of PLS
R	=	number of measurement replicates
R^2	=	coefficient of determination
R^2_{adj}	=	adjusted coefficient of determination
$\mathbf{r}$	=	vector of volumetric reaction rates
S	=	number of genetically modified reactions
$S_{t't''}$	=	similarity index between data at time points t' and t''
SPE_{lim}	=	confidence limit of the SPE
SSE	=	sum of squared error
$SSX_{\mathrm{exp},v}$	=	explained variance of variable v
$SSX_{\mathrm{res},v}$	=	residual variance of variable v
SSY_a	=	amount of the **Y** variability explained by the a-th LV
$\mathbf{S}$	=	stoichiometric matrix of GSMMs
s_e	=	error factor for prediction interval calculation
T	=	number of time points
T_n^2	=	Hotelling's T^2 for the n-th observation
T^2_{lim}	=	confidence limit of the Hotelling's T^2
$\mathbf{T}$	=	PCA or PLS score matrix
$\mathbf{t}_n$	=	score vector of PCA or PLS for the n-th observation
$\mathbf{t}_{\mathrm{NEW}}$	=	score corresponding to a new observation
$\mathbf{t}_{\mathrm{DES}}$	=	score vector associated with $\mathbf{y}_{\mathrm{DES}}$
U	=	number of metabolic reactions
$\mathbf{u}$	=	vector of controlled inputs
u_j	=	nutrient concentration profiles planned with DoDE
$u_{j,\mathrm{max}}$	=	maximum value of the profile of nutrient j
$u_{j,\mathrm{min}}$	=	minimum value of the profile of nutrient j
V	=	number of regressor variables
V_{c}	=	volume of the cell culture
V_{E}	=	number of extracellular metabolites
V_{I}	=	number of intracellular metabolites
V_{P}	=	number of process variables
VIP_v	=	VIP of the v-th variable
VIP_{LCL}	=	VIP lower confidence limit
VIP_{R}	=	relative VIP
$\mathbf{W}$	=	PLS weight matrix
$\mathbf{W}^*$	=	PLS weight star matrix
w_{va}	=	weight of the v-th regressor variable on the a-th LV

X_v	=	viable cell concentration
$\mathbf{X}$	=	matrix of regressors
$\mathbf{X}^t$	=	regressor matrix a the t-th time point
$\mathbf{X}_t$	=	regressor matrix unfolded from time point 1 to the t-th time point
$\underline{\mathbf{X}}$	=	multidimensional regressor matrix
$\mathbf{X}_\mathrm{new}$	=	matrix of new observations
$\mathbf{X}^\mathrm{A}$	=	augmented regressor dataset
$\mathbf{X}_\mathrm{train}$	=	training regressor dataset
$\mathbf{X}_\mathrm{val}$	=	validation regressor dataset
$\mathbf{x}_n$	=	regressor vector of the n-th observation
$\hat{\mathbf{x}}_n$	=	regressor vector of the n-th observation reconstructed or predicted
$\mathbf{x}_\mathrm{NEW}$	=	new observation (i.e., sample or experiment)
$\hat{\mathbf{x}}_\mathrm{NEW}$	=	new observation predicted or reconstructed
$\mathbf{x}_\mathrm{lb}$	=	lower bounds for the regressors
$\mathbf{x}_\mathrm{ub}$	=	upper bounds for the regressors
$\mathbf{x}^\mathrm{opt}$	=	optimal subfactors
$\mathbf{x}_\mathrm{art}$	=	artificially generated observation (i.e., sample or experiment)
$\mathbf{x}_\mathrm{NEW}^\mathrm{extra}$	=	extracellular part of $\mathbf{x}_\mathrm{NEW}$
$\mathbf{x}_\mathrm{NEW}^\mathrm{intra}$	=	intracellular part of $\mathbf{x}_\mathrm{NEW}$
$x_{n,v}$	=	value of the v-th original variable for the n-th observation
$\hat{x}_{n,v}$	=	value of the v-th original variable for the n-th observation reconstructed or predicted
$x_\mathrm{NEW}^\mathrm{intra}$	=	estimated intracellular flux values for a generic intracellular reaction
$x_\mathrm{NEW}^\mathrm{extra}$	=	estimated intracellular flux values for a generic extracellular exchange reaction
x_i	=	dynamic subfactors
$\bar{x}_e$	=	average value of e-th effect in the RSM
$Y_{x,\mathrm{gln}}$	=	cell yield on glutamine
$\mathbf{Y}$	=	response matrix
$\hat{\mathbf{Y}}$	=	predicted response matrix
$\mathbf{Y}^\mathrm{A}$	=	augmented response dataset
$\mathbf{Y}_\mathrm{d}$	=	response matrix with dummy variables
$\mathbf{y}_\mathrm{DES}$	=	desired response variable in model inversion
$\hat{\mathbf{y}}_\mathrm{DES}$	=	desired response variable predicted by the model
$\hat{\mathbf{y}}_\mathrm{E}$	=	ensemble of predicted responses
$\mathbf{y}_\mathrm{lb}$	=	lower bounds for the response
$\mathbf{y}_m$	=	vector of the m-th response variable
$\mathbf{y}_n$	=	response of the n-th observation

$\hat{\mathbf{y}}_n$	=	predicted response of the n-th observation
$\mathbf{y}_{\mathrm{train}}$	=	training response vector
$\hat{\mathbf{y}}_{\mathrm{train}}$	=	predicted response vector of the training dataset
$\mathbf{y}_{\mathrm{ub}}$	=	upper bounds for the response
$\mathbf{y}_{\mathrm{val}}$	=	validtion response vector
$\hat{\mathbf{y}}_{\mathrm{val}}$	=	predicted validation response vector
$y_{m,o}$	=	value of the m-th response variable for the o-th new observation
$\hat{y}_{m,o}$	=	predicted value of the m-th response variable for the o-th new observation
y_n^{mAb}	=	original mAb productivity of the n-th cell line
y_n^{biom}	=	original biomass of the n-th cell line
y_0^{mAb}	=	productivity bias
$\mathbf{z}$	=	stoichiometric vector indicating how different intracellular fluxes are combined to form the GSMM objective function
z	=	normalized dynamic variable

Greek letters

α	=	confidence level
α_{x}	=	cellular carrying capacity
β_k	=	first order parameter of the RSM
$\hat{\beta}_k$	=	estimated first order parameter of the RSM
γ	=	random coefficient for linear combination
$\mathbf{\Gamma}$	=	matrix of weights for PLS inversion
δ_m	=	weight for intracellular reaction constraints inclusion
$\Delta_{k,k}$	=	higher order parameter of the RSM
$\hat{\Delta}_{k,k}$	=	estimated higher order parameter of the RSM
η	=	learning rate
λ	=	eigenvalue of PCA or PLS
λ_{mAb}	=	increase factor from mAbs
λ_{reg}	=	regularization coefficient
$\mathbf{\Lambda}^{-1}$	=	diagonal matrix with the inverse PCA or PLS eigenvalues
μ	=	specific growth rate
μ_{max}	=	maximum specific growth rate
μ_{SPE}	=	average of the SPE distribution
$\boldsymbol{\mu}$	=	specific production/consumption rates
$\boldsymbol{\mu}_{\mathrm{max}}$	=	maximum specific rates of production/consumption for each culture variable
$\boldsymbol{\nu}$	=	vector of intracellular reaction rates (i.e., intracellular fluxes)

ν_u = flux of the u-th metabolic reaction

$\nu_u^{\min}$ = lower bound of the u-th metabolic reaction

$\nu_u^{\max}$ = upper bound of the u-th metabolic reaction

$\nu_{\mathrm{MIN}}^{\mathrm{intra}}$ = genetic intracellular flux lower bound given by the GSMM

$\nu_{\mathrm{MAX}}^{\mathrm{intra}}$ = genetic intracellular flux upper bound given by the GSMM

$\nu_{\mathrm{MIN},n}^{\mathrm{extra}}$ = generic flux lower bound for extracellular exchange reactions given by the extracellular metabolite uptake rates of the n-th cell line

$\nu_{\mathrm{MAX},n}^{\mathrm{extra}}$ = generic flux upper bound for extracellular exchange reactions given by the extracellular metabolite uptake rates of the n-th cell line

$\nu_{\mathrm{MIN},v}$ = flux lower bound for reaction v given by the GSMM

$\nu_{\mathrm{MAX},v}$ = flux upper bound for reaction v given by the GSMM

σ = general standard deviation

σ_v = standard deviation of the v-th process variable

σ_{SPE}^2 = variance of the SPE distributionτ=dimensionless culture time

$\mathbf{\Phi}$ = matrix for prediction interval calculation

$\mathbf{\Psi}$ = matrix for prediction interval calculation

$\omega_{h,v}$ = weight of the h neurons associated with the v-th input variable

ω_h^0 = bias of the h-th neuron

$\boldsymbol{\omega}$ = matrix/vector of ANN weights

Appendix symbols

a_{d} = constant for death rate

c_{H} = concentration of heavy chains

c_{L} = concentration of light chains

$c_{\mathrm{H_2L}}$ = concentration of heavy and light chain intermediates

$c_{\mathrm{H_2}}$ = concentration of heavy and light chain intermediates

$c_{\mathrm{H_2L_2}}^{\mathrm{ER}}$ = concentration of heavy and light chain intermediates

$c_{\mathrm{H_2L_2}}^{\mathrm{G}}$ = concentration of heavy and light chain intermediates

f_{inh} = inhibition factor

K_{A} = assembly rate constant

K_{d} = inverse saturation constant for the specific death rate

$K_{\mathrm{d,amm}}$ = ammonia constant for cell death

$K_{\mathrm{d,gln}}$ = constant for glutamine degradation

K_{ER} = rate for ER-to-Golgi transportK_{G}=rate for Golgi-to-medium antibody transport

K_{lac} = Monod constant for lactate consumption

K_{RNA} = heavy and light chain mRNA decay rate

KI_{lac} = Monod constant for lactate

KI_{amm}	=	Monod constant for ammonia
m_{gln}	=	maintenance coefficient of glutamine
m_{H}	=	heavy chain mRNA concentration
m_{L}	=	light chain mRNA concentration
N_{H}	=	heavy chain gene copy number
N_{L}	=	light chain gene copy number
Q_{amm}	=	specific ammonia production rate
Q_{gln}	=	specific glutamine consumption rate
Q_{lac}	=	specific lactate production rate
$Q_{\mathrm{lac,cons}}$	=	specific lactate consumption rate
Q_{mAb}	=	specific mAb production rate
R_{H}	=	rates of heavy chain consumption in assembly
R_{L}	=	rates of light chain consumption in assembly
S_{H}	=	heavy chain gene specific transcription rates
S_{L}	=	light chain gene specific transcription rates
T_{H}	=	heavy chain specific translation rates
T_{L}	=	light chain specific translation rates
X_{t}	=	total cell concentration
$Y_{\mathrm{amm,gln}}$	=	yield of ammonia from glutamine
$Y_{\mathrm{lac,glc}}$	=	yield of lactate from glucose
$Y_{\mathrm{mAb,glc}}$	=	yield coefficient of mAb from glucose
$Y_{x,\mathrm{glc}}$	=	yield of cells on glucose
$Y_{x,\mathrm{gln}}$	=	yield of cells of glutamine
$Y_{\mathrm{x,lac}}$	=	yield coefficient of cells from lactate consumption

Greek letters

α_1, α_2	=	contestants of glutamine maintenance coefficient
γ_1, γ_2	=	constants for antibody production
ε_1	=	ER glycosylation efficiency factor
ε_2	=	Golgi apparatus glycosylation efficiency factor
μ_{d}	=	specific death rate
$\mu_{\mathrm{d,max}}$	=	maximum specific death rate
ξ_{mAb}	=	molecular weight of mAbs

Chapter 1

State of the art and motivation

This Chapter provides the background and the motivations of this Dissertation. First, the current state of the biopharmaceutical industry is presented. Then, monoclonal antibodies (mAbs), cell cultures and their development pipeline are introduced. Furthermore, the state-of-the-art mathematical modeling in the biopharmaceutical industry 4.0 is presented and critically discussed, focusing on the subdivision between data-driven and knowledge-driven models. Finally, the main challenges in the application of mathematical modeling in the biopharmaceutical industry are pinpointed to introduce how they have been addressed in this Dissertation.

1.1 Biopharmaceutical industry: an overview

Biopharmaceuticals (also known as biologicals or biologics) are therapeutics and drugs synthetized or extracted from living organisms (such as microbial, animal, or human cells) used for the treatment and the prevention of diseases (Hong et al., 2018). The living organisms used in the biopharmaceutical industry are genetically modified to specifically produce the desired biologicals with therapeutic effects.

Biologicals, being produced by living cells, have different characteristics with respect to traditional pharmaceuticals that are chemically synthetized. In fact, the synthesis from biological source allows the production of more complex drugs than traditional pharmaceutics (Rader, 2008). For example, a traditional drug may count dozens of atoms, while a typical biological may be 100 or 1000 times larger. Furthermore, on the therapeutic side, biologicals have more specific targets, thus producing fewer side effects (Kesik-Brodacka, 2018).

The main biopharmaceutical products are:

- vaccines;
- cells, such as stem cells;
- biological tissues;
- recombinant proteins, such as monoclonal antibodies;
- gene therapy medicinal products.

In the last decade, the biopharmaceutical industry has considerably grown and nowadays the biopharmaceutical market covers the largest portion of the pharmaceutical sector (Tripathi &

Shrivastava, 2019). In 2017, the global annual sales of the biopharmaceutical compartment were estimated at 188 billion $ (Walsh, 2018), showing a growth of 8.3% per year between 2010 and 2015 (Smietana et al., 2016). Other authors report estimated global sales for 2019 of 450 billion $ (Hong et al., 2018) and 227 billion $ (Epifa, 2021), while the estimates for 2021 are 401 billion $ (MordorIntelligence, 2021). The global sales of the biopharmaceutical compartment are forecasted to achieve 534 billion $ in 2027 with an expected yearly growth of 7.32% (MordorIntelligence, 2021). In terms of biopharmaceutical products, 155 new biologicals were approved between 2014 and 2018 (Walsh, 2018), with more than 7000 drugs in the development pipeline in 2019 (Hong et al., 2018).

Table 1.1 *The 10 top-selling biopharmaceutical products in 2017. Adapted from Walsh (2018).*

Rank	Product	Type	Sales 2017 ($ billions)	Cumulative sales, 2014-2017 ($ billions)	Approval year	Company
1	Humira	mAb	18.94	62.6	2002	AbbVie, Elsai
2	Enbrel	recombinant protein	8.34	35.4	1998	Amgen, Pfizer, Takeda Pharmaceutics
3	Rituxan	mAb	7.78	29.1	1997	Roche, Biogen Idec
4	Remicade	mAb	7.77	35.6	1998	Johnson & Johnson, Merck, Mitsubishi Tababe Pharma
5	Herceptin	mAb	7.39	27.1	1998	Roche
6	Avastin	mAb	7.04	27.0	2004	Roche
7	Lantus	insulin	6.72	27.4	2000	Sanofi
8	Eylea	recombinant protein	5.93	18.0	2011	Regeneron, Bayer
9	Opdivo	mAb	5.79	11.4	2014	Bristol-Meyers Squibb, Ono Pharmaceuticals
10	Neulasta	recombinant protein	4.53	20.1	2002	Amgen, Kyowa Hakko Kirin

In the biopharmaceutical market, monoclonal antibodies (mAbs) are the biggest selling class of biologicals (Hong et al., 2018), with 123 billion $ of estimated annual sales in 2017 (Walsh, 2018). Monoclonal antibodies are a class of anti-viral and anti-cancer biologicals whose sales grew by 9.8% per year between 2010 and 2015 (Smietana et al., 2016), and are expected to achieve 138.6 billion $ sales by 2024 (O. Yang et al., 2020), favored by the new advancements in large-scale recombinant protein production (Tripathi & Shrivastava, 2019). A list of the top selling biologicals is reported in Table 1.1. In 2017 mAbs represented 53% of the overall new approval in the biopharmaceutical sector (Walsh, 2018), with more than 1500 new mAbs in the development pipeline (Hong et al., 2018). Nowadays, the preferred method for recombinant protein production is mammalian cell cultures, accounting for 84% of the overall production (Walsh, 2018). Among these, the 84% of produced mAbs were synthetized by Chinese Hamster Ovary (CHO) cells (Walsh, 2018). Other recently approved mAbs were synthetized by *E. Coli* and *Saccharomyces Cerevisiae* (Tripathi & Shrivastava, 2019).

1.2 Monoclonal antibodies

Monoclonal antibodies (mAbs) are therapeutic proteins currently utilized for the treatment of autoimmune diseases, cancers, and infectious diseases (Kesik-Brodacka, 2018). The main pathologies treated with mAbs are rheumatoid arthritis, Chron's disease, leukemia, colorectal cancer, and metastatic breast cancer, hepatitis A and B viruses, HIV-1 infection, and SARS-CoV-2 (Castelli et al., 2019).

Monoclonal antibodies (or immunoglobulins, Ig) are large Y-shaped proteins (Figure 1.1), with two identical heavy chains and two identical light chains, connected via disulfide bunds (Castelli et al., 2019; Chartrain & Chu, 2008; Chiu et al., 2019).

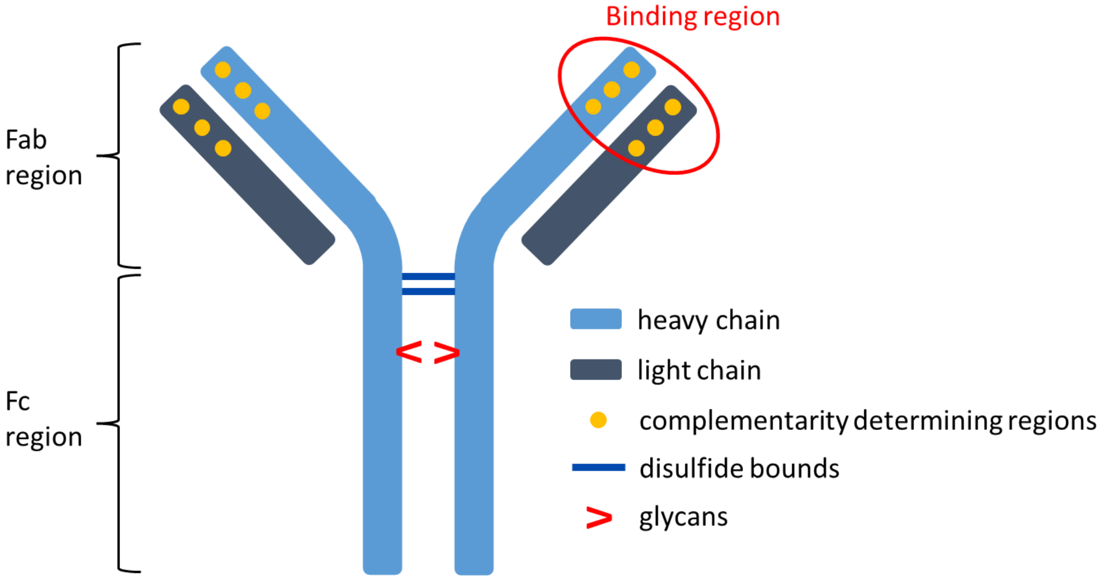


Figure 1.1 *Structure of monoclonal antibodies. Adapted from Chartrain and Chu (2008).*

Monoclonal antibodies are divided in the Fc region and the antibody binding region (Fab) with specialized sites, called complementarity determining regions, which dictate the specificity of each mAb through their amino acid sequence (Chartrain & Chu, 2008; Gaughan, 2016; Kang & Lee, 2021).

The Fc region is glycosylated with N-linked glycans (i.e., polysaccharides) with two N-acetylglucosamine residues connected to three bisecting mannose residues (i.e., bi-antennary structure) and variable terminal sugar composition, affecting the activity of antibodies (Batra & Rathore, 2016; Sha et al., 2016).

The main role of antibodies in living organisms is to clear the host from invading pathogens and external molecules. Antibodies binds to very specific targets, called antigens, forming a complex that is recognized and cleared by specialized components or cells of the immune system of the host organism (Castelli et al., 2019; Chartrain & Chu, 2008).

In living organisms, mAbs are mainly produced by secretory B-cells, a component of the cell immune system (Gaughan, 2016). The secretion of monoclonal antibodies follows a specific mechanism, which involves light and heavy chain translation, folding, and glycosylation in the endoplasmic reticulum and Golgi apparatus (Gutierrez et al., 2020; Kontoravdi et al., 2005, 2007, 2010). More detail on monoclonal antibodies structure, functions, and production are reported in Appendix A.1.

1.3 Cell cultures

Cell cultures are the main industrial means used to produce mAbs, which at industrial level are, nowadays, organized in two sections (Shukla & Thömmes, 2010):

- upstream: section in which mAbs are produced by mammalian cells in large bioreactors;
- downstream: section aimed at the purification of mAbs to reduce product and non-product related impurities to acceptable levels.

Additional details on cell cultures are reported in Appendix A.2. However, an overview of the mAb production process helps in better understanding the steps and the data available along the mAb development.

The advent of Industry 4.0 open the era of Big Data, characterized by an extremely easy access to massive amounts of data, advanced modeling tools and computational resources (Qin, 2014; Sansana et al., 2021). The availability of Big Data can lead to new solutions for problems not addressed in the past. The biopharmaceutical filed has started to acknowledge the opportunities arising by exploiting the large amount of physical, chemical, and biological data available along the development of new biologicals. This is leading to an increasing interest towards the transition to the biopharmaceutical industry 4.0 era, which is also favored by the awareness of the regulatory agencies of the huge impact of Big Data (Banner et al., 2021; Food and Drug Administration, 2022; ICH Harmonised Tripartite Guideline, Guidance for Industry, Q8 Pharmaceutical Development, 2009; Silva et al., 2020). Furthermore, this transition is largely supported by the introduction of high-throughput systems which allow to collect massive amounts of data and are at the basis for the advanced modeling of bioprocesses. In mAbs cell culture the main available data types are:

- process information;
- omics data.

1.3.1 Upstream process

In the upstream section of biopharmaceutical processes, biopharmaceuticals are produced in bioreactors, where cells responsible for production grow.

The main bioreactors used for the production of mAbs are stainless steel bioreactors and disposable ones (Chartrain & Chu, 2008; Gaughan, 2016; Rodrigues et al., 2009a). Stainless steel bioreactors are stirred tanks with volume ranging from 1000 L to 25000 L, while disposable bioreactors are polymeric bags with a volume ranging between 50 L and 2000 L and some forms of mixing strategy.

The cells growing in the bioreactor are highly sensitive to culture conditions, making product yield and quality dependent on bioreactor operating parameters (Birch & Racher, 2006; Chartrain & Chu, 2008; F. Li et al., 2010; Rodrigues et al., 2009a; Shukla & Thömmes, 2010). The most critical parameters for cell health are temperature and pH, which have a direct impact on growth and productivity. Dissolved gasses, such as O_2 and CO_2, are other important operating parameters, since O_2 is required to produce energy from carbon sources and CO_2 to maintain pH and regulate cell activities. Other bioreactor operating parameters are osmolarity, which affects the duration of the exponential growth phase, and agitation rate, which is strictly related to the control of the dissolved gasses.

In order to survive, grow, and produce, cells require the presence of a culture medium and some nutrients (Birch & Racher, 2006; Chartrain & Chu, 2008; Gaughan, 2016; Rodrigues et al., 2009a; Shukla & Thömmes, 2010). Cell media contain all the growth supporting molecules, such as amino acids, vitamins, nucleosides, trace elements, metals, inorganic salts, lipids and insulin or insulin-like growth factors. The main cell nutrients are glucose and glutamine, being the primary sources of carbon. The metabolism of such nutrients leads to the production of toxic by-products, such as lactate and ammonia, which strongly inhibit growth and productivity, and reduce product quality, when they accumulate in the culture. For this reason, an appropriate feeding strategy to maximize productivity and growth, and to minimize the formation of undesirable by-products is required (Chartrain & Chu, 2008).

In relation to nutrient management, bioreactors can be operated in three modes (Chartrain & Chu, 2008; Gaughan, 2016; Rodrigues et al., 2009b; Shukla & Thömmes, 2010): *i*) batch, *ii*) fed-batch, and *iii*) perfusion. In batch operating mode, the bioreactor is initially loaded with medium, nutrients, and cells, which are allowed to grow with no further nutrient additions or withdrawals. In fed-batch mode, nutrients are periodically added with fresh medium to increase culture longevity, typically 2 weeks, maintain nutrient sufficiency, and limit the effect of nutrient depletion, but without avoiding the accumulation of growth-inhibiting by-products. In perfusion bioreactors, fresh medium is continuously added to the culture at very low rate, while an equal amount of spent medium with the product is removed from the culture, leading to very stable operations lasting for long periods of time, even 35-40 days.

1.3.2 Downstream process

In the biopharmaceutical production of mAbs, the downstream process has become widely established to purify the product and reduce all the impurities to acceptable levels (Birch & Racher, 2006; Gronemeyer et al., 2014; Shukla & Thömmes, 2010). The downstream process is mainly divided into three steps: *i*) protein A affinity chromatography, *ii*) polishing chromatography, and *iii*) viral filtration. In protein A affinity chromatography antibodies are separated from host cell proteins, DNA and other impurities, and in polishing chromatography a further separation of DNA and high-molecular-weight aggregates is performed, while in virus filtration viruses are inactivated and removed.

1.3.3 Data in cell cultures

During cell cultures a large amount of data is typically collected. Those data are the basis of all modeling activities performed in the biopharmaceutical industry. The main data types available in mammalian cell cultures are:

- process data;
- biological data.

1.3.3.1 Process data

Process data provides information on the macroscopic behavior of the cell culture. They are the measurements of the main process parameters and chemical properties, typically monitored to follow the culture growth and cell metabolism, and identify the root cause of any decline in cell health (F. Li et al., 2010; Rodrigues et al., 2009b).

Measurements of process parameters are categorized according to their type as: *i*) on-line, and *ii*) off-line. Some process parameters, such as temperature, pH, and dissolved gases, are measured on-line, while others, such as osmolarity, viable cell concentration (VCC), product and metabolite concentrations, are measured off-line. VCC, being the most critical measurement to evaluate culture physiology as response to the culture conditions, is typically measured by taking daily samples from the bioreactor. Similarly, metabolite concentrations, such as glucose, glutamine, ammonia, lactate, and glutamate, but also amino acids, are routinely measured by taking periodic samples from the bioreactor. The typically measured process parameters, their measurement type and analytical instrument is reported in Table 1.2.

Table 1.2 *Typically measured process parameters with measurement type and analytical instrument.*

Parameters	Type	Instrument
temperature	on-line	thermocouples or resistance temperature devices
pH	on-line off-line verification for drifts and reduced sensibility	autoclavable probes
dissolved oxygen (DO)	on-line	electrodes (Clark type)
dissolved CO_2	on-line off-line in some cases	specific sensors and mass-spectrometry
osmolarity	off-line	commercial analyzers through freezing-point depression osmometry
viable cell concentration	off-line	automated analyzers based on image analysis
metabolite concentrations	off-line	high-performance liquid chromatography
product concentration	off-line	enzyme-linked immunosorbent assay, Western blot, protein A liquid chromatography, or bioassays

1.3.3.2 Biological data

Biological data available in cell cultures regards the internal microscopic characteristics and behavior of the living organisms used in bioprocesses. The biological data concerns the flow of information (Figure 1.2) that in all living organisms moves from DNA to mRNA, proteins and metabolites, finally expressing itself in the cell phenotypes (Reel et al., 2021). Specifically, the information encoded in the DNA is transcribed in the mRNA, which contains the information that is then translated into proteins. The proteins, in the form of enzymes, catalyze almost all metabolic reactions among metabolites. The concentration of metabolites is strictly connected to the cell phenotype that can be observed. This informative flow of biological information is captured through omics techniques, which can be categorized according to the source of the information. The main types of omics techniques are:

- genomics;
- transcriptomics;
- proteomics;
- metabolomics;
- fluxomics.

Genomics concerns the analysis of the biological information of cells at the DNA level (Reuter et al., 2015). It provides the data on the actual DNA sequence, on single nucleotide polymorphisms, rare variants, and variations in the copy number. Typically, genomics data are generated with sequencing methods followed by several post-processing.

Transcriptomics concerns the analysis of the biological information at the mRNA level (Lowe et al., 2017). In this way, all the information that is recorded in the DNA and expressed through transcription in the transcriptome, the collection of all the transcribed mRNA, can be collected. Transcriptomics provide information on the genes that are expressed and their expression level. Measurements are performed with sequencing techniques, microarray, and mass spectrometry.

Proteomics concerns the analysis of the biological information at the protein level (Aslam et al., 2017). It allows to identify and quantify all the proteins that are in a cell, providing information on the expressed proteins and their abundance. Proteomics measurements are typically performed with mass spectrometry and liquid chromatography - mass spectrometry.

Metabolomics concerns the analysis of the biological information at the metabolite level (B. Zhou et al., 2012). Metabolomics identifies and quantifies all the small molecules involved in metabolic reactions of a given system, called metabolites. Accordingly, it provides information on the metabolites identified in a system and their abundance. Metabolomic measurements are typically carried out through mass spectrometry, liquid chromatography - mass spectrometry (LC-MS), gas chromatography - mass spectrometry (GC-MS).

Fluxomics concerns the analysis of the biological information at the reaction flux level (Winter & Krömer, 2013). Fluxomics is the detailed quantification of the intracellular metabolic fluxes with ^{13}C isotope labeling experiments. Specifically, isotopically enriched substances are provided to cell cultures, which uses them as carbon sources. Then mass spectrometry or nuclear magnetic resonance are used to quantify the carbon enrichment in the target metabolites. Finally, specific metabolic models are required to balance those metabolites and obtain an estimate of the intracellular fluxes.

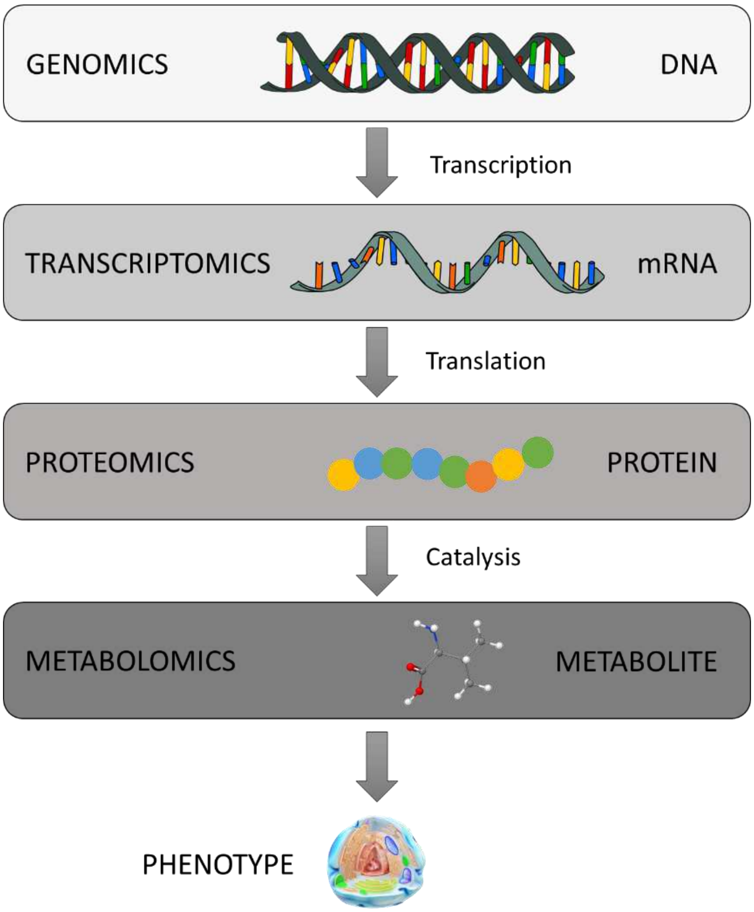


Figure 1.2 *Flow of biological information from DNA to phenotype.*

This brief explanation elucidates how genomics, transcriptomics and proteomics provide an understanding of the biological characteristics of cell, while metabolomics and fluxomics are excellent indicators of cell activity and are the closest information to cell phenotype (Reel et al., 2021).

The presented omics techniques are not the only ones. In fact, additional minor omics techniques such as lipidomics (i.e., study at the lipid level) and glycomics (i.e., study at the glycan level) exist. Details on these techniques can be found in Reel et al. (2021).

1.4 Bioproduct development

The development of mAbs is a very intensive and time-consuming process that follows two parallel paths: one concerning the development of the product, from the molecule discovery to the drug approval, and the second concerning the development of the production process, from the generation of the cell lines to the manufacturing process.

1.4.1 Product development

The product development consists in a series of steps required to bring a new drug product from its discovery to the market (Figure 1.3). The initial steps of the development pipeline consist of all the research activities required for the new product, which are:

- drug discovery: this step may last up to 6 years and has a very low success rate, typically less than 0.01% (Epifa, 2021; IFPMA, 2022). At this stage, the target antigen for the mAb must be selected based on the knowledge of the biological processes involved in a specific disease. This allows to design the mechanism of action, binding specificity, affinity, kinetics and the isoform of the antigen (Mould & Meibohm, 2016). At this stage, the patent application for the new drug product is submitted.
- pre-clinical trials: this stage is aimed at testing the safety and efficacy of the new drug through *in vitro* and *in vivo* tests. Initially, mAbs with undesirable properties are excluded to identify the ones with specific affinity properties. *In vivo* tests are typically conducted on model animals, which should be appropriately selected to have the closest representation of the human effect. These typically include pharmacodynamics and pharmacokinetics test, which, unfortunately, are not always predictive of the immunogenicity in humans. Other tests concern the toxicity and repeated dose toxicity, which allow to predict the risk of adverse events in human and select a safe starting dose (Mould & Meibohm, 2016). These pre-clinical trials may last up to 3 years (Epifa, 2021; IFPMA, 2022).

The initial research for the development of new drugs is resource intensive, and is estimated to require ~15.7% of the overall research and development budget (Epifa, 2021).

Once the pre-clinical trials are completed, the new drug undergoes to a series of clinical trials (i.e., in humans). The clinical trials are divided in the three phases, and the new drug must subsequently pass all the phases to be approved and launched on the market. The phases are:

- Phase I: this step is aimed at understanding the efficacy of and safety of drugs on humans. Specifically, the suitable mAb dose in humans is determined, together with potential interactions with others drugs, that are typically small for mAbs (Mould & Meibohm, 2016). This trial involves a small number of healthy volunteers (typically 20-100), with a range of success of 57% (IFPMA, 2022). However, sometimes mAbs for cancer treatments are tested directly on patients (Mould & Meibohm, 2016).
- Phase II: this stage is aimed at testing the final dose regime, the efficacy, and the presence of side effects in humans. In mAbs their efficacy and safety must be evaluated on the target patients in order to get pharmacokinetics and pharmacodynamics data in case of different disease severity (Mould & Meibohm, 2016). Phase II typically involves 100-500 volunteers,

and has a much smaller rate of success (39%) because inefficacy or toxic effects are usually discovered at this stage (IFPMA, 2022).

- Phase III: this stage is aimed at testing the efficacy of the drug comparing it with other available treatments. It involves a larger number of volunteers (1000-5000) and represents the most costly and time-consuming phase, but has a high success rate about 68% (IFPMA, 2022).

The clinical trials require up to 6-7 years and ~48% of the overall research and development budget (Epifa, 2021; IFPMA, 2022).

Once a drug positively passes all clinical trials, it can be submitted to the regulatory agency for the approval. The approval step may last 2 years and requires up to 4% of the overall research and development budget (Epifa, 2021; IFPMA, 2022).

Overall the development of a new mAb may last 12-13 years and has an estimated cost for the entire research and development phase of more than 2.55 billion $ (Epifa, 2021). This high development and manufacturing cost, together with the typically low amount of produced products and the fact that companies have only few remaining years to return the investment before patent expiration, which lasts for 20 years from the initial application (Epifa, 2021), drastically increases the price of biopharmaceuticals. This has a large impact on the costs that patients and national health care systems must sustain to access potentially life-saving drugs.

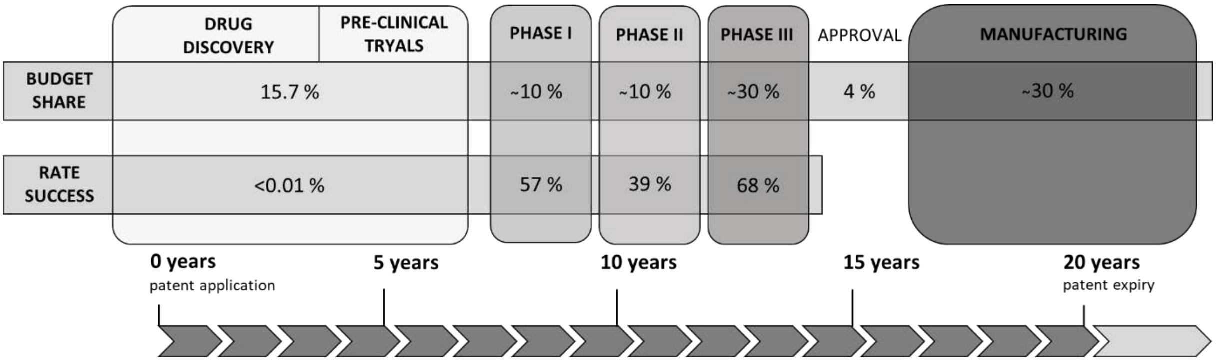


Figure 1.3 *Product development phases. Adapted from Destro and Barolo (2022).*

1.4.1 Process development

The process development comprises all the activities aimed at the large-scale production of a biopharmaceutical product. Process development is resource-intensive and time-consuming, and is typically divided into several step (Figure 1.4): host cell generation and engineering (4-8 months), cell line selection (8-12 months), process characterization (4-8 months), media and feed optimization, and process optimization (6-12 months) (Chartrain & Chu, 2008; Gronemeyer et al., 2014; Tripathi & Shrivastava, 2019). Furthermore, the process development is typically divided into early and late stages. During early stages the process is rapidly developed to produce material for Phase I and II clinical trials and toxicology studies. With phase III, the development moves to the late stages in which the best cell lines are selected, the process, media and feeding are optimized, together with the bioreactor operating conditions (F. Li et al., 2010). This Section will only consider the development of the upstream process, being the one requiring the hugest effort.

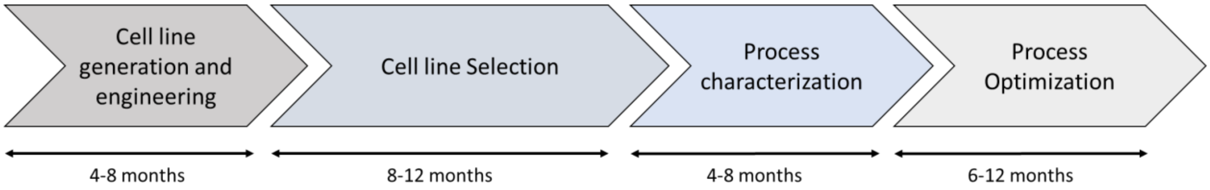


Figure 1.4 *Process development phases. Adapted from Chartrain and Chu (2008).*

The large improvements obtained after years of research in host cell engineering, screening methods, medium and feed development, and process engineering allows today's processes to achieve high yield and be economically viable (Wurm, 2004). However, a major aspect for the cost effectiveness of the process development relies on reducing the long timelines required for a new biopharmaceutical to reach the market (F. Li et al., 2010). For this reason, biopharmaceutical companies are looking for innovative science-based solutions to support and accelerate the various stages of process development.

1.4.1.1 Cell generation and engineering

The first step of the process development is the generation of the cell lines responsible for the production of the desired mAb. After the selection of the host cell, transfection is performed to make cell lines produce the antibody of interest. In transfection, a plasmid bearing the antibody light and heavy chain genes is used to insert the information into the cells. After transfection, cell lines are initially screened for the expression of the desired protein with high productivity and stability, and desired product quality (Gronemeyer et al., 2014; Tripathi & Shrivastava, 2019).

At this stage host cells are genetically engineered to improve or modify product quality, growth, productivity, and robustness (F. Li et al., 2010). The main strategy to genetically engineer host cells is performed by acting on gene regulation, such as downregulating, upregulation or knocking out the expression of specific genes (Lai et al., 2013). Several cellular functions are targeted to improve the functions of CHO cells such as apoptosis, autophagy, proliferation, regulation of cell cycle, protein folding, protein secretion and metabolites production. However, one of the main areas of improvement concerns the cell viability and productivity. For example, a combination of anti-apoptosis and secretion genetic engineering led to 60% increased antibody titers (Lai et al., 2013).

Glycosylation control has also received a lot of attention because antibody glycosylation pattern has a major impact on the bioactivity of a mAb (Gronemeyer et al., 2014; F. Li et al., 2010). Other genetic engineering targets concern metabolic improvements typically aimed at reducing ammonium and lactate accumulation, and the improvement of the therapeutic efficiency, such as the antibody dependent cell mediated cytotoxicity (Lai et al., 2013).

The genetic engineering of host cells can strongly benefit from the advances in the genome-wide *in silico* modeling of mammalian system, especially if coupled with advanced omics tools. This can lead to the creation of fully optimized mammalian cell lines through multiple genetic modifications (Lai et al., 2013).

1.4.1.2 Cell lines selection and scale-up

Cell line screening and selection is one of the most important steps in upstream process development, because the selection of inappropriate cell lines have major consequences on the entire development process. In fact, a change in the production cell lines during late development stages is considered a major process change, requiring additional clinical studies. Consequently, the production cell line should be identified before Phase III trials, but, preferably, even during Phase I (F. Li et al., 2010).

After cell generation, the obtained pool of candidate cell lines is highly heterogeneous, because the random integration of the recombinant gene into the host cell genome produces very different expression levels. Typically, highly productive cell lines are rare and they often show a low growth rate because a significant portion of the metabolic resources are diverted for protein synthesis (Lai et al., 2013; F. Li et al., 2010). For this reason, an extremely large number of cell lines (thousands) are screened to isolate the ones with desired characteristics, making this cell selection stage very time-consuming (6-12 months) and labor- and capital-intensive.

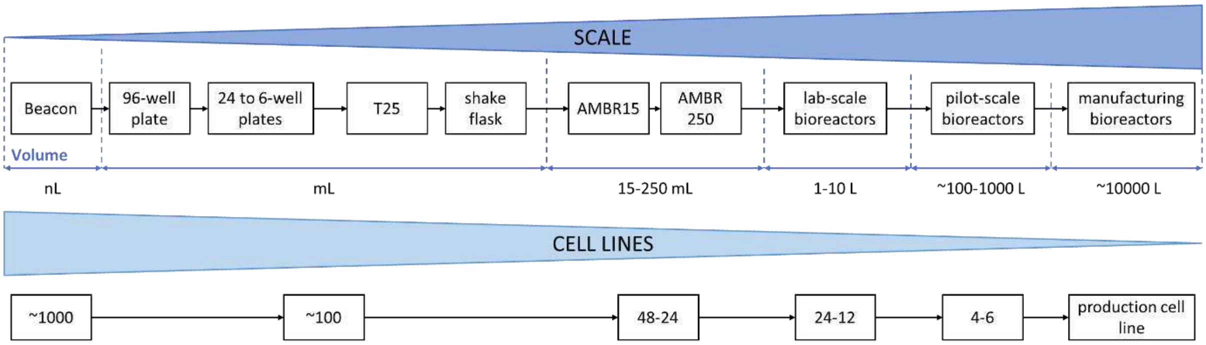


Figure 1.5 *Schematic representation of the cell selection and scale-up phase. Adapted from Facco et al. (2020).*

The main criteria for cell lines selection (i.e., critical quality attributes, CQAs) are based on: growth, cell specific productivity, stability, glycosylation profiles, aggregate formation, protein sequence heterogeneity, robustness, high viability, metabolic characteristics, low lactate and ammonium generation (Gronemeyer et al., 2014; F. Li et al., 2010; Wurm, 2004). The evaluation of such CQAs is not straightforward, because they are cell line dependent and many of them are affected also by the specific processing conditions (F. Li et al., 2010). Since the CQAs that can be quantified and the number of cell lines that can be screened are specific for each process scale, cell line selection is performed at different process scales (Figure 1.5). After transfection, several hundreds of cell lines are screened at different static deep well plate scales (96, 24, 12, and 6) up to 5 mL to assess growth and productivity. More recently, nanofluidic technologies (Le et al., 2018), such as the Beacon platform, are utilized to simultaneously screen thousands of transfected cell lines at the nanoliter scale. After this initial screening, only the cell line meeting the desired CQAs are progressed to shaken scales, such as T25 and shake flasks, with larger volumes (25-500 mL). At this stage typically 24-48 cell lines meeting the desired CQAs are identified and progressed to high-throughput multi-parallel bioreactors (i.e., AMBR15^{TM} and AMBR250^{TM}). These systems have 12-48 single-use parallel bioreactors with a working volume of 15-250 mL. The AMBR^{TM} bioreactors are equipped with an internal impeller and a gas sparger, allowing them to mimic the larger bioreactors of 3, 5, and 200 L (Rameez et al., 2014). The AMBR^{TM} systems together with lab scale bioreactors (1-10 L) allow the isolation of 4-6 highly performing cell lines that are further evaluated at pilot scale (F. Li et al., 2010). Finally, the last tests on the pilot scale leads to the selection of the production cell line which will be used for the entire life-span manufacturing of the mAb.

Several advancements in the analytical technologies, such as the introduction of high-throughput systems, allow a faster and a wider screening of cell lines, consistently improving the identification of promising cell lines. However, the selection of high performing cell lines in a limited time frame is still a major challenge (Lai et al., 2013; F. Li et al., 2010). Furthermore, a better understating of cell biology from omics techniques, such as genomics, transcriptomics, proteomics, and metabolomics, would definitely improve the selection of highly performing cell lines and reduce the overall cost of process development (F. Li et al., 2010).

1.4.1.3 Process characterization

The process characterization is an important step because it is required for the drug approval, as part of the biologic license application (ICH Harmonised Tripartite Guideline, Guidance for Industry, Q8 Pharmaceutical Development, 2009; F. Li et al., 2010). Process characterization usually takes place after the completion of the Phase III trials, when no significant changes are introduced in the manufacturing process. However, part of the process characterization is also performed early on during the development process, because it provides valuable information for the completion of other steps such as process and media optimization (F. Li et al., 2010).

Process characterization is aimed at understanding the impact of process operating parameters (i.e., critical process parameters, CPPs) on the product characteristics (i.e., CQAs), establishing the acceptable ranges for operating parameters, and demonstrating process robustness. This is necessary to understand how the quality specification required by the regulatory agencies can be met and ensured.

Process characterization is typically performed through DoE in scale-down models (i.e., processes), which allow to accelerate the experimentation and reduce costs (F. Li et al., 2010), since process characterization at manufacturing scale is not feasible because of the high costs and time requirements for this kind of experimentation. However, to demonstrate the consistency of the process performance between the scale-down models and the commercial scale, full scale validation is required (F. Li et al., 2010).

Process characterization in the biopharmaceutical industry is still behind the one in the traditional pharmaceutical sector. The main bottleneck in bioprocess characterization is the limited combination of high-throughput systems with high-resolution quality analytics, because of the long timelines for analyses, cost, and high device complexity. The availability of high-throughput systems with real time CQA determination would greatly benefit process characterization and the entire bioprocess development (Guerra et al., 2019). Furthermore, characterization of the relationship between the variability of metabolic states and heterogeneity in process behavior is still an open issue and only few publications are focused on that (Guerra et al., 2019). Advancements in omics technologies, such as genomics, proteomics, and metabolomics, and methodologies to understand how cell biology is affected by process conditions and vice versa can lead to a better process characterization, even allowing to adjust the process conditions according to biological difference in host cells (F. Li et al., 2010).

1.4.1.4 Media and feed optimization

Another important step in upstream process development is the optimization of the media and feeding schedule, which is essential to balance cell growth and productivity, and to achieve adequate product quality.

The culture media must be optimized for every cell line, because of their intrinsic behavior diversity. The optimized media must consider the differences in cell metabolism, nutrient consumption, by-products formation, and the balance between growth and productivity (Gronemeyer et al., 2014; Tripathi & Shrivastava, 2019). Typically, media optimization is performed through Design of Experiments (DoE), but the recent advent of genome-wide *in silico* modeling will be precious to this purpose (Gronemeyer et al., 2014).

The optimization of the feeding schedule is extremely important to correctly balance growth, productivity, product quality and accumulation of growth-inhibiting by-products. To reduce the production of growth-inhibiting by-products, such as lactate and ammonia, the concentration of glucose and glutamine is usually maintained at a low level (F. Li et al., 2010). This is achieved by feeding cell cultures with frequent boluses, at specific time periods, of glucose and glutamine, which are the main carbon sources for mammalian cells. The optimization of the feeding schedule deals with the identification of the best way of providing nutrients over time. The optimization of the feeding schedule is typically done with high-throughput scale-down equipment and statistical DoE (Gronemeyer et al., 2014; F. Li et al., 2010). However, DoE only considers static factors, but can be extended to dynamic ones by assigning a different DoE factors to the feeding action at each day (Mora et al., 2019). This rapidly leads to extended designs with several dozens of experiments, especially in the case of long and multiple dynamic variables to consider. For this reason, strategies to optimize dynamic variables , such as the Design of Dynamic Experiments (DoDE; Georgakis, 2013), are required. In this case the experiments are specifically designed to deal with time-varying factors while minimizing the total number of required experimental runs, in such a way as to obtain the maximum information content from the experiments. These characteristics make DoDE a good candidate for the optimization of the feeding schedule (Wang & Georgakis, 2017). Unfortunately, even in DoDE the number of experimental runs increases with the number of dynamic variables and the complexity of the dynamic profiles. Due to the high cost and time required by each biopharmaceutical experimental run, the strategies selected for the optimization of the feeding should be able to deal with data scarcity without requiring too many experimental runs.

1.4.1.5 Process optimization

The optimization of process operating parameters (i.e., CPPs) in another essential step to achieve stable and high expression of protein and adequate product quality.

Temperature, pH, agitation, aeration, dissolved oxygen (DO), CO_2, and hydrodynamic shear, osmolality, redox potential, and addition of cell culture additives are the CPPs that are typically optimized (Gilgunn & Bones, 2018; Gronemeyer et al., 2014; F. Li et al., 2010; Tripathi & Shrivastava, 2019). At this stage, optimal temperature and pH shifts are typically determined, together with the required gas exchange rate (Tripathi & Shrivastava, 2019).

This process optimization is typically performed though experimental methodologies, which leads to cell cultures with high cell growth, enhanced productivity and better product quality, such as a reduced content of host cell proteins (Gilgunn & Bones, 2018; Tripathi & Shrivastava, 2019).

Recently, a lot of effort is being made in the optimization of disposable bioreactors. However, work is still necessary to optimize the aeration system, develop standardized methods, and solve some performance issues. Furthermore, the development of validation strategies for assessing the risk of leachables and extractables from the disposable plastic is still an open issue.

1.4.2 Regulatory aspects of bioprocess development

The biopharmaceutical process development, as well as the entire biopharmaceutical industry, is highly regulated and is subject to the supervision of regulatory agencies, which aim at protecting and improving human health. U. S. Food and Drug Administration (FDA) and the European Medicine Agency (EMA) are two of the most important regulatory agencies, but other regulatory bodies are spread all over the world. Many of these regulatory agencies share an aligned position on important regulations for the drug products that are on the market through the International Council on Harmonization (ICH). ICH guidelines define the main recommendations that pharmaceutical and biopharmaceutical companies follow. For example, the framework of Quality by Design (QbD) was presented by ICH guidelines (ICH Q8, 2009). QbD promotes the need of building the quality into the drug product through a systematic, scientific and risk-based approach since the early stages of product conception and development. Differently from quality by testing, which is oriented to determine the product quality in an a-posteriori fashion, QbD promotes an enhanced understanding of the product and process, the in-depth analysis of the relationship between raw material and process parameters and the CQAs, and all the sources of variability affecting the product quality (Destro & Barolo, 2022; Rathore, 2014; Rathore & Winkle, 2009). Accordingly, following QbD a product is initially formulated to meet the desired clinical performance, while the process is designed to consistently deliver a product that meets the necessary quality attributes (Destro & Barolo, 2022; Rathore & Winkle, 2009).

The implementation of QbD goes through a sequence of 6 activities:

1) identification of the quality target product profile (QTPP): identification of the quality characteristics and their limits, ranges, and distributions that the drug must achieve to ensure the desired safety and efficacy;
2) product design space: identification of all the in-process, drug substance and drug product CQAs to guarantee a product with desired performance.
3) process design space: identification of the CPPs and critical material attributes (CMAs), and their relationship and impact on product CQAs. This activity is typically performed through DoE methodologies (Rathore & Winkle, 2009);
4) definition of the control strategy: definition of the planned set of controls that ensure the product quality through risk assessment accounting for process capability. In this step, all procedural controls, in-process controls, lot release testing and process monitoring procedures are defined;
5) process validation: evaluation of the process capability to deliver a product with the desired QTPP when operated into the design space;
6) life-cycle management: monitoring of the process to ensure stability of operation within the design space, and possible process changes to preserve process consistency.

The fulfillment of these activity is required because all the information gathered at these stages are required for filing the application to the regulatory agencies. In particular, the filing incudes: product and process design space description, control strategy plans, validation activities performed, and process monitoring equipment.

An important regulatory initiative to enable the building of quality into the product is Process Analytical Technology (PAT; Food and Drug Administration, 2004). To ensure the quality of a drug product, PAT encourage the use of methodologies from chemistry, control theory, and mathematical and statistical modeling aiming at *i*) achieving timely measurements, *ii*) process monitoring and prediction, *iii*) providing real time understanding, *iv*) accurate managing of process variability, and *v*) prediction of the final product quality form process parameters (Glassey et al., 2011; Maruthamuthu et al., 2020; Rathore, 2014)

In the biopharmaceutical industry, the application of PAT is still ongoing due to the complexity of the bioprocess and the perception that the reward for the manufacturer is not commensurate with the risk. For this reason, PAT applications are solely focused on the monitoring of the process with no application for process control (Rathore, 2014).

1.5 Mathematical modeling in Biopharma Industry 4.0

The novel technologies arising in the bioprocessing field and the increasing availability of data can be exploited thanks to digitalization tools, determining the transition to the Biopharmaceutical Industry 4.0. One important tools to support digitalization is mathematical modeling (Destro & Barolo, 2022; Qin, 2014; Reis & Gins, 2017; Sansana et al., 2021).

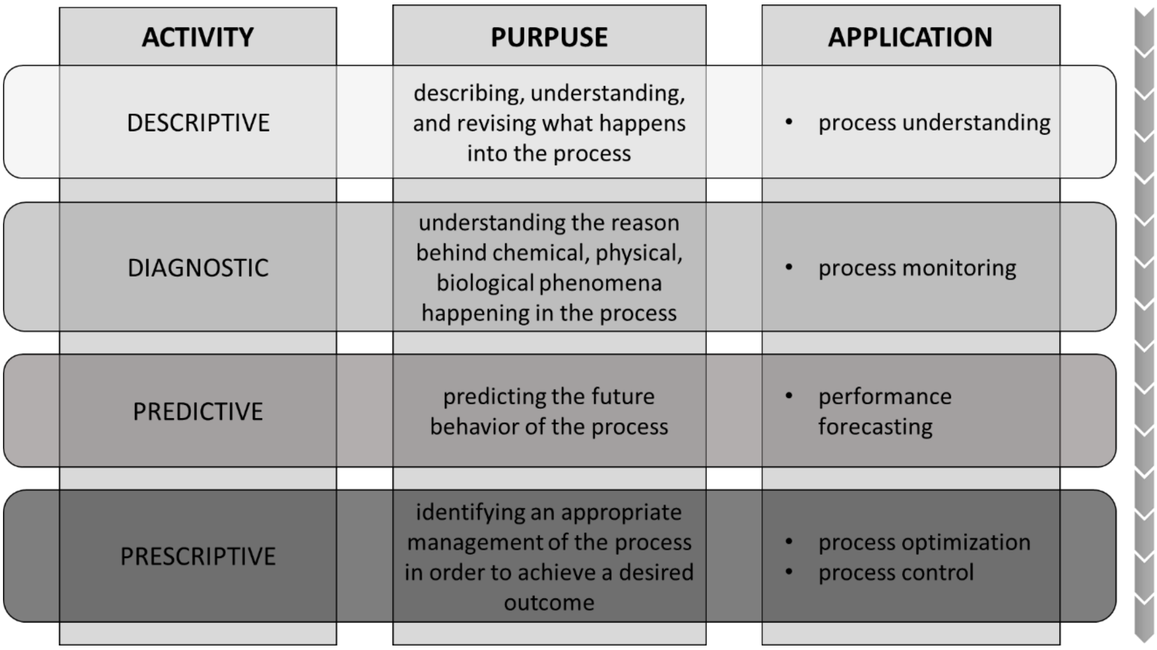
Mathematical modeling can be classified according to three criteria:

- model *type*;
- available *data* (Section 1.3.3);
- model *scope* (i.e., the final purpose of model implementation).

The model *type* describes the amount of knowledge embedded within the model. Models can be classified as data-driven (also known as black-box), first principles (knowledge driven or mechanistic models, also known as white-box), and hybrid (also known as gray-box).

Data-driven models do not require a-priori knowledge of the physical mechanisms behind the system, but rely on parametric equations for prediction, classification, and unsupervised tasks based on a given set of data. First principles models, instead, are based on the fundamental knowledge of the chemical, physical and biological mechanisms behind the system, which are described in mathematical terms by systems of equations (i.e., mass and energy balances). Hybrid models take advantage of both first principles models for the accurate description of the known physical mechanisms and data-driven models for learning the unknown complex variability of the system under study.

In the biopharmaceutical industry different types of models are used to deal with specific data types (Section 1.3.3). In particular, process data are typically analyzed through first principles, hybrid, and data-driven models, depending on *data* complexity and model scope. Biological data are typically analyzed with data-driven or first principles models, while the development of hybrid models solely based on biological data is still an open issue (Teixeira, Carinhas, et al., 2007).



ACTIVITY	PURPUSE	APPLICATION
DESCRIPTIVE	describing, understanding, and revising what happens into the process	• process understanding
DIAGNOSTIC	understanding the reason behind chemical, physical, biological phenomena happening in the process	• process monitoring
PREDICTIVE	predicting the future behavior of the process	• performance forecasting
PRESCRIPTIVE	identifying an appropriate management of the process in order to achieve a desired outcome	• process optimization • process control

Figure 1.6 *Modeling activities with their purpose and applications, organized with an increasing complexity and business value.*

The model *scope* describes the purpose of implementation and the application of the models. Modeling activities can be divided in four categories based on the purpose of the implementation, their complexity, and the respective business value (Figure 1.6):

- descriptive activities: describing, understanding, and revising what happens into the biopharmaceutical process;
- diagnostic activities: understanding the reason behind chemical, physical, biological phenomena happening in the biopharmaceutical process;
- predictive activities: forecasting the future behavior of the biopharmaceutical process;
- prescriptive activities: identifying an appropriate management of the biopharmaceutical process in order to achieve a desired outcome.

Within this framework, modeling activities in the biopharmaceutical industry are divided according to the specific application (Figure 1.6):

- process understanding: uncovering the mechanisms behind the process, the relationship between CPPs, CMAs and CQAs and their importance through exploratory analysis, information mining, variability analysis (Yu et al., 2014);
- process monitoring: determining whether a process is operating under normal (standard) conditions and diagnosing the reasons of any deviation from that (Destro & Barolo, 2022);
- performance forecasting: predicting the process performance or CQAs from CPPs or CMAs, classifying the process outcomes (e.g., in/out of specifications, high/low cell productivity), and forecasting the culture status;

- process optimization: identifying the best set of CPPs to enhance desired CQAs, defining the best culture media and feeding strategy, and proposing genetic improvement of host cell performance;
- process control: developing strategies that automatically adjust process parameters to maintain the process output within a desired range (Q13 - Continuous Manufacturing of Drug Substances and Drug Products, 2021; Sinner et al., 2020).

1.5.1 Data-driven modeling

Data-driven models are widely applied in the biopharmaceutical industry because of their fast development and implementation, their capability of handling problems with high complexity and dimensionality (Badr & Sugiyama, 2020) and their ability to deal with the high intrinsic variability of the living materials.

1.5.1.1 Descriptive activities

Process data

Descriptive activities in biopharmaceutical process development are mainly oriented to process understanding. In this context, data-driven models are typically applied to process data for:

- the identification of the important CPPs, CMAs and their inter-relationship;
- the understanding of the relationship between CPPs or CMAs and CQAs;
- the study of process and CQAs variations at different scales.

Among the data-driven techniques, multivariate (MV) methods, such as Principal Component analysis (PCA; Wold et al., 1987) and Partial Least-Squares (PLS; Wold et al., 2001), are the preferred strategies for process understanding based on process data mainly due to their interpretability and ease of use.

The main technique to study the inter-relationship in CPPs and CMAs and identify the most important ones is PCA. In fact, PCA applied on CPPs averaged over the culture time identified the main effect of pH and DO on batch variability (Sokolov et al., 2015), while PCA found how amino acid concentrations change with different pH and DO, when used to correlate process measurements and amino acid composition data (Green & Glassey, 2015). PCA was even able to follow the evolution of CQAs towards the desired optima in experiments performed at different scales (Sokolov et al., 2018).

When dealing with the understanding of the relationship between CPPs or CMAs and CQAs both PCA and PLS are typically used. For example, PCA identified the CPPs time evolution leading to the desired CQAs and early inferred the CQAs based on score trajectory in an evolving fashion (Facco et al., 2020), while when coupled with decision trees was applied during process design to select continuous and categorical CPPs and CMAs leading to the improvement of multiple CQAs (Sokolov et al., 2017). PLS, instead, is the preferred choice when the study of the relationship between CPPs and CQAs is associated with the identification of the important factors for process variability. For example, when applied to 12 CPPs PLS identified the large correlation of VCC, glutamate, glutamine and lactate with product titer (Sokolov et al., 2015), while in the cell-free production of mAbs PLS identified the CPPs affecting the process yield (Duran-Villalobos et al., 2021). Similarly, PLS combined with a genetic algorithm identified *i*) the impact of media factors on fragments quantity and of process variables (i.e., temperature and pH shift, and nitrogen flow) on product titer and glycoforms (Sokolov et al., 2017), and *ii*) studied the relationship between the dynamic changes in process variables and CQAs at AMBR15^{TM} scale (Sokolov et al., 2018).

Finally, the study of process and CQA variation across scales is typically tackled with PCA, but also PLS is used with lower frequency. For example, PCA was used to guide the selection of clones and media factors in the execution of the experimental campaign at different scales (Sokolov et al., 2018), such as deep well plates, AMBR15^{TM}, 3.5L and 300L bioreactors. Furthermore, similarity scores based on PCA were used to assess the similarity between variables and batches at different static, shaken and stirred process scales, leading to a better understanding of how cell lines may respond differently to the scale-up (Facco et al., 2020). In the regards of studying the process at different scales, both PLS Discriminant Analysis (PLS-DA; Barker and Rayens, 2003) and Orthogonal PLS-DA (OPLS-DA; Trygg and Wold, 2002) were used to study the similarity of scaled-down models (shake flasks and 3L bioreactor) with a 15000L production scale bioreactor, and improved the 3L scaled-down model by acting on the aeration rate which caused different behaviors of pH and pCO_2 (Ahuja et al., 2015).

Biological data

In the context of descriptive activities in biopharmaceutical processes, especially process understanding, data-driven methods have been typically applied to biological data to:

- better understand the metabolic state of cells associated with different characteristics, culture conditions, and CQAs;
- identification of cellular features (i.e., biomarkers) associated with cell characteristics and CQAs.

Biological data are typically large, complex, extremely correlated and often characterized by nonlinearities. Despite their linearity, MV methods, especially PCA and PLS, are the preferred choice for process understanding because their ease of interpretation and capability of extracting valuable information.

When it comes to metabolomic data for a better understanding of the metabolic state of cells, PCA is the most commonly used MV methods. For example, PCA on intracellular GC-MS metabolomics of mammalian cells differentiated cells based on their age, cell source and reactor scale (Chrysanthopoulos et al., 2010), while in CHO cell lines identified a metabolic shift moving form growth to production phase (Sumit et al., 2019). PCA also analyzed the intracellular metabolomic profiles change associated with the steady state VCC value in perfusion cell cultures, showing that nucleotide sugar donors, coenzyme A, and metabolites precursors of phospholipids concentrations varies between VCCs (Karst et al., 2017). In regard to the understanding of cell metabolic state, PCA has also been used to better understand the impact of different culture media and its time change on cell metabolism. In fact, PCA identified differences in the metabolomic profiles of CHO cells cultured in different media and how these differences correlate to different growth rates and VCC (Dietmair, Hodson, Quek, Timmins, Chrysanthopoulos, et al., 2012). Furthermore, PCA showed that media has a strong impact on growth and productivity in CHO cell lines (Mohmad-Saberi et al., 2013), and it was able to study the time changes of extracellular metabolites in CHO cultures finding metabolites that are accumulating, such as amino acid derivates, and depleting, likely media components (William P K Chong et al., 2009). Other applications of PCA on biological data concern the study of the different amino acid content in different mammalian cell lines (Selvarasu et al., 2012), the identification of nutrient supplement enhancing mammalian cell productivity (Richardson et al., 2015), and to assess the effect of processing conditions on proteomics data (Strasser et al., 2021).

Despite the metabolomic data are often used for process understanding and correlated to culture (i.e., process) conditions, a comprehensive framework to integrate process and biological information is still missing, and no precise indications are typically followed.

The identification of cellular biomarkers is typically performed through PLS, mainly because of its capability of identifying features in the biological data associated with desired characteristics or CQAs. For example, PLS was used to identify metabolites associated to growth rate (Dietmair, Hodson, Quek, Timmins, Chrysanthopoulos, et al., 2012), and with an enhancing and inhibiting effect on product titer, allowing to adjust media composition with those metabolites achieving enhanced antibody productivity (Morris et al., 2020). Similarly, OPLS-DA on LC-MS intracellular metabolomics identified metabolites changing with each culture phase, but it did not find process scale specific metabolic states (Vodopivec et al., 2019), and identified extracellular metabolites differentiating wildtype and transfected cells in the study of mutations that can lead to protein accumulation potentially causing Alzheimer (Chang et al., 2015). In this case, the identified metabolites were used to assess the effect of different treatments in reducing protein accumulation.

In this regard, the enhanced understanding achieved through MV modeling of biological data is typically used to characterize the process and to improve the design of media and processes. The direct exploitation of the wealth of biological information to aid selection of high performing cell lines is still missing. Furthermore, as can be noticed, the use of dynamic metabolomic data for process understanding is still an open issue.

When dealing with the identification of biological features associated with cell characteristics or CQAs, appropriate strategies, named feature selection methods, must be coupled to data-driven models, because biological datasets are often characterized by an extremely large number of variables. For example, in metabolomics, data on several thousand metabolites are often available and this can be repeated in several time points along the culture, leaving with several tens of thousands of metabolites.

Feature selection strategies change with the selected data-driven method and improve both model interpretability and performance. For example, PLS-based feature selection methods are divided into three families: *i*) filter methods, *ii*) wrapper methods, and *iii*) embedded methods. In filter methods, the outputs of the PLS are directly used to identify the most important features, while in wrapper methods the important features identified by filter methods iteratively give feedback to re-fit the PLS model. Finally, in embedded methods the feature selection is an integrated part of the PLS algorithm. A shortlist of different PLS-based feature selection methods is reported in Table 1.3, while a comprehensive explanation of the PLS-based feature selection strategy is reported in Mehmood et al. (2012).

Table 1.3 *Applications of PLS-based feature selection methods.*

Reference	Strategy	Method	Application
Wiklund et al. (2008)	filter	Covariance + correlation (S-plot)	Identification of metabolites differentiating transgenic and wild type polar trees
Bryan et al. (2008)	filter	variable importance in projection (VIP) + Pearson’s correlation	Automated tool for feature selection in metabolomic data
Clarke et al. (2011)	wrapper	Iterative backward elimination of uninformative variables	Identification of genes associated to cell productivity from CHO transcriptomic data through PLS
Dietmair et al. (2012)	filter	VIP	Identification of metabolites associated with growth rate in CHO cells
Chong et al. (2012)	filter	VIP + Pearson’s correlation + *t*-test	Identification of metabolites characterizing high- and low-productive cell lines
Afanador et al. (2013)	filter	VIP bootstrap	Tool for robust feature selection
Pujos-Guillot et al. (2013)	filter	VIP	Identification of metabolites characterizing human intake of citrus fruit for biomedical applications
Chang et al. (2015)	filter	VIP	Identification of metabolites differentiating cell lines with a mutation inducing protein accumulation potentially leading to Alzheimer
Morris et al. (2020)	filter	VIP	Identification of metabolites enhancing or inhibiting product titer
Zürcher et al. (2020)	filter	VIP	Identification of metabolites associated to product quality in CHO cell cultures

Other feature selection strategies were applied with other data-driven methods, such as elastic-net (Badsha et al., 2016), univariate filtering (correlation or mutual information criterion) coupled with embedded methods in support vector machine (SVM), random forest or ANOVA (Grissa et al., 2016), permutation on random forest or gradient boosting (Jr, 2014), and genetic algorithm prior MV modeling (Davis et al., 2006). In other cases, a graph based variable selection coupled with random forest and SVM was used to overcome the limitation of multiple annotations and missing values in metabolomic data (Cai et al., 2017).

In many cases, data-driven methods have been applied to biological data thanks to third party software which allows less experienced practitioners to perform such analyses but lacks customizability. Specifically, SIMCA-P (Sartorius, Goettingen, Germany) was often used to build the models and perform feature selection, MetaboAnalyst to perform exploratory analysis, classification, feature selection and functional analysis of metabolites (J. Chong et al., 2019), and Matlab® (MathWorks, Natick, MA, US) or R libraries, such as: PLS_Toolbox (Eigenvector Research Inc, Wenatchee, WA, US) or Classification and Regression Training package. The adoption of easy-to-use software, especially in PLS-based analysis, leads to the use of very simple filter methods (i.e., VIP) or feature selection which are often less performing than wrapper or embedded methods.

1.5.1.2 Diagnostic activities

Process data

Diagnostic activities in biopharmaceutical processes development are mainly oriented to process monitoring. In this context, data-driven models are typically applied to process data for:

- fault detection and diagnosis
- identification of measured variables that can be used for monitoring CQAs.

In process monitoring for the biopharmaceutical industry, MV methods are the only ones used because of the availability of diagnostics which allow to build effective charts for monitoring and fault detection with moderate effort.

Concerning fault detection and diagnosis, both PCA and PLS models, have been used to exploit their monitoring charts as hypothesis testing tools. For examples, PCA-based monitoring charts have been used with online measurements to monitor in real-time the deviations from the Normal Operating Conditions (NOC) and identify the faulty batches in a penicillin fermentation process (Goldrick et al., 2019). In a similar process, PCA-based monitoring charts exploited online measurements and a hybrid strategy with Extended Kalman Filter (EKF) based on a simple first principles model for monitoring and fault detection (Destro et al., 2020). The addition of the estimated states and EKF parameters provided meaningful information about phenomena involved in the fault and allowed to diagnose its root causes, much better than using online measurement alone. In regard to monitoring and fault detection, PLS-based diagnostics and monitoring charts allowed the real time monitoring of the final product concentration in a fed-batch fermentation from batch-wise unfolded process measurement and the identification of faulty batches and their root causes (Gregersen & Jørgensen, 1999). Similarly, Multiway PLS (MPLS; Nomikos and MacGregor, 1995) was applied to the production of mAbs in cell-free environments for the anticipated monitoring of the final yield, allowing also to identify faulty batches with unusual changes in pH and temperature (Duran-Villalobos et al., 2021). However, biopharmaceutical process development is characterized by the availability of few process batches, especially at pilot scale where only a handful of experimental runs are available due to their high cost and long duration (F. Li et al., 2010). To deal with the problem of data scarcity, PLS was coupled with a data augmentation strategy based on high frequency resampling and Gaussian Process for fault detection (Tulsyan et al., 2018, 2019). In this case, 50 *in silico* batches originated starting from 2 or 3 available process batches were used to build a PLS based monitoring strategy, which correctly identified faults in new batches through monitoring charts. The use of data augmentation strategies, as in this case, would greatly benefit various stages of biopharmaceutical process development, especially the ones involving few available runs, to reduce the experimental burden and accelerate the timelines. Unfortunately, the application of data augmentation methods, especially model-based ones, in the biopharmaceutical industry is still an open issue.

In regard to the identification of process variable for the monitoring of CQAs, PLS is the main technique. In particular, PLS based on DoE data was used to identify the joint importance of temperature, pH and pCO_2 in monitoring the product quality in mammalian cell cultures (Goldrick et al., 2017).

Biological data

In the context of diagnostic activities in biopharmaceutical processes, biological data are not often exploited for process monitoring due to their off-line measurement nature. However, data-driven methods, especially MV ones, which are the preferred choice also in this case, have been applied to:

- monitor CQAs based on biological features;
- build soft sensors.

In the context of monitoring CQA, PCA is the most used method, which has been used with metabolomic data to identify indicators of good cell health (such as asparagine) and indicators of apoptosis (such as ornithine and lysine) that help to monitor the state of the culture (Mohmad-Saberi et al., 2013). Furthermore, PCA applied to intracellular metabolomic data identified specific metabolic changes with each culture phase and could be used to monitor cell physiology and phase transition in perfusion cell cultures (Vernardis et al., 2013).

In this regard, the exploitation of time-varying biological data in a quasi-real time manner for the monitoring of cell behavior and CQAs is still missing and would probably allow to early infer any variations or anomalies in cell cultures.

Concerning the implementation of soft sensors for process monitoring, PLS is the typical used methodology. In particular, Raman spectroscopy based soft sensors thorough PLS allowed to implement a real time monitoring system for product glycosylation, estimating the concentration of glycosylated and non-glycosylated antibodies (M.-Y. Li et al., 2018), and monitor the penicillin yield in a simulated fermentation process (Goldrick et al., 2019).

1.5.1.3 Predictive activities

Process data

Predictive activities in biopharmaceutical processes development are mainly oriented to performance forecasting. Data-driven models are typically applied to process data for the prediction of CQAs from CPPs.

In this context, a large variety of data-driven, from MV to deep learning methods, have been applied. Data-driven methods are categorized in *i*) linear, and *ii*) non-linear. In linear models, such as MV ones, inputs are linearly combined with the model parameters to represent the output, making the model easily interpretable. In non-linear model, such as neural networks, instead, the inputs are combined in a non-linear fashion to represent the output. This allows models to represent more complex relationship between inputs and outputs, but strongly limits their interpretability. Several applications of data-driven methods in the biopharmaceutical industry has been reviewed by Rathore et al. (2022).

Concerning the prediction of CQAs, PLS regression is the most applied MV method because it allows to easily predict multiple responses and identify at the same time the most influential factors affecting the prediction among regressors. For example, PLS was used to predict the VCC at each process scale and it was compared to a Joint Y-PLS (JY-PLS; Garcıa Munoz et al., 2005) that used data from multiple scales to predict the VCC (Facco et al., 2020). In this case, PLS and JY-PLS preformed similarly, but JY-PLS outperforms PLS in predicting the VCC for scales where only few calibration data are available. Applied to amino acid concentrations, PLS accurately predicted CQAs, such as product titer and glycan profile (Green & Glassey, 2015). Furthermore, PLS coupled with a genetic algorithm accurately predicted CQAs from media composition and the dynamic evolution of process variables and identified the media or process variables with the largest influence on prediction (Sokolov et al., 2018). Similarly, CQAs, such as product titer and glycoforms, were predicted from media factors and process history at AMBR15^{TM} scale using an Evolving PLS (Ramaker et al., 2005) coupled with a genetic algorithm (Sokolov et al., 2017). This strategy allowed to identify how CQAs predictability changes in time and to understand media and process factors affecting the studied CQAs. Finally, PLS, applied to mammalian cell cultures for predicting the time evolution of the main process variables (Narayanan et al., 2019), showed lower prediction accuracy than more complex hybrid models (which will be covered in Section 1.5.2).

In recent years, more advanced non-linear machine learning methodologies, such as SVM, and deep learning methods, such as artificial neural networks (ANN), have been used for their predictive power despite their lack of interpretability. For example, the best model structure to predict the mAbs concentration at harvest was identified by means of SVM and five feature selection methods to rank variables (Gangadharan et al., 2021). Furthermore, ANNs predicted better than the fully mechanistic model the glycan distribution in CHO cell cultures from the nucleotide sugar donors which are calculated by a first principles model of cell metabolism (Kotidis & Kontoravdi, 2020). In this case, the glycan distribution was predicted form nucleotide sugar donors and enzyme concentration by means of ANN, allowing to study the effect of enzyme concentration changes achieved through genetic engineering on the glycan distribution. In this regard, the application of SVM and ANN on bioprocesses is still an open research topic mainly because of the lower interpretability than MV models, and the reduced generalizability with the (relatively limited) amount of data typically available in biopharmaceutical development. Here, the introduction of data augmentation techniques can strongly support the use of such powerful methodologies, as was done in other fields such as image processing.

Biological data

In the context of predictive activities in biopharmaceutical processes, even when dealing with performance forecasting based on biological data, data-driven methods are the preferred choice to:

- predict CQAs from biological features (i.e., metabolism or genotype);
- discriminate between different cell conditions.

A large variety of data-driven methods have been applied for performance forecasting, but MV methods are preferred because of their simplicity and capability to handle the large, complex and correlated nature of biological datasets.

Considering the prediction of CQAs from biological features, PLS is typically used. For example, MPLS on dynamic metabolomic data was used to predict the phenylalanine yield in *E. Coli* cultures, also allowing to understand how the metabolites associated to yield change in time by inspecting the PLS regression coefficients (Rubingh et al., 2009). Additionally, extracellular metabolites combined with process data allowed to predict the glycan profile in CHO cell cultures (Zürcher et al., 2020). In this case, the VIP index allowed to maximize validation performance and identify the metabolic features associated to the product quality prediction. Furthermore, PLS was the first model built on gene expression data to predict CHO specific productivity (Clarke et al., 2011). In this case, the model correctly predicted specific productivity in new experiments and identified about 300 genes associated to productivity through a multi-loop iterative backward elimination of uninformative variables.

Moving to the discrimination of different cell conditions, linear discrimination methodologies, such as PLS-DA, are typically adopted. They combine both accurate classification performance and interpretability of the results. For example, PLS-DA has been used on MS metabolomic data collected from deep well plate cultures to discriminate CHO cells according to their productivity level measured in 10L bioreactors (Povey et al., 2014). This example showed that the metabolic features are intrinsic characteristics of the host cell and can be predictive of a phenotype even at a different process scale, and that this can be done even for cell lines producing different antibodies. Similarly, OPLS-DA was also used with CHO LC-MS intracellular metabolomic data to discriminate between low and high productive cell lines, identifying at the same time the fingerprint of high productive cell lines in increased level of electron carriers and nucleotide sugar donors through the combination of VIP, Pearson's correlation and t-test (William Pooi Kat Chong et al., 2012).

In regard to the prediction of CQAs and the discrimination of cell characteristics, dynamic biological data are often available, but, due to their limited number, the dynamics of biological information is often disregarded. Considering the dynamics of the biological data would allow to study how the biological features associated to prediction (or discrimination) change in time, in such a way as to improve the understanding of the system. Furthermore, the exploitation of biological data to directly support the selection of promising cell lines is still an open issue and it would greatly improve the mAb development process.

More recently, advanced data-driven methods, such as ANN, have been applied to biological data, thanks to the reduced cost of analytical analysis. For example, cell phenotype was predicted from genotype in *E. Coli* cultures by means of ANN (Guo et al., 2017). In this three-layer ANN, each layer corresponded to a biological level (genotype, proteins, phenotype) and the connection between neurons reflected the known gene-proteins and proteins-phenotype relationships. A two-step training was used, consisting in an unsupervised training with an autoencoder to reconstruct the transcriptomic data, followed by a fine-training using genotype-phenotype data. The results showed that transcriptomic data predicted cell phenotype in an accurate manner, even without any biological knowledge embedded into the network connections; however, an initial unsupervised training was necessary to learn robust features and weights.

In this regard, application of advanced data-driven methods, such as ANN, is still limited in the biopharmaceutical sector mainly because of the fact that the very high number of biological variables is typically coupled with a limited number of experiments, which limits the generalizability of the models.

1.5.2 Hybrid modeling

Hybrid models combine first principles and data-driven models and take advantage of both modeling strategies, providing good generalization capabilities, learning complex relationships, and generally performing better than first principles or data-driven models alone. For example, in a polymethyl methacrylate polymerization reactor hybrid models described the complex kinetics resulting more accurate than the first principles model (Ghosh et al., 2019). In mammalian cell cultures, hybrid models achieved better prediction performance than a data-driven PLS model, even providing better extrapolation capability (Narayanan et al., 2019).

In hybrid models, the most commonly used first principles models are material and energy balances, thermodynamics and kinetic equations. First principles models are very robust and can be generalized very well, because they are based on a detailed description of the physical and chemical phenomena taking place into the system. However, they require long time to be developed. Furthermore, in complex systems the mechanistic knowledge is not always that accurate or available (S. Yang et al., 2020). Note that the first principles models typically used in hybrid modeling are not similar to the first principles models of cell metabolism explained in this Dissertation (Section 1.5.3).

The data-driven part of hybrid models compensates for the lack of mechanistic knowledge, since it learns the complex and unknown relationships from data, effectively managing the wide variability which is typical of the biopharmaceutical applications. The main data-driven methods used in hybrid models are artificial neural networks (Narayanan et al., 2019; Oliveira, 2004; Psichogios & Ungar, 1992), recurrent neural networks (Smiatek et al., 2021), PLS (Carvalho et al., 2022), SVM, and extended Kalman filter (Destro et al., 2020; Ghosh et al., 2019).

Hybrid models have been extensively reviewed in previous works (Sansana et al., 2021; von Stosch et al., 2014; S. Yang et al., 2020). Here, we summarize some details on their most important features, such as model structure, training procedure, and degree of hybridization.

Hybrid models can have two types of structure (Figure 1.7; Sansana et al., 2021; von Stosch et al., 2014; S. Yang et al., 2020):

- parallel (Figure 1.7a): when a sufficiently good first principles model is available, the data-driven model is used to correct the outputs and to improve estimation of the first principles model by combining the outputs of the two models. In this case, the data-driven part receives the same inputs as the first principles model and describes the mismatch between the first principles model and the experimental data;
- serial: when the first principles model describes the conservation laws, the data-driven part can be used to represent the underlying kinetic and transport terms, which are sometimes unknown or extremely complex. In this case, the data-driven part receives the input from the environment and feeds its output to the first principles model which calculates the final outputs of the system (Figure 1.7b, serial structure). The inverse structure can be found as well (Figure 1.7c, serial structure).

Some attempts have been made to compare the performance of the different model structures, but the outcomes result to be case dependent. In fact, it was observed that the structure minimizing the validation error for pharmaceutical unit operations such as feeder and blender is the parallel one, while the serial is more appropriate in the case of a CSTR reactor with complex kinetics (Y. Chen & Ierapetritou, 2020). Therefore, the appropriate structure must be selected according to the characteristics of: the problem under study, the available data, and the available first principles model. However, it must be considered that, especially when applied to mammalian cell cultures, the parallel structure seems to be more accurate, but the serial one is easier to build (Vande Wouwer et al., 2004).

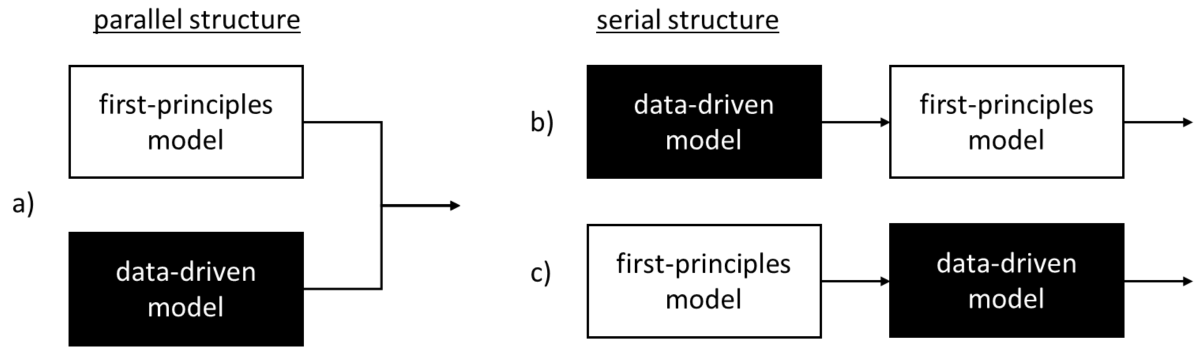


Figure 1.7 *Hybrid model structures.*

Due to the combination of fist principles and data-driven models, the training of hybrid model is not always straightforward, and appropriate techniques are required. Two main strategies are used to train the data-driven part of the hybrid model (Sansana et al., 2021; von Stosch et al., 2014; S. Yang et al., 2020):

- direct approach;
- indirect approach.

In the direct approach, the standard training of the data-driven methods (e.g., Section 2.2.1 for ANN) is performed using the known inputs-outputs pairs from experimental data. This approach can be used either in the parallel structure, where the mismatch between the first principles model and the experimental data can be easily calculated, or in the serial structure, by calculating the data-driven outputs from the experimental data using the first principles model. In the latter, the calculated outputs might be biased or imprecise and can pass this bias to the data-driven model (von Stosch et al., 2014).

In the indirect approach, the data-driven model is trained indirectly through the first principles one using the experimental data. In this case, the sensitivity equation method (Galvanauskas & Simutis, 2007; Oliveira, 2004) is used to backpropagate the effect that the data-driven part has on the hybrid model outputs. In indirect training any imprecision or bias is absorbed by the data-driven outputs to maximize the representation of the experimental data (von Stosch et al., 2014). Finally, if the first principles model contains parameters to be estimated from experimental data, a multi-step procedure has been proposed to reliably estimate the parameters of both model compartments (A. Yang et al., 2011).

One last aspect to consider in serial hybrid models is the degree of hybridization (Narayanan, Luna, et al., 2021), which accounts for the amount of information that is described by the first principles model. Accordingly, a fully data-driven model will have a 0% degree of hybridization, while a pure first principles model will have a 100% degree of hybridization. Increasing the degree of hybridization improves the prediction performance and the extrapolation capability of the model, but high degrees of hybridization are not always beneficial for model performance. This is mainly due to the complexity of the model which requires a large number of training experiments to achieve robust performance. Accordingly, a tradeoff between the model complexity and the number of experiments for model training is required. However, a structured procedure to understand the optimal degree of hybridization is still missing (Narayanan, Luna, et al., 2021).

In biopharmaceutical processes, the interest in hybrid models started in the '90s, with the first hybrid model for a fed-batch bioreactor developed in 1992, combining a first principles model and an ANN (Psichogios & Ungar, 1992). Other examples follow shortly after, with a hybrid model to simulate the production of a recombinant protein (Dors et al., 1996).

In the last two decades, applications of hybrid models to the biopharmaceutical industry have considerably grown in number, and field of applications, such as bacteria and yeast fermentations (Ferreira et al., 2014; Oliveira, 2004) , mammalian cell cultures (Smiatek et al., 2021; Vande Wouwer et al., 2004), and even in the downstream processing (Narayanan, Seidler, et al., 2021). However, research on this topic has to progress to allow a consistent applicability of this technology in the biopharmaceutical industry.

In this context, hybrid models are typically applied to process data, but there is a growing interest in hybrid models combining process and biological data (Teixeira, Carinhas, et al., 2007), which is still an open problem. The main applications of hybrid models concern process optimization, such as the optimization of the media and feeding schedule, but they can still be used for process understanding and monitoring, performance forecasting, and process control.

1.5.2.1 Descriptive activities

Descriptive activities in biopharmaceutical processes mainly concern the understanding of the process in the operating region, while limiting the number of exploratory experiments required to achieve a good understanding as much as possible. In this context, hybrid models have been mainly used to assess their applicability for process understanding compared to traditional experimental methods and DoE strategies.

In this regard, the advantages of hybrid models rely on a faster, cheaper and better understanding than the traditional methodologies used in the field. For example, a hybrid model trained on DoE experiments from a *E. Coli* fermentation process was able to correctly characterize biomass and product formation at different processing conditions of temperature, feed rate, and pH, allowing to assess the effect of temporal variations in the processing condition on CQAs (von Stosch et al., 2016). In a similar context, hybrid models achieved a better process characterization than the traditional DoE-based response surface methodology (RSM), providing better characterization of product and biomass concentration and better representation of the process variables time profiles (Bayer, Stosch, et al., 2020). Hybrid models trained on intensified DoE[1] data (5 experimental runs) well described the entire region spanned by a full factorial DoE experimental campaign (9 experimental runs), showing that accurate process understanding is achieved using a smaller number of experiments (von Stosch et al., 2016). Furthermore, a hybrid model trained on intensified DoE experiments achieved prediction performance which are similar to those of a model trained on a full-factorial DoE, providing a 66% reduction in the required number of experiments, thus accelerating bioprocess characterization (Bayer, Striedner, et al., 2020).

In regards of bioprocess understanding, research is still needed for a consistent applicability of hybrid models, especially to determine the appropriate degree of hybridization. Furthermore, the application of hybrid model for mammalian cell culture understanding is still an open issue.

1.5.2.2 Diagnostic activities

For what concerns the diagnostic activities in the biopharmaceutical industry, data-driven models are sometimes not sufficient in modeling the temporal evolution of noisy data due to the lack of mechanistic knowledge, while process data alone are often not sufficient to establish accurate process monitoring. Hybrid models can represent a viable alternative for:

- soft sensing;
- fault detection and diagnosis.

In regard to the implementation of soft sensors, hybrid models combined with an Extended Kalman Filter were used as a soft sensors for the online monitoring of the glucose concentration (Narayanan et al., 2020). In this case, the good prediction performance of the hybrid model and the noise rejection capability of the Kalman filter allowed to correctly assess the glucose addition requirements of the cell culture to avoid CHO cell starvation.

For fault detection and diagnosis, a hybrid model based on Extended Kalman filter was used in a penicillin fermentation process (Destro et al., 2020). This demonstrated that hybrid models which acquire online data and work in combination with multivariate statistical methods improved the identification of faulty batches and their root causes than using solely online data. Research in bioprocess monitoring through hybrid models is still ongoing, and applications in mammalian cell cultures are very limited mainly due to the scarce availability of online data.

1.5.2.3 Predictive activities

In the context of predictive activities, accurate methodologies to forecast process performance are welcomed to further support various stages of bioprocess development and manufacturing. Hybrid models are a viable alternative to purely data-driven ones for the prediction of CQAs, such as biomass or product concentration, and the temporal profiles of process variables.

[1] In intensified DoE the factors are varied according to a classical DoE in a fixed number of stages during each experiment, to reduce the total number of required experiments.

In this regard, hybrid models trained on DoE and intensified DoE experiments were used in *E. Coli* fermentation processes to predict:

- biomass and product concentration (Bayer, Stosch, et al., 2020; Bayer, Striedner, et al., 2020; von Stosch & Willis, 2017);
- time profiles of process variables (Simutis & Lübbert, 2017; von Stosch et al., 2016).

In CHO cultures, instead, hybrid models were used to predict *i*) VCC and antibody titer, providing better prediction accuracy than a traditional data-driven approach (Narayanan et al., 2020), *ii*) the temporal evolution of nutrients (Narayanan et al., 2020), such as glucose, and even *iii*) all culture variables (Narayanan et al., 2019). In other cases, VCC, glucose, and antibody concentration were accurately predicted from data on medium composition through a parallel hybrid model combining a first principles model and PLS, which performed better that the first principles model alone and could be used for media optimization (Carvalho et al., 2022).

Regarding performance forecasting, the main researches focused on proving the superior performance of hybrid models than purely first principles or data-driven models, while the application of these predictions to solve important bioprocess problems is still underexplored.

1.5.2.4 Prescriptive activities

Prescriptive activities in biopharmaceutical processes are mainly oriented to process optimization, such as the identification of the best feeding schedule, while bioprocess control is still underexplored. In this regard, prescriptive activities are typically based on experimentation. However, strategies to perform these tasks in a virtual fashion could be extremely beneficial to accelerate the experimentation and reduce experimental burden.

Hybrid models can be used as virtual copies of the real process to perform iterative techniques aimed at finding the media or feeding strategy that maximize a desired CQA (Dors et al., 1996). For example, a batch-to-batch optimization strategy was used to identify the feeding strategy giving improved production (Ferreira et al., 2014; Teixeira et al., 2006). In *Pichia Pastoris* fermentation, a hybrid model trained on an initial set of 5 exploratory experiments was used to optimize the feeding strategy. Then, the optimal experiment is iteratively executed on the process and used to retrain the hybrid model, until convergence between the predicted and measured protein concentration. In this case, 5 iterations were needed to identify the optimal feeding strategy (Ferreira et al., 2014). Batch-to-batch optimization was also used in three different fermentation processes, where after a single initial experiment the feeding schedule was optimized in 5 iterations (Teixeira et al., 2006). A similar strategy was used in *E. Coli* fermentation processes, to identify the optimal CPPs with the smallest number of experimental runs (Bayer et al., 2021). Here, a preliminary fractional-factorial DoE for planning 5 exploratory runs allowed to identify the process optimum with only 4 additional sequential experimental runs, while the exploratory runs planned through an intensified DoE did not allow the hybrid model to identify the process optimum even with a large number of additional experiments. In other cases, a bootstrap aggregated method was used to improve the prediction of the hybrid model, allowing to optimize the CPPs by correctly representing the response surface of the system (J. Pinto et al., 2019). In this method, hybrid models were trained with different splitting of the same training dataset, and the predictions of all these models were averaged to improve the robustness and the accuracy of predicted profiles. This strategy showed better accuracy that a single hybrid model in 3 datasets simulated running different DoE campaigns.

In *E. Coli* fermentations, complex dynamic metabolic models (explained in Section 1.5.3) were substituted by hybrid models, making them as a viable alternative for bioprocess optimization (Setoodeh et al., 2012). In this work, the data-driven section of the hybrid model was trained on data obtained from a Genome-scale Metabolic Model (GSMM) to estimate growth rate and product exchanges, which were then fed to material balances of the extracellular species. The hybrid model achieved comparable performance with the fully first principles GSMM model and was used to consistently accelerate the process optimization.

Even in mammalian cell cultures, the best feeding strategy was identified through hybrid models. For example, different hybrid model structure were compared and the best one was used to identify the optimal profile of glucose and glutamine for the mammalian cell culture (Teixeira et al., 2005). Interestingly, a first attempt to consider the metabolic network of mammalian cell lines into hybrid models was made through the use of Elementary flux Modes (Teixeira, Alves, et al., 2007). These elementary flux modes represent a pathway into the cell metabolism that connects an input (i.e., nutrient) to one or more outputs (i.e., byproducts). In this case, the nutrient profile was optimized and used to assess how the fluxes within the cell changed along the culture. In the last example, the concept of prediction risk was introduced (Teixeira et al., 2005; Teixeira, Alves, et al., 2007), as well. Specifically, it was possible to quantify the risk of using the model for predictions in a defined region of the experimental space, according to the distance to the training experiments.

Regarding the optimization of the feeding schedule, a lot of research focused on demonstrating the applicability of hybrid models for process optimization and proposed different strategies to perform this task. However, proof of their advantages over traditional experimental methods to support bioprocess development is still an open issue.

In biopharmaceutical upstream processes, control applications based on hybrid models are still underexplored, mainly due to limited availability of online data and sensors, which hinder the capability of hybrid model in establishing online process control strategies.

However, an early attempt of model based control through an hybrid model on a yeast fed-batch fermentation was found (Schubert et al., 1994). In this case, a controller based on an ANN trained on data generated by the hybrid model controlled the ethanol concentration allowing to keep a desired biomass profile. Furthermore, the ANN-based controller outperformed a traditional PID controller since it captured the complex nonlinear relationship between ethanol and biomass concentration. Accordingly, additional research is required to consolidate the application of hybrid model for bioprocess control.

1.5.3 First principles modeling of cell metabolism

The first principles modeling of cell metabolism, one of the main research areas of systems biology, is a complex task. Cell metabolism is typically modelled through:

- kinetic modeling;
- stoichiometric modeling.

Kinetic models of cell metabolism describe several metabolic mechanisms, such as thermodynamics, kinetics, enzyme regulation, and reaction stoichiometry. This is achieved through ordinary differential equations which express the metabolic flux as function of metabolite concentration, enzyme concentration and enzyme kinetic parameters (Hendry et al., 2020). To build these models, a significant amount of data carrying information on metabolomics and fluxomics is required to correctly estimate model parameters; this strongly limits the applicability of kinetic metabolic models for the description of complex organisms, such as mammalian cells. Despite that, some attempts have been made to use kinetic metabolic models in CHO cell cultures (Robitaille et al., 2015). In this work, the central pathways (i.e., glycolysis, tricarboxylic acid (TCA) cycle and pentose phosphate pathway), energy production and amino acid metabolism were modeled through material balances and Michaelis-Menten kinetic expressions for a better understanding of the host cells. In particular, it was found that a single model accurately describes cells in different phases and cultured in different media. Furthermore, in the studied cells the TCA cycle was mainly fed by amino acids instead of glucose, leading to the formation of undesired metabolites (i.e., ammonia) from the amino acid decomposition, while glucose was mainly converted in lactate. The accumulation of lactate and ammonia provided a considerable inhibition of cell growth.

Stoichiometric models rely on stoichiometric information to describe the mass balance of each metabolite at steady state through a system of linear equations. These models are based on a pseudo-steady state assumption, which assume no accumulation of metabolites through the metabolic network considering the very low velocity of biological reactions during batch and fed-batch cultures (Quek et al., 2010). Accordingly, these models are much simpler than kinetic metabolic models, but cannot capture metabolite concentrations, enzyme saturation and gene regulatory effects.

GSMMs are stoichiometric models that collect all metabolic reactions for a given organism, containing information on stoichiometry, directionality, gene-protein-reaction association of all known metabolic reactions, as well as biomass and product composition information. Gene-protein-reaction associations contain information about the relationship between a reaction, the enzyme required to catalyze that reaction, and the gene/genes that encode for a given enzyme.

An important information embedded into GSMMs is the composition of the biomass for a given organism, accounting for all the biological components required to produce biomass.

In biopharmaceutical processes, GSMMs of prokaryotic organisms, such as bacteria and yeasts, have been available for some years, while GSMMs for mammalian cell cultures have been developed recently, despite stoichiometric modeling of mammalian cells started while before. In fact, Hefzi et al. (2016) developed the first complete reconstruction of the CHO cell GSMM with a simplified product secretion in 2016, while the accurate description of antibody secretion mechanism and its energy cost have been released very recently (Gutierrez et al., 2020) and the first application to CHO industrial culture just one year prior (Calmels et al., 2019). Despite that, practical applications of GSMMs to mammalian cell cultures are still limited because of the large complexity of these models and research is needed to allow a wider applicability (Richelle et al., 2020).

In order to apply and analyze GSMMs, appropriate computational methods are required, such as flux balance analysis (FBA; Maranas and Zomorrodi, 2016). In FBA, the system of equations describing the material balances of all metabolites contained in the model is solved. To be solved, this system of equation requires the knowledge of all cell inputs and outputs, which are usually the rates at which cell carbon substrates, biomass precursors and by-products are absorbed or secreted from the extracellular environment. Despite that, the number of measured extracellular fluxes (i.e., which can be some dozens) and the total number of metabolites (i.e., which can be thousands) is smaller than the number of model reactions (i.e., which can be several thousands), making the system of material balances underdetermined. Because of that, the solution of the material balances is performed through an optimization, which requires the definition of an appropriate objective function which is typically maximized. The most common objective function involves the biomass production and assumes that the cell metabolic fluxes are distributed to maximize the production of biomass and cell growth, accordingly. This is typically verified in the exponential growth phase, but this is not always true since sometimes cells allocate resources for several different tasks (Gutierrez et al., 2020). For this reason, other objective function can be found, such as the minimization of enzyme cost, the minimization of Gibbs energy dissipation, the minimization of the reactive oxygen species, and the minimization of nicotinamide adenine dinucleotide (*NADH*) regeneration (S.-M. Schinn et al., 2021).

Despite the complexity of the formulation of the optimization problem, an infinite number of flux vectors can satisfy the steady-state material balances, which requires to define additional constraints to reduce the possible solution space. This is usually achieved by defining lower and upper bounds for the fluxes of each reaction. The values of these boundaries are typically defined through biological and thermodynamic considerations (Feist et al., 2007), and enzyme activity and turnover (Maranas & Zomorrodi, 2016). Other strategies to set additional constraints and reduce the solution space exist (Figure 1.8), such as parsimonious enzyme usage

FBA (pFBA) and carbon constraining FBA (ccFBA). In pFBA (Lewis et al., 2010), the main assumption is that cells use the minimum amount of resource (i.e., enzymes) to maximize the given objective. This translates in finding the maximum value of the objective function while minimizing the total sum of intracellular fluxes. In ccFBA (Lularevic et al., 2019), the intracellular fluxes are limited by the total amount of carbon that is taken by cells from the extracellular environment, thus limiting the total number of carbon atoms that are circling within the cell. A comparison between these methods in the prediction of biomass showed that FBA and pFBA typically achieve very similar performance, but overpredict biomass with respect to the experimental measurement, while ccFBA strongly limits the availability of resources for growth causing the underprediction of biomass (Antonakoudis et al., 2021). In this study, a nitrogen-constrained FBA was also proposed in which, similarly to ccFBA, the total amount of nitrogen flowing in the cell is limited by its uptake, but it performed similarly to ccFBA. In this regard, constraining methods are typically based on assumption rather than on experimental data. Hence, the development of constraining methodologies for GSMM based on experimental data, especially cheap and easily available one, is still an open problem.

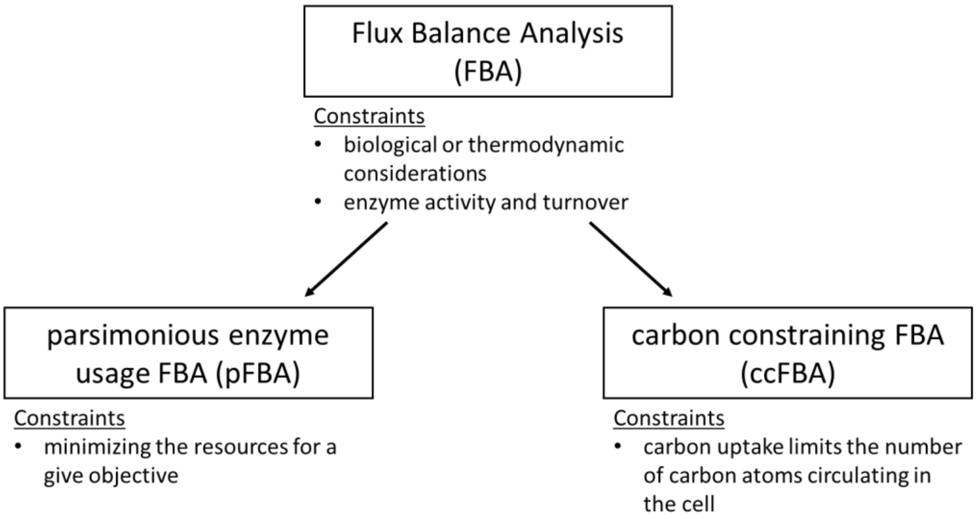

Flux Balance Analysis (FBA)
Constraints
• biological or thermodynamic considerations
• enzyme activity and turnover
parsimonious enzyme usage FBA (pFBA)
Constraints
• minimizing the resources for a give objective
carbon constraining FBA (ccFBA)
Constraints
• carbon uptake limits the number of carbon atoms circulating in the cell

Figure 1.8 *GSMM constraining strategies.*

In order to apply GSMMs, several model elements must be selected for a good representation of the intracellular metabolic state. The main model elements describe the cell dry weight (i.e., which is required to calculate the cell fluxes) and its dynamic change along the culture, the biomass composition, the objective function, the cell death rate, the cost for protein secretion and amino acid catabolism. It was observed (in different reactor conditions, for several genetic traits and different types of mAbs) that the objective function, the biomass composition, and the dynamic changes of cell dry weight have a major impact on GSMMs prediction (S.-M. Schinn et al., 2021). Accordingly, accurate selection of these elements is required.

Some attempts were made to use GSMMs in a dynamic fashion by coupling them with dynamic material balances. In one case, the GSMM was used to estimate the uptake and secretion rates of extracellular metabolites which were fed to material balances (Setoodeh et al., 2012). In another cases, the pseudo-steady state assumption was removed and the dynamic material balances for all species were solved by reducing it to a single optimization problem (Martínez et al., 2015). This model suggested that the temperature shift in CHO cell improves productivity because it allows a prolonged cell growth and higher viability.

1.5.3.1 Descriptive activities

In the context of descriptive activities in biopharmaceutical processes, an accurate understanding of the cell metabolism in response to different stimuli has a great impact on cell culture design and management, and host improvement. Stoichiometric models and GSMMs are a valuable tool to achieve a better understanding of cell metabolism. In fact, they have been applied in the biopharmaceutical industry for (Gopalakrishnan et al., 2020):

- the understanding of the relationship between cell metabolism and phenotype, even in the presence of genetic perturbation;
- the understanding of the relationship between cell metabolism and culture conditions;
- the understanding of the relationship between cell metabolism and medium composition;
- the generation of context specific models.

The understanding of cell metabolism is typically achieved by *i*) exploiting GSMMs alone, or *ii*) integrating GSMMs with biological data.

GSMMs for metabolic understanding

In regard to the study of the relationship between cell metabolism and phenotypes, many studies focused on the relationship between cell behavior in response to mAbs production. For example, cell lines expressing variable amounts of mAbs in different growth phases when subject to a specific treatment were studied thanks to a CHO metabolic model describing the central and amino acid metabolism, extracellular transport and biomass synthesis generated from nuclear magnetic resonance data (Carinhas et al., 2013). The model showed that the treatment produces a more sustained nutrient consumption especially during stationary phase resulting in a more pronounced mAbs production. Other studies focused on understanding the metabolic differences between high and low productive cell lines through GSMMs (Huang & Yoon, 2020b; Popp et al., 2016). High productive cell lines are characterized by upregulated oxidative phosphorylation, TCA cycle, and amino and nucleotide sugars metabolism, while low productive cell lines showed downregulation of the TCA cycle and pentose phosphate pathway (Huang & Yoon, 2020b). Similarly, cells with variable levels of productivity showed a significant difference in the metabolism of amino acids, such as glutamate, aspartate and glutamine, resulting in high productive cells that consume or produce few lactate, and low productive cells that produce a substantial amount of lactate (Popp et al., 2016). Recently, the integration of the secretory pathway into the GSMM and related cost of protein production allowed to understand that CHO cells expressing mAbs reduce the production of other non-essential complex host cell proteins (Gutierrez et al., 2020).

In the same context, the cell metabolic state during growth and non-growth phases was also characterized by means of GSMMs. For example, a metabolic model was curated using CHO DNA to improve reactions, mAb and biomass composition, and supported a metabolomics analysis in finding that lower amount of glycerophospholipids, the main components of cellular membranes, is associated with limited growth (Selvarasu et al., 2012).

Concerning the understanding of cell metabolic behavior in different culture conditions, the metabolic changes during the shift between lactate producing and lactate consuming phases were studied through a metabolic model of CHO cells describing the central pathways and amino acid metabolism (Martínez et al., 2013). In this case, similar fluxes through the lower part of the TCA cycle were observed, but lactate-consuming cells showed a much higher energy efficiency, with a 6 times greater amount of ATP produced per mole of carbon substrate.

In regard to the understanding of the relationship between cell metabolism and culture media, GSMM can really help the design tailored media to enhance desired host characteristics. In fact, the different metabolic states of cell cultured in different media were studied thanks to a reduced stoichiometric model describing central pathways, amino acid metabolism and biomass formation using the elementary flux mode (Hagrot et al., 2017). In other cases, the strong impact of leucine and valine concentration in the media on cell productivity was highlighted by GSMMs (Huang et al., 2020).

Finally, a better understanding of specific cellular strains can be achieved through the generation of context specific GSMMs by means of transcriptomic data. In this regard, the transcription levels are used to identify available genes and clean the GSMM of unavailable reactions (Gutierrez et al., 2020; Hefzi et al., 2016; Huang & Yoon, 2020a). These context-specific GSMMs are typically generated through the GIMME algorithm (Becker & Palsson, 2008). Unfortunately, the deactivation of a gene which is not present in the transcriptomic data is not always correct, because it can lead to delete reactions that are not captured in the transcriptome for lack of analytical sensitivity (Hyduke et al., 2013).

Despite the number of applications of GSMMs on CHO cells, the research on this topic is still at the beginning with respect to the metabolic modeling of bacterial species, mainly because of the complex mammalian metabolic network. Furthermore, one big limitation is given by the difficulty to assess the goodness of model metabolic representation.

GSMMs and biological data for metabolic understanding

Attempts to overcome the problem of assessing the quality of the model metabolic representation were made by integrating GSMMs with omics data, such as transcriptomic, proteomics, metabolomics and fluxomics, to compare model predictions and identify points of similarity or discord with the reality (Hyduke et al., 2013). For example, transcriptomic were used to confirm the high fluxes of certain reactions found in the GSMM by assuming that high transcription levels are associated to high enzyme activity and high fluxes (Huang et al., 2020; Huang & Yoon, 2020a).

Fluxomics represents one of the best solutions to obtain representative stoichiometric models from experimental data. In fluxomics, stoichiometric models are necessary to calculate the metabolic fluxes from ^{13}C isotope labelling metabolite concentration and extracellular uptakes data through a procedure called metabolic flux analysis (MFA; Quek et al., 2009). Fluxomics and labelling experiments allow studying the metabolic state of cells through the direct measurement of intracellular fluxes in response to different conditions.

In regard to the understanding of the relationship between cell metabolism and phenotypes, non-expressing and high productive apoptosis-resistant cells were compared by means of ^{13}C labelling data showing that high productive cells typically consume lactate and have high fluxes in TCA cycle and enhanced oxidative metabolism (Templeton, Smith, et al., 2017). Similarly, the difference between cells with non-induced and induced protein synthesis highlighted a more efficient utilization of glucose which leads to larger flux of pyruvate into the TCA cycle (Sheikholeslami et al., 2013), while high-growing and high-producer cell were characterized by a more efficient replenishment of TCA cycle with metabolites from glucose at later culture stages, and a more robust oxidative phosphorylation (Dean & Reddy, 2013). Furthermore, ^{13}C labelling data showed that high growth was correlated with low TCA cycling and lactate consumption during peak production, while antibody production is associated to high flux in the pentose phosphate pathway and high energy production from oxidative phosphorylation (Templeton et al., 2013).

Concerning the understanding of the cell behavior in different conditions, ^{13}C labelling experiments showed that higher lactate uptake during the lactate consuming phase led to higher cell growth and viability in CHO cells with overexpressed anti-apoptotic proteins (Templeton et al., 2014). Furthermore, metabolic difference in cells cultured in fed-batch and perfusion mode was highlighted as a higher gross cell growth associated with higher cell death rate in perfusion cultures, and a larger productivity in fed-batch cultures (Templeton, Xu, et al., 2017).

Regarding the better understanding of the metabolic response of cells to different media and feed composition, ^{13}C labelling experiments are valuable assets. In fact, they highlighted that the adjustment of glutamine, aspartate, glutamate, and serine concentrations can reduce ammonia production, improving cell growth without alter bioenergetic fluxes (McAtee Pereira et al., 2018). Furthermore, the use of low glutamine feeding favor the glycolytic flux and a higher percentage of pyruvate directed towards the TCA cycle rather than to lactate production, resulting in lower cell concentration but higher specific productivity (Sheikholeslami et al., 2014).

The use of fluxomics and ^{13}C isotope labelling experiments allow to achieve by means of stoichiometric models an accurate understanding of cell metabolism as response of various stimuli. However, ^{13}C isotope labelling experiments are very expensive strongly limiting the number of research on this field. Furthermore, simplified metabolic network are often used to deal with the limited number of metabolites monitored in these experiments. Accordingly, strategies to use the complete GSMMs in an accurate fashion would be beneficial for a more consistent application of GSMMs in host understanding during bioprocess development.

1.5.3.2 Predictive activities

In the context of predictive activities in the biopharmaceutical industry, one of the main interests is the prediction of phenotype when cells are subjected to different stimuli. GSMMs represents a solution to associate culture conditions, intracellular fluxes, and cell phenotypes. In fact, they have been often used to predict the main cell phenotypes, such as growth rate and productivity. For example, growth rate (Calmels et al., 2019; Feist et al., 2007; Gutierrez et al., 2020; Hefzi et al., 2016) and productivity (Calmels et al., 2019; Gutierrez et al., 2020) were predicted with good agreement with experimental data thanks to GSMMs. Furthermore, the growth rate, predicted by means of context specific GSMMs generated with time varying transcriptomic data, showed that gene dynamics does not perturb growth rate predictions, even when different set of reactions are excluded as unavailable in the cells (Huang & Yoon, 2020a). Other phenotypes were predicted through GSMMs, such as amino acid consumptions, essential genes for cell survival (Feist et al., 2007), and the metabolic features with large impact on protein secretion (Gutierrez et al., 2020).

GSMM prediction are not often completely accurate and a certain mismatch with experimental data is typically observed, especially with large and complex models (Calmels et al., 2019). This can be mainly due to inconsistencies in the metabolic network, due to unknown reaction routes, and the constraining method. General constraints are often loose, allowing an excessive freedom of the intracellular fluxes which reduces prediction accuracy. Other methods, such as pFBA and ccFBA, improve the constraining, but are still not enough representative of the real situation, while ^{13}C isotope labeling experiments are typically too expensive to be utilized as standard constraining method. Hence, the improvement of the constraining methods, such as the introduction of data-based solutions for GSMM constraining, would be extremely valuable to improve the prediction accuracy.

In the recent years, GSMMs have been coupled with machine learning methods to improve the prediction capability of the metabolic models. Applications combining GSMMs and machine learning have been recently reviewed by many authors (Antonakoudis et al., 2020; Khaleghi et al., 2021; Zampieri et al., 2019). In particular, the main applications concerned the prediction of growth rate, amino acids concentration and product quality. In fact, the intracellular fluxes from an *E. Coli* GSMM were fed to a LASSO regularized multinomial logistic regression to predict growth conditions, leading to increased accuracy in the predictions and the identification of the essential biological reactions related to the growth rate (Sridhara et al., 2014). Other linear regression models were used to predict the time-course variations in amino acid concentrations from VCC, product titer and the uptakes of main nutrients and by-products in CHO cell cultures (S. Schinn et al., 2021). In this case, the synergism between GSMMs and linear regression techniques allowed to improve the predictions with respect to the ones provided by the GSMM alone.

Even complex, non-linear deep-learning methods (i.e., ANN) were coupled to GSMMs. For example, the N-glycosylation of proteins in CHO cells was estimated from the flux of nucleotide sugars donors provided by a GSMM by means of an ANN (Antonakoudis et al., 2021). This work showed that the combination of these techniques produced very accurate predictions of the glycan distribution, allowing to substitute and simplify a complex section of the model. Other studies tried different methods to integrate GSMM fluxes and transcriptomic data to predict experimental cell growth in yeast fermentations (Culley et al., 2020). In this case, a multiview ANN, which combines features extracted from multiple inputs, showed the best prediction performance and was positively validated with samples having a different knockout pattern in their transcriptome than the training dataset. Finally, 29 intracellular fluxes measured in ^{13}C labelling experiments were predicted from 16 extracellular uptakes and other categorical features through SVM (Wu et al., 2016). In this work, ^{13}C data from 120 paper on different species and cultivation methods were used to train the machine learning models, whose prediction were adjusted using some stoichiometric constraints to satisfy material balances.

Despite the many applications are available in the Literature, research on consistent ways of combining machine learning and GSMMs is still ongoing.

1.5.3.3 Prescriptive activities

The optimization of both the culture environment and the host cell is one of the main tasks during bioprocess development. In this regard, GSMMs have been mainly used to:

- optimize media and feed composition;
- suggest genetic modification strategies leading to the overproduction of a desired chemical compound.

Optimization of media and feed composition

The optimization of the media composition through GSMMs is typically not automated and relies on the use of GSMMs to preliminary test the cell response to changes in the composition of the medium. In fact, through the knowledge of the metabolic impact of media components, it is possible to adjust the medium composition in order to drive the metabolism to obtain a desired objective. For example, the understanding of metabolic differences between CHO cells cultured in different media through GSMMs integrated with transcriptomic data led to the improvement of specific productivity (Huang et al., 2020). Specifically, the addition of valine and leucine in the medium showed to improve cell productivity both in the GSMM and in validatory cultures. Similarly, the behavior of CHO cells in different media was studied through stoichiometric models and ^{13}C labelling experiments leading to the identification of essential and non-essential amino acids and the design of an optimized media (Deshpande et al., 2009). Furthermore, in CHO cell cultures the feed composition was optimized to enhance the desired metabolic characteristics by means of GSMMs. For example, the accumulation of reactive oxygen species in high seed density CHO cultures was identified as cause of reduced viability through GSMMs. (Brunner et al., 2021). In this case, the reformulation of the feed through the addition of lactate and cysteine showed a reduction in the reactive oxygen species formation leading to enhanced viability and 47% increase in antibody titer.

In this regard, the optimization of media and feeding composition relies on accurate predictions of the cell phenotype to achieve a good culture design, which is often limited by the available constraining methods. Hence, the development of constraining method resulting in more accurate prediction would be beneficial for a better culture design.

Genetic engineering of host cells

GSMMs coupled with optimization techniques are able to suggest gene regulation that enhance the expression of a desired biopharmaceutical target. These techniques aim at identifying multiple gene regulations, either upregulation, downregulation, or knockout, that maximally couple the cellular objective (i.e., cellular growth) with an industrially imposed production target, for example, in terms of product secretion. This is typically achieved with a bilevel optimization through Mixed-Integer Linear Programming algorithms, in which the inner level maximizes the cellular objective (e.g., through FBA) given the genetic regulations that allows to maximize the desired production target at the outer level. Examples of these techniques are OptKnock (Burgard et al., 2003), which identifies optimal genes deletion, OptReg (Pharkya & Maranas, 2006), which identifies optimal genes activation, inhibitions and deletions, and OptForce (Ranganathan et al., 2010), which identifies the intracellular flux deviation from the wild-type that allows overproduction. Many other algorithms, which slightly modify these ideas, can be found in the Literature and have been reviewed in Antonakoudis et al. (2020).

These algorithms have been used to genetically modify the metabolism of several host cells to enhance a desired production objective. For example, in yeast and bacterial fermentations the production of aromatic amino acids as polymer precursors (Suástegui et al., 2017), octanoic acid from renewable feedstocks (Tan et al., 2018), and sugar precursors for glycan production (Wayman et al., 2019) were enhanced by means of these optimization algorithms. Similarly, in *Pichia Pastoris* cultures for the production of recombinant proteins, single gene knockouts that led to improved production while also minimizing the intracellular flux deviations from the wild-type were identified through GSMMs and optimization algorithms (Saitua et al., 2017). In this case the knockout of genes led to increased formation of cysteine and tryptophan resulting in increased productivity.

In CHO cell cultures, GSMMs have been used to heuristically support the genetic engineering of CHO cells. For example, the partial knockout of a gene was virtually simulated by means of GSMMs, which predicted a 24% productivity increase with the total knockout of the gene (Gutierrez et al., 2020). In this case, the experimental evidence showed a 14% productivity increase in real cell cultures. In another case, regulation of multiple genes to reduce the production of host cell proteins, a process related impurity during therapeutic recombinant protein production, was assisted by GSMMs (Kol et al., 2020). In this case, the energetic cost of host cell protein production was simulated and determined through GSMMs, leading to a reduction in host cell proteins by 40-70% and an increase productivity and cell growth.

In this regard, optimization methods for genetic engineering purposes are complex to implement (i.e., requiring a bilevel optimization) and extremely time-consuming, especially in large and complex metabolic networks. For this reason, the GSMM-based genetic engineering of mammalian cells, characterized by extremely large metabolic networks, is still an open issue. In this context, the introduction of faster and simpler methods to suggest genetic modification would be extremely beneficial to support the development of better and optimized cells.

1.6 Objective of the research

Despite the wide application of mathematical modeling in the context of Biopharmaceutical Industry 4.0, several limitations have been found in the current Literature. Accordingly, research in those areas can provide valuable improvement to the biopharmaceutical sector, and specifically to the development of monoclonal antibodies. The main challenges can be summarized into:

- a wide portion of the analyzed literature is focused on the exploitation of data to improve process understanding as fulfillment of the regulatory requirements of QbD. In particular, biological data, such as metabolomics, have often been used for process host understanding. However, an integrated workflow for the metabolomics analysis and the fusion of metabolomics and process data for a broad understanding of the relationship between metabolism and process CQAs is still missing. This is particularly emphasized when considering industrial data: process data and metabolomic data are rarely available at the same time. Furthermore, due to the batch nature of the process, those data are intrinsically dynamic, while process and metabolic dynamics are often disregarded.
- A large part of the presented Literature is focused on the exploitation of data and modeling methodologies to accelerate decision making in such a way as to reduce development times and costs. This is often achieved through the forecasting of process performance. In the case of mAbs, there is a lack of new types of data and innovative modeling strategies to

accelerate the development process. In particular, the exploitation of metabolomic data dynamics to accelerate cell lines selection is missing.

- The development of mAbs is typically characterized by the reduced availability of data (i.e., a handful of experiments), especially at large process scales, such as the pilot one, because of their high cost and long duration. In fact, a single experiment can cost tens of thousands $ and last several weeks. Multivariate statistical methods can be beneficial in studying data from bioprocesses, but they cannot reliably capture the main correlations when the number of data is limited, due to sample underrepresentation and large biological variability. This leads to the use of univariate methods to model an intrinsically multivariate one, which can lead to misleading conclusions. In this context, the use of science-based methodologies that can handle the multivariate nature of the process can be supported by data augmentation strategies, which are often used in other sectors in scenario with limited available data. However, in the biopharmaceutical field, application of data augmentation methods, especially model-based ones, is still an open issue.
- New modeling strategies, such as hybrid models, have been developed for mAb cell cultures and applied for various activities. Despite the ferment of the research community about this topic, the proposal of innovative hybrid methodologies is typically limited to the consistent application of hybrid model for process understanding and monitoring. For what concerns process optimization, the use of hybrid models has been proved to be effective for the identification of optimal feeding schedule. However, a proof of the advantages of using hybrid model for feeding schedule optimization over traditional experimental strategies is missing.
- GSMMs have been widely used in bacterial fermentations, but their application on mammalian cells is still limited and at its early stages. This is mainly due to the high complexity of mammalian cells and the inaccuracies that are still present in the metabolic models. Such high complexity leads to an accuracy of the model which is not fully satisfactory and strongly limits the applicability of GSMMs on mammalian cells. Specifically, a reliable, accurate and cheap method based on experimental data to properly define the intracellular constrains for accurate modeling of cell metabolism is missing. The available methodologies are either too expensive (i.e., ^{13}C labeling experiments) or not enough accurate (i.e., FBA, pFBA, ccFBA).
- GSMMs have been widely applied in fermentation process to identify targets for the genetic improvement of the host cells. However, this still relies on optimization methods that are complex to implement and computationally demanding and time-consuming. This limits the application of such optimization algorithms for the genetic improvement of CHO cell lines, which are typically characterized by large metabolic networks. Accordingly, an alternative method based on simple data-based mathematical methodology to exploit GSMMs for the identification of genetic engineering target is missing and would be

extremely beneficial for the genetic improvement of the mammalian cell in mAb development.

In view of the above, the objective of this Dissertation is to develop digital models (data-driven, hybrid, and first principles) to support and accelerate the monoclonal antibody development process.

The innovative contributions that can be found in this Dissertation are:

- **Integration of process and biological information to aid the selection of performing cell lines**. Metabolomics is typically related to a single CQA or process parameter, and the association with multiple ones is missing. In this Dissertation a framework for the integration of multiple dynamic information from process and metabolomics will be presented. This correlates the changes in cell metabolic states to the dynamic process behaviors, allowing to better understand and manage the cultivation process, and increase the confidence in the selection of performing cell lines.
- **Understanding of the metabolic changes occurring over the cultivation process and their influence on cell phenotype from the explicit exploitation of the dynamic evolution of metabolomic data**. The dynamic evolution of untargeted metabolomics is typically not considered explicitly, especially in machine learning applications. Part of this Dissertation is focused on the explicit exploitation of dynamic metabolomic data to support and accelerate cell lines selection during mAbs development. This can be exploited to: *i*) understand of the sequence of metabolic changes occurring over the cultivation process; *ii*) study the influence of cell metabolic and physiological changes on cell phenotype.
- **Identification of high performing cell lines in scenarios with limited available data through *in silico* data augmentation**. In biopharmaceutical processes, which are characterized by a limited number of experiments, especially at pilot scale, the use of *in silico* data augmentation methods is typically underexplored. This would allow the use of science-based solutions, based on multivariate latent variable regression methods, to support various stages of mAb development. In this Dissertation, a proof of the applicability of model-based *in silico* data generation is given to improve the identification of high performing cell lines through multivariate methods in scenarios with limited available data, typical of the biopharmaceutical sector.
- **Feeding schedule optimization through hybrid models**, with particular focus on demonstrating the advantages of applying hybrid models over DoE methodologies. Despite hybrid models have been applied to mammalian cell cultures with different purposes, a proof of their applicability to accelerate process optimization is missing. Part of the research effort in this Dissertation is focused on proving that the optimization of the feeding schedule through hybrid models can be accelerated with respect to experimental strategies. This can be exploited to conduct *in silico* experimental campaign for the optimization of the feeding

schedule, providing a substantial reduction in the experimental effort and time required during process development.

- **Prediction of GSMMs intracellular constraints from cheap and easily available data through deep learning models**. Typical GSMM constraining methods (i.e., pFBA and ccFBA) provide only wide constrains often resulting in inaccurate predictions, while experimental measurement of intracellular fluxes with ^{13}C labelling experiments are expensive and time-consuming. Part of this Dissertation is aimed at developing a deep learning model to estimate GSMMs constraints from cheap, fast, and routinely available measurements. This can be exploited to decrease the width of intracellular bounds providing more accurate predictions of both metabolic state and phenotypes of cells, which lead to a better description of cell metabolism.
- **Identification of genetic engineering targets to improve CHO cell productivity through latent variable regression model inversion and GSMMs**. Integration of machine learning methods and GSMMs focusing on the genetic engineering of cells are still missing, as well as applications to mammalian cell cultures. Part of this Dissertation aims at developing machine learning models to identify optimal sets of genetic modifications (i.e., gene regulations) for the improvement of CHO cell productivity. In particular, the inversion of latent variable models on metabolic flux data from GSMMs directly identifies the metabolic state of the improved cells, which will lead to the identification of several complementary scenarios with mAb overproduction. Furthermore, the method can be easily specialized to different organisms, cells, and processing conditions, and is able to identify the optimal set of genetic modifications for each specific host organism.

Industrial and simulated case studies will be presented throughout this Dissertation to prove the effectiveness of the proposed methodologies.

1.7 Dissertation roadmap

This Dissertation is organized following the six innovative contributions presented in the previous section. A schematic roadmap of the Dissertation is shown in Figure 1.9. In this Dissertation, descriptive, diagnostic, predictive, and prescriptive activities are performed to support various stages of the mAb development pipeline. In the first part, industrial case studies, performed in collaboration with the multinational pharmaceutical company GSK, are presented, while the second part focused on the simulated case studies, which allow to know the real behavior of the system under investigation and prove the effectiveness of the proposed methodologies by a direct comparison with the system under investigation.

This Dissertation is organized as follows.

Chapter 2 describes the bases of the main mathematical methodologies used in this Dissertation.

Chapter 3 mainly comprises descriptive and diagnostic activities. In particular, it focuses on the integration of dynamic information from process and metabolomic data to accelerate cell line selection during bioprocess development. In this case, a framework for the integration of industrial data from process and biology is presented. This is exploited to understanding the metabolic state changes occurring along the cultivation process, and how they are associated with process performance. In this case, the use of multivariate techniques allows to understand the metabolic differences along the culture, perform a quasi-real time monitoring of cell metabolic state and correlate the metabolic changes to the time evolution of several process variables, providing information on metabolites that can be exploited for the monitoring of desired process performance.

Chapter 4 mainly comprises predictive and diagnostic activities. It is focused on the identification of high productive cell lines from the dynamic evolution of the metabolomic data to accelerate cell lines selection in an industrial case study. In this case, the dynamic evolution of the metabolomic data is exploited for the early identification of high productive cell lines during the culture course, obtaining with very high accuracy through multivariate modeling and specifically proposed variable selection methodologies. The interpretation of the models allows to identify metabolites that are associated to cell productivity, which can be exploited for the anticipated identification of high productive cells. Furthermore, the metabolic functions associated to cell productivity change along the culture are identified, providing insights for the enhancement of cell productivity through genetic engineering.

Chapter 5 comprises diagnostic and predictive activities. It is focused on proving that the application of *in silico* data generation is helpful in the identification of high performing cell lines in scenarios with a limited data availability, typical of large-scale biopharmaceutical processes. In this case, *in silico* data generated by means digital models improves the accuracy of multivariate models in predicting the antibody titer when few process batches are available with respect to modeling only the available process batches. This leads to the improved identification of high performing cell lines even in scenarios with limited available process data.

Chapter 6 mainly comprises prescriptive activities. It is focused on comparing an *in silico* experimental campaign for the optimization of the feeding schedule in the development of biopharmaceutical processes through hybrid semi-parametric models with an experimental campaign on the process. This to evaluate if the *in silico* experimental campaign can accelerate the experimentation and reduce the experimental burden in the process development. In study the feeding schedule is optimized through Design of Dynamic Experiments methodologies examining the sensitivity to the number of experiments. The results of the experimental campaigns are then compared with the *in silico* optimization performed with the hybrid model and eventually comparted with the theoretical process optimum. The *in silico* experimental campaign identifies better process optimum, closer to the real process one, even reducing the number of experiments required to identify the best feeding schedule.

In Chapter 5 and 6, a simulated process for the production of mAbs is considered. Hence, these works represent a proof of concept for the application of the proposed methodologies.

Chapter 7 mainly comprises predictive activities. It is focused on the development of a deep learning model to predict intracellular flux constraints for GSMMs from cheap and easily available measurements. In this work, several tested model configurations accurately predicted the metabolic intracellular fluxes, allowing to determine flux constraints to apply on GSMMs. The application of the predicted constraints is briefly presented to show the application of the proposed methodology, despite they have been performed by collaborators at Imperial College London (UK).

Finally, Chapter 8 mainly comprises prescriptive and diagnostic activities. It is focused on developing a machine learning strategy exploiting GSMMs to identify genetic modifications that improve the productivity of CHO cells. The proposed methodology results faster and simpler than traditional model-based genetic engineering techniques relying on optimization. Furthermore, it suggests genetic modification concerning the metabolism of specific amino acids and mAb glycosylation to improve cell productivity.

The works presented in this Dissertation fulfill some of the regulatory requirements of QbD (Food and Drug Administration, 2004; ICH Harmonised Tripartite Guideline, Guidance for Industry, Q8 Pharmaceutical Development, 2009), such as the enhanced process understanding, the management of process variability, risk mitigation, and monitoring and prediction of CQAs (Figure 1.10). The enhanced process understanding is achieved by knowing the impact and the functional relationship among process factors and CQAs, while the risk mitigation consists in the in-depth understanding of the factors with the largest impact on CQAs, in such a way as to reduce the probability of poor-quality products. Regulatory agencies require also to identify, understand, and control all the sources of variability with an impact on the CQAs, and to monitor and predict CQAs from other information over the entire design space.

In this Dissertation, the enhanced process understanding and high level of scientific knowledge by collecting and analyzing process and biological data is achieved by *i*) the integration of process and biological information (Chapter 3), *ii*) the identification of relevant process parameters with few available process batches (Chapter 5), and *iii*) the improved description of cell metabolism (Chapter 7). The identification and explanation of some critical sources of variability observed along the process is achieved by *i*) the identification of the cellular functions associated with cell productivity (Chapter 4), and *ii*) the improved description of cell metabolism (Chapter 7), the variability of the host cells and process is partially managed by developing of methodologies for genetic engineering of cells (Chapter 8), which allows to generate pools of cell lines with generally higher and less variable mAb productivity. Furthermore, the risk of obtaining a product with poor quality is reduced by *i*) early identifying of high productive cell lines (Chapter 4), *ii*) optimizing feeding schedule (Chapter 6), leading to a less probable selection of poorly productive cell lines and a feeding schedule ensuring high performance, and iii) identifying genetic engineering target to improve productivity (Chapter 8), which guarantees high product quality since cell generation Finally, the monitoring and prediction of CQAs or process endpoints is achieved by *i*) estimating the product titer time profile (Chapter 3), *ii*) developing the software for metabolomics analysis (Chapter 4), and *iii*) identifying high performing cell lines (Chapter 5). A better process understanding, management of the process variability, prediction, and risk management allow a higher degree of flexibility to improve the process and to discover potential weaknesses of the product and process.

Appendix A contains additional information on mAbs and industrial cell cultures. Appendix B, C, and D, E, and F comprise additional material associated with Chapter 3, 4, and 5, 6, and 7, respectively.

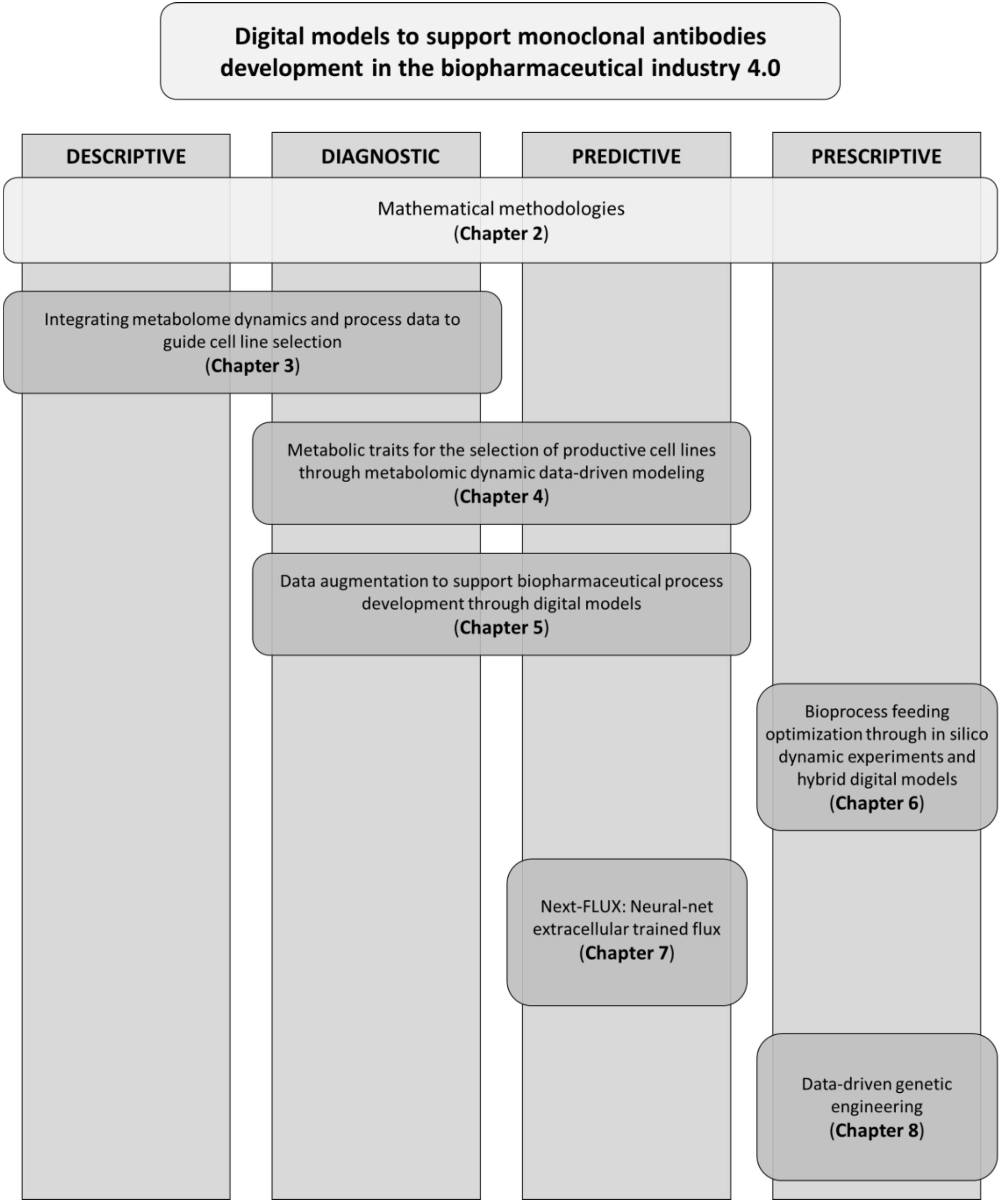

Digital models to support monoclonal antibodies development in the biopharmaceutical industry 4.0
DESCRIPTIVE
DIAGNOSTIC
PREDICTIVE
PRESCRIPTIVE
Mathematical methodologies (Chapter 2)
Integrating metabolome dynamics and process data to guide cell line selection (Chapter 3)
Metabolic traits for the selection of productive cell lines through metabolomic dynamic data-driven modeling (Chapter 4)
Data augmentation to support biopharmaceutical process development through digital models (Chapter 5)
Bioprocess feeding optimization through in silico dynamic experiments and hybrid digital models (Chapter 6)
Next-FLUX: Neural-net extracellular trained flux (Chapter 7)
Data-driven genetic engineering (Chapter 8)

Figure 1.9 *Dissertation roadmap.*

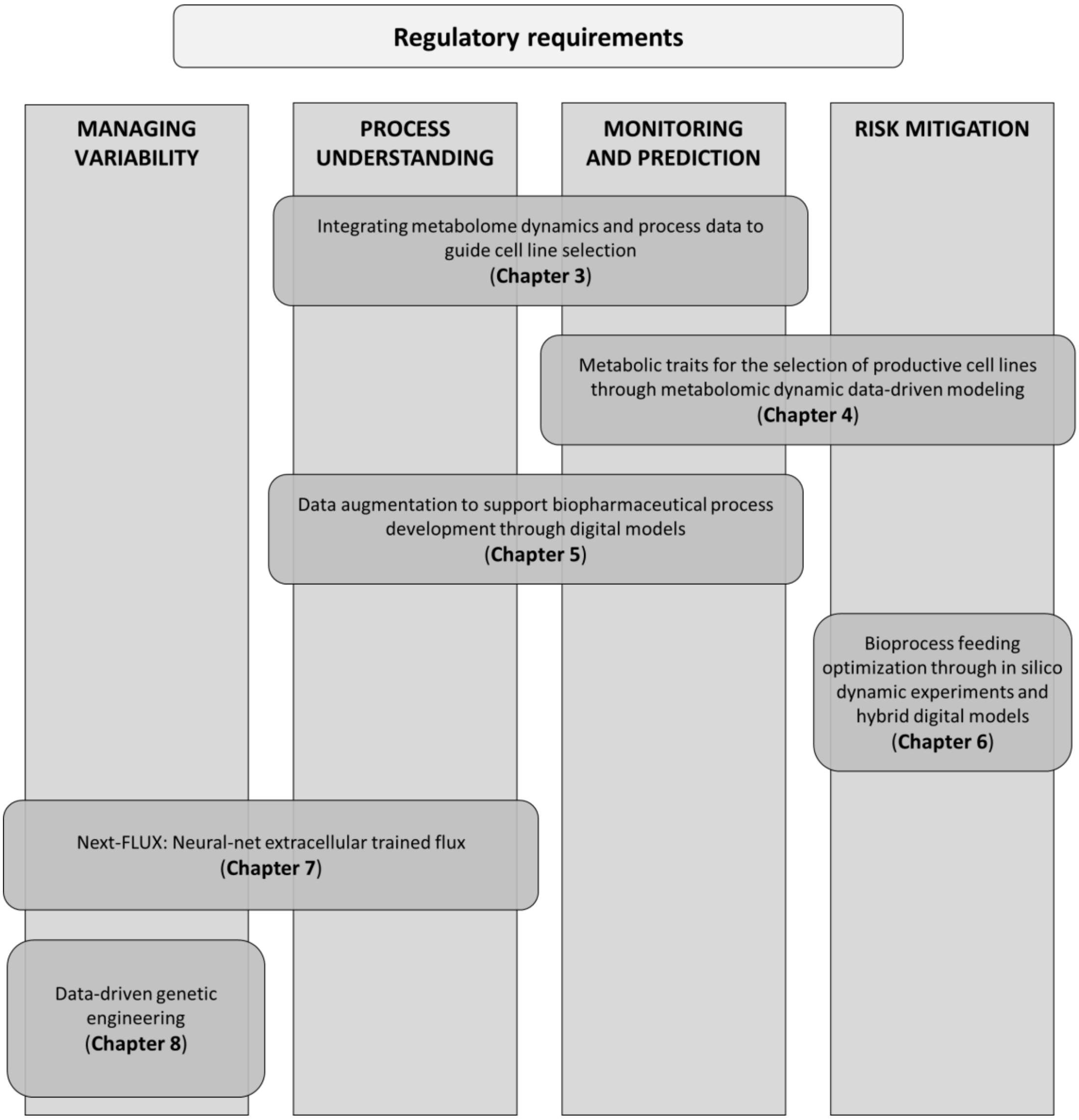

Regulatory requirements
MANAGING VARIABILITY
PROCESS UNDERSTANDING
MONITORING AND PREDICTION
RISK MITIGATION
Integrating metabolome dynamics and process data to guide cell line selection (Chapter 3)
Metabolic traits for the selection of productive cell lines through metabolomic dynamic data-driven modeling (Chapter 4)
Data augmentation to support biopharmaceutical process development through digital models (Chapter 5)
Bioprocess feeding optimization through in silico dynamic experiments and hybrid digital models (Chapter 6)
Next-FLUX: Neural-net extracellular trained flux (Chapter 7)
Data-driven genetic engineering (Chapter 8)

Figure 1.10 *Roadmap of the regulatory requirement addressed in this Dissertation.*

Chapter 2

Mathematical methodologies

In this Chapter, the theory behind the mathematical methodologies used in this Dissertation is presented. In particular, multivariate models, such as Principal Component Analysis (PCA) and Partial Least-Squares (PLS), are initially explained, followed by Artificial Neural Networks (ANN). The theoretical basis of Genome-Scale Metabolic Models (GSMMs) is finally presented.

2.1 Multivariate modeling

Multivariate models are statistical models that are used for dimensionality reduction, data interpretation and visualization, correlation analysis, and regression/classification.

The dataset analyzed by multivariate models are typically preprocessed to remove scale and biases from the data: dataset are mean-centered (i.e., by removing the column-wise mean value) or autoscaled (i.e., by removing the column-wise mean value and scaling for the column-wise standard deviation), but other types of scaling exist (Eriksson et al., 2006).

2.1.1 Principal Component Analysis

Principal Component Analysis (PCA; Jolliffe, 2002) is a multivariate technique used for dimensionality reduction and information extraction. PCA decomposes a scaled dataset $\mathbf{X}\,[N \times V]$, of N samples or observations and V variables, in A independent (i.e., orthogonal) principal components (PCs), which describe the direction of maximum variability in $\mathbf{X}$ and capture the correlation between the V variables. PCA decomposes the dataset as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad , \tag{2.1}$$

where $\mathbf{T}\,[N \times A]$ is the score matrix, $\mathbf{P}\,[V \times A]$ is the loading matrix, the superscript T indicates the transpose, and $\mathbf{E}\,[N \times V]$ is the residual matrix, which is minimized in the least-squares sense. The scores represent the projection of observations in the PC space and describe the relationship between the N observations, while loadings describe the correlation structure between the V variables.

The computation of model scores and loadings (i.e., calibration) can be performed through a single value decomposition or through the Nonlinear Iterative Partial Least-Square algorithm (NIPALS; Geladi and Kowalski, 1986).

The number of PCs (i.e., the dimension of the reduced space) is typically selected through: *i*) scree test, *ii*) eigenvalues, and *iii*) cross-validation. In the scree test (Jackson, 1991), the number of PCs is set at the value where the variance explained by PCs stabilizes to an almost constant value, which indicates that any additional PC describes noise. According to the eigenvalue method (Mardia et al., 1979), PCs are discarded when the associate eigenvalue is smaller than one. This relies on a rule of thumb for which the eigenvalue roughly represents the number of original variables whose variability is captured by a PC. In cross-validation (Svante Wold, 1978), the number of PCs is selected as the one minimizing the reconstruction error, typically in terms of root mean squared error (RMSE) in a bootstrapping/jackknifing procedure. This method is the most robust.

The main model diagnostics to assess the performance of the model are the RMSE and the coefficient of determination R^2. The RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\mathbf{x}_n - \hat{\mathbf{x}}_n)^2}{N}} \quad , \tag{2.2}$$

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

$$R^2 = 1 - \frac{\sum_{v=1}^{V}\sum_{n=1}^{N}\left(x_{n,v} - \hat{x}_{n,v}\right)^2}{\sum_{v=1}^{V}\sum_{n=1}^{N}\left(x_{n,v} - \bar{x}_{v}\right)^2} \quad , \tag{2.3}$$

where $x_{n,v}$ is the value of the v-th original variable for the n-th sample, $\hat{x}_{n,v}$ is the value of the v-th original variable for the n-th sample reconstructed by the PCA model, and $\bar{x}_v$ is the average value of the v-th original variable. The coefficient of determination calculated over new samples is typically referred as $\ ^2$.

Once a PCA model is calibrated, a new observation $\mathbf{x}_{\text{NEW}}$ $[1 \times V]$ can be projected into the PCA model to study its relationship with the calibration observations and assess whether $\mathbf{x}_{\text{NEW}}$ conform to them. The projection is performed as:

$$\mathbf{t}_{\text{NEW}} = \mathbf{x}_{\text{NEW}}\mathbf{P} \quad , \tag{2.4}$$

where $\mathbf{t}_{\text{NEW}}$ $[1 \times A]$ is the score vector of the new observation.

Sample diagnostics, namely Hotelling's T^2 and squared prediction error (SPE), can be calculated to assess how well an observation is described by the model, identify potential outliers, and assess the influence of an observation on the overall model.

The Hotelling's T^2 quantifies the distance between the projection of an observation and the origin of the reduced space. It is typically used to quantify the deviation of a given sample from the average conditions of the calibration dataset. The Hotelling's T^2 for a given observation n is defined as:

$$T_n^2 = \mathbf{t}_n \mathbf{\Lambda}^{-1} \mathbf{t}_n^{\mathrm{T}} \quad , \tag{2.5}$$

where $\mathbf{t}_n$ is the score vector of the n-th observation, and $\mathbf{\Lambda}^{-1}$ $[A \times A]$ is a diagonal matrix collecting the inverse eigenvalues.

The SPE quantifies the mismatch between an observation and its reconstruction through the PCA model. It is used to identify observations with a correlation structure different than the others. The SPE for a given observation n is defined as:

$$SPE_n = \mathbf{e}_n\mathbf{e}_n^{\mathrm{T}} \quad , \tag{2.6}$$

where $\mathbf{e}_n = \mathbf{x}_n - \hat{\mathbf{x}}_n$ is the residual vector of the n-th observation $\mathbf{x}_n$, and $\hat{\mathbf{x}}_n = \mathbf{t}_n\mathbf{P}^{\mathrm{T}}$ is the reconstructed observation by the PCA model.

Confidence limits can be set for both Hotelling's T^2 and SPE (Nomikos & MacGregor, 1995a) to identify possible outlier values. The calculations of these statistics are based on the assumption that data used to build the model are independent and identically distributed, which leads to multi-normally distributed scores, and white-noise residuals.

The confidence limit on the Hotelling's T^2, T^2_{lim}, is calculated as:

$$T^2_{\mathrm{lim}} = \frac{A(N-1)}{N-A} F_{A,N-\mathrm{A},\alpha} \quad , \tag{2.7}$$

where $F_{A,N-\mathrm{A},\alpha}$ is the critical value of a F-distribution with A and $N-A$ degrees of freedom at the significance level α.

The confidence limit on the SPE, SPE_{lim}, is calculated as:

$$SPE_{\mathrm{lim}} = \frac{\sigma^2_{\mathrm{SPE}}}{2\mu_{\mathrm{SPE}}} \chi^2_{2\mu^2_{\mathrm{SPE}}/\sigma^2_{\mathrm{SPE}},\alpha} \quad , \tag{2.8}$$

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

2.1.1.2 Multiblock PCA

Multi-Block PCA (MB-PCA; Westerhuis et al., 1998) is an unsupervised multi-block method which relates different blocks of variables. MB-PCA correlates different blocks of variables measured for the same observation n by finding a common latent space. This methodology is particularly useful to improve interpretability of multivariate models and to correlate data from different sources.

MB-PCA can be performed through a standard PCA. The available data blocks are horizontally concatenated, placing the same observation along a row, prior the decomposition through a standard PCA (Eq. 2.1). The data blocks can be appropriately scale after the concatenation.

2.1.2 Partial Least-Squares regression

Partial Least-Squares regression (PLS; Wold et al., 2001, 1993) is a linear multivariate regression technique that is used to explain the joint correlation between a regressor matrix and a response one, and predict a new response given a set of new regressors. PLS identifies the direction of maximum covariance between a scaled regressor matrix $\mathbf{X}\,[N \times V]$ and a scaled matrix $\mathbf{Y}\,[N \times M]$ of M responses. PLS projects both $\mathbf{X}$ and $\mathbf{Y}$ in a reduced space of A latent variables (LVs) accoring to

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{2.9}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{2.10}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1} \quad , \tag{2.11}$$

where $\mathbf{P}\,[N \times A]$ and $\mathbf{Q}\,[M \times A]$ are the loading matrices, $\mathbf{T}\,[N \times A]$ is the score matrix, $\mathbf{E}\,[N \times V]$ and $\mathbf{F}\,[N \times M]$ are the residual matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively (minimized in a least-square sense), and $\mathbf{W}\,[N \times A]$ is the weight matrix used for the calculation of the scores. The weights are required to preserve the orthogonality among LV scores and identify the direction of maximum correlation among the scaled versions of $\mathbf{X}$ and $\mathbf{Y}$. The terms $\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}$ in Eq. (2.11) is often defined as $\mathbf{W}^{*}$ (Svante Wold et al., 2001), which is typically used for prediction. It is worth noting that the scores $\mathbf{T}$ and loadings $\mathbf{P}$ of a PLS model do not coincide with the ones of a PCA model. However, this notation is kept consistent with the general Literature on the topic.

The calculation of model scores, loadings, and weights (i.e., calibration) is typically done through iterative methods. The most common ones are the NIPALS algorithm (Svante Wold et al., 2001) and the SIMPLS algorithm (S. De Jong, 1993).

PLS can be used for predicting a response variable $\hat{\mathbf{y}}\,[1 \times M]$ from a set of new predictors $\mathbf{x}_{\mathrm{NEW}}\,[1 \times V]$ according to:

$$\hat{\mathbf{y}} = \mathbf{x}_{\mathrm{NEW}}\mathbf{W}^{*}\mathbf{Q}^{\mathrm{T}} \quad . \tag{2.10}$$

The number of LVs can be selected similarly to PCA (Section 2.1.1) with the scree test, eigenvalue method, and cross-validation. The cross-validation is the one ensuring to construct the best possible model for the desired use.

In PLS, the importance of each predictor variable (i.e., in $\mathbf{X}$) for the prediction of $\mathbf{Y}$ can be quantified through the variable importance in projection (VIP; Wold et al., 1993) index, which is defined for a variable v as:

$$VIP_v = \sqrt{\frac{V\sum_{a=1}^{A} w_{va}^2 SSY_a}{\sum_{a=1}^{A} SSY_a}} \quad , \tag{2.11}$$

where w_{va} is the weight of the v-th regressor variable on the a-th LV, and $SSY_a = \mathbf{q}_a^2\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a$ is the amount of the $\mathbf{Y}$ variability explained by the a-th LV, $\mathbf{q}_a$ and $\mathbf{t}_a$ are the column vectors referred to the a-th LV of $\mathbf{Q}$ and $\mathbf{T}$, respectively. The VIP value is proportional to the influence of a variable in the prediction of $\mathbf{Y}$. Typically, a variable with a VIP_v greater than one is considered as valuable predictor for the response $\mathbf{Y}$ (i.e., important variable).

2.1.2.1 PLS Discriminant Analysis

PLS Discriminant Analysis (PLS-DA; Barker and Rayens, 2003) is a classification technique, used to classify observation in different classes according to the regressors. Similarly to PLS, PLS-DA identifies the direction of maximum covariance (i.e., A orthogonal LVs) between the regressors $\mathbf{X}$ $[N \times V]$ and matrix $\mathbf{Y}_\mathrm{d}$ $[N \times B]$, which defines the attributions to B classes through B dummy variables, where a 0 is attributed to column b in row n if the sample n does not belong to the b-th class, while a 1 is attributed if the sample n belongs to the b-th class.

PLS-DA is performed over $\mathbf{X}$ and $\mathbf{Y}_\mathrm{d}$ through a standard PLS (Section 2.1.2). Hence, it outputs a real number, which can be used to determine the class attribution probability. A cumulative density function is fit based on the $\mathbf{Y}_\mathrm{d}$ of the calibration dataset to identify the probability of belonging to a specific class (Fawcett, 2006). This density function is then used on new observations to calculate their attribution probability.

2.1.3 Multiway modeling

Multiway multivariate modeling (Nomikos & MacGregor, 1994, 1995b) is used to deal with multidimensional matrices, where one dimension is usually associated with time (i.e., data has a time variability).

Multiway modeling consists in properly unfolding the multidimensional data $\underline{\mathbf{X}}$ $[N \times V \times T]$ (where T is the number of time instants in which V variables are collected for N batches) followed by the decomposition with a standard multivariate model. The data unfolding procedure is schematically shown in Figure 2.1. Data collected at different time instants (e.g., $\mathbf{X}^t$ $[N \times V]$ with $t = 1, 2, \dots, T$) are horizontally concatenated to generate a matrix $\mathbf{X}\,[N \times V \cdot T] = [\mathbf{X}^1 \quad \mathbf{X}^2 \quad \cdots \quad \mathbf{X}^T]$, which is the batch-wise unfolded version of $\underline{\mathbf{X}}$.

In multiway multivariate modeling, the loadings describe the correlation structure between $\mathbf{X}$ variables at different time instants, allowing to understand how the dynamics of different variables are cross correlated.

In this Dissertation, the multiway version of PCA (MPCA) and PLS (MPLS) have been used. In this case, the unfolding is followed by a standard PCA modeling, for MPCA, and a standard PLS modeling, for MPLS.

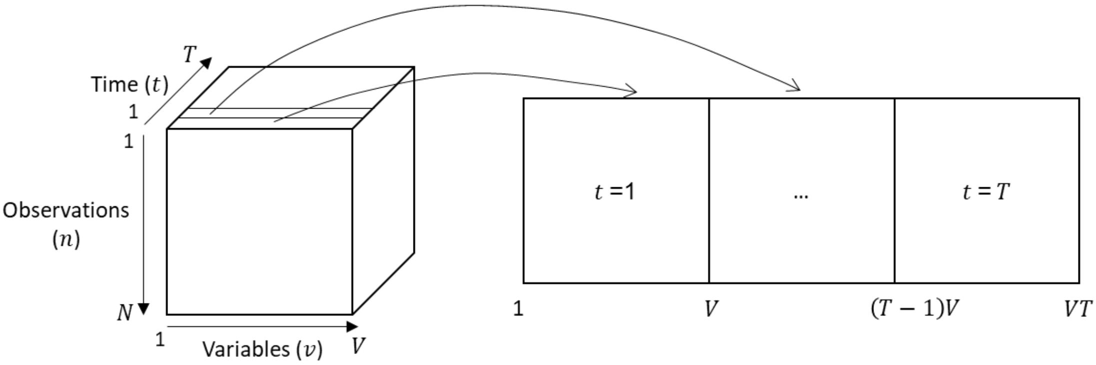


Figure 2.1 *Batch-wise unfolding procedure for multiway multivariate modeling.*

2.1.4 Evolving modeling

Evolving multivariate modeling (Ramaker et al., 2005) is a multi-model strategy that exploits partial dynamic information of time-varying data to accomplish the modeling. Evolving methodologies retain enlarging information on the past history of a dynamic observation to accomplish the multivariate modeling. Specifically, at each time instant t (with $t = 1, 2, \ldots, T$) a multivariate model is built on the batch-wise unfolded data considering the first t time instant, namely, the time instant from the beginning and the considered one t, $\mathbf{X}_t = [\mathbf{X}^1 \quad \mathbf{X}^2 \quad \cdots \quad \mathbf{X}^t]$. The matrix $\mathbf{X}_t$ is progressively enlarged as time progresses $t = 1, 2, \ldots, T$ and a new multivariate model is built, until the entire dynamics of available data is considered in the modeling. A schematic representation of the evolving strategy is shown in Figure 2.2.

The evolving version of the main multivariate modeling techniques (i.e., PCA and PLS) exists: in this Dissertation an evolving PLS-DA (PLS-DA; Barker & Rayens, 2003; Ramaker et al., 2005) was used.

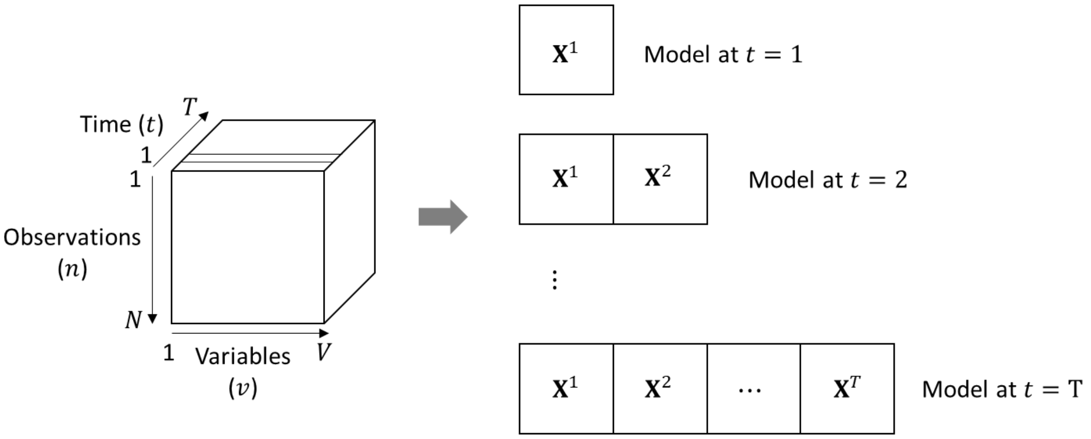


Figure 2.2 *Multi-model evolving strategy.*

2.2.5 Inversion of latent variable models

Multivariate model inversion (Jaeckle & MacGregor, 2000) is typically used for prescriptive and optimization purposes in product formulation and process design. Multivariate regression models (e.g., PLS) can be inverted to estimate a new set of inputs $\mathbf{x}_{\text{NEW}}$ $[1 \times V]$ corresponding to a desired response variables $\mathbf{y}_{\text{DES}}$ $[1 \times M]$ according to the learned correlations. The inversion of PLS models consists of the following steps (Tomba et al., 2012):

1. building a multivariate model between preprocessed regressors **X** and responses **Y** (Section 2.1.2);
2. determining the desired response variable $\mathbf{y}_{\text{DES}}$, which should be assigned as a predefined value (i.e., equality constraints) or as a one- or two-sided constraints (i.e., inequality constraints);

3. determining constraints for the new set of inputs $\mathbf{x}_{\mathrm{NEW}}$ according to physical bounds and the nature of the problem. Constraints can be given as equality or inequality constraints;
4. if $\mathbf{y}_{\mathrm{DES}}$ has only equality constraints, the model can be inverted only if it is valid for the desired $\mathbf{y}_{\mathrm{DES}}$. To assess that, $\mathbf{y}_{\mathrm{DES}}$ is projected into the latent space, calculating its $SPE_{\mathbf{y}_{\mathrm{DES}}}$ according to Eq. (2.6) by substituting $\mathbf{e}_n$ with $\mathbf{f}_{\mathbf{y}_{\mathrm{DES}}} = \mathbf{y}_{\mathrm{DES}} - \hat{\mathbf{y}}_{\mathrm{DES}}$, and comparing it with the SPE of historical observations through the 95% confidence limit (Eq. 2.8). If the $SPE_{\mathbf{y}_{\mathrm{DES}}}$ is largely different from the one of the historical observations or greater than the 95% confidence limit is not recommended to perform the inversion;
5. inverting the PLS model through the appropriate formulation defined according to the selected constraints at step 2 and 3.

The formulation of the PLS inversion problem depends on the constraints that are set on the desired response and acceptable regressors. In the most generic frameworks 4 different scenarios are defined (Tomba et al., 2012), but in this Dissertation only two scenarios are considered: *i*) unconstrained regressors and only equality constraints on $\mathbf{y}_{\mathrm{DES}}$, and *ii*) constrained regressors and some inequality constraints on $\mathbf{y}_{\mathrm{DES}}$.

In Scenario 1, when the regressors are unconstrained and only equality constraints are set for $\mathbf{y}_{\mathrm{DES}}$, the model inversion can be directly solved by calculating the score associated with $\mathbf{y}_{\mathrm{DES}}$ as:

$$\mathbf{t}_{\mathrm{DES}} = (\mathbf{Q}^{\mathrm{T}}\mathbf{Q})^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{y}_{\mathrm{DES}}^{\mathrm{T}} \quad . \tag{2.12}$$

This score is then used to calculate the estimated new input as:

$$\hat{\mathbf{x}}_{\mathrm{NEW}} = \mathbf{t}_{\mathrm{DES}}\mathbf{P}^{\mathrm{T}} \quad , \tag{2.13}$$

where $\hat{\mathbf{x}}_{\mathrm{NEW}}$ is the reconstructed version of $\mathbf{x}_{\mathrm{NEW}}$ and belongs to the model space.

In Scenario 2, when the regressors are constrained and some inequality constraints are set for $\mathbf{y}_{\mathrm{DES}}$, the model inversion is solved through an optimization problem, which is formulated as:

$$\min_{\mathbf{x}_{\mathrm{NEW}}}\left[(\hat{\mathbf{y}}_{\mathrm{NEW}} - \mathbf{y}_{\mathrm{DES}})\boldsymbol{\Gamma}(\hat{\mathbf{y}}_{\mathrm{NEW}} - \mathbf{y}_{\mathrm{DES}})^{\mathrm{T}} + g_1 T^2 + g_2 SPE_{\mathbf{x}_{\mathrm{NEW}}}\right] \quad , \tag{2.14}$$

subject to

$$\hat{\mathbf{y}}_{\mathrm{NEW}} = \mathbf{t}\mathbf{Q}^{\mathrm{T}} \quad , \tag{2.15}$$

$$\hat{\mathbf{x}}_{\mathrm{NEW}} = \mathbf{t}\mathbf{P}^{\mathrm{T}} \quad , \tag{2.16}$$

$$\mathbf{t} = \mathbf{x}_{\mathrm{NEW}}\mathbf{W}^{*} \quad ,. \tag{2.17}$$

$$SPE_{\mathbf{x}_{\mathrm{NEW}}} = (\hat{\mathbf{x}}_{\mathrm{NEW}} - \mathbf{x}_{\mathrm{NEW}})(\hat{\mathbf{x}}_{\mathrm{NEW}} - \mathbf{x}_{\mathrm{NEW}})^{\mathrm{T}} \leq g_3 SPE_{\mathrm{lim},\mathbf{x}} \quad , \tag{2.18}$$

$$\mathbf{x}_{\mathrm{NEW}} \in [\mathbf{x}_{\mathrm{lb}},\ \mathbf{x}_{\mathrm{ub}}] \quad ,. \tag{2.19}$$

$$\mathbf{y}_{\mathrm{NEW}} \in [\mathbf{y}_{\mathrm{lb}},\ \mathbf{y}_{\mathrm{ub}}] \quad ,. \tag{2.20}$$

where $\hat{\mathbf{y}}_{\mathrm{NEW}}$ is the predicted response by the PLS model, $\mathbf{t}$ the associated score vector, $SPE_{\mathbf{x}_{\mathrm{NEW}}}$ is the SPE calculated for $\mathbf{x}_{\mathrm{NEW}}$, $SPE_{\mathrm{lim,x}}$ is the SPE 95% confidence limit calculated for the regressors, $\mathbf{x}_{\mathrm{lb}}$ and $\mathbf{x}_{\mathrm{ub}}$ are the vectors of lower and upper bounds set for the regressors (note that they may also be equality constraints), $\mathbf{y}_{\mathrm{lb}}$ and $\mathbf{y}_{\mathrm{ub}}$ are the vectors of lower and upper bounds set for the desired response, and g_1, g_2, and g_3 are some corrective constants.

$\mathbf{\Gamma}$ is a matrix that defines through its diagonal elements the weight given to meet the specified $\mathbf{y}_{\mathrm{DES}}$ constrains. Large weight may be given to important variables for the specific studies, otherwise the fraction of each response variable m variability explained by the model can be used as weight.

2.2 Neural Networks

Neural networks are a machine learning method inspired to the structure of human brains, where neurons are connected to each other and exchange information. Artificial neural networks are a class of neural networks mainly used for regression based on a set of predictor variables.

Artificial neural networks (ANN; Rosenblatt, 1958) learns the relationship between inputs $\mathbf{X}\ [N \times V]$ and outputs $\mathbf{Y}\ [N \times M]$ from examples. This relationship learned during the training step is used to predict a new output given the inputs, through the mathematical interaction of interconnected neurons.

An ANN is composed by several layers of interconnected neurons (Figure 2.3), which are the fundamental units of the neural networks (Goodfellow et al., 2016). Typically, the ANN architecture has an input layer, several hidden layers, and an output layer. The input layer receives the input information and pass it to the first hidden layer for processing. It has the same number of neurons of the input data $\mathbf{x}_n\ [1 \times V]$. The hidden layers receive information from the previous layer and process it, before passing it to the following layer. The output layer receives information from the last hidden layer and provides the neural networks outputs. It has the same number of neurons of the output data $\mathbf{y}_n\ [1 \times M]$.

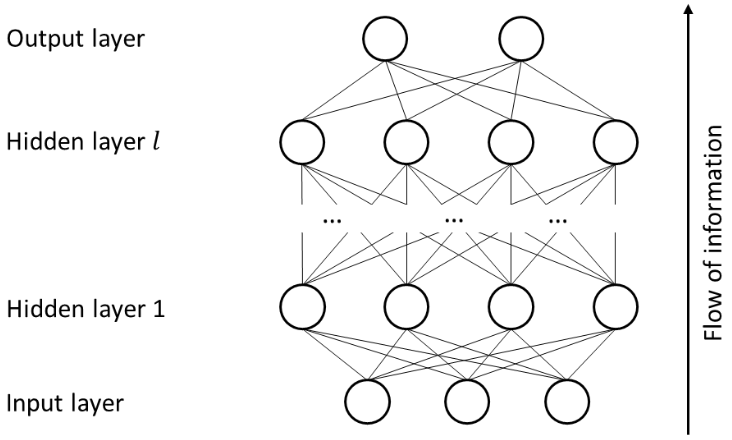


Figure 2.3 *General architecture of an ANN.*

Neurons are mathematical entities which receive information from other neurons and process it before passing to the other neurons. Considering the first hidden layer, it receives V input variables $\mathbf{x}_n = [x_1, x_2, \dots, x_V]$ from the input layer and performs linear and non-linear transformations on them to obtain the neuron's output a_h (i.e., activation), where $h = 1, 2, \dots, H$ is the number of the neuron in the layer. The mathematical transformation performed by the neurons is (Goodfellow et al., 2016):

$$a_h = f\left(\omega_h^0 + \sum_{v=1}^{V} x_v \omega_{h,v}\right) \quad , \tag{2.21}$$

where ω_h^0 is the bias of the h-th neuron, x_v is the v-th input variable, $\omega_{h,v}$ in the weight of the h neurons associated to the v-th input variable, and $f()$ is the non-linear activation function. Several activation functions, such as hyperbolic tangent, sigmoid, and rectified linear unit (ReLU; Fukushima, 1975), can be used. The *reLu* activation function is defined as:

$$f() = \max(0, a_h') \quad , \tag{2.22}$$

while the hyperbolic tangent activation function is defined as:

$$f() = \frac{e^{a_h'} - e^{-a_h'}}{e^{a_h'} + e^{-a_h'}}, \tag{2.23}$$

where a_h' is the output of the h-th neuron prior the application of the activation function.
The development of an ANN follows three phases: the selection of the ANN architecture, the training, and the testing. The architecture of the ANN should be selected accordingly to the specific application to maximize performance. Some rules of thumb indicates that 1 or 2 layers are sufficient for regression purposes. The training (Section 2.2.1) is the most time-consuming phase, in which weights are adjusted to match the input-output pattern of the training data. Finally, in the testing phase the ANN is presented with data not seen during training to assess its performance.

2.2.1 Training of the neural networks

The training of the neural networks is required to represent the mapping between the inputs and outputs coherently with the training data, by adjusting the values of the weights to maximize the neural network performance. Performance is typically evaluated through the minimization of a cost function, for example, the mean squared error (MSE) for ANN, defined as:

$$\mathcal{L}(\boldsymbol{\omega}) = \frac{1}{N}\sum_{n=1}^{N} L(\mathbf{y}_n, \hat{\mathbf{y}}_n) = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{y}_n - \hat{\mathbf{y}}_n)^2 \quad , \tag{2.24}$$

where $\boldsymbol{\omega}$ the matrix collects all the ANN weights, L is a loss function, and $\hat{\mathbf{y}}_n$ is the output predicted by the neural networks for the n-th sample. Several loss function can be used according to the specific application, but the most common ones are the mean squared error (Eq. 2.24), for regression tasks, and the categorical cross-entropy, for classification tasks (Gentiluomo et al., 2019; Krizhevsky et al., 2017).

The optimal weights are the ones minimizing the cost function $\mathcal{L}(\boldsymbol{\omega})$. Any numerical optimization algorithm can be used to calculate the optimal weights, but gradient-based ones are typically used in neural networks because provide the best results. In gradient-based optimization algorithms, the weights are updated according to:

$$\boldsymbol{\omega}^{(it+1)} \leftarrow \boldsymbol{\omega}^{(it)} - \eta \frac{\partial \mathcal{L}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\bigg|_{it} , \tag{2.25}$$

where $\boldsymbol{\omega}^{(it)}$ are the ANN weights at the it-th training iteration, η is the learning rate, and $\partial \mathcal{L}(\boldsymbol{\omega})/\partial \boldsymbol{\omega}|_{it}$ is the gradient of the cost function with respect the weight calculated at the it-th iteration, which needs to be calculated. The learning rate defines the velocity at which the weights are updated and is typically reduced along the training to aid convergence. The initial learning rate is typically selected according to the specific application.

Several optimization algorithms are available, which use a more complex formulation of Eq. (2.25). Nowadays, the most common and reliable algorithm for neural networks training is the ADAM algorithm (Kingma & Ba, 2015).

The gradient of the cost function is calculated through backpropagation, consisting of the propagation of the errors from the output to the inputs (Baughman & Liu, 1995; Goodfellow et al., 2016). In backpropagation, the chain rule is used to propagate the errors backward through the network for each weight. Considering a very simple neural networks with one hidden neuron (Figure 2.4) the gradient of the cost function can be calculated as:

$$\frac{\partial \mathcal{L}(\omega)}{\partial \omega_1} = \frac{\partial \mathcal{L}(\omega)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathrm{a}_1} \frac{\partial \mathrm{a}_1}{\partial \omega_1} \quad \text{and} \tag{2.26}$$

$$\frac{\partial \mathcal{L}(\omega)}{\partial \omega_2} = \frac{\partial \mathcal{L}(\omega)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \omega_2} , \tag{2.27}$$

where ω_1 and ω_2 are the weights of the hidden and output neurons, $\hat{y}$ is the predicted output of this simplified neural networks, and a_1 is the activation (i.e., output) of the hidden neuron.

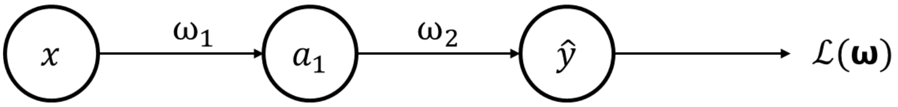


Figure 2.4 *Schematic representation of a simple ANN with one hidden layer.*

The training of the neural networks is composed by several steps:

1. initialization of the neural network weights;
2. prediction of the neural network outputs $\hat{\mathbf{Y}}$ $[N \times M]$ from the available data $\mathbf{X}$ $[N \times V]$;
3. calculation of the cost function $\mathcal{L}(\boldsymbol{\omega})$ (Eq. 2.24) and its gradient through backpropagation;
4. updating of the weights according to the selected optimization algorithm, similarly to Eq. (2.25);
5. iteration to step 2 until convergence.

Several methods can be used to initialize the weights in step 1, such as from a random normal distribution, with a variance that can be arbitrarily low or defined according to the number of neurons of the layers (Glorot & Bengio, 2010; He et al., 2015).
A validation dataset is typically used to stop the training and to avoid overfitting, ensuring the generalizability of the neural network predictions.
The training of the neural networks often requires long time, especially for deep neural networks (i.e., large number of hidden layers). However, the introduction of graphics processing unit (GPU) computations is helping to speed up the training. Furthermore, large datasets are required for a good learning process, because the reliability of results is not guaranteed with limited training data, in which the learning might be biased by local minima of the cost function (Goodfellow et al., 2016).

2.3 Genome-scale Metabolic Models

Genome-scale metabolic models are at the basis of the mathematical modeling of cell metabolism (Maranas & Zomorrodi, 2016). GSMMs are stoichiometric-based models, in which the pseudo-steady state assumption is exploited to describe the mass conservation of each metabolite through a system of equations. The pseudo-steady state assumption holds true since the time constants of metabolic reactions are typically much smaller than the other cellular process, such as transcriptional regulation and cellular growth. According to that, the model can be expressed in the form of:

$$\mathbf{S} \cdot \boldsymbol{\nu} = 0 \quad , \tag{2.28}$$

where $\mathbf{S}\ [D \times U]$ is the model stoichiometric matrix for D metabolites and U intracellular reactions, and $\boldsymbol{\nu} = [\nu_1 \quad \nu_2 \quad \cdots \quad \nu_U]$ is the vector of U intracellular reaction rates (i.e., intracellular fluxes). Large mammalian cell GSMMs typically contain a couple of thousand metabolites and several thousands of reactions. The stoichiometric matrix $\mathbf{S}$ links metabolites to the respective metabolic reactions providing all the structural information of the metabolic network. GSMMs typically contain the link between metabolic reactions and the associated genes, since reactions are introduced in the GSMM if the enzyme catalyzing the reaction is encoded in the genome of the studied organism. The stoichiometric matrix $\mathbf{S}$ contains a
fictitious biomass reaction inventorying and draining all the biomass precursors (amino acids, lipids, carbohydrates, and energy as ATP) in an appropriate ratio. The rate of this biomass reaction ν_{biomass} is indicative of the growth rate per amount of resources taken from the extracellular environment. Furthermore, the stoichiometric matrix $\mathbf{S}$ also encodes the cell inputs (e.g., carbon sources, oxygen, etc.) and outputs (e.g., CO_2, ammonia, acetic acid, etc.) through artificial reactions carrying metabolites through the system's boundaries, named exchange reactions (Figure 2.5).

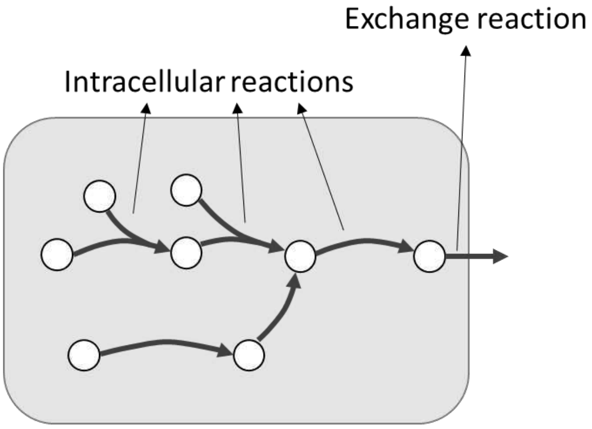


Figure 2.5 *Schematic representation of a GSMM with intracellular and exchange reactions.*

2.3.1 Flux balance analysis

Flux Balance Analysis (FBA; Duarte et al., 2007), a constraint-based analysis and reconstruction method, is the most widely applied method for the analysis and the solution of GSMMs. In FBA the GSMM model is solved, consisting in solving Eq. (2.28). Unfortunately, the number of metabolites D is smaller than the number of reactions ($D < U$), because most metabolites participate in several metabolic reactions. For this reason, the system of equation (Eq. 2.28) is underdetermined, and an infinite number of flux vectors $\boldsymbol{\nu}$, which are in the null space of $\mathbf{S}$, satisfy the steady state mass conservation.

To reduce the solution space to physiologically plausible metabolic states only, constraints are typically imposed on the intracellular flux vector $\boldsymbol{\nu}$ as:

$$\nu_u^{\min} \leq \nu_u \leq \nu_u^{\max} \quad , \tag{2.29}$$

where ν_u is the intracellular flux value of the u-th metabolic reaction from $\boldsymbol{\nu}$, $\nu_u^{\min}$ and $\nu_u^{\max}$ are the lower and upper bounds set for the u-th metabolic reaction, respectively. Flux constraints contain information on reaction reversibility (i.e., reversible or irreversible), which is typically determined from Gibbs free energy data or literature sources. In particular, if $\nu_u^{\min} = 0$ or $\nu_u^{\max} = 0$ the reaction is irreversible in the forward or backward direction, respectively, while if $\nu_u^{\min} < 0$ and $\nu_u^{\max} > 0$ the reaction is reversible. Constraints of exchange reactions depends on the metabolites available in the culture medium and are typically derived from measurement of the extracellular metabolite uptake and secretion rates. For intracellular reactions, the upper bound is typically set to a large value (e.g., 1000 or -1000 according to reaction directionality and reversibility), but constraints can also be derived based on enzyme activity and turnover, or very complex ^{13}C isotope labelling experiments (Maranas & Zomorrodi, 2016).

Despite the constraining, an infinite number of intracellular flux vectors satisfy the material balance under the given system conditions (Eq. 2.28 and the given constraints). To overcome this issue, in FBA it is assumed that the most plausible metabolic state of this system is given when an objective function, which defines the aim of the organism, is maximized or minimized. In this way, the solution of the material balance of Eq. (2.28) can be formulated as an optimization problem as:

$$\max_{\mathbf{v}} \mathbf{z}^{\mathrm{T}}\mathbf{v} \tag{2.30}$$

subject to the model equations (Eq. 2.28) and the defined constraints (Eq. 2.29), where $\mathbf{z}$ is a stoichiometric vector indicating how different intracellular fluxes are combined to form the objective function. Biomass is the most typical objective function, since it assumes that organisms evolved to efficiently convert resources into components and energy supporting cellular growth. Other objective function can be used, which are reported in Section 1.5.3 and in the cited Literature.

Parsimonious enzyme usage FBA (pFBA; Lewis et al., 2010) is an extension of FBA, which assumes the maximum stoichiometric efficiency of metabolic pathways, achieved by producing the maximum amount of biomass per unit of flux. Accordingly in pFBA, the solution of the material balances (Eq. 2.28) is computed by maximizing biomass (Eq. 2.30), while minimizing the sum of all intracellular reaction fluxes at the same time, as:

$$\min_{\mathbf{v}} \sum_{u=1}^{U} \mathrm{v}_u \quad , \tag{2.31}$$

2.3.2 Flux sampling

FBA provides a flux vector $\mathbf{v}$ within the feasible solution space that satisfies the optimality, but this optimal vector is not always predictive for the actual intracellular fluxes. Flux sampling is a possible solution to overcome this problem and obtain the probability distribution of the attainable fluxes for each metabolic reaction in the given conditions. In flux sampling (Bordel et al., 2010) random samples are drown from the feasible solution space of the model, providing a population of likely flux values within the given conditions defined by the constraints. Flux sampling only requires defining the model constraints, while no objective function has to be used, making the solution free form the assumption of a cellular objective.

The algorithms used for flux sampling (Megchelenbrink et al., 2014) typically follow a hit-and-run logic, where solution points are selected within the solution space in a sequential manner by slightly moving in an arbitrary direction defined by the solution space boundaries.

Chapter 3

Integrating metabolome dynamics and process data to guide cell line selection[*]

This Chapter studies the industrial development of monoclonal antibodies at micro-bioreactor scale (AMBR15^{TM}) and aims at accelerating the selection of the better performing cell lines. To that end, we apply a machine learning approach to integrate time-varying process and biological information (i.e., metabolomics), explicitly exploiting their dynamics.
Strikingly, cell line performance during the cultivation can be predicted from early process timepoints by exploiting the gradual temporal evolution of metabolic phenotypes. Furthermore, product titer is estimated with good accuracy at late process timepoints, providing insights into its relationship with underlying metabolic mechanisms and enabling the identification of biomarkers to be further investigated. The biological insights obtained through the proposed machine learning approach provide data driven metabolic understanding allowing early identification of high performing cell lines. Additionally, this analysis offers the opportunity to identify key metabolites which could be used as biomarkers for industrially relevant phenotypes and onward fit into our commercial manufacturing platforms.

3.1 Introduction

In the recent years, the pharmaceutical industry has invested heavily in the research, development and manufacturing of biopharmaceutical products to face the issues of increasing cost of traditional drug development, patent expiration, and market erosion through generic drugs. Recombinant proteins, such as monoclonal antibodies (mAbs), are the highest selling class of biotechnological medicines (Hong et al., 2018) with over 1500 new drugs in the development pipeline in 2016 expecting $138 billion global sales by 2024 (O. Yang et al., 2020).
Chinese Hamster Ovary (CHO) cell cultures are nowadays the preferred host platform to produce mAbs. The development of a successful biopharmaceutical molecule starts with the selection of high performing cell lines, which are scaled-up from the laboratory to the

[*] Barberi, G., Benedetti, A., Diaz-Fernandez, P., Sévin, D. C., Vappiani, J., Finka, G., Bezzo, F., Barolo, M., Facco, P. (2022). Integrating metabolome dynamics and process data to guide cell line selection in biopharmaceutical process development. *Metabolic Engineering*, **72**, 353-364.

manufacturing scale. Since scale-up is a multi-stage, expensive, resource-intensive and time-consuming process, pharmaceutical companies are looking at effective solutions to accelerate the development of mAbs, while preserving the desired product quality (Le et al., 2018). In fact, the reduction of time to market through an early identification of commercial cell lines that can be rapidly progressed to commercial-scale production has a major impact on the economics of biopharmaceutical drug development (Rameez et al., 2014).

Selection of commercial cell lines is performed through the screening of thousands different cell lines because the biological difference between cell lines usually has a major impact on the final product quality attributes (QAs) and process performance. Accordingly, developing a science-based strategy to identify the most productive and stable cell lines is a critical aspect of bioprocess development, which is usually carried out using a limited number of QAs (Facco et al., 2020). Cell growth, specific productivity, cell stability, and product titer are among the most important QAs that determine the quality target product profile in mAbs, and they have to be optimized along the development process (F. Li et al., 2010). However, only limited biological information on a handful of extracellular metabolites (Facco et al., 2020; Sokolov et al., 2017) are usually exploited for this purpose. The cell selection, as well as the in-depth process understanding and the optimization of the cultivation process could benefit from the extraction of the wealth of information retained in the biological profiling, such as in metabolomics, which identifies the intracellular or extracellular metabolites related to cell metabolism through liquid chromatography–mass spectrometry (B. Zhou et al., 2012).

The integration of data from the cultivation process and the biological profiling can effectively relate cell physiological state to industrially relevant phenotypes. In fact, data analytics and machine learning on metabolomics and culture information are widely used to relate cell physiological state to culture information or QAs. For instance, they have been demonstrated to provide valuable insight into the physiological difference between the cell lines producing the protein of interest and the parental ones (Dietmair, Hodson, Quek, Timmins, Gray, et al., 2012), the discrimination between cultures with different cell densities (Karst et al., 2017), and the characterization of the basal physiological state of cells during fed-batch operation at different bioreactor scales (Vodopivec et al., 2019). Furthermore, metabolomics has been linked to CHO process performance indicators and QAs to attain a better process understanding and for prediction purposes. Specifically, metabolomics has been used to discriminate CHO cell productivity and study the metabolic differences between high and low producing cells (William Pooi Kat Chong et al., 2012), to predict the glycan profile using a limited number of experiments (Zürcher et al., 2020), and to forecast product titer in such a way as to identify metabolites promoting or inhibiting the product titer (Morris et al., 2020). However, two main limitations are identified in the present literature, whose exploration would bring great advantages to bioprocess development. In the first place, the dynamics of untargeted metabolomics is often unconsidered explicitly, especially in machine learning applications to mAbs production and development. Examining the dynamics of metabolomics can provide a better understanding of the sequence of metabolic changes occurring over the cultivation process, allowing to determine the point along the culture time course providing the largest information for the identification of industrially relevant phenotypes. Secondly, metabolomics is typically related to a single QA or process parameter, but the integrated analysis of process (i.e., culture) variables and biological data (i.e., metabolomics) is typically not considered. This would provide greater insight into the cultivation process, also permitting a better management of the process. Furthermore, the study of the correlation between the specific cell physiological state and process behavior allows to improve the host cell lines through metabolic engineering and to develop a more informed and robust cell selection strategy. Accordingly, this study is aimed at integrating process and biological information from CHO cultivation explicitly exploiting the dynamics of the available process and biological data to accelerate cell line selection during biopharmaceutical process development. An industrial case study concerning development of mAbs and cell selection at micro-bioreactor scale (AMBR15$^{\text{TM}}$) is considered in this work. The proposed methodology will allow to deeply understand the cell lines behavior at the early development stages, to identify how the biological phenomena occurring in the culture change during the time course, and to predict the culture QAs, identifying at the same time the most important biomarkers associated to the observed cellular behavior.

3.2 Materials and methods

In this Section the data and the mathematical methodologies used in this Chapter will be briefly presented.

3.2.1 Available data

Data of two experimental runs performed in the AMBR15$^{\text{TM}}$ miniature bioreactor system (Sartorius Stedim Biotech, Sartorius AG, Goettingen, Germany) are available from the cultivation process. These data refer to $N = 96$ CHO clonal cell lines expressing the same therapeutic antibody that are cultured for 15 days in the 48 parallel 15 mL bioreactors of AMBR15$^{\text{TM}}$ (i.e., experimental batches). All production runs in both experiments were run using the GSK proprietary platform process. This process is performed in fed-batch using glucose as the main carbon source. Process conditions such as bolus feeding addition, pH and temperature were the same for all microbioreactors.

To allow for a larger number of clones to be screened in parallel, no biological replicates were performed in these experiments.

A total of $V_{\text{P}} = 7$ variables were measured in $T = 7$ time instants during the experimental batch ($t = 1, 2, \dots, T$, namely 0, 3, 5, 8, 10, 13 and 15 days): *viable cell concentration* (*VCC*), *product titer*, and nutrients and byproducts such as extracellular *glucose*, *glutamine*, *glutamate*, *lactate* and *ammonium*. The process variables were arranged in a three-dimensional array $\underline{\mathbf{X}}_{\mathrm{P}}[N \times V_{\mathrm{P}} \times T]$ = [96 cell lines × 7 process variables × 7 time instants], which is thereafter defined as process dataset.

Metabolomic data referring to the same experimental runs performed in the AMBR15$^{\mathrm{TM}}$ for the same CHO cell lines are available, as well. Metabolites were harvested from both cell pellets and culture supernatants (intra- and extracellular, respectively) and analyzed by flow injection liquid chromatography–mass spectrometry (LC-MS; Fuhrer et al., 2011). LC-MS measurements were performed in negative ionization mode with a scan range of mass over charge (*m/z*) 50-1000. Raw LC-MS data were preprocessed through an in-house pipeline prior to the statistical analysis (Frederick et al., 2020; Perrin et al., 2020). Of note, detected ions are tentatively annotated as metabolites solely based on accurate mass, with the inherent limitation that isomers or metabolites with masses within the annotation tolerance cannot be distinguished and some ions are ambiguously annotated as multiple tentative metabolites.

Metabolomic profiling was performed in $R = 2$ replicates in the same T time instants as in the culture analysis. Note that the intracellular metabolomic profiles are missing at time instant $t = 2$, because the number of cells in the cultures was insufficient to perform the analytical testing at that stage of the culture growth. In the following, the dynamic evolution of the metabolomic profiles will be referred as to metabolomic profiles dynamics. The preprocessed intra- and extra-cellular metabolomic profiles, consisting in intensities of $V_{\mathrm{I}} = 4587$ and $V_{\mathrm{E}} = 4489$ ions, respectively, were arranged in two four-dimensional arrays $\underline{\mathbf{X}}_{\mathrm{ic}}\ [N \times V_{\mathrm{I}} \times (T-1) \times R]$ and $\underline{\mathbf{X}}_{\mathrm{ec}}\ [N \times V_{\mathrm{E}} \times T \times R]$ = [96 cell lines × no. of ions × 7 time instants × 2 replicates]. The ions with more than 20% of missing intensities were excluded from the analysis, while the remaining missing data were imputed with a missing data replacement technique (Troyanskaya et al., 2001). This method assigns missing values as the weighted average intensity of the K_{miss} metabolites with the intensity profiles that are more similar to the metabolite of interest. This method was selected because it showed robust imputation performance in a high-dimension dataset even with large percentages of missing data. In this study, K_{miss} =15 metabolites were used to minimize the imputation error. Metabolites highlighted in this work were checked to ensure that the amount of missing data was largely below the 20% threshold to avoid any artifacts produced by data replacement technique.

3.2.2 Methodology to integrate process and biology

In this work, the integration of process and biological dynamic data is exploited for: *i)* process understanding, *ii)* studying the time course changes in process and biology, and *iii)* QAs and cell performance estimation and metabolic phenotypes identification. To systematize the integration of process and biological data through data-driven techniques, a multistep procedure (Figure 3.1) is used to conduct any analysis based on the application of data-driven techniques. First of all, some preliminary steps are required to correctly set up the analysis. Specifically, the aim of the analysis and the possible industrial advantages have to be defined (step 1), the available process and biological data have to be identified and prepared for the analysis (step 2), and finally the most appropriate (multivariate) mathematical/statistical technique has to be selected (step 3, Section 3.2.3). Once the preliminary steps are completed, the statistical model is built (step 4) and its performance is evaluated to understand the goodness of data representation (step 5). The last step consists in the interpretation of the model outcomes (step 6), which comprises the in-depth understanding of the samples and their relationships, the identification of the relevant process and biological behaviors, and a joint study of them.

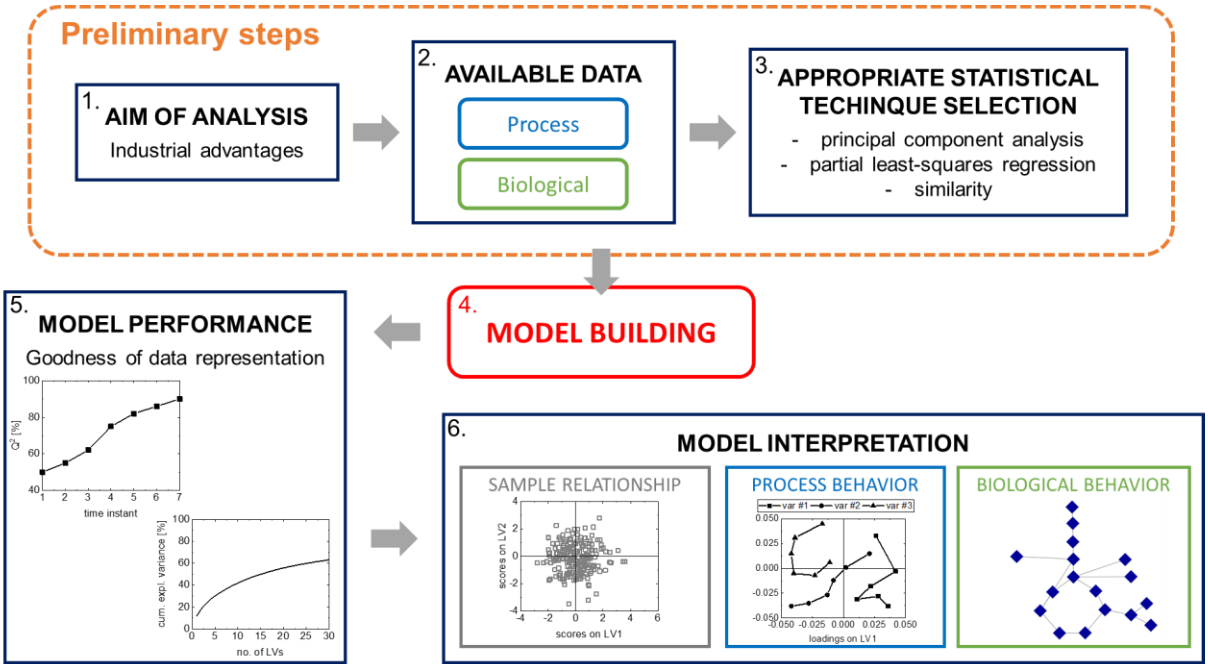


Figure 3.1 *Procedure to carry out data-driven activities aimed at integrating process and biological dynamic information.*

3.2.3 Multivariate statistical analysis

Prior to statistical analysis, data were unfolded to account for measurement replicate variability (Appendix B.1). Process data were autoscaled to zero mean and unit variance to account for differences in measurements units, while metabolomic data were pareto scaled (Eriksson et al., 2006), dividing each ion's intensity by the square root of its standard deviation, to avoid amplification of noisy measurements.

The most appropriate mathematical methodology is selected according to the objective of the work. This paper is primarily aimed at better understanding the relationship between QA or process performance and biological data. Furthermore, it is intended to provide a more confident cell selection through the understanding of the biological function (e.g., metabolic pathways) occurring during the culture course. Accordingly, we adopted multivariate statistical techniques. Despite being linear models, they allow an in-depth understanding of the high variability and the correlation structures among the variables of the system under study allowing a straightforward interpretation of the results. Furthermore, they permit to effectively monitor cell line performance (R. C. Pinto, 2017). All these tasks can be performed even when the amount of explained data variability is small (Kjeldahl & Bro, 2010). In fact, multivariate statistical techniques are consolidated methodologies in metabolomics studies (Paul & de Boves Harrington, 2021; Worley & Powers, 2013) because of their capability of producing robust and reliable models while handling highly dimensional, noisy and collinear data often comprised by a small number of samples (Trygg et al., 2007). In fact, the multivariate approach allows identifying the metabolic features that are common to all cell lines with desired culture behavior, accounting for the variability provided by the entire available dataset. Finally, multivariate statistical techniques can handle dynamic data and capture the time-course correlations in variables (Boccard & Rudaz, 2014; Smilde et al., 2010). Specifically in this paper, we adopted multiway principal component analysis to map the metabolomics profile dynamics in relation to process variables and to anticipately infer the process behavior of cell lines from the changes in their metabolomic profiles. Similarity analysis was utilized to study the changes in the metabolomic profiles along the culture time course through a straightforward metric. Multi-block multiway principal component analysis was considered to correlate the dynamics of metabolomic profiles and process variables and to understand how the time course variation in specific metabolites correlates to the desired process behaviors. Multiway partial least-squares regression was utilized to correlate directly metabolomic profiles dynamics and QAs by estimating their time course trajectory.

Multiway principal component analysis (MPCA; Nomikos and MacGregor, 1994) is a dimensionality reduction technique dealing with data dynamics (Appendix B.2), in which properly unfolded data (e.g., as $\mathbf{X}_{\mathrm{I}}\,[N \cdot R \times V_{\mathrm{I}} \cdot (T-1)]$ or $\mathbf{X}_{\mathrm{E}}\,[N \cdot R \times V_{\mathrm{E}} \cdot T]$) are decomposed in principal components (PCs). MPCA loadings capture the correlation between original variables (e.g., metabolites) and how variables are auto-correlated in time and cross-correlated with the dynamics of other variables. Furthermore, MPCA allows the real time mapping of observations by iteratively decomposing the data up to each instant t (with $t = 1, 2, \dots, T$) and completing the missing measurements for the remaining part of the experimental batch (from $t+1$ to $t = 7$) with the respective average values calculated over the calibration data used to build the model (Ramaker et al., 2005).

The similarity factors (Facco et al., 2020; Krzanowski, 1979) compare the direction of maximum variability of two datasets, namely the metabolomic profiles at different time instants, allowing to assess the similarity in their major driving forces (Appendix B.3). The similarity factor is bounded between 0 and 1, indicating 0 the absence of similarity in the data driving forces and 1 the same dataset driving forces. This similarity factor provides a metric to assess the similarity among the metabolomics profiles of all cell lines at different time instants.

Multi-block principal component analysis (MB-PCA; Westerhuis et al., 1998) is a dimensionality reduction multi-block technique dealing with different types of data organized in separated blocks, such as process and metabolomic data (Appendix B.4). MB-PCA captures in the loadings the correlation between variables in the same blocks and the cross-correlation between variables in different blocks. Since dynamic data are available in this study, MPCA and MB-PCA are combined in multi-block multiway principal component analysis (MB-MPCA) to consider the correlation in time between variables which pertain to different data blocks. MP-MPCA was performed by horizontally concatenating process and metabolomic data and performing a standard MPCA. Process data were autoscaled and metabolomic data ware Pareto scaled; no block scaling was performed to avoid underestimating the importance of the metabolomics block which comprises a large number of variables (Westerhuis et al., 1998).

Multiway partial least-squares (MPLS; Nomikos and MacGregor, 1995a) is a multivariate statistical regression technique which deals with data dynamics and can be used for estimation, prediction and classification. MPLS identifies the direction of maximum covariance between properly unfolded regressors (e.g., $\mathbf{X}_{\mathrm{E}}\ [N \cdot R \times V_{\mathrm{E}} \cdot T]$) and a matrix $\mathbf{Y}\ [N \cdot R \times M]$ of M responses (e.g., any vertical slice of $\underline{\mathbf{X}}_{\mathrm{P}}$), and decomposes $\mathbf{X}_{\mathrm{E}}$ and $\mathbf{Y}$ into a reduced space of A latent variables LVs (Appendix B.5). Model performance was evaluated through a 250-iterations Monte Carlo cross-validation, in which samples are randomly split in calibration and validation sets (88% of samples is for calibration). External validation cell lines were randomly selected from the initial dataset (12 cell lines) and were used to assess model robustness and generalization performances.

Relevant metabolites for the regression models were identified through a bootstrap procedure (Afanador et al., 2013) on the Variable Importance in Projection index (VIP; Eriksson et al., 2006) (Appendix B.6). Through bootstrap over the VIP, the lower 90% confidence limit (VIP_{LCL}) of each ion at a time instant was calculated and the top 5% of ranked variables were selected. This threshold was selected based on a sensitivity analysis in such a way as to guarantee a good compromise between the prediction performance, namely the amount of explained $\mathbf{Y}$ variability, and the selection of a reduced number of ions, which permits the model interpretability. A new PLS model was built after the variable selection, showing improved prediction performance, whose variables with a $VIP_{\mathrm{LCL}} > 1$ are largely important for the $\mathbf{Y}$ response estimation.

All software were implemented in Matlab® 2019b (MathWorks, Natick, Massachusetts, USA) using in-house developed codes and the PLS Toolbox (Eigenvector Research Inc, Wenatchee WA, U.S.A.).

3.3 Results and Discussion

3.3.1 Process understanding

In this Section, the integration of process and biological data is presented for a deeper process understanding. For this purpose, the metabolome dynamics of cell lines will be mapped according to process performance and the time course similarity in biological phenomena will be studied.

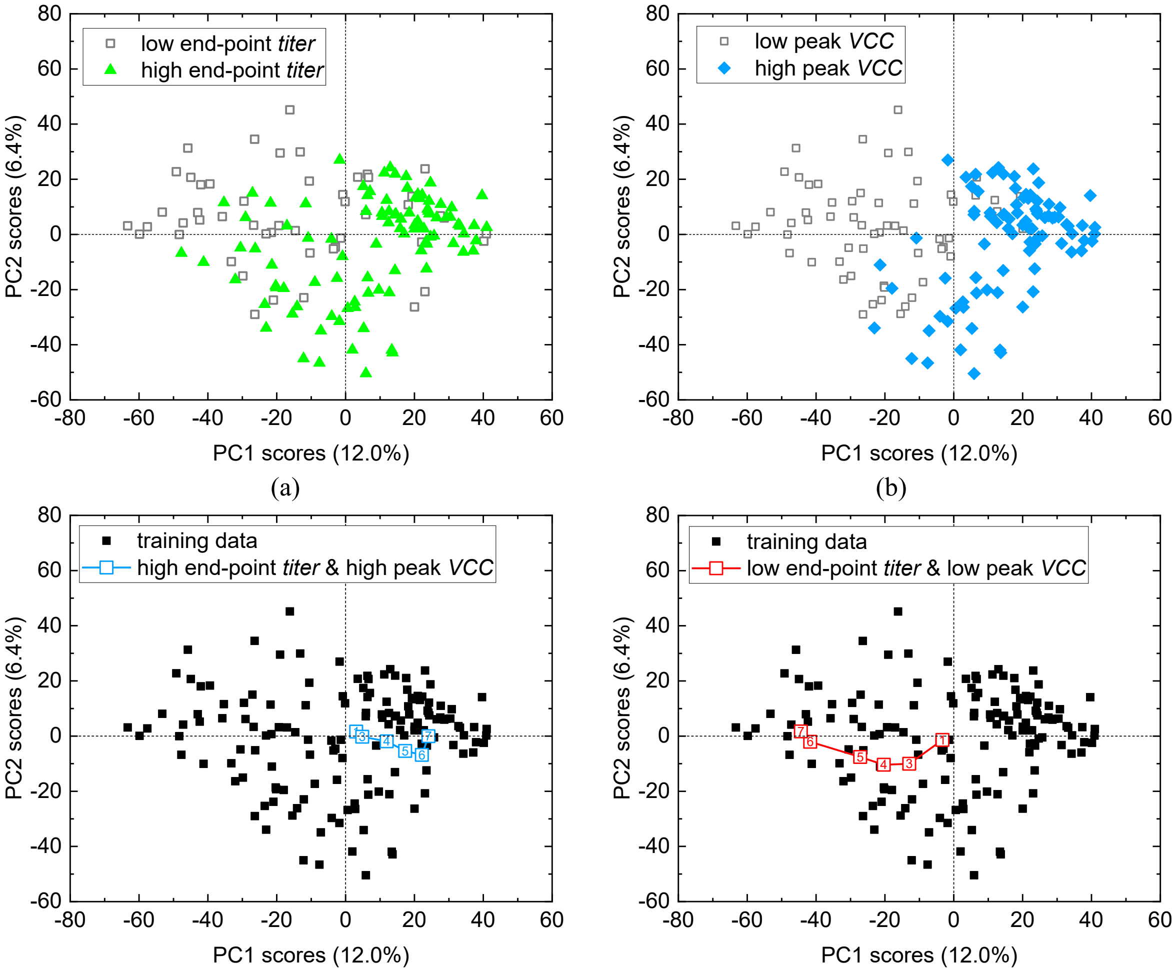


Figure 3.2 *Score space of PC1 and PC2 for the MPCA model on* $\mathbf{X}_1$*: (a) mapping of high end-point titer cell lines; (b) mapping of high peak VCC cell lines; (c) quasi-real-time mapping of a high end-point titer high peak VCC cell line; (d) quasi-real-time mapping of a low end-point titer and low peak VCC cell line. The numbers in Figure (c) and (d) refer to the time instant* t *(*t *= 1, 3, …, 7) at which the real-time mapping is performed.*

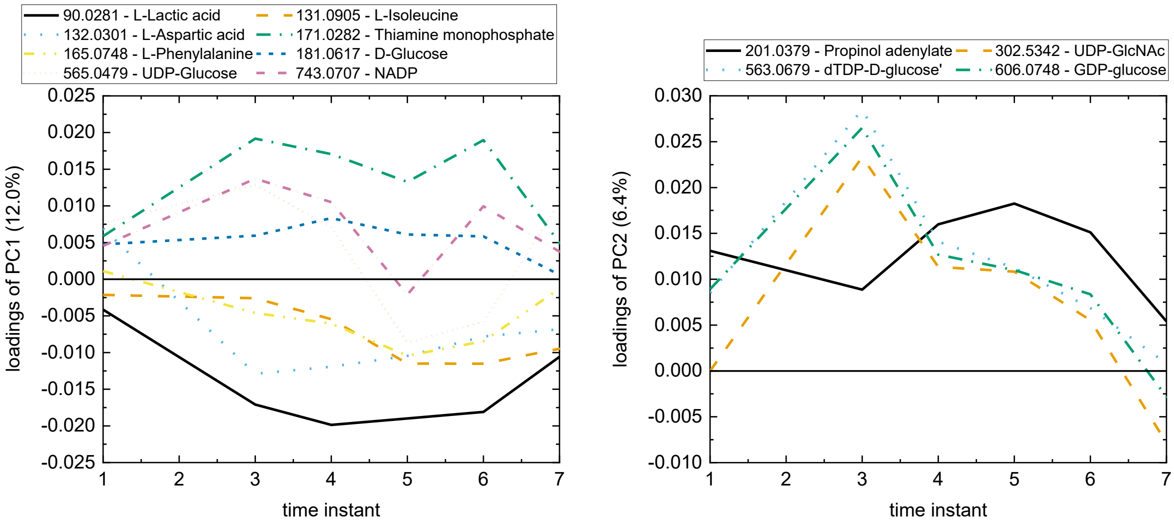
3.3.1.1 Metabolic mapping of cell lines according to process performance

A better comprehension of the correlations in metabolomics profile dynamics and their relations with process variables and QAs can provide insights on the underlying biological differences between cells in the production of different mAbs and enhance confidence in cell selection. For this reason, in this Section process and biological information are integrated by mapping the metabolomic dynamics of the cell lines in relation to the process variables for the purpose of an in-depth process understanding, process monitoring and early cell line screening. MPCA Appendix B.2; Nomikos and MacGregor, 1994) on intracellular metabolomic data $\mathbf{X}_\mathrm{I}$ is adopted to map cell lines performance according to end-point *product titer* and peak *VCC* to observe if a clear fingerprint of the end-point *product titer* and peak *VCC* is left on the intracellular metabolism. This methodology also allows the quasi-real-time quality monitoring of the culture.

The MPCA model captures about 40% of the metabolomic profile dynamics variability with 9 PCs, the first PC explaining 12% of the total variability. The relatively low captured variability indicates that metabolomic data dynamics is influenced by a complex series of independent chemical/biological phenomena. However, the low explained variability is not a concern since it does not imply a more accurate and descriptive model, while the use of more complex and non-linear modeling strategies does not necessarily guarantee to obtain more descriptive and interpretable models (Kjeldahl & Bro, 2010).

The score plot of PC1 and PC2 of Figures 3.2a and 3.2b shows the mapping of cell lines according to end-point *product titer* and peak *VCC*, respectively, where each point summarizes the dynamic evolution of cell metabolomic profiles along the entire culture. Accordingly, the distance between points is a metric of the differences between the dynamics of cell metabolomic profiles. In particular, Figure 3.2a shows that a certain degree of separation is evident between cell lines with either high or low end-point *product titer*. In fact, cell lines resulting in high end-point *product titer* have higher density in the space of positive PC1. On the other side, negative values of PC1 identify a space in which low end-point *product titer* cell lines have higher density. Accordingly, cell lines with high or low end-point *product titer* are characterized by different dynamics of metabolomic profiles.

Similarly, the PC1 vs. PC2 score space effectively maps the cell lines according to peak *VCC* (Figure 3.2b), where high peak *VCC* cell lines are typically located in the space of positive PC1. According to these results, the metabolomic profiles dynamics contains useful information that is strongly related to the cell lines performance during cultivation process and the QAs.



(a) (b)

Figure 3.3 *Loadings time trajectory of ions form the MPCA model on* $\boldsymbol{X_1}$*: (a) PC1; and (b) PC2. Ions are reported with the m/z value and a tentatively annotated metabolite.*

The dynamics of metabolomic profiles which are typical of cell lines showing industrially relevant phenotypes can be easily tracked and understood by inspecting the loadings of the MPCA model. Figure 3.3 shows the loadings time trajectory of ions captured by the MPCA model. Cell lines located in the part of the score space with positive PC1 (Figure 3.3a) are characterized by low amounts along the entire experimental batch of ion *m/z* 90.0281 (tentatively annotated to *L-lactic acid*), ion *m/z* 132.0310 (tentatively annotated to *L-aspartic acid*), ion *m/z* 165.0748 (tentatively annotated to *L-phenylalanine*), and ion *m/z* 131.0905 (tentatively annotated to *L-isoleucine*), and large amounts along the entire experimental batch of ion *m/z* 171.0282 (tentatively annotated to *thiamine monophosphate*) and ion *m/z* 181.0671 (tentatively annotated to *D-glucose*). Positive values of PC1 also describes cell lines showing in the first half of the experimental batch large amounts of ion *m/z* 565.0479 (tentatively annotated to *UDP-glucose*) and ion *m/z* 743.0707 (tentatively annotated to *NADP*). Differently, cell lines located in the part of the score space with positive PC2 (Figure 3.3b) are characterized by large amounts of ion *m/z* 302.5342 (tentatively annotated to *UDP-GlcNAc*), ion *m/z* 563.0679 (tentatively annotated to *dTDP-D-glucose*), and ion *m/z* 606.0748 (tentatively annotated to *GDP-glucose*) in the first half of the experimental batch, and ion *m/z* 201.0379 (tentatively annotated to *propinol adenylte*) for the entire experimental batch. Accordingly, cell lines with high peak *VCC* and end-point *titer* show high consumption or low availability of *L-aspartic acid*, *L-phenylalanine*, and *L-isoleucine*, and low production of *L-lactic acid*, while higher concentration of *D-glucose*, *UDP-glucose*, and *NADP* are maintained especially in the first half of the experimental batch. These results highlight the importance of understanding the impact of key amino acids, such as *L-aspartic acid*, *L-phenylalanine*, and *L-isoleucine*, which influence both cell growth and productivity and applying appropriate feeding strategies. Previous studies suggested careful monitoring of *L-phenylalanine* and *L-isoleucine* due to their high consumption rates (Ritacco et al., 2018) while appropriate feeding to find the right balance of *D-glucose* and *L-lactic acid* concentrations, has a strong correlation with cell performance. Furthermore, most of the cell lines assessed in this study (i.e., the ones located in the first quadrant) show a sustained presence of *UDP-GlcNAc*, *dTDP-D-glucose*, *GDP-glucose*, and *propinol adenylte* which is correlated to protein production.

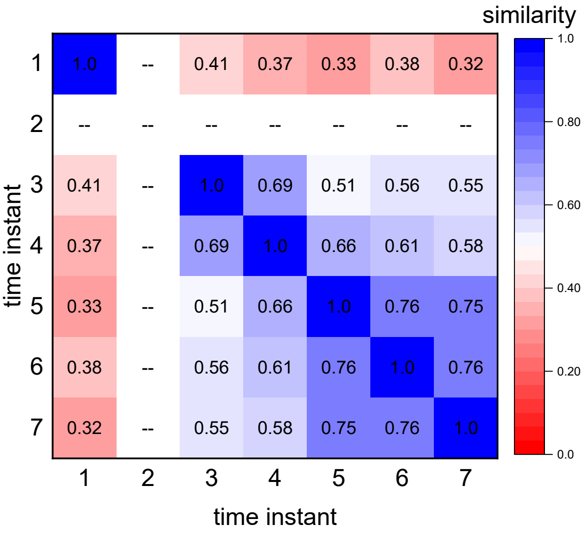
Since good mapping of cell lines according to process performance is achieved, the score space can be effectively used to track in real-time the cell line trajectories to monitor their performance while being processed. For this purpose, the previously developed MPCA model is used to real-time map 19 validation cell lines randomly excluded from the training dataset. Each validation cell line is projected onto the model space at each time instant t (t = 1, 3, …, 7), replacing the future missing measurement from time $t + 1$ to the end of the culture with the average values of the respective calibration dataset (Appendix B.2). Figures 3.2c and 3.2d show the projection on the score space of two validation cell lines as the experimental batch progresses. The validation cell line shown in Figure 3.2c evolves toward the part of the score space with positive PC1 where cell lines resulting in high end-point *titer* and peak *VCC* lie. Differently, Figure 3.2d shows the case of a cell line evolving toward negative values of PC1, the zone of the score space where low end-point *titer* and peak *VCC* cell lines are located. Note that, in both cases, the culture status can be correctly inferred already at time instant $t = 4$, namely after approximately one half of the entire experimental batch length, providing an anticipated indication of cells with desired performance and additional confidence for their selection. This result shows that the length of experimentation in the AMBR15$^{\text{TM}}$ equipment can be potentially reduced without losing information on cell line performance and screening capabilities.

Furthermore, the time course changes of metabolomic profiles can be identified by the “shift” of the respective score point trajectory. For example, in Figure 3.2c since cell line shows small shifts in score space, the evolution during culture of the metabolomic profile is characterized by limited metabolic changes. On the contrary, in Figure 3.2d the cell line shows large shifts, meaning that the metabolomic profile changes significantly in different time instants during the experimental batch.

The analysis performed on extracellular metabolomic data $\mathbf{X}_{\mathrm{E}}$ through MPCA provides similar mapping of cell lines.

3.3.1.2 Time course similarity in biological phenomena

In this Section, the similarity between metabolomic profiles at different time instants is analyzed to understand when the main changes in the cell physiological state occur and to identify the metabolic phenomena characterizing each cell growth phase Such information is crucial to pinpoint at what stage the cell metabolism is stable during the experimental batch time course and could provide a better characterization of process performance and QAs. This can be a valuable information to reduce the analytical burden required to perform metabolomic measurements. For this purpose, the similarity factors (Facco et al., 2020) are used to compare the main biological driving forces in the intracellular metabolomic profiles in $\mathbf{X}_\mathrm{I}$ at different time instants (Appendix B.3). To calculate the similarity factors, PCA models are built at each time instant t (t = 1, 3, …, 7), each one capturing ~50% of the metabolomic profiles variability with 6 PCs at each time instant t. It is worth noting that these PCA models capture a larger portion of variability with less PCs than the model presented in Section 3.3.1.1, because they are built on the metabolomic profiles at each time instant rather than on the entire metabolomic profiles dynamics.



time instant	1	2	3	4	5	6	7
1	1.0	--	0.41	0.37	0.33	0.38	0.32
2	--	--	--	--	--	--	--
3	0.41	--	1.0	0.69	0.51	0.56	0.55
4	0.37	--	0.69	1.0	0.66	0.61	0.58
5	0.33	--	0.51	0.66	1.0	0.76	0.75
6	0.38	--	0.56	0.61	0.76	1.0	0.76
7	0.32	--	0.55	0.58	0.75	0.76	1.0

Figure 3.4 *Similarity pattern between metabolomic profiles at different time instant during cell culture: intracellular data. Data at time instant t=2 are missing.*

The similarity factors between metabolomic profiles at different time points are shown in the heatmap of Figure 3.4. Each row reports the similarity between metabolomic profiles at one time instant and all other time instants. Metabolomic profiles at time instant $t = 1$ show low similarity with later time instants ($S_{t't''} < 0.42$). From time instant $t \geq 3$ metabolomic profiles show larger similarity ($S_{t't''} > 0.6$) with the contiguous time instants. Furthermore, all metabolomic profiles from time instant $t = 5$ to the end of the culture (i.e., time instant $t = 7$) show high similarity ($S_{t't''} > 0.7$). Since the major driving forces (i.e., the main variability directions into the data) are similar in each cell growth phase, the main biological phenomena underlying the metabolomic data remain the same in each growth phase while change in the subsequent phases. Specifically, the cell physiological states vary moving from the exponential growth phase (time instant $t = 1$) to the stationary phase (time instant $t = 3$), while remain quite stable along the stationary phase (time instants $t = 3$ and $t = 4$). Similarly, changes are detected in the decline phase (time instant $t = 5$ to $t = 7$), while higher metabolic stability is detected throughout the entire decline phase.

The biological phenomena (i.e., metabolic pathways) preserving high importance in the metabolomic profiles of each phase are identified according to the ions responsible for the similarity across different time instants (Appendix B.3). The metabolic phenomena with high importance in each culture phase are reported in Table 3.1. It is worth noticing that important pathways are obtained from the tentatively annotated metabolites; hence, further investigation is required to confirm their identity.

Similar results are observed for the extracellular metabolomic data $\mathbf{X}_\mathrm{E}$.

Table 3.1 *Metabolic phenomena (i.e., metabolic pathways) with high importance in each culture phase identified though the similarity analysis. These important pathways are obtained from the tentatively annotated metabolites.*

Culture phase	Key metabolic pathways
Exponential growth phase (time instant t=1)	*Pyrimidine metabolism* [cge00240] *Tryptophan metabolism* [cge00380]
Stationary phase (time instant t= 3, 4)	*Alanine, aspartate and glutamate metabolism* [cge00250] *Ascorbate and aldarate metabolism* [cge00053] *Cysteine and methionine metabolism* [cge00270] *Tyrosine metabolism* [cge00350] *Vitamin B6 metabolism* [cge00750] *Pyrimidine metabolism* [cge00240] *Tryptophan metabolism* [cge00380]
Decline phase (time instant t=5, 6, 7)	*Amino sugar and nucleotide sugar metabolism* [cge00520] *Cysteine and methionine metabolism* [cge00270] *Glutathione metabolism* [cge00480] *Phenylalanine metabolism* [cge00360] *Purine metabolism* [cge00230]

As a result, we have shown that the physiological state of cells does not change significantly in the second half of the experimental batches and specifically in the decline phase. Accordingly, this stability can be exploited to infer the process performance at the beginning of the decline phase (i.e., time instant $t = 5$) for anticipating the selection of promising cell lines.

3.3.2 Time course changes in process and biology

In this Section, the time trajectories of process data and the dynamics of metabolomic data are integrated to gain a better process understanding on how time course changes in process and biological data are related. The understanding of the relation between biological phenomena and process variables and QAs accelerates the bioprocess scale-up allowing an anticipated and more informed cell selection, by generating insight into the relation of metabolites with industrially relevant phenotypes. For this purpose, a MB-MPCA (Appendix B.4) is built on intracellular metabolomic data ($\mathbf{X}_\mathrm{I}$) and batch-wise unfolded process data $\underline{\mathbf{X}}_\mathrm{P}$ (as $\mathbf{X}_\mathrm{P}$ $[N \cdot R \times V_\mathrm{P} \cdot T]$) and captures 39.1% of data variability with 9 PCs. The limited captured variability is due to the vast number of available variables and several correlations that are simultaneously taken into consideration (i.e., process-metabolomics correlation, block variable correlation, correlation in time, and biological and measurement variability). MB-MPCA allows studying correlations at two levels: 1) correlation between variables inside each block; 2) correlation between variables of different blocks.

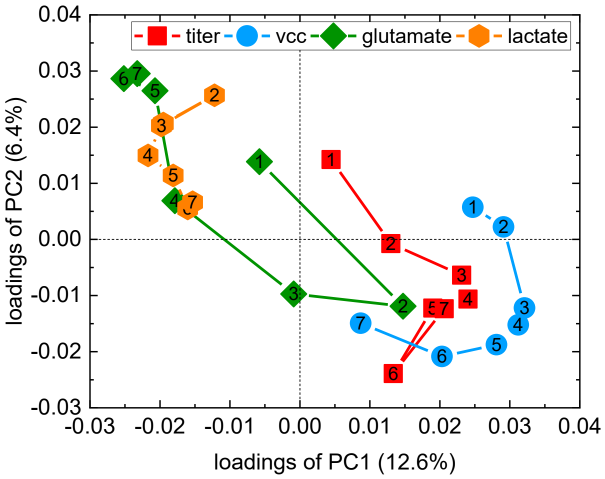
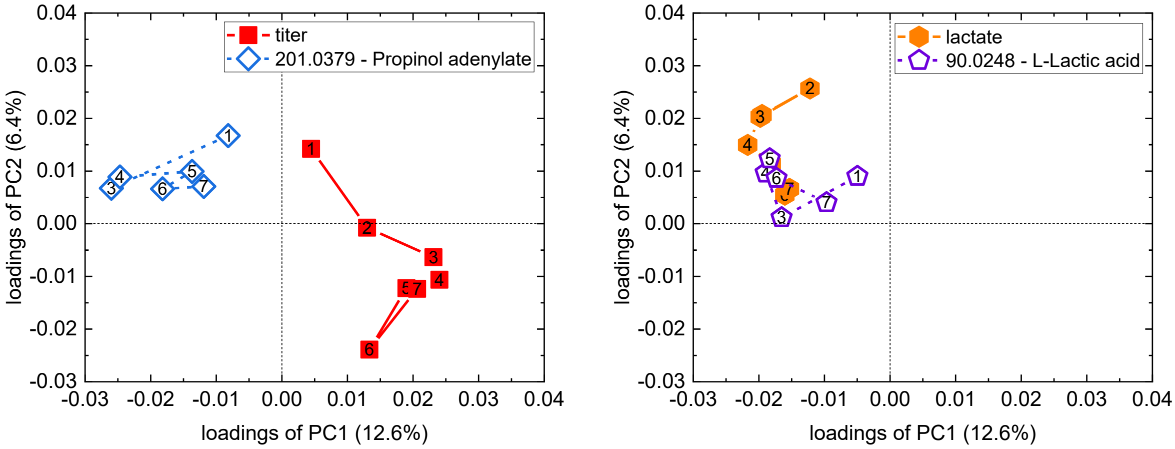


Figure 3.5 *MB-MPCA model on* $\mathbf{X}_I$*: loadings plot showing the correlation between process variables time trajectories. The numbers refer to the time instant.*

At the first level, the correlation structure between the main culture variables is described by the MB-MPCA loadings of the process block in Figure 3.5. The model captures the anticorrelation of *product titer* and *VCC* with *lactate* and *glutamate*, which represents 12.6% of the total data variability on the first PC, and the time course trajectory of process variables such as *product titer*, *glutamate* and *lactate*, which represents 6.4% of the total variability on the second PC. *Product titer* and *VCC* evolve similarly from the first to the fourth quadrant in the first half of the culture (i.e., $t \leq 4$), indicating that the evolution of these variables is strictly connected and related to similar metabolites (and accordingly to similar metabolic reactions). In the second half of the culture, *product titer* and *VCC* remain located in the fourth quadrant with a different time evolution, indicating a variation in the cell behavior. This fact is probably due to cells entering the decline phase, in which the *product titer* continues to increase while the *VCC* starts to decrease. *Lactate* evolution in time is located in the second quadrant, resulting anticorrelated to *product titer* and *VCC* for the entire cell culture. This indicates that batches with a low *lactate* concentration usually exhibit higher *product titer* and *VCCs*. Similarly, at the end of the culture glutamate evolves mainly in the second quadrant and is anticorrelated to *product titer* and *VCC* from $t = 4$ on. This confirms that cell lines showing low concentration of *glutamate* are usually the ones exhibiting higher *titers* and *VCCs*. This relation is expected as high concentration of *glutamate* enhance its metabolism which produces *ammonium* a known inhibitor of cell growth (Ozturk et al., 1992). However, at $t = 2$ *glutamate* is positively correlated with *product titer*, indicating that cell cultures with high *glutamate* at this time instant exhibit in general higher *product titers*. As can be noticed, moving from positive to negative values of PC2 *product titer*, *VCC* and *lactate* generally increase together during culture, while *glutamate* usually decreases along the culture. This is a reasonable outcome, because mAbs and *lactate* are secreted by the cells while *glutamate* is fed to the culture and is consumed by cells. The same loading plot identifies that *glucose* is correlated with *titer* and *VCC* at time instants $t < 3$ and anticorrelated thereafter, while *glutamine* is anticorrelated to *glutamate* in the second half of the culture (Appendix B.7; Figure B.1). Accordingly, cells with high *glucose* concentration in the initial part of the experimental batches and consuming more *glucose* are likely to exhibit higher *product titer* and *VCC*. The relation between *glutamine* and *glutamate* as cells convert the *glutamate* into *glutamine* during their metabolism.

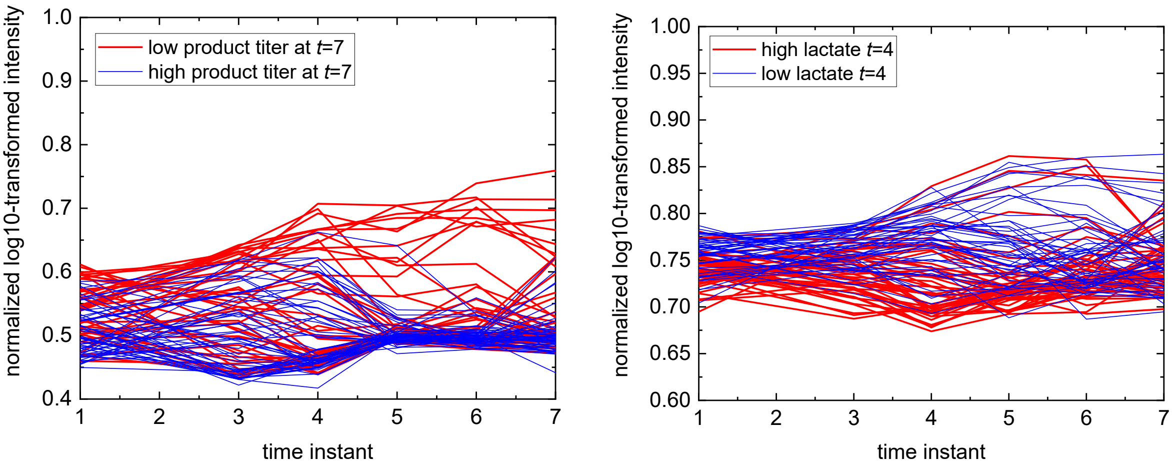


(a) (b)

Figure 3.6 *MB-PCA on* $\mathbf{X}_I$ *loadings plot reporting the correlation between process variables and intracellular metabolites: (a) ion anti-correlated with product titer, and (b) ion correlated with lactate. The numbers refer to the time instant. Ions are reported with the m/z value and a tentatively annotated metabolite.*

At the second level of analysis, the joint inspection of the blocks loadings relates the time course evolution of process variables to the metabolomics profiles dynamics. In particular, the metabolite dynamics mostly impacting the variability of the main process variables are shown in Figure 3.6, where some MB-MPCA loadings of the metabolomics and process blocks are reported.

Figure 3.6a shows the negative correlation between the ion *m/z* 201.0379, tentatively annotated as *propinol adenylate* (*propanoate metabolism* [cge00640]) or *4'-phosphopantothenoyl-L-cysteine* (*pantothenate and CoA biosynthesis* [cge00770]) with *product titer*. The dynamic trajectory of this metabolite is located in the second quadrant, indicating that it is anticorrelated to *product titer*, especially in the central part of the batch ($t = 3$, $t = 4$ and $t = 6$). *Propinol adenylate* is an intermediate in the production of *propanoyl-CoA* from *propanoate*, while *4'-phosphopantothenoyl-L-cysteine* is an intermediate in the *CoA* production form *L-cysteine*. Accordingly, cell cultures showing low amount of this ion (especially in the central part of the batch), probably due to low production of *propanoyl-CoA*, whose accumulation is known to produce several metabolic disfunction in humans (Wongkittichote et al., 2017), or more a sustained production *CoA*, an essential cofactor involved in several metabolic reactions, generally exhibit high *product titers* (Figure 3.7a).



(a) (b)

Figure 3.7 *Normalized log10 transformed ions intensity time profiles of all available cell lines: (a) ion m/z 201.0379 (tentatively annotated to propinol adenylate), and (b) ion m/z 90.0248 (tentatively annotated to L-lactic acid).*

Figure 3.6b shows a large correlation between the ion *m/z* 90.0248 tentatively annotated as *L-lactic acid* or its isomer *3-hydroxypropionic acid* (*propanoate metabolism* [cge00640]) and culture *lactate*. In the central part of the experimental batches the correlation is maximum, indicating that high intensity of *L-lactic acid* is generally found in cultures exhibiting high *lactate* and low *product titers* (Figure 3.7b). According to these correlations, this inspected ion is with high probability originated from *L-lactic acid*, which is the intracellular molecule that is excreted by cells as *lactate*.

Similar correlations between culture variables are observed for extracellular data.

This analysis demonstrates to be important for both the process understanding and the identification of metabolic characteristics of cell with desired phenotypes, which can be exploited for a more informed cell selection. However, despite being general and applicable to all type of biological data, the adopted unsupervised technique might be difficult to interpret. Hence, an easier understanding or the relation between process and metabolomics can be achieved through supervised methods, such as partial least-squares regression.

3.3.3 QAs and cell performance estimation and biological phenomena identification

In this Section, we integrate the process and biological information by estimating a QA from metabolomic dynamic data to achieve a better understanding of the metabolic behavior of cells showing promising performance and linking relevant cell phenotypes to several metabolites (i.e., biomarkers). This would provide precious insights for the improvement of the host cell through metabolic engineering and for a more informed and robust cell line selection, which can even be anticipated by extracting identified biomarkers through targeted metabolomic analysis at earlier development stages. For this purpose, a MPLS (Appendix B.5) is built to predict the *product titer* time trajectory $\mathbf{Y}\,[N \cdot R \times 1 \cdot T]$ (obtained from $\underline{\mathbf{X}}_{\mathrm{P}}$) from the intracellular metabolomic profiles dynamics ($\mathbf{X}_{\mathrm{I}}$). Since a large number of ions in time are available, the less informative ones are eliminated through VIP bootstrap (Appendix B.6) to reduce the number of regressors and improve model performance.

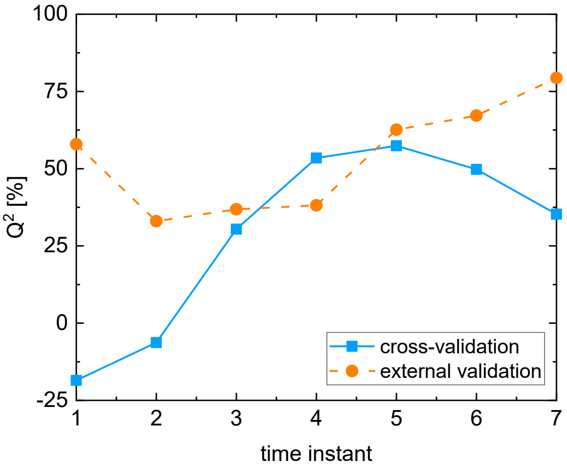
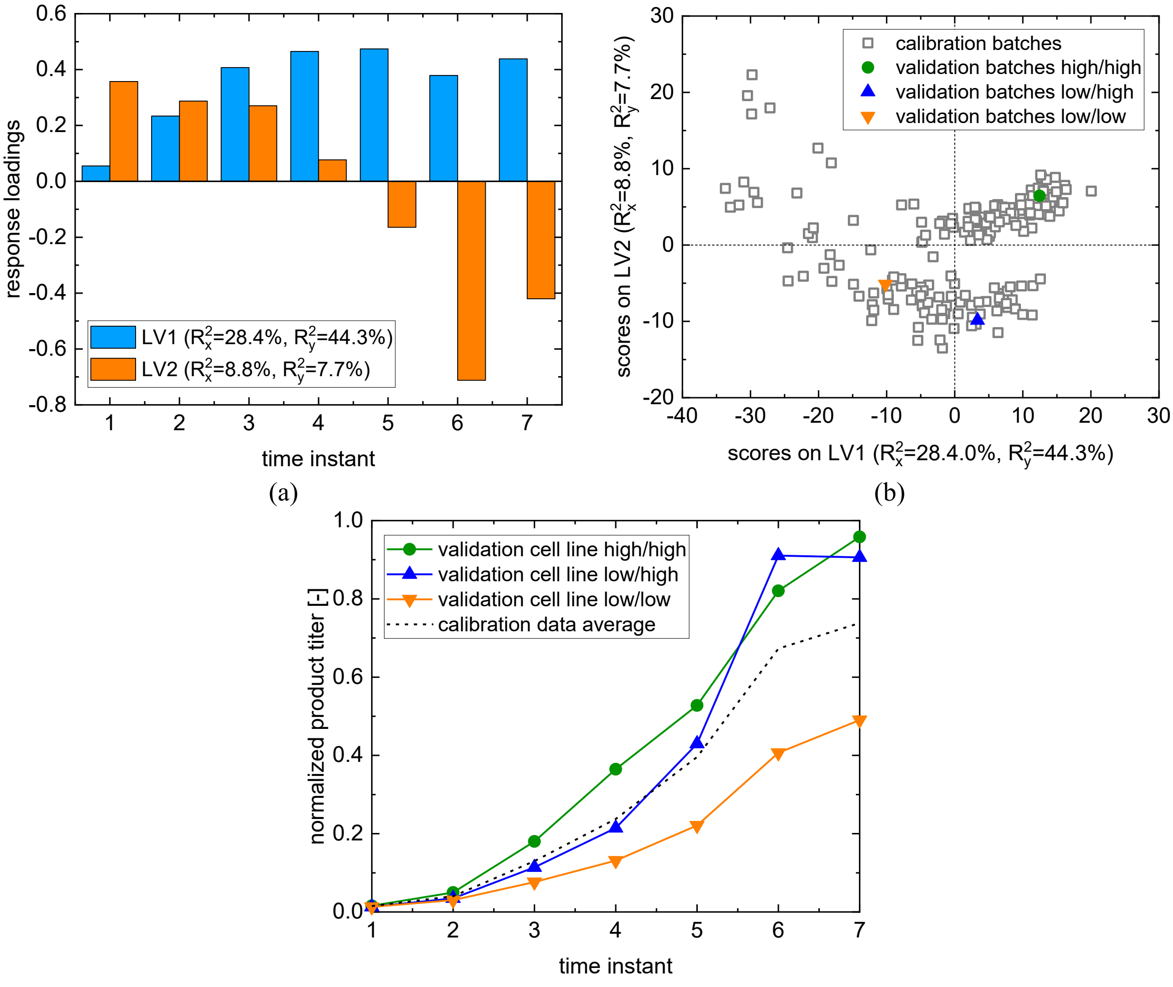


Figure 3.8 *MPLS model for the estimation of product titer time trajectory: coefficient of determination in cross-validation (cyan squares) and external validation (orange circles).*

The coefficient of determination in cross-validation (Q^2) of the developed two-LVs MPLS model (capturing 51.9% of $\mathbf{Y}$ variability with 37.3 % of $\mathbf{X}_{\mathrm{I}}$ variability) is shown in Figure 3.8 as cyan squares. *Product titer* estimation is acceptable in the second half of the experimental batch (time instant $t > 3$), in which $Q^2 > 40\%$ is observed, while in the initial days of culture, namely time instants $t = 1$ and $t = 2$, *titer* is poorly estimated ($Q^2 < 40\%$). The estimation performance suggests that information regarding the *product titer* at time instants $t = 1$ and $t = 2$ is hidden by nonsystematic variability due to measurement noise or other biological phenomena.

External validation cell lines (orange circles in Figure 3.8) are estimated with $Q^2 > 30\%$ at all time instants. Specifically, low Q^2 (~30%) is observed at time instant $t \leq 4$, while $Q^2 > 60\%$ is observed in the second half of the experimental batch, namely at t ≥5. Accordingly, the model is sufficiently robust to provide a good description of the *product titer* in the final part of the experimental batches, while it provides a limited explanation of the *product titer* in the initial day of culture. In fact, since during the exponential growth phase (time instants t=1 and t=2) cell lines metabolism promotes growth, rather that protein production, the changes in the metabolic state (i.e., variations in ions intensities) of cells are more related to cell growth than to protein production. This leads to a poor performance in *product titer* estimation. Conversely, during the stationary and decline phases, cells shift their metabolic behavior to protein production, allowing for an accurate estimation of the product titer.

The response loadings of the first two LVs of the MPLS model (which explain the largest part of the *product titer* variability, Figure 3.9a) report the autocorrelation in the *product titer* along culture time course and allows to connect *product titer* with metabolomics profile dynamics.



(a) (b)

(c)

Figure 3.9 *MPLS model on* $\mathbf{X}_1$ *for the estimation of product titer time profile: (a) response loadings, (b) score space, and (c) product titer time trajectory of three external validation cell lines (normalized between 0 and 1).*

Two main independent phenomena are identified in the *product titer* time trajectory, which are related to the ions captured by each LV. A positive autocorrelation in the *product titer* time trajectory, especially in the second half of the batch (after time instants $t \geq 3$, blue bars) explains the cell lines whose *product titer* remains high for positive values of PC1 (or low for negative values of PC1) for the entire experimental batch. Secondly, a positive autocorrelation in *product titer* time trajectory at time instants $t \leq 3$ and a negative one at the time instant $t \geq 5$ captures cell lines changing behavior between the first and second half of the experimental batch, meaning that they exhibit low *product titer* in the initial part of the batch, but high *product titer* in the final (or vice versa).

The MPLS score space (Figure 3.9b) relates the cell lines to product titer time trajectory and metabolomic profile dynamics. According to the response loading, high *product titer* cell lines for the entire experimental batch are located in the first quadrant, while cell lines with low *titer* in the initial part of the batch, but high titer in the final part, are located in the fourth quadrant. The second and fourth quadrant describes cell lines with low final *titer* despite the *titer* in the initial part of the culture is high or low.

External validation cell lines are correctly mapped into the score space according to *product titer* time trajectory behavior. In fact, a cell line with *product titer* above the average for the entire experimental batch is correctly mapped in the first quadrant (green circle), while cell lines with *product titer* below the average in the initial part of the culture are correctly characterized by negative values of LV2. This correct mapping is proved by the normalized *product titer* time trajectory of these validation cell lines (Figure 3.9c).

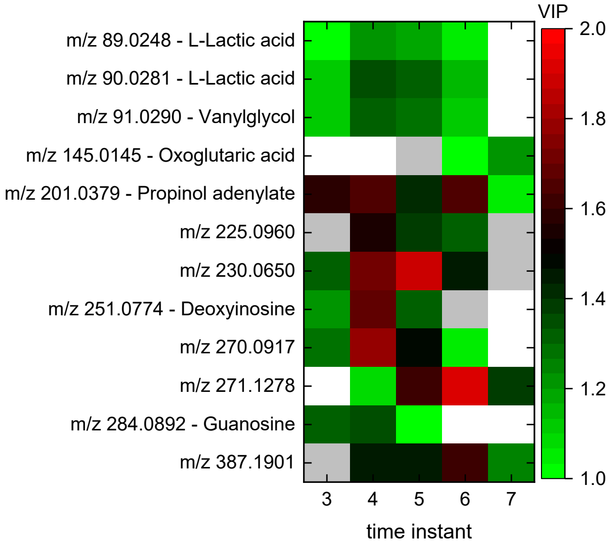


Figure 3.10 *MPLS model on* $\mathbf{X}_I$ *for the estimation of product titer time profile:* VIP_{LCL} *score of relevant ions. No ion at* t*=1 is found relevant by the model. White - ions not retained during the variable selection procedure; gray – ions with* VIP_{LCL} *< 1. Ions are reported with the m/z value and a tentatively annotated metabolite.*

The VIP_{LCL} identifies the ions with high importance for estimating the *product titer* time trajectory. Figure 3.10 shows the time course changes of VIP_{LCL} for some important ions. The ions *m/z* 201.0379 tentatively annotated to *propinol adenylate* (*propanoate metabolism* [cge00640]) or *4'-Phosphopantothenoyl-L-cysteine* (*pantothenate and CoA biosynthesis* [cge00770]) is important for the prediction of *product titer* time profile, showing a $VIP_{\mathrm{LCL}} > 1.6$ in the central part of the experimental batches. This result is in full accordance with what was observed through the MB-PCA model (Section 3.3.2). The ion *m/z* 90.0281 tentatively annotated to *L-lactic acid* (*Glycolysis/Gluconeogenesis* [cge00640]) or *3-hydroxypropionic* *acid* (*propanoate metabolism* [cge00640]) shows mild importance throughout the culture for the prediction of the *product titer* time profile, with peak importance at $t = 4$, $t = 5$ ($VIP_{\mathrm{LCL}} > 1.2$). Furthermore, another ion *m/z* 89.0248 tentatively annotated to *L-Lactic acid* or *3-hydroxypropionic acid* shows a similar importance profile during culture confirming a strict relation between these two ions. This result agrees with the MB-MPCA model, in which a negative correlation with *product titer* is found. Accordingly, this study confirmed from a metabolomic prospective the well-known relation between *lactate* and *product titer* (Facco et al., 2020; Sokolov et al., 2015). Furthermore, the possible presence of *3-hydroxypropionic acid* with *propinol adenylate* might indicate a strict relation between *product titer* and *propanoate metabolism* [cge00640].

Similar performance is achieved in the estimation of *product titer* time trajectory from the extracellular metabolomic profiles dynamics $\mathbf{X}_{\mathrm{E}}$.

The developed methodology is general and could be applied to relate other QAs or process performance indicators to metabolomic dynamics, even with the possibility of correlating more QAs and process variables at the same time. Supervised techniques, such as MPLS, identify a more direct relation between biological data and QAs and important biomarkers can be easier identified with respect to unsupervised techniques, such as MPCA and MB-PCA. Additionally, the variables selection method applied to metabolomics ensures statistical robustness, because only statistically relevant ions which demonstrates to be predictive for several different splitting of the dataset are retained.

3.4 Conclusions

In this study, process and biological time-varying (i.e., dynamic) data were integrated through data-driven techniques to accelerate cell line selection during biopharmaceutical process development. Both unsupervised and supervised multivariate models were used to improve process understanding on the relation between changes in cell physiological state and cultivation process behavior, to monitor the culture time course and to provide insights into biomarkers related to QAs which can be monitored at specific times along cell culture.

We showed how the dynamics of metabolomic profiles relate to cell culture performance, providing the identification of cell lines with relevant phenotypes in the first half of AMBR15$^{\mathrm{TM}}$ runs. We also discussed ways to understand how metabolomic profiles changes during time and how the dynamics of certain metabolites relates to the dynamics of QAs and culture variables. Furthermore, we presented how the *product titer* time trajectory can be estimated from metabolomic data, and how this relation provides insights into the cell physiological state useful for a more informed cell selection. A statistically robust variable selection methods, introduced to metabolomic data, gives insights on possible biomarkers of *product titer* which should be investigated further. However, since *product titer* is strictly related to viable cell concentration, in this study primarily focused on methodology we cannot obtain insights on biological phenomena strictly and solely related to antibody production. To glean further biological insights on antibody production, the relation between dynamic biological data and cell specific productivity will be explored in further publications.

From an industrial perspective, the proposed methodology provides a deeper understanding of the biological pathways and metabolites correlated with commercially relevant phenotypes. This methodology could be applied to smarter data driven cell line selection and commercial manufacturing platform fit to reduce onward development timelines and resources. Furthermore, the proposed methodology can be extended to different biopharmaceutical products and may benefit from the integration of other data types, such as transcriptomics and proteomics.

Chapter 4

Metabolic traits for the selection of productive cell lines through metabolomic dynamic data-driven modeling*

In this Chapter, we propose multiway and multivariate statistical techniques exploiting dynamic metabolomic data from the AMBR15^{TM} scale to assist the selection of high productive cell lines during bioprocess development and scale-up. The wealth of information contained in the metabolomic profiles dynamics allows to identify the cell lines with high productivity, already from the early stages of experimentation. Moreover, the developed models allow to identify the biomarkers that are mostly related to cell productivity and to study how the important metabolic pathways for the discrimination of cell productivity vary along the cultivation. Specifically, tricarboxylic acid (TCA) cycle related pathways demonstrate to be predominant in the early stages of the cultivation process, while amino and nucleotide sugar pathways are impactful in the late stage of the culture.

4.1 Introduction

Recently the development of monoclonal antibodies (mAbs) has gained a central role either against infectious diseases, such as SARS-CoV-2 (Taylor et al., 2021), or human immunological and oncological diseases. Monoclonal antibodies are a class of recombinant proteins which are typically produced in mammalian cell cultures, because they require enzymatic post-translational modifications, such as glycosylation, for correct activity and safety in humans (Sha et al., 2016). Nowadays, Chinese Hamster Ovary (CHO) cells are the most common expression host for mAbs, accounting for 84% of approved mAb products (Tripathi and Shrivastava, 2019).

A major challenge during the development of mammalian cell cultures is the selection of cell lines with the desired characteristics of productivity and stability because performance is strongly dependent on the cell line selected (Facco et al., 2020). Accordingly, product quality,

* Barberi, G., Benedetti, A., Diaz-Fernandez, P., Sévin, D. C., Vappiani, J., Finka, G., Bezzo, F., Facco, P. (2022). Identification of CHO platform metabolic traits for the selection of productive cell lines in biopharmaceutical process development through metabolomic dynamic data-driven modeling. *Submitted to Metabolic Engineering*

productivity and stability are the main cell line attributes considered for cell line selection, and should be optimized from the early stages of the bioprocess development (Li et al., 2010). In particular, cell specific productivity (Q_{P}) sharply varies in a population of transfected CHO cells, partially as an effect of the mAb heavy and light chain gene copies ratio and their chromosomal insertion sites (Jiang et al., 2006). For this reason, the development of new mammalian cell cultures and the selection of the best performing cell lines requires extensive experimentation and multiple stages, with great expenditure of time and resources. Biopharmaceutical companies are therefore trying to develop effective solutions for selecting the best performing cell lines in order to accelerate the development of mAbs and to reduce the drug time to market (Rameez et al., 2014).

Metabolomics, the comprehensive analysis of all the metabolites in a biological system typically performed with liquid chromatography coupled with mass spectrometry (Zhou et al., 2012), is gaining interest in mammalian cell cultures for the in-depth understanding of the physiological processes within the host platform. Data analytics has proved to be a precious tool to mine the vast amount information generated by metabolomic studies (Gorrochategui et al., 2016). Data analytics, in fact, can be effectively applied on metabolomics to deeply understand the cell physiological state along the cultivation process and to gain insight into the relationship between the cell physiological characteristics and the phenotypes searched during cell line selection. Specifically, metabolomics studies revealed through data analytics the metabolic differences between parental and transfected cell lines (Dietmair et al., 2012b), between cultivations at different cell densities (Karst et al., 2017) and cell age (Chrysanthopoulos et al., 2010), identified specific inhibitors of cell growth (Alden et al., 2020) and understanding of the physiological state of cell lines cultured at 10, 100 and 1000 L bioreactor scales (Vodopivec et al., 2019). Furthermore, data analytics were used to identify the main metabolites and metabolic characteristics of high productive cell lines (Chong et al., 2012), and provided insight on specific biomarkers of the desired glycan profile (Zürcher et al., 2020) and metabolites promoting or inhibiting product titer (Morris et al., 2020). In these studies, metabolomic measurements repeated along the cultivation are often available; however, the content of information deriving from their dynamic evolution is not explicitly exploited. To our knowledge, the work by Rubingh et al. (2009) was the only one exploiting dynamic metabolomic data to predict bacterial productivity in phenylalanine-producing *E. Coli* cultures, and to identify the changes in the prediction importance of metabolites. However, an extensive study on the impact of the cell metabolic and physiological state changes on cell phenotype for mammalian cell cultures, and on the use of dynamic metabolomic data to support cell line selection is still missing.

This work aims at bridging this gap, demonstrating how dynamic metabolomic data can be exploited through data analytics to support and accelerate the selection of high producing cell lines during bioprocess development and scale-up. The exploitation of dynamic metabolomic data allows to identify specific metabolites characterizing the desired phenotype (e.g., cell productivity) during the course of the culture process, and provides insights to engineer the host cell lines for enhancing a desired phenotype. The monitoring of metabolites characterizing cellular phenotypes can provide valuable information for a more informed selection of the best performing cell lines, and at the same time a significant reduction in time and resources required for the analytical testing. Furthermore, in this work we analyzed industrial data from the development of a therapeutic monoclonal antibody, comprising a large number of cell lines. This aspect can be advantageous for the bioprocessing field because the main scientific works on this topic lack the use of industrial data and are often based on lab-scale experiments, which comprise a limited number of experimental runs covering a limited portion of the possible cell metabolic states. Instead, the analysis of a large number of cell lines allows to have a broader overview of many possible different physiological states that cell lines can express along the cultivation process.

In this work, we specifically use (Figure 4.1) the dynamics of metabolomic profiles to discriminate between low productive and high productive cell lines through multivariate statistical techniques (Barker and Rayens, 2003; Nomikos and MacGregor, 1995), even anticipating this identification at the early culture stages. Performance and interpretability of the developed models are improved through variable selection, by excluding uninformative metabolites (Fernández Pierna et al., 2009). Then, the outcomes of the developed models are used for biological understanding through different state-of-the-art methodologies (Conesa et al., 2008; Goel et al., 2014; Goeman et al., 2004; Subramanian et al., 2005; Wiklund et al., 2008) to improve the robustness of the interpretation. In this way we identify how the dynamic changes in metabolites and cell functions (i.e., metabolic pathways) are related to cell productivity.

This Chapter is structured as follow: in Section 2 we will present the available data and briefly discuss the data analytics methods applied, and in Section 3 we will discuss the main results of the identification of high productive cell lines, how it can be anticipated to early culture stages, and the biological understanding.

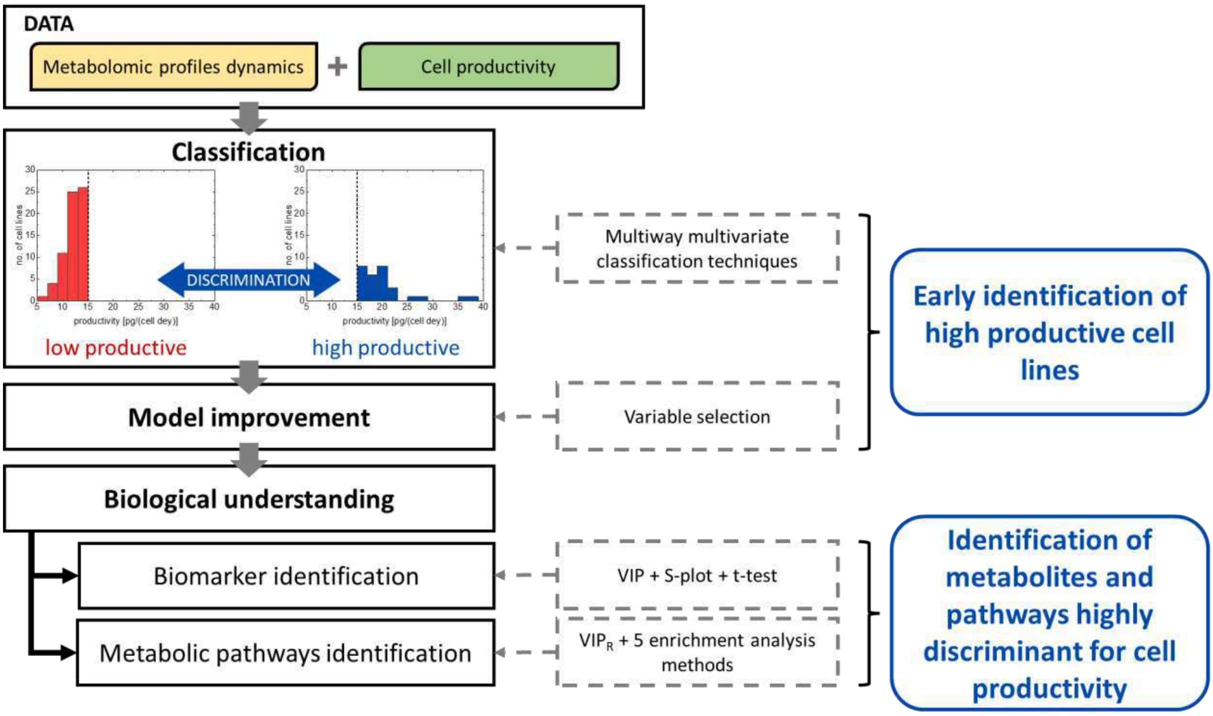

DATA
Metabolomic profiles dynamics
+
Cell productivity
Classification
no. of cell lines
productivity [pg/(cell dey)]
DISCRIMINATION
low productive
high productive
Model improvement
Biological understanding
Biomarker identification
Metabolic pathways identification
Multiway multivariate classification techniques
Variable selection
VIP + S-plot + t-test
VIP$_R$ + 5 enrichment analysis methods
Early identification of high productive cell lines
Identification of metabolites and pathways highly discriminant for cell productivity

Figure 4.1 *Detailed procedure of the methodologies utilized to accelerate cell selection and biological understanding.*

4.2 Materials and methods

In this Section the data and the mathematical methodologies used in this Chapter are briefly presented.

4.2.1 Cell culture data

A total of $N = 96$ CHO clonal cell lines were cultured for 15 days in the AMBR15$^{\text{TM}}$ miniature bioreactor system (Sartorius Stedim Biotech, Sartorius AG, Goettingen, Germany) to produce a therapeutic mAb. All production runs were performed using the GSK proprietary platform process. The process is run in fed-batch mode using glucose as main carbon source, while process conditions (i.e., bolus feeding addition, pH and temperature) were kept constant for all microbioreactors. No biological replicates were performed to allow for a larger number of cell lines to be screened in parallel.

Viable cell concentration (*VCC*) and *product titer* were measured at $T = 7$ time instants along each experimental batch (t with $t = 1, 2, \dots, T$, namely 0, 3, 6, 8, 10, 13 and 15 days): *VCC* (in cells/L) was measured with ViCell Cell Viability Analyzer (Beckman Coulter Inc., Brea, California, US); *product titer* (in mg/L) was measured with IgG Cedex Bio HT analyzer (Roche Diagnostic Corporation, Indianapolis, US).

Specific cell productivity (Q_{P}) is the target quality attribute and was calculated as the ratio between the titer at harvest $titer|_{t=7}$ and the integral of *VCC* over the entire experimental batch:

$$Q_{\mathrm{P}} = \frac{titer|_{\mathrm{t=7}}}{\int_0^{t=7} VCC\, dt} \quad . \tag{4.1}$$

Metabolomic data, available from both cell pellets and culture supernatant (i.e., intracellular and extracellular metabolites, respectively), were analyzed by flow injection liquid chromatography-mass spectrometry (LC-MS) at the same 7 time instants t (with $t = 1, 2, \dots, T$) in $R = 2$ replicates. Missing data were present in the dataset: metabolomics measurements of intracellular cytoplasm at $t = 2$ are missing because the number of cells in the culture was insufficient to perform the analytical testing. LC-MS analysis was performed in negative ionization mode under a scan range of mass over charge (*m/z*) 50-1000 (Fuhrer et al., 2011): $V_{\mathrm{I}} = 4587$ ions were collected from the cell pellets and $V_{\mathrm{E}} = 4489$ ions were collected from the culture supernatant, characterized by their *m/z* and intensity values. LC-MS analytical measurements were pre-processed following an in-house pipeline (Frederick et al., 2020; Perrin et al., 2020) consisting of peak detection, global alignment of scans, and metabolite annotation. Detected ions were tentatively annotated as metabolites based on accurate mass, with the limitation that isomers of metabolites with masses within the annotation tolerance cannot be distinguished and some ions are ambiguously annotated as multiple tentative metabolites.

Intracellular and extracellular metabolomic data were then organized into two 4-dimensional arrays $\underline{\mathbf{X}}_{\mathrm{I}}$ $[N \times V_{\mathrm{I}} \times T \times R]$ and $\underline{\mathbf{X}}_{\mathrm{E}}$ $[N \times V_{\mathrm{E}} \times T \times R]$, respectively.

Two productivity classes were also defined as a class indicator **Y**: *i*) low productive cell lines ($Q_{\mathrm{P}} \leq 15$ pg/(cell · day)), and *ii*) high productive cell lines ($Q_{\mathrm{P}} > 15$ pg/(cell · day)). The class indicator **Y** $[N \times 2]$ has the form of a dummy variable.

4.2.2 Multiway and multivariate data analysis

Prior to analysis, metabolomic data were mean centered and Pareto scaled (Eriksson et al., 2006) (i.e., each ion intensity is divided by the square root of its standard deviation). Noisy ions with a relative standard deviation > 0.25 and ions with more than 20% of missing data were excluded from the analysis. The remaining missing data were imputed with a missing data replacement technique (Troyanskaya et al., 2001), which assigns missing values as the weighted average of the K_{miss} metabolites with intensity profiles that are more similar to the metabolite of interest. In this study, K_{miss} =15 was proved to minimize the imputation error. Metabolites highlighted in this work were checked to ensure that the amount of missing data were largely below the 20% threshold to avoid any artifacts produced by data replacement technique.

Multiway partial least squares discriminant analysis (MPLS-DA) (Barker & Rayens, 2003; Nomikos & MacGregor, 1995b) was used to discriminate cell lines according to their specific productivity level. MPLS-DA is a multivariate latent-variable (LV) classification method which was adopted to discriminate the cell lines according to productivity from dynamic metabolomic data. In MPLS-DA the datasets are unfolded in such a way to deal with the multidimensional (i.e., multi-way) data structure (Nomikos & MacGregor, 1994), resulting in the matrices $\mathbf{X}_\mathrm{I}$ $[N \cdot R \times V_\mathrm{I} \cdot T]$ and $\mathbf{X}_\mathrm{E}$ $[N \cdot R \times V_\mathrm{E} \cdot T]$ for intracellular and extracellular metabolomic data, respectively. $\mathbf{X}_\mathrm{I}$ and $\mathbf{X}_\mathrm{E}$ are thereafter defined as metabolomic profiles dynamics. Then, a PLS-DA model is built on the unfolded dataset. The PLS-DA (Barker & Rayens, 2003) model reduces the $V_\mathrm{E} \cdot T$-dimensional space of the metabolomic profiles dynamics to a smaller space A orthogonal LVs, which captures in this case the dynamics of metabolites mostly related to the discrimination of cell productivity. In MPLS-DA, the model scores are used to describe the relationship between cell lines according to their metabolomic profile dynamics, while the loadings are used to describe how the dynamics of metabolites and their correlations are related to the discrimination of the cell productivity.

Evolving MPLS-DA (E-MPLS-DA) (Barker & Rayens, 2003; Ramaker et al., 2005) was used to discriminate cell lines according to their specific productivity level early during the culture. In this case, a single MPLS-DA model is built at each time instant t with $t = 1, 2, \dots, T$, considering the metabolomic profiles dynamics up to t (namely, from instant 1 to instant t). This method retains information on the entire past history of the experimental batch to accomplish the classification in each time instant in which metabolomic data are available along the culture time course.

For both modeling strategies, the dataset was randomly split into a calibration set (90 cell lines) and an external validation set (6 cell lines). The number of LVs was selected through a 8-venetian blinds cross-validation (Geladi & Kowalski, 1986b) over the calibration set, while calibration performance were evaluated through 250-iteration Monte Carlo cross-validation, which consists of randomly splitting of the calibration dataset in 79 cell lines for building the model and 11 for validation. In any splitting of the data, measurement replicates were included in the same set.

An iterative variable selection technique was used to retain only meaningful predictors (i.e., ions) and improve model performance. The most important ions for productivity discrimination were then selected through a robust and computationally intensive backward iterative uninformative variable elimination procedure (Fernández Pierna et al., 2009; Mehmood et al., 2012), where three importance metrics are used to identify the uninformative variables: *i*) Variable importance in Projection index (VIP) (S Wold et al., 1993), *ii*) selectivity ratio (SR) (Kvalheim & Karstang, 1989), and *iii*) regression coefficients. A MPLS-DA or E-MPLS-DA model was then built with the retained ions, showing improved classification performance.

Additional information on the multivariate methodologies used in this work are reported in Appendix C.

4.2.3 Important biomarker and biological function identification

The developed models were used to improve the biological understanding of selected host cell and to identify the metabolic characteristics of high productive cell lines. Specifically, a multi-step procedure was used to identify tentative metabolites to consider as productivity biomarkers and to assess how they change along the culture process, while multivariate and enrichment analysis methods were used to identify how metabolic pathways for productivity discrimination change along the experimental batch.

4.2.3.1 Biomarker identification

Tentative productivity biomarkers were identified with three methods sequentially applied to improve the robustness of the selection and to exploit the advantages of the different techniques. A bootstrap procedure (Afanador et al., 2013) was used on the model developed in Section 4.2.2 to identify only the most robust ions for productivity discrimination. Only ions whose VIP score remains high independently of the available subset of samples in a cross-validation were retained. In particular, the results of the $it_{\text{max}} = 250$ iterations Monte Carlo cross-validation (Section 2.2) were used to calculate lower VIP 90% confidence limit VIP_{LCL} as:

$$VIP_{\text{LCL}} = \overline{VIP_v} - \hat{\sigma}_{VIP_v} t_{1-\alpha/2, it_{\text{max}}-1} \quad , \tag{4.2}$$

where $\overline{VIP_v}$ and $\hat{\sigma}_{VIP_v}$ are the average value and the standard deviation of the VIP score of the ion v over the it_{max} iterations, and $t_{1-\alpha/2, it_{\text{max}}-1}$ identifies the lowest 5% confidence threshold of a t-distribution with $(it_{\text{max}} - 1)$ degrees of freedom. Additional details on this bootstrap procedure are reported in the Appendix C.5.

Among the ions with VIP_{LCL}>1, a S-plot (Wiklund et al., 2008) was used to identify the ones with large covariance and correlation with the LV1 of the E-MPLS-DA model, which captures the main differences between the two productivity classes. This method was used because the specific information on covariance and correlation cannot be retrieved directly from the VIP score. Only ions largely covariant and correlated with the cell specific productivity were retained.

Finally, a Student's t-test over the ion intensities was used to assess if statistically meaningful variations are present between the two productivity classes. Specifically, the Student's t-test was applied to each ion to prove the null hypothesis that the two classes of cell productivity have intensity coming from distribution with equal mean and variance. A Bonferroni correction for multiple hypothesis testing was used to control the family-wise error rate at a confidence level < 0.05. Only ions with statistically meaningful variations between the classes are retained. This procedure allows to identify robust biomarkers with high importance for the discrimination of the cell productivity, highly covariant and correlated with it, which have also statistically significant variations in their intensities between the two productivity classes.

4.2.3.2 Metabolic pathway identification

The metabolic pathways highly important for the discrimination of the cell productivity were identified from the outcome of the 250-iteration Monte Carlo cross-validation through an importance index named VIP_{R} which quantifies the importance of each metabolic pathway. The VIP_{R} is obtained as the ratio between the VIP_{LCL} averaged over the ions belonging to a metabolic pathway and the VIP_{LCL} averaged over all the other available ions. Accordingly, VIP_{R} marks metabolic pathways with ions showing an average importance greater than the average importance of all other ions.

Additionally, to confirm the importance of metabolic pathways the outcomes of five independent methods were combined into a single p-value through the Z-scores (Zaykin, 2011) In this way, we improve the robustness of the identified pathways and mitigate the disadvantages of each method. The methods considered in this work are:

i) metabolite set enrichment analysis (A. Subramanian et al., 2005), which is based on correlation with the desired phenotype and is particularly suitable when metabolites in a pathway are strongly cross-correlated, but loses performance with pathways containing a small number of metabolites;

ii) enrichment analysis based on hypergeometric test, which strongly relies on the method to select metabolites from the background distribution (i.e., ions selected by the multivariate models in this case) and does not explicitly consider the correlation with a desired phenotype;

iii) global test (Goeman et al., 2004), which assesses the predictive effect of the metabolites in a pathway on the desired phenotype through a random effects model without considering the possible interactions with metabolites in other pathways;

iv) multivariate projection method (Conesa et al., 2008), which assesses if the variability of metabolites in each pathway is predictive of the desired phenotype through multivariate methods. However, this relationship might be hidden in small portion of the variability of the metabolites in a pathway, that is not always captured by this method which focuses on the largest sources of variability in a pathway;

v) multivariate inference of pathway activity (Goel et al., 2014), which employs multivariate methods and five different metrics to assess the magnitude and the significance of the pathway activity. However, nothing ensures that the multivariate method identifies as primary sources of variability the ones mostly related to the desired phenotype.

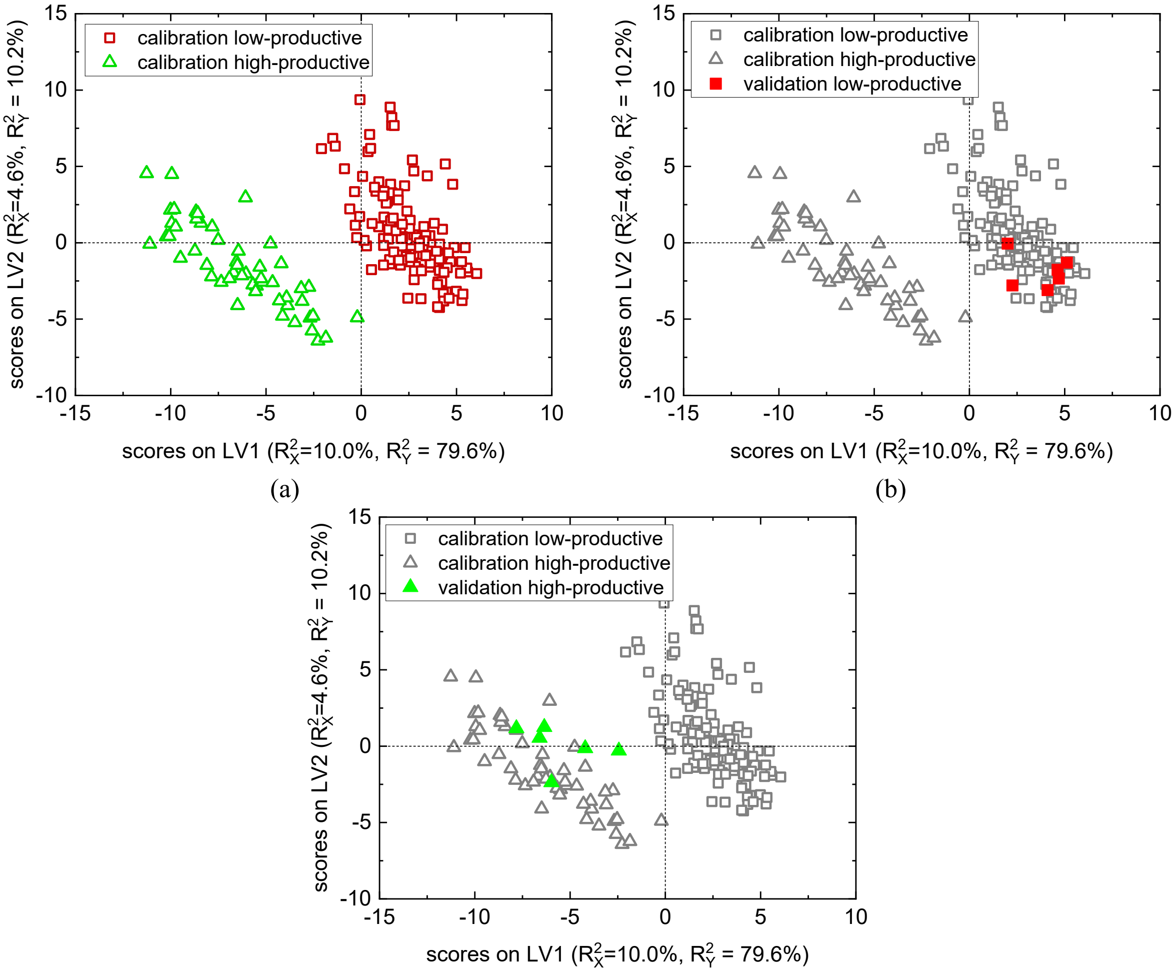
Metabolic pathways with a $VIP_{\mathrm{R}} > 1$ and a composed p-value < 0.05 were considered significantly important for productivity discrimination and highly related to it.

The CHO metabolic pathway network was visualized with Cytoscape 3.8.2 (Shannon et al., 2003).

4.3 Results and Discussion

4.3.1 Discrimination analysis to study the correlation between cell productivity and metabolomic profiles dynamics

In this Section we study the correlation between the cell productivity and the metabolomic profiles dynamics to understand the relation between the cell physiology changes along the culture and the cell line productivity. The identification of strong correlations allows to analyze how changes in specific metabolites intensity and variations in the cellular functions (i.e., metabolic pathways) characterize cell productivity. To this purpose, cell productivity is discriminated from the entire metabolomic profiles dynamics through a MPLS-DA. Furthermore, backward elimination of uninformative variables is used to discard metabolites poorly related to cell productivity and improve the classification performance of the models.



(a)

(b)

(c)

Figure 4.2 *Score space of the MPLS-DA model built on the intracellular metabolomic data for the discrimination of cell productivity: (a) calibration samples, (b) low productive external validation cell lines, and (c) high productive external validation cell lines.*

The model built on $\mathbf{X}_\mathrm{I}$ with 2 LVs reaches 99.4% overall classification accuracy in cross-validation with 14.6% of the metabolomics variability (meaning that a lot of redundancy is present in metabolomics dynamics and a large amount of the metabolomic data is not related to cell productivity). To our knowledge, this good classification performance is not due to an obvious surrogate for titer or productivity present in the metabolomic profiles, because LC-MS being performed up to *m/z*=1000 is not able to directly measure antibody concentration. Additionally, similarly good discrimination performance from intracellular metabolomic data has been reported in previous studies (William Pooi Kat Chong et al., 2012). Accordingly, a good correlation between the cell productivity and the metabolomic profiles dynamics exists, meaning that changes of metabolomic profiles during the experimental batches contain systematic information for the discrimination of cell productivity.

A better understanding of the relation among cell lines and of the amount of information related to productivity discrimination is achieved through the interpretation of the model scores. Figure 4.2a shows the score space of the model built on $\mathbf{X}_\mathrm{I}$, in which each point summarizes the behavior of the 542 ions retained from the entire metabolomic profiles dynamics (i.e., the entire duration of an experimental batch). The model explains 89.9% of the productivity class difference through 14.6% of $\mathbf{X}_\mathrm{I}$ variability: the main difference between productivity classes is explained along LV1 that retains 10.0% of $\mathbf{X}_\mathrm{I}$ variability (i.e., the content of information of approximatively 50 ions). A good separation between low and high productive cell lines is found despite productivity being a continuous variable. This is due to the variable selection that retains only the ions that are most effective at discrimination. A permutation test performed on the class labels showed a clear reduction in the model discrimination performance, thus demonstrating that the observed good separation and model performance are not due to overfitting.

Specifically, in Figure 4.2a, low productive cell lines (red squares) are located in the first and fourth quadrants (positive LV1), while high productive cell lines (green triangles) are mainly located in the third quadrant (negative LV1 and LV2, jointly). Accordingly, cell lines showing higher intensities of the ions defined by LV1 are likely to be low productive, while cell lines showing lower intensities of the ions defined by both LVs are likely to be high productive.

The external validation cell lines are used to assess the robustness of the captured correlation and the generalizability of the ions retained form the metabolic profiles dynamics by the backward elimination procedure. All external validation cell lines, which are producing the same therapeutic antibody and were randomly sampled from the available data prior the analysis, are correctly classified. Hence, the model captures a robust correlation between metabolomic profiles dynamics and cell productivity, which is generalizable to unknown cell lines producing a specific product. Figure 4.2b and 4.2c, in particular, show the projection onto the score space of the low productive external validation cell lines and the high productive ones, respectively. The model correctly maps the cell lines into specific subspace characterizing each productivity class. In fact, low productive cell lines are mapped into the fourth quadrant (Figure 4.2b), while high productive ones are correctly mapped mainly in the third quadrant (Figure 4.2c).

The model on $\mathbf{X_E}$ achieves with 3 LVs a comparable cross-validated classification accuracy (97.8%), capturing 21.0% variability of the 2005 ions retained from the entire metabolomics profiles dynamics by the backward elimination of uninformative variables, indicating that many ions of the extracellular metabolism are necessary to achieve a correct discrimination of cell productivity. This model achieves 83.3% validation accuracy in the classification of external validation cell lines; this lower classification performance in external validation indicates that the correlation between cell productivity and extracellular metabolomic profiles dynamics is weaker and not as fully generalizable as for intracellular data, because the retained ions do not discriminate cell productivity of all cell lines producing the same product. Similarly, the score space of the model built on $\mathbf{X_E}$ (Appendix C; Figure C.3) shows a reduced separation between low and high productive cell lines, despite only the most discriminant ions were retained in the model. This reduced separation further proves that extracellular ions are less correlated to differences in the productivity level, and further supports the usage of intracellular metabolomics.

The lower discrimination capability of extracellular metabolomics is probably due to the fact that it contains only information on the substances that are unused or excreted by the cells, while intracellular metabolomics provides a comprehensive view of all the substances available inside the cells, thus, resulting richer of information. Similar observations have been already reported in previous studies (Dietmair, Hodson, Quek, Timmins, Chrysanthopoulos, et al., 2012).

The abovementioned results highlight that a deep and well-established correlation between the dynamics of both extracellular and intracellular metabolomic profiles and cell productivity exists. However, the intracellular metabolomics profiles dynamics provides better and more robust information for the characterization of CHO cell productivity. Furthermore, the adopted multivariate modeling and variable selection strategies prove to be useful to correlate the cell physiological state to its phenotype, allowing a better understanding of this relation.

4.3.2 Anticipated discrimination of cell productivity

In this Section we study if it is possible to anticipate the discrimination of cell productivity at the early stages of cultures exploiting the information stored into the metabolomic profiles dynamics. To this purpose an E-MPLS-DA model is built to discriminate cell productivity by subsequently enlarging the section of the metabolomic profiles dynamics used for classification. Furthermore, model classification performance is improved retaining only highly explanatory ions through backward elimination of uninformative variables.

The performance for the anticipated discrimination of cell productivity using the retained ions from the intracellular metabolomics $\mathbf{X}_\mathrm{I}$ is reported in Table 4.1. The model captures at all the time instants a large portion of the productivity variability (≥77%), and achieves high classification accuracy (>93%) at all time instants, retaining a reduced number of ions. In particular, the model achieves very high classification accuracy (~99%) in the second half of the experimental batches, namely form $t = 4$, indicating that from that time instant the metabolomic profiles contain information for an accurate discrimination of cell productivity. The model at these time instants misclassifies only samples with a productivity which is very close to the threshold between the classes. In the first part of the experimental batches, a lower, but still high, classification accuracy is observed (~93%), indicating that the metabolomic profiles in the first days of the culture contain less, but still enough information for an accurate discrimination of cell productivity. Accordingly, in most of the cases the cell productivity is clearly characterized by cell physiological states in the first part of the experimental batches.

Table 4.1 *Performance of the E-MPLS-DA multi-model in the anticipated discrimination of cell productivity from intracellular metabolomic data: the number of LVs, the explained response variance (R_y^2), the number of retained ions and the accuracy in cross validation and external validation are reported for the model built at each time instant.*

time instant	selected number of LVs	R_y^2 [%]	number of selected ions	accuracy in cross-validation [%]	external validation accuracy [%]
1	2	77.0	110	93.4	75
3	2	76.8	292	93.6	100
4	2	85.6	150	98.7	100
5	2	88.1	239	99.1	100
6	2	89.4	307	99.3	100
7	2	89.9	542	99.4	100

External validation cell lines are then used to verify the robustness of the correlation found by the model and the generalizability of the retained ions for the discrimination of cell productivity. All validation samples are correctly classified at time instants$t \geq 3$, namely at all the time instants apart from $t = 1$ in which the classification accuracy is 75%. Accordingly, a robust correlation between intracellular metabolomics profiles and cell productivity can be found already at time instant $t = 3$, meaning that the subsets of retained ions are explanatory of cell productivity and contains its fingerprint. Furthermore, since external validation cell lines are correctly discriminated, these subsets of retained ions are generalizable to cell lines producing the same product. Differently, at time instant $t = 1$ the correlation is not completely generalizable, because the subset of retained ions does not correctly discriminate the productivity of all the external validation cell lines. This fact could be due to differences in the physiological state of some cell lines at the beginning of the culture, and that cell lines require some time to stabilize their behavior, and accordingly their metabolism, after seeding.

The model built on $\mathbf{X_E}$ achieves > 97% cross-validation accuracy at all time instants, while the external validation cell lines are classified with lower accuracy (< 84%) especially in the first half of the experimental batches in which validation accuracy is < 70% (Appendix C; Table C.1). As previously observed, the correlation between intracellular metabolomics profiles and cell productivity is not robust as for intracellular data. Furthermore, metabolomics profiles at early culture stages contain less generalizable information related to cell productivity.

In this section, we showed that models on both intra- and extra-cellular data discriminate with high accuracy the cell productivity at all time points, being the model on intracellular data more generalizable and accurate especially in the first half of the culture. Accordingly, the model on intracellular data can be integrated in the cell selection process, since it identifies the optimal cell lines to further process already in the first days the experimental batches. For this reason, the developed model can be used for risk management, because it reduces the risk of excluding potentially high performing cell lines during the selection process. Furthermore, the early identification of the high productive cell lines allows to speed up the bioprocess development process, because high performing cell lines can be early scaled-up before the end of the experiment at AMBR15$^{\text{TM}}$ scale.

4.3.3 Identification of cell productivity biomarkers

The aim of this Section is the identification of ions that, at each time instant, are highly important for the discrimination of cell productivity (i.e., productivity biomarkers). The identification of few metabolites well characterizing cell productivity allows to improve the cell selection procedure, providing more confidence for the selection of optimal cell lines and anticipating it at the early culture stages. Furthermore, few biomarkers can be measured without running the entire metabolomics analysis and easily interpreted, providing the capability of making timely decision during process development.

To this purpose, the outcomes of the E-MPLS-DA model presented in Section 4.3.2 are utilized for the identification of the most important ions for the discrimination of cell productivity through the procedure explained in Section 4.2.2. An example of the S-plot utilized for this analysis is reported in Figure 4.3. The S-plot is built at each time instant on the ions with a $VIP_{\mathrm{LCL}} > 1$. This screening excludes in general the largest part of the ions (e.g., as shown in Figure 4.3 in which the ions retained are identified by triangles). Through the S-plot, the ions highly covariant and correlated with the LV1 of the E-MPLS-DA LV are selected having an absolute covariance > 0.05 and, at the same time, an absolute correlation > 0.6 (e.g., colored triangles in Figure 4.3). A Student's t-test is then applied to these retained ions to prove the null hypothesis that the two classes of cell productivity have intensity coming from distribution with equal mean and variance. In general, all the ions identified through the S-plot as highly covariant and correlated with the cell productivity show significantly different intensities between the two productivity classes.

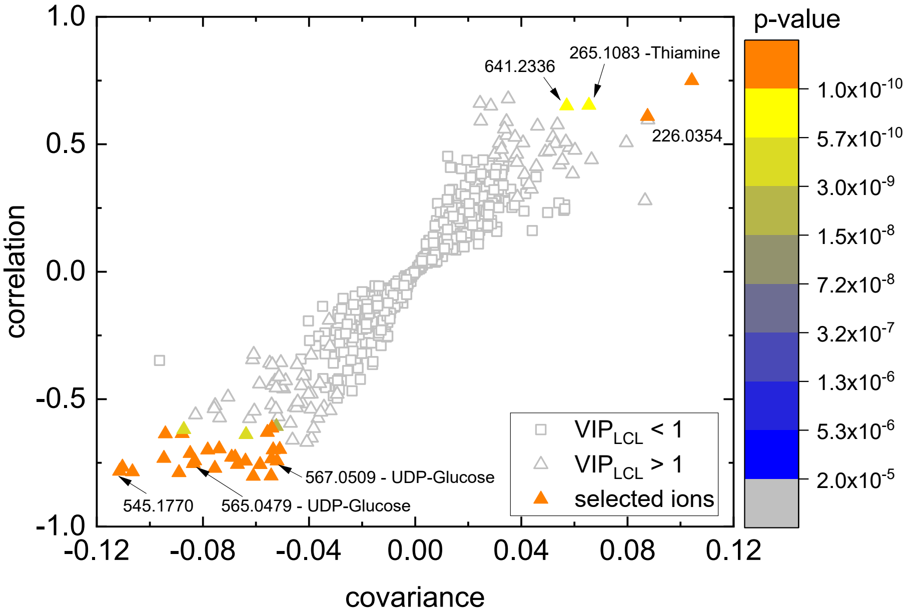


Figure 4.3 *S-plot used for the identification of ions highly covariant and correlated with cell productivity from the model at time instant* $t = 7$*. Ions with absolute covariance > 0.05 and absolute correlation > 0.6 are selected and a Student's t-test is used to calculate the reported p-value. Ions m/z value and the tentatively annotated metabolites are reported.*

Table 4.2 summarizes the results of the abovementioned analysis performed for each time instant t of the E-MPLS-DA model. For each of them, the most important ions for cell productivity are reported with the respective *m/z* value, the tentatively annotated metabolite, the ion measurement time instant, the VIP_{LCL}, the S-plot covariance, the S-plot correlation and the Student's *t*-test *p*-value.

Results show that up to $t = 4$ the ions measured in the first time instant of the experimental batches ($t = 1$) are relevant for the discrimination of cell productivity, indicating that the intracellular metabolic state at the beginning of the culture is an important indicator of cell productivity. In particular, in the model the ions *m/z* 191.0197 and 192.0231 tentatively annotated to *Citric acid* (*TCA cycle* and *alanine, aspartate and glutamate metabolism*) are identified as biomarkers of cell productivity at $t = 1$ and $t = 3$. *Citric acid* results an important indicator of productivity ($VIP_{\mathrm{LCL}} > 1.31$) and negatively correlated to cell productivity (i.e., positively correlated to LV1), indicating that high concentrations of *Citric acid* are typical of low productive cell lines. Accumulation of intracellular *Citric acid* suggests that cell lines with reduced activity of the *TCA cycle* in the early batch stages are likely to show lower productivity, probably due to reduced ATP availability for antibody synthesis. This result is in accordance with previous studies, in which *Citric acid* was highlighted as determinant for both growth and productivity in CHO cells (Dickson, 2014), and *TCA cycle* was found downregulated in low-productive CHO cells through transcriptomic analysis (Huang & Yoon, 2020b). Furthermore, since *Citrate* could be present in the bioreactor also as a media or feed component, this result proves the importance of feeding optimization for the enhancement of cell productivity. Other ions *m/z* 145.0873 and 471.2891 (possibly related to sterol lipids or fatty acids) with no tentative annotation are identified as productivity biomarkers in the first half of the experimental batches. Additional studies are required to identify the possible metabolites originating these ions.

In the second half of the experimental batches, namely from $t = 4$ to $t = 6$, a shift in the physiological state of cells is observed and different ions are identified to be highly important for the discrimination of cell productivity. Ions measured at $t = 5$ and $t = 6$ are preferentially identified as productivity biomarkers, meaning that the intracellular metabolic state in these instants is an important indicator of the cell productivity. That is probably related to the fact that at these time instants the largest production rate is usually observed, indicating that the metabolomic profiles in the production phase are extremely informative of the productivity level. Furthermore, it seems reasonable that the most informative part of the culture is found from the mid cell stationary phase.

Table 4.2 *Intracellular ions retained by the E-MPLS-DA model at different time instants showing a significant relation with cell productivity. These ions show a VIP_{LCL} > 1, high covariance and correlation with cell productivity and a significantly different intensity values for each productivity class. The time instant column refers to the time instant in which the ion was measured with LC-MS.*

m/z	Metabolite name	Measurement time point	$\mathrm{VIP_{LCL}}$	covariance	correlation	p-value
		$t = 1$				
145.0873		1	1.39	-0.112	-0.668	$3.0{\cdot}10^{-7}$
191.0197	Citric acid	1	1.37	0.121	0.672	$1.1\cdot 10^{-7}$
192.0231	Citric acid	1	1.31	0.118	0.612	$4.1\cdot 10^{-6}$
471.2891		1	1.66	-0.171	-0.704	$2.6\cdot 10^{-7}$
		$t = 3$				
145.0873		1	1.45	-0.073	-0.679	$3.0\cdot 10^{-7}$
191.0197	Citric acid	1	1.39	0.072	0.620	$1.1\cdot 10^{-7}$
471.2891		1	1.71	-0.106	-0.679	$2.6\cdot 10^{-7}$
		$t = 4$				
471.2891		1	1.53	-0.127	-0.616	$2.6\cdot 10^{-7}$
		$t = 5$				
438.9425		5	1.73	-0.102	-0.663	$1.2\cdot 10^{-8}$
628.9785		5	1.89	-0.121	-0.683	$8.5\cdot 10^{-10}$
		$t = 6$				
226.0354		6	1.65	0.102	0.612	$3.2\cdot 10^{-12}$
253.0315		6	2.11	0.164	0.638	$6.4\cdot 10^{-10}$
265.1083	Thiamine	6	1.43	0.074	0.641	$2.1\cdot 10^{-10}$
438.9425		5	1.28	-0.060	-0.605	$1.2\cdot 10^{-8}$
545.1770		6	2.16	-0.133	-0.812	$7.7\cdot 10^{-14}$
565.0479	UDP-glucose, UDP-galactose	6	1.86	-0.101	-0.788	$3.9\cdot 10^{-19}$
567.0509	UDP-glucose, UDP-galactose	6	1.47	-0.062	-0.771	$6.3\cdot 10^{-16}$
641.2334		6	1.48	0.069	0.676	$2.4\cdot 10^{-10}$
		$t = 7$				
226.0354		6	1.87	0.088	0.610	$3.2\cdot 10^{-12}$
253.0315		6	2.35	0.137	0.616	$6.4\cdot 10^{-10}$
265.1083	Thiamine	6	1.64	0.065	0.653	$2.1\cdot 10^{-10}$
545.1770		6	2.40	-0.111	-0.783	$7.7\cdot 10^{-14}$
565.0479	UDP-glucose, UDP-galactose	6	2.07	-0.084	-0.753	$3.9\cdot 10^{-19}$
567.0509	UDP-glucose, UDP-galactose	6	1.64	-0.052	-0.742	$6.3\cdot 10^{-16}$
641.2334		6	1.67	0.057	0.650	$2.4\cdot 10^{-10}$

The identified productivity biomarkers in the second half of the experimental batches are the ions *m/z* 565.0479, 566.0508, and 567.0509 tentatively annotated to *UDP-glucose* (*UDP-Glc*) or *UDP-galactose* (*UDP-Gal*) and the ion *m/z* 265.1083 tentatively annotated to *thiamine* (*thiamine metabolism*) measured at $t = 6$. *UDP-Glc/UDP-Gal* results highly relevant for the discrimination ($VIP_{\text{LCL}} > 1.4$) and positively correlated to cell productivity (i.e., being negatively correlated to LV1); this indicates that high levels of *UDP-Glc/UDP-Gal* are typically found in high productive cell lines. High levels of *UDP-Glc/UDP-Gal* have been previously observed in high productive cell lines (William Pooi Kat Chong et al., 2012). Furthermore, this result seems reasonable also because UDP sugars are key component in the glycosylation of mAbs (Kochanowski et al., 2006), and in CHO cells the nucleotide sugar donors (which UDP sugars belong to) allocated for mAbs glycosylation outweigh the ones allocated for host cell proteins glycosylation (Jimenez del Val, Polizzi, et al., 2016). However, our result points out that *UDP-Glc/UDP-Gal* is an important productivity indicator only at the end of the stationary phase (corresponding to $t = 6$), when cell lines are mainly focused on protein production, and not along the entire culture. *Thiamine* results negatively correlated to cell productivity (i.e., being correlated to LV1), indicating that high levels are typically found in low productive cell lines. Accumulation of *thiamine* in the second half of the experimental batches might be originated from a disfunction in its metabolism in the early stages of experimental batches, this indicates that a correct consumption of *thiamine* in the initial part of the culture is a key aspect for high-productive cell lines. In fact, a the addition of vitamins to CHO cell cultures has shown to improve the mAbs productivity (D. Y. Kim et al., 2005; Ritacco et al., 2018).

A similar analysis is performed over the model built on extracellular metabolomic data. No common ions with intracellular result are found. However, the ions identified as relevant for productivity discrimination are the ion *m/z* 145.0622 tentatively annotated to *L/D-glutamine* in the initial part of the experimental batches, the ions *m/z* 96.9700 and 132.0304 tentatively annotated to *phosphoric acid* and *L/D-aspartic acid*, respectively, in the final part of the experimental batches, and the ion *m/z* 611.1422 tentatively annotated to *oxidized glutathione* for almost the entire culture. The observed higher level of *L/D-glutamine* in high productive cell lines is probably related to a more efficient production of glutamine from glutamate, which results in a higher consumption of ammonia required for this reaction. Similarly, the observed higher level *L/D-aspartic acid*, which is involved in the glutamate-glutamine conversion, in high productive cell lines is probably related to similar phenomena and supports the fact that *L/D-aspartic acid* starvation results in low mAbs production (Ritacco et al., 2018). The observed lower level of *oxidized glutathione* in high productive cell lines indicates that enhanced productivity is related to a reduced accumulation of glutathione in the culture media, which has been observed as growth-limiting factor in previous studies (William P K Chong et al., 2009; Vodopivec et al., 2019).

This analysis highlights how the in-depth interpretation of multivariate models can improve the understanding of the important biological phenomena that characterize the cell phenotypes (productivity in this case). However, an approach focused on metabolic pathways rather than the single metabolites, could provide a better understanding of the cellular function mostly related to cell productivity.

4.3.4 Identification of important biological functions in the discrimination of cell productivity

The aim of this Section is the identification of the cellular functions (i.e., metabolic pathways) which are mostly related to the discrimination of cell productivity and the study on how they change along the culture. This is intended to understand when a cell function is important for the studied phenotype, providing a better understanding of the host cell. Furthermore, this analysis provides a robust basis for cell selection and insights into possible targets to improve cell performance through host engineering. To this purpose, the VIP_{R} is calculated for each pathway from the outcomes of the E-MPLS-DA, and the five additional tests (Section 4.2.3) are applied to confirm and integrate the obtained results (Section 4.2.3.2). To note, this analysis is based on ions that are tentatively annotated to metabolites (as explained in Section 4.2.1) and similarly tentatively associated to a metabolic pathway. Accordingly, further testing is required to confirm the accurate identity of the identified ions and metabolic pathways.

The heatmap of Figure 4.4 shows the evolution during culture of pathway importance in term of VIP_{R} for metabolic pathways significantly related to productivity (i.e., with a composed p-value of the five selected methodologies < 0.05). A sharp change in the cell physiological state is visible at $t = 5$, in which the metabolic pathways with high importance for the discrimination of cell productivity changes. This behavior is probably due to the cell reaching the decline phase, as shown by the VCC reaching its peak, and focusing on protein production rather than replication. In the first part of the culture, the metabolic pathways with large importance for productivity discrimination are *alanine, aspartate and glutamate metabolism* (cge00250), *TCA cycle* (cge00020), *glyoxylate and dicarboxylate metabolism* (cge00630), and *purine metabolism* (cge00230), while at $t = 5$ the important pathways shift to *amino sugar and nucleotide sugar metabolism* (cge00520), *pyrimidine metabolism* (cge002400), *oxidative phosphorylation* (cge00190), *tryptophan metabolism* (cge00380). Two metabolic pathways, such as *ascorbate and aldarate metabolism* (cge00053) and *pentose and glucuronate interconversions* (cge00040) are very important for cell productivity discrimination along the entire culture.

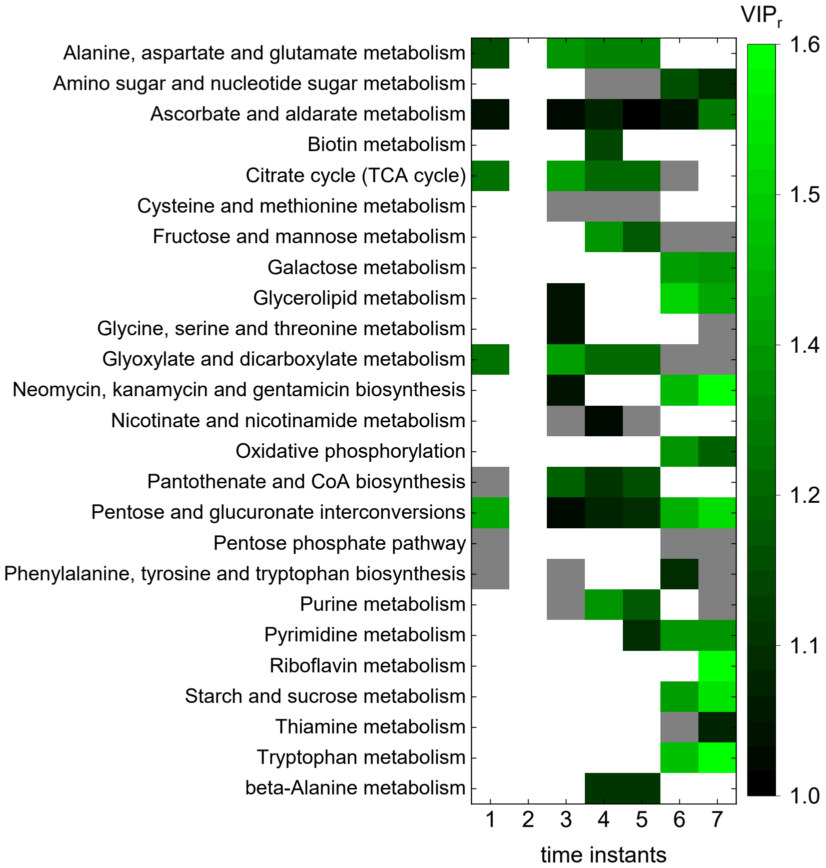
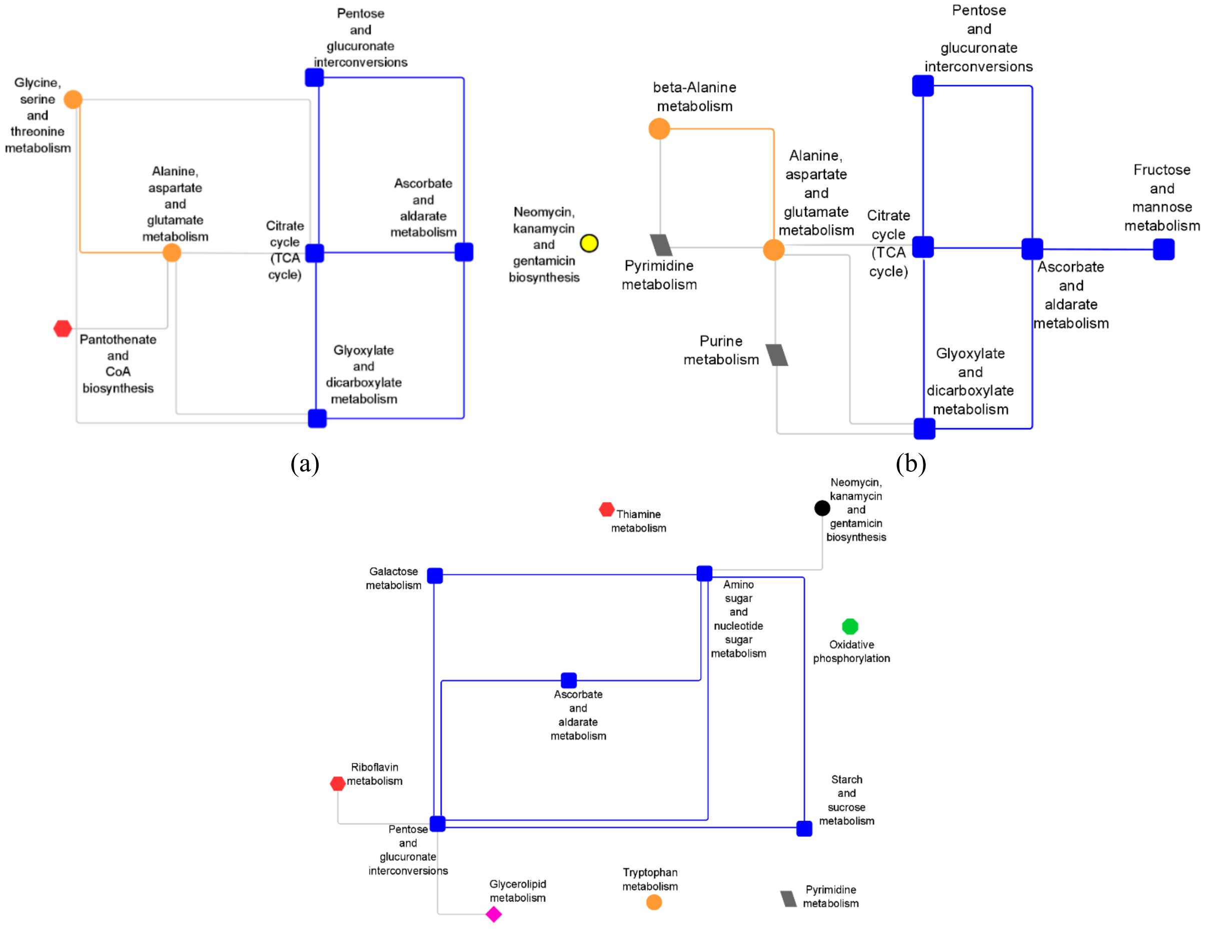


Figure 4.4 *VIP_R value during culture for pathways significantly related to cell productivity. White – pathway is not significant at that time instant; gray – pathway is significant but VIP_R < 1.*

In exponential growth and subsequent stationary phase (i.e., from $t = 1$ to $t = 5$), metabolic pathways connected to energy production and DNA replication are found to be important for cell productivity. This seems reasonable, because cells are generally consuming energy both for replication and for protein production. Similarly, the metabolism of amino acids, such as *alanine, aspartate and glutamate metabolism*, is dominant until the end of the stationary phase. In fact, high flux of *alanine, aspartate* and *glutamate* pathway (as well as other amino acids) is usually observed in the initial part of the culture until the complete depletion of those amino acids during the stationary phase (Sellick et al., 2011). Furthermore, *alanine, aspartate and glutamate metabolism* was found to be upregulated in low-productive cell lines (Huang & Yoon, 2020b), confirming our finding.

A better understanding of the inter-relationships between different cellular functions and cell productivity is achieved by inspecting the CHO metabolic network. The network of important metabolic pathways in the exponential growth phase, namely at $t = 3$ is shown in Figure 4.5a. The physiological state at $t = 3$ is very similar to the one at $t = 1$, showing that all the considerable pathways are primarily connected to *TCA cycle* and also to *alanine, aspartate and glutamate metabolism*. At the end of the stationary phase, namely at $t = 5$, a similar pattern in the cell physiological state is found (Figure 4.5b). However, in this phase nucleotide metabolisms are gaining relevance, as well as *fructose and mannose metabolism*, which were found upregulated in high-productive cell lines (Huang & Yoon, 2020b). In the decline phase (i.e., from $t = 6$) the cell physiological state is totally connected to the metabolism of nucleotide and other sugars (Figure 4.5c). Enhanced nucleotide sugar metabolism in the decline phase seems reasonable since nucleotide sugars were also found of primarily importance for protein glycosylation and productivity (William Pooi Kat Chong et al., 2012). Similarly to the result of Figure 4.5c, *amino sugar and nucleotide sugar metabolism*, *galactose metabolism*, *starch and sucrose metabolism*, and *glycerolipid metabolism* were found upregulated in high-productive cell lines, while *oxidative phosphorylation* was found downregulated (Huang & Yoon, 2020b).


Glycine, serine and threonine metabolism
Pentose and glucuronate interconversions
Alanine, aspartate and glutamate metabolism
Citrate cycle (TCA cycle)
Ascorbate and aldarate metabolism
Neomycin, kanamycin and gentamicin biosynthesis
Pantothenate and CoA biosynthesis
Glyoxylate and dicarboxylate metabolism
beta-Alanine metabolism
Pentose and glucuronate interconversions
Alanine, aspartate and glutamate metabolism
Citrate cycle (TCA cycle)
Fructose and mannose metabolism
Pyrimidine metabolism
Ascorbate and aldarate metabolism
Purine metabolism
Glyoxylate and dicarboxylate metabolism
Thiamine metabolism
Neomycin, kanamycin and gentamicin biosynthesis
Galactose metabolism
Amino sugar and nucleotide sugar metabolism
Oxidative phosphorylation
Ascorbate and aldarate metabolism
Riboflavin metabolism
Starch and sucrose metabolism
Pentose and glucuronate interconversions
Glycerolipid metabolism
Tryptophan metabolism
Pyrimidine metabolism

(a)

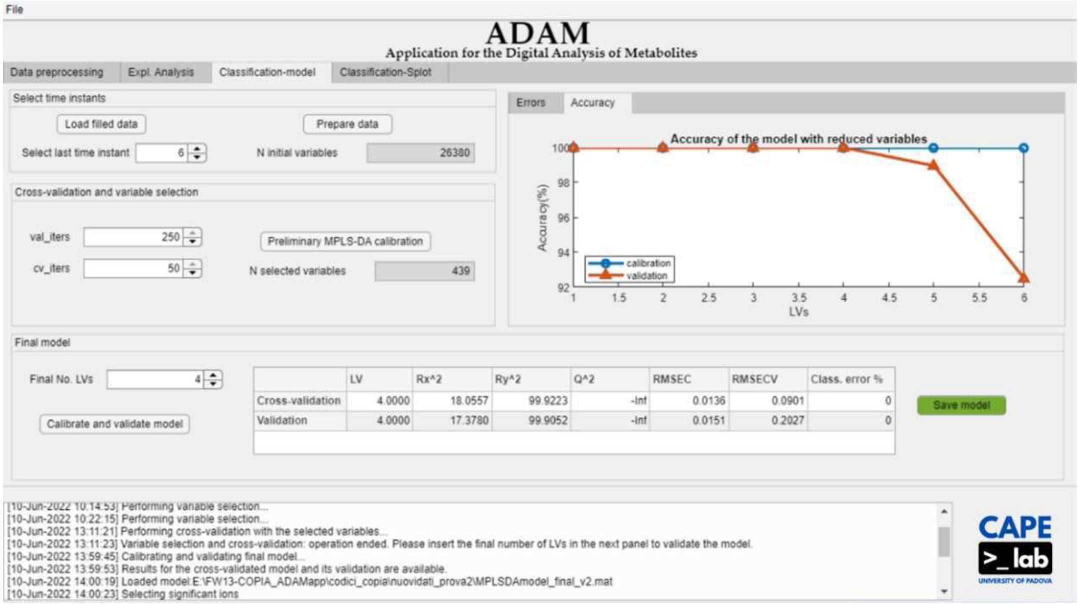
(b)

(c)

Figure 4.5 *Network of metabolic pathways significantly associated to CHO cell productivity at different time instants during culture: (a) $t = 3$, (b) $t = 5$, and (c) $t = 7$.*

Extracellular data does not show the physiological shift in the decline phase. Pathway as *arginine biosynthesis*, *glutathione metabolism*, *pantothenate and CoA biosynthesis*, and *vitamin B6 metabolism* are important for the entire duration of the experimental batches. Differently, *glyoxylate and dicarboxylate metabolism* is important in the initial part of the culture, while *beta-alanine metabolism* and *oxidative phosphorylation* are important in the final part of the experimental batches. These results highlight the importance of arginine and vitamins (Ritacco et al., 2018), and glutathione (Orellana et al., 2015; Pereira et al., 2018) for a sustained productivity.

This analysis highlights how a multivariate statistical analysis provides a better understanding on the biological functions related to CHO cell productivity. However, since this analysis is based on a priori knowledge of the metabolic network, no new hypothesis on relations between metabolites or reaction routes can be generated. Only a network approach could provide additional insights on the metabolic reactions mostly relevant for cell productivity.



Select time instants

Select last time instant 6

N initial variables 26380

Cross-validation and variable selection

val_iters 250

cv_iters 50

N selected variables 439

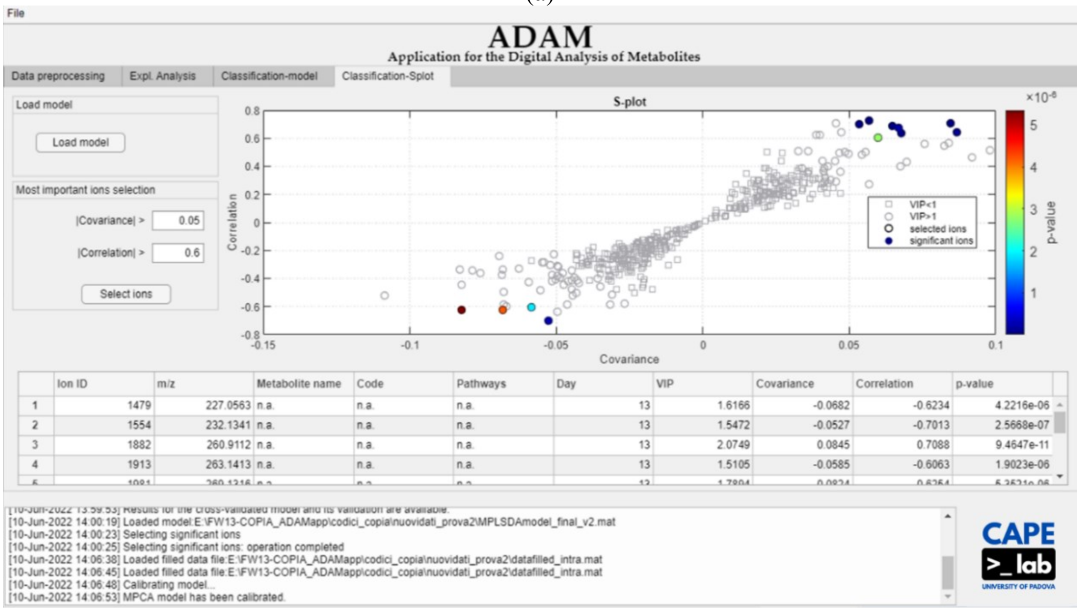
Final model

Final No. LVs 4

	LV	Rx^2	Ry^2	Q^2	RMSEC	RMSECV	Class. error %
Cross-validation	4.0000	18.0557	99.9223	-Inf	0.0136	0.0901	0
Validation	4.0000	17.3780	99.9052	-Inf	0.0151	0.2027	0

```
[10-Jun-2022 10:14:53] Performing variable selection...
[10-Jun-2022 10:22:15] Performing variable selection...
[10-Jun-2022 13:11:21] Performing cross-validation with the selected variables...
[10-Jun-2022 13:11:23] Variable selection and cross-validation: operation ended. Please insert the final number of LVs in the next panel to validate the model.
[10-Jun-2022 13:59:45] Calibrating and validating final model...
[10-Jun-2022 13:59:53] Results for the cross-validated model and its validation are available.
[10-Jun-2022 14:00:19] Loaded model:E:\FW13-COPIA_ADAMapp\codici_copia\nuovidati_prova2\MPLSDAmodel_final_v2.mat
[10-Jun-2022 14:00:23] Selecting significant ions
```

(a)



Load model

Most important ions selection

|Covariance| > 0.05

|Correlation| > 0.6

	Ion ID	m/z	Metabolite name	Code	Pathways	Day	VIP	Covariance	Correlation	p-value
1	1479	227.0563	n.a.	n.a.	n.a.	13	1.6166	-0.0682	-0.6234	4.2216e-06
2	1554	232.1341	n.a.	n.a.	n.a.	13	1.5472	-0.0527	-0.7013	2.5668e-07
3	1882	260.9112	n.a.	n.a.	n.a.	13	2.0749	0.0845	0.7088	9.4647e-11
4	1913	263.1413	n.a.	n.a.	n.a.	13	1.5105	-0.0585	-0.6063	1.9023e-06
5	1981	269.1316	n.a.	n.a.	n.a.	13	1.7804	0.0824	0.6254	6.3521e-06

```
[10-Jun-2022 13:59:53] Results for the cross-validated model and its validation are available.
[10-Jun-2022 14:00:19] Loaded model:E:\FW13-COPIA_ADAMapp\codici_copia\nuovidati_prova2\MPLSDAmodel_final_v2.mat
[10-Jun-2022 14:00:23] Selecting significant ions
[10-Jun-2022 14:00:25] Selecting significant ions: operation completed
[10-Jun-2022 14:06:38] Loaded filled data file:E:\FW13-COPIA_ADAMapp\codici_copia\nuovidati_prova2\datafilled_intra.mat
[10-Jun-2022 14:06:45] Loaded filled data file:E:\FW13-COPIA_ADAMapp\codici_copia\nuovidati_prova2\datafilled_intra.mat
[10-Jun-2022 14:06:48] Calibrating model...
[10-Jun-2022 14:06:53] MPCA model has been calibrated.
```

(b)

Figure 4.6 *View of ADAM user interface: (a) phenotype classification window, and (b) biomarker identification window.*

4.5 ADAM: Application for the Digital Analysis of Metabolites

The work presented in this Chapter resulted in the development of a software: ADAM - Application for the Digital Analysis of Metabolites, that will be used by GlaxoSmithKline (U.K.) to perform metabolomics analysis in the context of biopharmaceutical process development. ADAM reproduces the work presented in this Chapter in a simple, intuitive, and automated manner, making metabolomics analysis available to large group of practitioners. ADAM allows to:

- perform exploratory analysis of metabolomic data in order to understand how the dynamics of metabolites is associated to a desired phenotype or process behavior (as Section 3.3.1.1);
- classify samples according to their phenotype from the dynamics of metabolites (as Section 4.3.1 and 4.3.2);
- identify the biomarkers mostly associated with the studied phenotype through the three steps procedure presented in this Chapter (as Section 4.3.3).

A view of the user interface is reported in Figure 4.6.

4.4 Conclusions

In this study, dynamic metabolomics at the AMBR15^{TM} scale of a monoclonal antibody production process development was analyzed through data analytics to identify high productive cell lines to progress in the process scale-up, to examine the most important biomarkers and to find cellular functions for the characterization of cell productivity along the experimental batch evolution. Specifically, multivariate classification provided an accurate discrimination of cell productivity (98.7% calibration and 100% external validation accuracy) in the early stages of the experimental batches (approximately 8 days over 2 weeks of culture). Furthermore, the multivariate model outcome allowed to identify that the metabolic characteristics of optimal cell lines are related to *TCA cycle* and *alanine, aspartate and glutamate metabolism* until the stationary growth phase, and to the metabolism of nucleotide and other sugars in the decline phase.

These developed models identify the cell lines with the desired phenotype in the early culture stages, allowing to accelerate bioprocess development by progressing those cell lines to larger scales.

Moreover, the identification of few productivity biomarkers, which can be easily analyzed and interpreted in real-time without running an entire metabolomic study, allows to make timely decisions on process development. All the acquired knowledge could be exploited for the implementation of a more robust and confident cell selection protocol and to mitigate the risk of progressing to larger scale poorly performing cell lines. Furthermore, the identified relevant cellular functions provide insight on targets that can be manipulated though host engineering or process optimization to increase the frequency of obtaining high productive cell lines.

The result reported in this work are specific to the analyzed cell lines and product. The generalization of these results across different therapeutic antibodies will be object of further studies. However, the adopted modeling strategy is general and could be applied to any other bioprocess in which dynamic biological information is available.

As future direction of this study, a network level approach could provide insights into new routes in the metabolic network which might better characterize a desired phenotype.

Chapter 5

Data augmentation to support Biopharmaceutical process development through Digital Models*

In this Chapter, we propose the use of digital models to generate *in silico* data and augment the amount of data available from real (i.e., *in vitro*) experimental runs, accordingly. In particular, we propose two strategies for *in silico* data generation to estimate the endpoint antibody titer in mAbs manufacturing: one based on a first principles model and one on a hybrid semi-parametric model. As a proof of concept, the effect of *in silico* data generation is investigated on a simulated biopharmaceutical process for the production of mAbs. We obtained very promising results: the digital model effectively supports the identification of high-productive cell lines (i.e., high mAb titer) even when a very low number of real experimental batches (two or three) is available.

5.1 Introduction

Monoclonal antibodies (mAbs) are a class of recombinant proteins utilized against human immunological and oncological diseases, which are typically produced at the industrial level in fed-batch cultures of mammalian cells, engineered to secrete the protein of interest (Tripathi & Shrivastava, 2019). In the last few years, mAbs are gaining a lot of interest: they comprise over one-half of the biopharmaceutical approvals by regulatory agencies, and their market passed the threshold of 120 billion $ in annual sales (Walsh, 2018) expecting to reach 140 billion $ in 2024 (O. Yang et al., 2020). However, the development of new monoclonal antibodies is a time-consuming and resource-intensive procedure (F. Li et al., 2010; Tripathi & Shrivastava, 2019), which usually requires many years and large investments from biopharmaceutical companies (Epifa, 2021; Farid et al., 2020). In fact, experiments on mammalian cells may last several weeks and cost tens of thousands of dollars each. For this reason, the number of performed experimental runs is often limited. Furthermore, while scaling up the process, the number of experiments gradually decreases because the cost of a single experimental run increases with

* Botton, A., Barberi, G., Facco, P. (2022). Data augmentation to support biopharmaceutical process development through digital models—A proof of concept. *Processes*, **10**, 1796

the process volume. Hence, the number of experimental runs decreases from several dozens, if not hundreds, at the milliliter scales to 12–24 at a shake-flask scale, while a couple of runs only are typically performed at the pilot scale (F. Li et al., 2010).

Following the wave of digitalization in Industry 4.0, large amounts of data (e.g., culture variables from high throughput technologies (Rameez et al., 2014), and omics data such as transcriptomics (Clarke et al., 2011) or metabolomics (Barberi et al., 2022)) are usually collected from all the stages of the scale-up. The wealth of information contained in the experimental data can be extracted to support the mAbs development through machine learning (Barberi et al., 2022; Facco et al., 2020). In particular, different data-driven techniques demonstrated to be effective to: (*i*) understand the similarity among bioreactors at different scales and improve the similarity between scales in the scaled-down models (Ahuja et al., 2015); (*ii*) predict the mAbs concentration at harvest allowing to identify the parameters that promote or suppress production (Goldrick et al., 2017); (*iii*) estimate the mAbs quality and interpret the relationship between process and product when coupled with genetic algorithms (Sokolov et al., 2017); and (*iv*) capture very complex biological relationships through neural networks coupled with first principles models of the culture environment and accurately predict the mAbs quality attributes (Kotidis & Kontoravdi, 2020). Despite their efficacy, data-driven methods suffer when the number of available data is limited (Kjeldahl & Bro, 2010). In this case, the main driving forces and correlations in the data cannot be reliably captured due to sample underrepresentation and the large biological variability. Furthermore, the estimation performance of data-driven models degrades with few data, and models become prone to overfitting and sensitive to outliers (Tulsyan et al., 2019). For this reason, the industrial practice is to switch to univariate modeling (Tulsyan et al., 2019). Since biopharmaceutical processes are intrinsically multivariate, univariate techniques provide only a poor representation of the system under investigation and may fail to understand the complex correlation among critical process parameters (CPPs) and critical quality attributes (CQAs) (Mercier et al., 2014). For this reason, elaborating alternative strategies to overcome the limitation of a restricted amount of data from few experiments is of paramount importance to accelerate the process/product development without increasing the experimental burden.

In this respect, the generation of *in silico* data is a possible solution to the limited data problem. For example, data augmentation was successfully applied in the fields of artificial intelligence and image processing (Maharana et al., 2022; Shorten & Khoshgoftaar, 2019), and to industrial microelectronic and chemical processes (Z.-S. Chen et al., 2017; Rato et al., 2020). *In silico* data may be generated artificially either by perturbing the available data points or by combining them if no prior knowledge of the process is available. The data augmentation by means of perturbation can be performed simply by adding Gaussian noise to the available data points (Lee, 2000; Xie et al., 2020). Furthermore, the available data can be linearly combined to generate new artificial samples (Chawla et al., 2002). As an alternative, prior process knowledge can be exploited for the purpose of data augmentation and *in silico* batch generation by building a digital version of the process. For example, a hybrid mechanistic–empirical model was built to explore different settings and scenarios for a large-scale fed-batch mammalian cell culture producing a therapeutic antibody (O'Brien et al., 2021). A Gaussian process state-space model coupled with a resampling from the high-frequency acquisition system was used to generate *in silico* samples and improve the multivariate monitoring of biopharmaceutical batch processes (Tulsyan et al., 2018). Moreover, generative adversarial neural networks were used to generate *in silico* single-cell RNA sequence data for biomedical research (Marouf et al., 2020).

Despite considerable effort being made to solve the problem of limited data availability, the research and application of *in silico* model-based data generation in the biopharmaceutical industry is still an open issue. In this field, overcoming the limited availability of data in a digital manner can significantly reduce the experimental burden and development timelines, allowing for a reduction in the cost of life-saving drugs and making them available to patients earlier.

In this work, we show how, in the development of monoclonal antibodies, the application of different strategies for *in silico* batch generation can improve the identification of cell lines with the desired CQA (i.e., high mAb titer) in the scenario of limited available data. Specifically, we propose the use of two approaches based on the following digital models: a first principles model (Jimenez del Val, Fan, et al., 2016), and a hybrid semi-parametric model (Narayanan et al., 2019). The proposed methods for data augmentation will be applied to the case study of a simulated process for mammalian cell culture (Kontoravdi et al., 2010) for the purpose of improving the estimation of mAb titer.

The rest of the Chapter is organized as follows: Section 5.2 describes the general framework of the proposed procedure for *in silico* data generation, the (simulated) process, the digital models used for *in silico* batch generation, and the multivariate modeling used for the estimation of mAb titer at harvest; Section 5.3 reports the mAb titer estimation performance in a data-poor scenario and the capability of understanding the process evaluated for both *in silico* data generation strategies; and Section 5.4 contains the final remarks and future perspectives of this study.

5.2 Materials and Methods

5.2.1 Methodological procedure

The methodological procedure for the *in silico* data augmentation with digital models (Figure 5.1) goes through the following steps:

- Step 1 – Experimental campaign on the mAbs production process: batch data are obtained from experiments performed on the development scale of the process under study according to the availability of resources. In this work, we consider a simulated process for the production of mAbs at the shake-flask scale (Section 5.2.2);
- Step 2 – *In silico* batch generation from a digital model: data on real batches are utilized through digital models of the process to drive the generation of *in silico* batches with a wider variety of behaviors. In particular, two alternative modeling strategies are adopted in this work: a first principles digital model (Section 5.2.3) and a hybrid digital model (Section 5.2.4);
- Step 3 – Multivariate data-based modeling: all the available data (both the ones from the process and the ones generated *in silico*) are fed to a data-based model to support the process development and scale-up. In this work, process and *in silico* generated batches are regressed to estimate a CQA (i.e., mAb titer at harvest) through multivariate latent variable modeling (Section 5.2.5). In this way, the multivariate models exploit the data of a few process batches and the additional process knowledge extracted from the *in silico* generated batches, to make estimations of cell behavior for new samples from the culture variable time trajectories. Such estimations, especially in the presence of biological variability in the batches, are not feasible with the digital models of the process, which can only estimate the culture variable trajectories when the inputs (i.e., process initial conditions, feed composition, and scheduling) are manipulated given the biological characteristics already hardcoded in the digital model parameters.

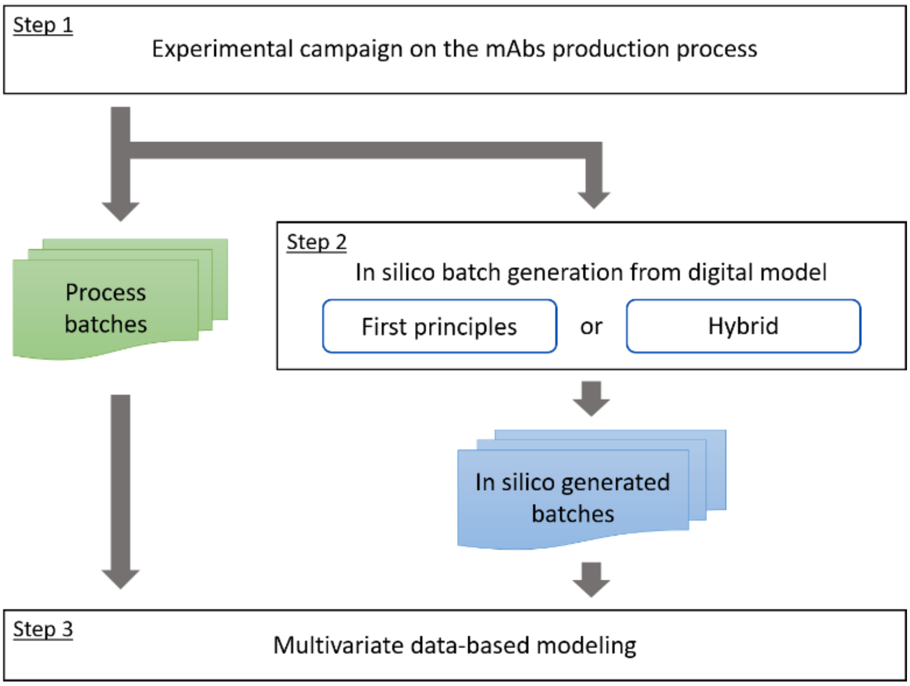

Step 1
Experimental campaign on the mAbs production process
Process batches
Step 2
In silico batch generation from digital model
First principles
or
Hybrid
In silico generated batches
Step 3
Multivariate data-based modeling

Figure 5.1 *Methodological procedure for in silico data augmentation from digital models.*

5.2.2 Process for the production of monoclonal antibodies

We consider a simulated cell culture process for the production of mAbs in fed-batch mode at a shake-flask scale. The process is based on the well-established human embryonic kidney (HEK) cell first principles model (Kontoravdi et al., 2010). The available culture variables are: viable cell concentration (VCC); nutrients (i.e., glucose and glutamine); by-products concentrations (i.e., lactate and ammonia); and mAbs titer (i.e., antibody concentration).

The variability among batches lies in the different cultured cell lines, which are simulated to display different specific productivity, $Q_{\mathrm{P}} = c_{\mathrm{mAb}}(T) / \int_0^{t=T} X_{\mathrm{v}} \mathrm{d}t$, where $c_{\mathrm{mAb}}(T)$ is the mAbs titer at harvest and X_{v} is the viable cell concentration along the batch (cell/L). For this purpose, the HEK model parameters are sampled from normal distributions with mean and standard deviation reported in Appendix D.1 Table D.1. These values are adjusted from the reference parameters found in the HEK model reference (Kontoravdi et al., 2010) in such a way as to obtain a variability of the batch time trajectories that mimic the dynamic behavior of real experimental batches at that scale. Furthermore, measurement error is simulated by adding ~6% white noise to the culture variables' profiles, accounting for the typical measurement uncertainty of analytical equipment.

An experimental campaign is carried out in 0.2 L cultures with an inoculation seed density of $2 \cdot 10^8$ cell/L. The initial media composition is set to 25.1 mM of glucose and 5.1 mM of glutamine. Feeding is performed every 20 h starting from 10 h after cell seeding by feeding 0.00875 L in 10 min. The feed composition is set to 50 mM of glucose and 10 mM of glutamine. The measurement sampling is performed prior to the feeding through a 0.0015 L withdrawal from the culture in 10 min, resulting in 10 measurement sampling points during the batch.

All the considered experimental batches satisfy the following conditions: (*i*) the final mAbs titer is below 5000 mg/L; (*ii*) the peak of VCC is reached after 50 h; and (*iii*) the specific productivity is in the range 0–20 pg/(cell·day).

The available data are concerned with 100 batches, which are organized in: matrix $\underline{\mathbf{X}}_{\mathrm{PC}}$ = [100 batches×5 variables×10 time points] that contains the time profiles of all the culture variables; and vector $\mathbf{y}_{\mathrm{PC}} = [100 \times 1]$ that contains the mAbs titer at harvest (time point 10). These data are used in different ways to calibrate digital and multivariate models. Similarly, 10 validation batches are available and organized in the matrix $\underline{\mathbf{X}}_{\mathrm{PV}} = [10 \times 5 \times 10]$ for process data and vector $\mathbf{y}_{\mathrm{PV}} = [10 \times 1]$ for mAbs titer. These validation batches are used to test the estimation performance of the multivariate models.

In this study, a simulated process is selected, not only because it reduces the time and cost of the experimental campaign, but also because it allows a full knowledge of the relationship between CPPs and CQAs, and better control of both the process behavior and the biological diversity in the experiments.

5.2.3 Modeling strategy 1: First Principles Digital Model

The first principles digital model (FPDM) is a modified version of the simplified mathematical model proposed by Jimenez del Val et al. (2016) describing a fed-batch mAbs production process (Jimenez del Val, Fan, et al., 2016). The culture variables described by the FPDM are VCC, glucose, lactate, and mAbs titer. The FPDM is modified with respect to the original model to better resemble the process. In fact, in the original model (Jimenez del Val, Fan, et al., 2016) cells grow until glucose is available in the culture and this causes a substantial difference between the behavior of the model and the process. This makes the original model unusable for the generation of batches that conform to the ones of the process. Accordingly, we added a simplified material balance for glutamine, and introduced growth limitation at low glutamine concentration and glucose consumption limitation at reduced cell growth.

The simplified material balance for glutamine is defined as:

$$\frac{\mathrm{d}(V_\mathrm{c} c_\mathrm{gln})}{\mathrm{d}t} = -\left(\frac{\mu}{Y_{x,\mathrm{gln}}}\right) X_\mathrm{v} V_\mathrm{c} \quad , \tag{5.1}$$

where V_c is the liquid volume in the culture (L), c_gln is the glutamine concentration (mM), μ is the specific growth rate (h^{-1}), and $Y_{x,\mathrm{gln}}$ is the cell yield on glutamine (cell/mmol).

In order to account for the effect of glutamine on cell growth, a limiting factor f_lim is added to the specific growth rate expression:

$$\mu = \mu_\mathrm{max}\left(\frac{c_\mathrm{glc}}{K_\mathrm{glc}+c_\mathrm{glc}}\right) - \frac{X_\mathrm{v}}{\alpha_\mathrm{x}} f_\mathrm{lim} \quad , \tag{5.2}$$

where μ_max is the maximum specific growth rate (h^{-1}), K_glc is the Monod constant for the growth on glucose (mM), α_x is the cellular carrying capacity (cell/mmol), and c_glc is the glucose concentration (mM). The limiting factor f_lim is defined as:

$$f_\mathrm{lim} = \frac{c_\mathrm{gln}}{c_\mathrm{gln}+K_\mathrm{gln}} \quad , \tag{5.3}$$

where K_gln is the Monod constant for glutamine (mM). The limiting factor f_lim decreases with the glutamine concentration, reducing cell growth when the glutamine decreases.

To limit the glucose consumption with reduced cell growth, the glucose material balance is modified as:

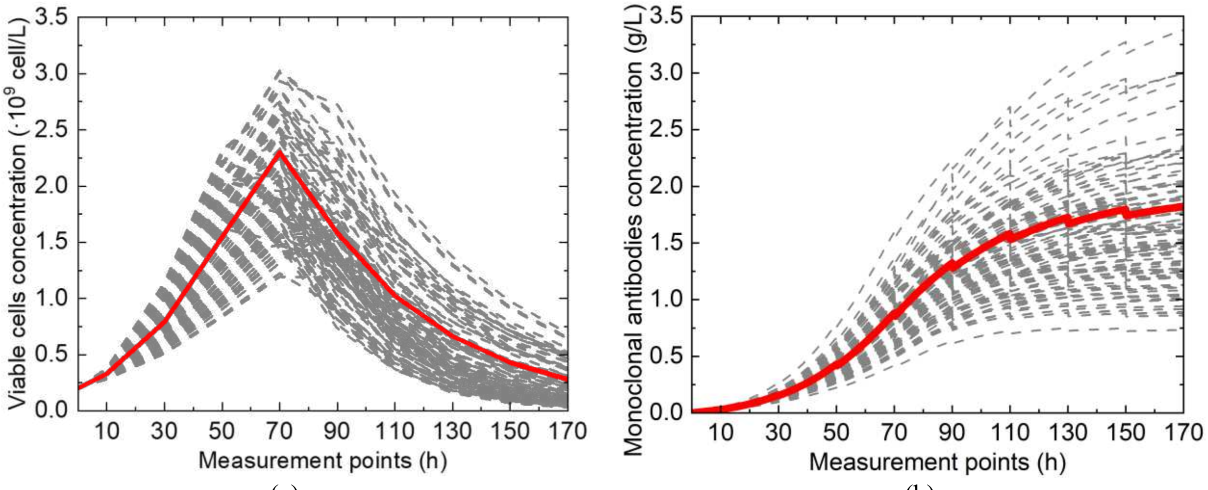
$$\frac{\mathrm{d}(V_\mathrm{c} c_\mathrm{glc})}{\mathrm{d}t} = F_\mathrm{in} c_\mathrm{glc,in} - F_\mathrm{out} c_\mathrm{glc} - Q_\mathrm{glc} X_\mathrm{v} V_\mathrm{c}\left(f_\mathrm{lim} + m_\mathrm{glc}\right) \quad , \tag{5.4}$$

where $c_\mathrm{glc,in}$ is the glucose concentration in the feeding stream, F_in and F_out are the inlet and outlet flow rates of the bioreactor (L/h), respectively, Q_glc is the specific glucose consumption rate (mmol/(cell·h)) and m_glc (-) is the glucose maintenance constant. The complete FPDM used for the generation of *in silico* batches is reported in Appendix D.2.

5.2.3.1 *In silico* batch generation through First Principles Digital Model

The FPDM is used for *in silico* batch generation. The reference parameters for FPDM are estimated from the reference process batch (i.e., obtained using the reference process parameters from Kontoravdi et al. 2010). *In silico* batches are generated by sampling the parameter values from a normal distribution with mean and standard deviation reported in Appendix D.2 Table D.2. These distributional parameters are heuristically determined to generate batches with a variability slightly larger than the one observed in the process batches. An example of *in silico* generated batches is reported in Figure 5.2: Figure 5.2a shows the time profile along the entire batch duration for viable cells concentration, and Figure 5.2b shows the time profile along the entire batch duration for mAbs titer.

This strategy is used to generate 100 *in silico* batches. The generated variables profiles are subsampled in the same 10 time points in which the process measurements are available. The resulting data are organized in matrix $\underline{\mathbf{X}}_{\text{FPDM}} = [100 \times 4 \times 10]$, which contains the time profiles of the culture variables, and vector $\mathbf{y}_{\text{FPDM}} = [100 \times 1]$, which contains the mAbs titer at harvest.



(a) (b)

Figure 5.2 *Example of batches generated in silico through the FPDM: (a) VCC profiles and (b) mAbs titer profiles for 100 batches. The thick red continuous lines represent the reference batch estimated from the process while the grey dashed lines represent the simulated ones.*

5.2.4 Modeling strategy 2: Hybrid Digital Model

The hybrid digital model (HDM) is a hybrid semi-parametric model (Narayanan et al., 2019; Oliveira, 2004; Teixeira et al., 2005; von Stosch et al., 2014), whose considered culture variables are VCC, glucose, glutamine, lactate, ammonia, and mAbs titer.

The HDM has the serial structure (Sansana et al., 2021; S. Yang et al., 2020) reported in Figure 5.3, with a mechanistic section describing the material balances of the chemical species and an artificial neural networks (ANN; Rosenblatt, 1958) to estimate the complex and unknown kinetic expressions from cell culture experimental data.

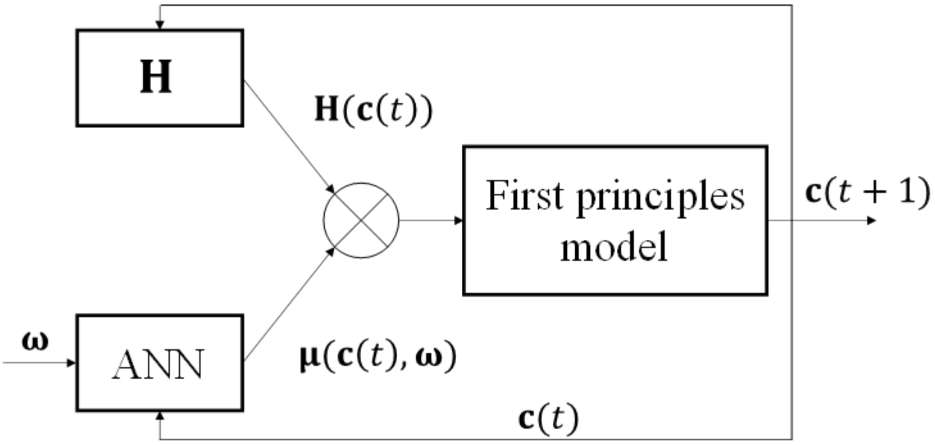

H
$\mathbf{H}(\mathbf{c}(t))$
First principles model
$\mathbf{c}(t+1)$
$\boldsymbol{\omega}$
ANN
$\boldsymbol{\mu}(\mathbf{c}(t),\boldsymbol{\omega})$
$\mathbf{c}(t)$

Figure 5.3 *Structure of the hybrid digital model to generate in silico batches.*

The HDM comprises the material balances for the culture variables of interest $\mathbf{c}\,[V \times 1] = [X_\text{v} \quad c_\text{glc} \quad c_\text{gln} \quad c_\text{lac} \quad c_\text{amm} \quad c_\text{mAb}]$ as:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} X_\text{v} \\ c_\text{glc} \\ c_\text{gln} \\ c_\text{lac} \\ c_\text{amm} \\ c_\text{mAb} \end{bmatrix} = \boldsymbol{\mu}_\text{max} \begin{bmatrix} X_\text{v} & 0 & 0 & 0 & 0 & 0 \\ 0 & -X_\text{v} & 0 & 0 & 0 & 0 \\ 0 & 0 & -X_\text{v} & 0 & 0 & 0 \\ 0 & 0 & 0 & X_\text{v} & 0 & 0 \\ 0 & 0 & 0 & 0 & X_\text{v} & 0 \\ 0 & 0 & 0 & 0 & 0 & X_\text{v} \end{bmatrix} \begin{bmatrix} \mu_{X_\text{v}} \\ \mu_\text{glc} \\ \mu_\text{gln} \\ \mu_\text{lac} \\ \mu_\text{amm} \\ \mu_\text{mAb} \end{bmatrix} = \boldsymbol{\mu}_\text{max}\mathbf{H}(\mathbf{c})\boldsymbol{\mu}(\mathbf{c}^*,\boldsymbol{\omega}) \quad , (5.5)$$

where $\boldsymbol{\mu}_\text{max}$ (Teixeira, Alves, et al., 2007) is the vector of the maximum specific rates of production/consumption for each culture variable (reported in Appendix D.3 Table D.3), $\boldsymbol{\mu}(\mathbf{c}^*,\boldsymbol{\omega}) = [\mu_{X_\text{v}} \quad \mu_\text{glc} \quad \mu_\text{gln} \quad \mu_\text{lac} \quad \mu_\text{amm} \quad \mu_\text{mAb}]$ is the vector of the specific production/consumption rates estimated by the ANN, $\mathbf{H}(\mathbf{c}) = [V \times V]$ contains the known kinetic expressions, and $\mathbf{c}^* = [X_\text{v} \quad c_\text{glc} \quad c_\text{gln} \quad c_\text{lac} \quad c_\text{amm}]$ is the reduced concentration vector used as input for the ANN.

The matrix $\mathbf{H}(\mathbf{c})$ contains all the known mechanistic information for the calculation of the reaction rates in Equation (5.5). In this work, the known mechanistic part of the reaction rates, $\mathbf{H}(\mathbf{c})$, has no fitted parameters. $\mathbf{H}(\mathbf{c})$ accounts for the dependence of the reaction rates on the cell concentration X_v and the apparent stoichiometric coefficient, which indicates if a metabolite is produced or consumed. The maximum specific rates of production/consumption $\boldsymbol{\mu}_\text{max}$ are constant parameters heuristically set in preliminary studies to appropriately scale the ANN outputs in the desired experimental ranges.

The vector of specific production/consumption rates is modeled through an artificial neural network. In particular, a two-layer ANN is used to estimate the specific production and consumption rates from the reduced concentration vector $\mathbf{c}^*$. The selected ANN has a 10-neurons hidden layer with a hyperbolic-tangent activation function and a linear output layer with 6 neurons (i.e., given by the dimension of $\boldsymbol{\mu}$). The mathematical expression of the ANN is:

$$\boldsymbol{\mu}(\mathbf{c}^{*}, \boldsymbol{\omega}) = \boldsymbol{\omega}^{(2)} \tanh\left(\boldsymbol{\omega}^{(1)} \mathbf{c}^{*} + \boldsymbol{\omega}_0^{(1)}\right) + \boldsymbol{\omega}_0^{(2)} \quad , \tag{5.6}$$

where $\boldsymbol{\omega}$ is the weight vector, $\boldsymbol{\omega}_0$ is the bias vector and the superscript (1) and (2) refer to the hidden and output layer, respectively. In this case, it is assumed that the specific production/consumption rates do not depend on the mAbs titer while depending on the number of cells, nutrients, and by-products in the culture.

The hybrid model identification is performed through the sensitivity method (Oliveira, 2004; Sansana et al., 2021), by backpropagating the errors in the concentration space through the model. In this work, the normalized sum of squared errors (SSE) between the measured concentrations c_v and the ones calculated by the HDM, $\hat{c}_v$, is directly minimized as:

$$\operatorname{argmin}(SSE) = \operatorname{argmin}\left(\sum_{t=1}^{T} \sum_{v=1}^{V} \frac{(\hat{c}_v(t) - c_v(t))^2}{\sigma_v} + \lambda_{\text{reg}} \|\boldsymbol{\mu}\|\right) \quad , \tag{5.7}$$

where σ_v is the standard deviation of the v-th process variable calculated over the training data, $c_v(t)$ is the measured concentration of the v-th culture variable at the t-th time instant, $\hat{c}_v(t)$ is the calculated concentration of the v-th culture variable at the t-th time instant, and $\lambda_{\text{reg}} = 0.05$ is a regularization term (A. Yang et al., 2011), which is added to aid training convergence. In this work, the error backpropagation is performed by calculating the gradient of the concentration errors (i.e., SSE) with respect to the ANN weights, because the hybrid model does not contain any mechanistic parameter to be fitted and the only adjustable parameters are the ANN weights. The gradient of the concentration errors (i.e., SSE) with respect to the ANN weights is calculated as:

$$\frac{\partial SSE}{\partial \boldsymbol{\omega}} = \sum_{t=1}^{T} \sum_{v=1}^{V} (\hat{c}_v(t) - c_v(t))^2 \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\omega}}\right)_t \quad , \tag{5.8}$$

where $(\partial \mathbf{c}/\partial \boldsymbol{\omega})_t$ is the gradient of the concentrations with respect to the ANN weights, calculated with the sensitivity method (Oliveira, 2004).

An Adam optimizer (Kingma & Ba, 2015) is then used to adjust the ANN parameters according to the calculated gradient, because it is nowadays one of the most effective algorithm to train ANNs. The hybrid model is trained for 400 iterations with a learning rate $\eta = 10^{-3}$, and subsequent 300 iterations with a learning rate $\eta = 10^{-4}$. Prior to the training, the ANN weights are initialized by sampling from a normal distribution $N(0, \sigma)$ where $\sigma = 0.01$.

The integration of the HDM is performed stepwise between the feeding time points. A bolus feeding of glucose and glutamine (consistent with the training batches) is simulated by updating the initial concentration after the feeding according to (Narayanan et al., 2019):

$$c_j(t^{+}) = c_j(t^{-}) + \Delta c_j(t) \quad , \tag{5.9}$$

where $c_j(t^+)$ is the concentration of nutrient j after the feeding, $c_j(t^-)$ is the concentration of the nutrient j before the feeding and $\Delta c_j(t)$ is the change in concentration of the nutrient j (i.e., glc or gln) due to the feeding at time instant t.

5.2.4.1 *In silico* batch generation through Hybrid Digital Model

The HDM is used for *in silico* batch generation as an alternative to FPDM. First, the HDM is trained on 10 batches (Narayanan, Luna, et al., 2021), which are selected from $\underline{\mathbf{X}}_{\mathrm{PC}}$ to cover a sufficient range of process variability. Then, *in silico* batches are generated by changing the maximum specific rate of production/consumption $\boldsymbol{\mu}_{\max}$, which is kept constant during training.

The values of $\boldsymbol{\mu}_{\max}$ are sampled from a normal distribution with mean and standard deviation reported in Appendix D.3 Table D.3. These values are heuristically selected, based on preliminary tests, to cover a sufficiently large variability around the process batch profiles, while preserving similarity with them. An example of the batches generated by the HDM is shown in Figure 5.4.

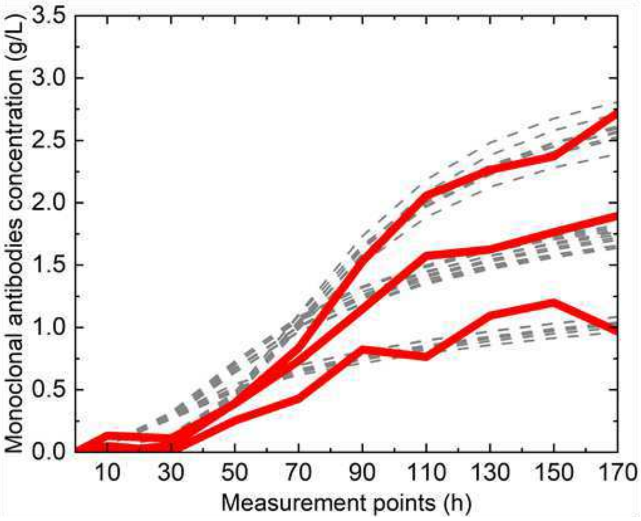


Figure 5.4 *Example of batches generated through HDM: mAbs titer profiles. In this example, 10 batches are generated from 3 training batches taken from the process. The thick red continuous lines represent the training batches while the grey dashed lines represent the simulated ones.*

This strategy is used to generate 100 *in silico* batches, 10 from each batch used to train the HDM. The generated variables profiles are subsampled in the same 10 time points in which the process measurements are available. The resulting data are organized in $\underline{\mathbf{X}}_{\mathrm{HDM}} = [100 \times 5 \times 10]$, which contains the time profiles of the culture variables, and vector $\mathbf{y}_{\mathrm{HDM}} = [100 \times 1]$, which contains the mAbs titer at harvest.

5.2.5 Multivariate predictive modeling

In this study, multi-way partial least squares regression (MPLS; Nomikos and MacGregor, 1995) is used to estimate the CQA, namely, the mAbs titer at harvest, from the multi-dimensional datasets of the correlated culture variables (both real and generated *in silico*) time trajectories.

MPLS consists of a proper unfolding of the data followed by standard PLS modeling.

Batch-wise unfolding is performed in this study to capture the correlation between the culture variables' time profiles and the response together with the cross-correlation between culture variables at different time points. In batch-wise unfolding, the two-dimensional slices at each time point $t = 1, 2., \dots, T$ of the matrix $\underline{\mathbf{X}} = [N \times V \times T]$, $\mathbf{X}^t = [N \times V]$, where N is the number of batches and V is the number of variables, are horizontally concatenated, resulting in two-dimensional matrix $\mathbf{X} = [N \times V \cdot T] = [\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^t, \dots, \mathbf{X}^T]$. Accordingly, the matrices $\underline{\mathbf{X}}_{\text{PC}}$, $\underline{\mathbf{X}}_{\text{PV}}$, $\underline{\mathbf{X}}_{\text{FPDM}}$, and $\underline{\mathbf{X}}_{\text{HDM}}$ are unfolded in the bidimensional matrices: $\mathbf{X}_{\text{PC}} = [100 \times 5 \cdot 10]$, $\mathbf{X}_{\text{PV}} = [10 \times 5 \cdot 10]$, $\mathbf{X}_{\text{FPDM}} = [100 \times 4 \cdot 10]$, and $\mathbf{X}_{\text{HDM}} = [100 \times 5 \cdot 10]$.

Partial least squares regression (PLS) (Svante Wold et al., 2001) is a multivariate statistical linear regression technique that identifies the directions of maximum covariance between a regressor matrix $\mathbf{X} = [N \times V \cdot T]$ and a response matrix $\mathbf{Y} = [N \times M]$, where M is the number of response variables. PLS decomposes both the regressor and response matrices into a common latent space of orthogonal latent variables (LVs). In this study, **X** and **Y** are auto-scaled to zero mean and unit variance (i.e., by subtracting to each column its mean value and dividing each column by its standard deviation). PLS decomposes the auto-scaled matrices **X** and **Y** as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad , \tag{5.10}$$

and

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \quad , \tag{5.11}$$

with

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} \quad , \tag{5.12}$$

where $\mathbf{T} = [N \times A]$ is the scores matrix that captures the relationships among batches according to the features of the covariance between **X** and **Y**; $\mathbf{P} = [V \cdot T \times A]$ and $\mathbf{Q} = [M \times A]$ are the loadings matrices which capture the relationships among the variables' dynamics in **X** and variables in **Y**, respectively; $\mathbf{E} = [N \times V \cdot T]$ and $\mathbf{F} = [N \times M]$ are the residual matrices for **X** and **Y**, respectively, which contain the information that is not described by the model; $\mathbf{W}^{*}$ is the weights matrix, which directs the scores to be the most predictive for the response **Y**; A is the number of selected LVs; and the superscript T represents the transpose operation. In this work, the selected number of latent variables is $A = 2$ which minimizes the estimation error of the responses in cross-validation (Valle et al., 1999).

PLS is used to estimate the response variable $\hat{\mathbf{Y}}$ for a set of O new batches, whose predictors $\mathbf{X}_{\text{new}} = [O \times V \cdot T]$ are known, from:

$$\hat{\mathbf{Y}} = \mathbf{X}_{\text{new}}\mathbf{W}^{*} \quad {}^{\mathrm{T}} \quad , \tag{5.13}$$

To improve PLS estimations, variable selection (Barberi et al., 2022; Mehmood et al., 2012) is used, in such a way as to identify and retain in the model only the variables with the largest information content on the mAbs titer and exclude the other variables. Variable importance is assessed through the variable importance in projection (VIP; Eriksson et al., 2006) index:

$$VIP_v = \frac{\sqrt{V \cdot T \cdot \sum_{a=1}^{A} R_{\mathrm{Y},a}^2 w_{va}^2}}{\sqrt{\sum_{a=1}^{A} R_{\mathrm{Y},a}^2}} \quad , \tag{5.14}$$

where $R_{\mathrm{Y},a}^2$ is the $\mathbf{Y}$ variance captured by the a-th latent variable and w_{va} is the weight corresponding to the a-th LV and v-th $\mathbf{X}$ variable. In this work, the selection of variables with $VIP > 1$ is performed over a 100-iteration Monte Carlo cross-validation; only variables with high selection frequency (i.e., 80% of the iterations with $VIP > 1$) are considered informative for the estimation and used to recalibrate the model.

The mAbs titer estimation performances are evaluated through the mean absolute prediction error (MAPE):

$$MAPE_m = \frac{\sum_{o=1}^{O} |y_{m,o} - \hat{y}_{m,o}|}{O} \quad , \tag{5.15}$$

where $\hat{y}_{m,o}$ is the estimation of the m-th response variable for the o-th batch and $y_{m,o}$ is the measured value.

When process data only are utilized for the titer estimation, the model calibration matrices $\mathbf{X}$ and $\mathbf{Y}$ are obtained from $\mathbf{X}_{\text{PC}}$ and $\mathbf{y}_{\text{PC}}$. As an alternative, when few data are available from process and *in silico* generated batches are used in PLS modeling to augment the calibration dataset, data from the digital model $\mathbf{X}_{\text{FPDM}}$ and $\mathbf{y}_{\text{FPDM}}$ (or $\mathbf{X}_{\text{HDM}}$ and $\mathbf{y}_{\text{HDM}}$) are vertically concatenated to the available process data in $\mathbf{X}_{\text{PC}}$ and $\mathbf{y}_{\text{PC}}$ to create augmented matrices $\mathbf{X}$ and $\mathbf{Y}$. Hence, the number of batches that is used for the model calibration is much larger than the number of batches available from the process. Autoscaling is applied as a data normalization preprocessing directly on the augmented matrices $\mathbf{X}$ and $\mathbf{Y}$. Note that in this study the process and the digital models have very similar statistical characteristics and separate preprocessing did not improve model performance. However, if *in silico* generated data showed different statistical characteristics with respect to process data, separate and specific preprocessing would be required.

5.3 Results and Discussion

The results are organized as follows. First of all, the mAbs titer at harvest estimation performance is presented when only the process batches are available and then compared to the performance when the *in silico* generated batches are present. Furthermore, the ability to identify the most influential CPPs for mAbs productivity is discussed critically for both the model on the process batches and the improved models with augmented data.

5.3.1 Monoclonal antibodies titer estimation

In this section, we analyze the performance of an MPLS that estimates the mAbs titer at harvest when only the process batches are used (i.e., base case). Then, this model is compared to the one in which process data are augmented with the *in silico* batch data generated through the digital models.

5.3.1.1 Titer estimation performance and sensitivity to the available number of process calibration batches

Here, we analyze the estimation performance of the MPLS model and assess the sensitivity of its estimation performance to the number of process batches available for calibration.

For this purpose, we iteratively increase the number of calibration batches from 3 to 50, by randomly extracting them from $\mathbf{X}_{\mathrm{PC}}$ and $\mathbf{y}_{\mathrm{PC}}$. This extraction is repeated 20 times for each number of calibration batches. At each step, a 2 LVs MPLS model is built with the available calibration batches and validated with $\mathbf{X}_{\mathrm{PV}}$ and $\mathbf{y}_{\mathrm{PV}}$. The titer estimation performance for the validation dataset and its sensitivity to the number of calibration batches are examined in terms of MAPE (averaged over the 20 iterations) as a function of the number of batches used to calibrate the MPLS model (Figure 5.5). As expected, MAPE (black dashed line in Figure 5.5) decreases with an increasing number of calibration batches. In particular, with more than 20–25 calibration batches, the MAPE average stabilizes around 210 mg/L, which is a good estimation performance, because it is comparable to the measurement error of ~150 mg/L. The MAPE increases when less than 20 batches are used for calibration and reaches large values when the number of batches is lower than 10 (MAPE > 230 mg/L). Note that a substantial increase of the estimation error is observed exactly in the range of experimental runs typically performed at the shake-flask scale, which spans between 12 and 24. Due to these inaccurate estimations, the identification of cell lines meeting the target mAb titer to be progressed in the scale-up becomes much more difficult, especially when the number of available experimental runs approaches 10. Furthermore, it should be highlighted that with less than 10 batches the model performance is inaccurate.

Furthermore, the MAPE variability (in terms of the 95% confidence region of the Gaussian MAPE distribution over different iterations, grey shaded area in Figure 5.5) increases with a low number of available batches. This indicates that the lower the number of calibration batches, the more the estimation performances are erratic and depend on the batches included in the model. In fact, if the model is calibrated on a small number of batches, the limited portion of the wide process variability captured by the model is insufficient to correctly describe new batches whose operating conditions may be far from the ones of a limited calibration dataset.

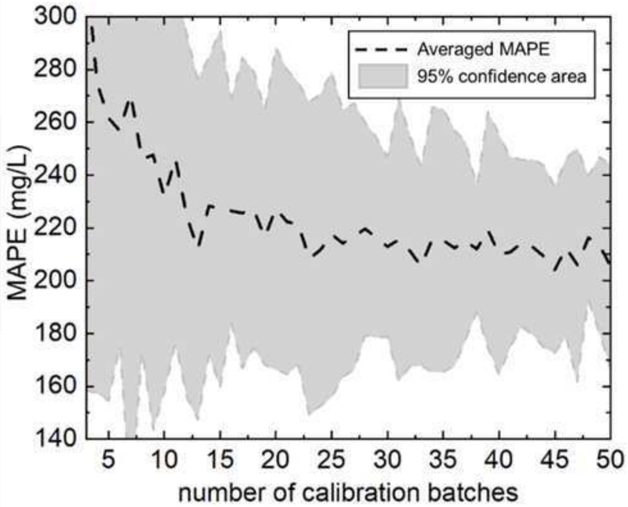


Figure 5.5 *MPLS performance sensitivity to the available process calibration batches in the estimation of mAbs titer at harvest. Black dashed line—average validation MAPE (averaged over the 20 random selections of the calibration batches from the set of process batches) as a function of the number of calibration batches; grey area—95% confidence area of the distributions.*

For this reason, the generation of *in silico* batches could be valuable to widen the variability in calibration data and cover new portions of variability which cannot be included in a limited set of calibration batches. This will eventually improve the estimation performance, providing an invaluable benefit to the selection of high productive cell lines, especially when the number of available batches is lower than 10. Since this case is often encountered in the biopharmaceutical industry at the scales of shake flasks and stirred bioreactors, where the typical number of available batches ranges between 1 and 8, in the following, we will focus on this range of batches available from the process.

5.3.1.2 Effect of data augmentation on the estimation performance

In this section, we assess the sensitivity of the MPLS estimation performance to the number of calibration process batches when data are augmented through *in silico* batches.

For this purpose, we iteratively increase the number of process calibration batches from 1 to 8, randomly extracted from a subset of 10 batches contained in $\mathbf{X}_{\mathrm{PC}}$ and $\mathbf{y}_{\mathrm{PC}}$ (the same 10 batches used for the training of the HDM, Section 2.4.1) to inspect the range of available process batches in which unsatisfactory performances were observed in the base case (Section 3.1.1). The extraction is repeated 20 times for each number of process batches. At each step, a 2 LVs MPLS model is built with the available process batches concatenated either: (*i*) with 30 FPDM *in silico* generated batches randomly extracted from $\mathbf{X}_{\mathrm{FPDM}}$ and $\mathbf{y}_{\mathrm{FPDM}}$; or (*ii*) with the 10 HDM *in silico* batches from $\underline{\mathbf{X}}_{\mathrm{HDM}}$ corresponding to each process batch used in MPLS calibration. The number of *in silico* batches is selected to increase the variability in batch behavior without overwhelming the information provided by process data. At each repetition, the MPLS models are then validated with $\mathbf{X}_{\mathrm{PV}}$ and $\mathbf{y}_{\mathrm{PV}}$. In all the cases, only the most important variables for the estimation are included in the models. Details about the selected variables will be given in the next section.

We compare the MAPE distributions in the estimation of the mAbs titer at harvest obtained through MPLS models built on: (*i*) process batches; (*ii*) process batches plus FPDM *in silico* generated batches; and (*iii*) process batches plus HDM *in silico* generated batches.

The MAPE distributions in the 20 repetitions are reported in Figure 5.6 as a function of the number of calibration process batches through boxplots. The boxes represent the 25° and 75° percentile with the median value; the dots represent the mean value of the MAPE; the error bars represent the 95% confidence intervals; and the diamonds represent errors outside the 95% confidence intervals. In Figure 5.6, green boxes represent the error distribution of the base case; red boxes represent the error distribution of the FPDM data augmentation strategy; and blue boxes represent the error distribution of the HDM data augmentation strategy.

In the base case, MAPE decreases with the number of available process batches, reaching ~180 mg/L when 8 process batches are used for model calibration (note that this value differs from Section 5.3.1.1 because the variable selection is applied here, indicating that variable selection improves the estimation performance).

When more than 5 process batches are available, both data augmentation strategies show similar performance (170 < MAPE < 200 mg/L), even if the lowest average error values are obtained in the process base case (down to ~150 mg/L). The addition of *in silico* batches considerably reduces the variability of the estimation error with respect to the process base case, independently of the augmentation strategy. This indicates that the augmented number of batches helps to increase the estimation robustness and reduces the sensitivity of the performance to the specific calibration batches. However, the average estimation error slightly increases because the *in silico* batches present some differences from the process and add variability to the dataset.

By contrast, when 4 or 5 process batches are available, the addition of the simulated batches is highly beneficial. In fact, both FPDM and HDM augmentation strategies improve the estimation performance and reduce error variability (170 < MAPE < 220 vs. 150 < MAPE < 300 mg/L). In this case, the FPDM augmentation strategy provides the largest improvement. When even less than 4 process batches are available, the FPDM augmentation strategy is very helpful for the mAbs titer estimation, because it allows better performances than both the base case and the HDM generation strategy. Good models can even be built when a very reduced number of process batches is available, namely fewer than 3 (190 < MAPE < 250 mg/L). In this case, the HDM augmentation strategy does not improve the estimation performance (results not shown) and provides errors that are similar to the ones of the base case (300 < MAPE < 500 mg/L). This is due to the high similarity between the process batches and the ones generated *in silico* through HDM.

These results show that the FPDM generation strategy allows to properly mimic the behavior of the process batches and identify the batches with high mAb titer to be progressed in the scale-up. This is because it improves the multivariate regression model estimation performance and increases the captured variability independently on process batches availability. The HDM augmentation strategy provides very good estimation performance when more than 4 or 5 process batches are available and allows to represent the behavior of the process batches more accurately than the FPDM, which makes the HDM unhelpful when the number of calibration batches is extremely small.

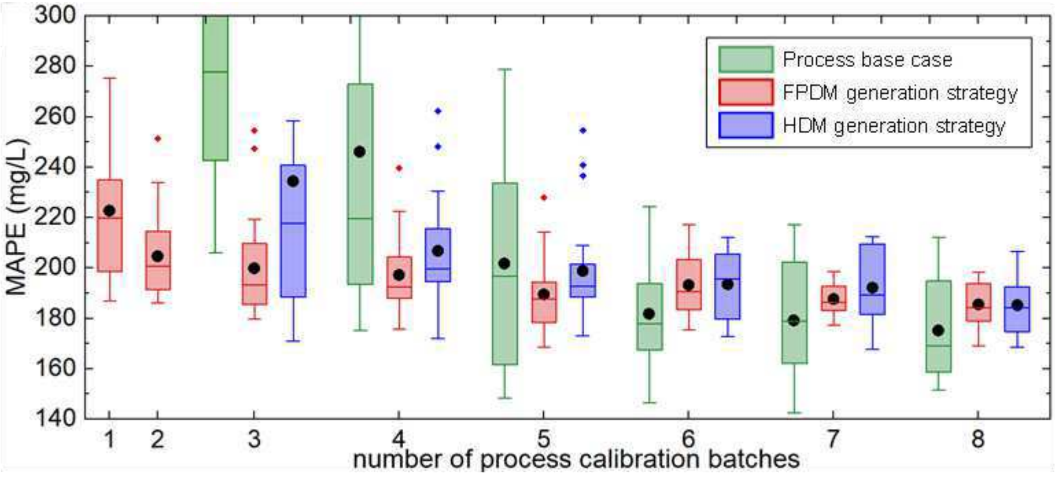


Figure 5.6 *Validation estimation performance comparison: MAPE distribution profiles from a 20-repetitions validation in the estimation of mAbs titer at harvest through MPLS. Green boxes—process base case; red boxes—FPDM data augmentation strategy; blue boxes—HDM data augmentation strategy. The boxes represent the 25° and 75° percentile and the median value, the dots represent the mean value of the MAPE, the error bars represent the 95% confidence intervals, and the diamonds are errors outside the 95% confidence intervals.*

5.3.2 Process understanding for mAbs titer estimation

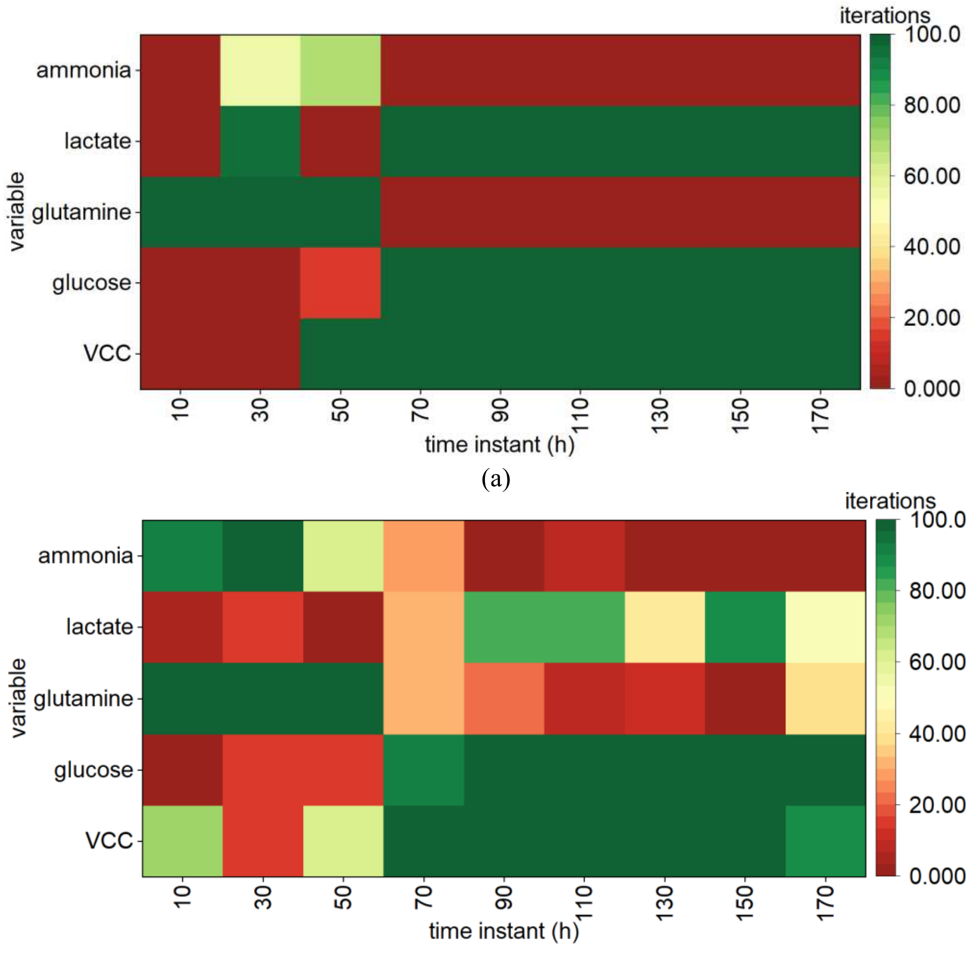
In this section, we analyze the most important CPPs (i.e., culture variables) for the estimation of mAbs titer at harvest when the data from the process are used alone and when they are combined with the batches generated through the digital models.

5.3.2.1 Process understanding with process batches only

In this section, we compare the identification of the most important culture variables for the estimation of mAbs titer at harvest in two scenarios: Scenario 1, rich in available data from the process (i.e., a high number of available batches, $N_{\mathrm{P}} = 80$ batches), and Scenario 2 with only limited data (i.e., number of available batches $N_{\mathrm{P}} = 80$).

For this purpose, two MPLS models are built on 2 LVs to estimate the mAbs titer at harvest, one for each scenario. The models are built 100 times using batches randomly extracted from $\mathbf{X}_{\mathrm{PC}}$ and $\mathbf{y}_{\mathrm{PC}}$. At each iteration, the VIP index is calculated for the model variables, and the importance of each culture variable is assessed by selection absolute frequency, namely the number of iterations in which a variable has $VIP > 1$.

The importance of the culture variables at each time point is shown in Figure 5.7 through a heatmap of the selection absolute frequency: the green color represents variables that are important with high frequency (>75–80) for the estimation of the mAbs titer at harvest, while the red one represents variables which are important only in few iterations (<20–25).



(a)

(b)

Figure 5.7 *Process understanding for mAbs titer estimation through MPLS variable importance at each time point: selection frequency for a) MPLS model calibrated with 80 process batches, and b) MPLS model calibrated with 8 process batches, randomly extracted from* $\boldsymbol{X}_{PC}$ *and* $\boldsymbol{y}_{PC}$.

In Scenario 1, when MPLS is calibrated with $N = N_{\mathrm{P}} = 80$ batches (Figure 5.7a), glucose, VCC, and lactate show high importance for mAbs titer estimation in the second half of the batch

(70 to 170 h), while glutamine shows high importance in the first half (10 to 50 h). Other variables at other time points have a very low selection frequency, except for ammonia on the second and third day of culture. As expected, the most important factors for the estimation are the concentration of viable cells (VCC) at later culture stages, which represents the number of cells that can produce mAbs, and the available glucose, which represents the available nutrient for growth and mAbs production. Similarly, glutamine, which is the limiting nutrient in the initial part of the batch and remains constant after the initial few days, is identified as particularly important within the first 50 h of the experimental batches. Lactate, instead, significantly limits cell growth only above a certain concentration, confirming its importance only in the second half of the batch. Moreover, a high concentration of ammonia in the initial part of the batch increases cell death causing a reduction in the number of producing cells, hence limiting mAbs production. Accordingly, ammonia shows moderate importance only in the first few days of culture.

The variable importance obtained in Scenario 2, the model calibrated with 8 batches (Figure 5.7b), indicates that glucose and VCC are important in the second half of the batch, while glutamine is important in the first half. However, this model fails in the identification of the importance of lactate and ammonia. In fact, their importance is not always significant as in the previous case. Furthermore, the model identifies as mildly important variables that were completely uninfluential in the previous case (see ammonia, glutamine, and VCC).

According to these results, the limited availability of batches does not allow completely reliable identification of the CCPs that are most related to the CQAs. This spoils the process understanding that can be achieved through the multivariate latent variable model. For this reason, the generation of *in silico* batches could be a valuable strategy to improve process understanding and performance.

5.3.2.2 Process understanding supported by FPDM *in silico* data augmentation

Here, we study the impact that the number of available batches has on the identification of the most important process factors for the estimation of the mAbs titer at harvest when *in silico* batch generation is performed by means of the FPDM.

The procedure utilized here is similar to the one used in Section 3.2.1. We build a 2 LVs MPLS with $N_{\mathrm{P}} = 8$ process batches plus $N_{\mathrm{FP}} = 80$ FPDM generated *in silico* to estimate the mAbs titer at harvest. The model building is repeated 100 times randomly selecting the process calibration batches from a subset of 10 batches contained in $\mathbf{X}_{\mathrm{PC}}$ and the FPDM calibration batches from $\mathbf{X}_{\mathrm{FPDM}}$. The importance of each culture variable is assessed similarly to Section 5.3.2.1. In this case, it is worth noticing that, since the ammonia is not modeled by the FPDM, it is not present in this MPLS model.

The importance of the culture variables at each time point is shown in Figure 5.8 in terms of selection frequency. VCC, glucose, and lactate are important for the estimation of mAbs titer from the second day of culture. This is coherent with important variables identified in Scenario 1, when a large number of process batches are available. Differently from the previous case, the importance of the glutamine in the first half of the batch is not identified. This is due to the simplified nature of the glutamine balance, which has not a relevant impact on the first principles model.

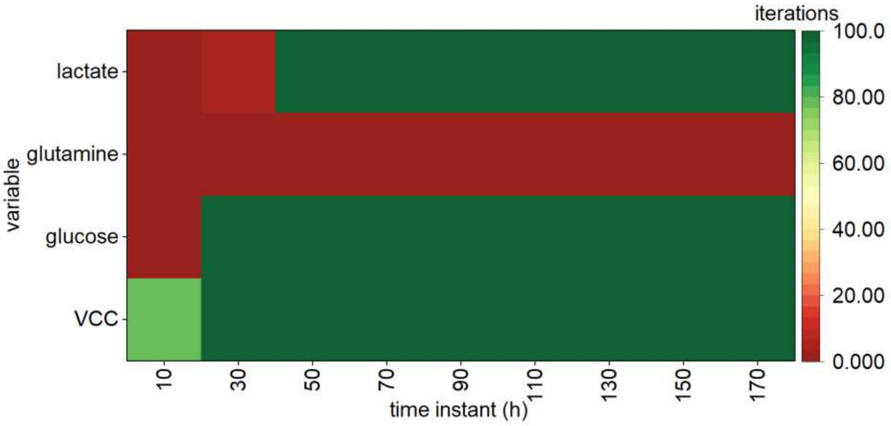


Figure 5.8 *Process understanding for mAbs titer estimation through MPLS variable importance at each time point: selection frequency for MPLS model calibrated with 8 process batches, randomly extracted from a subset of 10 batches contained in* $\boldsymbol{X}_{PC}$ *and* $\boldsymbol{y}_{PC}$, *and 80 FPDM in silico generated batches from* $\boldsymbol{X}_{FPDM}$ *and* $\boldsymbol{y}_{FPDM}$.

This result shows that the generation of *in silico* batches through the FPDM model provides an improved identification of the important variables, even if a limited number of process batches is available. In fact, the addition of *in silico* batches allows identifying more clearly the variables that are important for the estimation than process Scenario 2 (Section 5.3.2.1), having the same availability of process batches. However, this improved understanding strongly relies on the effectiveness of the model used for batch generation. In fact, the *in silico* batches do not allow correct identification of the glutamine importance, due to the simplified nature of its equations. Despite that, in absence of additional process information, the FPDM *in silico* batch generation is helpful to improve process understanding, even when a simplified model is available.

5.3.2.3 Process understanding supported by HDM *in silico* data augmentation

In this section, we study the impact of the HDM *in silico* batch generation on the identification of the most predictive variables for the mAb titer at harvest. The procedure is analogous to the one presented in the previous section, but here HDM *in silico* batches are combined with process ones. The 10 HDM batches corresponding to each training process batch are used for the augmentation.

The importance of the culture variables for the titer estimation at each time point is shown in Figure 5.9. VCC, glucose, and lactate in the second half of the batch (70–170 h) are identified to be the most important variables for mAbs titer estimation. This result is in accordance with the important variables identified in Scenario 1, when a large number of process batches are available. However, lactate shows an average selection frequency (~60), meaning that the identified relationship between lactate and mAbs titer is not as strong as it appears from the process batches. Furthermore, similarly to Scenario 1, glutamine is correctly identified as important in the first half of the batch (10–50 h) and irrelevant in the second half, while ammonia as mildly important only in the first half of the batch. However, glutamine at 10 h has a relatively low selection frequency (~40), indicating that its importance is not correctly identified. Finally, several variables that result to be uninfluential from the process data (Scenario 1) show an average selection frequency (~50).

This result shows that HDM *in silico* batch generation does not identify the important process factors which are completely faithful to the one provided by process batches. In fact, the identification performance is not better than process Scenario 2 when only the reduced number of process batches is used. This is probably due to the high representation accuracy of the HDM, resulting in *in silico* batches very similar to the training ones. For this reason, the HDM does not increase the amount of information contained in the augmented data, providing less accurate identification of the important factors.

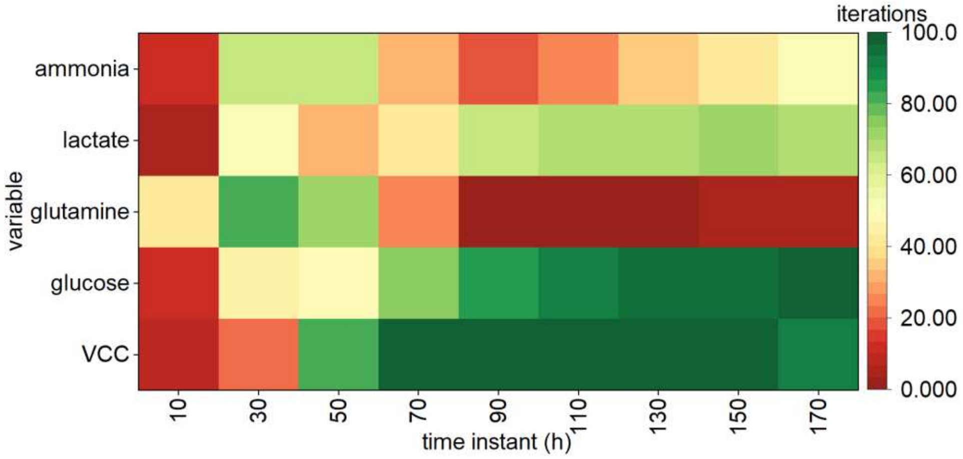


Figure 5.9 *Process understanding for mAbs titer estimation through MPLS variable importance at each time point: selection frequency for MPLS model calibrated with 8 process batches, randomly extracted from a subset of 10 batches contained in* $\boldsymbol{X}_{PC}$ *and* $\boldsymbol{y}_{PC}$*, and 80 HDM in silico generated batches extracted from* $\boldsymbol{X}_{HDM}$ *and* $\boldsymbol{y}_{HDM}$*, 10 for each of the corresponding calibration batches.*

5.4 Conclusions

In this work, we investigated the utility of *in silico* data augmentation through digital models to support the development of monoclonal antibodies in scenarios when only a few experiments can be carried out at a given scale. In particular, we investigated two strategies for *in silico* data generation: a first principles digital model and a hybrid digital model. We applied these strategies to increase the number of available data used in multivariate regression models to estimate the antibody titer at harvest in a simulated process for the production of monoclonal antibodies on a shake-flask scale.

Both *in silico* data generation strategies demonstrated to be very effective. In particular, the first principles digital model augmentation strategy allowed a significant improvement in the estimation performance especially when the number of available process batches is extremely limited (1-5), providing a low estimation error of the antibody titer at harvest, comparable with the typical measurement errors (~150–200 mg/L). Furthermore, the first principles digital model improved process understanding. In fact, it allowed to clearly provide process understanding and identify the most important CPPs for the CQA (namely, the mAbs titer at harvest), even when the availability of process batches is limited (<10). The hybrid digital model generation strategy, instead, did not allow an equivalent identification of the important CPPs. Nonetheless, it improved the estimation performance when the number of available process batches is greater than 4. It should be highlighted that the success of *in silico* data generation relies on the quality of the digital model and its representativeness of the process.

In silico data generation could provide great advantages at different scales of the product and process development, especially at the stirred bioreactor scales, where the number of available batches is typically between 2 and 10.

This study is a proof of concept for the use of *in silico* data generation in the biopharmaceutical field and further studies will be oriented to adapt the investigated strategies to *in vitro* applications. Specifically, different ways of combining and pre-processing process and *in silico* data will be studied. Furthermore, strategies to estimate the parameters for *in silico* data generation from the experimental batches will be developed.

Chapter 6

Bioprocess feeding optimization through in silico dynamic experiments and hybrid digital models[*]

In this Chapter, we compare the feeding schedule optimization of mammalian cell cultures performed by means of an *in silico* experimental campaign on a hybrid digital model and an experimental campaign on the process. This to show if the *in silico* experimental campaign permits to accelerate the experimentation and to reduce the experimental burden. As a proof of concept, the proposed methodology is applied on a simulated process for the production process of monoclonal antibodies at 1-L shake flask scale. Design of Dynamic Experiments (DoDE) is used to design optimal experiments that are then utilized to train a hybrid semi-parametric digital model. Despite the hybrid digital model requires only a very limited number of experiments to be accurately trained (i.e., 9), it outperforms the results obtained by the experimental campaigns planned with DoDE on a much larger number of experiments (i.e., 31), achieving a 2.8% higher antibody titer than the DoDE campaigns and a 34.9% improvement in the antibody titer with respect to the experimental campaign used to train the hybrid model

6.1 Introduction

Monoclonal antibodies (mAbs) are biological drugs which are gaining great interest for the treatment of autoimmune, oncological and infectious diseases (Castelli et al., 2019): in 2018 they represented 53% of the overall biopharmaceutical approvals by the regulatory agencies and 65.6 % of the entire biopharmaceutical sales (Walsh, 2018). At the industrial scale, mAbs are produced in fed-batch cultures of mammalian cells, which are appositely generated to secrete the desired product (O'Flaherty et al., 2020; Wurm, 2004).

The development of mAbs is multi-step process which requires a lot of resources in terms of time and capital investments, because it usually lasts several years and costs billions of dollars (Epifa, 2021; Farid et al., 2020). The upstream development of mAbs starts with cell line generation, screening and selection, and process characterization. At this stages, a large pool of

[*] Barberi, G., Giacopuzzi, C., Facco, P. Bioprocess feeding optimization through in silico dynamic experiments and hybrid digital models - A proof of concept. *In preparation*

producing cell lines is generated and tested at different process scales (Barberi et al., 2022; Facco et al., 2020) to identify the ones meeting the desired performance in terms of growth, productivity and product quality (Gronemeyer et al., 2014; Tripathi & Shrivastava, 2019). Furthermore, the relationship between critical process parameters (CPP) and critical quality attributes (CQA) is studied for regulatory compliance and for the following process optimization phase. During process optimization, the bioreactor operating parameters, in terms of temperature, pH, agitation, dissolved oxygen, etc., are adapted to the specific host system to enhance cell growth and specific productivity (Gronemeyer et al., 2014; F. Li et al., 2010; Tripathi & Shrivastava, 2019). Similarly, an appropriate optimization of the medium and feeding strategy is required to balance cell growth, productivity and product quality (S. H. Kim & Lee, 2009; Ling et al., 2015; Tripathi & Shrivastava, 2019).

High-throughput scaled-down equipment and statistical Design of Experiments (DoE) are the most common methodologies to optimize media and feeding strategy in a systematic way (F. Li et al., 2010; Mora et al., 2019; W. Zhou et al., 1997). Typically, cell cultures are fed with frequent boluses of glucose and glutamine to maintain a low concentration, which minimize the production of by-products, such as lactate and ammonia (F. Li et al., 2010). Hence, the optimization of the feeding strategy requires to determine the best way of providing feed boluses over time. However, DoE only deals with "static" factors. To deal with the batch process dynamics, DoE can be exploited by assigning a different DoE factor to the feeding action at each day (Mora et al., 2019), but this results in a design with too many factors that requires several dozens of experiments. An appropriate solution to this issue is the adoption of Design of Dynamic Experiments (DoDE), which guarantees to optimize time-varying factors while minimizing the number of experimental runs (Georgakis, 2013). In fact, DoDE utilizes dynamic subfactors to code the time-varying factors' profiles, and then build a Response Surface Model (RSM) to correlate the factors' dynamic profile to the CQA. Research on DoDE application in the bioprocessing field is still ongoing, but some applications on *in silico* fermentation processes (Klebanov & Georgakis, 2016) and simulated mammalian cell cultures (Wang & Georgakis, 2017) are available in the Literature.

However, despite being designed to maximize the content of information obtained by the experiments while minimizing the number experimental runs, the number of experiments designed by DoDE rapidly increases with the number of dynamic variables and the complexity of their dynamic profiles, leading to high numbers of required experimental runs. Since each experimental run can last for several weeks and cost tens of thousands of dollars, the duration and cost of large experimental campaigns limits the applicability of DoDE in the biopharmaceutical industry. Accordingly, strategies to limit the allocation of resources for the experimental campaigns are of paramount importance.

Hybrid semi-parametric digital models represent an innovative solution to reduce experimental requirements. They combine mechanistic knowledge of the system under investigation with data-driven methods, which learn complex and possibly unknown relationship among the system variables from experimental data (Sansana et al., 2021; von Stosch et al., 2014; S. Yang et al., 2020).

Hybrid semi-parametric models were widely applied to the bioprocess development for tasks such as prediction, process understanding, and process and quality monitoring. For example, an improved understanding of the relationship of biomass and productivity with the process parameters in microbial cell culture was achieved through hybrid semi-parametric models (von Stosch et al., 2016), while good prediction accuracy was attained by hybrid models trained on intensified DoE data (von Stosch & Willis, 2017), allowing to accelerate upstream process characterization (Bayer, Striedner, et al., 2020). In mammalian cell cultures, the prediction performance of hybrid models were tested in interpolation and extrapolation scenarios (Narayanan, Luna, et al., 2021), while compared to purely multivariate techniques the prediction of the main culture variables through hybrid models resulted more accurate (Narayanan et al., 2019). In the same context, hybrid semi-parametric models coupled with Extended Kalman Filter was used to monitor the glucose concentration in bioreactors, suggesting the appropriate timing for the feeding action to avoid cell starvation (Narayanan et al., 2020).

Hybrid semi-parametric models were also used for bioprocess optimization. For example, the optimal processing conditions (Ferreira et al., 2014) and glucose feeding strategy (Teixeira et al., 2006) for microbial cell cultures were identified through an iterative batch-to-batch strategy based on hybrid models: the optimal condition identified by the hybrid model at each step was used to retrain the model for further optimizations. A similar strategy identified static process parameters improving the product yield in *E. Coli* cultures by means of 9 experimental runs, only, 5 from the initial exploratory campaign, and 4 suggested in the batch-to-batch optimization (Bayer et al., 2021). Furthermore, the feeding schedule of mammalian cell culture was optimized by means of hybrid semi-parametric models (Teixeira et al., 2005; Teixeira, Alves, et al., 2007), showing the applicability of these new methodologies in mammalian cell culture optimization.

Despite hybrid models were applied for the bioprocess optimization, and their added value on the optimization of mammalian feeding schedule was proven, the advantages of using hybrid semi-parametric models in feeding schedule optimization during bioprocess development are underexplored, and research is still needed to allow a consistent applicability of hybrid models in bioprocess optimization.

This study compares an *in silico* experimental campaign for the optimization of the feeding schedule in mammalian cell cultures through hybrid digital models with an experimental campaign on the process to evaluate if the *in silico* experimental campaign can accelerate the experimentation and reduce the experimental burden in the process development. In particular, we use a hybrid semi-parametric model calibrated on the experiments designed through DoDE in such a way as to identify the time profiles of fed glucose and glutamine which maximize the antibody titer. The proposed methodology is tested on a well-established simulated process for the production of mAbs at a shake flask scale.

6.2 Materials and methods

In this Section, the mathematical methodologies used in this Chapter are presented.

6.2.1 Proposed methodology

In this work, an *in silico* experimental campaign (Strategy #1) for the optimization of the feeding schedule of mammalian cell cultures is proposed (Figure 6.1a). The adopted procedure comprises five steps:

1. DoDE experiments planning: initially, experiments are planned according to a DoDE (Section 6.2.2) on 2 dynamic factors, namely, the time profiles of glucose and glutamine concentrations, and a response, namely, the antibody titer at harvest;
2. experiments execution: planned experiments are executed on a simulated process (Section 6.2.3). In this study, we used a simulated process because it allows to know exactly the relationship between nutrients and antibody titer which can be exploited for identifying the optimal feeding schedule, which is the reference to evaluate the performance of the proposed optimization strategy. Furthermore, it allows to follow in real-time the entire time profiles of the culture variables, whose measurements in a real case are available only at a much lower frequency (every few hours);
3. training of the hybrid model: a hybrid semi-parametric model (Section 6.2.4) is trained on the experiments executed at step 2;
4. optimization: a genetic algorithm (Section 6.2.5) is used to identify the feeding schedule maximizing the antibody titer at harvest. This algorithm exploits the hybrid model to simulate *in silico* experiments and predict the resulting antibody titer given the profiles of glucose and glutamine. In this work, genetic algorithm was used to avoid the complex calculation of the objective function gradients and accelerate the optimization step by requiring only the evaluation of the objective function;
5. execution of the confirmatory experiment at the optimal conditions: once the optimal nutrient profiles (i.e., feeding schedule) are identified, they are executed on the process to assess the antibody titer that the process can achieve and the reliability of the predicted values.

Optimization Strategy #1 is compared with optimization Strategy #2 (Figure 6.1b), namely, a standard experimental campaign for the optimization of the feeding schedule carried out directly on the process. The experimental campaign has the same steps 1 and 2 as the ones of Strategy #1, but continues with step 3-5 as follows:

3. response surface modeling: a RSM is built with the experiments executed at step 2 according to the DoDE theory. The model is used to predict the antibody titer at harvest from the DoDE dynamic subfactors, after being updated by excluding the effects with low influence on the response (Section 6.2.2);
4. optimization: in this case, the genetic algorithm exploits the RSM to predict the antibody titer given the profiles of glucose and glutamine;
5. execution of the confirmatory experiment at the optimal conditions: similar to Strategy #1.

The confirmatory experiments obtained at step 5 of both optimization Strategies are then compared with the process optimum, which is known in this study because the process is simulated. In the next Sections, details on the DoDE, the process, the hybrid model, and the techniques used for the experimentation and the optimization are presented.

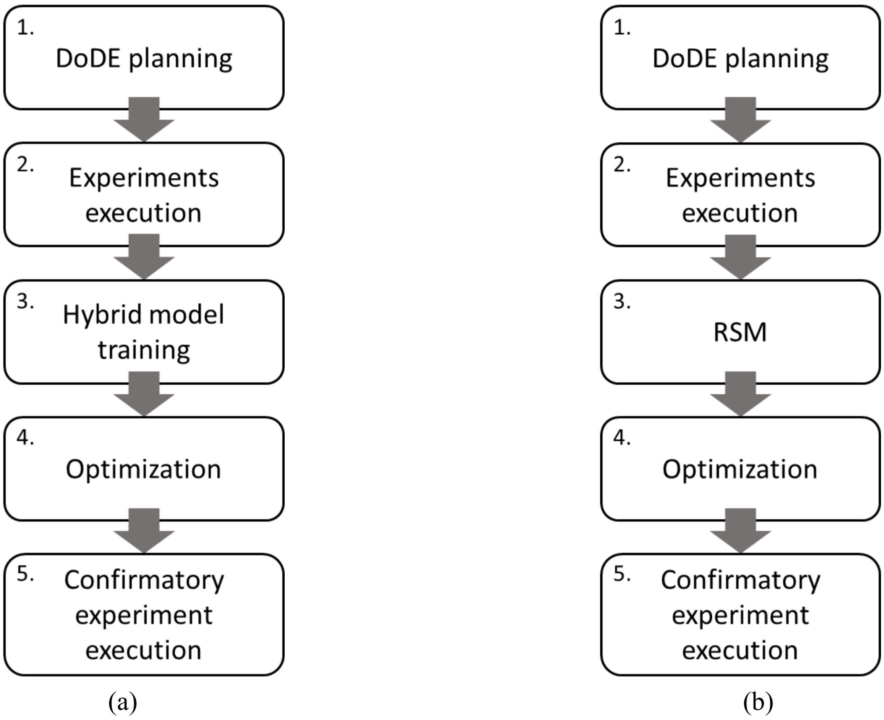


Figure 6.1 *Proposed methodology: (a) optimization Strategy #1 (in silico), and (b) optimization Strategy #2 (experimental).*

6.2.2 Design of Dynamic Experiments

Design of Dynamic Experiments (Georgakis, 2013) is used in this work to plan the experimental campaign for the optimization of the glucose and glutamine profile in the cell culture. In DoDE, the time-varying factors (i.e., manipulated variables) are expressed as normalized dynamic variables $z(\tau)$, which varies between -1 and 1. Normalized dynamic variables are the sum of orthogonal time-varying profiles weighted by dynamic subfactors x_i, which are equivalent to the Design of Experiment factors. The normalized dynamic variables are defined as:

$$z(\tau) = \sum_{i=1}^{I} x_i P_{i-1}(\tau) \quad , \tag{6.1}$$

where $P_{i-1}(\tau)$ is a shifted Legendre polynomial of degree $i-1$, and $\tau = t/t_b$ is the dimensionless culture time (i.e., the % of experimental batch completion), with t_b being the culture duration. Details on the expression of the Legendre polynomials can be found in the original reference by Georgakis (2013).The number of subfactors defines the maximum degree of the $z(\tau)$ profile. In this work, to have independent profiles for each nutrient with second degree curvature, and avoid an excessive number of factors, $I = 3$ dynamic subfactors are used for each nutrient, summing up to a total of $K = 6$ dynamic subfactors: specifically, subfactors $x_1^{\text{glc}}, x_2^{\text{glc}}, x_3^{\text{glc}}$ refer to the glucose profile and $x_1^{\text{gln}}, x_2^{\text{gln}}, x_3^{\text{gln}}$ to the glutamine one. Independently on the specific nutrient, x_1 (Figure 6.2a) controls the initial value of the profile (e.g., 1 correspond to the top of the interval, while -1 to the bottom), x_2 (Figure 6.2b) controls the overall increasing or decreasing tendency of the profile (e.g., 1 correspond to fully increasing profiles, while -1 to fully decreasing), and x_3 (Figure 6.2c) controls the concavity of the profile (e.g., 1 upward and -1 downward).

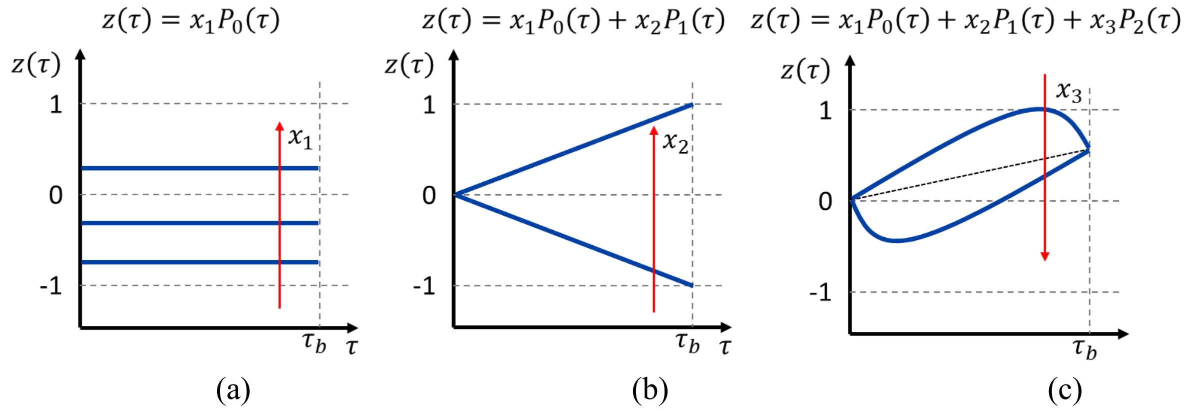

$z(\tau) = x_1 P_0(\tau)$
$z(\tau)$
1
0
-1
x_1
τ_b
τ
(a)
$z(\tau) = x_1 P_0(\tau) + x_2 P_1(\tau)$
$z(\tau)$
1
0
-1
x_2
τ_b
τ
(b)
$z(\tau) = x_1 P_0(\tau) + x_2 P_1(\tau) + x_3 P_2(\tau)$
$z(\tau)$
1
0
-1
x_3
τ_b
(c)

Figure 6.2 *Effect of the dynamic subfactors on the normalized dynamic variable z(τ) for a 3 subfactors design: (a) x_1, (b) x_2, and (c) x_3. The red arrows indicate the direction of increasing subfactors.*

To ensure that $-1 \leq z(\tau) \leq 1$, the dynamic subfactors must satisfy the following constraints:

$$-1 \leq x_1^{\text{glc}} \pm x_2^{\text{glc}} \pm x_3^{\text{glc}} \leq 1 \quad , \tag{6.2}$$

$$-1 \leq x_1^{\text{gln}} \pm x_2^{\text{gln}} \pm x_3^{\text{gln}} \leq 1 \quad , \tag{6.3}$$

and the value of each subfactor must be bounded, as well:

$$-1 \leq x_i \leq 1 \quad . \tag{6.4}$$

The glucose and glutamine concentration profiles planned through the DoDE can be determined from the respective $z(\tau)$ according to the relation:

$$u_j(\tau) = u_{j,0} + z_j(\tau)\Delta u_j \quad \text{with } j = \text{glc or gln}, \tag{6.5}$$

where:

$$u_{j,0} = \frac{u_{j,\max}+u_{j,\min}}{2} \quad \text{and} \tag{6.6}$$

$$\Delta u_j = \frac{u_{j,\max}-u_{j,\min}}{2} \quad , \tag{6.7}$$

being $u_{j,\max}$ and $u_{j,\min}$ the maximum and minimum values in which the profile of each nutrient j is allowed to vary. In this work, glucose and glutamine are assumed to vary in the ranges $[20, 50]$ mM and $[2, 10]$ mM, respectively. These values are selected to remain in proximity of the concentration at which the process operates (Kontoravdi et al., 2010).

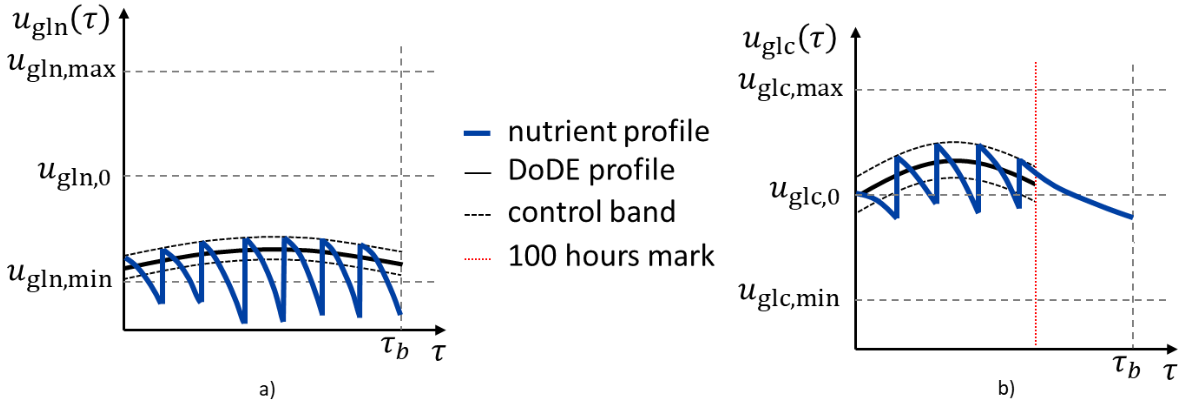


Figure 6.3 *Schematic representation of (a) glutamine and (b) glucose profiles (blue lines) with the profile determined by the DoDE (black line), the 110/90% control band (black dashed lines), and the 100 hours limit for glucose feeding (red dotted line).*

Since the nutrients are both manipulated and observed, their concentrations vary because of both the cell consumption and the feeding. In this work, we simulate to have only off-line measurements, because advanced monitoring strategies, such as on-line monitoring and control systems, are not standard in industrial mammalian cell cultures yet. Furthermore, the measurements and feeding actions are performed once every 24 hours. For these reasons, the nutrients profile cannot precisely follow the one proposed by DoDE. To deal with this issue, we introduce a specific procedure replicate the profiles indicated by the DoDE during the experiments, which is schematically represented in Figure 6.3. The proposed procedure consists of:

- defining a 10% band around the DoDE profile which is intended to control the feeding actions (Figure 6.3a, black dashed lines);

2 performing the feeding of the nutrients only if their concentration in the culture < 90% of the concentration defined by the designed experiment (Figure 6.3a, lower black dashed line);

- the feeding is performed using a predefined amount of fresh medium with a nutrient concentration allowing to achieve 110% of the concentration defined by the designed experiment (Figure 6.3a, upper black dashed line).

The feeding is visible in the nutrient profiles (Figure 6.3a) as the vertical jumps in the blue line where the nutrient concentration is brought to the 110% the one defined by the designed experiment. Furthermore, since the glucose consumption is slow and hardly decreases in the final part of the batch, the glucose cannot follow sharply decreasing profiles, hence it is controlled (and the feeding performed, accordingly) only in the first 100 hours of the batch (Figure 6.3b, red dotted line). After this point the glucose is fed only to compensate for any dilution effect due to glutamine addition. This is shown in Figure 6.3b, where after 100 hours (red dotted line) the feeding is not performed, and nutrient concentration decreases because of cell consumption. Accordingly, the glucose profile after 100 hours has no controllable effect on the antibody titer; hence, the glucose profile in the final part of the culture is not taken in consideration in the analysis.

In this work, the DoDE nutrient profiles are designed by means of a D-optimal design of experiment (de Aguiar et al., 1995) applied to the $K = 6$ subfactors. Once the experiments are executed on the process, a RSM (Montgomery, 2007) is fitted on the experimental data obtained from the designed experimental campaign through multiple linear regression. A second-order RSM (typically used for optimization) is defined to predict the antibody titer at harvest from the dynamic subfactors:

$$\hat{y} = \beta_0 + [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_K]\mathbf{x} + \mathbf{x}^{\mathrm{T}} \begin{bmatrix} \Delta_{1,1} & \Delta_{1,2} & \cdots & \Delta_{1,K} \\ \Delta_{2,1} & \Delta_{2,2} & \cdots & \Delta_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{K,1} & \Delta_{K,2} & \cdots & \Delta_{K,K} \end{bmatrix} \mathbf{x} \ , \tag{6.8}$$

where $\hat{y}$ is the predicted antibody titer, $\mathbf{x} = \begin{bmatrix} x_1^{glc} & x_2^{glc} & x_3^{glc} & x_1^{gln} & x_2^{gln} & x_3^{gln} \end{bmatrix}$ is the column vector of the dynamic subfactors for a single experiment, β_k and $\Delta_{k,k}$ are the first order and higher order parameters of the RSM, respectively. The model parameters are estimated minimizing the residual error in a least-square manner. Each addend of in the Equation (6.8) defines an effect, namely, the way in which each subfactor determines the variability of the response.

The RSM is affected by uncertainty. The uncertainty of the estimated parameter $\hat{\beta}_e$ (i.e., $\hat{\beta}_k$ or $\hat{\Delta}_{k,k}$) for the effect e is determined through the parameter confidence intervals:

$$\hat{\beta}_e \pm t_{1-{\alpha}/{2},N-E} \sqrt{\frac{\frac{1}{N-E}\sum_{n=}^{N}(y_n-\hat{y}_n)^2}{\sum_{n=1}^{N}\left(x_{n,e}-\bar{\mathbf{x}}_e\right)^2}} \tag{6.9}$$

where y_n is the measured response of the n-th experiment, $\hat{y}_n$ is the response of the n-th experiment predicted by Eq. (6.8), $x_{n,e}$ is the value of e-th effect for the n-the experiment, $\bar{x}_e$ is the average value of e-th effect, N is the total number of experiments, E is the total number of effects, and $t_{1-{\alpha}/{2},N-E}$ is the critical value of Student's t distribution with $N - E$ degrees of freedom calculated at the confidence level $\alpha = 0.05$. The effects whose confidence interval crosses the 0 are affected by too high uncertainty and, accordingly, removed from the model.

The uncertainty on the parameter determines also uncertainty in the predictions, which, for a validatory experiment with subfactors $\mathbf{x}_{\mathrm{NEW}}$ is assessed through the 95% prediction interval (Wang & Georgakis, 2019):

$$PI = t_{1-\alpha/_{2},N-E}\sqrt{s_e^2(1+\mathbf{x}_{\mathrm{NEW}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_{\mathrm{NEW}})} \quad , \tag{6.10}$$

where $s_e^2 = SSE/(N-E)$ and SSE is the sum squared error of the model, and $\mathbf{X}$ is the matrix containing the subfactors vectors for all the N designed DoDE experiments placed along the rows, and $t_{1-\alpha/2,N-E}$ is the critical value of Student's t distribution with $N-E$ degrees of freedom calculated at the confidence level $\alpha = 0.05$. The real response y of a confirmatory experiment is expected to lie in the interval $\hat{y} - PI \le y \le \hat{y} + PI$ with a confidence of 95%. To assess the extent of process improvement that can be achieved planning a different number of experiments, DoDE is adopted to design two alternative experimental campaigns: experimental campaign A and B. Experimental campaign A is a complete campaign for process optimization, used to assess the process improvement that can be achieved with an extended experimental campaign. A second-order with pairwise interaction RSM (as Eq. 6.8) is fitted with data from 31 experiments planned by assigning the values of the dynamic subfactors through a D-optimal Design of Experiment (Appendix E.1, Table E.1). Among the 31 experiments, 28 are required to fit a the RSM for the 6 dynamic subfactors, while the 3 remaining experiments are used to estimate the model lack-of-fit (Georgakis, 2013). Experimental campaign B is used to assess the process improvement that can be achieved with a small number of experiments. Data from 9 experiments planned by assigning the values of the dynamic subfactors through a D-optimal Design of Experiment (Appendix E.1, Table E.2) are used to fit a first-order RSM:

$$\hat{y} = \beta_0 + [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_K]\mathbf{x} \quad . \tag{6.11}$$

Among the 9 experimental runs, 7 are used to fit the RSM for the 6 dynamic subfactors, while the 2 remaining experiments are used to estimate the model lack-of-fit (Georgakis, 2013).

6.2.3 Process for the production of monoclonal antibodies at 1-L shake flasks scale

A simulated process for the production of monoclonal antibodies at 1-L shake flasks scale (Kontoravdi et al., 2010) is considered in this work. It models the dynamic behavior of the Viable Cell Density (VCD, X_v), and the concentration of the main nutrients and by-products, such as glucose (c_{glc}), glutamine (c_{gln}), lactate (c_{lac}), and ammonia (c_{amm}). Additionally, RNA, and light and heavy chain balances in the cytosol and Golgi apparatus are considered to simulate protein synthesis and model the dynamic behavior of antibody titer (c_{mAb}). Details on the model and the respective parameters can be found in the work by Kontoravdi et al. (2010).

The total duration of a batch is $t_b = 168$ hours, with an initial volume of 200 mL and inoculation cell density of $0.2 \cdot 10^6$ cell/mL (Kontoravdi et al., 2010). Measurement sampling was simulated every 24 hours through the withdrawn of 2.5 mL from the culture. Feed of glucose and/or glutamine is performed after the sampling through the addition of 20 mL of concentrated medium in 10 minutes, to simulate a bolus addition without causing a too severe concentration change. The concentration of glucose and glutamine is determined at any feeding addition in such a way as to reach the nutrient concentration profiles planned by DoDE. The model is integrated between each sampling time instant through a variable-step variable-order solver with a maximum order of 5. A 6% white noise is added as measurement error.

6.2.4 Hybrid model

A serial hybrid semi-parametric model is used (Oliveira, 2004; Teixeira et al., 2005; von Stosch et al., 2014) to capture the behavior of mammalian cell cultures producing mAbs. This digital model combines a mechanistic model, which embeds the knowledge of the system, and an artificial neural networks (ANN), which accounts for the unknown dependences in the system under study.

The mechanistic knowledge of the cell culture is described by the concentration balances for the main culture variables, organized in the column vector $\mathbf{c} = [X_\mathrm{v} \quad c_\mathrm{glc} \quad c_\mathrm{gln} \quad c_\mathrm{lac} \quad c_\mathrm{amm} \quad c_\mathrm{mAb}]$:

$$\frac{d\mathrm{c}(t)}{dt} = \mathbf{r}(\mathbf{c}^*(t), \boldsymbol{\omega}) - D_V\mathbf{c}(\mathrm{t}) + \mathbf{u} \quad , \tag{6.12}$$

where $\mathbf{r}(\mathbf{c}^*(t), \boldsymbol{\omega})$ $[V \times 1] = [6 \times 1]$ is the vector of volumetric reaction rates for the V culture variables, $\mathbf{c}^* = [X_\mathrm{v} \quad c_\mathrm{glc} \quad c_\mathrm{gln} \quad c_\mathrm{lac} \quad c_\mathrm{amm}]$ is the column vector of culture variables with the exclusion of the antibody titer, $\boldsymbol{\omega}$ is the vector of the ANN parameters (weights and biases), D_V is the dilution factor, and $\mathbf{u}$ $[6 \times 1]$ is the vector of controlled inputs. The volumetric reaction rates can be expressed as combination of the specific production/consumption rate and the viable cell concentration X_v:

$$\mathbf{r}(\mathbf{c}^*(t), \boldsymbol{\omega}) = \mathbf{S}\, X_\mathrm{v}\, \boldsymbol{\mu}(\mathbf{c}^*(t), \boldsymbol{\omega}) \quad , \tag{6.13}$$

where $\mathbf{S}$ is the stoichiometric matrix with -1 and 1 on the diagonal for consumed and produced components, respectively, and $\boldsymbol{\mu}(\mathbf{c}^*(t), \boldsymbol{\omega})$ $[6 \times 1]$ is the vector of the specific production/consumption rates. Here, we assume that the production/consumption rates do not depend on antibody titer, because it is expected to have no impact on other culture variables (Narayanan, Luna, et al., 2021).

The relationship between specific production/consumption rates and culture variables is typically very complex and accurate mechanistic expressions are not typically available. This lack of knowledge is compensated with a data-driven model which captures the relationship f between specific production/consumption rates and culture variables, $\boldsymbol{\mu} = f(\mathbf{c}^*(t), \boldsymbol{\omega})$, learning it from experimental data. A single hidden layer ANN with 5 neurons and hyperbolic tangent activation functions are used in this work to capture the nonlinear relationship between culture variables and specific production/consumption rates:

$$\boldsymbol{\mu}(\mathbf{c}^*(t), \boldsymbol{\omega}) = \boldsymbol{\mu}_{\max} \circ \boldsymbol{\omega}^{(2,1)}\tanh\left(\boldsymbol{\omega}^{(1,1)}\mathbf{c}^*(t) + \boldsymbol{\omega}^{(1,2)}\right) + \boldsymbol{\omega}^{(2,2)} \quad , \tag{6.14}$$

where $\boldsymbol{\omega}^{(1,1)}$ and $\boldsymbol{\omega}^{(2,1)}$ are the weights, and $\boldsymbol{\omega}^{(1,2)}$ and $\boldsymbol{\omega}^{(2,2)}$ are the biases of the hidden and output layer, respectively, $\boldsymbol{\mu}_{\max}$ $[6 \times 1]$ is the vector of the maximum production/consumption rates, and ○ represents the Hadamard product. The vector of maximum production/consumption rates, $\boldsymbol{\mu}_{\max}$, is used to scale the output of the ANN at different magnitudes (Teixeira, Alves, et al., 2007) and is heuristically determined in preliminary tests. The number of hidden neurons was selected as the one maximizing the Bayesian Information Criterion (BIC; Schwarz, 1978; von Stosch and Willis, 2017).

The model is integrated between each feeding action, which is simulated by appropriately changing the controlled input vector $\mathbf{u}$. Similarly to the process, 20 mL of fresh medium with the nutrients is added in 10 minutes

The model parameters $\boldsymbol{\omega}$ are estimated from experimental data, containing the measured culture variable and the values of the controlled inputs required to follow the nutrient profiles planned by DoDE (as explained in Section 6.2.2). The hybrid model is trained with the 9 experiments of experimental campaign B (Section 6.2.2) with a stepwise decreasing learning rate (from 0.005 to 0.0001). The same 9 experiments with the addition of 2.5% white noise are used as internal validation experiments to stop the training procedure and make the model robust to noise. The model parameters $\boldsymbol{\omega}$ are estimated through the Adam optimization algorithm (Kingma & Ba, 2015). Adam algorithm was used as one of the best gradient-based algorithms, which are the standard and most effective ways to train ANNs. A norm-two regularized (A. Yang et al., 2011) weighted sum of squared error is used as objective function:

$$\mathcal{L} = \sum_{t=0}^{t_b}\sum_{n=1}^{N}\sum_{v=1}^{V}\frac{\left(c_{n,v}(t) - \hat{c}_{n,v}(t)\right)^2}{\sigma_v^2(t)} + \lambda_{\text{reg}}\sum_{t=0}^{t_b}\|\boldsymbol{\mu}(t)\|^2 \quad , \tag{6.15}$$

where $c_{n,v}(t)$ and $\hat{c}_{n,v}(t)$ are the measured and predicted v-th culture variable for the n-th experiment at the time instant t, respectively, $\sigma_v^2(t)$ is the variance of the v-th measured culture variable at time instant t, and λ_{reg} is the regularization coefficient. In this work, λ_{reg} was heuristically set to 0.05. The gradients of the objective function with respect to the ANN parameters required by the Adam algorithm are calculated as:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\omega}} = \left(\frac{\partial\mathcal{L}}{\partial\mathbf{c}}\right)_t\left(\frac{\partial\mathbf{c}}{\partial\boldsymbol{\omega}}\right)_t \quad , \tag{6.16}$$

where the matrix $(\partial\mathbf{c}/\partial\boldsymbol{\omega})_t$ is calculated through the sensitivity equations (Oliveira, 2004). The sensitivity equations are integrated together with the model equations Eq. (6.12) starting from the initial condition $(\partial\mathbf{c}/\partial\boldsymbol{\omega})_{t=0} = 0$. To avoid local minima, $N_{\text{models}} = 20$ hybrid models are trained starting from initial parameters randomly initialized with values form a normal distribution with variance σ_0^2, randomly selected in the interval $[0.05, 0.001]$. The model showing the smallest sum of squared error over the internal validation experiments is selected and used for the analysis.

The uncertainty of hybrid model prediction is calculated from the $N_{\text{models}} = 20$ trained models by exploiting the prediction intervals determined by means of the population of predicted antibody titer values for the training experiments. The half width of the prediction interval is calculated as:

$$PI_H = t_{1-\alpha/2, N \cdot N_{\text{models}}} \sqrt{\sigma_H^2 \left(1 + \frac{1}{N \cdot N_{\text{models}}}\right)} \quad , \tag{6.17}$$

where σ_H^2 is the standard deviation of the hybrid models errors in predicting the antibody titer of the training experiments, $N \cdot N_{\text{models}}$ is the total number of predicted values, and $t_{1-\alpha/2, N \cdot N_{\text{models}}}$ is the critical value of Student's t distribution with $N \cdot N_{\text{models}}$ degrees of freedom calculated at the confidence level $\alpha = 0.05$.

The hybrid model is used to perform an *in silico* experimental campaign. It receives as input both the initial viable cell concentration and the culture volume, which are required to simulate the entire experimental run. Feeding is simulated by adjusting the appropriate value of the controlled input vector $\mathbf{u}$. In the *in silico* experimental campaign, we use a DoDE approach to define the feeding schedule. In particular, the hybrid model is controlled to follow the DoDE profiles, defined by the values of the dynamic subfactors, in the same way as in the optimization Strategy #2 (Section 6.2.2).

6.2.5 Feeding optimization

The optimal profile for glucose and glutamine is determined as the one maximizing the antibody titer at harvest through an optimization problem. Since the shape of the nutrient profiles is defined by the value of the dynamic subfactors according to Eq. (6.1), the optimization problem is formulated considering the DoDE dynamic subfactors x_i as:

$$\max_{x_i} \hat{y}(x_i) \quad , \tag{6.18}$$

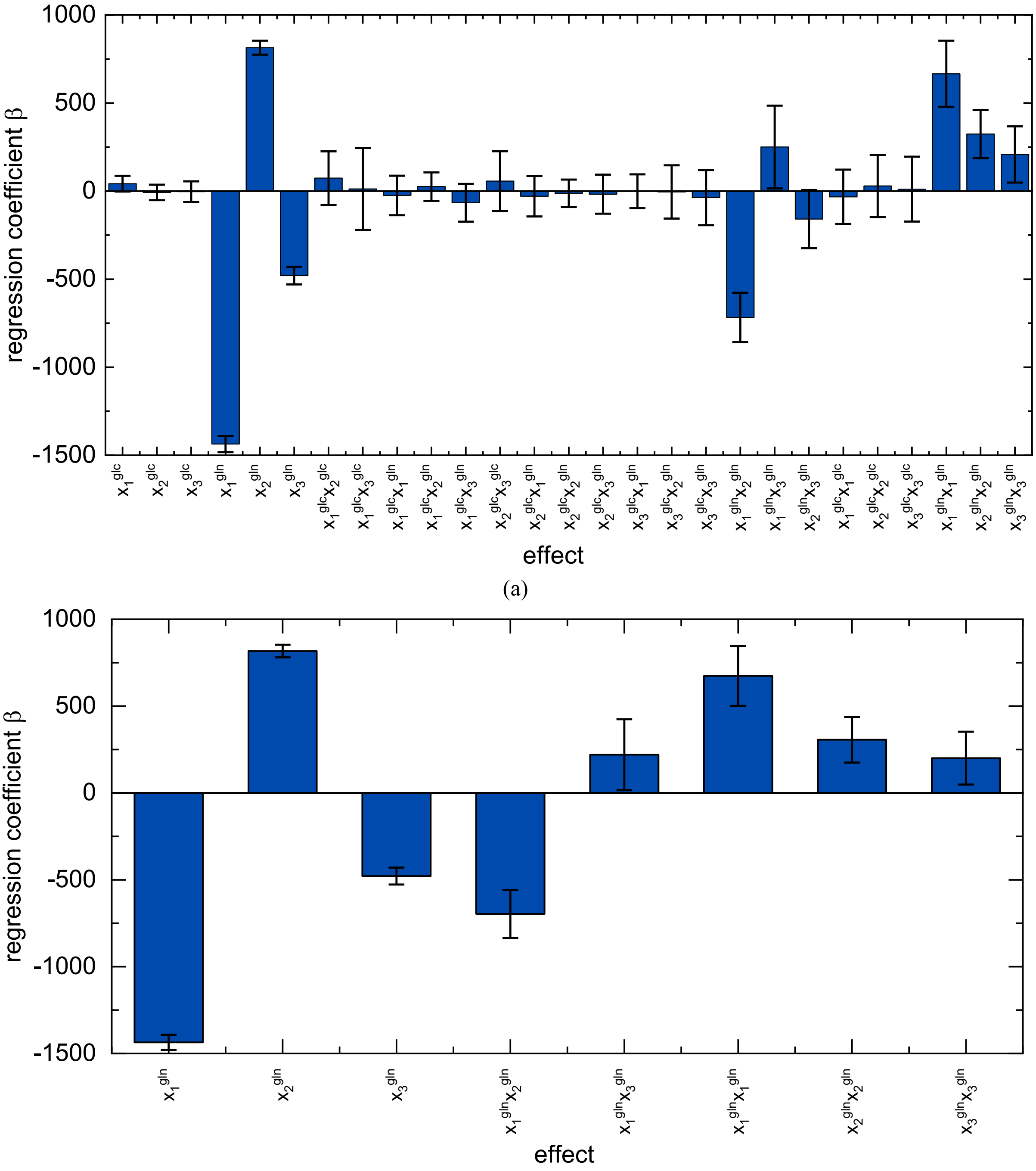
subject to the constraints of Eq. (6.2-6.4).

The antibody titer at harvest $\hat{y}(x_i)$ is predicted either by the RSM in optimization Strategy #2, or directly by the hybrid model in optimization Strategy #1. The optimization problem of Eq. (6.2-6.4, 6.18) is solved through a genetic algorithm (Sivanandam & Deepa, 2008) with a starting population of 200 individuals

All the simulations described in this work are performed in Matlab® 2020b, through the optimization toolbox and in house developed routines.

6.3 Results

The results of the optimization Strategy #2 on the experimental campaigns A and B, followed by the optimization Strategy #1 are presented here. These results are then compared with the process optimum, which is known because a simulated process is considered in this work.



(a)

(b)

Figure 6.4 *RSM regression coefficient with respective 95% confidence interval ($R^2 = 0.999$): (a) complete, and (b) updated. For a given nutrient j, x_i^j is the effect of the subfactor, $x_{i'}^j x_{i''}^j$ is the effect of the pairwise interaction between subfactors, and $x_i^j x_i^j$ is the effect of second-order interaction.*

6.3.1 Experimental campaign A for feeding schedule optimization

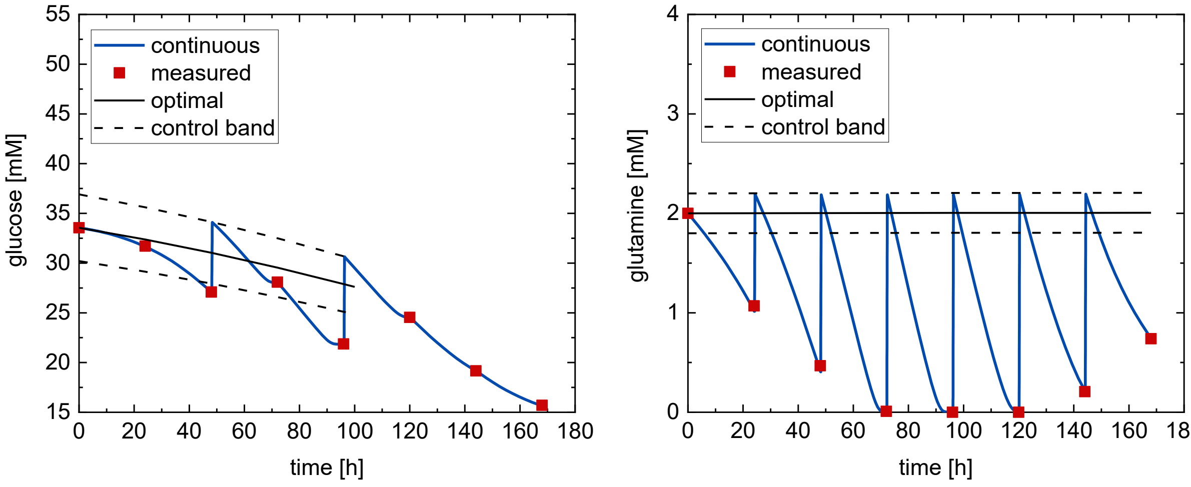
This section is aimed at identifying the optimal nutrient profile that maximizes the antibody titer at harvest through the application of DoDE by performing an experimental campaign directly on the process under study and state-of-the-art response surface modeling.

To this purpose, the experimental campaign A with 31 experiments planned through DoDE is used. The values of the dynamic subfactors $x_1^{\mathrm{glc}}, x_2^{\mathrm{glc}}, x_3^{\mathrm{glc}}$ and $x_1^{\mathrm{gln}}, x_2^{\mathrm{gln}}, x_3^{\mathrm{gln}}$ for glucose and glutamine, respectively, and the antibody titer at harvest obtained by experimental campaign A are used to fit a second-order RSM. The values of the dynamic subfactors affect the DoDE nutrient profiles as explained in Section 6.2.2.

The RSM shows a very high coefficient of determination $R^2 = 0.999$ (where the adjusted coefficient of determination is $R_{\mathrm{adj}}^2 = 0.999$), indicating the that the model provides an optimal fitting of the data. Figure 6.4a shows the RSM regression coefficient with their 95% confidence interval related to the uncertainty for all the dynamic subfactor x_i^j, their interactions $x_{i'}^{j'} x_{i''}^{j''}$, and second-order terms $x_{i'}^j x_{i'}^j$, where i' and i'' are the factor number and j, j', or $j'' = \mathrm{glc}, \mathrm{gln}$ is the nutrient. The effects showing high uncertainty (namely, the ones whose error bars cross 0 in Figure 6.4) are considered uninfluential for the model and excluded from the updated RSM. The updated RSM (Figure 6.4b) shows optimal fitting of the data as well, with $R^2 = 0.997$ ($R_{\mathrm{adj}}^2 = 0.997$).

Recalling that the subfactors define the shape of the nutrient profile, and specifically x_1 defines the position of the initial value, x_2 defines the increasing or decreasing tendency of the profile, and x_3 defines the concavity of the profile, the glucose profile results to have a very limited influence on the antibody concentration at harvest. In fact, the effects of all first-order and second-order glucose terms have confidence interval crossing zero and are not included in the updated RSM (Figure 6.4b). Furthermore, the profile of the two nutrients do not determine any interaction. In fact, the effects of all interactions between nutrients $x_{i'}^{\mathrm{glc}} x_{i''}^{\mathrm{gln}}$ are affected by large uncertainty and are not included in the updated RSM (Figure 6.4b). Differently, the glutamine profile has a large and strongly nonlinear effect on the antibody titer at harvest, since all glutamine first and second order terms are significant for the model and are included in the updated RSM (Figure 6.4b). Specifically, x_1^{gln} and x_3^{gln} show negative values, while x_2^{gln} and all second order terms show positive values. Accordingly, antibody titer is expected to be higher when the glutamine profile has a small initial value and shows an increasing tendency with a downward (negative) concavity. However, the initial glutamine value and shape of the profile are not independent and must be carefully tuned, since the effects of the interaction terms $x_1^{\mathrm{gln}} x_2^{\mathrm{gln}}$ and $x_1^{\mathrm{gln}} x_3^{\mathrm{gln}}$ are significant for the model and included in the updated RSM (Figure 6.4b). The negative effect of the interaction $x_1^{\mathrm{gln}} x_2^{\mathrm{gln}}$ means that low initial value of glutamine should be associated with a profile having an increasing tendency to induce and increase of the antibody titer at harvest, while the positive effect of the interaction $x_1^{\mathrm{gln}} x_3^{\mathrm{gln}}$ means that high initial value of glutamine should be associated with a positive (upward) concavity to increase the final titer.

According to these results, the antibody titer will not change much as response of different glucose profiles which can be arbitrarily set within the factor ranges. The glutamine profile, instead, is extremely important to achieve high antibody titer and must be carefully optimized.



(a) (b)

Figure 6.5 *Confirmatory experiment - optimal nutrient profile, determined from the DoDE experimental campaign A with 31 experiments, executed on the process: (a) glucose, and (b) glutamine. Red dots – process measurements; black line – optimal nutrient profile; black dashed line – control band; blue line – continuous measurement.*

The RSM is then used for process optimization to determine the nutrient profile providing the highest possible antibody titer at harvest. A genetic algorithm (Section 6.2.5) is adopted for the optimization. The optimal nutrient profiles (black line) and the profiles of the confirmatory experiment at the optimal conditions executed on the process (red points – process measurements) are shown in Figure 6.5. In the Figure the continuous measurement (blue line) of the nutrient profile is also reported. Considered that at shake flask scale this measurement is typically not available in real time; in this case is available since the process is simulated. In general: *i*) the optimal glucose profile (Figure 6.5a) starts at around half of its possible range (33.6 mM) and follows a decreasing profile with a very small downward concavity; *ii*) the optimal glutamine profile (Figure 6.5b) starts at the minimum value of its possible range (2 mM) and follow an almost constant profile for the entire culture. The optimal values of the glucose and glutamine subfactors are $\mathbf{x}_{\mathrm{A}}^{\mathrm{opt}} = [-0.439 \quad -0.385 \quad -0.04 \quad -0.999 \quad -0.0006 \quad -0.0002]$. As expected, the continuous measured profiles of glucose and glutamine (blue lines) cannot adhere perfectly to the respective optimal profiles, because nutrients are continuously consumed by cells, while nutrients are fed in boluses once a day as typically done at industrial level. Since the nutrients are fed only when the measured value (the red dot) goes below the control band (black dashed line), a sawtooth time profile of the variables is generated. However, this behavior is common to all experimental runs, and can be seen natural experimental variability. Furthermore, the lack of feeding in the final part of the batch does not produce negative effects on antibody titer because in this phase the viable cell concentration decreases and the available glucose, which is usually high, is sufficient to avoid cell starvation.

In experimental campaign A, with the abovementioned optimal nutrients profiles the RSM predicts a very high antibody titer at harvest of $\hat{y}_{\mathrm{A}} = 3530.0 \pm 54.6$ mg/L, while the confirmatory experiment at the optimal conditions executed on the process results in an antibody titer at harvest of $y_{\mathrm{A}} = 3118.2$ mg/L. The experimental antibody is outside the prediction interval and the RSM shows an error of 13.2%, meaning that it has a limited predictive capability, despite describing very well the calibration data ($R^2 = 0.997$). This is due to the highly nonlinear nature of the relationship between the subfactor values (i.e., the shape of the nutrient profiles) and the product titer at harvest, which cannot be captured by the second-order model.

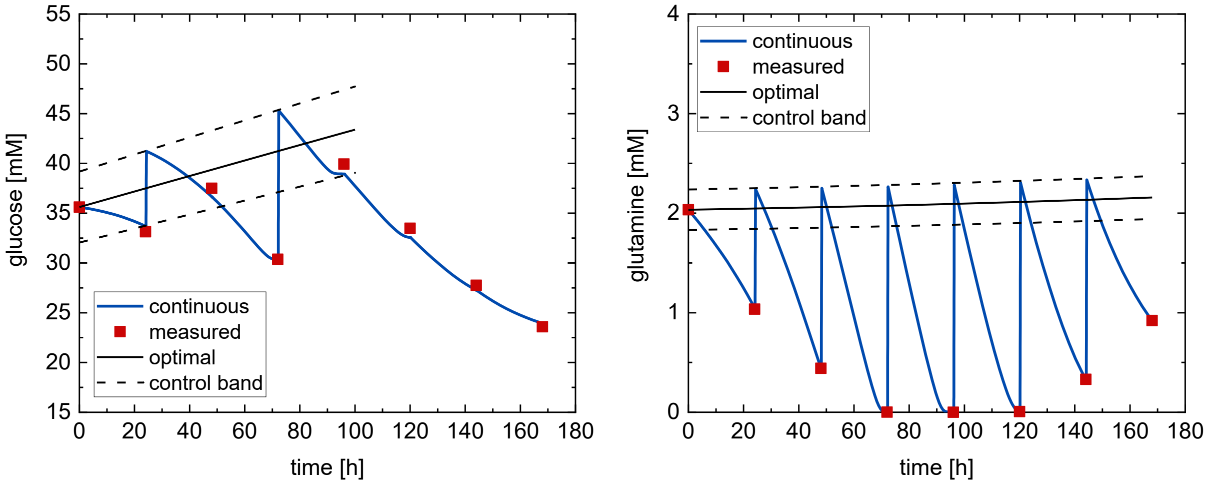
Based on this result, we will introduce additional methods, such as hybrid models, to describe the relationship between the nutrient profiles and the antibody titer at harvest.

6.3.2 Experimental campaign B for feeding schedule optimization

This Section is aimed at identifying the optimal nutrient profile that maximizes the antibody titer at harvest using only a limited set of experiments planned through the DoDE. This is intended to describe how the optimal nutrient profiles identified through the DoDE change when the number of performed experimental runs is low.

To this purpose, the experimental campaign B with 9 experiments planned through the DoDE is used. The values of the dynamic subfactors $x_1^{\mathrm{glc}}, x_2^{\mathrm{glc}}, x_3^{\mathrm{glc}}$ and $x_1^{\mathrm{gln}}, x_2^{\mathrm{gln}}, x_3^{\mathrm{gln}}$ for glucose and glutamine, respectively, and the antibody titer at harvest is then used to fit a first-order RSM. The RSM fitted on the process data shows a coefficient of determination $R^2 = 0.999$ ($R^2_{\mathrm{adj}} = 0.998$), indicating the that calibration data are well captured by the model. Similarly, the updated RSM describe the calibration data very well, with $R^2 = 0.996$ ($R^2_{\mathrm{adj}} = 0.994$), The model coefficients are similar to the linear terms shown in Figure 6.4a, hence, they are not shown for sake of conciseness. In this case, the initial glucose value results to have a small positive impact, indicating that only the initial glucose concentration slightly influences antibody titer, while the shape of the profile has no relevant effect. Glutamine, instead, shows a strong effect, having negative x_1^{gln} and x_3^{gln}, and positive x_2^{gln}. Accordingly, as previously observed, the antibody titer increases with glutamine profile having low initial value and an increasing profile with downward (negative) concavity.

The RSM is then used for process optimization to determine the nutrient profile giving the highest possible antibody titer at harvest by means of a genetic algorithm.



(a) (b)

Figure 6.6 *Confirmatory experiment - optimal nutrient profile, determined from the DoDE experimental campaign B with 9 experiments, executed on the process: (a) glucose, and (b) glutamine. Red dots – process measurements; black line – optimal nutrient profile; black dashed line – control band; blue line – continuous measurement.*

The resulting optimal nutrient profiles (black lines) and the profiles of the confirmatory experiment at the optimal conditions executed on the process (red points – process measurements) are shown in Figure 6.6, with the continuous measurement (blue lines). The optimal glucose profile (Figure 6.6a) starts at around half of its possible range and follow a linearly increasing profile with almost no concavity. The optimal glutamine profile (Figure 6.6b), instead, shows a constant profile along the culture at the minimum value of its possible range. The optimal values of the glucose and glutamine subfactors are $\mathbf{x}_{\mathrm{B}}^{\mathrm{opt}} = [-0.476 \quad -0.433 \quad -0.002 \quad -0.978 \quad 0.015 \quad 0.002]$.

Through the optimal nutrient profiles, the RSM predicts an antibody titer at harvest of $\hat{y}_{\mathrm{B}} = 3021.8 \pm 112.6$ mg/L, which is lower that the value predicted by the second-order RSM built in experimental campaign A on 31 experiments. The confirmation experiment executed on the process with the optimal feeding strategy shows an antibody titer at harvest of $y_{\mathrm{B}} = 3136.3$ mg/L. The experimental antibody is slightly outside the prediction interval and the RSM shows an error of 3.8%. In this case, the error between predicted and experimental value is lower than in the case of the second-order RSM (Section 6.3.1) built on a large number of experiments, indicating that the second-order model slightly overfit the calibration data providing worse prediction than a first-order one, which demonstrates better generalization capabilities. Despite the better prediction performance, the predicted value is still outside the prediction intervals, probably due to the highly nonlinear relationship between nutrient profiles and antibody titer, which cannot be captured by a first-order model.

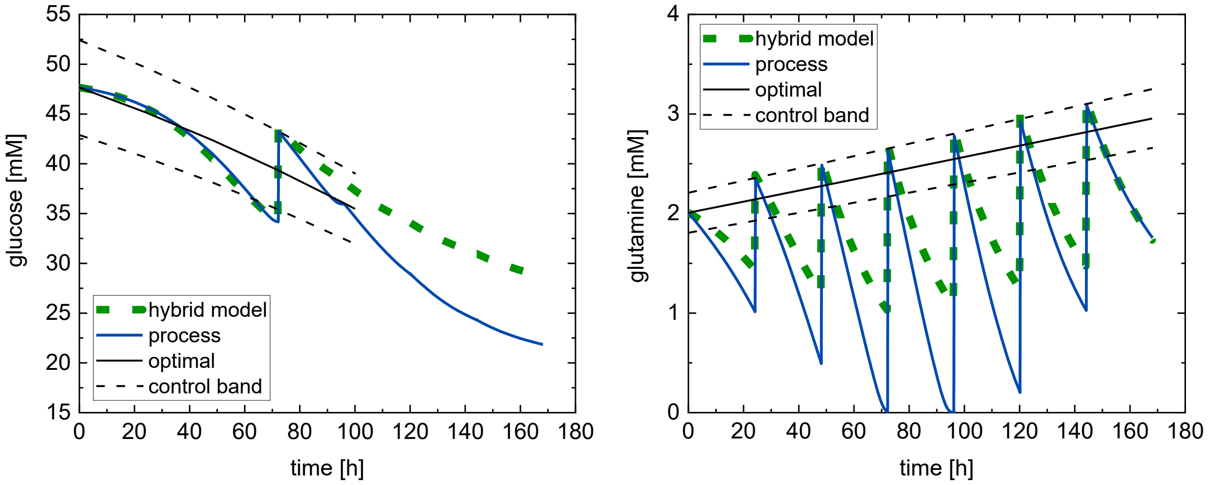
6.3.3 In silico cell culture optimization though hybrid model

This Section shows the optimization of the nutrient profiles by performing an *in silico* experimental campaign through a hybrid model. This will serve as proof of concept to understand the applicability and the advantage of conducting virtual experimental campaigns for the optimization of cell culture quality attributes through hybrid models.

To this purpose, a hybrid model (Section 6.2.4) is trained data collected during experimental campaign B planned through the DoDE (Section 6.3.2), which consists of 9 experiments. The hybrid model is exploited to perform an *in silico* experimental campaign following a DoDE strategy (Section 6.2.4). In the *in silico* experimental campaign a genetic algorithm guides the experiments to execute by suggesting the values of the dynamic subfactors defining the nutrient profiles.

The optimal nutrient profiles (black lines), the simulated ones though the hybrid model (green dashed lines), and the profiles of the confirmation experiment at the optimal conditions executed on the process (blue lines) are reported in Figure 6.7. The initial value of optimal glucose profile (Figure 6.7a) is close to the upper bound of the glucose range (47.7 mM) and follow a monotonically decreasing profile with slight downward concavity. Instead, the initial value of the optimal glutamine profile (Figure 6.7b) starts at the lower bound of its span range and follows an increasing profile with small slope and almost no concavity. The optimal values of the glucose and glutamine subfactors are $\mathbf{x}_{\mathrm{HM}}^{\mathrm{opt}} = [0.146 \quad -0.772 \quad -0.074 \quad -0.880 \quad 0.118 \quad 0.001]$. The glucose predicted by the hybrid model (Figure 6.7a, green dashed line) matches the simulated process profile before the first feeding action, while overestimates the process profile in the final part of the batch. This suggests that the addition of the glucose bolo drives the culture state to a region only partially explored by the training samples resulting in an underestimation of the glucose consumption rate and a reduction in the prediction performance. Instead, the overall glutamine profile (Figure 6.7b) is better predicted throughout the entire culture, showing however a slight underestimation of the glutamine consumption.

Through the optimal nutrient profiles, the hybrid model predicts a maximum antibody titer at harvest of $\hat{y}_1 = 2624.6 \pm 353.9$ mg/L, while a confirmatory experimental run at the optimal conditions performed on the process provided an antibody titer at harvest of $y_1 = 3222.8$ mg/L, which is outside the prediction intervals. The wide prediction interval is mainly due to the variability of the parameters estimated during the different hybrid model trainings, which derives from the typical drawback of the neural networks used in the hybrid model. According to these results, the hybrid model underpredicts the antibody titer by 18.6%, confirming that the hybrid model does not accurately predict the correct numerical value of the antibody titer. Despite that, the antibody titer predicted by the hybrid model is higher than the ones observed in the experiments used to train the hybrid model (experimental campaign B), indicating that the model correctly captures the correlation between nutrients and antibody titer and identifies the region of experimental domain with the highest antibody titer. However, it cannot accurately predict the antibody titer value because it is trained on data which are far from that region.



(a) (b)

Figure 6.7 *Confirmatory experiment - optimal nutrient profile, determined from the in silico experimental campaign through the hybrid model trained with the 9 experiments of experimental campaign B: (a) glucose, and (b) glutamine. Green dashed line – hybrid model simulation; black line – optimal nutrient profile; black dashed line – control band; blue line – process continuous measurement.*

6.3.4 Real optimal feeding schedule

In this Section, the real optimum of the process is presented to understand how well the investigated methodologies can identify the optimal feeding schedule for the cultures. The optimum of the process is known because a simulated process is considered, this information would not be available in a real scenario. The genetic algorithm (presented in Section 6.2.5) is applied directly on the process to find the optimal feeding conditions.

The optimal nutrient profiles of the process are reported in Figure 6.8 (black lines – target profile; blue line – continuous process measurement; red dots –measurements). The initial value of the glucose profile is in the middle if its possible range and monotonically grows with an upward concavity (Figure 6.8b). The initial value of the glutamine profile (Figure 6.8c) is at the lower bound of its range and it follows a slightly increasing profile with a small downward concavity.

The optimal nutrient profiles allow the process to achieve an antibody titer $y_{\mathrm{P}} = 3228.8$ mg/L.

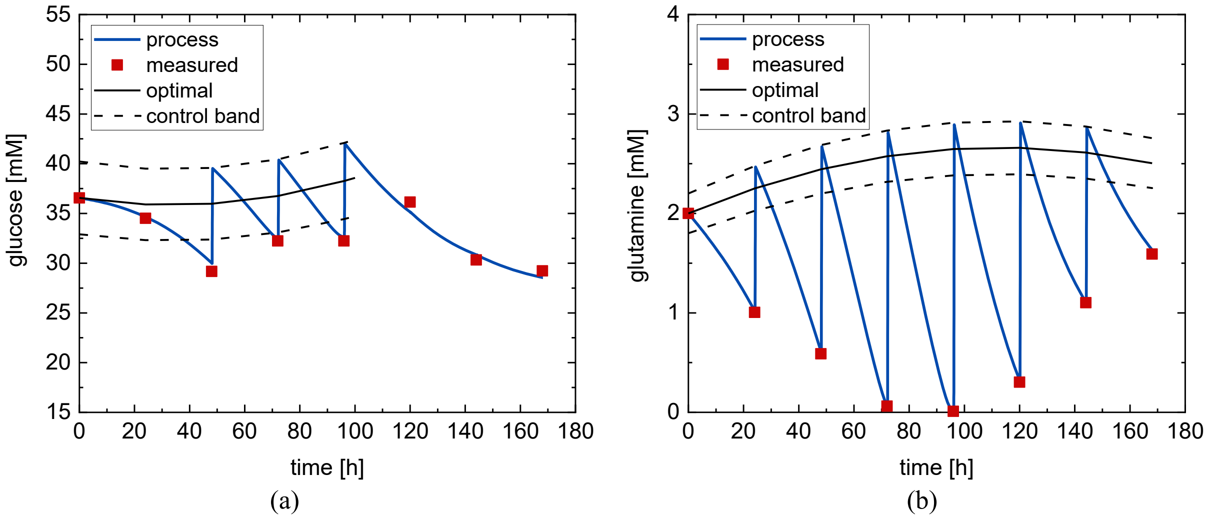


Figure 6.8 *Optimal nutrient profile of the process: (a) glucose, and (b) glutamine. Blue line – continuous process measurement; red dots – process measurements; black line – optimal profile; black dashed line – control band.*

6.4 Discussion

In this Section, the optimal feeding schedule of the process is compared with the ones obtained through the optimization Strategies #1 and #2. At the end, the antibody titer obtained in the confirmatory experiment at the optimal conditions are used to identify the best optimization strategy.

6.4.1 The optimal feeding schedule

The optimal feeding schedule of the process is characterized by an initial glucose concentration at approximately the average value in the range of possible concentrations, which allows a sustained cell growth in the initial part of the culture, and an increasing profile, which maintains high the cell growth even at high viable cell concentration. The low initial glutamine concentration provides enough nutrient for a sustained growth and at the same time determines a reduced formation of ammonia, which is detrimental because it limits cell growth and favors cell death. Furthermore, the downward concavity of the glutamine profile is coherent with the necessity of providing more glutamine when the viable cell concentration is higher (i.e., in the central part of the culture) while limiting ammonia formation at the same time. These results are coherent with previous studies (Teixeira et al., 2005), which recommended to limit the availability of glutamine in the initial growth phase and increase it later on in the culture. For what concern glucose, differently from our work previous studies recommended a low concentration of glucose along the entire culture, possibly decreasing it later in the culture (Teixeira, Alves, et al., 2007). Even if low glucose concentration is reasonable to limit lactate production, it should be highlighted that feeding enough glucose (i.e., as in our case) is of paramount importance to avoid cell starvation which negatively affects cell growth, productivity, and product quality (Fan et al., 2015; Narayanan et al., 2020).

6.4.2 Comparison of the optimal feeding schedule

In the optimization Strategy #2 (i.e., experimental campaigns planned through DoDE), the need of low level of glutamine (required to limit ammonia production) throughout the entire duration of the culture is identified in both experimental campaign A and B. However, the slight increasing amount required in the central part of the culture to compensate for the increased viable cell concentration is not identified by both approaches (i.e., experimental campaigns A, with 31 experiments, and B, with 9 experiments). Regarding the glucose, only in experimental campaign B (with 9 experiments) a profile similar to the process optimum is identified, showing no overfeeding at the beginning of the culture and an increasing glucose concentration at higher viable cell concentration. In experimental campaign A, instead, a glucose profile with high initial concentration and a decreasing profile is identified, that leads to a more sustained production of lactate especially in the initial part of the culture.

In optimization Strategy #1 (*in silico* experimental campaign), a correct behavior of the glutamine concentration, which starts at low level and increase along the culture, is identified. The optimal glucose profile instead has a high initial concentration and decreases along the culture, showing some similarity with experimental campaign A of the optimization Strategy #2.

These differences in the optimal glucose profiles are due to the small influence that glucose has on the antibody titer in the process. In fact, if glucose is not limiting, the growth rate (which also determines the productivity) is only controlled by the glutamine level and by the produced ammonia, which leads glucose to have only a minor role in cell productivity. For this reason, both modeling strategies capture the glucose behavior in a limited way. In particular, the second-order RSM is not affected by glucose and does not capture the relationship between glucose, lactate and reduced cell growth. Similarly, the hybrid model underestimates the impact that lactate has on cell growth. This leads both modeling strategies to suggest high levels of glucose at the beginning of the culture.

6.4.3 The best optimization strategy

The predicted antibody titer of the abovementioned strategies is compared to the one achieved in the optimal experiment executed on the process and results are summarized in Table 1. In the Table, the titer column reports the antibody titer achieved in the confirmatory experiment executed on the process at the optimal conditions identified by each strategy, while the predicted titer column reports the antibody titer predicted by the RSM (for experimental campaign A and B) and the hybrid model (for the *in silico* experimental campaign).

RSM and the planning of experiments through DoDE demonstrated to be applicable in mammalian cell cultures to optimize the feeding schedule, providing a simple and roust science-based strategy to improve the yield in antibody. In fact, in optimization Strategy #2, experimental campaign A (3118.2 mg/L) and B (3136.3 mg/L) both achieved an improved yield of antibody when the optimal feeding schedule is carried out on the process. In particular, experimental campaign B achieved with only 9 experiments a higher yield than experimental campaign A with 31 experiments, showing a 31.2% improved antibody titer with respect to the initial experiments executed during experimental campaign B. However, optimization Strategy #2 achieved antibody titer consistently lower than the real process optimum y_P (3228.8 mg/L). Despite the good yield improvement, the predictions of the antibody titer performed by the two RSMs of optimization Strategy #2 are inaccurate. The second-order RSM fitted on the 31 experiments from experimental campaign A shows a 13.2% prediction error, much greater than the 3.8% error shown by the first-order RSM trained on the 9 experiments of campaign B. These results suggest that the RSM does not completely capture the complex relationship between nutrients and product titer independently on model complexity. Furthermore, since the first-order RSM (used in experimental campaign B) allows the identification of a better feeding schedule than experimental campaign A, the use of a large number of samples (such as in experimental campaign A) is not always beneficial. For this reason, the planning of a large number of experiments must be coupled with models that effectively handle such information. However, the first-order RSM can only capture a linear relationship between nutrients and antibody titer. Furthermore, the robustness and generalizability of the models should be carefully tested though validation experiments in order to avoid overfitting issues.

Table 6.1 *Optimal nutrient profiles obtained with different strategies: subfactors value, simulated experimental antibody titer, predicted antibody titer and 95% confidence interval of the predicted antibody titer.*

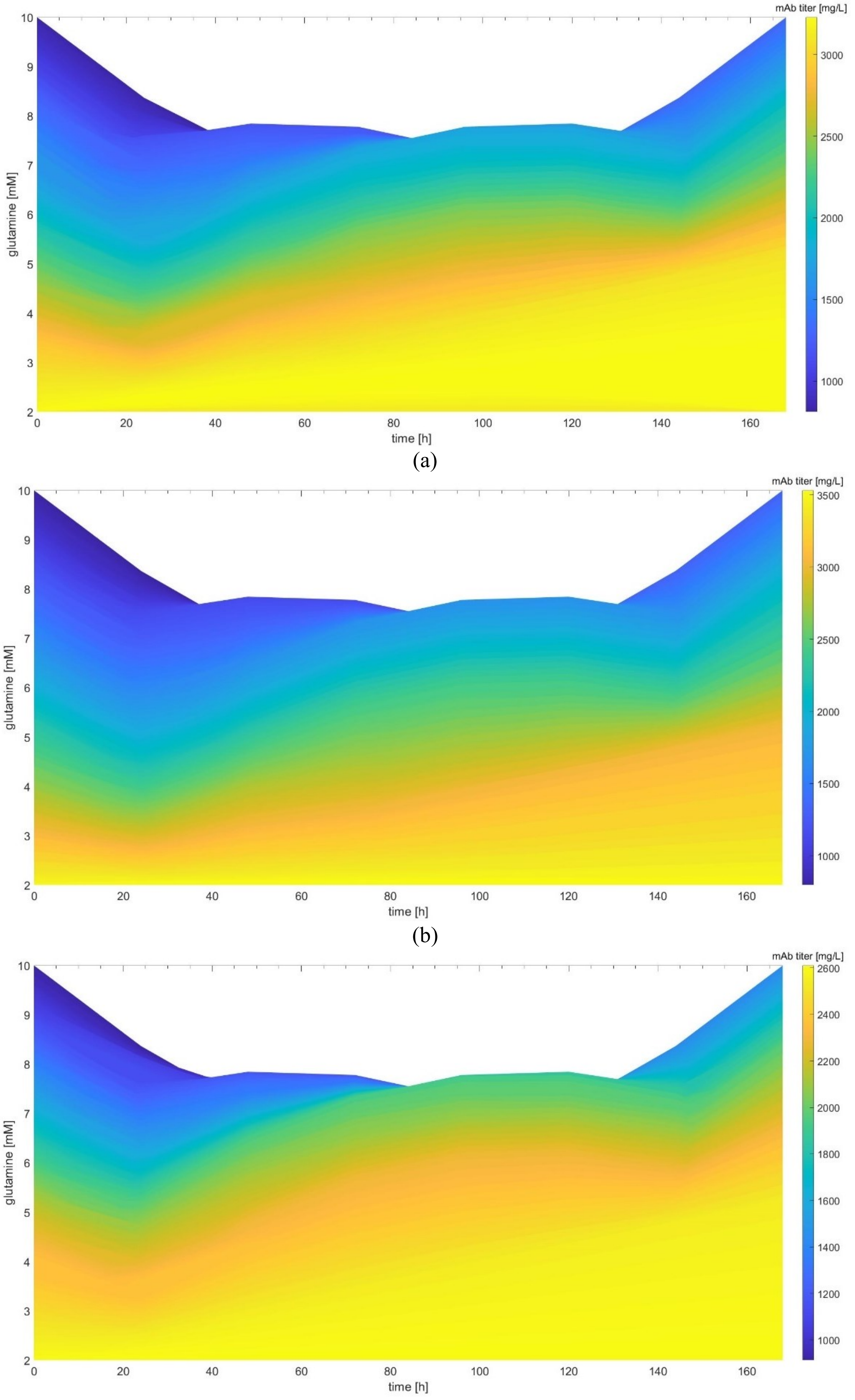
Strategy	Campaign	# of experiments	Titer [mg/L]	Predicted titer [mg/L]	CI [mg/L]
	process	-	3228.8		
1	*in silico*	9	3222.8	2624.6	353.9
2	A	31	3118.2	3530.0	54.6
2	B	9	3136.3	3021.8	112.6

Hybrid semi-parametric models are powerful tools which allow performing *in silico* experimental campaigns, since they provide a good representation of the system even when built on a reduced number of runs. In fact, the optimal feeding schedule identified by optimization Strategy #1 applied on the process achieved a very high antibody titer (3222.8 mg/L), which results very close to the real process optimum y_P (3228.8 mg/L). Optimization Strategy #1 improved the antibody titer by 34.9% with respect to the experiments of campaign B and provided a 2.8% increase in antibody titer with respect optimal antibody obtained through the experimental campaign B of optimization Strategy #2 (3222.8 mg/L vs. 3136.3 mg/L).

However, the antibody titer y_1 predicted in optimization Strategy #1 (2624.6 mg/L) is the lowest observed value, showing the largest prediction error (18.6%). This can be easily observed in the response surface of Figure 6.9, which shows the maximum antibody titer at harvest achieved with a certain glutamine concentration at a specific time, when the glucose subfactors are fixed at $[0.256 \quad 0.349 \quad 0.195]$, which are the real optimal glucose subfactor values. Note that the color scale in Figure 6.9a, b, and c is not the same to highlight differences in the shape of the profiles rather than the actual antibody titer value. The hybrid model in optimization Strategy #1 (Figure 6.9c) underpredicts the real antibody titer (Figure 6.9a) especially in the region at low glutamine and high antibody titer, while the RSM fitted in experimental campaign A (Figure 6.9b) overpredicts the antibody titer but shows a lower prediction error. Despite the lower prediction accuracy, the hybrid model better resembles the real relationship between the glutamine profiles and the antibody titer, than the RSM. In fact, the hybrid model identifies the positive effect that an increasing glutamine profile along the culture duration has on the antibody titer (compare the slope of the region with low glutamine after 100 hours in Figure 6.9). Furthermore, the hybrid model captures the effect that a downward concavity has on the antibody titer, predicting the right sharpe increase of titer in the region with glutamine ~5 mM and 50-120 hours, as shown in Figure 6.9. Differently, the RSM predicts a lower antibody titer when increasing the glutamine along the culture, and captures a lower positive effect of a curved glutamine profile. Accordingly, the hybrid model captures a relationship between nutrients and antibody titer which is more similar to the real one than the RSM, succeeding in the identification of region of experimental domain with highest antibody titer. However, since the region with low glutamine during the entire culture is not well represented in the calibration data, the hybrid model cannot accurately predict the antibody titer value. Despite that, these results indicate that hybrid models are powerful methods for *in silico* experimental campaign and can be used to virtually simulates the execution of experiments because of their underling mechanistic knowledge, but the predicted values are not completely representative because hybrid models are not always accurate in extrapolation, which is the typical drawback of data-driven models.

It is extremely important to point out that optimization Strategy #1 guarantees that the antibody titer obtained in the confirmatory experiment at the optimal conditions identified through the hybrid model is close to the one of the real optimal feeding schedule. Furthermore, the optimal feeding is identified by performing only 9 experiments (i.e., used to train the hybrid model) on the process. Accordingly, the hybrid model correctly learns and generalizes the relationship between nutrients and antibody titer and captures the cross-correlation between them, even if it is trained from a limited number of experiments. This is somehow expected because hybrid models combine the knowledge of the biological phenomena involved in cell cultures with the capability of learning complex relationship of the data-driven models.

It is also interesting to notice that, as previously proven, the selected hybrid model structure is the best in terms of number of samples required for training and extrapolation with different feeding scheduling (Narayanan, Luna, et al., 2021). In fact, the improvement of the model structure by introducing additional mechanistic knowledge, improves the description of the system, but requires a larger number of training samples to achieve comparable prediction performance. Accordingly, a tradeoff is required between model effectiveness and the model complexity (which requires a higher number of training samples).



(a)

(b)

(c)

Figure 6.9 *Maximum antibody titer at harvest with a certain value of glutamine in time: (a) achieved in the process, (b) predicted by the RSM fit in experimental campaign A, and (c) predicted by the hybrid model. The glucose subfactors are fixed at* $[0.256 \quad 0.349 \quad 0.195]$.

6.5 Conclusions

In this work we compared an *in silico* experimental campaign to optimize the feeding schedule of a mammalian cell culture with an experimental campaign on the process in such a way as to assess if the *in silico* experimentation can accelerate the process development and reduce the experimental burden. To conduct the *in silico* experimentation, we used a combination of Design of Dynamics Experiments (DoDE) and a hybrid semi-parametric model to identify the optimal time profile of glucose and glutamine profile in a virtual manner. The optimal nutrient profiles were compared with the ones obtained through two experimental campaigns planned with DoDE: experimental campaign A (31 experiments) and experimental campaign B (9 experiments).

The optimal antibody titer achieved through experimental campaign B (3136.3 mg/L) is 31.2% higher than the titer observed in the experiments executed during the experimental campaign. Experimental campaign A done with 31 experiments provided a lower optimal antibody titer (3118.2 mg/L) than experimental campaign B done with 9 experiments. Despite being able to improve the antibody titer, the experimental campaigns planned with DoDE could not achieve values similar to the real optimum of the process.

The *in silico* experimental campaign, which required only 9 experimental runs to train the hybrid digital model, provided a 34.9% overall improvement in the antibody titer with respect to the experiments used to train the model, a 2.8% improvement with respect to experimental campaign A and B, and reached a titer very close to the process optimum. The hybrid model accurately captures the relationship between nutrient profiles and antibody titer, but underpredicts the numerical value of the antibody titer. Accordingly, hybrid semi-parametric models are promising tools and can be used to conduct *in silico* experimental campaigns, providing high representation performance, and reducing the experimental burden and time required to perform the feeding schedule optimization in biopharmaceutical process development.

The testing of the proposed framework on a real process must be carried out to confirm our findings.

Chapter 7

Next-FLUX: Neural-net extracellular trained flux[*]

In this Chapter, a deep learning strategy for constraining genome-scale metabolic models from easily available and cheap data is proposed. A brief introduction to the available data and model implementation is first given. Then, the constraints prediction capability of the machine learning model in different configurations is presented. The results of the application of the predicted constraints on the genome-scale metabolic model will conclude the Chapter.

7.1. Materials and Methods

In this Section, the data and the main mathematical methodologies used in the development of the deep learning strategy for constraining GSMMs are presented. In particular, the available data, the structure of the deep learning model predicting the constraints, and the procedure to apply the predicted constraints on the GSMM are explained.

7.1.1 ^{13}C isotope labeling dataset

A ^{13}C isotope labelling dataset of CHO cells assembled from 8 different publicly available sources is used in this work. The dataset contains 31 different experiments (i.e., cell line) for which 3 sets of measurements are available, corresponding to the upper, lower, and median values of experimentally observed intracellular fluxes in central carbon metabolism. Specifically, each set comprises the measurements of 24 extracellular metabolite uptake rates, and 59 intracellular reaction rates, calculated from the measured ^{13}C isotope concentrations (Nomikos & MacGregor, 1995c). The available intracellular reactions, whose fluxes are available in the ^{13}C isotope labeling dataset, are reported in Appendix F. The experimental datasets span a wide variety of CHO cell types, phases, and processing conditions, which are reported in Table 7.1. In this work, the ^{13}C labeling fluxes and extracellular metabolite uptake rates are converted into the units of $mmol/g_{DCW}h$, through the dry cell weight experimentally measured for each cell line, to comply with the typical fluxomics unit of measure.

[*] This work is a collaboration with Imperial College London (UK). Please refer to the foreword for the complete disclosure statement.

The resulting dataset, comprising 93 experimental datapoints, is organized in the matrix $\mathbf{X}\,[N \times V] = [93 \times 24]$, which contains the extracellular metabolite uptake rates, and $\mathbf{Y}\,[93 \times 59]$, which contains the intracellular reaction rates. In the rest of this Chapter, the intracellular reaction rates will be referred to as intracellular fluxes.

Table 7.1 *^{13}C isotope labeling data sources and data conditions.*

Source	Cell type	# of experiments	Culture phase	Names
Templeton et al. (2013)	CHO-S producer	4	Early exponential, late exponential, stationary, and decline	early, late, stat, decline
Sheikholeslami et al. (2013)	CHO-Cum2 producer	2	Late exponential	ind, nonind
Templeton et al. (2014)	CHO-S non-producer	6	Early and late exponential	CLP, LELP, HELP, CLC, LELC, HELC
Sheikholeslami et al. (2014)	CHO-Cum2 producer	2	Late exponential	lowgln, highgln
Nicolae et al. (2014)	CHO-K1 non-producer	1	Late exponential	nicolae
Templeton et al. (2017a)	CHO-S producer and non-producer	11	Stationary	SV, SVGS, SVM1, SVM2, SVM3, SVM4, BCL2, BCL2-M1, BCL2-M2, BCL2-M3, BCL2-M4
Templeton et al. (2017b)	CHO-S producer	2	Perfusion and stationary	fed-batch, perfusion
McAtee Pereira et al. (2018)	CHO-S producer	3	Late exponential	CM, LA, LAplus

7.1.2 Prediction of intracellular fluxes from extracellular metabolite uptake rates through artificial neural networks

7.1.2.1 Data management

The available ^{13}C isotope labeling data are incomplete, since certain intracellular fluxes are not available in some data sources. Because of that, only the intracellular reactions with at least 75% of measurements are considered in this work. Accordingly, the matrix of intracellular fluxes considered in the rest of this Chapter is $\mathbf{Y}\,[N \times M] = [93 \times 47]$. The remaining missing data are not artificially imputed to avoid the introduction of additional artifacts in the data. Accordingly, only the available measurements of each intracellular reaction are considered to construct the artificial neural networks (ANN) model.

To mitigate overfitting issues and improve the robustness of the ANN parameters due to the reduced number of available datapoints, the SMOTE method and gaussian noise addition are used to artificially increase the number of training observations. In SMOTE (Chawla et al., 2002) a new artificial datapoint is generated as linear combination of two experimental datapoints. For each experimental datapoint $\mathbf{x}_n$, one among the 10 nearest neighbor experimental points $\mathbf{x}_{n'}$ is randomly selected and used to generate the artificial datapoint $\mathbf{x}_{\mathrm{art}}$. Then, the two selected experimental datapoints $\mathbf{x}_n$ and $\mathbf{x}_{n'}$ are linearly combined through a random coefficient, γ, as

$$\mathbf{x}_{\mathrm{art}} = \mathbf{x}_n + \gamma(\mathbf{x}_{n'} - \mathbf{x}_n) \quad . \tag{7.1}$$

Both extracellular metabolite uptake rates **X** and intracellular fluxes **Y** were augmented through Equation (7.1) using the same γ value and selected neighbor. Through this procedure, three artificial datapoints are generated from each experimental datapoint in **X** and **Y** by randomly selecting a different neighbor and γ value, resulting in $3 \cdot N$ artificial datapoints. Once generated, all the artificial datapoints $\mathbf{x}_{\mathrm{art}}$ are vertically concatenated to the original datasets **X** and **Y** to form the augmented matrices $\mathbf{X}^{\mathrm{SMOTE}}$ and $\mathbf{Y}^{\mathrm{SMOTE}}$.

Gaussian noise addition is used to further increase the number of training experiments and improve the robustness of the ANN model. Five new artificial datapoints are generated for each datapoint in $\mathbf{X}^{\mathrm{SMOTE}}$ and $\mathbf{Y}^{\mathrm{SMOTE}}$, adding 3% white noise to $\mathbf{X}^{\mathrm{SMOTE}}$ and 1% white noise to $\mathbf{Y}^{\mathrm{SMOTE}}$. This noise is selected as a tradeoff between a low distortion of the original values and improved prediction performance of the ANN evaluated in preliminary studies (not shown here for the sake of conciseness). Once generated the new artificial datapoints are vertically concatenated to $\mathbf{X}^{\mathrm{SMOTE}}$ and $\mathbf{Y}^{\mathrm{SMOTE}}$, to produce the final matrices for ANN training $\mathbf{X}^{\mathrm{A}}$ and $\mathbf{Y}^{\mathrm{A}}$.

7.1.2.2 Data-driven modeling strategy

Artificial neural networks (Section 2.2) are used to predict intracellular fluxes from extracellular metabolite uptake rates. The ANN is selected since it outperformed other regression methodologies in preliminary studies (not shown here for sake of conciseness). One ANN is built for each intracellular reaction m, to predict its reaction rate $\mathbf{y}_m$ (a vertical slice of **Y**) from all extracellular metabolite uptake rates **X**. Prior to ANN modeling, all data are scaled to 0 mean and unit variance by autoscaling.

The structure of the ANN comprised one input, one output layers, and two fully connected hidden layers. The activation function is selected for each intracellular reaction between *reLu* and *tanh* as the one maximizing validation accuracy in preliminary studies. The ANN is trained with an Adam algorithm (Kingma & Ba, 2015) using the mean squared error (MSE) as loss function. The Adam algorithm is nowadays one of the most effective algorithms for the training of ANN.

Because of the relatively small number of available datapoints, instead of splitting the data in training and validation subsets, a 15-fold cross-validation is coupled with a grid search to determine the optimal number of neurons for each hidden layer and the learning rate. The hyperparameters are selected as the ones minimizing the MSE in cross-validation. Similarly, the number of training epochs is determined through a 100-iteration Monte Carlo cross-validation as the average of the training epochs observed over the cross-validation iterations. In this case, 5% of the datapoints are randomly extracted and used as internal validation set, while the remaining 95% are used from model training. This internal validation set is used to stop the ANN training when the validation MSE starts to increase, thus determining the best number of training epochs. In all cross-validation operations, the measurements corresponding to the same experiment and all the augmented data generated from them are place together either in the training or the internal validation set.

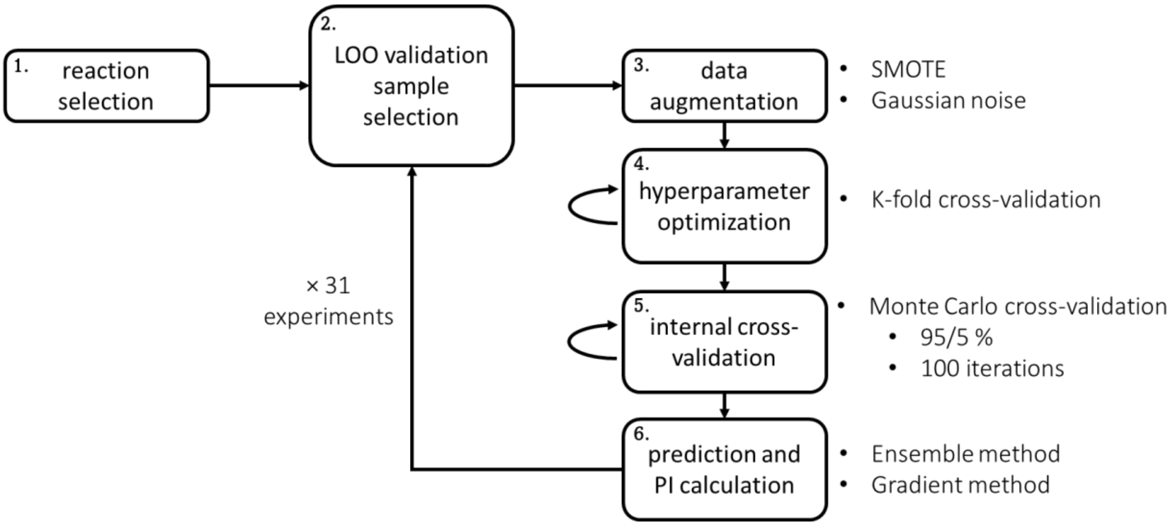

1. reaction selection
2. LOO validation sample selection
3. data augmentation
4. hyperparameter optimization
5. internal cross-validation
6. prediction and PI calculation
× 31 experiments
• SMOTE
• Gaussian noise
• K-fold cross-validation
• Monte Carlo cross-validation
• 95/5 %
• 100 iterations
• Ensemble method
• Gradient method

Figure 7.1 *Neural networks leave-one-out validation procedure.*

The ANN models were validated through a leave-one-out procedure to assess in the best possible way the prediction performance given the small number of available experimental measurements. The leave-one-out validation procedure is organized as follows (Figure 7.1):

1. Reaction selection: initially, the intracellular reaction m to consider is selected and its reaction rates $\mathbf{y}_m$ $[93 \times 1]$ are extracted from $\mathbf{Y}$. All the following steps are performed separately for each intracellular reaction.
2. Leave-one-out validation sample selection: at this step the leave-one-out validation experiment n is selected, and its three sets of measurements are extracted from $\mathbf{y}_m$ and $\mathbf{X}$, generating the training dataset $\mathbf{X}_{\text{train}}$ $[90 \times 24]$ and $\mathbf{y}_{\text{train}}$ $[90 \times 1]$, and the validation one $\mathbf{X}_{\text{val}}$ $[3 \times 24]$ and $\mathbf{y}_{\text{val}}$ $[3 \times 1]$. The validation dataset $\mathbf{X}_{\text{val}}$ and $\mathbf{y}_{\text{val}}$ are kept aside until the validation step (step 6). All the following steps are repeated for all the available experiments.
3. Data augmentation: once the leave-one-out validation sample is extracted, data augmentation (Section 7.1.2.1) is performed on the training experiments $\mathbf{X}_{\text{train}}$ and $\mathbf{y}_{\text{train}}$ to generate the training matrices $\mathbf{X}_{\text{train}}^{\text{A}}$ $[1350 \times 24]$ and $\mathbf{y}_{\text{train}}^{\text{A}}$ $[1350 \times 1]$.
4. Hyperparameter optimization: the learning rate and the number of neurons of the two hidden layers are determined through 15-fold cross-validation as previously explained.

5. Internal cross-validation: at this step, the training number of epochs are determined through a Monte Carlo cross-validation as previously explained.
6. Reaction rate prediction and prediction interval calculation: at this step, the ANN is trained on augmented training experiments $\mathbf{X}_{\text{train}}^{\text{A}}$ and $\mathbf{y}_{\text{train}}^{\text{A}}$. Then, the intracellular fluxes of the left-out experiment are predicted ($\hat{\mathbf{y}}_{\text{val}}$) from $\mathbf{X}_{\text{val}}$ and the prediction interval is calculated with two different methods: the ensemble method (explained in Section 7.1.2.3) and the gradient method (explained in Section 7.1.2.4). After this step, the procedure is concluded, and another leave-one-out validation sample should be selected at step 2.

The prediction performance of the model is evaluated through the MSE and the coefficient of determination Q^2 in validation, which is calculated at the end of the leave-one-out validation joining all predicted values $\hat{\mathbf{y}}_{\text{val}}$ together to have a comprehensive metric for the effectiveness of the model.

In this work, we proposed an innovative strategy to constrain the GSMMs. Specifically, we used the prediction intervals of intracellular fluxes (whose calculation methods are available in the Literature and reported in Section 7.1.2.3 and Section 7.1.2.4) estimated through the ANN to constrain intracellular reactions in GSMMs.

7.1.2.3 Ensemble prediction interval calculation method

The ensemble method is one of the most straightforward methods for the calculation of the prediction intervals. In the ensemble method (Lakshminarayanan et al., 2017) the ANN model is trained multiple times in parallel with different randomly initialized weights on randomly shuffled training data $\mathbf{X}_{\text{train}}^{\text{A}}$. The model ensemble is used for prediction, providing a population of predicted values $\hat{\mathbf{y}}_{\text{E}}$, whose differences is only due to the uncertainty in network weight estimation.

The ensemble predicted value is defined as the mean value of the population of predictions $\hat{\mathbf{y}}_{\text{E}}$. The uncertainty prediction interval is calculated as the 95% confidence interval of a Student's t distribution build on the population $\hat{\mathbf{y}}_{\text{E}}$:

$$\text{PI} = \sigma_{\hat{\mathbf{y}}_{\text{E}}}\, t_{1-\alpha/2, N_{\text{models}}-1} \quad , \tag{7.2}$$

where PI is the half width prediction interval, $\sigma_{\hat{\mathbf{y}}_{\text{E}}}$ is the standard deviation of the population of ensemble predictions $\hat{\mathbf{y}}_{\text{E}}$, and $t_{1-\alpha/2, N_{\text{models}}-1}$ identifies the confidence threshold of a t-distribution with $N_{\text{models}} - 1$ degrees of freedom calculated with $\alpha = 0.05$, and N_{models} is the total number of trained models. In this work, $N_{\text{models}} = 25$ models are trained in parallel for the estimation of the prediction intervals.

7.1.2.4 Gradient prediction interval calculation method

The gradient method directly exploits the gradient through the ANN calculated with backpropagation to estimate the prediction intervals. In this case, the prediction interval for a new datapoint $\mathbf{x}_{\mathrm{NEW}}$ is calculated exploiting the trained ANN (De Veaux et al., 1998; Khosravi et al., 2011) as:

$$\mathrm{PI} = s_e\, t_{1-\alpha/2,N-p^*} \sqrt{1 + \mathbf{g}_{\mathrm{NEW}}^{\mathrm{T}}\, \boldsymbol{\Phi}(\mathbf{J}^{\mathrm{T}}\mathbf{J})\boldsymbol{\Phi}\mathbf{g}_{\mathrm{NEW}}}\,, \tag{7.3}$$

where $t_{1-\alpha/2,\mathrm{N}-p^*}$ identifies the confidence threshold of a t-distribution with $N - p^*$ degrees of freedom calculated with $\alpha = 0.05$, N is the total number of training experiments, p^* is defined in the following, $\mathbf{g}_{\mathrm{NEW}} = \partial L(\mathbf{x}_{\mathrm{NEW}})/\partial\boldsymbol{\omega}$ is the gradient vector of the loss function calculated in the new datapoint, $L(\mathbf{x}_{\mathrm{NEW}})$, over the ANN weights $\boldsymbol{\omega}$, $\mathbf{J} = \partial L(\mathbf{X}_{\mathrm{train}}^{\mathrm{A}})/\partial\boldsymbol{\omega}$ is the matrix of the gradient vectors of the loss function calculated for all training experiments $\mathbf{X}_{\mathrm{train}}^{\mathrm{A}}$ over the network weights, and T indicates the transpose operation. The matrix $\boldsymbol{\Phi}$ is defined as:

$$\boldsymbol{\Phi} = \left(\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda_{\mathrm{reg}}\mathbf{I}\right)^{-1} \quad , \tag{7.4}$$

where λ_{reg} is a regularization coefficient and $\mathbf{I}$ is the identity matrix. In this work, a regularization coefficient $\lambda_{\mathrm{reg}} = 0.05$ is used. The s_e factor is calculated as:

$$s_e = \frac{\left(\mathbf{y}_{\mathrm{train}}^{\mathrm{A}} - \hat{\mathbf{y}}_{\mathrm{train}}^{\mathrm{A}}\right)^{\mathrm{T}}\left(\mathbf{y}_{\mathrm{train}}^{\mathrm{A}} - \hat{\mathbf{y}}_{\mathrm{train}}^{\mathrm{A}}\right)}{N-p^*} \quad , \tag{7.5}$$

where $\hat{\mathbf{y}}_{\mathrm{train}}^{\mathrm{A}}$ are the ANN predicted values of the training dataset, $p^* = \mathrm{tr}(2\boldsymbol{\Psi} - \boldsymbol{\Psi}^2)$, where tr defines the trace, and the matrix $\boldsymbol{\Psi}$ is calculated as:

$$\boldsymbol{\Psi} = \mathbf{J}\boldsymbol{\Phi}\mathbf{J}^{\mathrm{T}} \quad . \tag{7.6}$$

7.1.3 Genome-scale Metabolic model[2]

The proposed methodology was tested on a comprehensive CHO cell GSMM, iCHO2441, that couples the protein secretory pathway of iCHO2048 (Gutierrez et al., 2020) to the recently updated iCHO2291 (Yeo et al., 2020). To our knowledge, this is the most complete CHO GSMM to date, with the highest number of annotated genes and gene per reaction ratio. The GSMM comprises $D = 4174$ metabolites, $U = 6337$ metabolic reactions, and 2441 genes.

The constraining of the GSMM with the bounds predicted by the ANN is not straightforward and comprises four steps:

1. Constraining of the GSMM with the extracellular uptake rates form the ^{13}C labelling experiments (Section 7.1.1);
2. Reaction mapping;
3. Maximization of the feasible constraints;

[2] The GSMM and the methodology to apply the ANN predicted constrains on the GSMM has been developed by the research group of Prof. Cleo Kontoravdi at Imperial College London (U.K.).

4. GSMM solution.

7.1.3.1 Reaction mapping

The intracellular reactions predicted by the ANN, because available from the ^{13}C labelling data, lump several GSMM reactions, making a one-to-one mapping impossible. For this reason, the reactions predicted by the ANN must be mapped to the ones in the GSMM. This is achieved by considering the GSMM reactions as an electric circuit, with parallel and serial connections. The fluxes of parallel reactions were summed and subsequently treated as single serial reaction. For serial reactions, the overall flux is determined as the minimum of all reaction in series.

7.1.3.2 Maximization of the feasible constraints

GSMM reactions are constrained with the bounds predicted by the ANN (i.e., prediction intervals) following the mapping determined in Section 7.1.3.1. However, due to over constraining issues, the constraining of certain combinations of mapped reactions is not feasible. To ensure model feasibility, the optimum set of reaction to constraint is found through a Mixed Integer Linear Programming problem before the actual constraining of the GSMM. The GSMM is iteratively solved with the objective of finding the feasible combination of intracellular flux constraints that maximizes the number of included ANN predicted bounds. The optimization problem is defined as:

$$\max \textstyle\sum_{m=1}^{M} \delta_m \quad , \tag{7.7}$$

subject to the solution of the GSMM through FBA (Section 2.3.1), the ANN predicted constraints, and the extracellular metabolites uptakes rates form ^{13}C labeling data; where $\delta_m = 1$ if the m-th reaction is constrained with the ANN predicted bounds, and $\delta_m = 0$ if the m-th reaction is not constrained with the ANN predicted bounds.

7.1.3.3 GSMM solution

The GSMM is solved though flux sampling. In flux sampling, intracellular fluxes are calculated by averaging 5 million possible flux solutions within the GSMM solution space (Section 2.3.2). The prediction of cell phenotypes, such as biomass, are performed by maximizing biomass through FBA.

7.1.3.4 Performance evaluation

The performance of the GSMM constrained with the ANN predicted bounds are compared with the base case, which is typical state-of-the-art method for GSMM solution. In the base case, the GSMM is only constrained with the extracellular uptake rates form the ^{13}C labelling experiments, while the constraints of intracellular reactions are left as the default ones (0 – 1000 or -1000 – 1000 according to reaction reversibility).

Intracellular flux estimation performance is evaluated through the Pearson correlation coefficient between the ^{13}C labelling experimentally measured and GSMM calculated intracellular fluxes. Note that the same mapping presented in Section 7.1.3.1 is used to associate ^{13}C labeling and GSMM reactions. The Pearson correlation is calculated for each available intracellular reaction in the ^{13}C labelling experiments over all available experiments as:

$$\rho_m = \frac{\sum_n (\mathrm{v}_{n,m} - \bar{\mathrm{v}}_m)(y_{n,m} - \bar{y}_m)}{\sqrt{\sum_n (\mathrm{v}_{n,m} - \bar{\mathrm{v}}_m)^2 \sum_n (y_{n,m} - \bar{y}_m)^2}} \quad , \tag{7.8}$$

where $\mathrm{v}_{n,m}$ is the m-th calculated intracellular flux for the n-th experiment through the GSMM, $\bar{\mathrm{v}}_m$ is the average value of the m-th intracellular flux calculated through the GSMM, $y_{n,m}$ is the m-th ^{13}C intracellular flux for the n-th experiment from $\mathbf{y}_m$, and $\bar{y}_m$ is the average value of the m-th intracellular flux $\mathbf{y}_m$. The Pearson correlation coefficient was analyzed based on its distribution over all available intracellular reactions M.

All the codes used to obtain the results of this Chapter are developed in Python 3.10, using Tensorflow 2.8 (Abadi et al., 2016) and COBRApy (Ebrahim et al., 2013).

7.4 Prediction of intracellular fluxes

In this Section, the intracellular flux prediction performance of the developed ANN is presented for the proposed methods for the calculation of the prediction interval and for different dimensions of the grid search for optimal number of neurons.

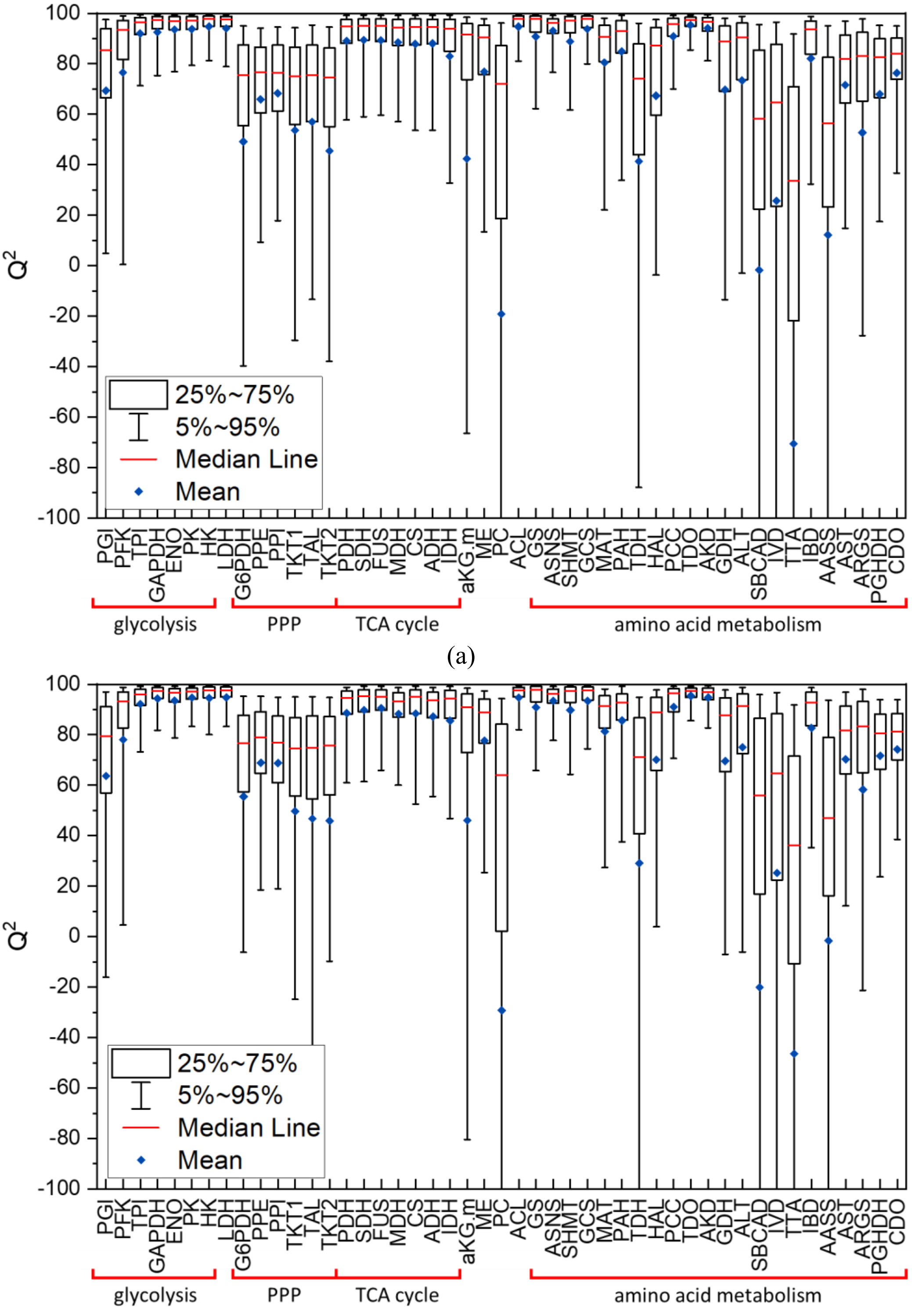
7.2.1 Performance of different prediction interval calculation methods

This Section is aimed at assessing the differences in the ANN prediction performance between the two prediction interval calculation methods tested in this study. This examines how different methods predict the actual value of intracellular fluxes and the respective prediction intervals (i.e., which will be further used to constrain the GSMM). To this purpose the Q^2 in the internal cross-validation and in the leave-one-out validation is inspected, because it gives a better idea of the overall prediction performance of the models than absolute or relative metrics (e.g., root mean squared error, mean absolute error, percentage error, etc.).

The distribution of Q^2 in cross-validation is reported in Figure 7.2 through a box plot because a single value, such as the mean, would not give complete information on the model behavior. Note that in this box plot, the percentiles of the model performance in cross-validation are reported, these are not the prediction intervals used in the following Section for GSMM constraining. The cross-validated Q^2 quantifies the robustness of the model to different splitting of the dataset. Specifically, a low Q^2 median value[3] indicates that the predictions are generally

[3] In this case, the median value is more meaningful than the average, because it gives indication of the trends in the prediction performance without being affected by outliers (i.e., low Q^2 values, which are unbounded in the negative direction).

inaccurate independent on the calibration-validation splitting, while high Q^2 median value indicates accurate prediction performance; furthermore, a wide Q^2 distribution means that the prediction performance are strongly dependent on the specific internal calibration-validation splitting, while a small dispersion Q^2 distribution means that the prediction performance are repeatable.



(a)

(b)

Figure 7.2 *Neural network* Q^2 *in cross-validation for different prediction intervals calculation methods: (a) ensemble and (b) gradient.*

Very similar trends in Q^2 can be observed for both prediction interval calculation methods. The main metabolic reactions involved in energy production, belonging to glycolysis and TCA cycle, have a very high median Q^2 with very narrow distribution, apart for reactions *PGI* and *PFK*, which are located in proximity of the bifurcation between glycolysis and pentose phosphate pathway (PPP). This indicates that the prediction of the flux through these energy producing reactions is very robust, and accurate predictions are always performed independently on the calibration-validation splitting. Differently, PPP reactions, which are parallel to glycolysis and involved in the production of reducing agents, and nucleotide and aromatic amino acid precursors, show a median Q^2 slightly above 70% with a wider distribution even reaching negative values. Accordingly, prediction of PPP fluxes, which alternative way to metabolize glucose, is less robust and slightly less accurate than the primary route of glucose, glycolysis and TCA cycle, but a satisfactory accuracy is still achieved. Furthermore, PPP flux predictions are more sensitive to the specific internal calibration-validation splitting, and their performance are likely to be experiment dependent. Differently, the metabolism of amino acids shows very variable performance. Reactions, such as *SBCAD*, *IVD*, *TTA* and *AASS*, which involve metabolites participating in many metabolic reactions, such as *AcCoA*, *aKG.c*, and *Glu.c*, show low median Q^2 and very wide distributions, whereas other reactions involving the conversion of amino acids into other metabolites, such as *GS*, *ASNS*, *GCS*, *PCC*, *TDO* and *AKD*, show high median Q^2 and narrow distribution. Other reactions concerning the amino acid metabolism show intermediate, but still satisfactory performance, with an average influence of the specific internal calibration-validation set.

With respect to the Q^2 in cross-validation, the only difference between the two prediction interval calculation methods relies on slight variations in the median, average and percentile values of the predicted fluxes. Only PPP reactions, *G6PDH*, *TAL* and *TKT2*, and the glycolysis reaction in proximity of the PPP bifurcation, *PGI*, show a large variation in the 5/95 percentile values, indicating a different sensitivity of the predictions to the specific internal calibration-validation set.

The Q^2 in leave-one-out validation for the entire available metabolic network is reported in Figure 7.3 and 7.4. The Q^2 in validation gives information on the overall performance of the ANN in predicting new and totally unknown samples, providing the best way to evaluate the ANN. Also in this case, the Q^2 is used, because other absolute and relative metrics would be more difficult to interpret in a comprehensive fashion.

Very similar trends in the Q^2 in leave-one-out validation can be observed for both prediction interval calculation methods. The main carbon route for energy production, involving glycolysis and TCA cycle, is predicted with high accuracy showing a Q^2 near 90%, apart for the reactions located in proximity of the bifurcation between glycolysis and PPP, *PGI* and *PFK*. Fluxes of the parallel route for metabolizing glucose, PPP, are predicted with lower accuracy than the primary glucose route, glycolysis and TCA cycle, whereas amino acid metabolism show very variable performance coherent with was previously observed. In this case, fluxes of some reactions in different regions of the metabolism, such as *PC*, *ARGS*, *TTA*, *PAH*, and *AASS* are predicted with low accuracy showing $Q^2 < 30\%$ in both prediction interval calculation methods.

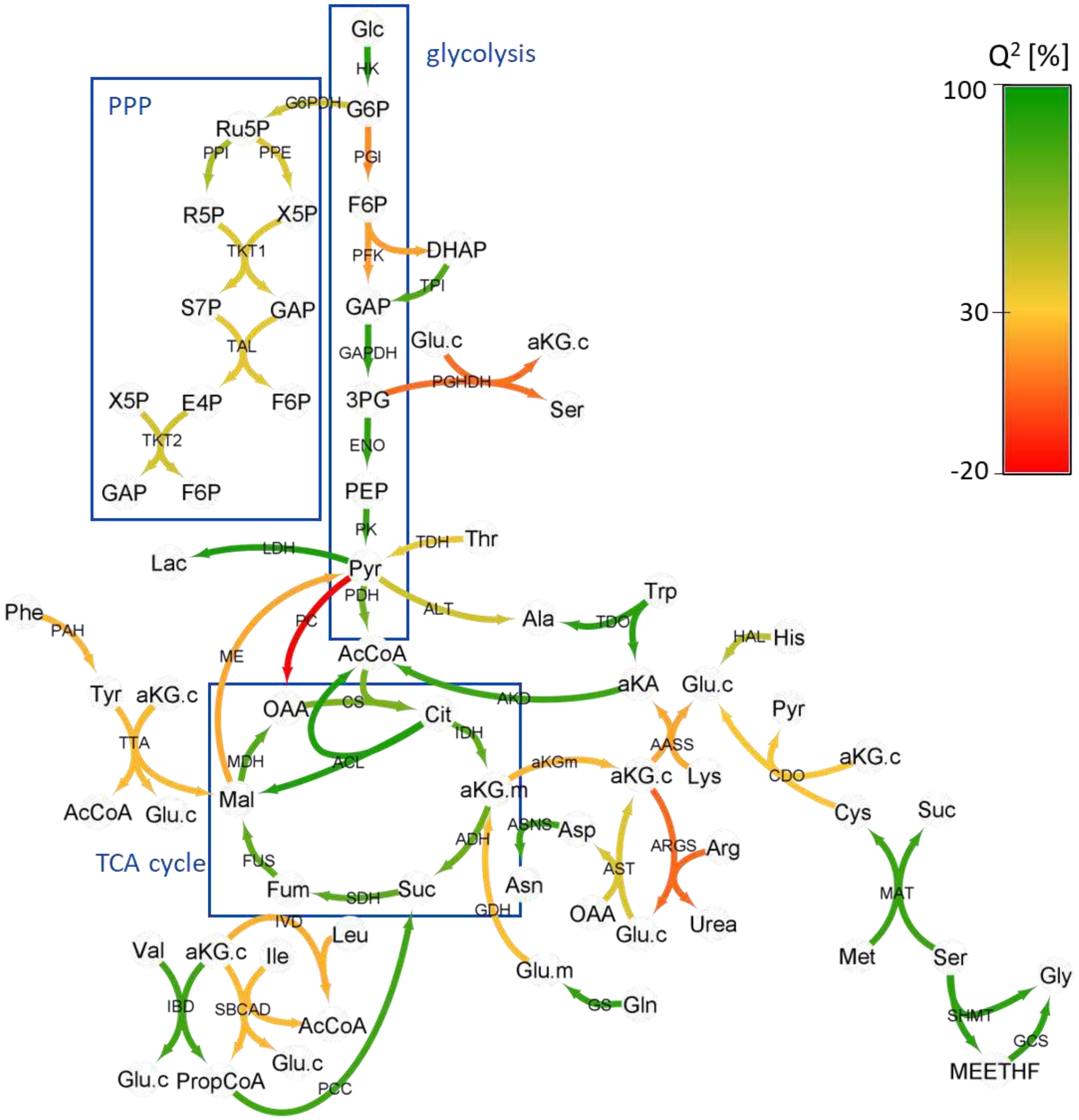


Figure 7.3 *Neural networks* Q^2 *in leave-one-out validation for the ensemble prediction intervals calculation method.*

Interestingly, low prediction performance is observed mainly in parallel/alternative reactions and reactions involving metabolites participating in many metabolic reactions. Specifically, the entire PPP is an alternative route to metabolize glucose and show lower prediction accuracy than the main route. Additionally, a deeper observation highlights that the reactions sharing metabolites between glycolysis and PPP (i.e., involving the metabolites *GAP* and *F6P*) show low Q^2. Similarly, almost all reaction involving *AcCoA*, *aKG.c* and *Glu.c*, which participate in many metabolic reactions, show low Q^2 in validation. This localized inaccuracy to specific reactions can be due to inconsistency in the available ^{13}C labelling data. In particular, the intracellular fluxes used for the training of the ANN are not directly measured experimentally but are calculated through a simplified metabolic model from the measured concentrations of the ^{13}C isotopes. This may lead to the introduction of inconsistency that are typical of metabolic models, such as in the case of parallel/alternative reactions, where fluxes are split in a somehow random way. This inconsistencies in the flux calculation might result in highly variable and badly distributed data that break the general relationship between inputs (extracellular metabolite uptake rates) and the intracellular flux value, leading to the inaccurate predictions of the neural networks.

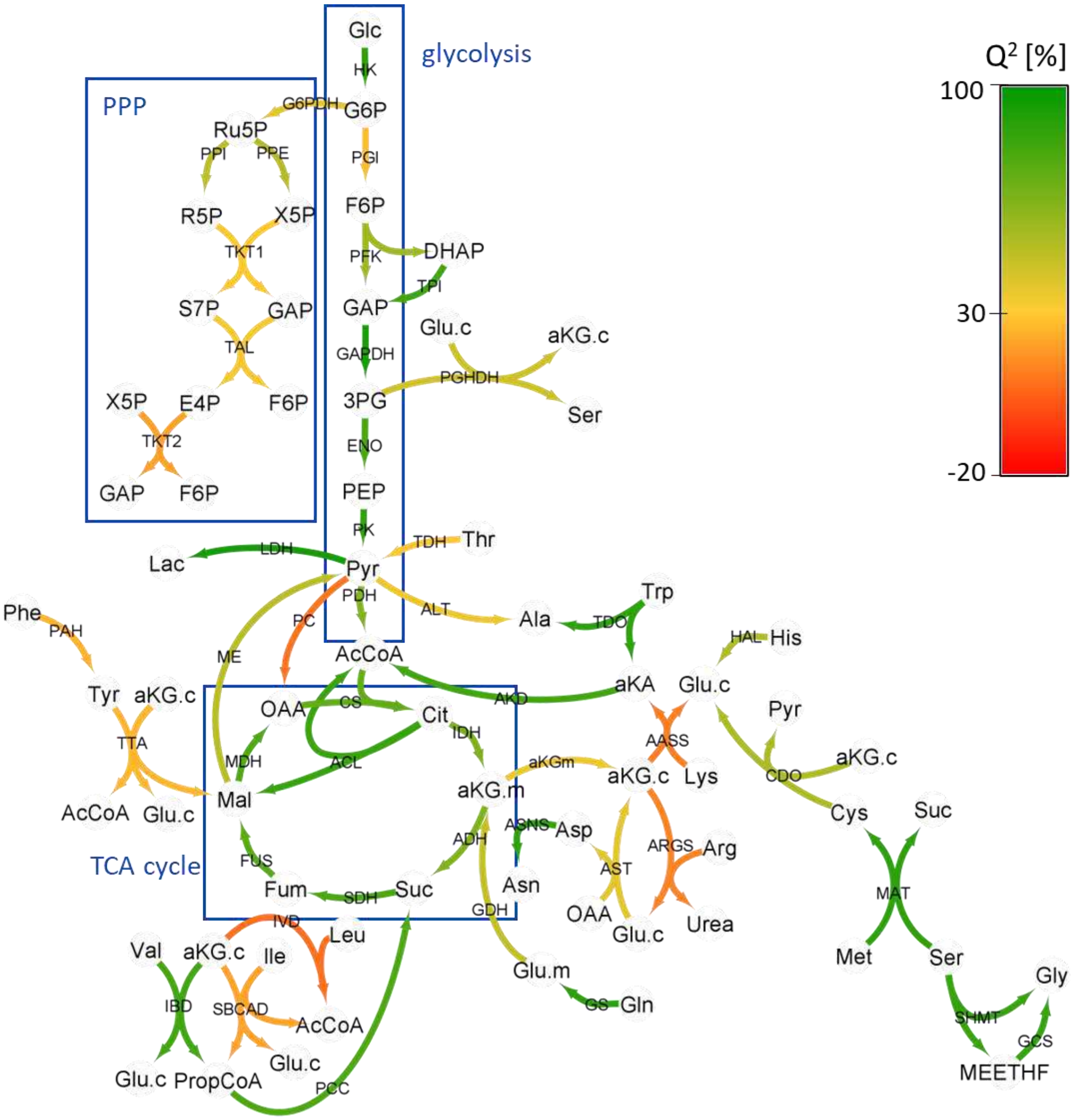


Figure 7.4 *Neural networks Q^2 in leave-one-out validation for the gradient prediction intervals calculation method.*

Further comparing the two prediction interval calculation methods (Figure 7.3 and 7.4), different prediction performance for some specific reactions can be observed. Reactions in the PPP, such as $TKT1$, TAL and $TKT2$, show higher Q^2 in the ensemble method, whereas glycolysis fluxes of reactions sharing metabolites with the PPP, such as PGI and PFK, are better predicted by the gradient method. Furthermore, reactions involving pyruvate (Pyr), such as ME, PC, and $PGHDH$, and the CDO reaction show higher Q^2 in the gradient method, while ALT, $AASS$, and $SBCAD$ reactions show higher Q^2 in the ensemble method.

Despite these small differences, both prediction interval calculation methods show good overall prediction performance, and no method outperforms the other. For this reason, both methods will be used to predict intracellular flux constraints and tested in the GSMM.

The constraints for intracellular reactions are estimated through the prediction intervals calculated with the ensemble and gradient methods. In both methods, the width of the prediction interval (PI) depends on the amount of uncertainty that characterize the prediction of a specific intracellular flux. Accordingly, a prediction affected by low uncertainty will show narrow PI, while a prediction affected by uncertainty will show wide PI. This prediction uncertainty is associated to each single experiment and depends on how the extracellular uptake rate values are positioned within the space generated by the uptakes of the training dataset.

An example of the PI estimated by the two methods is reported in Figure 7.5 for two different experiments, showing with the dots the estimated flux with their PIs (i.e., error bars) and the true ^{13}C flux with red dashed line. In both examples, the gradient method predicts the intracellular flux with high accuracy, as shown by the dot being closer to the red line. Furthermore, the estimated PIs change between samples and between PI calculation methods. An experiment with prediction affected by low uncertainty denoted by narrow PI is shown in Figure 7.5a, while an experiment with prediction affected by high uncertainty denoted by wide PI is shown in Figure 7.5b. In the case of low uncertainty, the ensemble method estimates very narrow PI, which are almost half in width of the one estimated by the gradient method. In the case of prediction affected by uncertainty, the ensemble method estimates very wide PI, almost double of the gradient method. The results shown in this example are generally valid for all experiments and intracellular reactions. Specifically, the ensemble method is more confident than the gradient method with predictions that result in low uncertainty, while being less confident in prediction characterized by high uncertainty. This results in narrower PIs for prediction with low uncertainty and wider PIs for prediction with uncertainty.

In general, predictions are required to be as accurate as possible, and the PIs as narrow as possible. This is true also in the specific application discussed here. However, currently it is not clear either if, in further constraining of the GSMM, PIs should be narrow or if wider constraints may be beneficial. For this reason, both PI methods will be tested on the GSMM in the following section.

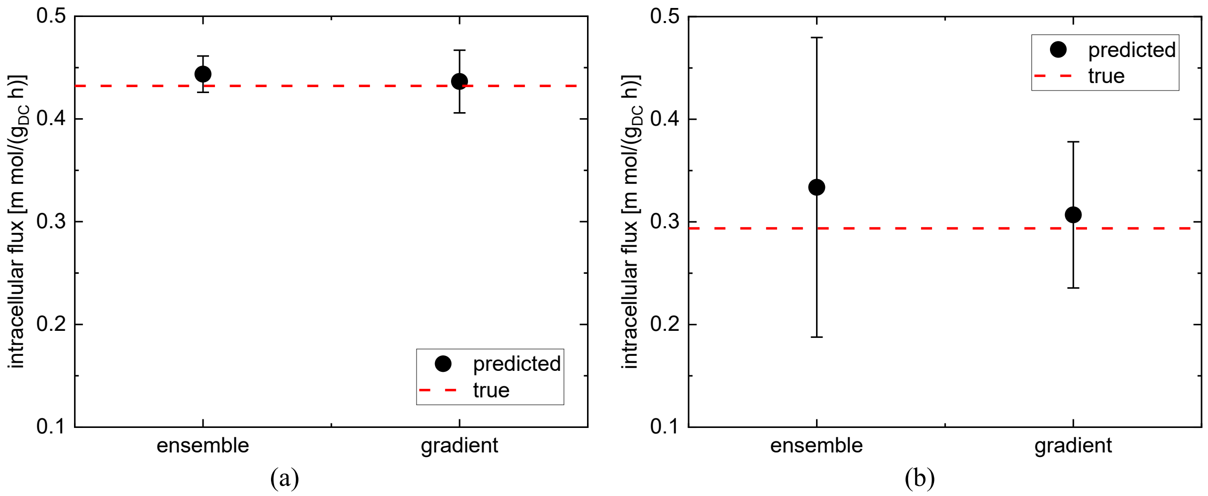


Figure 7.5 *Predicted intracellular flux of the ENO reaction with PI (error bars) for different experiments: (a) BCL2-M1 (prediction with low uncertainty), and (b) BCL2-M2 (prediction with uncertainty).*

7.2.2 Sensitivity to different number of network neurons

This Section is aimed at studying the sensitivity of the ANN prediction performance for different dimensions of the grid search for the optimal number of neurons. This is intended to better understand the impact that the number of neurons has on the predictions and how this reflects on the GSMM, allowing to select the best possible network.

To this purpose, the ANN was retrained multiple times selecting a different dimension of the grid search for optimizing the number of neurons. The tested grid search dimensions are: *i*) 0 to 25 (small size), *ii*) 0 to 50 (medium size), and *iii*) 0 to 100 (large size). The sensitivity on the different grid search dimensions is shown only for the gradient method, because it provides better performance on the GSMM (Section 7.3).

The Q^2 in leave-one-out validation for the three different grid search dimensions is reported in Figure 7.6. In general, medium and large size searches show slightly higher Q^2 that the small size search. In particular, fluxes of glycolysis and amino acid metabolism are predicted with similar accuracy in the three grid searches, the PPP fluxes are generally better predicted in the medium size search, while the TCA cycle fluxes are predicted with higher accuracy in the medium and large size searches. This indicates that low number of neurons is typically not enough to capture the general relationship between extracellular metabolite uptake rates and intracellular fluxes. Fluxes of some intracellular reactions, such as *aKG.m*, *PC*, *AST*, and *ARGS*, are badly predicted ($Q^2 < 0$) in the low size search, while medium and large size searches show higher accuracy and positive Q^2. However, fluxes of reactions *PGI*, *TPI*, *HK*, *PPI*, *TDH*, and *SBCAD* are predicted with the highest Q^2 in the small size search, indicating that in some cases the relationship between uptakes and intracellular fluxes is simple enough to be generalized by a small neural network.

Medium and large size searches show similar Q^2 values, with medium size search being slightly better. In fact, the prediction of the reactions *G6PDH*, *TKT1*, *TAL*, *TKT2*, *FUS*, *IDH*, *ACL*, *ALT*, *IVD*, *TTA*, and *AASS* are substantially better in the medium size search.

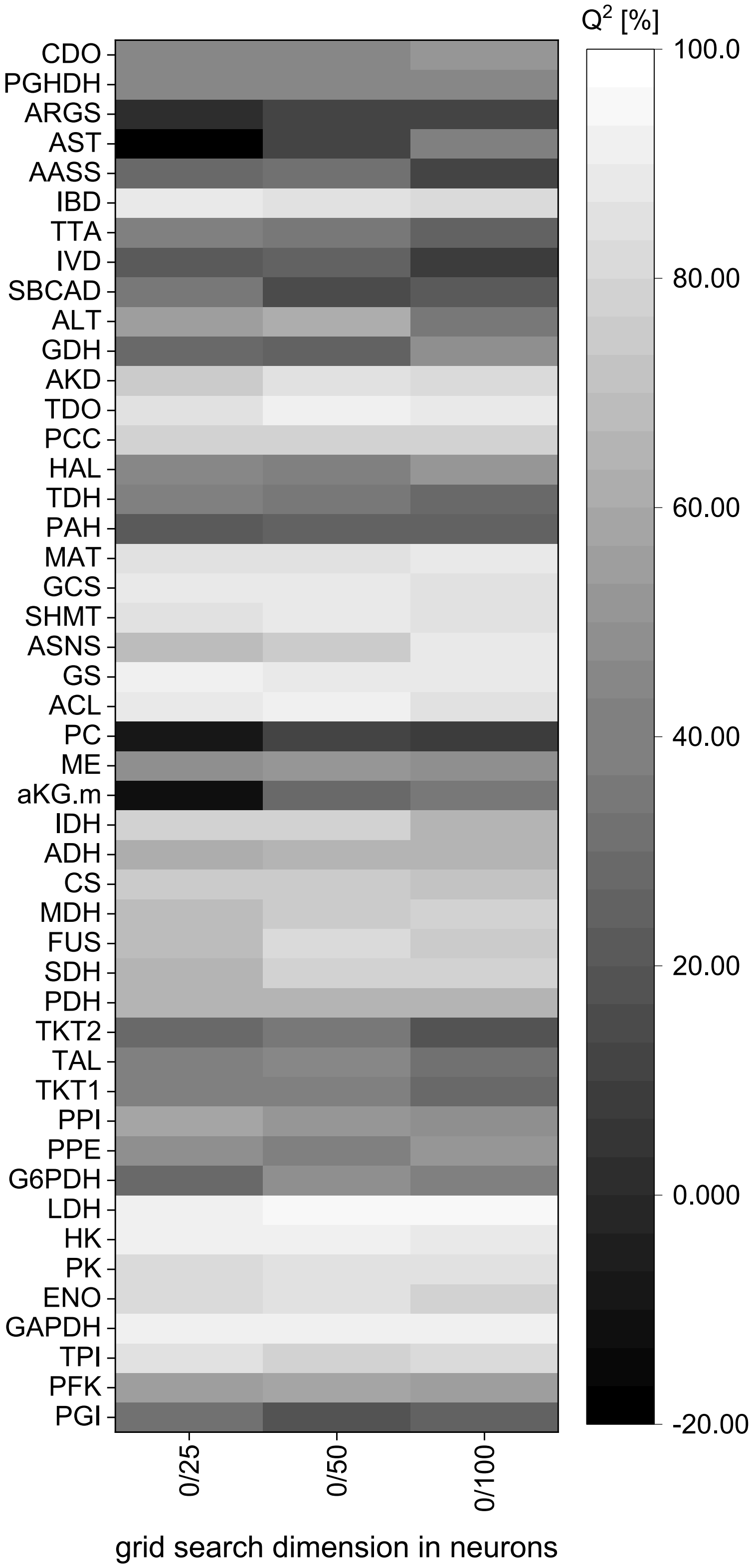


Figure 7.6 *Neural networks Q^2 in leave-one-out validation for the gradient PI calculation method using different grid search dimensions in the optimization of the number of neurons.*

In ANN, the use of a lower number of neurons is generally preferred to reduce overfitting and achieve a more robust weight estimation with the available training experiments. In this case, since the medium size search achieves similar and even slightly better prediction performance using a smaller number of neurons, this configuration is to be preferred with respect to the others. However, the constraints predicted by all three configurations will be tested on the GSMM to assess the best possible model structure.

7.3 Genome-scale metabolic model predictions

In this Section, the prediction performance of the GSMM when it is constrained with the neural network predicted bounds is analyzed to identify the best neural network configuration and understand the performance of the proposed methods with respect to the state-of-the-art.

To this purpose, the GSMM is constrained with the extracellular metabolite uptake rates **X** used by the ANN for prediction and the estimated PI are used as lower and upper bound to constrain intracellular reactions. In the base case (Section 7.1.3.4), the GSMM is constrained with the extracellular metabolite uptake rates, while intracellular reactions have default upper and lower flux bounds. In both cases, the flux sampling method is used to calculate the distribution of intracellular fluxes.

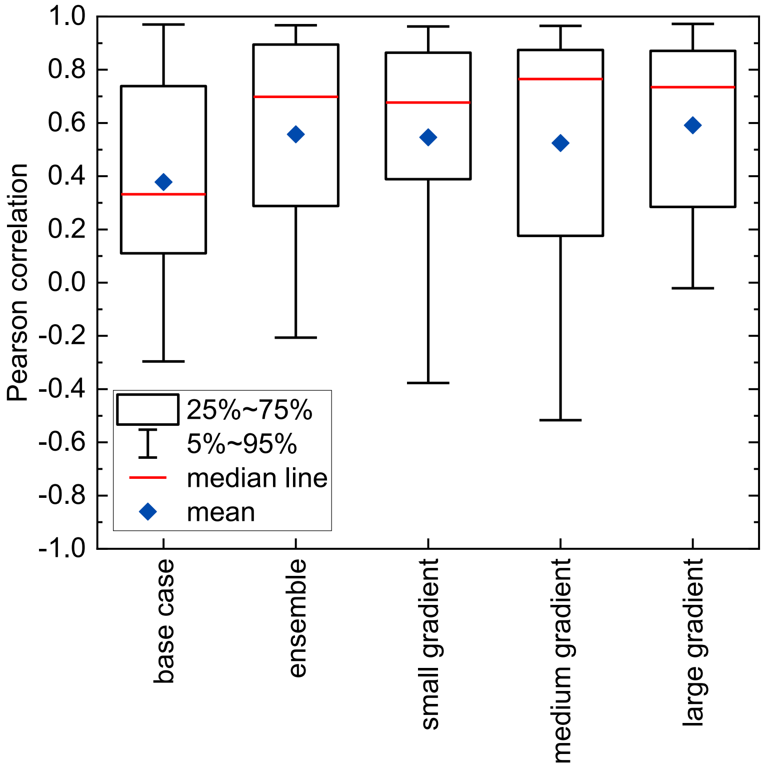


Figure 7.7 *Pearson correlation distributions of GSMM calculated intracellular fluxes for base case and the proposed ANN in different configurations.*

The performance of the GSMM in calculating intracellular fluxes is reported in terms of Pearson correlation (Section 7.1.3.4) in Figure 7.7. The represented distributions consider the Pearson correlation over all leave-one-out validation experiments and all experimentally available intracellular fluxes.

The proposed methodology of constraining the flux of intracellular reactions with the ANN prediction intervals outperforms the base case independently on the specific ANN configuration, showing higher mean, median, and 25/75 percentiles of the Pearson correlation distribution, and, in many cases, an overall narrower distribution. Accordingly, the proposed method largely improves the GSMM capability of calculating intracellular fluxes, making GSMM a reliable methodology to improve the description of cell metabolism. The proposed methodology outperforms also other state-of-the-art methods, such as pFBA and ccFBA. However, these results are excluded from this Dissertation, because they are related to the partner work.

When comparing different configuration of the proposed ANN models, all gradient methods show higher Pearson correlation than the ensemble method, i.e., higher prediction accuracy. This is not true for the small size search which shows a lower mean and median Pearson correlation than the ensemble method (median 0.677 vs. 0.698; mean: 0.546 vs. 0.557). Consequently, the generally wider PIs of the gradient method, together with the narrower ones for prediction affected by uncertainty, determine that gradient method outperforms the ensemble one. This is probably due to the fact that the wider bounds (of predictions not affected by large uncertainty) allow a sufficient freedom in adjusting the overall fluxes across the metabolic network, while the narrower ones (of predictions affected by uncertainty) avoid an excessive freedom which drives the calculated flux away from the experimental value.

Among different method configurations for the gradient PI calculation, the medium size search shows the highest median value (0.765), but a slightly lower mean that the large size search (0.525 vs. 0.591). Furthermore, the medium size search shows wider 25/75 percentiles than the large size one. This denotes that the majority of the intracellular fluxes have higher Pearson correlation (i.e., higher accuracy) than the large size search case, but in some cases the correlation can be lower (i.e., lower accuracy). These metrics indicate that the ANN in the medium and large size search configurations allows the best calculation of the intracellular fluxes in the GSMM, with the medium size configuration being slightly better. As previously explained, the medium size search configuration is to be preferred, because it results more robust and less prone to overfit new datapoints due to the lower number of neurons and allows GSMMs to achieve good performance in intracellular flux calculation.

The GSMM is also used to estimate the cell growth rate by maximizing the produced biomass through FBA. The proposed methodology (large size search configuration) outperforms the base case method in the prediction of the growth rate, with a $Q^2 = 61.6\%$ evaluated on the leave-one-out validation experiments against a $Q^2 = 13.3\%$ of the base case.

7.4 Deployment of Next-FLUX

The methodology proposed in this Chapter resulted in the deployment of Next-FLUX (Neural-net EXtracellular Trained flux) software. This software is used to predict intracellular fluxes and their PIs to constrain GSMMs starting from commonly available extracellular metabolite uptake rates.

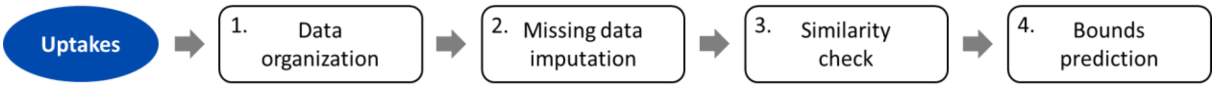


Figure 7.8 *Deployed software procedural steps.*

The software has the sequential structure of Figure 7.8. The ANN inputs are extracellular metabolite uptake rates (Section 7.1.3), which are required to predict intracellular fluxes. Next-FLUX has four sequential steps:

1. Data organization: initially, the provided uptakes are compared with ones required by the ANN and organized accordingly. Unnecessary uptakes are discarded, while any missing required uptake is recorded and will be treated in step 2.
2. Missing data imputation: scattered missing data and entirely missing uptakes are imputed at this step. The imputation is performed based on a PCA model calibrated on the available ^{13}C labelling dataset used for ANN training through the procedure proposed by Muñoz et al. (2004). In this method, the known data are used to estimate the score of the PCA model through the respective loadings. From the score, the unknown uptakes values are estimated using the respective loadings. This procedure is iterated until convergence of the score value, whose final converged value gives the imputed missing uptakes.
3. Similarity check: at this stage the similarity of the input datapoints with the training experiments is assessed through a PCA model. The same PCA model of step 2 is used to calculate the $T^2$and SPE diagnostics (Section 2.1). Based on these diagnostics the new datapoints are categorized as similar or different from the training experiments. Different rationales based on PCA diagnostics can be selected by the user to define similarity: *i*) inside confidence limits of both diagnostics, and *ii*) inside the confidence limit of a single diagnostic index. The user is informed of the new datapoints passing the similarity check, and only the similar datapoints are progressed to step 4. The user can force the software to progress all new datapoints to step 4 independently on the similarity check results.
4. Bounds prediction: at this step, the intracellular fluxes and their PIs are predicted for all the available intracellular reactions and datapoints provided by step 3. An ANN trained on all available ^{13}C labelling experiments is used for the prediction. The ANN is trained following the same procedure explained in Section 7.1.3 without excluding any experiment for validation.

The predicted PIs can be then transferred to a GSMM and used as constraints for intracellular reaction.

7.5 Concluding remarks and future work

In this Chapter, a deep learning strategy to predict GSMM intracellular constraints from easily available and cheap experimental data was proposed. The proposed strategy exploits an artificial neural networks, trained on ^{13}C isotope labeling experimental data, to predict the main intracellular fluxes from extracellular metabolite uptake rates and provide an estimation of the lower and upper flux bounds for the GSMM through the calculation of prediction intervals.

In a leave-one-out validation, the artificial neural networks accurately predicted with $Q^2 > 65\%$ most of the intracellular fluxes from the extracellular metabolite uptake rates, mainly reactions in glycolysis, TCA cycle, and some reactions in the amino acid metabolism, such as GS, $ASNS$, $SHMT$, GCS, MAT, TDO, AKD, IBD. However, slightly lower performance ($Q^2 < 35\%$), probably due to small inconsistency in the training data, was observed in parallel and alternative reactions, such as PPP ones, and reactions with metabolites participating in many metabolic reactions, such as PC, $ARGS$, TTA, and $AASS$. Two prediction interval calculation methods were compared, which showed similar prediction performance, but estimated prediction intervals with slightly different widths.

A GSMM constrained with the predicted bounds showed better performance in calculating intracellular fluxes than the base case and other state-of-the-art methods, and more accurate biomass predictions. The gradient prediction interval calculation method showed better performance in the GSMM, because of the estimated bounds show smaller width variability.

This work resulted in the deployment of an automated software for the prediction of GSMM intracellular constraints, named Next-FLUX, Neural-net EXtracellular Trained FLUX. This software automatically checks the consistency of the provided data with the historical ones and provided estimations of GSMM intracellular constraints.

In the future, Next-FLUX will form an integrated platform for GSMM analysis, embedding the application of data-driven constraints and the solution of the GSMM. Furthermore, Next-FLUX will be validated on independent ^{13}C isotope labelling datasets to assess its general applicability and robustness.

Chapter 8

Data-driven genetic engineering[*]

In this Chapter, a machine learning strategy to identify targets for the genetic engineering of host cells exploiting genome-scale metabolic models (GSMMs) is proposed. Initially, the machine learning strategy based on latent variable model inversion and the generation of the strain specific GSMMs is explained. Then, the outcomes of the machine learning model used for genetic engineering are presented and the identified genetic modifications are analyzed and discussed.

8.1. Material and Methods

In this Section, the data and the main mathematical methodologies used in this work are presented. Specifically, the available culture data, the GSMM used in this work and the generation of metabolic data is initially presented. Then, the machine learning strategy to identify the genetic modifications is detailed.

8.1.1 Available culture data

A CHO cell culture dataset assembled from 4 different publicly available sources is used in this work. The dataset contains experimental measurements of 24 extracellular metabolite uptake rates for $N = 23$ different cell lines. For each cell line, the upper, lower, and median value of the observed metabolite uptake rates are available. The experimental datasets span a wide variety of CHO cell types, phases, and processing conditions, which are reported in Table 8.1 with the name given to each cell line. Note that in the following cell lines will be referred with their specific names.

In this work the extracellular metabolite uptake rates are converted into units of $mmol/g_{DCW}h$, through the dry cell weight experimentally measured for each cell line, to be used in the GSMM.

[*] This work is a collaboration with Imperial College London (UK).

Table 8.1 *Experimental data sources and detail on cell line conditions.*

Source	Cell type	# of experiments	Culture phase	Cell line names
Templeton et al. (2013)	CHO-S producer	4	Early exponential, late exponential, stationary, and decline	early, late, stat, decline
Templeton et al. (2014)	CHO-S non-producer	6	Early and late exponential	CLP, LELP, HELP, CLC, LELC, HELC
Templeton et al. (2017a)	CHO-S producer and non-producer	11	Stationary	SV, SVGS, SVM1, SVM2, SVM3, SVM4, BCL2, BCL2-M1, BCL2-M2, BCL2-M3, BCL2-M4
Templeton et al. (2017b)	CHO-S producer	2	Perfusion and stationary	fed-batch, perfusion

8.1.2 Proposed strategy

In this Chapter, a strategy for the identification of genetic engineering targets (i.e., genetic modifications) that improve mAb productivity in GSMMs is proposed. The proposed strategy comprises four steps (Figure 8.1), which are presented in the following.

1. Metabolic data generation: a dataset of intracellular reaction rates (i.e., intracellular fluxes) and phenotypes (i.e., productivity and biomass) is generated through a GSMM (Section 8.1.3). This dataset is used as base to identify genetic modifications leading to an improved phenotype by calibrating a latent variables regression model.
2. Latent variable model inversion: a latent variable regression model is trained on the dataset generated in step 1. Despite the large number of metabolic reactions, this model retains all available intracellular fluxes because we are interested in correlating the entire intracellular flux distribution with cell phenotypes. The latent variables model is then inverted to estimate the intracellular flux distribution associated with desired and improved cell phenotypes.
3. Identification of genetic modifications: a specifically developed algorithm exploits a GSMM to identify a small set of genetic modifications that produce the desired phenotypes. Based on the intracellular flux distribution calculated in step 2, the algorithm identifies key metabolic reactions and associated genes to genetically modify.
4. Test on the GSMM: the rate of the metabolic reactions identified in step 3 is adjusted in the GSMM according to the intracellular flux distribution estimated in step 2. The solution of the GSMM tests if the identified genetic modification improves cell phenotypes and allows to understand the metabolic reasons behind the improved performance.

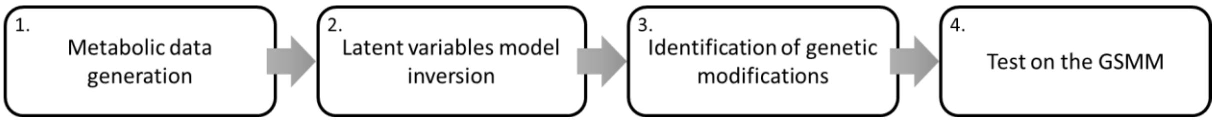

1.
Metabolic data generation
2.
Latent variables model inversion
3.
Identification of genetic modifications
4.
Test on the GSMM

Figure 8.1 *Proposed strategy for the identification of genetic engineering targets that improve mAb productivity.*

8.1.3 Strain specific Genome-scale Metabolic Model and metabolic data generation

A comprehensive CHO cell GSMM[4], iCHO2441, is used in this work. The GSMM couples the current CHO cell metabolic information of iCHO2291 (Yeo et al., 2020) to the entire protein secretory pathway of iCHO2048 (Gutierrez et al., 2020). To our knowledge, this is the most complete CHO GSMM to date, with the highest number of annotated genes and gene per reaction ratio.

This GSMM is used as base to generate a strain-specific CHO-S metabolic model using publicly available RNA-Seq data (Hefzi et al., 2016). The expression levels of 26795 genes are available at 10 time points throughout the culture and averaged in time. The procedure suggested by Hart et al. (2013) is used to identify the cut-off expression value for unexpressed genes. The Gene Inactivity Moderated by Metabolism and Expression (GIMME) algorithm (Becker & Palsson, 2008) is used to generate the strain specific model from the expression levels and the previously determined cut-off value. Strain specific models are created per cell line by constraining the GSMM with the extracellular metabolite uptake rates and using biomass as objective function with an optimality threshold of 0.9. The final CHO-S model used in this work is generated by joining all reactions retained in the strain specific models reconstructed for each cell line. The generated CHO-S GSMM comprises $D = 4069$ metabolites, $U = 5624$ metabolic reactions, and 2111 genes.

The CHO-S GSMM is initially used to generate a synthetic dataset to calibrate latent variable regression models. To generate data, the GSMM is constrained with the available upper and lower bounds of extracellular metabolite uptake rates (Section 8.1.1) and is solved by maximizing biomass through pFBA (Section 2.3.1).

The intracellular reaction rates (i.e., intracellular fluxes) calculated by the GSMM for all available experiments are organized in the matrix $\mathbf{X}\,[N \times V]$, where V is the number of intracellular reactions/fluxes considered in latent variable modeling. The direct secretory reactions, from protein translation to secretion, are excluded from the dataset, since their flux is almost equal to the produced mAbs, and we are not interested in genetic modification concerning the direct secretory pathway. Furthermore, reactions with zero flux in all N cell

[4] The GSMM used in this work has been developed Benjamin Strain, Ph.D. student at Imperial College London (U.K.).

lines, are excluded from the dataset. The resulting dataset of intracellular fluxes is $\mathbf{X}\,[23 \times 721]$.

The M cell phenotypes predicted by the GSMM for all available cell lines are organized in the matrix $\mathbf{Y}\,[N \times M] = [23 \times 2]$, containing cell specific mAb productivity and biomass (i.e., growth rate).

8.1.4 Latent variables regression model inversion

A PLS model (Section 2.1.2) is built to correlate the intracellular fluxes **X** to the cell phenotype **Y**. The number of LVs is selected through a leave-one-out cross-validation as the one minimizing the root mean squared error (RMSECV). The same cross-validation is used to assess the prediction performance of the model, since an external validation dataset is absent, and all the available experiment are needed for model calibration. Cross-validation performance is measured in terms of RMSECV and coefficient of determination (Q^2).

In this model variable selection is not applied despite the imbalance in **X** between the number of cell lines (i.e., observations) and metabolic reactions. This because we are interested in correlating the entire intracellular flux distribution with cell phenotypes for the further identification of genetic modifications among all possible metabolic reactions.

Furthermore, a physical constraint on the predicted productivity is imposed in the PLS model, which sets the predicted productivity to 0 in case of a negative predicted value.

The inversion of PLS models (Section 2.2.5) consists in the estimation of the new set of intracellular fluxes $\mathbf{x}_{\text{NEW}}$ corresponding to the desired phenotype $\mathbf{y}_{\text{DES}}$, while: *i*) minimizing the Hotelling's T^2, *ii*) minimizing the reconstruction error SPE and ensuring it smaller that its 95% confidence limit SPE_{lim}, *iii*) satisfying the PLS model equations, *iv*) and satisfying the constraints on $\mathbf{x}_{\text{NEW}}$ and $\mathbf{y}_{\text{DES}}$. In this case, given that there are no equality constraints on $\mathbf{y}_{\text{DES}}$ the general formulation of the PLS inversion (Equation 2.14) can be simplified as:

$$\min_{\mathbf{x}_{\text{NEW}}} [g_1 T^2 + \; g_2 SPE] \quad , \tag{8.1}$$

subject to the PLS model and SPE constraints:

$$\hat{\mathbf{y}}_{\text{DES}} = \mathbf{t}_{\text{DES}} \mathbf{Q}^{\text{T}} \quad , \tag{8.2}$$

$$\hat{\mathbf{x}}_{\text{NEW}} = \mathbf{t}_{\text{DES}} \mathbf{P}^{\text{T}} \quad , \tag{8.3}$$

$$\mathbf{t}_{\text{DES}} = \mathbf{x}_{\text{NEW}} \mathbf{W}^* \quad \text{, and} \tag{8.4}$$

$$SPE = (\hat{\mathbf{x}}_{\text{NEW}} - \mathbf{x}_{\text{NEW}})(\hat{\mathbf{x}}_{\text{NEW}} - \mathbf{x}_{\text{NEW}})^{\text{T}} \leq g_3 SPE_{\text{lim}} \quad , \tag{8.5}$$

and to the constraints on $\mathbf{x}^{\text{NEW}}$ (Eq. 8.6-8.7) and $\mathbf{y}_{\text{DES}}$ (Eq. 8.8-8.9), where $\hat{\mathbf{y}}_{\text{DES}} = [\hat{y}_{\text{mAb}} \quad \hat{y}_{\text{biom}}]$ is the PLS predicted phenotype, $\mathbf{t}_{\text{DES}}$ is the PLS score vector related to $\mathbf{x}_{\text{NEW}}$, **Q** is the response loading matrix, $\hat{\mathbf{x}}_{\text{NEW}}$ is the PLS reconstruction of $\mathbf{x}_{\text{NEW}}$, **P** is the PLS loading matrix, $\mathbf{W}^*$ is the PLS weight matrix, g_1, g_2, and g_3 are corrective constants, and T denotes the transpose operation. In this work all corrective constants are set to 1.

The constraints, set for the estimated intracellular fluxes $\mathbf{x}_{\mathrm{NEW}} = \left[\mathbf{x}_{\mathrm{NEW}}^{\mathrm{extra}}, \mathbf{x}_{\mathrm{NEW}}^{\mathrm{intra}}\right]$, require satisfying both the generic intracellular constraints of the GSMM and the extracellular constraints (i.e., measured extracellular metabolite uptake rates) of the experiment for which the intracellular fluxes are estimated, as:

$$\nu_{\mathrm{MIN}}^{\mathrm{intra}} \leq x_{\mathrm{NEW}}^{\mathrm{intra}} \leq \nu_{\mathrm{MAX}}^{\mathrm{intra}} \quad \text{and} \tag{8.6}$$

$$\nu_{\mathrm{MIN},n}^{\mathrm{extra}} \leq x_{\mathrm{NEW}}^{\mathrm{extra}} \leq \nu_{\mathrm{MAX},n}^{\mathrm{extra}} \quad , \tag{8.7}$$

where $x_{\mathrm{NEW}}^{\mathrm{intra}}$ and $x_{\mathrm{NEW}}^{\mathrm{extra}}$ are the estimated intracellular flux values for a generic intracellular and extracellular exchange reaction respectively, $\nu_{\mathrm{MIN}}^{\mathrm{intra}}$ and $\nu_{\mathrm{MAX}}^{\mathrm{intra}}$ are genetic intracellular flux bounds given by the GSMM, and $\nu_{\mathrm{MIN},n}^{\mathrm{extra}}$ and $\nu_{\mathrm{MAX},n}^{\mathrm{extra}}$ are generic flux bounds for extracellular exchange reactions given by the extracellular metabolite uptake rates of the n-th cell line (Section 8.1.1). In constraining $\mathbf{x}_{\mathrm{NEW}}$ within the measured extracellular metabolite uptake rate bounds, it is assumed that the culture conditions and overall extracellular behavior of cells are the same even after the genetic modification.

The inversion of the PLS model is performed for each available cell line, meaning that the measured extracellular metabolite uptake rates of each experiment are used to set the constraints on $\mathbf{x}_{\mathrm{NEW}}$ (Eq. 8.6 and 8.7) and specific constraints are set for the desired phenotype $\mathbf{y}_{\mathrm{DES}} = \left[y_{\mathrm{DES}}^{\mathrm{mAb}} \quad y_{\mathrm{DES}}^{\mathrm{biom}}\right]$. Specifically, the intracellular fluxes $\mathbf{x}_{\mathrm{NEW}}$ will be estimated for each one of the *N* available cell lines from the corresponding $\mathbf{y}_{\mathrm{DES}}$. For each cell line, the desired phenotype is an increased mAb specific productivity, while allowing a reduction in biomass to compensate for resource reallocation and sustain increased mAb production. This translates in the imposition of inequality constraints on the $\mathbf{y}_{\mathrm{DES}}$ as:

$$y_{\mathrm{DES}}^{\mathrm{mAb}} \geq \lambda_{\mathrm{mAb}}\, y_n^{\mathrm{mAb}} + y_0^{\mathrm{mAb}} \quad \text{and} \tag{8.8}$$

$$y_{\mathrm{DES}}^{\mathrm{biom}} \geq 0.9\, y_n^{\mathrm{biom}} \quad , \tag{8.9}$$

where y_n^{mAb} and y_n^{biom} are the original biomass and productivity of the n-th cell line, y_0^{mAb} is a productivity bias used to set a positive productivity value in non-productive experiments (i.e., when $y_n^{\mathrm{mAb}} = 0$), and λ_{mAb} is an increase factor, which was set between 1.5 and 2 to adjust the productivity requirement according to the specificity of each experiment. These values are arbitrarily set to induce a significant productivity increase in the further application on the GSMM.

8.1.5 Genetic modifications identification

The intracellular fluxes $\mathbf{x}_{\mathrm{NEW}}$ (i.e., intracellular flux distribution) estimated by the PLS model inversion are applied to the GSMM for testing *in silico* the mAb productivity that can be achieved by the new flux distribution. This simulates *in silico* the cell phenotype change resulting from the variation of certain metabolic reaction rates. To induce metabolic reaction rate variations in cells, the expression of the gene or genes associated to the specific metabolic reactions under consideration must be regulated. Gene expression can be:

- upregulated: induces an increased flux in the associated reaction;
- downregulated: induces a decreased flux in the associated reaction;
- knocked out: stops the flux through the associated reaction.

In GSMMs, gene regulation is simulated by directly acting on the flux through the associated metabolic reactions. These intracellular flux changes are achieved by acting on the constraints of metabolic reactions (i.e., upper and lower bounds of the flux value), which are set according to the intracellular flux values $\mathbf{x}^{\text{NEW}}$ estimated by PLS inversion. For each metabolic reaction v and a generic cell line, the intracellular constraints are adjusted according to type of regulation as:

$$k_{\text{reg}}\, \boldsymbol{x}_{\text{NEW},v} \leq \nu_v \leq \nu_{\text{MAX},v} \quad \text{if} \quad \boldsymbol{x}_{\text{NEW},v} > x_v \quad \text{(upregulation)}, \tag{8.10}$$

$$\nu_{\text{MIN},v} \leq \nu_v \leq k_{\text{reg}}\, \boldsymbol{x}_{\text{NEW},v} \quad \text{if} \quad \boldsymbol{x}_{\text{NEW},v} < x_v \quad \text{(downregulation), and} \tag{8.11}$$

$$0 \leq \nu_v \leq 0 \quad \text{if} \quad \boldsymbol{x}_{\text{NEW},v} = 0 \quad \text{(knockout)}, \tag{8.12}$$

where $\boldsymbol{x}_{\text{NEW},v}$ is the estimated flux of reaction v in the PLS inversion, ν_v is the flux of the v-th intracellular reaction, $\nu_{\text{MAX},v}$ is the flux upper bound for reaction v given by the GSMM, $\nu_{\text{MIN},v}$ is the flux lower bound for reaction v given by the GSMM, x_v is the intracellular flux of the v-th reaction for a generic cell line from the generated dataset **X** (Section 8.1.3), and k_{reg} is a scaling coefficient. Note that the constraints of exchange reactions are not modified, since $x_{\text{NEW}}^{\text{extra}}$ from the PLS inversion is within the measured extracellular metabolite uptake rates as set by the related inversion constraints (Eq. 8.7).

In PLS inversion, the estimated intracellular fluxes for a generic cell line $\mathbf{x}_{\text{NEW}}$ are varied with respect to the original ones (i.e., from **X**) in a multivariate fashion according to the correlations captured by the PLS model, meaning that the intracellular flux values of all the V metabolic reactions are varied together to achieve the desired increase in mAb productivity, requiring the execution of V genetic modifications. However, in *in vitro* applications it is not practical to perform a large number of genetic modifications, nor it is typically necessary to achieve improved phenotypes. Similarly, in GSMMs, a small set of constraint changes (i.e., simulated genetic modifications) is typically sufficient to achieve the desired intracellular fluxes and induce increased productivity.

Hence, an algorithm that identifies the minimum set of reactions to modify in such a way as to determine an increased productivity in GSMMs is developed. The proposed algorithm (Figure 8.2) comprises the following steps:

- Step #1 – Randomize the order in which reactions are considered: the order, in which the genetic modification of each metabolic reaction is tested, is randomized. Extracellular exchange reactions $\mathbf{x}_{\mathrm{NEW}}^{\mathrm{extra}}$ are not considered since they are constrained by the experimental metabolite uptake rates.
- Step #2 – Select a reaction v to genetically modify: a reaction whose flux should be modified, v, is selected according to the randomized order determined in step #1.
- Step #3 – Test the genetic modification of reaction v: the intracellular flux change suggested by the PLS inversion is applied to the selected reaction v. The constraints are changed according to Eq. 8.10, 8.11, or 8.12 if either an increased, a decreased, or a zero intracellular flux is required, respectively. Then, the GSMM is solved with pFBA (Section 2.3.1).
- Step #4 – Check feasibility of the genetic modification: if the GSMM is infeasible or produces a negative biomass or productivity the genetic modification of reaction v is not accepted and its constraints are set to the original values, otherwise the genetic modification of reaction v is accepted, and the new constraints set in step #3 are kept.
- Step #5 – Test genetic modification of all V reactions: a new reaction v is selected until the genetic modification of all V intracellular reactions is tested.
- Step #6 – GSMM with all feasible genetic modifications: once the genetic modification of all V intracellular reactions is tested (step #5), the estimated intracellular fluxes $\mathbf{x}_{\mathrm{NEW}}$ are applied to the GSMM by constraining all feasible reactions among the V ones, producing a multivariate set of genetically modified reactions S.
- Step #7 – Exclusion of genetic modifications: the genetic modification unnecessary to achieve an increased productivity in the GSMM are excluded in the next section of the algorithm. A counter it defines the number of times that this exclusion procedure is performed.
- Step #8 – Select a genetic modification s to exclude: one genetically modified reaction $s \in S$ is selected according to the initial randomized order (step #1).
- Step #9 – Test the exclusion of the genetic modification s: the constraints of the reaction s are set to the original value. The GSMM includes all genetic modifications that are kept during this exclusion section (step #10), while at the beginning of the exclusion section the GSMM with S genetic modifications (step #6). The GSMM is solved with pFBA.
- Step #10 – Check essentiality of the s genetic modification: if the solution of the GSMM shows positive and increased productivity (i.e., than the previously observed one), the genetic modification of reaction s is not necessary to achieve the improved productivity. Hence, the constraints of reaction s are kept to the original values as set in step #9. Otherwise, the genetic modification of reaction s is essential to achieve the improved productivity and must be kept. Hence, the constraints of reaction s are set again to the value estimated by the PLS inversion (set at step #3).

- Step #11 – Test exclusion of all S genetic modifications: a new genetically modified reaction s is selected until the exclusion of all S genetic modifications is tested.
- Step #12 – Iterate the exclusion of genetic modifications: once all S genetic modifications is tested, the counter it is updated $it = it + 1$ and the exclusion of genetic modification (steps #8-11) is performed again. This is done to obtain the smallest set of essential genetic modifications to achieve the improved mAb productivity. The exclusion section is performed until $it < it_{\max}$.
- Step #13 – Final set of genetic modifications: one the exclusion section is terminated, the smallest set of genetically modified metabolic reactions that allow to achieve improved productivity is obtained.

The set of metabolic reactions to genetically modify is tested for reaction essentiality through a sensitivity analysis. In this case, one reaction at a time is considered and its constraints are set to the original value. If a productivity reduction greater than 20% is observed, the genetic modification of a reaction is considered essential and must be kept, otherwise it is not essential to achieve improved productivity.

A complete run of the proposed algorithm produces a genetic engineering scenario, with the smallest set of essential reaction to modify to obtain an increased productivity in the considered experiment. In this work, 30 different scenarios are run, each one with a different order for testing the reactions (step #1) and scaling coefficient k_{reg} between 1 and 1.3 (Eq. 8.10 and 8.11).

Each genetic engineering scenario is then applied on a GSMM to quantify *in silico* the new improved mAb productivity, study the intracellular flux distribution and understand the reason behind the increased mAb production.

All codes used in this Chapter were developed in Python 3.10, using COBRApy (Ebrahim et al., 2013), PyPhi (https://github.com/salvadorgarciamunoz/pyphi), and Pyomo (W. E. Hart et al., 2011). All the metabolic reactions and genes mentioned in this Chapter can be found at http://bigg.ucsd.edu.

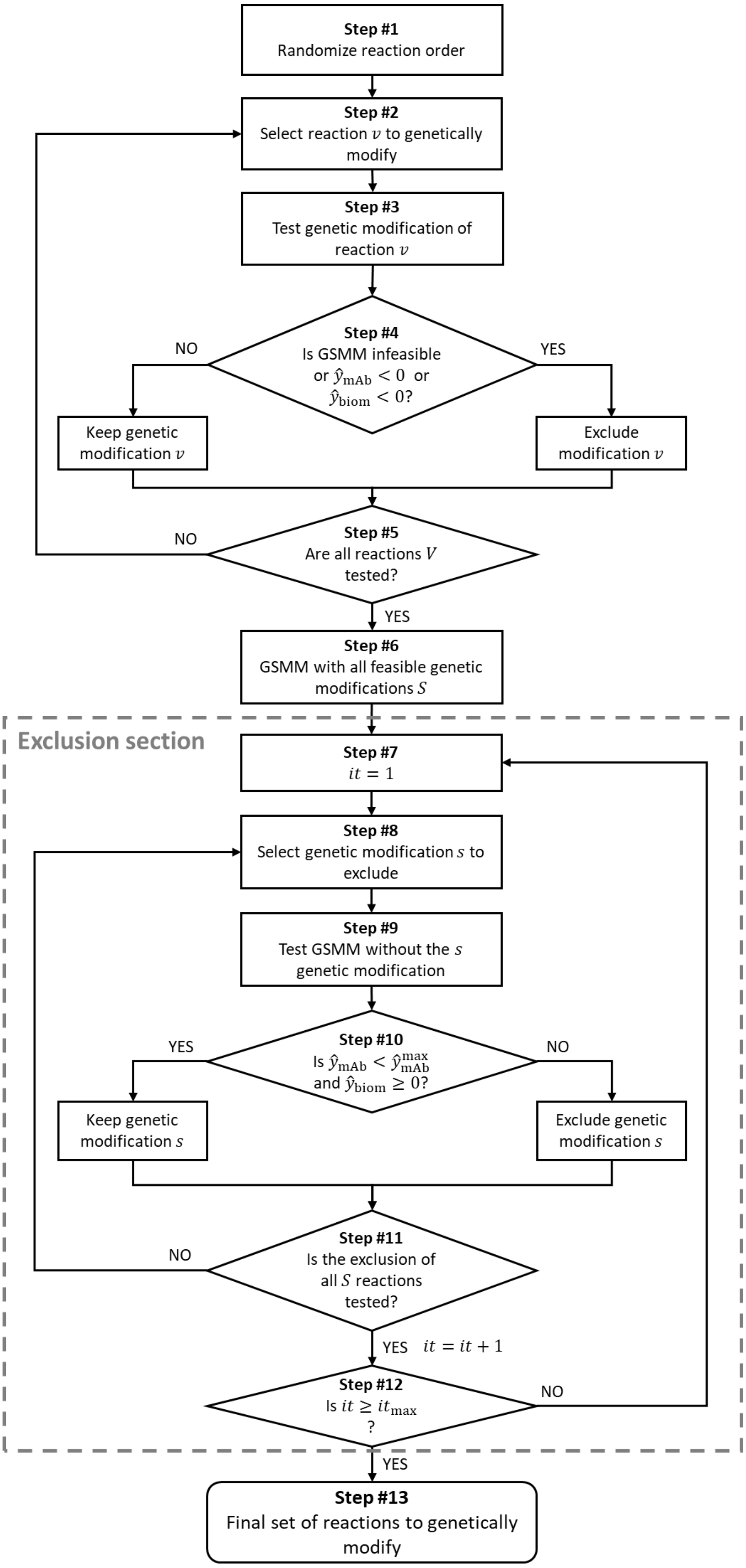

Step #1
Randomize reaction order
Step #2
Select reaction v to genetically modify
Step #3
Test genetic modification of reaction v
Step #4
Is GSMM infeasible or $\hat{y}_{\text{mAb}} < 0$ or $\hat{y}_{\text{biom}} < 0$?
NO
YES
Keep genetic modification v
Exclude modification v
Step #5
Are all reactions V tested?
NO
YES
Step #6
GSMM with all feasible genetic modifications S
Exclusion section
Step #7
$it = 1$
Step #8
Select genetic modification s to exclude
Step #9
Test GSMM without the s genetic modification
Step #10
Is $\hat{y}_{\text{mAb}} < \hat{y}_{\text{mAb}}^{\text{max}}$ and $\hat{y}_{\text{biom}} \geq 0$?
YES
NO
Keep genetic modification s
Exclude genetic modification s
Step #11
Is the exclusion of all S reactions tested?
NO
YES $it = it + 1$
Step #12
Is $it \geq it_{\text{max}}$?
NO
YES
Step #13
Final set of reactions to genetically modify

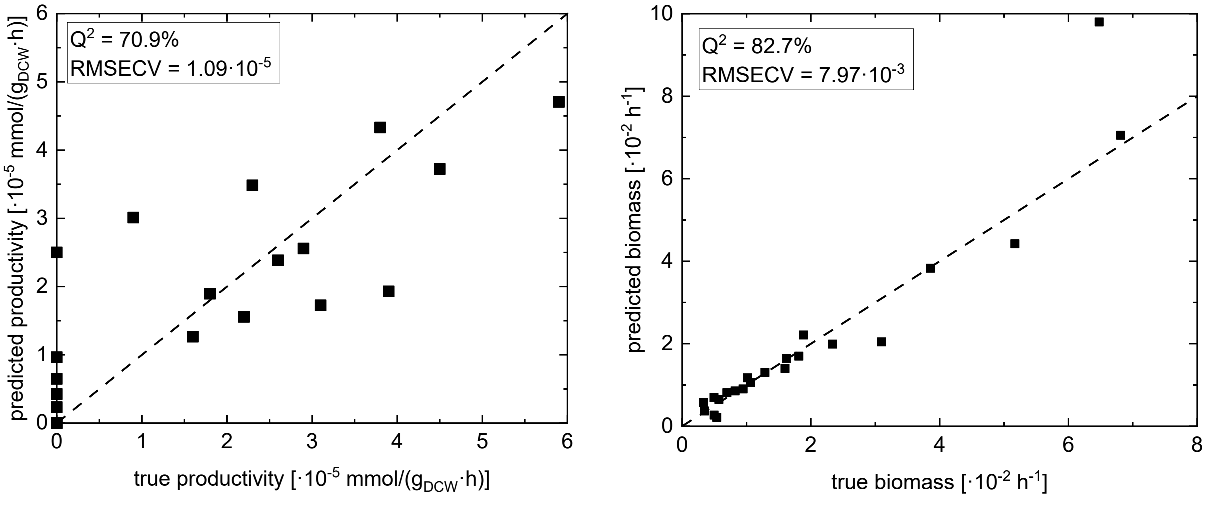
Figure 8.2 *Algorithm for the identification of the essential intracellular reactions to modify to achieve the desired flux distribution.*

8.2. Prediction of biomass and productivity

In this Section the outcomes of the prediction of specific mAb productivity and biomass growth rate from the intracellular fluxes are presented and critically discussed. This identifies the association between the phenotypes (i.e., specific productivity and biomass) and the flux value of each intracellular reaction and is used in the following Section to identify the required intracellular flux values that provide the desired phenotypes.

To this purpose, a PLS model is bult to predict biomass and productivity **Y** from the intracellular fluxes **X**. The reactions involved in the direct pathway from protein translation to mAb secretion are not included in **X**, as previously explained, because the fluxes are almost the same as the productivity and genetic modifications in this pathway are obvious ways to improve the productivity in GSMM, while this might not be possible in real cells.

A PLS model with 3 LVs describes 94.5% of **Y** variability through 67.1% of **X** variability. Describing a high percentage of **X** variability is essential for the further inversion, because it increases the accuracy in the estimation of the intracellular fluxes $\mathbf{x}_{\mathrm{NEW}}$ associated with the desired phenotype. In this case, the amount of **X** variability captured by the 3 LVs is considered enough for the model inversion. The addition of further LVs only increases the explained **X** variability by few percentage points but causes a substantial reduction in the prediction performance and robustness of the model, making it unsuitable for inversion.



(a) (b)

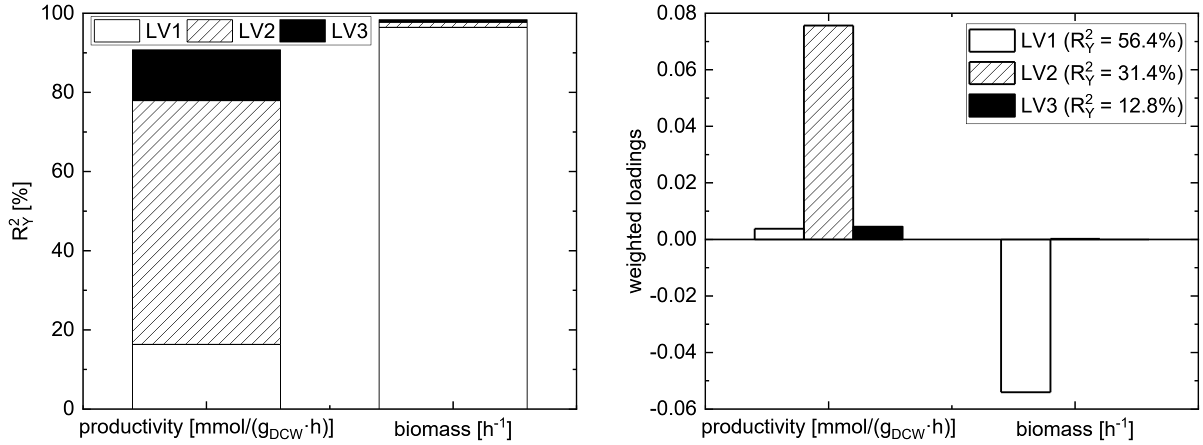
Figure 8.3 *PLS cross-validation parity plot: (a) productivity, and (b) biomass.*

The model is cross validated to assess its performance and robustness through a leave-one-out cross-validation procedure. In cross-validation, the model achieves a $Q^2 = 70.9\%$ for productivity and $Q^2 = 82.7\%$ for biomass, showing a slightly lower accuracy in the specific productivity prediction probably due to the complex correlations between the highly interconnected metabolic network and mAb productivity. This prediction performance is satisfactory despite the much larger number of variables in **X** (i.e., intracellular fluxes) than observations (i.e., cell lines). For this reason and to show the model that will be further inverted, variable selection is not applied in this work to improve performance nor interpretability.

In the leave-one-out cross-validation, the predictions for the *early* cell line (Table 8.1) are extremely inaccurate and much greater than the real value. This is due to diversity of *early* intracellular fluxes with respect to all other cell lines, resulting in diagnostics (i.e., T^2 and SPE) largely outside the confidence limits. This indicates that the *early* cell line is badly predicted because its specific conditions are underrepresented in the dataset making it unpredictable when it is left out.

The parity plot in cross-validation is reported in Figure 8.3 for productivity and biomass. In this plot, the closer a sample is to the diagonal (the dashed line) the higher the prediction accuracy since predicted and real values approach each other. For productivity (Figure 8.3a), cell lines are scattered around the diagonal with someone closer than others. However, for non-productive cell lines (i.e., with zero productivity) the predictions are rather scattered around 0, as results of the linear relationship captured by the PLS model. For biomass (Figure 8.3b) instead, cell lines are scattered very close to the diagonal, indicating that predicted values are very close to the observed ones in most of the experiments. In this case, the biomass of *CLP* cell line (Table 8.1) is overpredicted, indicating that *CLP* has a degree of specificity in its intracellular fluxes that differentiate it from other cell lines making it difficult to predict. Specifically, *CLP* cell line has an unusually high flux of the reaction *PYK7* (i.e., pyruvate kinase), probably due to the high growth rate.

These results show that the model captures the relationship between intracellular fluxes and the studied phenotypes, resulting, in most cases, in reasonably accurate predictions for unknown samples. For this reason, the PLS model can be interpreted and, furthermore, inverted to identify the intracellular flux values required to obtain a cell with the desired phenotype.



(a) (b)

Figure 8.4 *PLS model interpretation: (a) response explained variance and (b) response weighted loadings.*

The relevant parameters for PLS model interpretation, such as the response explained variance and the weighted loadings, are reported in Figure 8.4. According to the response explained variance (Figure 8.4a), the first LV captures almost the entire variability of the biomass and less than 20% of productivity variability. The second LV, instead, capture more than 60% of productivity variability and very small portion of biomass variability. Finally, the third LV captures ~12% of productivity variability and a negligible amount of biomass variability. Accordingly, the first LV captures most of the biomass behavior and the associated productivity one, while the second and third LVs capture the independent behavior of productivity.

The response loadings (Figure 8.4b) describe the correlation structure between the two phenotypes captured by the PLS model. In particular, the first LV captures the anticorrelation between biomass and productivity, meaning that cells with high productivity tend to have a lower growth rate. This is reasonable because part of the cell resources is diverted from biomass to protein production. The second and third LVs capture the independent variation of cell productivity with respect to biomass production. The inspection of the **X** weights associated to each LV (i.e., the intracellular fluxes that are most related to both productivity and biomass) can give information of the intracellular fluxes mostly related to these phenomena. This will be exploited in the model inversion phase to appropriately manipulate the intracellular fluxes to obtain the desired phenotypes.

8.3 Identification of the optimal flux distribution

In this Section, the results of the PLS model inversion are presented. This determines the productivity and biomass that can be achieved through genetic modification according to the correlation structure between fluxes and cell productivity and biomass explained by the PLS model. The PLS model built in Section 8.2 is inverted as explained in Section 8.1.4. In order to obtain the flux distribution associated with the desired specific productivity and biomass (i.e., optimal flux distribution) for each cell lines, the inversion of the PLS model is repeated considering the extracellular metabolite uptake rates of each single cell line. These uptakes set the extracellular constraints in the inversion as explained in Section 8.1.4.

The optimal biomass and productivity are compared with the original ones in Table 8.2. Here, the values are reported for each available cell line. The PLS inversion provided increased productivity in all cell lines, as can be easily observed in Table 8.2. For non-producer samples, a minimum productivity of $1.99 \cdot 10^{-5}\ mmol/(g_{DCW}\ h)$ is required during the inversion. In a real case, non-producer cells are not equipped for mAb production, making production impossible without transfection. However, in this synthetic analysis, we consider the non-producer case in such a way as to observe the genetic modifications that increase productivity in the corresponding producer cell with similar metabolic and extracellular characteristics.

Table 8.2 *Base biomass and productivity and ones obtained through the inversion of PLS model.*

Sample	Biomass $[1/h]$	Optimal biomass $[1/h]$	Productivity $[mmol/(g_{DCW}\ h)]$	Optimal productivity $[mmol/(g_{DCW}\ h)]$
SV	0.0102	0.0116	0	$1.99 \cdot 10^{-5}$
SVGS	0.0129	0.0159	0	$1.99 \cdot 10^{-5}$
SVM1	0.0107	0.0143	$2.95 \cdot 10^{-5}$	$4.42 \cdot 10^{-5}$
SVM2	0.0070	0.0122	$1.84 \cdot 10^{-5}$	$3.67 \cdot 10^{-5}$
SVM3	0.0095	0.0188	$4.49 \cdot 10^{-5}$	$6.73 \cdot 10^{-5}$
SVM4	0.0233	0.0211	$2.26 \cdot 10^{-5}$	$4.53 \cdot 10^{-5}$
BCL2	0.0386	0.0347	0	$1.99 \cdot 10^{-5}$
BCL2-M1	0.0162	0.0245	$5.87 \cdot 10^{-5}$	$8.80 \cdot 10^{-5}$
BCL2-M2	0.0160	0.0146	$0.89 \cdot 10^{-5}$	$2.81 \cdot 10^{-5}$
BCL2-M3	0.0181	0.0239	$6.29 \cdot 10^{-5}$	$9.43 \cdot 10^{-5}$
BCL2-M4	0.0188	0.0195	$3.79 \cdot 10^{-5}$	$5.69 \cdot 10^{-5}$
fed-batch	0.0033	0.0194	$3.07 \cdot 10^{-5}$	$4.61 \cdot 10^{-5}$
perfusion	0.0057	0.0098	$1.57 \cdot 10^{-5}$	$3.13 \cdot 10^{-5}$
early	0.0221	0.0207	$1.50 \cdot 10^{-5}$	$3.00 \cdot 10^{-5}$
late	0.0310	0.0279	$2.59 \cdot 10^{-5}$	$3.89 \cdot 10^{-5}$
stat	0.0082	0.0152	$3.94 \cdot 10^{-5}$	$5.91 \cdot 10^{-5}$
decline	0.0050	0.0121	$2.17 \cdot 10^{-5}$	$4.35 \cdot 10^{-5}$
CLP	0.0648	0.0583	0	$1.99 \cdot 10^{-5}$
LELP	0.0681	0.0613	0	$1.99 \cdot 10^{-5}$
HELP	0.0517	0.0465	0	$1.99 \cdot 10^{-5}$
CLC	0.0054	0.0077	0	$1.99 \cdot 10^{-5}$
LELC	0.0050	0.0075	0	$1.99 \cdot 10^{-5}$
HELC	0.0035	0.0073	0	$1.99 \cdot 10^{-5}$

The biomass estimated in PLS inversion (Table 8.2) increases for some cell lines, such as *SV*, *SVGS*, *SVM1* to *M3*, *BCL2-M1* to *M4*, *fed-batch*, *perfusion*, *stat*, *decline*, *CLC*, *LELC*, and *HELC*, and decreases in others, such as *SVM4*, *BCL2*, *early*, *late*, *CLP*, *LELP*, and *HELP*. This estimated biomass increase is not likely to happen in the GSMM and in a real cell, because cells need to divert resources from growth to other functions to improve their productivity. In this case, the biomass increase is due to the specific culture condition of each cell line (i.e., extracellular metabolite uptake rates) and the insufficiently strong anticorrelation between productivity and biomass captured by the PLS model in the available experiments. However, the recalibration of the PLS model introducing genetically engineered cell lines in the calibration data would improve the goodness and generalizability of the model. In this way, the model will lean the strongest anticorrelation between productivity and biomass.

The PLS inversion provides the new intracellular fluxes $\mathbf{x}_{\text{NEW}}$ that are required to obtain an increased productivity in each cell line. This information will be further used to identify the smallest set of genetic modifications that will improve mAb productivity in GSMMs.

8.4 Genetic modifications

In this Section, the genetic modifications suggested by the proposed machine learning approach are presented and analyzed. The suggested genetic modifications are categorized as *i*) metabolic modifications, and *ii*) secretory modifications.

To this purpose, the sets of intracellular fluxes $\mathbf{x}_{\text{NEW}}$ (Section 8.3) are fed to the developed algorithm to identify the smallest set of genetic modifications required to improve mAb productivity (Section 8.1.5). For each experiment, 30 scenarios are obtained by running the developed algorithm multiple times with a different order for testing the reactions and scaling coefficient k_{reg}.

8.4.1 Metabolic genetic modifications

In this Section genetic modifications that improve mAb productivity through the regulation of cell metabolism are presented and discussed, indicating the genes that should be regulated and the mechanism behind the increased mAb productivity.

The developed algorithm suggests 30 different genetic modification scenarios for each cell line, one for each run of the algorithm, containing repetitions and slight variation of the same main modifications. Hence, only the most common and potentially applicable genetic modifications are presented in this Section.

This analysis highlights that the regulation of amino acid metabolism, specifically of *L-valine* and *L-tryptophan*, which are building blocks of proteins and mAbs, increases cell productivity. The metabolism of some amino acids, such as *L-leucine* and *L-valine*, has been previously observed to be related to cell productivity, as their metabolism is typically downregulated in high productive cells (Huang & Yoon, 2020b).

8.4.1.1 Valine metabolism

The proposed machine learning strategy suggests that the metabolism of *L-valine* is associated with mAb productivity, and the regulation of this metabolism is a possible way to increase mAb productivity.

The *L-valine* related genetic modifications are suggested for the *CLP* cell line (Table 8.1). The phenotypes observed in both the base case and the genetically engineered one are reported in Table 8.3a. The genetic modifications allow a non-producer cell to achieve a substantial production, even larger than the one predicted by the PLS model inversion (Table 8.2). It is worth noticing that in the non-producer case, we are observing the impact of genetic modifications on a normally productive cell with similar metabolic conditions.

Table 8.3 *Genetic modification scenario for CLP cell line: (a) comparison of biomass and productivity achieved by the base case and the genetically engineered cell; (b) list of genetic modifications improving the mAb productivity.*

(a)

Condition	Biomass $[1/h]$	Productivity $[mmol/(g_{DCW}\,h)]$
Base case	0.0648	0
Genetically engineered	0.0151	$7.41 \cdot 10^{-5}$

(b)

Reaction	Short name	Regulation	Regulation coefficient
valine transaminase	VALTA_f	downregulate	0.73
glutaryl-CoA dehydrogenase	GLUTCOADHm	upregulate	1.01
2-aminomuconate reductase	AMCOXO	downregulate	0.68

The suggested reactions, their regulations and regulation coefficients are reported in Table 8.3b. The reaction *VALTA_f* is directly related to protein secretion, while the other reactions (i.e., *GLUTCOADHm* and *AMCOXO*) reduce cell growth allowing the reallocation of metabolic resources. In fact, *GLUTCOADHm* and *AMCOXO* reactions produce a ~50% reduction in the biomass when regulated alone, while *VALTA_f* produces only a 1% biomass reduction when regulated alone.

A section of the relevant metabolic network of the genetically engineered cell is shown in Figure 8.5 and 8.6. The downregulation of *VALTA_f* decreases the amount *L-valine* (*val_L[c]*) metabolized to *L-glutamate* (*glu_L[c]*), thus increasing the availability of *L-valine* for protein translation (*ICproduct_TRANSLATION_protein*). The increased availability of *L-valine* results in larger protein translation and mAbs synthesis when coupled with a reduction in the growth rate. The upregulation of *GLUTCOADHm* (Figure 8.6) increases the amount of *CoA* (*coa[m]*) diverted to *GLCOASYNT* and reduces the flux through *PDHbr*. Finally, the downregulation of *AMCOXO* reduces the overall flux through the pathway, which is connected to *2-Oxoglutarate* (*akg[c]*), an important link between the TCA cycle and the metabolism of amino acids.

The suggested genetic modifications can be achieved by regulating the expression of the genes associated with the identified reactions. In particular, *VALTA_f* is associated to the *Bcat1* gene and *GLUTCOADHm* to the *Gcdh* gene, while *AMCOXO* has no known associated gene. Fortunately, the downregulation of an upstream reaction *PCLAD*, associated to the *Acmsd* gene, produces the same increased productivity as the regulation of *AMCOXO*.

The suggested genes are not identified as essential in previous studies (Xiong et al., 2021) and an essentiality analysis[5] run on the GSMM. For this reason, this suggested genetic modification of the *L-valine* metabolism should be feasible in a real cell.

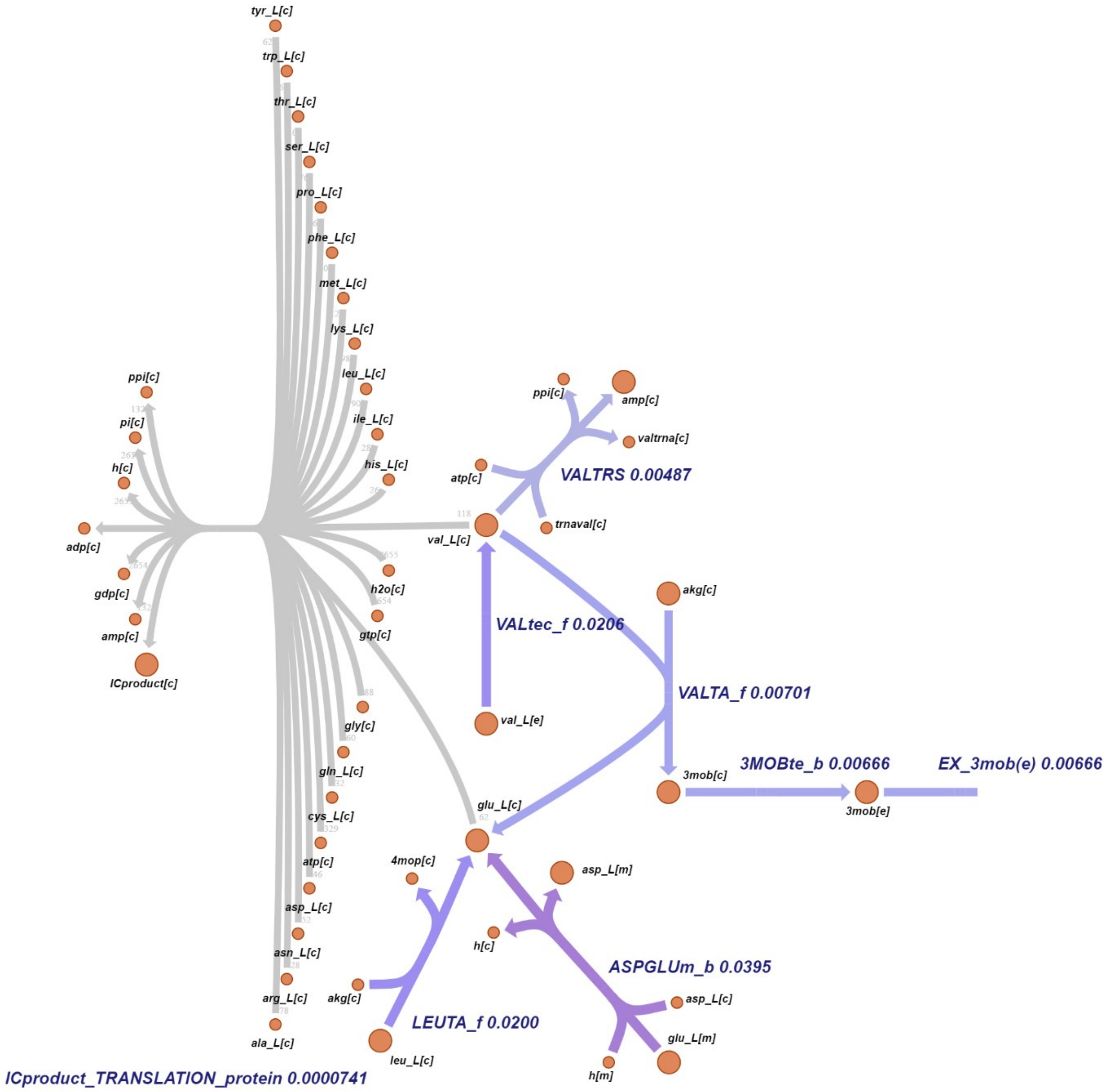


Figure 8.5 *Metabolic network section of CLP cell line related to reaction VALTA_f. Size and color of the arrows is connected to the intracellular flux value, which is indicated after the reaction name. Large orange circles refer to primary metabolite, while small orange circles refer to secondary metabolites.*

[5] In essentiality analysis, the GSMM is used to assess if the knockout of a gene does not prevent cell growth. Specifically, a gene in essential if its knockout prevent cell growth, while a gene is not essential if its knockout does not prevent cell growth.

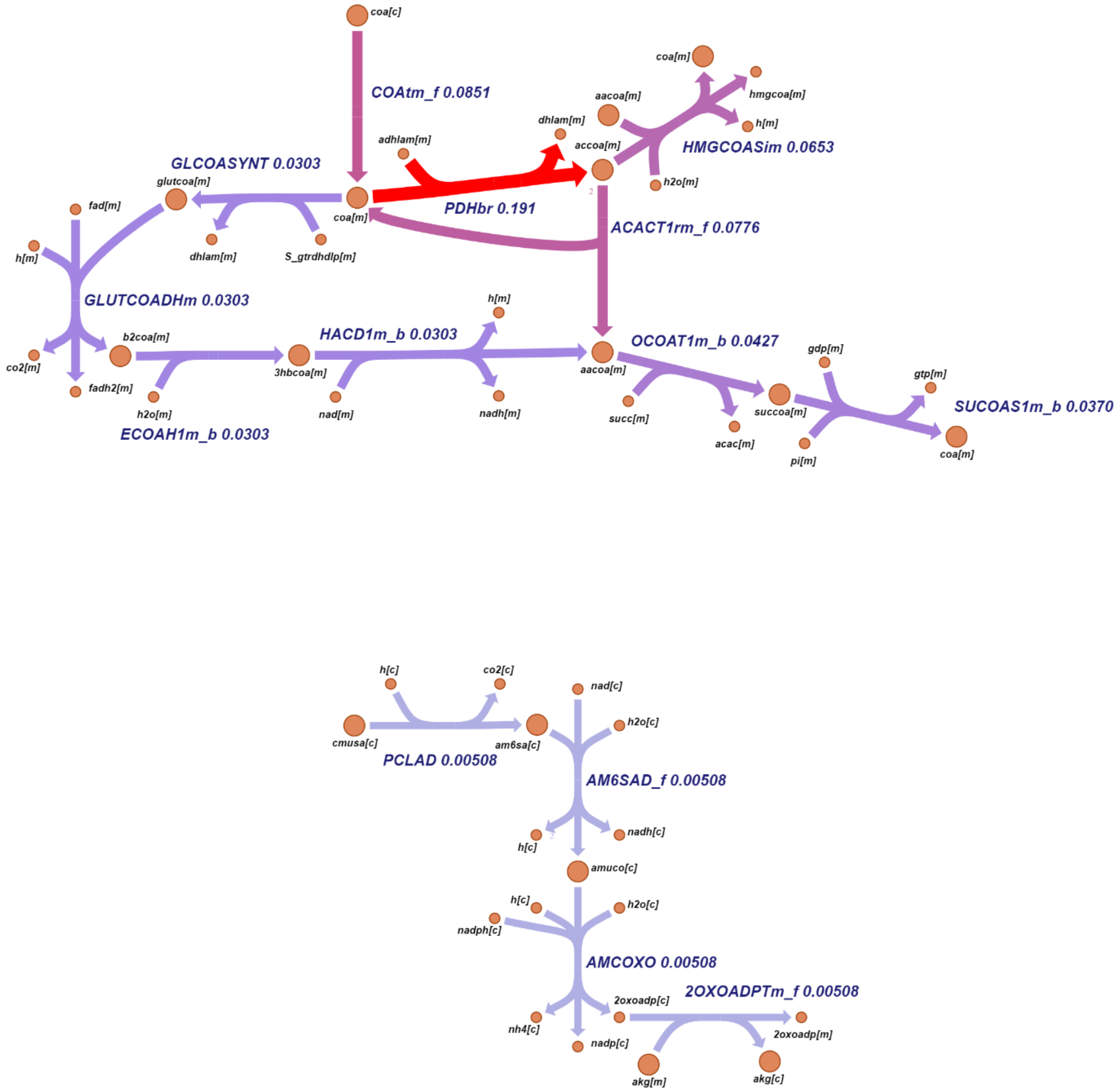


Figure 8.6 *Metabolic network section of CLP cell line related to reaction GLUTCOADHm and AMCOXO. Size and color of the arrows is connected to the intracellular flux value, which is indicated after the reaction name. Large orange circles refer to primary metabolite, while small orange circles refer to secondary metabolites.*

8.4.1.2 Tryptophan metabolism

The proposed machine learning strategy suggests that the metabolism of *L-tryptophan* is associated with mAb productivity, and the regulation of this metabolism is a possible way to increase mAb productivity.

The *L-tryptophan* related genetic modifications are suggested for the *BCL2-M4* cell line (Table 8.1). The phenotypes observed in both the base case and the genetically engineered one are reported in Table 8.4a. The suggested genetic modifications almost double mAb productivity at the price of reducing by half the cell growth rate. In this case, the flux regulations suggested by the PLS model inversion produce a substantially larger productivity in the GSMM than the predicted one ($7.24 \cdot 10^{-5}$ vs. $5.69 \cdot 10^{-5}$ $mmol/(g_{DCW}\, h)$). The large difference between the PLS predicted and GSMM observed productivity is probably due to underrepresentation of such highly productive cells in the calibration dataset. For this reason, the recalibration of the PLS model including genetically modified sample should improve the accuracy of the phenotype prediction during PLS inversion.

The suggested reactions, their regulations and regulation coefficients are reported in Table 8.4b. The downregulation of *TRPO2* alone produces an increase in cell productivity and a reduction in the growth rate.

Table 8.4 *Genetic modification scenario for BCL2-M4 cell line: (a) comparison of biomass and productivity achieved by the base case and the genetically engineered cell; (b) list of genetic modifications improving the mAb productivity.*

(a)

Condition	Biomass $[1/h]$	Productivity $[mmol/(g_{DCW}\, h)]$
Base case	0.0188	$3.79 \cdot 10^{-5}$
Genetically engineered	0.0089	$7.24 \cdot 10^{-5}$

(b)

Reaction	Short name	Regulation	Regulation coefficient
L-Tryptophanoxygen 2,3-oxidoreductase decyclizing	TRPO2	downregulate	0.79

The relevant metabolic network of the base case and genetically engineered cell is shown in Figure 8.7 and 8.8. Specifically, the downregulation of *TRPO2* (Figure 8.7) reduces the amount of *L-tryptophan* that is converted to *L-kynurenine* (*Lkynr[c]*), thus increasing the availability of *L-tryptophan* for protein translation. The increased availability of *L-tryptophan* results in larger protein translation and mAbs synthesis. This specific modification also decreases the flux through the reaction *TRPTRS*, which reduces the growth rate by limiting the production of a biomass component (*prot_prod[c]*). This can be observed by comparing the genetically engineered cell (Figure 8.7) and the base case (Figure 8.8). Finally, to achieve production, the increased availability of *L-tryptophan* is additionally supported by an increase uptake of *L-tryptophan* from the culture media. Accordingly, this genetic modification must be associated with an appropriate formulation of the culture medium.

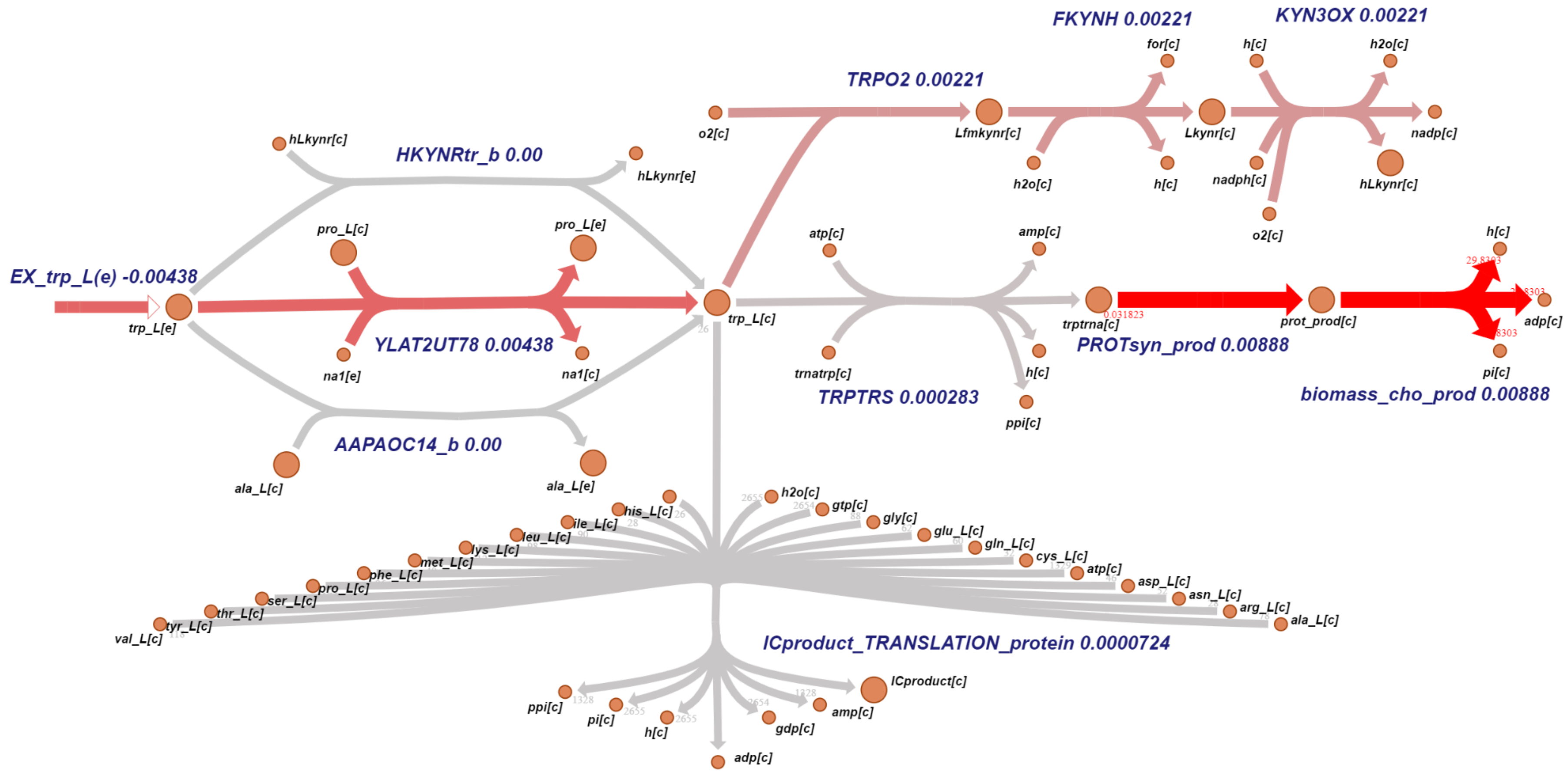


Figure 8.7 *Metabolic network section of BCL2-M4 cell related to reaction TRPO2: genetically engineered cell. Size and color of the arrows is connected to the intracellular flux value, which is indicated after the reaction name. Large orange circles refer to primary metabolite, while small orange circles refer to secondary metabolites.*

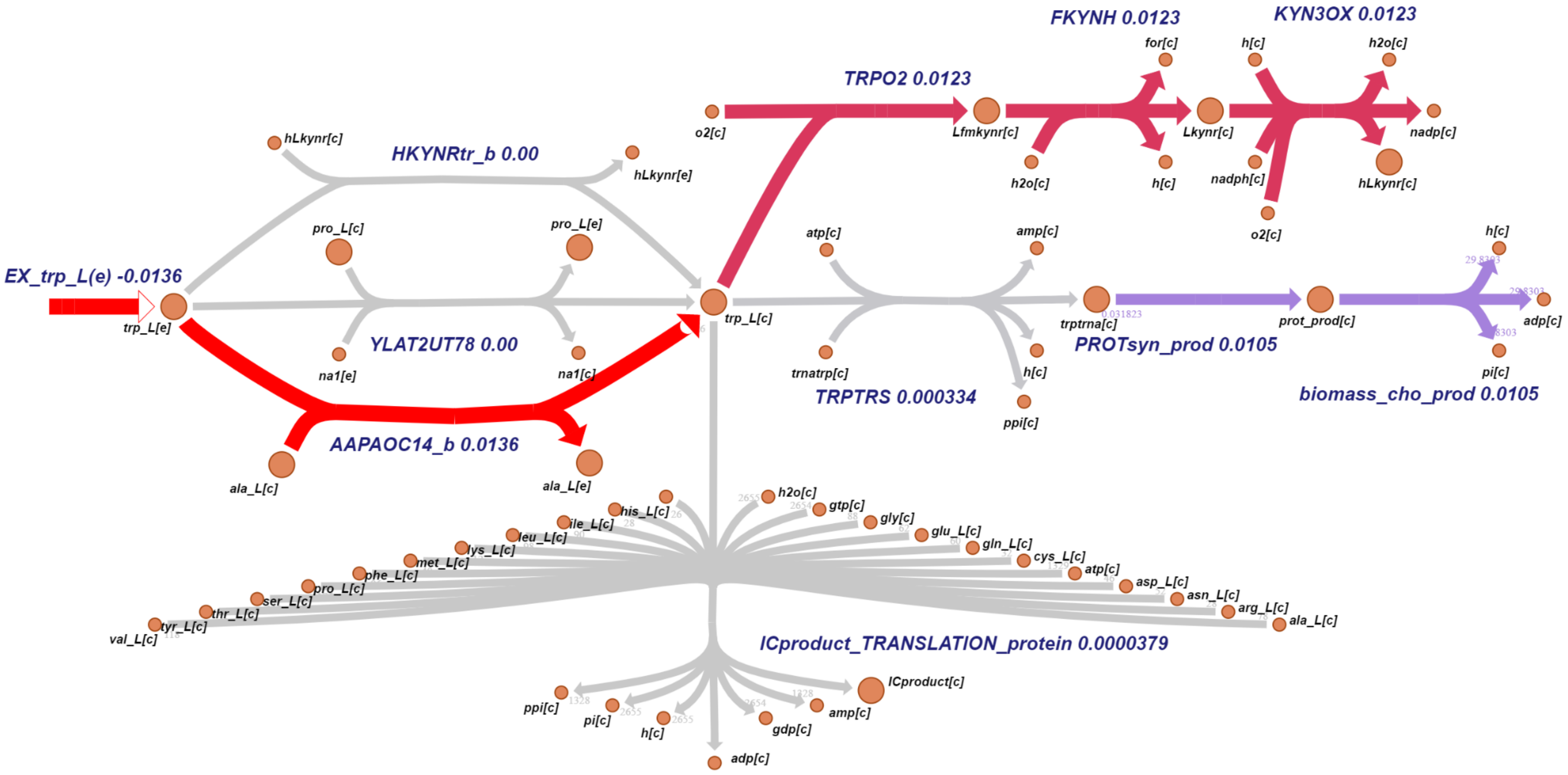


Figure 8.8 *Metabolic network section of BCL2-M4 cell related to reaction TRPO2: original cell. Size and color of the arrows is connected to the intracellular flux value, which is indicated after the reaction name. Large orange circles refer to primary metabolite, while small orange circles refer to secondary metabolites.*

The suggested genetic modifications can be achieved by regulating the expression of the genes associated with the identified reactions. In particular, *TRPO2* is associated with one or multiple genes among *Ido1*, *Ido2*, and *Tdo2*. Because of that, the genetic modification of the reaction *TRPO2* is complex and it unlikely to be feasible in the real cell. However, the two downstream reactions *FKYNH* and *KYN3OX* produce, when downregulated, the same effect as the *TRPO2* modification and are associated to single genes. Specifically, *FKYNH* is associated with gene *Afmid*, and *KYN3OX* with gene *Kmo*.

The genes associated to the suggested genetic modification are marked as essential by the essentiality analysis run on the GSMM, but they have not been identified as essential in previous *in vitro* studies (Xiong et al., 2021). Despite the fact that these reactions might be essential, the downregulation required to achieve increased production is small and it would not produce any negative effects on cells.

8.4.2 Secretory pathway genetic modifications

In this Section genetic modifications that improve mAb productivity through the regulation of the cell secretory pathway are presented and discussed, indicating the genes whose expression should be regulated and the mechanism behind the increased mAb productivity.

Similarly to the metabolic modifications, these sets of genetic modifications stem from the 30 scenarios obtained for each cell line. Only the most common and potentially applicable genetic modifications are presented in this Section.

The proposed machine learning strategy suggests that the pathway of *mannose* recirculation during early glycosylation is associated with mAb productivity, and the regulation of this pathway is a possible way to increase mAb productivity. These genetic modifications are suggested multiple times for almost all the available experiments. The modification suggested for the *BLC2-M1* cell line (Table 8.1) is presented here as a general explanation. The phenotypes observed in the base case and genetically engineered one are reported in Table 8.5a.

Table 8.5 *Genetic modification scenario for BCL2-M1 cell line: (a) comparison of biomass and productivity achieved by the base case and the genetically engineered cell; (b) list of genetic modifications improving the mAb productivity.*

(a)

Condition	**Biomass** $[1/h]$	**Productivity** $[mmol/(g_{DCW}\ h)]$
Base case	0.0162	$5.87 \cdot 10^{-5}$
Genetically engineered	0.0042	$9.84 \cdot 10^{-5}$

(b)

Reaction	**Short name**	**Regulation**	**Regulation coefficient**
Mannose efflux from Golgi apparatus	MANtg	upregulate	1.68

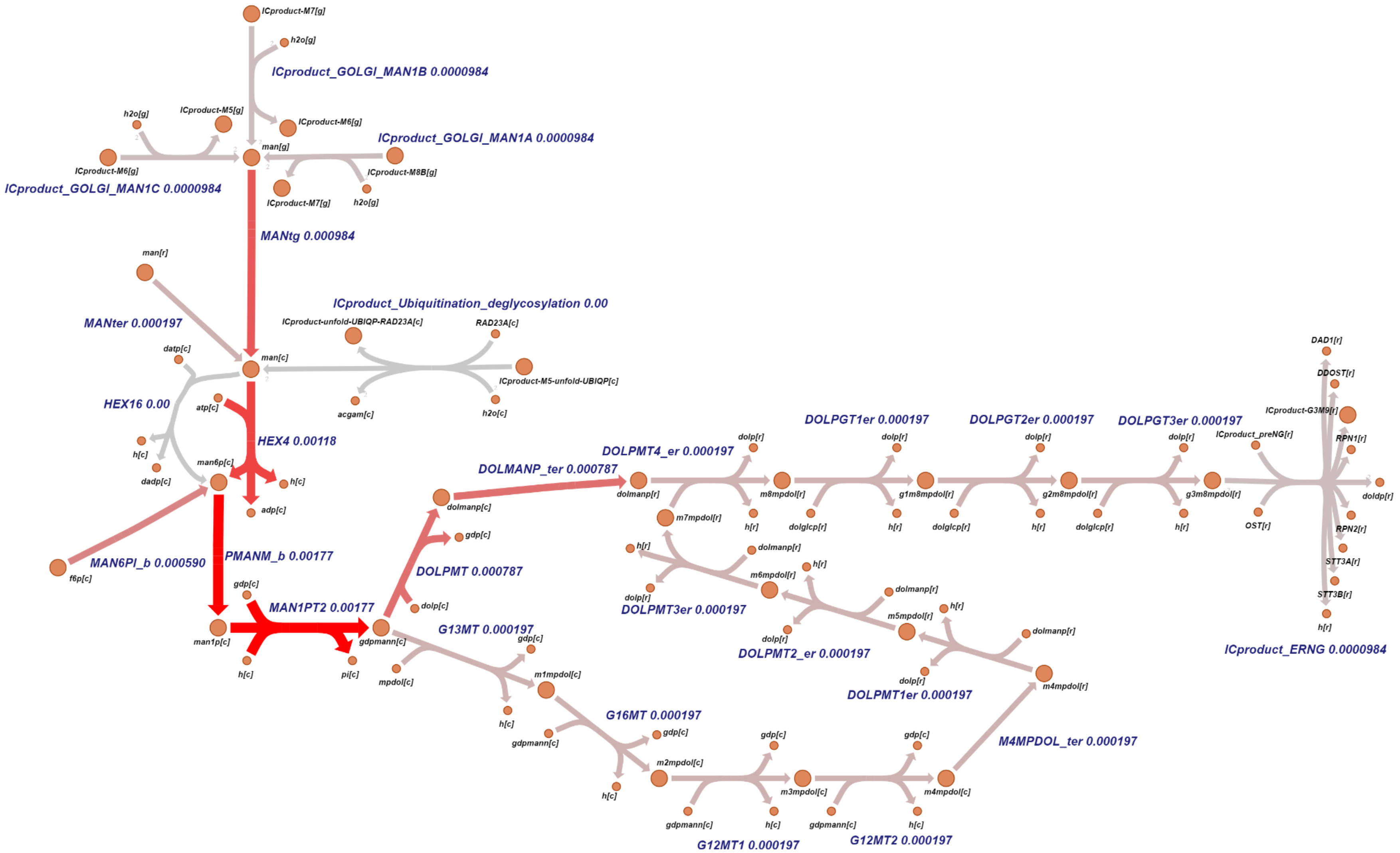


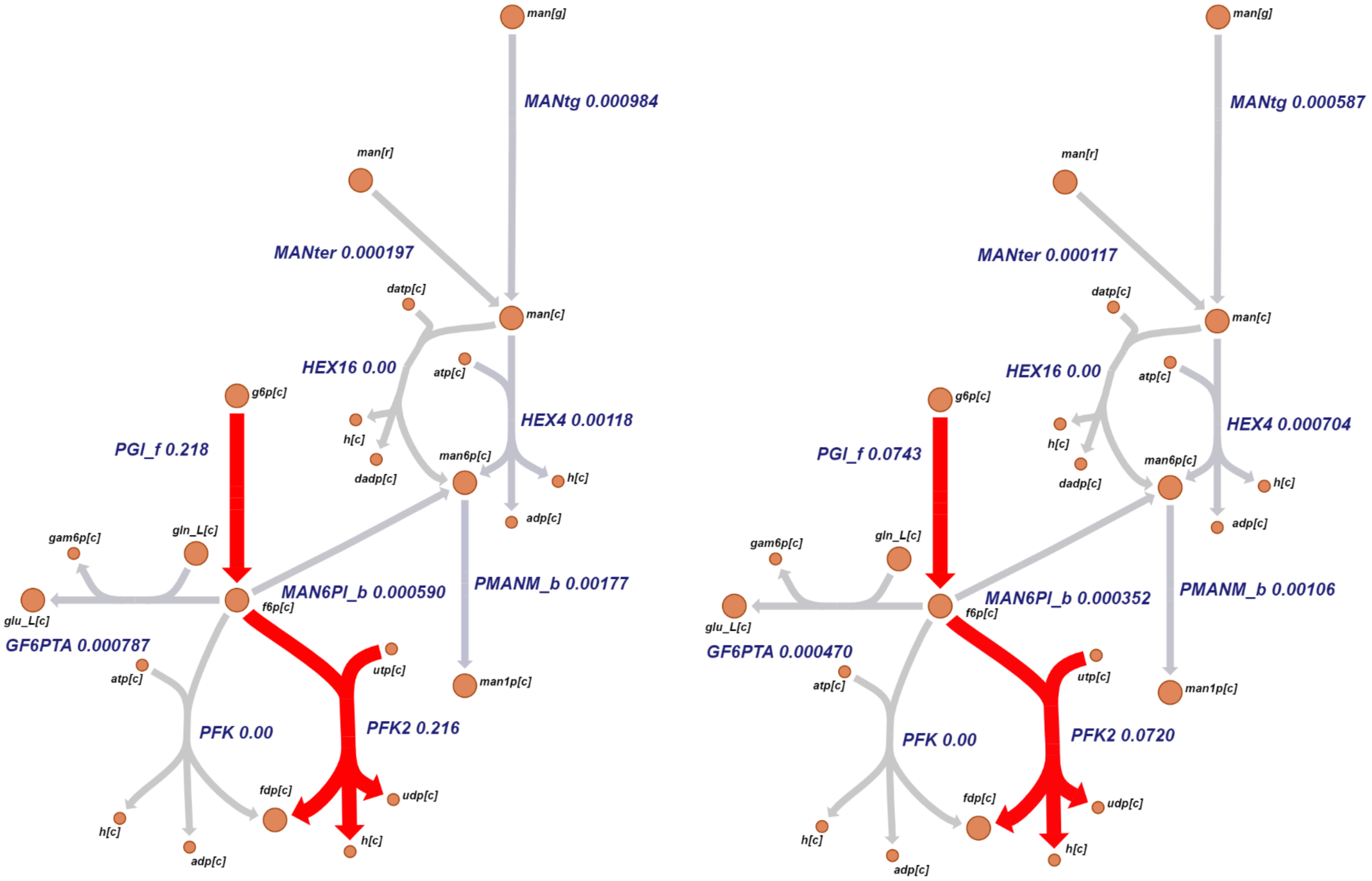
Figure 8.9 *Metabolic network section of the genetically engineered BCL2-M1cell related to reaction MANtg. Size and color of the arrows is connected to the intracellular flux value, which is indicated after the reaction name. Large orange circles refer to primary metabolite, while small orange circles refer to secondary metabolites.*

In this case, the suggested genetic modification increases cell productivity by 68%, being the strategy providing the highest mAb productivity among all experiments. Unfortunately, this is achieved with a considerable reduction in the growth rate (74%) due to resource reallocation.

The suggested reactions, their regulations and regulation coefficients are reported in Table 8.5b. Specifically, the machine learning strategy suggests upregulating the *MANtg* reaction, which involves the transport of *mannose* from the Golgi apparatus to the cell cytoplasm after *mannose* groups are detached before the last step of glycosylation from the high *mannose* glycans produced in the early stages of glycosylation (Sha et al., 2016). The relevant metabolic network of the base case and genetically engineered cell is shown in Figure 8.9. As said, the upregulation of *MANtg* increase the flux of *mannose* moving from Golgi apparatus to the cytoplasm, which results in an increased synthesis *mannose 6-phosphate* (*man6p[c]*), which leads to higher availability of this metabolite. *Mannose 6-phosphate* is required for the synthesis of high *mannose* glycans during the early stages of glycosylation in the Endoplasmic Reticulum; hence, its higher availability triggers a higher rate of glycan synthesis, which results in higher productivity. The increased productivity associated with the genetic modification of the *MANtg* reaction indicates that the cotranslational addition of the N-glycan block in the Endoplasmic Reticulum (*ICproduct_ERNG*) is a bottleneck for glycoprotein (such as mAbs) synthesis and secretion.

The *MANtg* reaction is involved not only in mAbs production, but also in the synthesis of all glycosylated proteins, which is not described by the GSMM. For this reason, the upregulation of *MANtg* would probably increase the overall protein production.

The increased productivity after *MANtg* upregulation could be the results of a mathematical artifact due to metabolic network inconsistencies (Sonnenschein et al., 2012). In fact, alternative routes, which might divert back the increased flux to the cell without leading in increased productivity, might still miss in the early glycosylation pathway involved in the *MANtg* genetic modification. In fact, the *mannose 6-phosphate* has no alternative route to be diverted back to cell metabolism, being either progressed to glycan production or lost in a demand reaction. Furthermore, glycosylation is assumed to be linearly correlated with mAb synthesis in the GSMM, which might not completely be true in real cells. Additional studies are required to better understand the relationship between this section of glycosylation and cell productivity.



(a) (b)

Figure 8.10 *Metabolic network section of BCL2-M1 cell related to reaction MANtg: (a) genetically engineered cell, and (b) original cell. Size and color of the arrows is connected to the intracellular flux value, which is indicated after the reaction name. Large orange circles refer to primary metabolite, while small orange circles refer to secondary metabolites.*

The upregulation of *MANtg* causes alone the diversion of cell resources from growth to production. One of the main effects is shown in Figure 8.10, where genetically engineered (Figure 8.10a) and original (Figure 8.10b) cells are compared. The genetically engineered cell shows a significantly larger glycolytic flux (*PGI_f* and *PFK2* reactions) than the original cell. In fact, reaction *PGI_f* and *PFK2* show a flux of 0.218 and 0.216 $mmol/(g_{\mathrm{DCW}} \cdot h)$, respectively, in the genetic engineered cell, while they show a flux of 0.0743 and 0.0720 $mmol/(g_{\mathrm{DCW}} \cdot h)$, respectively, in the original cell. Furthermore, despite the higher recirculation of *mannose*, a higher conversion rate of *fructose 6-phosphate* (*f6p[c]*) to *mannose 6-phosphate* (0.00059 vs. 0.000352 $mmol/(g_{DCW} \cdot h)$) is required to sustain the increased glycosylation rate in the Endoplasmic Reticulum.

A possible explanation for the reduced growth rate is connected to the increased *ATP* consumption for mAb synthesis and the increased flux through reaction consuming *ATP*, which reduce the energy available for cell growth.

Unfortunately, the suggested modification cannot be implemented since no known gene is associated with *MANtg*. For this reason, additional strategies to achieve the same objective must be found.

Many other reactions involved in the *mannose* recirculation pathway are suggested in several experiments. Among these, the upregulation of the reaction *G13MT* provides very similar phenotypes as the *MANtg* modification (biomass 0.0026 $1/h$ and productivity $9.20 \cdot 10^{-5}$ $mmol/(g_{DCW}\ h)$). As can be observed, both growth rate and productivity are slightly lower than the previous case but are still satisfactory. The suggested reactions, their regulations and regulation coefficients are reported in Table 8.6.

Table 8.6 *List of genetic modifications improving the specific productivity in BCL2-M1 cell.*

Reaction	Short name	Regulation	Regulation coefficient
Alpha-1,3-mannosyltransferase	G13MT	upregulate	2.75
Carbamoyl-phosphate synthase (glutamine-hydrolysing)	CBPS	downregulate	0.16

In this case, the regulation of a single reaction in the *mannose* recirculation pathway does not induce increased production because it is not able to consistently divert resources from growth to protein secretion. In fact, the reaction *G13MT* only achieves a 19% biomass reduction and no productivity when upregulated alone. Hence, it must be coupled with a growth-inhibiting reaction. In particular, the *CBPS* reaction is required to reduce growth, being able to reduce growth by 95% when downregulated.

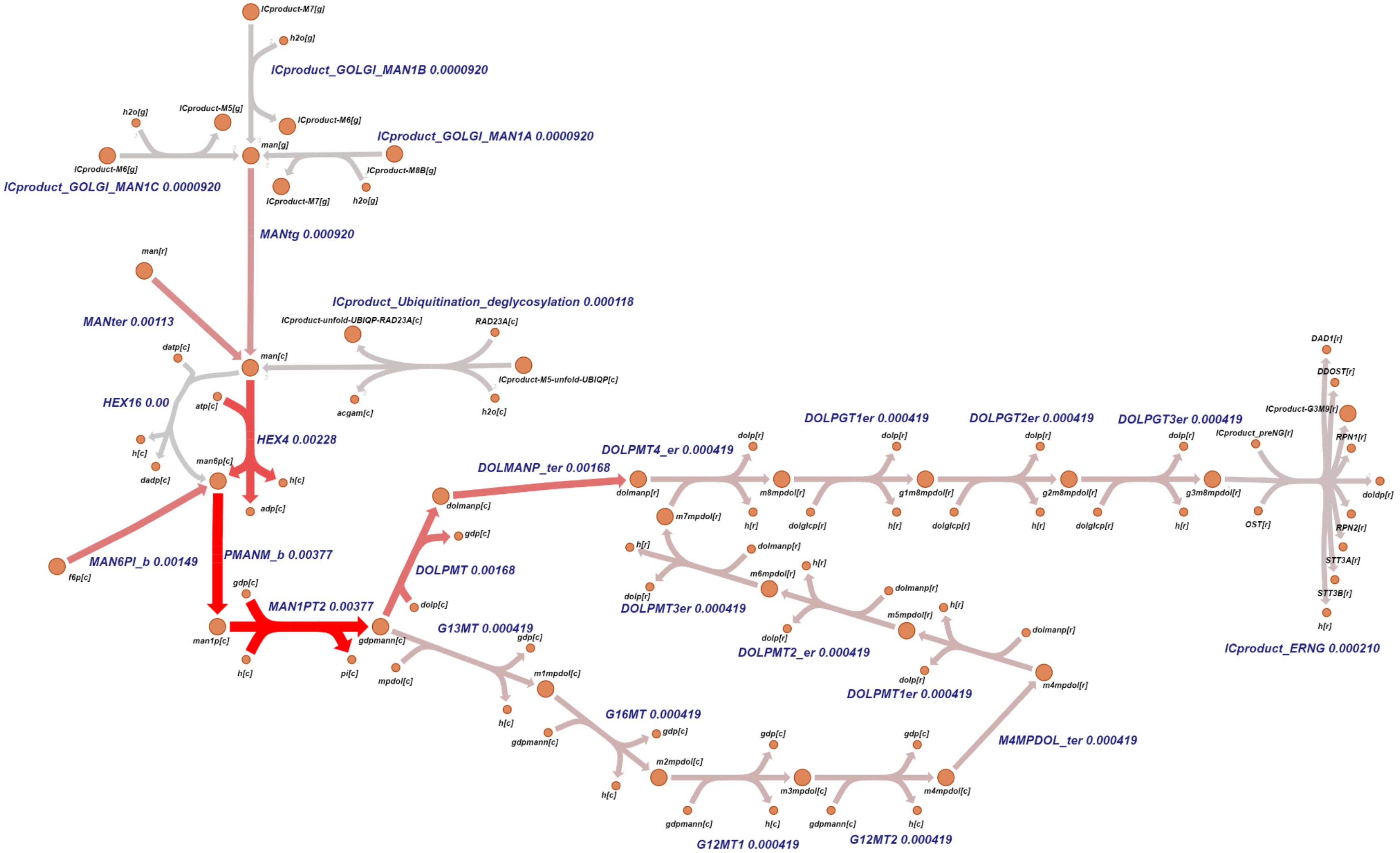


Figure 8.11 *Metabolic network section of the genetically engineered BCL2-M1 cell related to reaction GMT13. Size and color of the arrows is connected to the intracellular flux value, which is indicated after the reaction name. Large orange circles refer to primary metabolite, while small orange circles refer to secondary metabolites.*

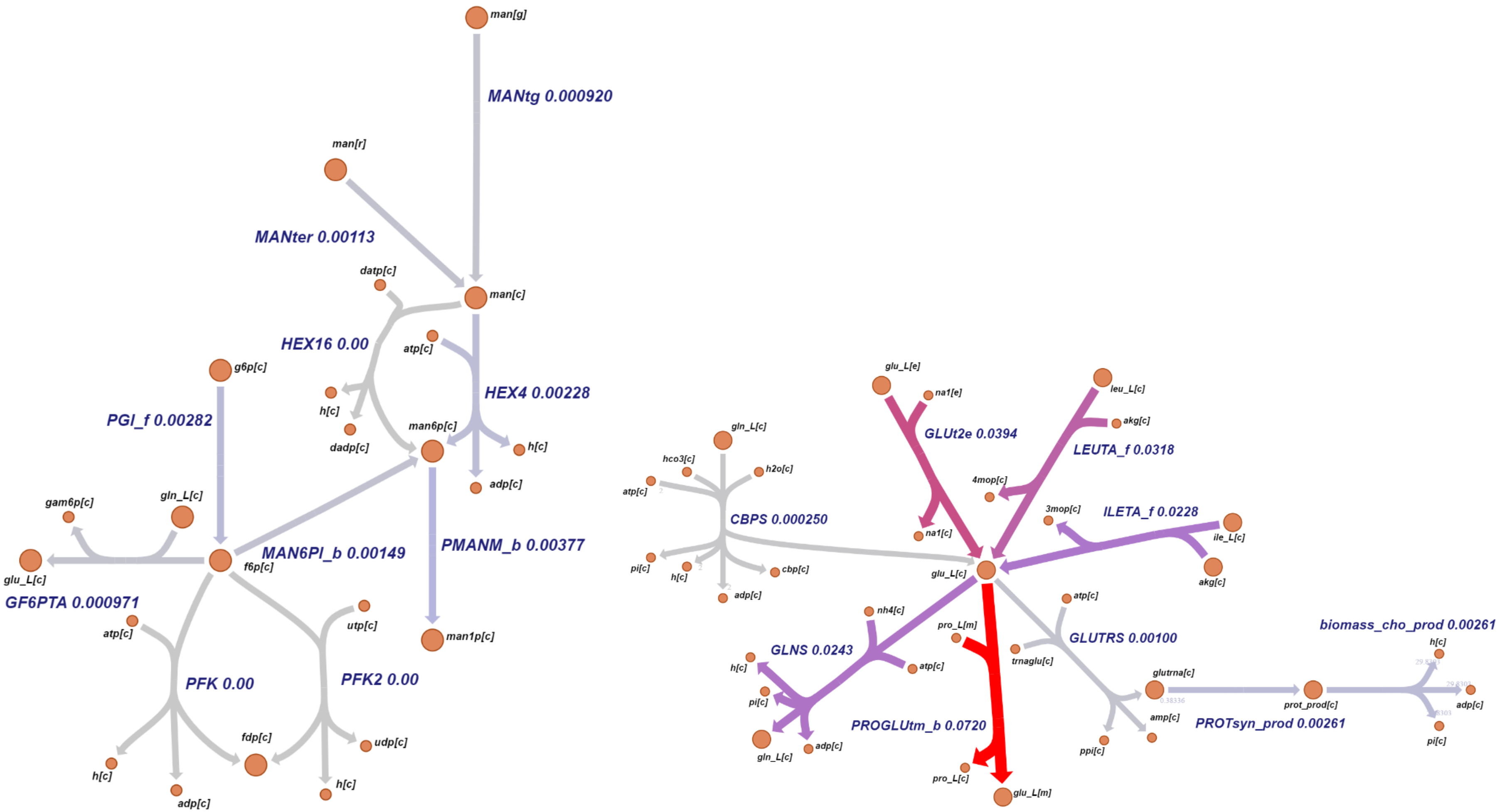
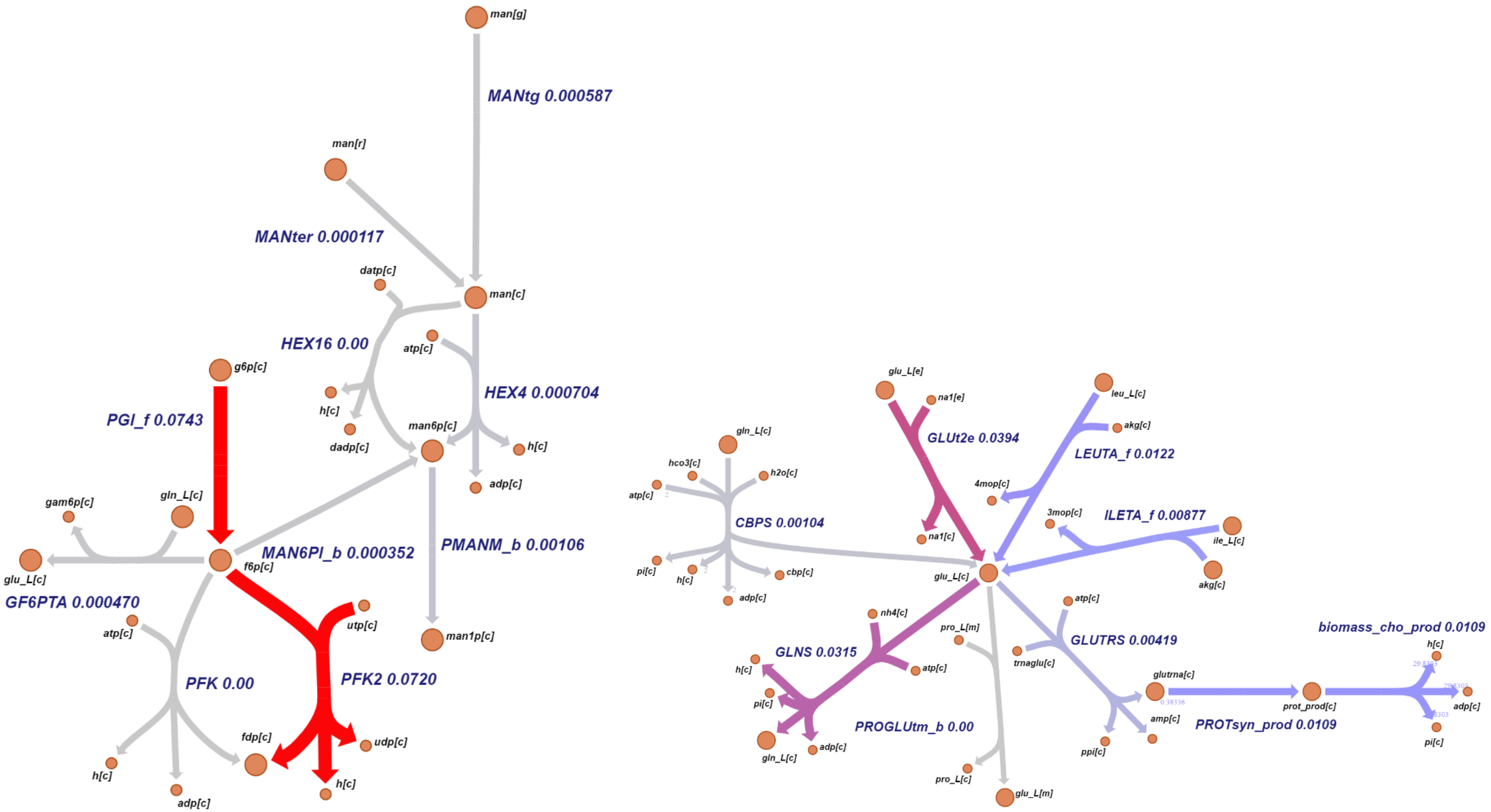


Figure 8.12 *Metabolic network section of BCL2-M1 cell related to reaction CBPS: genetically engineered cell. Size and color of the arrows is connected to the intracellular flux value, which is indicated after the reaction name. Large orange circles refer to primary metabolite, while small orange circles refer to secondary metabolites.*



(b)

Figure 8.13 *Metabolic network section of BCL2-M1 cell related to reaction CBPS: original cell. Size and color of the arrows is connected to the intracellular flux value, which is indicated after the reaction name. Large orange circles refer to primary metabolite, while small orange circles refer to secondary metabolites.*

The relevant metabolic network of both the base case and the genetically engineered cell is shown in Figure 8.11. The upregulation of *G13MT* works generate the same outcome than *MANtg* by increasing the glycosylation rate in the Endoplasmic Reticulum.

Differently form the upregulation of *MANtg*, the upregulation of *G13MT* does induce reduction in the glycolytic flux, with the reaction *PFK2* carrying no flux (comparing Figure 8.12 and 8.13), but requires the modification of other reactions to divert resources to mAb synthesis. The downregulation of *CBPS* reduces the conversion of *L-glutamine* to *L-glutamate* (Figure 8.11). This results in a different distribution of *L-glutamate* in the metabolic network, reducing the flux through *GLUTRS*, which limits the production of a biomass component and reduces cell growth.

The suggested genetic modification can be achieved by regulating the expression of the gene *Alg2* associated to *G13MT*, and the gene *Cad* associated to *CBPS*. These genes have been identified as essential by previous *in vitro* studies (Xiong et al., 2021), while only *Cad* is identified as essential by the essentiality analysis run on the GSMM. Because of that, the downregulation of the *Cad* should be done carefully to avoid negative effects on the cell, or the modification of other growth-inhibiting reactions might be associated to *G13MT*.

As previously mentioned, the regulation of other reactions in the *mannose* recirculation pathway is suggested in several experiments. The main ones are: *HEX4*, *DOLPMT4_er*, *G12MT1*, and *HEX16*. In order to achieve increased productivity, reactions in the *mannose* recirculation pathway are always coupled with other growth-inhibiting reaction, such as *IMPD*, *PRPPS*, *CGTPtn*, *RNDR1*, *RNDR4*, *NDPK7*, *TRDR*, *MI3PS_f*, *TMDS*, *ACCOAC*, *ADSS*, *G3PD1_b*, *CHOCK*, *ETHAK*, *AIRCr_f*, and *ACGAMK*.

8.5 Conclusions

In this Chapter, a machine learning strategy suggesting genetic modifications to increase monoclonal antibody productivity is presented. The proposed strategy exploits genome-scale metabolic models and the inversion of latent variable regression models to suggest the set of genetic modifications that improve the phenotype in a desired way through the knowledge of the relationship between intracellular reaction rates and cell phenotype.

In the proposed strategy, a latent variable regression model, which predicted both mAb specific productivity and biomass from the intracellular fluxes, was inverted to find how intracellular fluxes should be varied to obtain a desired phenotype (i.e., higher productivity). An algorithm exploited the suggested changes in the intracellular fluxes and a GSMM to identify sets of few genetic modifications that increase mAb specific productivity on the GSMM.

Genetic modifications involving the cell metabolism and the secretory pathway were identified by the proposed methodology.

Concerning the metabolic modifications, the genetic regulation of *L-valine* and *L-tryptophan* related reactions allowed to increase mAb productivity when coupled with growth-inhibiting reactions.

Concerning the secretory modifications, the regulation of reactions involved in the *mannose* recirculation in the early stages of glycosylation provided a substantial increase of mAb productivity in GSMM. These genetic modifications are theoretically feasible in real cells, but, since they are involved in the synthesis of all proteins, they are likely to increase the overall production of cell glycosylated proteins.

This study provided a novel and faster methodology than state-of-the-art methods to identify genetic modifications improving a desired phenotype in GSMMs. Furthermore, this demonstrated to be able to give valuable insights on cellular reactions and genes that can be targeted to improve the performance of cells. However, the suggested modification should be tested on real cell to assess their actual effectiveness. In the case of this study, the suggested genetic regulation of the *mannose* recirculation in the early stages of glycosylation will be analyzed in *in vitro* experiments at Imperial College London.

Conclusions and future perspectives

Monoclonal antibodies (mAbs) are biopharmaceutical drugs used for the treatment of autoimmune, oncological, and infectious diseases, which are typically produced in cultures of genetically modified mammalian cells. The development of new mAbs is a multi-step process, requiring large investments from biopharmaceutical companies and several years of research and testing. Typically, the entire development of a new drug requires more than 10 years and costs more than 2 billion dollars. The long timelines and large investments required for the development of new mAbs are pushing biopharmaceutical companies at looking for innovative and science-based solutions to support and accelerate the development of new drugs in all its phases: cell line generation, cell line selection, process characterization, and process optimization.

In this Dissertation, digital models were developed to support and accelerate all the phases of monoclonal antibody development process. Specifically, the contributions presented in this Dissertation were:

- supporting cell lines selection by integrating process and biological information;
- accelerating cell lines selection by exploiting dynamic biological information;
- identifying high performing cell lines in scenarios with limited available data through data-based models and *in silico* data augmentation;
- accelerating feeding schedule optimization through hybrid models;
- prediction of intracellular constraints of genome-scale metabolic models (GSMMs) from cheap and easily available data through deep learning models;
- identification of genetic engineering targets to improve CHO cell productivity exploiting GSMMs and latent variable model inversion.

Supporting cell lines selection by integrating process and biological information

The first study was aimed at integrating dynamic process and biological information from CHO cultivation to support cell line selection during biopharmaceutical process development, exploiting industrial data concerning development of mAbs provided by the multinational pharmaceutical GlaxoSmithKline. The main results of the work are listed in the following.

- The dynamics of metabolomic data allowed cell lines to be mapped according to process performance (e.g., viable cell concentration and antibody titer). This demonstrated to be an effective tool: *i*) to provide valuable information on the variations of metabolites associated with process behaviors; and *ii*) to infer the performance of new cell lines before the end of the experiment through a quasi-real time monitoring.

- The time course changes in biological phenomena were correlated to the process behavior by means of multiblock multivariate methods, allowing to identify how the time evolution of specific metabolites correlates to the process. In particular, a metabolite, *propinol adenylate*, was found to be anticorrelated to the antibody titer, while the metabolite, *L-lactic acid*, was found to be highly correlated with the lactate concentration, a typical by-product that reduce cell growth and productivity.
- The antibody titer time profile was estimated from dynamic metabolomic data by means of multivariate regression methods. A good estimation accuracy was achieved in cross-validation ($Q^2 > 40\%$) and external validation ($Q^2 > 60\%$) especially in the second half of the experimental batch. The model allowed also to identify the metabolites highly associated with the antibody titer over time. Specifically, *propinol adenylate* and *L-lactic acid* were shown to be associated to antibody titer especially in the central part of the experimental batch.

These results provided a deeper understanding of the metabolic states (i.e., biological pathways and metabolites) correlated with commercially relevant phenotypes. This methodology could be applied to increase the confidence in the selection of the most performing cell lines, allowing to reduce onward development timelines and resources. This work also fulfilled some of the regulatory requirements of Quality by Design, such as enhanced process understanding, and the monitoring and prediction of CQAs.

Accelerating cell lines selection by exploiting dynamic biological information

This work was aimed at demonstrating how dynamic metabolomic data can be exploited through data analytics to support and accelerate the selection of high productive cell lines during industrial bioprocess development. In this work, data from cell selection process at AMBR15$^{\text{TM}}$ scale provided by the multinational pharmaceutical GlaxoSmithKline was used. The main results of the work are listed in the following.

- High productive cell lines were identified by means of evolving multivariate multi-model strategies in the early stages of the cultivation process. In particular, both high and low productive cell lines could be discriminated with 100% accuracy in validation with 6 new cell lines.
- The metabolites associated with cell productivity were identified by means of a three-step method proposed in this work. Specifically, the metabolites associated with cell productivity were identified as *citric acid* in the initial part of the culture, and *UDP-glucose* and *thiamine* in the final part of the culture, which are associated with energy production, metabolism regulation, and protein glycosylation.

- The model allowed to identify the biological functions (i.e., metabolic pathways) associated with cell productivity over time. In exponential growth and stationary phases, the biological functions associated with productivity were related to energy production and DNA replication, whereas in the decline phase the biological functions associated with productivity were related to the metabolism of nucleotide and other sugars.

The developed models identified the cell lines with the desired phenotype in the early culture stages, allowing to accelerate bioprocess development by progressing those cell lines to larger scales. Moreover, the identification of few productivity biomarkers, which can be easily analyzed and interpreted in real-time without running an entire metabolomic study, allowed to make timely decisions on process development. All the acquired knowledge could be exploited for the implementation of a more robust and confident cell selection protocol and to mitigate the risk of progressing to larger scale poorly performing cell lines. Furthermore, the identified cellular functions provided insight on targets that can be manipulated though host engineering or process optimization to increase the frequency of obtaining high productive cell lines. This work also fulfilled some regulatory requirements of Quality by Design, such as the management of process variability, which is typically very large in biopharmaceutical applications, the monitoring and prediction of CQAs, and the mitigation of the risk of poor-quality products. The methodologies developed in this work were implemented in a software named ADAM, which is internally used by GlaxoSmithKline for the analysis of metabolomic data.

Identifying high performing cell lines in scenario with limited available data through data-driven models and *in silico* data augmentation

This work was aimed at applying different strategies for *in silico* batch generation to improve the identification of cell lines with the desired critical quality attributes (CQAs) (i.e., high mAb titer) by means of multivariate methods in scenarios of limited available data. The proposed method was tested in a simulated process for the production of mAbs to have a full knowledge of the relationship between process parameters and CQAs, and a better control of both the process behavior and the biological diversity in the experiments. Specifically, two approaches for *in silico* data generation were proposed: a first principles digital model, and a hybrid semi-parametric digital model. The main results of the work are listed in the following.

- A multivariate model to estimate the antibody titer was built on different numbers of process batches. The model accuracy sharply decreased when less than 10 process batches were used for model calibration (error $\gg$ 230 mg/L).
- A new model to estimate the antibody titer was built on different numbers of available process batches with the addition of the *in silico* data generated using the two proposed data augmentation strategies. *In silico* generated batches reduced the error variability with respect to process batches alone when 6-8 process batches are available, whereas improves the model accuracy when < 6 process batches are available. In fact, the model errors (170-

230 mg/L) were comparable with the measurement uncertainty (~150 mg/L). Furthermore, the first principles digital model improves the estimation accuracy even when the number of available batches is ≤ 3.

- The addition of the *in silico* generated batches from the first principle digital model allowed the correct identification of the important process parameters even when less than 10 process batches are available, which was not possible with only the process batches.

The use of *in silico* generated batches allowed the identification of high performing cell lines through data-based multivariate regression models even in scenarios where the number of available process data is limited, which is a typical situation encountered in mAbs development. Specifically, this could provide great advantages at different scales of the product and process development, especially at the stirred bioreactor scales, where the number of available batches is typically between 2 and 10. This improved identification of high performing cell lines reduced the experimental burden together with cost and timelines for process development. This work also fulfilled the regulatory requirements of Quality by Design concerned with the enhanced process understanding, and the monitoring and prediction of CQAs.

Accelerating feeding schedule optimization through hybrid models

This work compared an *in silico* experimental campaigns for the optimization of the feeding schedule in mammalian cell cultures through hybrid digital models with an experimental campaign on the process. This was intended to evaluate if the *in silico* experimental campaign can accelerate the experimentation and reduce the experimental burden in the process development. The proposed method was tested in a simulated process for the production of mAbs, which allowed knowing the exact relationship between nutrients and antibody titer and identifying the optimal feeding schedule of the process. The main results of the work are:

- Two experimental campaigns were planned by means of Design of Dynamic Experiments (DoDE): campaign A with 31 experiments and campaign B with 9 experiments. A response surface model, based on multiple linear regression, was built on the experiments to identify the feeding schedule maximizing the antibody titer at harvest. Experimental campaign A achieved an antibody titer of 3118.2 mg/L and campaign B 3136.3 mg/L, which did not approach the optimal antibody titer of the process of 3228.8 mg/L. Furthermore, the models predicted the antibody titer with limited accuracy, showing an error between 3.8% to 13.2%.
- The hybrid model trained on the 9 experiments of campaign B was used to conduct an *in silico* experimental campaign. This identified the feeding schedule maximizing the antibody titer at harvest. In particular, the *in silico* campaign achieved an antibody titer of 3222.8 mg/L, which approached the process optimum of 3228.8 mg/L. The hybrid model captured the trend of the relationship between nutrients and antibody titer better than the response surface models, even if the prediction of the actual antibody titer showed a relatively large error.

The *in silico* experimental campaign on a hybrid digital model achieved a better optimum than experimental campaigns designed through DoDE, providing an increase of 2.8% of the optimal antibody titer. Furthermore, the *in silico* experimental campaign reduced the experiments requirement to optimize the feeding schedule, because only 9 experiments were necessary for the hybrid model training. These results suggested that *in silico* experimental campaign can be powerful tools to reduce the experimental burden and timelines for process optimization. This work also fulfilled some regulatory requirements of Quality by Design, such as the mitigation of the risk of poor-quality products.

Prediction of intracellular constraints of GSMMs from cheap and easily available data through deep learning models

A deep learning strategy was developed to predict the constraints of GSMMs from easily available and cheap data in order to improve how GSMMs describe the metabolism of CHO cell lines. This work was a collaboration with Imperial College London (U.K.). The main results of the work are reported in the following.

- An artificial neural networks (ANN) was developed to predict intracellular fluxes from extracellular metabolite uptake rates, which are routinely measured in cell cultures. A 2 hidden layer ANN was used to predict one intracellular flux (among 47 total) from 24 extracellular metabolite uptake rates. Data augmentation based on SMOTE and gaussian noise addition was used to increase the number of training experiments and improve the robustness of the model.
- The ANN predicted with good accuracy ($Q^2 > 65\%$) most of the intracellular fluxes, especially reactions in glycolysis, TCA cycle, and some reactions in the amino acid metabolism, such as *GS*, *ASNS*, *SHMT*, *GCS*, *MAT*, *TDO*, *AKD*, *IBD*. The ANN predicted the fluxes of some reactions with lower performance ($Q^2 < 35\%$), mainly parallel and alternative reactions, such as PPP ones, and reactions involving metabolites that participate in many metabolic reactions, such as *PC*, *ARGS*, *TTA*, and *AASS*. The inaccurate predictions are probably due to inconsistencies in the available ^{13}C labeling data.
- The ANN estimated for each intracellular fluxes the 95% prediction interval. We proposed to use these prediction intervals to set lower and upper bounds of reactions in the GSMM.
- The GSMM constrained with the predicted bounds showed higher accuracy in calculating intracellular fluxes (Person correlation between GSMM calculated and experimental intracellular fluxes: 0.765 vs. 0.343) and predicting biomass ($Q^2 = 61.6\%$ vs. $Q^2 = 13.3\%$) than the base case and other state-of-the-art methods.

These results allowed a more effective use of GSMM to describe the metabolism of mammalian cells, providing a more accurate representation, better biological understanding, and an improved capability of culture design and optimization. This work also fulfilled some regulatory requirements of Quality by Design, such as enhanced understanding of the system, and the management of variability sources.

Identification of genetic engineering targets to improve CHO cell productivity exploiting GSMMs and latent variable model inversion

An innovative and efficient latent variable regression model inversion strategy was proposed to identify genetic engineering targets that improve mAb productivity by exploiting GSMMs. The proposed method was applied to CHO cells lines to improve mAb productivity. The main results are:

- A multivariate latent variable regression model was built to predict cell productivity and growth rate from the intracellular fluxes calculated by means of a GSMM. The model predicted with good accuracy ($Q^2 > 70\%$) both productivity and growth rate.
- The latent variable model was inverted to identify the intracellular fluxes associated to a desired and increased mAb productivity. In this specific case, a 50-100% increased productivity was desired. An algorithm exploiting GSMMs and the calculated intracellular fluxes was specifically developed to identify, among all the possible solutions, few reactions that must be genetically modified to obtain the desired productivity improvement.
- The strategy suggested that the downregulation of reactions involved in the metabolism of *L-valine* (*VALTA_f*) and *L-tryptophan* (*TRPO2*) improved the productivity of mammalian cells in GSMMs. The modification of these reactions must be associated with other growth-inhibiting modifications to divert resources from growth to mAbs production. The proposed genetic modifications were tested on a GSMMs and provided almost doubled productivity.
- The strategy suggested that the upregulation of reactions involved in the *mannose* recirculation pathways in the early glycosylation of mAbs improved cell productivity in GSMMs. The proposed genetic modifications, tested on a GSMMs, provided a productivity of $9.84 \cdot 10^{-5}$ against the initial $5.87 \cdot 10^{-5}$ a $mmol/(g_{DCW}\ h)$.

This study provided a novel and faster methodology than state-of-the-art methods to identify genetic modifications improving a desired phenotype in GSMMs. Furthermore, it demonstrated to be able to give valuable insights on cellular reactions and genes that can be targeted to improve the performance of cells. However, since these genetic modifications were observed in the GSMMs, the suggested genetic modifications should be tested *in vitro* to assess their actual effectiveness. This work also fulfilled some regulatory requirements of Quality by Design, such as the management of the variability sources, and the reduction of the risk of poorly productive cell lines, by guaranteeing high quality since cell generation.

Future perspectives

In view of the above conclusions, some possible ways to expand and improve the current work can be identified. Some future perspectives for this work are listed below.

- In Chapter 3 and 4, the use dynamic biological and process information was proven effective to support and accelerate the cell selection process. However, in the work only some critical quality attributes (CQAs) that are relevant form mAbs were considered. In future extensions of the work the methodology will study the relationship between biological information and antibody quality, such as glycan profile, in order to provide additional information for cell selection and eventually improve the understanding of the relationship between metabolic state and antibody quality.
- In Chapter 5, it was proven that the use of *in silico* data generation supports the identification of high productive cell lines even in the typical biopharmaceutical development scenario, where less than 10 experimental runs are available. However, this was proven on a simulated process. Hence, the proposed methodology will be tested on a real process. Furthermore, different modeling strategies and different solutions can be found to generate *in silico* data with a better balance between similarity with the original data and increased coverage of process variability.
- In Chapter 6, an *in silico* experimental campaign on a hybrid digital model was effectively used to optimize the feeding schedule of mammalian cells, achieving a higher antibody titer than experimental campaigns requiring a lower number of experiments. However, this was tested on a simulated process. For this reason, the proposed methodology will be tested on a real scenario, to assess the appropriate number of experiments for hybrid model training and improvement that can be achieved. Furthermore, different machine learning strategies could be exploited for the optimization of the feeding schedule exploiting the capability of hybrid models to simulate *in silico* an entire experimental campaign.
- In Chapter 7, Next-FLUX was proposed to predict the intracellular flux constraints of GSMMs form cheap and easily available measurements in order to make GSMM more accurate in describing the cell metabolism. The Next-FLUX was tested on the available ^{13}C isotope labeling dataset. For this reason, the proposed method will be further validated on new and independent ^{13}C isotope labeling data to assess its applicability in general and industrial scenarios. Furthermore, a software embedding the prediction of the intracellular constraints, their application on the GSMM, and its solution will be developed and released.
- In Chapter 8, a novel strategy to identify target genetic modifications exploiting GSMM and latent variable model inversion was proposed to enhance the cell lines productivity. The method identified meaningful genetic modifications that improved the productivity of mammalian cells in the GSMM. For this reason, the suggested genetic modifications will be tested on CHO cells at Imperial College London, to assess if the suggested modifications are able to improve cell productivity in a real scenario.

Appendix A

Monoclonal antibodies and cell cultures

In this Appendix, details on monoclonal antibodies structure, action, and production are presented. Furthermore, cell culture types, operations, and future trends will be explained.

A.1 Monoclonal antibodies

Monoclonal antibodies (mAbs) are therapeutic proteins commonly utilized for the treatment of autoimmune diseases, cancers, and infectious diseases (Kesik-Brodacka, 2018).

In the treatment of autoimmune diseases, mAbs targets different components of the immune system to suppress the acute responses typical of these diseases. For example, antibodies have been used for the treatment of rheumatoid arthritis, psoriatic arthritis, Chron's disease, ulcerative colitis, psoriasis and ankylosis spondylitis (Castelli et al., 2019).

Monoclonal antibodies have also been used for the treatment of both hematologic and solid tumors, such as leukemia, colorectal cancer, and metastatic breast cancer. In the treatment of the oncological diseases, mAbs act by: *i*) targeting some tumor antigens, such as growth factor receptors or hematopoietic differentiation antigens, to kill cancer cells, *ii*) delivering some radioisotopes to cancer cells in a selective way, and *iii*) targeting immune cells to enhance antitumor immune responses (Castelli et al., 2019).

In infectious diseases, mAbs have been used as prophylaxis and/or treatment, through the inhibition of viral replication. Monoclonal antibodies are available for several infectious diseases, but their development is slower in comparison to the treatments for oncological and autoimmune diseases. However, mAbs are available for the treatment of cytomegalovirus, hepatitis A and B viruses, HIV-1 infection, and SARS-CoV-2, while mAbs for the treatment of Ebola virus, hepatitis C and herpes simplex virus are under development (Castelli et al., 2019).

Apart from these common applications, mAbs have been also utilized in therapies for cardiovascular diseases, organ transplantations, respiratory diseases, and ophthalmologic diseases (Kesik-Brodacka, 2018).

A.1.1 Monoclonal antibody structure and function

Monoclonal antibodies (or immunoglobulins, Ig) are large Y-shaped proteins (Chapter 1; Figure 1.1), whose structure has been extensively reviewed in Chiu et al. (2019). They are composed of two identical heavy chains (~50 kDa each) and two identical light chains (~25 kDa each) with a molecular weight of ~150 kDa (Castelli et al., 2019). Heavy and light chains are connected and folded via intra and inter-chain disulfide bounds (Chartrain & Chu, 2008). The light chain is composed of one variable and one constant domain. The heavy chain, instead, is composed of one variable and 3 constant domains. The constant region, which defines the antibody class (i.e., IgM, IgG, IgD, IgA and IgE), has a nearly identical amino acid sequence in all antibodies of the same class (Kang & Lee, 2021). Currently, only IgGs are used as therapeutic mAbs because of their large circulating half-life and ease of production (Castelli et al., 2019). The variable region, instead, is the same for all the mAbs produced by a single cell clone.

The variable region of each heavy and light chain has three specialized sites, called complementarity determining regions, which dictate the specificity of each mAb through their amino acid sequence (Chartrain & Chu, 2008; Kang & Lee, 2021).

Monoclonal antibodies are divided in the antibody binding region (Fab) and the Fc region. The Fab region is composed of the two light and two heavy chains. The Fc region, instead, is composed of two heavy chains and binds to various receptors on effector cells of the immune system (Gaughan, 2016), which is the main action mechanism of mAbs. The Fc region is glycosylated at ASN-297 with N-linked glycans (i.e., polysaccharides) with a bi-antennary structure (Chartrain & Chu, 2008; Kang & Lee, 2021). This glycans have two N-acetylglucosamine (GlcNAc) residues connected to three bisecting mannose residues and can have a broad variety of terminal sugar composition, which greatly affect the activity of antibodies in terms of their inactivation. Additional details on antibody glycosylation, glycan composition and structure can be found in the Literature (Batra & Rathore, 2016; Sha et al., 2016).

The main role of antibodies in living organisms is to clear the host from invading pathogens and external molecules. Antibodies have very specific targets, called antigens, and are able to recognize them and bind exclusively to a small region (epitope) of a given antigen. The mAb uses the complementarity determining regions on the Fab to bind to the epitope on the antigen (Chartrain & Chu, 2008). The mAb binds to an antigen forming a complex that is recognized and cleared by specialized components or cells of the immune system of the host organism (Castelli et al., 2019; Chartrain & Chu, 2008). This is done in two main ways:

- antibody dependent cell-mediated cytotoxicity;
- complement-dependent cell cytotoxicity.

In antibody dependent cell-mediated cytotoxicity, killer cells (NK and NKC cells) recognize the mAb-target cell complexes and trigger the lysis and destruction of the pathogenic cells. In complement-dependent cell cytotoxicity, a protein complex binds to the mAb Fc region leading to lysis of the target cell.

A.1.2 Production of monoclonal antibodies

In living organisms, immunoglobulins are mainly produced by secretory B-cells, a component of the cell immune system (Gaughan, 2016). The secretion of monoclonal antibodies follows a specific mechanism (Gutierrez et al., 2020; Kontoravdi et al., 2005, 2007, 2010):

- DNA transcription: in the nucleus, the DNA genes encoding for the mAb light and heavy chains are transcribed into mRNA, which is responsible to carry the genetic information in the site of process synthesis;
- light and heavy chains translation: in the ribosomes, each nucleotide triplet in the mRNA, called codon, encodes for an amino acid and are translated into proteins thanks to a set of small RNA molecules called transfer RNA (tRNA);
- travel in the endoplasmic reticulum (ER): the light and heavy chains travel to the ER and two copies of each light and heavy chain are folded and combined to form the mAb. A control step is also performed in the ER in which cells try to correct protein misfolding and eventually degrade the mAb if the misfolded state is sustained for too long. Traveling through the ER the mAb is further glycosylated with N-glycans;
- travel in the Golgi apparatus: the mAb travel through the Golgi apparatus where it is further glycosylated through the bounding of N- and O-glycans.
- secretion: the complete mAb is secreted into the extracellular space through vesicles.

In the travel through the ER and Golgi, some glycosylation errors might happen leading to the secretion of mAbs that are not fully glycosylated with a degraded structural conformation. This undermines the bioactivity of mAb which may not be completely functional.

A.2 Cell cultures

A.2.1 Upstream process

In the context of the upstream section of biopharmaceutical processes, the operating modes, the bioreactor types, the operating parameters, the required elements for cell survival (i.e., medium and nutrients), and the upstream phases have to be considered.

A.2.1.1 Operating modes

Cell cultures for mAbs have three main operating modes (Chartrain & Chu, 2008; Gaughan, 2016; Rodrigues et al., 2009b):

- batch;
- fed-batch;
- perfusion.

Batch cultures are the simplest possible model of bioreactor operation, being often used in the past decade for many industrial applications (Rodrigues et al., 2009b). In batch operating mode, the bioreactor is initially loaded with medium and nutrients, then, cells are inoculated. The cells are allowed to grow until a determined cell density and product concentration with no further nutrient additions or withdrawals. Because of that, the nutrient concentration in the culture gradually decreases while product and by-products accumulate in the culture further limiting cell growth, allowing to reach a maximum viable cell concentration of ~10^6 cells/mL (Rodrigues et al., 2009b). A balance between the amount of nutrient to reduce growth limitation and the acceptable level of toxic by-products must be found for the correct operation of batch bioreactors.

Fed-batch cultures are nowadays the preferred choice for the production of mAbs. After the startup of the bioreactor, nutrients are periodically fed with fresh media to increase culture longevity, maintain nutrient sufficiency and limit the effect of nutrient depletion. However, the accumulation of growth-inhibiting by-products is not avoided. Furthermore, a frequent feeding allows to better control the growth rate of cells through the flow rate of feed and medium. In this way, fed-batch operations allow to reach cell specific productivity of over 20 pg/cell/day, antibody titers up to 10 g/L and viable cell concentration over $20 \cdot 10^6$ cells/mL (Gaughan, 2016). The product is harvested only at the end, when the viability of cells drops below a target value, typically after 2 weeks.

In perfusion bioreactors, fresh medium is continuously added to the culture at a very low rate, while an equal amount of spent medium with the product is removed from the culture (Birch & Racher, 2006). Each perfusion reactor is typically connected to a filtration or centrifugation unit for the separation of viable cells (Gaughan, 2016). In some cases, even viable cells are extracted from the bioreactor to avoid sterility issues over long periods of time (Shukla & Thömmes, 2010). Perfusion bioreactors provide very stable operations, with constant glucose/lactate concentration, pH and DO, lasting for long periods of time, even 35-40 days. Furthermore, by-products are constantly removed from the culture, increasing viable cell concentrations (~10^7 cells/mL) and specific productivity (Rodrigues et al., 2009b), generally requiring smaller cultivation vessels and smaller factories than fed-batch reactors and, accordingly, less space for the production of a similar amount of mAbs (Gaughan, 2016).

A.2.1.2 Bioreactor types and operating parameters

Bioreactors are the main equipment used in upstream processes, in which cells grow and produce mAbs. Several types of bioreactors are used for the production of mAbs; details on all the possible choices of bioreactor types can be found in Rodrigues et al. (2010). However, the main types currently used are two (Chartrain & Chu, 2008; Gaughan, 2016; Rodrigues et al., 2009a):

- stainless steel bioreactors;

- disposable bioreactors.

Stainless steel bioreactors are stirred tanks with baffles and impellers and a volume ranging from 1000 L to 25000 L. They allow high and flexible operating volumes and modes, high mass/gas transfer coefficients, applicability to several cell and product types, making them the preferred choice in the past decades. Disposable bioreactors are polymeric bags that can be placed on rocking plates for convection mixing or have their own internal impeller, usually with a volume ranging between 50 L and 2000 L. Recently, disposable, single-use bioreactors are the preferred choice in the biopharmaceutical industry because they have substantially decreased preparation times, eliminating at the same time all the issues of cleaning, sterilization and cross-contamination risks. Furthermore, they provide a significant reduction in the capital investments associated with stainless steel reactors.

To achieve high product yield and acceptable product quality the bioreactor operation must be optimized. The main bioreactor operating parameters for process optimization are (Birch & Racher, 2006; Chartrain & Chu, 2008; F. Li et al., 2010; Rodrigues et al., 2009a; Shukla & Thömmes, 2010):

- temperature: temperature is the most critical variable for healthy cultures. Cells are typically cultivated at 37 °C to favor cell growth. However, this temperature is not the most favorable for mAbs production, hence, once a desired viable cell concentration is reached the temperature is set to 30-35°C. This temperature shift allows the cells to redirect their metabolism away from growth toward mAb production, allowing higher specific productivity.
- pH: pH is another critical variable for mammalian culture, having an impact on cell growth, productivity, cell metabolism and protein glycosylation even with small variations. Cells are typically cultured with pH near 7.4, but, after reaching a desired viable cell concentration, pH is usually reduced to 6.7-7.0 to limit cell growth and increase specific productivity. Commonly, pH is controlled by the addition of CO_2 to the culture headspace or the addition of bicarbonate base.
- O_2: mammalian cells require oxygen to produce energy from carbon sources. However, the dissolved oxygen (DO) does not negatively affect cell growth and productivity while in physiological ranges, but can have an effect on protein glycosylation (Rodrigues et al., 2009b). DO is typically set between 30% and 60% of air saturation.
- CO_2: in mammalian cells, CO_2 should be kept at physiological levels because it is required to maintain the pH and to regulate many cellular activities while an excessive accumulation can have inhibitory effects. Partial pressures of CO_2 (pCO_2) are typically set at 120-150 mmHg.
- osmolarity: osmolarity has an influence on mAbs production, cell growth and death, and the duration of the exponential growth. In fact, high osmolarity, triggered by the addition

of base, glucose or concentrated medium, causes decreased growth rates and viable cell concentrations. The osmolarity is typically set and controlled at 270-330 mOsm/kg.

- agitation rate: the agitation rate provides mixing of cells and together with the gas flow rate is used to control the DO and the pCO_2.

A.2.1.3 Elements for cell survival

Apart from the bioreactor operating parameters, medium and nutrients are other two main parameters to be controlled for cell survival (Birch & Racher, 2006; Chartrain & Chu, 2008; Gaughan, 2016; Rodrigues et al., 2009a; Shukla & Thömmes, 2010).

The media for the cultivation of mammalian cells are highly complex, contain several components and cost more than 20$ per liter (Chartrain & Chu, 2008). These media contain all the growth supporting molecules, such as amino acids, vitamins, nucleosides, trace elements, metals, inorganic salts, lipids and insulin or insulin-like growth factors. In the past, bovine serum was the standard, but recently companies are avoiding animal ingredients in biopharmaceutical production to avoid the introduction of adventitious agents. As a consequence, fully chemically defined media have been developed using hydrolysates from yeast or plant sources which allow to effectively replace the use of serum. Media can be purchased by vendors, but many companies are starting to produce their own media to avoid shortage problems and to specifically design it for each purpose.

In mammalian cell cultures, glucose and glutamine are the most limiting nutrients, being the main carbon sources. The metabolism of such nutrients leads to the production of by-products, such as lactate and ammonia, whose accumulation in the culture can inhibit cell growth and mAb productivity, affecting also mAb glycosylation (Rodrigues et al., 2009b). For this reason, the feeding strategy should be optimized to maximize productivity and growth and minimize the formation of undesirable by-products (Chartrain & Chu, 2008). In order to achieve that it is extremely important to have quantitative understanding of cells nutritional requirements (Birch & Racher, 2006).

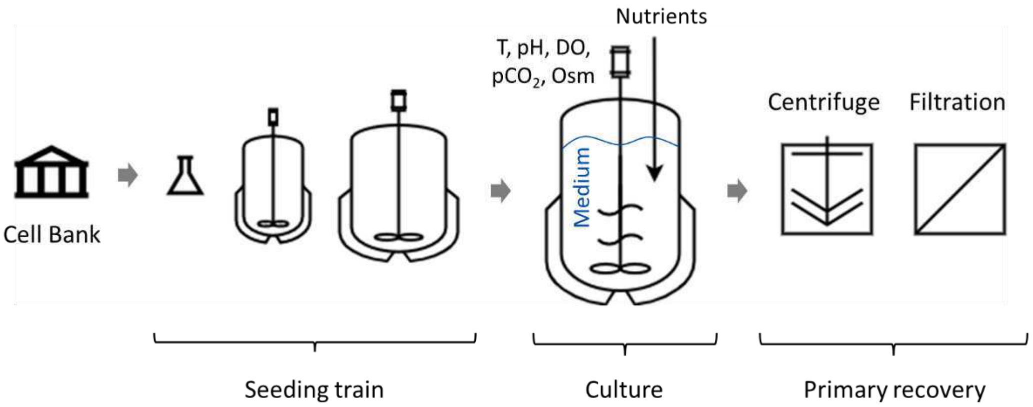


Figure A.1 *Monoclonal antibody industrial production: upstream process phases. Adapted from Shukla and Thömmes (2010).*

A.2.1.4 Upstream phases

The upstream process is mainly divided into three phases (Figure A.1):

- seeding;
- culture;
- primary recovery.

Selected cells for the production of a specific biopharmaceutical are typically stored in two cell banks, called Master Cell Bank and Working Cell Bank, which ensure a long term supply of cells for the entire expected life of a biopharmaceutical product (Chartrain & Chu, 2008).

The seeding step aims at generating enough cells to inoculate the final production bioreactor by serial expansions (Birch & Racher, 2006). Frozen vials, extracted from the Working Cell Bank, are initially expanded in shake flasks or spinner flasks, which progressively increase in volume (Shukla & Thömmes, 2010). The cells are inoculated into a series of seeding bioreactors, which are typically operated for many months. A procedure known as rolling seed train is used to speed up the expansion phase, which can span over more than one month due to the low average doubling time of mammalian cells (i.e., about one day). In this procedure, a substantial volume is drained from the bioreactor every few days and used to seed the production bioreactor. An equal volume of fresh medium is added to the bioreactor to preserve its operation and allow the cells to double again.

In the culture phase, cells are inoculated in the production bioreactor and grow (typically in fed-batch mode) until the desired conditions are met, leading to 2 weeks typical culture lengths (Birch & Racher, 2006).

After harvesting, the culture media with products and cells goes to the primary recovery, in which cells and cell debris are removed through a centrifuge and membrane filters prior being purified in the downstream process (Birch & Racher, 2006; Shukla & Thömmes, 2010)

A.2.2 Downstream process

In the biopharmaceutical production of mAbs, the downstream process has become widely established to purify the product and reduce all the impurities to acceptable levels (Shukla & Thömmes, 2010). Nowadays, growing attention is given to the downstream process because it has become the limiting step for process throughput (Birch & Racher, 2006). The downstream process is mainly divided into three steps:

- protein A affinity chromatography: protein A affinity chromatography is the antibody capturing step and is based on the specific binding activity between the Fc region of mAbs and a so-called protein A ligand. It has a capacity ranging from 15–100 g mAb/L and an extremely high selectivity towards IgG which allows host cell proteins, DNA and other impurities to be separated, achieving $>$ 95% purity in one step (Birch & Racher, 2006; Gronemeyer et al., 2014; Shukla & Thömmes, 2010).

- polishing chromatography: up to three polishing chromatographic separations are used to further reduce impurities, such as host cell proteins, DNA and high-molecular-weight aggregates, to acceptable levels. To this purpose, ion exchange chromatography is typically used (Birch & Racher, 2006; Gronemeyer et al., 2014).
- viral filtration: at least two steps are used to remove and inactivate viruses. These are typically based on filtration, low pH treatments and solvents. After that, the process is typically terminated with ultrafiltration/diafiltration operation to reduce storage volume (Birch & Racher, 2006; Shukla & Thömmes, 2010)

A.2.3 Future manufacturing trends: continuous production

In the recent years, a large demand of biopharmaceutical products is emerging, making the typical fed-batch processes currently used often insufficient to fulfill the product demand. Continuous manufacturing is a promising methodology to overcome this problem, because it shortens the production cycle, increases the equipment utilization, and enables constant product quality, providing a reduction in the overall costs (Schofield, 2018). Furthermore, regulatory agencies are promoting the transition to continuous manufacturing through guidelines (Q13 - Continuous Manufacturing of Drug Substances and Drug Products, 2021). However, guidelines with a framework for continuous manufacturing is not available yet (Manser & Glenz, 2022).

The transition to continuous manufacturing started with the development of perfusion bioreactors, which substantially increased productivity and mAb titers, making the downstream process, especially the protein A affinity chromatography, which is intrinsically batch, the main bottleneck in biopharmaceutical processes (Birch & Racher, 2006; Gerstweiler et al., 2021). The main solution is to use the perfusion reactors to prepare highly concentrated solution for the further purification steps. However, further steps must be undertaken to examine ways to integrate individual continuous unit operations for developing a fully integrated process (Gerstweiler et al., 2021). Recently, several progresses have been made in this context, especially on the substitution of chromatographic processes with crystallization, precipitation, and membrane technologies, and on the automated control of the entire downstream process (Thakur et al., 2022). However, the development of fully continuous processes for the production of mAbs is still an open issue.

Appendix B

Additional details on the integration of metabolome dynamics and process data

This Appendix collects additional mathematical details and results of the paper Barberi et al. (2022) that is at the basis of Chapter 3.

B.1 Measurement replicate unfolding

To account for measurement replicates variability, data are replicate-wise unfolded by vertically concatenating the R measurement replicates, producing the matrices $\underline{\mathbf{X}}_\mathrm{I}$ $[N \cdot R \times V_\mathrm{I} \times (T-1)]$ and $\underline{\mathbf{X}}_\mathrm{E}$ $[N \cdot R \times V_\mathrm{E} \times T]$. Two copies of process data $\underline{\mathbf{X}}_\mathrm{p}$ are vertically concatenated to account for R measurement replicates in $\underline{\mathbf{X}}_\mathrm{P}$ $[N \cdot R \times V_\mathrm{P} \times T]$, which are not available for process data. In any further splitting of the data, measurement replicates are considered as a single sample and extracted together.

B.2 Multiway principal component analysis

Multiway principal component analysis (MPCA; Nomikos and MacGregor, 1994) is used for the exploration of data in the case of multidimensional matrices, when usually one of the dimensions is related to data variability in time (i.e., data dynamics). MPCA consists in a PCA (Wold et al., 1987) on properly unfolded multiway data. In particular, data collected at different time instants (e.g., $\mathbf{X}_\mathrm{I}^t$ $[N \cdot R \times V_\mathrm{I}]$ with $t = 1, 3, \dots, T$) are horizontally concatenated to generate $\mathbf{X}_\mathrm{I}$ $[N \cdot R \times V_\mathrm{I} \cdot (T-1)]$, which is the cell-wise unfolded version of $\underline{\mathbf{X}}_\mathrm{I}$. The extracellular data $\underline{\mathbf{X}}_\mathrm{E}$ are unfolded in the same way in $\mathbf{X}_\mathrm{E}$ $[N \cdot R \times V_\mathrm{E} \cdot T]$. When required, also the process data $\underline{\mathbf{X}}_\mathrm{P}$ are cell-wise unfolded in $\mathbf{X}_\mathrm{P}$ $[N \cdot R \times V_\mathrm{P} \cdot T]$.

PCA is a multivariate statistical technique which decomposes a dataset (e.g., pareto scaled $\mathbf{X}_\mathrm{E}$ $[N \cdot R \times V_\mathrm{E} \cdot T]$) of $N \cdot R$ observations on $V_\mathrm{E} \cdot T$ variables into A independent principal components (PCs), which describe the direction of maximum variability of $\mathbf{X}_\mathrm{E}$ and capture the correlation structure between the $V_\mathrm{E} \cdot T$ original variables. This decomposition is performed according to:

$$\mathbf{X}_\mathrm{E} = \mathbf{T}\mathbf{P}^\mathrm{T} + \mathbf{E} \quad , \tag{B.1}$$

where $\mathbf{T}$ $[N \cdot R \times A]$ is the score matrix, $\mathbf{P}$ $[V_\mathrm{E} \cdot T \times A]$ is the loading matrix, the superscript T indicates the transpose, and $\mathbf{E}$ $[N \cdot R \times V_\mathrm{E} \cdot T]$ is the residual matrix, which is minimized in the least-square sense. The loadings describe not only the correlation structure among original variables (e.g., metabolites), but also how variables are auto-correlated in time and cross-correlated with the dynamics of other variables (when the data are cell-wise unfolded). Scores represent the projection of observations in the subspace of PCs and describe the relation between different observations according to the patterns of the time profiles of the considered variables.

A new observation $\mathbf{x}_\mathrm{NEW}$ $[1 \times V_\mathrm{E} \cdot T]$ can be projected onto a PCA model space through:

$$\mathbf{t}_\mathrm{NEW} = \mathbf{x}_\mathrm{NEW}\mathbf{P} \quad , \tag{B.2}$$

where $\mathbf{t}_\mathrm{NEW}$ $[1 \times A]$ is the score vector of the new observation. The projection of a new observation is used to assess whether a new observation is similar or conform to the ones used to build the PCA model.

The real-time mapping of a new observation is performed similarly by projecting an observation at each time instant t (with $t = 1, 2, \dots, T$) and completing the missing measurements (from t+1 to t=7) with the respective average values calculated over the calibration data used to build the model (Ramaker et al., 2005).

B.3 Similarity analysis

The similarity factor (Facco et al., 2020; Krzanowski, 1979) compares the correlation structure captured by two PCA models built on matrices $\mathbf{X}_\mathrm{I}^{t'}$ $[N \cdot R \times V_\mathrm{I}]$ and $\mathbf{X}_\mathrm{I}^{t''}$ $[N \cdot R \times V_\mathrm{I}]$ (for example), the slices of the three-dimensional matrix $\underline{\mathbf{X}}_\mathrm{I}$ at time instants t' and t'', using the same number of PCs. It compares the direction of maximum variability in the two datasets, therefore their major driving forces. The similarity factor $S_{t't''}$ is defined as:

$$S_{t't''} = \frac{trace\left(\mathbf{P}_{t'}^{*\mathrm{T}}\mathbf{P}_{t''}^{*}\mathbf{P}_{t''}^{*\mathrm{T}}\mathbf{P}_{t'}^{*}\right)}{\sum_{a=1}^{A}\lambda_{t',a}\lambda_{t'',a}} \quad , \tag{B.3}$$

where $\lambda_{t',a}$ is the eigenvalue of the PCA model built on $\mathbf{X}_\mathrm{I}^{t'}$ for the a-th PC, and $\mathbf{P}_{t'}^{*} = \mathbf{P}_{t'}\mathbf{L}_{t'}$, where $\mathbf{P}_{t'}$ is the loading matrix of the model built on $\mathbf{X}_\mathrm{I}^{t'}$ and $\mathbf{L}_{t'}$ $[A \times A]$ is the diagonal matrix of the square root of $\lambda_{t',a}$. The notation for $\mathbf{X}_\mathrm{I}^{t''}$ is similar.

The similarity analysis gives a quantitative information on how much the driving metabolic phenomena change along the culture time course. Specifically, high values of the similarity factor (values close to 1) between metabolomic data at different time instants indicate that a significant portion of the ions shows similar variation across all the samples. Accordingly, the metabolic phenomena driving these variations are likely to be the same. Conversely, low values of the similarity factor (values close to 0) between metabolomic data at different time instants indicates that a significant portion of the ions show very different variation across all the samples. Hence, the driving metabolic phenomena are likely to be different.

The variables mostly responsible for the similarity are identified as the ones having high and similar loading values (within 5% difference) in $\mathbf{P}_{t'}$ and $\mathbf{P}_{t''}$.

B.4 Multi-block principal component analysis

Multi-block PCA (MB-PCA; Westerhuis et al., 1998) is an unsupervised multi-block method which relates different blocks of variables (e.g., process and biological data). In MB-PCA, the available data blocks (with the same observations on the rows) are horizontally concatenated, prior the decomposition through a standard MPCA (Appendix B.2). Process data were autoscaled to zero mean and unit variance, while metabolomic data were pareto scaled (Eriksson et al., 2006). No additional block scaling was performed to avoid reducing the importance of the metabolomics block which comprises a large number of variables. In this study, since culture and metabolomic dynamic data are available, a multi-block-multiway-PCA (MB-MPCA) was applied to relate the dynamic variations of metabolomic profiles and process variables.

B.5 Multiway partial least-squares regression

Multiway partial least-squares (MPLS; Nomikos and MacGregor, 1995) consists in a batch-wise unfolding followed by partial least-squares (PLS; Wold et al., 2001) modeling. PLS is a linear multivariate regression technique which identifies the direction of maximum covariance between regressors (e.g., pareto scaled $\mathbf{X}_{\mathrm{E}}$ $[N \cdot R \times V_{\mathrm{E}} \cdot T]$) and an autoscaled matrix $\mathbf{Y}$ $[N \cdot R \times M]$ of M responses (e.g., a slice of $\underline{\mathbf{X}}_{\mathrm{P}}$). PLS projects $\mathbf{X}_{\mathrm{E}}$ and $\mathbf{Y}$ into a reduced space of A latent variables LVs as explained in Section 2.1.2. The selection of the appropriate number of LVs is performed through a 9-fold cross-validation (Geladi & Kowalski, 1986b).

Model performance is evaluated through a 250-iterations Monte Carlo cross-validation, in which samples are randomly split in calibration ad validation sets (88% of samples is for calibration). External validation cell lines are randomly selected from the initial dataset (12 cell lines) and are used to assess model robustness and generalization performances.

B.6 Variable selection

The selection of the relevant variables is performed through a bootstrap procedure (Afanador et al., 2013) on the variable importance in projection (VIP; Eriksson et al., 2006) index. The VIP score of a generic ion at a specific time instant, $vt = 1,2, \dots, V \cdot T$, is defined as:

$$VIP_{vt} = \frac{\sqrt{V \cdot T \sum_{a=1}^{A} R_{Y,a}^2 w_{vt,a}^2}}{\sqrt{\sum_{a=1}^{A} R_{Y,a}^2}} \quad , \tag{B.7}$$

where $R_{Y,a}^2$ is the variance of the response explained by the a-th LV of the model, and $w_{vt,a}$ is the weight of the vt-th ion in the cell-wise unfolded version of the data and a-th LV.

The bootstrap procedure selects the most influential variables for the prediction of the PLS response **Y**, because it retains only ions whose VIP index remain high independently from the subset of sample selected for validation in the specific iteration. The variable selection is performed on $it_{\max} = 250$ Monte Carlo iterations following three steps:

1. calculation of the VIP index standard deviation over the $it_{\max}$ iterations for each ion, $\hat{\sigma}_{\mathrm{VIP}_{vt}}$ (Afanador et al., 2013);
2. calculation of the 90% confidence interval of the VIP index distribution for each ion, under the assumption that the VIPs are distributed according to a Student's t distribution. The confidence interval is calculated through: $\hat{\sigma}_{\mathrm{VIP}_{vt}} t_{1-\alpha/2, it_{\max}-1}$, where $t_{1-\alpha/2, it_{\max}-1}$ identifies the confidence threshold of a t-distribution with $(it_{\max} - 1)$ degrees of freedom calculated with $\alpha = 0.1$;
3. selection of the top-ranked 5% variables that guarantee the largest values of the lower 90% confidence limit (LCL). The selection of the 5% of variables provides good model performance allowing an easier interpretation of the model outcomes.

This method guarantees that only variables having a VIP > 1 with a 95% confidence are selected as important for the estimation of the response variable **Y**. After the variable selection a new PLS model is built, showing improved prediction performance. In the new PLS model the variables with $VIP_{\mathrm{LCL}} > 1$ are considered highly related and predictive for the response **Y**.

B.7 Additional information on culture variable correlation

Details on Figure B.1 are briefly discussed in Section 3.3.2.

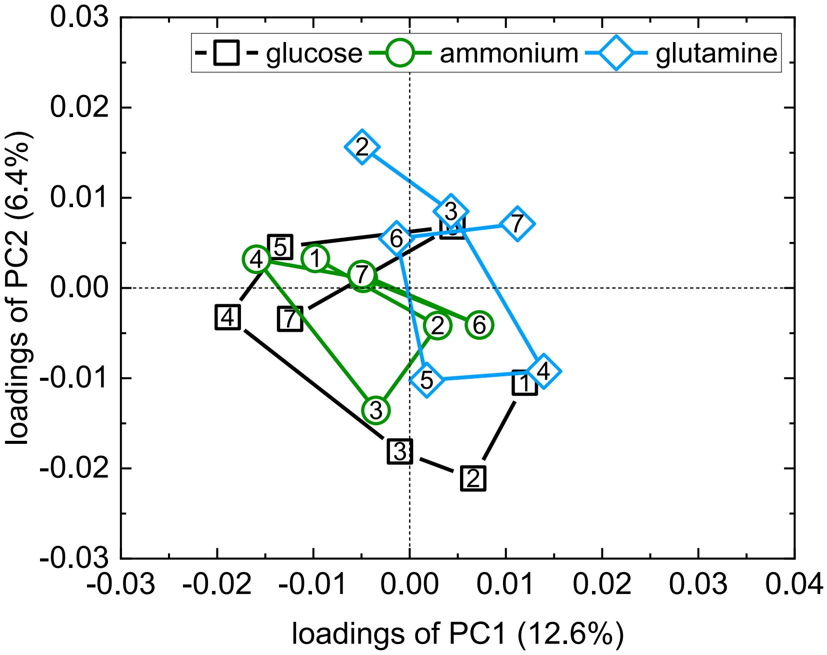


Figure B.1 *MB-MPCA model on* $\boldsymbol{X}_1$*: loadings plot showing the correlation between additional process variables dynamics.*

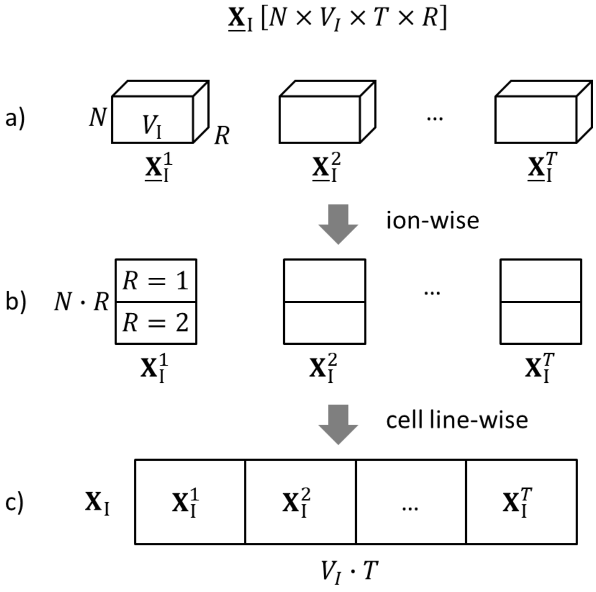
Appendix C

Mathematical details of for the selection of high productive cell lines

This Appendix collects additional mathematical information and results of Chapter 4.

C.1 Data unfolding

MPLS-DA (Barker & Rayens, 2003; Nomikos & MacGregor, 1995b) requires a proper data unfolding to deal with multidimensional data (i.e., with dynamic information and measurement replicates). The data unfolding procedure is schematically shown in Figure C.1. The multiway dataset $\underline{\mathbf{X}}_\mathrm{I}$ $[N \times V_\mathrm{I} \times T \times R]$ (for example) is firstly unfolded ion-wise to average the effect of different measurement replicates and then cell line-wise unfolded to take into consideration the dynamics of metabolomic data. The data collected at different time instants $\underline{\mathbf{X}}_\mathrm{I}^{\boldsymbol{t}}$ $[N \times V_\mathrm{I} \times R]$ with $t = 1, 2, \dots, T$ are isolated (Figure C.1a) and ion-wise unfolded by vertically concatenating the measurement replicates (Figure C.1b) to generate $\mathbf{X}_\mathrm{I}^{\boldsymbol{t}}$ $[N \cdot R \times V_\mathrm{I}]$. Then, the data collected at different time instants $\mathbf{X}_\mathrm{I}^{\boldsymbol{t}}$ are by horizontally concatenated (i.e., cell line-wise unfolded) to generate the matrix $\mathbf{X}_\mathrm{I}$ $[N \cdot R \times V_\mathrm{I} \cdot T]$ (Figure C.1c).



$\underline{\mathbf{X}}_\mathrm{I}$ $[N \times V_I \times T \times R]$
a)
N
V_I
R
$\underline{\mathbf{X}}_\mathrm{I}^1$
$\underline{\mathbf{X}}_\mathrm{I}^2$
...
$\underline{\mathbf{X}}_\mathrm{I}^T$
ion-wise
b)
$N \cdot R$
$R = 1$
$R = 2$
...
$\mathbf{X}_\mathrm{I}^1$
$\mathbf{X}_\mathrm{I}^2$
$\mathbf{X}_\mathrm{I}^T$
cell line-wise
c)
$\mathbf{X}_\mathrm{I}$
$\mathbf{X}_\mathrm{I}^1$
$\mathbf{X}_\mathrm{I}^2$
...
$\mathbf{X}_\mathrm{I}^T$
$V_I \cdot T$

Figure C.1 *Schematic representation of the data unfolding procedure.*

C.2 PLS-DA

The PLS-DA model (Barker & Rayens, 2003) reduces the $V_I \cdot T$-dimensional space of the metabolomic profiles dynamics to a smaller space A orthogonal LVs, which captures, in this case, the dynamics of metabolites mostly related to the discrimination of cell productivity. The PLS-DA model is built on the unfolded dataset (Appendix C.1) as:

$$\mathbf{X}_{\mathrm{I}} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad , \tag{C.1}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \quad , \tag{C.2}$$

$$\mathbf{T} = \mathbf{X}_{\mathrm{I}}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1} \quad , \tag{C.3}$$

where $\mathbf{P}\ [V_I \cdot T \times A]$ and $\mathbf{Q}\ [2 \times A]$ are the loading matrices, $\mathbf{T}\ [N \cdot R \times A]$ is the score matrix, $\mathbf{E}\ [N \cdot R \times V_I \cdot T]$ and $\mathbf{F}\ [N \cdot R \times 2]$ are the residual matrices of $\mathbf{X}_{\mathrm{I}}$ and $\mathbf{Y}$, respectively (minimized in a least-square sense), and $\mathbf{W}\ [V_I \cdot K \times A]$ is the weight matrix. The model scores describe the relationship between cell lines according to their metabolomic profile dynamics, while the loadings and the weights describe how the dynamics of metabolites and their correlations are related to the discrimination of the cell productivity.

C.3 E-MPLS-DA

E-MPLS-DA (Barker & Rayens, 2003; Ramaker et al., 2005) is a multi-model strategy that exploits partial dynamic information for classification. Specifically, this method retains information on the entire past history of the experimental batch to accomplish the classification in each time instant in which data are available along the culture course. At each time instant t with $t = 1, 2, \dots, T$ a MPLS-DA model (Section 4.2.2) is built on the matrix $\mathbf{X}_{\mathrm{I},t}\ [N \cdot R \times V_I \cdot t] = [\mathbf{X}_{\mathrm{I}}^1, \mathbf{X}_{\mathrm{I}}^2, \dots, \mathbf{X}_{\mathrm{I}}^t]$ which contains the ion-wise unfolded metabolomic data up to the time instant t. The matrix $\mathbf{X}_{\mathrm{I},t}$ is progressively enlarged, while new MPLS-DA models are built, until the entire available dynamics of data is considered. A schematic representation of the E-MPLS-DA model is shown in Figure C.2.

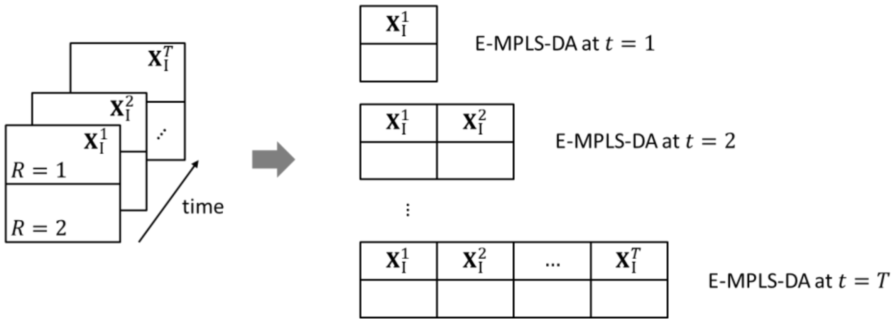

$\mathbf{X}_{\mathrm{I}}^T$
$\mathbf{X}_{\mathrm{I}}^2$
$\mathbf{X}_{\mathrm{I}}^1$
$R = 1$
$R = 2$
time
$\mathbf{X}_{\mathrm{I}}^1$
E-MPLS-DA at $t = 1$
$\mathbf{X}_{\mathrm{I}}^1$ $\mathbf{X}_{\mathrm{I}}^2$
E-MPLS-DA at $t = 2$
$\mathbf{X}_{\mathrm{I}}^1$ $\mathbf{X}_{\mathrm{I}}^2$ ... $\mathbf{X}_{\mathrm{I}}^T$
E-MPLS-DA at $t = T$

Figure C.2 *E-MPLS-DA multi-model building procedure.*

C.4 Variable selection

The most important ions for productivity discrimination are selected through a robust and computationally-intensive backward iterative elimination of the uninformative variables (Fernández Pierna et al., 2009; Mehmood et al., 2012), where three importance metrics are used to identify the uninformative variables:

i) Variable importance in Projection index (VIP) (S Wold et al., 1993) defined for the v-th ion as:

$$VIP_v = \sqrt{\frac{v \sum_{a=1}^{A} w_{va}^2 SSY_a}{\sum_{a=1}^{A} SSY_a}} \quad , \tag{C.4}$$

where SSY_a is the sum of squares of $\mathbf{Y}$ explained by the a-th LV, w_{va} is the weight of the v-th ion and a-th LV, V is the total number of ions and A the total number of LVs.

ii) selectivity ratio (SR) (Kvalheim & Karstang, 1989) defined for the v-th ion as:

$$SR_v = \frac{SSX_{\mathrm{exp},v}}{SSX_{\mathrm{res},v}} \quad , \tag{C.5}$$

where $SSX_{\mathrm{exp},v}$ is the explained variance, and $SSX_{\mathrm{res},v}$ is the residual variance for ion v.

iii) regression coefficients defined as:

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{Q}^{\mathrm{T}} \quad , \tag{C.6}$$

where $\mathbf{B}$ $[V_I \cdot T \times 2]$ is the matrix of regression coefficients.

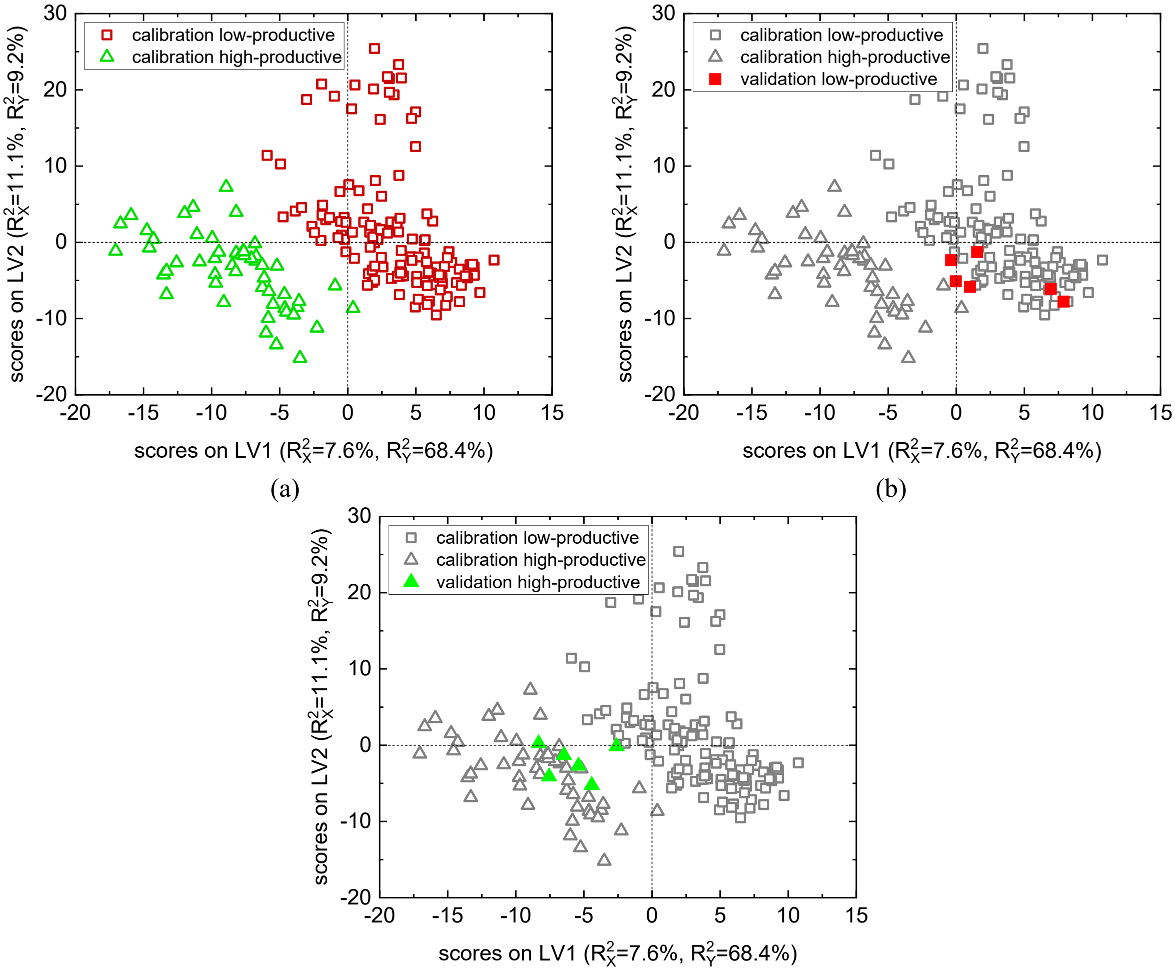
A MPLS-DA or E-MPLS-DA model is then built with the retained ions, showing improved classification performance.

C.5 VIP bootstrap

A bootstrap procedure (Afanador et al., 2013) is used on the model developed in Section 4.2.2 to identify the most robust ions for productivity discrimination. Only ions whose VIP score remains high independently of the available subset of samples available in a cross-validation are retained. In particular, the results of the $it_{\max} = 250$ iterations Monte Carlo cross-validation (Section 4.2.2.1) are used for:

1. calculation of the VIP standard deviation of each ion v over the $it_{\max}$ iterations, $\hat{\sigma}_{VIP_v}$;
2. calculation of the VIP 90% confidence limit for each ion, under the assumption that the VIPs are distributed according to a Student's t distribution. The 90% confidence limit ($\alpha = 0.1$) is calculated as: $\hat{\sigma}_{VIP_v} t_{1-\alpha/2, it_{\max}-1}$, where $t_{1-\alpha/2, it_{\max}-1}$ identifies the lowest 5% confidence threshold of a t-distribution with $(it_{\max} - 1)$ degrees of freedom;
3. ions with the lowest VIP 90% confidence limit (VIP_{LCL}) > 1 are retained. The VIP_{LCL} is calculated as: $\overline{VIP_v} - \hat{\sigma}_{VIP_v} t_{1-\alpha/2, it_{\max}-1}$, where $\overline{VIP_v}$ defines the average value of the VIP score of the ion v over the $it_{\max}$ iterations.

C.6 Discrimination of high productive cell lines through extracellular metabolomic data



(a)

(b)

(c)

Figure C.3 *Score space of the MPLS-DA model built on the extracellular metabolomic data for the discrimination of cell productivity: (a) calibration samples, (b) low productive external validation cell lines, and (c) high productive external validation cell lines.*

C.7 Anticipated discrimination of high-productive cell lines through extracellular metabolomic data

Table C.1. *Performance of the E-MPLS-DA multi-model in the anticipated discrimination of cell productivity from extracellular data: the number of LVs, the explained response variance ($\boldsymbol{R_y^2}$), the number of retained ions and the accuracy in cross validation and external validation are reported for the model built at each time instant.*

time instant	number of selected LVs	R_y^2 [%]	number of selected ions	accuracy in cross-validation [%]	validation accuracy [%]
1	3	94.6	184	98.0	66.7
2	2	94.0	224	97.9	58.3
3	2	95.3	357	99.6	50.0
4	3	94.6	588	99.4	66.7
5	3	93.5	684	98.5	75.0
6	3	92.0	2188	97.3	83.3
7	3	93.6	2005	97.8	83.3

Appendix D

Additional detail on data augmentation

D.1 Model of mammalian cell cultured producing mAbs

The HEK model (Kontoravdi et al., 2010) used in this work is a first principles mathematical model which simulates batches for the production of mAbs. It is composed of 3 main parts: cell growth and death, cell metabolism, and mAbs synthesis and secretion which are described by 28 equations and 31 parameters in total.

The overall culture material balance is given by:

$$\frac{\mathrm{d}V_\mathrm{c}}{\mathrm{d}t} = F_\mathrm{in} - F_\mathrm{out} \quad , \tag{D.1}$$

The growth and death of the cells part models the life of the cells influenced by nutrients (i.e., glucose and glutamine) and by-products (i.e., lactate and ammonia). It is described by:

$$\frac{\mathrm{d}(V_\mathrm{c}X_\mathrm{v})}{\mathrm{d}t} = \mu V_\mathrm{c}X_\mathrm{v} - \mu_\mathrm{d}V_\mathrm{c}X_\mathrm{v} - F_\mathrm{out}X_\mathrm{v} \quad , \tag{D.2}$$

$$\frac{\mathrm{d}(V_\mathrm{c}X_\mathrm{t})}{\mathrm{d}t} = \mu V_\mathrm{c}X_\mathrm{v} - F_\mathrm{out}X_\mathrm{t} \quad , \tag{D.3}$$

$$\mu = \mu_\mathrm{max}f_\mathrm{lim}f_\mathrm{inh} \quad , \tag{D.4}$$

$$f_\mathrm{lim} = \left(\frac{c_\mathrm{glc}}{K_\mathrm{glc}+c_\mathrm{glc}}\right)\left(\frac{c_\mathrm{gln}}{K_\mathrm{gln}+c_\mathrm{gln}}\right) \quad , \tag{D.5}$$

$$f_\mathrm{inh} = \left(\frac{KI_\mathrm{lac}}{KI_\mathrm{lac}+c_\mathrm{lac}}\right)\left(\frac{KI_\mathrm{amm}}{KI_\mathrm{amm}+c_\mathrm{amm}}\right) \quad , \tag{D.6}$$

$$\mu_\mathrm{d} = \frac{\mu_\mathrm{d,max}}{1+\left(K_\mathrm{d,amm}/c_\mathrm{amm}\right)^{a_\mathrm{d}}} \quad \text{with } a_\mathrm{d} > 1, \tag{D.7}$$

The cell metabolism part models the consumption of nutrients and their conversion into by-products. It is described by:

$$\frac{\mathrm{d}(V_\mathrm{c}c_\mathrm{glc})}{\mathrm{d}t} = -Q_\mathrm{glc}V_\mathrm{c}X_\mathrm{v} + F_\mathrm{in}c_\mathrm{glc,in} - F_\mathrm{out}c_\mathrm{glc} \quad , \tag{D.8}$$

$$Q_\mathrm{glc} = \frac{\mu}{Y_{x,\mathrm{glc}}} + m_\mathrm{glc} \quad , \tag{D.9}$$

$$\frac{\mathrm{d}(V_\mathrm{c}c_\mathrm{gln})}{\mathrm{d}t} = -Q_\mathrm{gln}V_\mathrm{c}X_\mathrm{v} - K_\mathrm{d,gln}V_\mathrm{c}c_\mathrm{gln} + F_\mathrm{in}c_\mathrm{gln,in} - F_\mathrm{out}c_\mathrm{gln} \quad , \tag{D.10}$$

$$Q_{\mathrm{gln}} = \frac{\mu}{Y_{x,\mathrm{gln}}} + m_{\mathrm{gln}} \quad , \tag{D.11}$$

$$m_{\mathrm{gln}} = \frac{\alpha_1 c_{\mathrm{gln}}}{\alpha_2 + c_{\mathrm{gln}}} \quad , \tag{D.12}$$

$$\frac{\mathrm{d}(V_{\mathrm{c}} c_{\mathrm{lac}})}{\mathrm{d}t} = Q_{\mathrm{lac}} V_{\mathrm{c}} X_{\mathrm{v}} - F_{\mathrm{out}} c_{\mathrm{lac}} \quad , \tag{D.13}$$

$$Q_{\mathrm{lac}} = Y_{\mathrm{lac,glc}} Q_{\mathrm{glc}} \quad , \tag{D.14}$$

$$\frac{\mathrm{d}(V_{\mathrm{c}} c_{\mathrm{amm}})}{\mathrm{d}t} = -Q_{\mathrm{amm}} V_{\mathrm{c}} X_{\mathrm{v}} + K_{\mathrm{d,gln}} V_{\mathrm{c}} c_{\mathrm{gln}} - F_{\mathrm{out}} c_{\mathrm{amm}} \quad , \tag{D.15}$$

$$Q_{\mathrm{amm}} = Y_{\mathrm{amm,gln}} Q_{\mathrm{gln}} \quad , \tag{D.16}$$

Finally, the synthesis and secretion of mAbs part is a structured one that models the kinetics of the amino acid chains assembly to create mAbs. It is described by:

$$\frac{\mathrm{d}m_{\mathrm{H}}}{\mathrm{d}t} = N_{\mathrm{H}} S_{\mathrm{H}} - K_{\mathrm{RNA}} m_{\mathrm{H}} \quad , \tag{D.17}$$

$$\frac{\mathrm{d}m_{\mathrm{L}}}{\mathrm{d}t} = N_{\mathrm{L}} S_{\mathrm{L}} - K_{\mathrm{RNA}} m_{\mathrm{L}} \quad , \tag{D.18}$$

$$\frac{\mathrm{d}c_{\mathrm{H}}}{\mathrm{d}t} = T_{\mathrm{H}} m_{\mathrm{H}} - R_{\mathrm{H}} \quad , \tag{D.19}$$

$$\frac{\mathrm{d}c_{\mathrm{L}}}{\mathrm{d}t} = T_{\mathrm{L}} m_{\mathrm{L}} - R_{\mathrm{L}} \quad , \tag{D.20}$$

$$R_{\mathrm{H}} = \frac{2}{3} K_{\mathrm{A}} c_{\mathrm{H}}^2 \quad , \tag{D.21}$$

$$R_{\mathrm{L}} = 2K_{\mathrm{A}} c_{\mathrm{H_2}} c_{\mathrm{L}} + K_{\mathrm{A}} c_{\mathrm{H_2L}} c_{\mathrm{L}} \quad , \tag{D.22}$$

$$\frac{\mathrm{d}c_{\mathrm{H_2}}}{\mathrm{d}t} = \frac{1}{3} K_{\mathrm{A}} c_{\mathrm{H}}^2 - 2K_{\mathrm{A}} c_{\mathrm{H_2}} c_{\mathrm{L}} \quad , \tag{D.23}$$

$$\frac{\mathrm{d}c_{\mathrm{H_2L}}}{\mathrm{d}t} = 2K_{\mathrm{A}} c_{\mathrm{H_2}} c_{\mathrm{L}} - K_{\mathrm{A}} c_{\mathrm{H_2L}} c_{\mathrm{L}} \quad , \tag{D.24}$$

$$\frac{\mathrm{d}c_{\mathrm{H_2L_2}}^{\mathrm{ER}}}{\mathrm{d}t} = K_{\mathrm{A}} c_{\mathrm{H_2L}} c_{\mathrm{L}} - K_{\mathrm{ER}} c_{\mathrm{H_2L_2}}^{\mathrm{ER}} \quad , \tag{D.25}$$

$$\frac{\mathrm{d}c_{\mathrm{H_2L_2}}^{\mathrm{G}}}{\mathrm{d}t} = \varepsilon_1 K_{\mathrm{ER}} c_{\mathrm{H_2L_2}}^{\mathrm{ER}} - K_{\mathrm{G}} c_{\mathrm{H_2L_2}}^{\mathrm{G}} \quad , \tag{D.26}$$

$$\frac{\mathrm{d}(V_{\mathrm{c}} c_{\mathrm{mAb}})}{\mathrm{d}t} = (\gamma_2 - \gamma_1 \mu) Q_{\mathrm{mAb}} V_{\mathrm{c}} X_{\mathrm{v}} - F_{\mathrm{out}} c_{\mathrm{mAb}} \quad , \tag{D.27}$$

$$Q_{\mathrm{mAb}} = \varepsilon_2 \xi_{\mathrm{mAb}} K_{\mathrm{G}} c_{\mathrm{H_2L_2}}^{\mathrm{G}} \quad . \tag{D.28}$$

Table D.1 reports the list of the parameters with the corresponding mean and standard deviations used for process batch generation.

Table D.1 *Mean (reference) and standard deviation values of the parameters used in process batch generation. Missing standard deviations represent that the parameter is kept constant at the reference value.*

Parameter	Kontoravdi et al. (2010) [mean]	Standard deviation
μ_{max} (h^{-1})	0.058	0.0068
$\mu_{\mathrm{d,max}}$ (h^{-1})	0.03	0.0025
K_{glc} (mM)	0.75	-
K_{gln} (mM)	0.075	-
KI_{lac} (mM)	171.76	-
KI_{amm} (mM)	28.48	-
$K_{\mathrm{d,amm}}$ (mM)	1.76	0.4253
a_{d} (-)	2	-
$Y_{x,\mathrm{glc}}$ (cell/mmol)	2.6×10^{8}	3.1×10^{7}
m_{glc} (mmol/(cell h))	4.9×10^{-14}	-
$Y_{x,\mathrm{gln}}$ (cell/mmol)	8.0×10^{8}	1.6×10^{8}
α_1 (mmol L/(cell h))	3.4×10^{-13}	-
α_2 (mM)	4.0	-
$Y_{\mathrm{lac,glc}}$ (mmol/mmol)	2.0	-
$Y_{\mathrm{amm,gln}}$ (mmol/mmol)	0.45	0.0825
$K_{\mathrm{d,gln}}$ (h^{-1})	9.6×10^{-3}	0.003
N_{H} (gene/cell)	100.0	-
S_{H} (mRNA/(gene h))	3000.0	-
K_{RNA} (h^{-1})	0.1	-
N_{L} (gene/cell)	100.0	-
S_{L} (mRNA/(gene h))	4500.0	-
K_{A} (cell/(molecule L))	1.0×10^{-6}	-
T_{H} (chain/(mRNA h))	17.0	-
T_{L} (chain/(mRNA h))	11.5	-
K_{ER} (h^{-1})	0.69	-
ε_1 (-)	0.995	0.1492
K_{G} (h^{-1})	0.14	-
γ_1 (-)	0.10	-
γ_2 (h)	2.0	0.333
ε_2 (-)	1.0	0.15
ξ_{mAb} (g/mol)	2.5×10^{-16}	-

D.2 First principles digital model

The FPDM model used in Chapter 5 to generate *in silico* batches is reported in the following. It is a modified version of the simplified mathematical model describing a fed-batch mAbs production process (Jimenez del Val, Fan, et al., 2016).

$$\frac{\mathrm{d}V_\mathrm{c}}{\mathrm{d}t} = F_\mathrm{in} - F_\mathrm{out} \quad , \tag{D.29}$$

The growth and death of the cells part models the life of the cells influenced by nutrients (i.e., glucose and glutamine) and by-products (i.e., lactate and ammonia). It is described by:

$$\frac{\mathrm{d}(V_\mathrm{c}X_\mathrm{v})}{\mathrm{d}t} = \mu V_\mathrm{c}X_\mathrm{v} - \mu_\mathrm{d}V_\mathrm{c}X_\mathrm{v} - F_\mathrm{out}X_\mathrm{v} \quad , \tag{D.30}$$

$$\mu = \mu_\mathrm{max}\left(\frac{c_\mathrm{glc}}{K_\mathrm{glc}+c_\mathrm{glc}}\right) - \frac{X_\mathrm{v}}{\alpha_\mathrm{x}}f_\mathrm{lim} \quad , \tag{D.31}$$

$$f_\mathrm{lim} = \frac{c_\mathrm{gln}}{c_\mathrm{gln}+K_\mathrm{gln}} \quad , \tag{D.32}$$

$$\mu_\mathrm{d} = \mu_\mathrm{d,max}\left(\frac{K_\mathrm{d}}{K_\mathrm{d}+\mu}\right) \quad , \tag{D.33}$$

The cell metabolism part models the consumption of nutrients and their conversion into by-products. It is described by:

$$\frac{\mathrm{d}(V_\mathrm{c}c_\mathrm{glc})}{\mathrm{d}t} = F_\mathrm{in}c_\mathrm{glc,in} - F_\mathrm{out}c_\mathrm{glc} - Q_\mathrm{glc}X_\mathrm{v}V_\mathrm{c}\left(f_\mathrm{lim} + m_\mathrm{glc}\right) \quad , \tag{D.34}$$

$$Q_\mathrm{glc} = \frac{\mu}{Y_{x,\mathrm{glc}}}\left(\frac{c_\mathrm{glc}}{K_\mathrm{glc}+c_\mathrm{glc}}\right) \quad , \tag{D.35}$$

$$\frac{\mathrm{d}(V_\mathrm{c}c_\mathrm{gln})}{\mathrm{d}t} = -\left(\frac{\mu}{Y_{x,\mathrm{gln}}}\right)X_\mathrm{v}V_\mathrm{c} \quad , \tag{D.36}$$

$$\frac{\mathrm{d}(V_\mathrm{c}c_\mathrm{lac})}{\mathrm{d}t} = Q_\mathrm{lac}V_\mathrm{c}X_\mathrm{v} - Q_\mathrm{lac,cons}V_\mathrm{c}X_\mathrm{v} - F_\mathrm{out}c_\mathrm{lac} \quad , \tag{D.37}$$

$$Q_\mathrm{lac} = Y_\mathrm{lac,glc}Q_\mathrm{glc} \quad , \tag{D.38}$$

$$Q_\mathrm{lac,cons} = \frac{1}{Y_\mathrm{x,lac}}\left(\frac{c_\mathrm{lac}}{K_\mathrm{lac}+c_\mathrm{lac}}\right) \quad , \tag{D.39}$$

Finally, the synthesis of mAbs is described by:

$$\frac{\mathrm{d}(V_\mathrm{c}c_\mathrm{mAb})}{\mathrm{d}t} = Q_\mathrm{mAb}V_\mathrm{c}X_\mathrm{v} - F_\mathrm{out}c_\mathrm{mAb} \quad , \tag{D.40}$$

$$Q_\mathrm{mAb} = Y_\mathrm{mAb,glc}Q_\mathrm{glc} \quad . \tag{D.41}$$

The lists of the parameters for *in silico* batch generation in the first principles digital model are shown in Table D.2.

Table D.2 *Reference, minimum and maximum values of the parameters used for first principles in silico batches generation. Missing ranges represent that the parameter is kept constant at the reference value.*

Parameter	Reference	Minimum	Maximum
μ_{max} (h^{-1})	0.073	0.058	0.09
K_{glc} (mM)	0.01	-	-
α_{x} (10^5 cell/mmol)	44704	-	-
$\mu_{\mathrm{d,max}}$ (h^{-1})	0.02	0.015	0.041
K_{d} (h^{-1})	0.635	-	-
$Y_{\mathrm{x,glc}}$ (10^5 cell/mmol)	65341	47700	80700
$Y_{\mathrm{lac,glc}}$ (mmol/mmol)	1.7	-	-
$Y_{\mathrm{x,lac}}$ (10^5 cell/mmol)	182050	-	-
K_{lac} (mM)	3.908	-	-
$Y_{\mathrm{mAb,glc}}$ (10^5 cell/mmol)	150.0	100	180
K_{gln} (mM)	0.02	0.02	0.05
$Y_{\mathrm{x,gln}}$ (10^5 cell/mmol)	8000	7000	11000
m_{glc} (-)	0.2	-	

D.3 Parameters for *in silico* data generation by hybrid digital model

The lists of the parameters for *in silico* batch generation in the hybrid digital model are shown in Table D.3.

Table D.3 *Mean (training) and standard deviation values of the parameters used for hybrid in silico batches generation.*

Parameter	Training (Mean)	Standard deviation
$\mu_{\mathrm{max,X_v}}$	2.0	0.13
$\mu_{\mathrm{max,glc}}$	8.0	0.27
$\mu_{\mathrm{max,gln}}$	3.0	0.10
$\mu_{\mathrm{max,lac}}$	8.0	0.53
$\mu_{\mathrm{max,amm}}$	2.0	0.13
$\mu_{\mathrm{max,mAb}}$	2.0	0.13

Appendix E

Experimental campaigns for feeding schedule optimization

E.1 Experimental campaigns

The values of the dynamic subfactors of the planned experiments in the experimental campaign A and B are reported in Table E.1 and E.2, respectively.

Table E.1 *Dynamic subfactor values of the experiments planned in experimental campaign A.*

Experiment	x_1^{glc}	x_2^{glc}	x_3^{glc}	x_1^{gln}	x_2^{gln}	x_3^{gln}
1	0.394	0.131	0.475	0.010	-0.475	0.515
2	-0.131	-0.636	-0.232	-0.253	-0.717	-0.030
3	0.212	-0.596	-0.192	0.152	-0.273	0.576
4	-0.131	-0.657	0.212	0.596	0.394	-0.010
5	0.434	0.556	0.010	-0.495	0.293	0.212
6	-0.596	0.374	-0.030	0.293	-0.556	-0.152
7	-0.051	0.273	-0.677	-0.192	-0.212	0.5956
8	-0.030	-0.677	-0.293	-0.313	0.657	0.030
9	-0.919	-0.051	-0.030	0.374	-0.192	0.434
10	0.7171	0.071	-0.212	0.596	-0.394	0.010
11	-0.697	-0.051	0.253	-0.091	-0.010	-0.899
12	-0.576	0.313	-0.111	-0.616	-0.253	0.131
13	-0.333	0.455	0.212	-0.051	0.636	-0.313
14	0.152	0.616	0.232	-0.192	-0.596	-0.212
15	0.636	-0.253	-0.111	-0.616	-0.293	0.091
16	0.091	0.313	0.596	0.7171	-0.212	-0.071
17	0.616	-0.273	0.111	0.172	0.818	0.010
18	-0.576	-0.111	0.313	-0.051	0.010	-0.111
19	-0.253	-0.576	0.172	-0.071	0.859	-0.07
20	0.232	-0.677	-0.091	-0.192	0.030	-0.778
21	-0.677	0.030	0.293	-0.030	-0.798	0.172
22	-0.091	0.010	-0.111	0.374	0.071	0.556
23	-0.030	-0.838	0.131	-0.455	0.091	0.455
24	0.152	0.778	-0.071	0.374	0.495	0.131
25	-0.051	-0.010	-0.939	0.596	0.232	-0.172
26	0.576	0.394	-0.030	0.051	0.313	-0.636
27	-0.172	0.232	0.596	-0.010	0.434	0.556

28	0.051	-0.657	0.293	0.030	-0.737	-0.232
29	0.051	0.232	-0.717	-0.657	0.071	-0.273
30	-0.717	0.273	-0.010	-0.010	0.717	0.273
31	0.172	-0.273	0.556	-0.697	0.273	-0.030

Table E.2 *Dynamic subfactor values of the experiments planned in experimental campaign B.*

Experiment	$\boldsymbol{x_1^{glc}}$	$\boldsymbol{x_2^{glc}}$	$\boldsymbol{x_3^{glc}}$	$\boldsymbol{x_1^{gln}}$	$\boldsymbol{x_2^{gln}}$	$\boldsymbol{x_3^{gln}}$
1	-0.495	0.455	0.0505	-0.152	-0.131	-0.717
2	0.697	-0.293	-0.010	-0.212	0.071	-0.717
3	0.737	0.152	-0.111	0.394	-0.333	0.273
4	-0.677	-0.293	0.030	-0.111	-0.434	-0.455
5	-0.333	-0.192	0.475	0.677	-0.111	-0.212
6	-0.576	-0.293	-0.131	-0.051	-0.434	0.515
7	0.354	-0.071	0.576	-0.010	0.758	0.232
8	0.374	0.293	0.333	-0.071	-0.535	0.394
9	-0.535	0.071	-0.394	0.192	0.717	0.091

Appendix F

^{13}C intracellular reactions and metabolites

F.1 Intracellular reactions

The intracellular reactions available in the ^{13}C isotope labeling dataset are listed in Table F.1.

Table F.1 *Metabolic reaction available in the ^{13}C labeling dataset: (a) glycolysis, (b) TCA cycle, (c) pentose phosphate pathway, (d) amino acid metabolism, and (e) other metabolic reactions.*

(a)

Reaction	Formula
HK	Glc → G6P
PGI	G6P ↔ F6P
PFK	F6P → DHAP + GAP
TPI	DHAP ↔ GAP
GAPDH	GAP ↔ 3PG
ENO	3PG ↔ PEP
PK	PEP → Pyr

(b)

Reaction	Formula
PDH	Pyr → AcCoA + CO_2
SDH	Suc ↔ Fum
FUS	Fum ↔ Mal
MDH	Mal ↔ OAA
CS	OAA + AcCoA → Cit
ADH	aKG.m → Suc + CO_2
IDH	Cit ↔ aKG.m + CO_2

(c)

Reaction	Formula
G6PDH	G6P → Ru5P + CO_2
PPE	Ru5P ↔ X5P
PPI	Ru5P ↔ R5P
TKT1	X5P + R5P ↔ GAP + S7P
TAL	S7P + GAP ↔ E4P + F6P
TKT2	X5P + E4P ↔ GAP + F6P

(d)

Reaction	Formula
GS	Gln ↔ Glu.m
ASNS	Asp ↔ Asn
SHMT	Ser ↔ Gly + MEETHF
GCS	CO_2 + MEETHF → Gly
MAT	Met + Ser → Cys + Suc
PAH	Phe → Tyr
TDH	Thr → Pyr + CO_2
HAL	His → Glu.c
PCC	PropCoA + CO_2 → Suc
TDO	Trp → 2 CO_2 + Ala + aKA
AKD	aKA → 2 CO_2 + 2 AcCoA
GDH	aKG.m ↔ Glu.m
ALT	Ala + aKG.c ↔ Pyr + Glu.c
SBCAD	Ile + aKG.c → AcCoA + CO_2 + PropCoA + Glu.c
IVD	Leu + aKG.c + CO_2 → CO_2 + 3 AcCoA + Glu.c
TTA	Tyr + aKG.c → CO_2 + Mal + 2 AcCoA + Glu.c
IBD	Val + aKG.c → Glu.c + 2 CO_2 + PropCoA
AASS	Lys + 2 aKG.c → 2 Glu.c + aKA
AST	OAA + Glu.c ↔ Asp + aKG.c
ARGS	Arg + aKG.c → 2 Glu.c + Urea
PGHDH	3PG + Glu.c → Ser + aKG.c
CDO	Cys + aKG.c → Pyr + Glu.c

(e)

Reaction	Formula
LDH	Lac↔ Pyr
aKG.m	aKG.m → aKG.c
ME	Mal → Pyr + CO_2
PC	Pyr + CO_2 → OAA
ACL	Cit → AcCoA + Mal

F.2 Metabolites

The metabolites available in the metabolic network used in the ^{13}C labeling experiments are listed in Table F.2.

Table F.2 *Metabolites available in the ^{13}C labeling dataset.*

Symbol	Name
3PG	3-Phosphoglyceric acid
AcCoA	acetyl coenzyme A
aKA	α-Ketoisocaproic acid
aKG.c	α-ketoglutarate in cytosol
aKG.m	α-ketoglutarate in mitochondria
Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
Cit	Citric acid
Cys	Cysteine
DHAP	Dihydroxyacetone phosphate
E4P	Erythrose 4-phosphate
F6P	Fructose 6-phosphate
Fum	Fumarate
G6P	Glucosio-6-fosfato
GAP	Glyceraldehyde 3-phosphate
Glc	Glucose
Gln	Glutamine
Glu.c	Glutamic acid in cytosol
Glu.m	Glutamic acid in mitochondria
Gly	Glycine
His	Histidine
Ile	Isoleucine
Lac	Lactate
Leu	Leucine
Lys	Lysine
Mal	Maltate
Met	Methionine
MEETHF	5,10-methylene-H_4folate
OAA	Oxaloacetic acid
PEP	Phosphoenolpyruvic acid
Phe	Phenylalanine
PropCoA	Propionyl coenzyme A
Pyr	Pyruvate
Ru5P	Ribulose 5-Phosphate
R5P	Ribose 5-phosphate
S7P	sedoheptulose 7-phosphate
Ser	Serine
Suc	Succinic acid
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
Urea	Urea
Val	Valine
X5P	Xylulose 5-phosphate

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *Network: Computation in Neural Systems*, *16*(2–3), 121–138. https://doi.org/https://doi.org/10.48550/arXiv.1603.04467

Afanador, N. L., Tran, T. N., & Buydens, L. M. C. (2013). Use of the bootstrap and permutation methods for a more robust variable importance in the projection metric for partial least squares regression. *Analytica Chimica Acta*, *768*(1), 49–56. https://doi.org/10.1016/j.aca.2013.01.004

Ahuja, S., Jain, S., & Ram, K. (2015). Application of multivariate analysis and mass transfer principles for refinement of a 3-L bioreactor scale-down model-when shake flasks mimic 15,000-L bioreactors better. *Biotechnology Progress*, *31*(5), 1370–1380. https://doi.org/10.1002/btpr.2134

Antonakoudis, A., Barbosa, R., Kotidis, P., & Kontoravdi, C. (2020). The era of big data: Genome-scale modelling meets machine learning. *Computational and Structural Biotechnology Journal*, *18*, 3287–3300. https://doi.org/10.1016/j.csbj.2020.10.011

Antonakoudis, A., Strain, B., Barbosa, R., Jimenez del Val, I., & Kontoravdi, C. (2021). Synergising stoichiometric modelling with artificial neural networks to predict antibody glycosylation patterns in Chinese hamster ovary cells. *Computers & Chemical Engineering*, *154*, 107471. https://doi.org/10.1016/j.compchemeng.2021.107471

Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., & Rasool, M. H. (2017). Proteomics: Technologies and Their Applications. *Journal of Chromatographic Science*, *55*(2), 182–196. https://doi.org/10.1093/chromsci/bmw167

Badr, S., & Sugiyama, H. (2020). A PSE perspective for the efficient production of monoclonal antibodies: integration of process, cell, and product design aspects. *Current Opinion in Chemical Engineering*, *27*, 121–128. https://doi.org/10.1016/j.coche.2020.01.003

Badsha, M. B., Kurata, H., Onitsuka, M., Oga, T., & Omasa, T. (2016). Metabolic analysis of antibody producing Chinese hamster ovary cell culture under different stresses conditions. *Journal of Bioscience and Bioengineering*, *122*(1), 117–124. https://doi.org/10.1016/j.jbiosc.2015.12.013

Banner, M., Alosert, H., Spencer, C., Cheeks, M., Farid, S. S., Thomas, M., & Goldrick, S. (2021). A decade in review: use of data analytics within the biopharmaceutical sector. *Current Opinion in Chemical Engineering*, *34*, 100758. https://doi.org/10.1016/j.coche.2021.100758

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

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*(3), 166–173. https://doi.org/10.1002/cem.785

Batra, J., & Rathore, A. S. (2016). Glycosylation of monoclonal antibody products: Current status and future prospects. *Biotechnology Progress*, *32*(5), 1091–1102. https://doi.org/10.1002/btpr.2366

Baughman, D. R., & Liu, Y. A. (1995). Neural Networks in Bioprocessing and Chemical Engineering. In *Neural Networks in Bioprocessing and Chemical Engineering*. Elsevier. https://doi.org/10.1016/c2009-0-21189-5

Bayer, B., Dalmau Diaz, R., Melcher, M., Striedner, G., & Duerkop, M. (2021). Digital Twin Application for Model-Based DoE to Rapidly Identify Ideal Process Conditions for Space-Time Yield Optimization. *Processes*, *9*(7), 1109. https://doi.org/10.3390/pr9071109

Bayer, B., Stosch, M., Striedner, G., & Duerkop, M. (2020). Comparison of Modeling Methods for DoE-Based Holistic Upstream Process Characterization. *Biotechnology Journal*, *15*(5), 1900551. https://doi.org/10.1002/biot.201900551

Bayer, B., Striedner, G., & Duerkop, M. (2020). Hybrid Modeling and Intensified DoE: An Approach to Accelerate Upstream Process Characterization. *Biotechnology Journal*, *15*(9), 2000121. https://doi.org/10.1002/biot.202000121

Becker, S. A., & Palsson, B. O. (2008). Context-Specific Metabolic Networks Are Consistent with Experiments. *PLoS Computational Biology*, *4*(5), e1000082. https://doi.org/10.1371/journal.pcbi.1000082

Birch, J. R., & Racher, A. J. (2006). Antibody production. *Advanced Drug Delivery Reviews*, *58*, 671–685. https://doi.org/10.1016/j.addr.2005.12.006

Boccard, J., & Rudaz, S. (2014). Harnessing the complexity of metabolomic data with chemometrics. *Journal of Chemometrics*, *28*(1), 1–9. https://doi.org/10.1002/cem.2567

Bordel, S., Agren, R., & Nielsen, J. (2010). Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Computational Biology*, *6*(7), 16. https://doi.org/10.1371/journal.pcbi.1000859

Brunner, M., Kolb, K., Keitel, A., Stiefel, F., Wucherpfennig, T., Bechmann, J., Unsoeld, A., & Schaub, J. (2021). Application of metabolic modeling for targeted optimization of high seeding density processes. *Biotechnology and Bioengineering*, *118*(5), 1793–1804. https://doi.org/10.1002/bit.27693

Bryan, K., Brennan, L., & Cunningham, P. (2008). MetaFIND: A feature analysis tool for metabolomics data. *BMC Bioinformatics*, *9*(1), 470. https://doi.org/10.1186/1471-2105-9-470

Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, *84*(6), 647–657. https://doi.org/10.1002/bit.10803

Cai, Q., Alvarez, J. A., Kang, J., & Yu, T. (2017). Network Marker Selection for Untargeted LC–MS Metabolomics Data. *Journal of Proteome Research*, *16*(3), 1261–1269. https://doi.org/10.1021/acs.jproteome.6b00861

Calmels, C., McCann, A., Malphettes, L., & Andersen, M. R. (2019). Application of a curated genome-scale metabolic model of CHO DG44 to an industrial fed-batch process. *Metabolic Engineering*, *51*(August 2018), 9–19. https://doi.org/10.1016/j.ymben.2018.09.009

Carinhas, N., Duarte, T. M., Barreiro, L. C., Carrondo, M. J. T., Alves, P. M., & Teixeira, A. P. (2013). Metabolic signatures of GS-CHO cell clones associated with butyrate treatment and culture phase transition. *Biotechnology and Bioengineering*, *110*(12), 3244–3257. https://doi.org/10.1002/bit.24983

Carvalho, M., Riesberg, J., & Budman, H. (2022). Hybrid model to predict the effect of complex media changes in mammalian cell cultures. *Biochemical Engineering Journal*, *186*(April), 108560. https://doi.org/10.1016/j.bej.2022.108560

Castelli, M. S., McGonigle, P., & Hornby, P. J. (2019). The pharmacology and therapeutic applications of monoclonal antibodies. *Pharmacology Research & Perspectives*, *7*(6), e00535. https://doi.org/10.1002/prp2.535

Chang, K. L., Pee, H. N., Tan, W. P., Dawe, G. S., Holmes, E., Nicholson, J. K., Chan, E. C. Y., & Ho, P. C. (2015). Metabolic Profiling of CHO-AβPP695 Cells Revealed Mitochondrial Dysfunction Prior to Amyloid-β Pathology and Potential Therapeutic Effects of Both PPARγ and PPARα Agonisms for Alzheimer's Disease. *Journal of Alzheimer's Disease*, *44*(1), 215–231. https://doi.org/10.3233/JAD-140429

Chartrain, M., & Chu, L. (2008). Development and Production of Commercial Therapeutic Monoclonal Antibodies in Mammalian Cell Expression Systems: An Overview of the Current Upstream Technologies. *Current Pharmaceutical Biotechnology*, *9*(6), 447–467. https://doi.org/10.2174/138920108786786367

Chawla, N. V, Bowyer, K. W., & Hall, L. O. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, Y., & Ierapetritou, M. (2020). A framework of hybrid model development with identification of plant-model mismatch. *AIChE Journal*, *66*(10), 1–16. https://doi.org/10.1002/aic.16996

Chen, Z.-S., Zhu, B., He, Y.-L., & Yu, L.-A. (2017). A PSO based virtual sample generation method for small sample sets: Applications to regression datasets. *Engineering Applications of Artificial Intelligence*, *59*(January), 236–243. https://doi.org/10.1016/j.engappai.2016.12.024

Chiu, M. L., Goulet, D. R., Teplyakov, A., & Gilliland, G. L. (2019). Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies*, *8*(4), 55. https://doi.org/10.3390/antib8040055

Chong, J., Wishart, D. S., & Xia, J. (2019). Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Current Protocols in Bioinformatics*, *68*(1), 1–128. https://doi.org/10.1002/cpbi.86

Chong, William P K, Goh, L. T., Reddy, S. G., Yusufi, F. N. K., Lee, D. Y., Wong, N. S. C., Heng, C. K., Yap, M. G. S., & Ho, Y. S. (2009). Metabolomics profiling of extracellular metabolites in recombinant Chinese Hamster Ovary fed-batch culture. *Rapid Communications in Mass Spectrometry*, *23*(23), 3763–3771. https://doi.org/10.1002/rcm.4328

Chong, William Pooi Kat, Thng, S. H., Hiu, A. P., Lee, D.-Y., Chan, E. C. Y., & Ho, Y. S. (2012). LC-MS-based metabolic characterization of high monoclonal antibody-producing Chinese hamster ovary cells. *Biotechnology and Bioengineering*, *109*(12), 3103–3111. https://doi.org/10.1002/bit.24580

Chrysanthopoulos, P. K., Goudar, C. T., & Klapa, M. I. (2010). Metabolomics for high-resolution monitoring of the cellular physiological state in cell culture engineering. *Metabolic Engineering*, *12*(3), 212–222. https://doi.org/10.1016/j.ymben.2009.11.001

Clarke, C., Doolan, P., Barron, N., Meleady, P., O'Sullivan, F., Gammell, P., Melville, M., Leonard, M., & Clynes, M. (2011). Predicting cell-specific productivity from CHO gene expression. *Journal of Biotechnology*, *151*(2), 159–165. https://doi.org/10.1016/j.jbiotec.2010.11.016

Conesa, A., Bro, R., García-García, F., Prats, J. M., Götz, S., Kjeldahl, K., Montaner, D., & Dopazo, J. (2008). Direct functional assessment of the composite phenotype through multivariate projection strategies. *Genomics*, *92*(6), 373–383. https://doi.org/10.1016/j.ygeno.2008.05.015

Culley, C., Vijayakumar, S., Zampieri, G., & Angione, C. (2020). A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proceedings of the National Academy of Sciences*, *117*(31), 18869–18879. https://doi.org/10.1073/pnas.2002959117

Davis, R. A., Charlton, A. J., Oehlschlager, S., & Wilson, J. C. (2006). Novel feature selection method for genetic programming using metabolomic 1H NMR data. *Chemometrics and Intelligent Laboratory Systems*, *81*(1), 50–59. https://doi.org/10.1016/j.chemolab.2005.09.006

de Aguiar, P. F., Bourguignon, B., Khots, M. S., Massart, D. L., & Phan-Than-Luu, R. (1995). D-optimal designs. *Chemometrics and Intelligent Laboratory Systems*, *30*(2), 199–210. https://doi.org/10.1016/0169-7439(94)00076-X

De Veaux, R. D., Schumi, J., Schweinsberg, J., & Ungar, L. H. (1998). Prediction Intervals for Neural Networks via Nonlinear Regression. *Technometrics*, *40*(4), 273–282. https://doi.org/10.1080/00401706.1998.10485556

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

Deshpande, R., Yang, T. H., & Heinzle, E. (2009). Towards a metabolic and isotopic steady state in CHO batch cultures for reliable isotope-based metabolic profiling. *Biotechnology Journal*, *4*(2), 247–263. https://doi.org/10.1002/biot.200800143

Destro, F., & Barolo, M. (2022). A review on the modernization of pharmaceutical development and manufacturing – Trends, perspectives, and the role of mathematical modeling. *International Journal of Pharmaceutics*, *620*(February), 121715. https://doi.org/10.1016/j.ijpharm.2022.121715

Destro, F., Facco, P., García Muñoz, S., Bezzo, F., & Barolo, M. (2020). A hybrid framework for process monitoring: Enhancing data-driven methodologies with state and parameter estimation. *Journal of Process Control*, *92*, 333–351. https://doi.org/10.1016/j.jprocont.2020.06.002

Dickson, A. J. (2014). Enhancement of production of protein biopharmaceuticals by mammalian cell cultures: The metabolomics perspective. *Current Opinion in Biotechnology*, *30*, 73–79. https://doi.org/10.1016/j.copbio.2014.06.004

Dietmair, S., Hodson, M. P., Quek, L.-E., Timmins, N. E., Gray, P., & Nielsen, L. K. (2012). A Multi-Omics Analysis of Recombinant Protein Production in Hek293 Cells. *PLoS ONE*, *7*(8), e43394. https://doi.org/10.1371/journal.pone.0043394

Dietmair, S., Hodson, M. P., Quek, L. E., Timmins, N. E., Chrysanthopoulos, P., Jacob, S. S., Gray, P., & Nielsen, L. K. (2012). Metabolite profiling of CHO cells with different growth characteristics. *Biotechnology and Bioengineering*, *109*(6), 1404–1414. https://doi.org/10.1002/bit.24496

Dors, M., Simutis, R., & Lübbert, A. (1996). Hybrid Process Modeling for Advanced Process State Estimation, Prediction, and Control Exemplified in a Production- Scale Mammalian Cell Culture. In K. R. Rogers, A. Mulchandani, & W. Zhou (Eds.), *Biosensor and Chemical Sensor Technology* (ACS Sympos). American Chemical Society.

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., & Palsson, B. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(6), 1777–1782. https://doi.org/10.1073/pnas.0610772104

Duran-Villalobos, C. A., Ogonah, O., Melinek, B., Bracewell, D. G., Hallam, T., & Lennox, B. (2021). Multivariate statistical data analysis of cell-free protein synthesis toward monitoring and control. *AIChE Journal*, *67*(6), 1–12. https://doi.org/10.1002/aic.17257

Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology*, *7*(1), 74. https://doi.org/10.1186/1752-0509-7-74

Epifa. (2021). *The Pharmaceutical Industry in Figures: Key data 2021*. https://www.efpia.eu/publications/downloads/efpia/the-pharmaceutical-industry-in-figures-2021/

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C., & Wold, S. (2006). *Multi-and megavariate data analysis*. Umetrics Ab.

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

Fan, Y., Jimenez Del Val, I., Müller, C., Wagtberg Sen, J., Rasmussen, S. K., Kontoravdi, C., Weilguny, D., & Andersen, M. R. (2015). Amino acid and glucose metabolism in fed-batch CHO cell culture affects antibody production and glycosylation. *Biotechnology and Bioengineering*, *112*(3), 521–535. https://doi.org/10.1002/bit.25450

Farid, S. S., Baron, M., Stamatis, C., Nie, W., & Coffman, J. (2020). Benchmarking biopharmaceutical process development and manufacturing cost contributions to R&amp;D. *MAbs*, *12*(1), 1754999. https://doi.org/10.1080/19420862.2020.1754999

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., & Palsson, B. Ø. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, *3*(1), 121. https://doi.org/10.1038/msb4100155

Fernández Pierna, J. A., Abbas, O., Baeten, V., & Dardenne, P. (2009). A Backward Variable Selection method for PLS regression (BVSPLS). *Analytica Chimica Acta*, *642*(1–2), 89–93. https://doi.org/10.1016/j.aca.2008.12.002

Ferreira, A. R., Dias, J. M. L., von Stosch, M., Clemente, J., Cunha, A. E., & Oliveira, R. (2014). Fast development of Pichia pastoris GS115 Mut+ cultures employing batch-to-batch control and hybrid semi-parametric modeling. *Bioprocess and Biosystems Engineering*, *37*(4), 629–639. https://doi.org/10.1007/s00449-013-1029-9

Food and Drug Administration. (2004). *Guidance for Industry, PAT-A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance* (Issue September). http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070305.pdf

Food and Drug Administration. (2022). *Computer Software Assurance for Production and Quality System - Draft Guidance for Industry and FDA staff.*

Frederick, D. W., McDougal, A. V., Semenas, M., Vappiani, J., Nuzzo, A., Ulrich, J. C., Becherer, J. D., Preugschat, F., Stewart, E. L., Sévin, D. C., & Kramer, H. F. (2020). Complementary NAD+ replacement strategies fail to functionally protect dystrophin-deficient muscle. *Skeletal Muscle*, *10*(1), 30. https://doi.org/10.1186/s13395-020-00249-y

Fuhrer, T., Heer, D., Begemann, B., & Zamboni, N. (2011). High-Throughput, Accurate Mass Metabolome Profiling of Cellular Extracts by Flow Injection–Time-of-Flight Mass Spectrometry. *Analytical Chemistry*, *83*(18), 7074–7080. https://doi.org/10.1021/ac201267k

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, *20*(3–4), 121–136. https://doi.org/10.1007/BF00342633

Galvanauskas, V., & Simutis, R. (2007). Software tool for efficient hybrid model-based design of biochemical processes. *Wseas Transactions On Biology And Biomedicine*, *4*(9), 136–144.

Gangadharan, N., Sewell, D., Turner, R., Field, R., Cheeks, M., Oliver, S. G., Slater, N. K. H., & Dikicioglu, D. (2021). Data intelligence for process performance prediction in biologics manufacturing. *Computers & Chemical Engineering*, *146*, 107226. https://doi.org/10.1016/j.compchemeng.2021.107226

Garcia Muñoz, S., Kourti, T., & MacGregor, J. F. (2004). Multivariate Forecasting of Batch Evolution for Monitoring and Fault Detection. *IFAC Proceedings Volumes*, *37*(9), 71–76. https://doi.org/10.1016/S1474-6670(17)31796-2

García Muñoz, S., MacGregor, J. F., & Kourti, T. (2005). Product transfer between sites using Joint-Y PLS. *Chemometrics and Intelligent Laboratory Systems*, *79*(1–2), 101–114. https://doi.org/10.1016/j.chemolab.2005.04.009

Gaughan, C. L. (2016). The present state of the art in expression, production and characterization of monoclonal antibodies. *Molecular Diversity*, *20*(1), 255–270. https://doi.org/10.1007/s11030-015-9625-z

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

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

Gentiluomo, L., Roessner, D., Augustijn, D., Svilenov, H., Kulakova, A., Mahapatra, S., Winter, G., Streicher, W., Rinnan, Å., Peters, G. H. J., Harris, P., & Frieß, W. (2019). Application of interpretable artificial neural networks to early monoclonal antibodies development. *European Journal of Pharmaceutics and Biopharmaceutics*, *141*(May), 81–89. https://doi.org/10.1016/j.ejpb.2019.05.017

Georgakis, C. (2013). Design of Dynamic Experiments: A Data-Driven Methodology for the Optimization of Time-Varying Processes. *Industrial & Engineering Chemistry Research*, *52*(35), 12369–12382. https://doi.org/10.1021/ie3035114

Gerstweiler, L., Bi, J., & Middelberg, A. P. J. (2021). Continuous downstream bioprocessing for intensified manufacture of biopharmaceuticals and antibodies. *Chemical Engineering Science*, *231*, 116272. https://doi.org/10.1016/j.ces.2020.116272

Ghosh, D., Hermonat, E., Mhaskar, P., Snowling, S., & Goel, R. (2019). Hybrid Modeling Approach Integrating First-Principles Models with Subspace Identification. *Industrial & Engineering Chemistry Research*, *58*(30), 13533–13543. https://doi.org/10.1021/acs.iecr.9b00900

Gilgunn, S., & Bones, J. (2018). Challenges to industrial mAb bioprocessing—removal of host cell proteins in CHO cell bioprocesses. *Current Opinion in Chemical Engineering*, *22*, 98–106. https://doi.org/10.1016/j.coche.2018.08.001

Glassey, J., Gernaey, K. V., Clemens, C., Schulz, T. W., Oliveira, R., Striedner, G., & Mandenius, C.-F. (2011). Process analytical technology (PAT) for biopharmaceuticals. *Biotechnology Journal*, *6*(4), 369–377. https://doi.org/10.1002/biot.201000356

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, *9*, 249–256.

Goel, G., Conway, K. L., Jaeger, M., Netea, M. G., & Xavier, R. J. (2014). Multivariate inference of pathway activity in host immunity and response to therapeutics. *Nucleic Acids Research*, *42*(16), 10288–10306. https://doi.org/10.1093/nar/gku722

Goeman, J. J., Van de Geer, S., De Kort, F., & van Houwellingen, H. C. (2004). A global test for groups fo genes: Testing association with a clinical outcome. *Bioinformatics*, *20*(1), 93–99. https://doi.org/10.1093/bioinformatics/btg382

Goldrick, S., Duran-Villalobos, C. A., Jankauskas, K., Lovett, D., Farid, S. S., & Lennox, B. (2019). Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process. *Computers & Chemical Engineering*, *130*, 106471. https://doi.org/10.1016/j.compchemeng.2019.05.037

Goldrick, S., Holmes, W., Bond, N. J., Lewis, G., Kuiper, M., Turner, R., & Farid, S. S. (2017). Advanced multivariate data analysis to determine the root cause of trisulfide bond formation in a novel antibody-peptide fusion. *Biotechnology and Bioengineering*, *114*(10), 2222–2234. https://doi.org/10.1002/bit.26339

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Gopalakrishnan, S., Dash, S., & Maranas, C. (2020). K-FIT: An accelerated kinetic parameterization algorithm using steady-state fluxomic data. *Metabolic Engineering*, *61*(January), 197–205. https://doi.org/10.1016/j.ymben.2020.03.001

Green, A., & Glassey, J. (2015). Multivariate analysis of the effect of operating conditions on hybridoma cell metabolism and glycosylation of produced antibody. *Journal of Chemical Technology & Biotechnology*, *90*(2), 303–313. https://doi.org/10.1002/jctb.4481

Gregersen, L., & Jørgensen, S. B. (1999). Supervision of fed-batch fermentations. *Chemical Engineering Journal*, *75*(1), 69–76. https://doi.org/10.1016/S1385-8947(99)00018-2

Grissa, D., Pétéra, M., Brandolini, M., Napoli, A., Comte, B., & Pujos-Guillot, E. (2016). Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. *Frontiers in Molecular Biosciences*, *3*(JUL), 1–15. https://doi.org/10.3389/fmolb.2016.00030

Gronemeyer, P., Ditz, R., & Strube, J. (2014). Trends in Upstream and Downstream Process Development for Antibody Manufacturing. *Bioengineering*, *1*(4), 188–212. https://doi.org/10.3390/bioengineering1040188

Guerra, A., von Stosch, M., & Glassey, J. (2019). Toward biotherapeutic product real-time quality monitoring. *Critical Reviews in Biotechnology*, *39*(3), 289–305. https://doi.org/10.1080/07388551.2018.1524362

Guo, W., Xu, Y., & Feng, X. (2017). *DeepMetabolism: A Deep Learning System to Predict Phenotype from Genome Sequencing*. 1–7. http://arxiv.org/abs/1705.03094

Gutierrez, J. M., Feizi, A., Li, S., Kallehauge, T. B., Hefzi, H., Grav, L. M., Ley, D., Baycin Hizal, D., Betenbaugh, M. J., Voldborg, B., Faustrup Kildegaard, H., Min Lee, G., Palsson, B. O., Nielsen, J., & Lewis, N. E. (2020). Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion. *Nature Communications*, *11*(1), 68. https://doi.org/10.1038/s41467-019-13867-y

Hagrot, E., Oddsdóttir, H. Æ., Hosta, J. G., Jacobsen, E. W., & Chotteau, V. (2017). 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*, *259*(November 2016), 235–247. https://doi.org/10.1016/j.jbiotec.2017.05.026

Hart, T., Komori, H., LaMere, S., Podshivalova, K., & Salomon, D. R. (2013). Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics*, *14*(1), 778. https://doi.org/10.1186/1471-2164-14-778

Hart, W. E., Watson, J.-P., & Woodruff, D. L. (2011). Pyomo: modeling and solving mathematical programs in Python. *Mathematical Programming Computation*, *3*(3), 219–260. https://doi.org/10.1007/s12532-011-0026-8

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classificatio. *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034.

Hefzi, H., Ang, K. S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C. A., Baycin-Hizal, D., Huang, Y., Ley, D., Martinez, V. S., Kyriakopoulos, S., Jiménez, N. E., Zielinski, D. C., Quek, L.-E., Wulff, T., Arnsdorf, J., Li, S., Lee, J. S., … Lewis, N. E. (2016). A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. *Cell Systems*, *3*(5), 434-443.e8. https://doi.org/10.1016/j.cels.2016.10.020

Hendry, J. I., Bandyopadhyay, A., Srinivasan, S., Pakrasi, H. B., & Maranas, C. D. (2020). Metabolic model guided strain design of cyanobacteria. *Current Opinion in Biotechnology*, *64*, 17–23. https://doi.org/10.1016/j.copbio.2019.08.011

Hong, M. S., Severson, K. A., Jiang, M., Lu, A. E., Love, J. C., & Braatz, R. D. (2018). Challenges and opportunities in biopharmaceutical manufacturing control. *Computers & Chemical Engineering*, *110*, 106–114. https://doi.org/10.1016/j.compchemeng.2017.12.007

Huang, Z., Xu, J., Yongky, A., Morris, C. S., Polanco, A. L., Reily, M., Borys, M. C., Li, Z. J., & Yoon, S. (2020). CHO cell productivity improvement by genome-scale modeling and pathway analysis: Application to feed supplements. *Biochemical Engineering Journal*, *160*(May), 107638. https://doi.org/10.1016/j.bej.2020.107638

Huang, Z., & Yoon, S. (2020a). Integration of Time-Series Transcriptomic Data with Genome-Scale CHO Metabolic Models for mAb Engineering. *Processes*, *8*(3), 331. https://doi.org/10.3390/pr8030331

Huang, Z., & Yoon, S. (2020b). Identifying metabolic features and engineering targets for productivity improvement in CHO cells by integrated transcriptomics and genome-scale metabolic model. *Biochemical Engineering Journal*, *159*(December 2019), 107624. https://doi.org/10.1016/j.bej.2020.107624

Hyduke, D. R., Lewis, N. E., & Palsson, B. Ø. (2013). Analysis of omics data with genome-scale models of metabolism. *Molecular BioSystems*, *9*(2), 167–174. https://doi.org/10.1039/C2MB25453K

ICH Harmonised Tripartite Guideline, Guidance for Industry, Q8 Pharmaceutical Development, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (2009).

Q13 - Continuous Manufacturing of Drug Substances and Drug Products, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (2021).

IFPMA. (2022). *Facts and Figures 2021: The Pharmaceutical Industry and Global Health*. https://www.ifpma.org/resource-centre/facts-and-figures-2022-the-pharmaceutical-industry-and-global-health/

Jackson, J. E. (1991). *A user's guide to principal components*. John Wiley & Sons, Inc.

Jaeckle, C. M., & MacGregor, J. F. (2000). Product transfer between plants using historical process data. *AIChE Journal*, *46*(10), 1989–1997. https://doi.org/10.1002/aic.690461011

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

Jimenez del Val, I., Polizzi, K. M., & Kontoravdi, C. (2016). A theoretical estimate for nucleotide sugar demand towards Chinese Hamster Ovary cellular glycosylation. *Scientific Reports*, *6*(1), 28547. https://doi.org/10.1038/srep28547

Jolliffe, I. T. (2002). *Principal Component Analysis* (Second Edi). Springer.

Jr, C. E. D. (2014). Optimal Algorithm for Metabolomics Classification and Feature Selection varies by Dataset. *International Journal of Biology*, *7*(1), 100–115. https://doi.org/10.5539/ijb.v7n1p100

Kang, S. H., & Lee, C. H. (2021). Development of Therapeutic Antibodies and Modulating the Characteristics of Therapeutic Antibodies to Maximize the Therapeutic Efficacy. *Biotechnology and Bioprocess Engineering*, *26*(3), 295–311. https://doi.org/10.1007/s12257-020-0181-8

Karst, D. J., Steinhoff, R. F., Kopp, M. R. G., Serra, E., Soos, M., Zenobi, R., & Morbidelli, M. (2017). Intracellular CHO Cell Metabolite Profiling Reveals Steady-State Dependent Metabolic Fingerprints in Perfusion Culture. *Biotechnology Progress*, *33*(4), 879–890. https://doi.org/10.1002/btpr.2421

Kesik-Brodacka, M. (2018). Progress in biopharmaceutical development. *Biotechnology and Applied Biochemistry*, *65*(3), 306–322. https://doi.org/10.1002/bab.1617

Khaleghi, M. K., Savizi, I. S. P., Lewis, N. E., & Shojaosadati, S. A. (2021). Synergisms of machine learning and constraint-based modeling of metabolism for analysis and optimization of fermentation parameters. *Biotechnology Journal*, *16*(11). https://doi.org/10.1002/biot.202100212

Khosravi, A., Nahavandi, S., Creighton, D., & Atiya, A. F. (2011). Comprehensive Review of Neural Network-Based Prediction Intervals and New Advances. *IEEE Transactions on Neural Networks*, *22*(9), 1341–1356. https://doi.org/10.1109/TNN.2011.2162110

Kim, D. Y., Lee, J. C., Chang, H. N., & Oh, D. J. (2005). Effects of supplementation of various medium components on Chinese hamster ovary cell cultures producing recombinant antibody. *Cytotechnology*, *47*(1–3), 37–49. https://doi.org/10.1007/s10616-005-3775-2

Kim, S. H., & Lee, G. M. (2009). Development of serum-free medium supplemented with hydrolysates for the production of therapeutic antibodies in CHO cell cultures using design of experiments. *Applied Microbiology and Biotechnology*, *83*(4), 639–648. https://doi.org/10.1007/s00253-009-1903-1

Kingma, D. P., & Ba, J. L. (2015). ADAM: A method for stochastic optimization. *ArXiv*.

Kjeldahl, K., & Bro, R. (2010). Some common misunderstandings in chemometrics. *Journal of Chemometrics*, *24*(7–8), 558–564. https://doi.org/10.1002/cem.1346

Klebanov, N., & Georgakis, C. (2016). Dynamic Response Surface Models: A Data-Driven Approach for the Analysis of Time-Varying Process Outputs. *Industrial & Engineering Chemistry Research*, *55*(14), 4022–4034. https://doi.org/10.1021/acs.iecr.5b03572

Kochanowski, N., Blanchard, F., Cacan, R., Chirat, F., Guedon, E., Marc, A., & Goergen, J. L. (2006). Intracellular nucleotide and nucleotide sugar contents of cultured CHO cells determined by a fast, sensitive, and high-resolution ion-pair RP-HPLC. *Analytical Biochemistry*, *348*(2), 243–251. https://doi.org/10.1016/j.ab.2005.10.027

Kol, S., Ley, D., Wulff, T., Decker, M., Arnsdorf, J., Schoffelen, S., Hansen, A. H., Jensen, T. L., Gutierrez, J. M., Chiang, A. W. T., Masson, H. O., Palsson, B. O., Voldborg, B. G., Pedersen, L. E., Kildegaard, H. F., Lee, G. M., & Lewis, N. E. (2020). Multiplex secretome engineering enhances recombinant protein production and purity. *Nature Communications*, *11*(1), 1908. https://doi.org/10.1038/s41467-020-15866-w

Kontoravdi, C., Asprey, S. P., Pistikopoulos, E. N., & Mantalaris, A. (2007). Development of a dynamic model of monoclonal antibody production and glycosylation for product quality monitoring. *Computers & Chemical Engineering*, *31*(5–6), 392–400. https://doi.org/10.1016/j.compchemeng.2006.04.009

Kontoravdi, C., Asprey, S. P., Pistikopoulos, S., & Mantalaris, A. (2005). Dynamic model of MAb production and glycosylation for the purpose of product quality control. In *Computer Aided Chemical Engineering* (Vol. 20, Issue C, pp. 307–312). https://doi.org/10.1016/S1570-7946(05)80173-7

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

Kotidis, P., & Kontoravdi, C. (2020). Harnessing the potential of artificial neural networks for predicting protein glycosylation. *Metabolic Engineering Communications*, *10*(February), e00131. https://doi.org/10.1016/j.mec.2020.e00131

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Krzanowski, W. J. (1979). Between-Groups Comparison of Principal Components. *Journal of the American Statistical Association*, *74*(367), 703–707. https://doi.org/10.1080/01621459.1979.10481674

Kvalheim, O. M., & Karstang, T. V. (1989). Interpretation of latent-variable regression models. *Chemometrics and Intelligent Laboratory Systems*, *7*(1–2), 39–51. https://doi.org/10.1016/0169-7439(89)80110-8

Lai, T., Yang, Y., & Ng, S. (2013). Advances in Mammalian Cell Line Development Technologies for Recombinant Protein Production. *Pharmaceuticals*, *6*(5), 579–603. https://doi.org/10.3390/ph6050579

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Occupational chronic exposure to organic solvents. *I31st Conference on Neural Information Processing Systems*.

Le, K., Tan, C., Gupta, S., Guhan, T., Barkhordarian, H., Lull, J., Stevens, J., & Munro, T. (2018). A novel mammalian cell line development platform utilizing nanofluidics and optoelectro positioning technology. *Biotechnology Progress*, *34*(6), 1438–1446. https://doi.org/10.1002/btpr.2690

Lee, S. S. (2000). Noisy replication in skewed binary classification. *Computational Statistics & Data Analysis*, *34*(2), 165–191. https://doi.org/10.1016/S0167-9473(99)00095-X

Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., Weitz, K. K., Eils, R., König, R., Smith, R. D., & Palsson, B. Ø. (2010). Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, *6*(1), 390. https://doi.org/10.1038/msb.2010.47

Li, F., Vijayasankaran, N., Shen, A. (Yijuan), Kiss, R., & Amanullah, A. (2010). Cell culture processes for monoclonal antibody production. *MAbs*, *2*(5), 466–479. https://doi.org/10.4161/mabs.2.5.12720

Li, M.-Y., Ebel, B., Paris, C., Chauchard, F., Guedon, E., & Marc, A. (2018). Real-time monitoring of antibody glycosylation site occupancy by in situ Raman spectroscopy during bioreactor CHO cell cultures. *Biotechnology Progress*, *34*(2), 486–493. https://doi.org/10.1002/btpr.2604

Ling, W. L. W., Bai, Y., Cheng, C., Padawer, I., & Wu, C. (2015). Development and manufacturability assessment of chemically-defined medium for the production of protein therapeutics in CHO cells. *Biotechnology Progress*, *31*(5), 1163–1171. https://doi.org/10.1002/btpr.2108

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLOS Computational Biology*, *13*(5), e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Lularevic, M., Racher, A. J., Jaques, C., & Kiparissides, A. (2019). Improving the accuracy of flux balance analysis through the implementation of carbon availability constraints for intracellular reactions. *Biotechnology and Bioengineering*, *116*(9), 2339–2352. https://doi.org/10.1002/bit.27025

Maharana, K., Mondal, S., & Nemade, B. (2022). A Review: Data Pre-Processing and Data Augmentation Techniques. *Global Transitions Proceedings*, 0–13. https://doi.org/10.1016/j.gltp.2022.04.020

Manser, B., & Glenz, M. (2022). Regulatory and Quality Considerations of Continuous Bioprocessing. In G. Subramanian (Ed.), *Process Control, Intensification, and Digitalisation in Continuous Biomanufacturing* (pp. 351–375). Wiley. https://doi.org/10.1002/9783527827343.ch10

Maranas, C. D., & Zomorrodi, A. R. (2016). *Optimization methods in metabolic netoworks*. Wiley.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. Academic Press Limited.

Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D. S., Krebs, C. F., & Bonn, S. (2020). Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature Communications*, *11*(1), 166. https://doi.org/10.1038/s41467-019-14018-z

Martínez, V. S., Buchsteiner, M., Gray, P., Nielsen, L. K., & Quek, L.-E. (2015). Dynamic metabolic flux analysis using B-splines to study the effects of temperature shift on CHO cell metabolism. *Metabolic Engineering Communications*, *2*, 46–57. https://doi.org/10.1016/j.meteno.2015.06.001

Martínez, V. S., Dietmair, S., Quek, L.-E., Hodson, M. P., Gray, P., & Nielsen, L. K. (2013). Flux balance analysis of CHO cells before and after a metabolic switch from lactate production to consumption. *Biotechnology and Bioengineering*, *110*(2), 660–666. https://doi.org/10.1002/bit.24728

Maruthamuthu, M. K., Rudge, S. R., Ardekani, A. M., Ladisch, M. R., & Verma, M. S. (2020). Process Analytical Technologies and Data Analytics for the Manufacture of Monoclonal Antibodies. *Trends in Biotechnology*, *38*(10), 1169–1186. https://doi.org/10.1016/j.tibtech.2020.07.004

McAtee Pereira, A. G., Walther, J. L., Hollenbach, M., & Young, J. D. (2018). 13 C Flux Analysis Reveals that Rebalancing Medium Amino Acid Composition can Reduce Ammonia Production while Preserving Central Carbon Metabolism of CHO Cell Cultures. *Biotechnology Journal*, *13*(10), 1700518. https://doi.org/10.1002/biot.201700518

Megchelenbrink, W., Huynen, M., & Marchiori, E. (2014). optGpSampler: An improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0086587

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010

Mercier, S. M., Diepenbroek, B., Wijffels, R. H., & Streefland, M. (2014). Multivariate PAT solutions for biopharmaceutical cultivation: current progress and limitations. *Trends in Biotechnology*, *32*(6), 329–336. https://doi.org/10.1016/j.tibtech.2014.03.008

Mohmad-Saberi, S. E., Hashim, Y. Z. H. Y., Mel, M., Amid, A., Ahmad-Raus, R., & Packeer-Mohamed, V. (2013). Metabolomics profiling of extracellular metabolites in CHO-K1 cells cultured in different types of growth media. *Cytotechnology*, *65*(4), 577–586. https://doi.org/10.1007/s10616-012-9508-4

Montgomery, D. C. (2007). *Design and Analysis of Experiments* (Fifth Ed.). John Wiley & Sons, Inc.

Mora, A., Nabiswa, B., Duan, Y., Zhang, S., Carson, G., & Yoon, S. (2019). Early integration of Design of Experiment (DOE) and multivariate statistics identifies feeding regimens suitable for CHO cell line development and screening. *Cytotechnology*, *71*(6), 1137–1153. https://doi.org/10.1007/s10616-019-00350-1

MordorIntelligence. (2021). *BIOPHARMACEUTICALS MARKET - GROWTH, TRENDS, COVID-19 IMPACT, AND FORECASTS (2022 - 2027).* https://www.mordorintelligence.com/industry-reports/global-biopharmaceuticals-market-industry

Morris, C., Polanco, A., Yongky, A., Xu, J., Huang, Z., Zhao, J., McFarland, K. S., Park, S., Warrack, B., Reily, M., Borys, M. C., Li, Z., & Yoon, S. (2020). Bigdata analytics identifies metabolic inhibitors and promoters for productivity improvement and optimization of monoclonal antibody (mAb) production process. *Bioresources and Bioprocessing*, *7*(1), 31. https://doi.org/10.1186/s40643-020-00318-6

Mould, D. R., & Meibohm, B. (2016). Drug Development of Therapeutic Monoclonal Antibodies. *BioDrugs*, *30*(4), 275–293. https://doi.org/10.1007/s40259-016-0181-6

Narayanan, H., Behle, L., Luna, M. F., Sokolov, M., Guillén-Gosálbez, G., Morbidelli, M., & Butté, A. (2020). Hybrid-EKF: Hybrid model coupled with extended Kalman filter for real-time monitoring and control of mammalian cell culture. *Biotechnology and Bioengineering*, *117*(9), 2703–2714. https://doi.org/10.1002/bit.27437

Narayanan, H., Luna, M., Sokolov, M., Arosio, P., Butté, A., & Morbidelli, M. (2021). Hybrid Models Based on Machine Learning and an Increasing Degree of Process Knowledge: Application to Capture Chromatographic Step. *Industrial & Engineering Chemistry Research*, *60*(29), 10466–10478. https://doi.org/10.1021/acs.iecr.1c01317

Narayanan, H., Seidler, T., Luna, M. F., Sokolov, M., Morbidelli, M., & Butté, A. (2021). Hybrid Models for the simulation and prediction of chromatographic processes for protein capture. *Journal of Chromatography A*, *1650*, 462248. https://doi.org/10.1016/j.chroma.2021.462248

Narayanan, H., Sokolov, M., Morbidelli, M., & Butté, A. (2019). A new generation of predictive models: The added value of hybrid models for manufacturing processes of therapeutic proteins. *Biotechnology and Bioengineering*, *116*(10), 2540–2549. https://doi.org/10.1002/bit.27097

Nicolae, A., Wahrheit, J., Bahnemann, J., Zeng, A.-P., & Heinzle, E. (2014). Non-stationary 13C metabolic flux analysis of Chinese hamster ovary cells in batch culture using extracellular labeling highlights metabolic reversibility and compartmentation. *BMC Systems Biology*, *8*(1), 50. https://doi.org/10.1186/1752-0509-8-50

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

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

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

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

O'Brien, C. M., Zhang, Q., Daoutidis, P., & Hu, W.-S. (2021). A hybrid mechanistic-empirical model for in silico mammalian cell bioprocess simulation. *Metabolic Engineering*, *66*(April), 31–40. https://doi.org/10.1016/j.ymben.2021.03.016

O'Flaherty, R., Bergin, A., Flampouri, E., Mota, L. M., Obaidi, I., Quigley, A., Xie, Y., & Butler, M. (2020). Mammalian cell culture for production of recombinant proteins: A review of the critical steps in their biomanufacturing. *Biotechnology Advances*, *43*, 107552. https://doi.org/10.1016/j.biotechadv.2020.107552

Oliveira, R. (2004). Combining first principles modelling and artificial neural networks: a general framework. *Computers & Chemical Engineering*, *28*(5), 755–766. https://doi.org/10.1016/j.compchemeng.2004.02.014

Orellana, C. A., Marcellin, E., Schulz, B. L., Nouwens, A. S., Gray, P. P., & Nielsen, L. K. (2015). High-Antibody-Producing Chinese Hamster Ovary Cells Up-Regulate Intracellular Protein Transport and Glutathione Synthesis. *Journal of Proteome Research*, *14*(2), 609–618. https://doi.org/10.1021/pr501027c

Ozturk, S. S., Riley, M. R., & Palsson, B. O. (1992). Effects of ammonia and lactate on hybridoma growth, metabolism, and antibody production. *Biotechnology and Bioengineering*, *39*(4), 418–431. https://doi.org/10.1002/bit.260390408

Paul, A., & de Boves Harrington, P. (2021). Chemometric applications in metabolomic studies using chromatography-mass spectrometry. *Trends in Analytical Chemistry*, *135*, 116165. https://doi.org/10.1016/j.trac.2020.116165

Pereira, S., Kildegaard, H. F., & Andersen, M. R. (2018). Impact of CHO Metabolism on Cell Growth and Protein Production: An Overview of Toxic and Inhibiting Metabolites and Nutrients. *Biotechnology Journal*, *13*(3), 1700499. https://doi.org/10.1002/biot.201700499

Perrin, J., Werner, T., Kurzawa, N., Rutkowska, A., Childs, D. D., Kalxdorf, M., Poeckel, D., Stonehouse, E., Strohmer, K., Heller, B., Thomson, D. W., Krause, J., Becher, I., Eberl, H. C., Vappiani, J., Sevin, D. C., Rau, C. E., Franken, H., Huber, W., … Bergamini, G. (2020). Identifying drug targets in tissues and whole blood with thermal-shift profiling. *Nature Biotechnology*, *38*(3), 303–308. https://doi.org/10.1038/s41587-019-0388-4

Pharkya, P., & Maranas, C. D. (2006). An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metabolic Engineering*, *8*(1), 1–13. https://doi.org/10.1016/j.ymben.2005.08.003

Pinto, J., de Azevedo, C. R., Oliveira, R., & von Stosch, M. (2019). A bootstrap-aggregated hybrid semi-parametric modeling framework for bioprocess development. *Bioprocess and Biosystems Engineering*, *42*(11), 1853–1865. https://doi.org/10.1007/s00449-019-02181-y

Pinto, R. C. (2017). Chemometrics Methods and Strategies in Metabolomics. In A. Sussulini (Ed.), *Metabolomics: From Fundamentals to Clinical Applications* (First, Vol. 965, pp. 163–190). Springer International Publishing. https://doi.org/10.1007/978-3-319-47656-8

Popp, O., Müller, D., Didzus, K., Paul, W., Lipsmeier, F., Kirchner, F., Niklas, J., Mauch, K., & Beaucamp, N. (2016). A hybrid approach identifies metabolic signatures of high-producers for chinese hamster ovary clone selection and process optimization. *Biotechnology and Bioengineering*, *113*(9), 2005–2019. https://doi.org/10.1002/bit.25958

Povey, J. F., O'Malley, C. J., Root, T., Martin, E. B., Montague, G. A., Feary, M., Trim, C., Lang, D. A., Alldread, R., Racher, A. J., & Smales, C. M. (2014). Rapid high-throughput characterisation, classification and selection of recombinant mammalian cell line phenotypes using intact cell MALDI-ToF mass spectrometry fingerprinting and PLS-DA modelling. *Journal of Biotechnology*, *184*, 84–93. https://doi.org/10.1016/j.jbiotec.2014.04.028

Psichogios, D. C., & Ungar, L. H. (1992). A Hybrid Neural Network-First Principles Approach to Process Modeling. *AIChE Journal*, *38*(10), 1499–1511.

Pujos-Guillot, E., Hubert, J., Martin, J.-F., Lyan, B., Quintana, M., Claude, S., Chabanas, B., Rothwell, J. A., Bennetau-Pelissero, C., Scalbert, A., Comte, B., Hercberg, S., Morand, C., Galan, P., & Manach, C. (2013). Mass Spectrometry-based Metabolomics for the Discovery of Biomarkers of Fruit and Vegetable Intake: Citrus Fruit as a Case Study. *Journal of Proteome Research*, *12*(4), 1645–1659. https://doi.org/10.1021/pr300997c

Qin, S. J. (2014). Process data analytics in the era of big data. *AIChE Journal*, *60*(9), 3092–3100. https://doi.org/10.1002/aic.14523

Quek, L.-E., Dietmair, S., Krömer, J. O., & Nielsen, L. K. (2010). Metabolic flux analysis in mammalian cell culture. *Metabolic Engineering*, *12*(2), 161–171. https://doi.org/10.1016/j.ymben.2009.09.002

Quek, L.-E., Wittmann, C., Nielsen, L. K., & Krömer, J. O. (2009). OpenFLUX: efficient modelling software for 13C-based metabolic flux analysis. *Microbial Cell Factories*, *8*(1), 25. https://doi.org/10.1186/1475-2859-8-25

Rader, R. A. (2008). (Re)defining biopharmaceutical. *Nature Biotechnology*, *26*(7), 743–751.

Ramaker, H.-J., van Sprang, E. N. M., Westerhuis, J. A., & Smilde, A. K. (2005). Fault detection properties of global, local and time evolving models for batch process monitoring. *Journal of Process Control*, *15*(7), 799–805. https://doi.org/10.1016/j.jprocont.2005.02.001

Rameez, S., Mostafa, S. S., Miller, C., & Shukla, A. A. (2014). High-throughput miniaturized bioreactors for cell culture process development: Reproducibility, scalability, and control. *Biotechnology Progress*, *30*(3), 718–727. https://doi.org/10.1002/btpr.1874

Ranganathan, S., Suthers, P. F., & Maranas, C. D. (2010). OptForce: An Optimization Procedure for Identifying All Genetic Manipulations Leading to Targeted Overproductions. *PLoS Computational Biology*, *6*(4), e1000744. https://doi.org/10.1371/journal.pcbi.1000744

Rathore, A. S. (2014). QbD/PAT for bioprocessing: moving from theory to implementation. *Current Opinion in Chemical Engineering*, *6*, 1–8. https://doi.org/10.1016/j.coche.2014.05.006

Rathore, A. S., Nikita, S., Thakur, G., & Mishra, S. (2022). Artificial intelligence and machine learning applications in biopharmaceutical manufacturing. *Trends in Biotechnology*. https://doi.org/10.1016/j.tibtech.2022.08.007

Rathore, A. S., & Winkle, H. (2009). Quality by design for biopharmaceuticals. *Nature Biotechnology*, *27*(1), 26–34.

Rato, T. J., Delgado, P., Martins, C., & Reis, M. S. (2020). First Principles Statistical Process Monitoring of High-Dimensional Industrial Microelectronics Assembly Processes. *Processes*, *8*(11), 1520. https://doi.org/10.3390/pr8111520

Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, *49*(March), 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

Reis, M., & Gins, G. (2017). Industrial Process Monitoring in the Big Data/Industry 4.0 Era: from Detection, to Diagnosis, to Prognosis. *Processes*, *5*(4), 35. https://doi.org/10.3390/pr5030035

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, *58*(4), 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Richardson, J., Shah, B., Bondarenko, P. V, Bhebe, P., Zhang, Z., Nicklaus, M., & Kombe, M. C. (2015). Metabolomics analysis of soy hydrolysates for the identification of productivity markers of mammalian cells for manufacturing therapeutic proteins. *Biotechnology Progress*, *31*(2), 522–531. https://doi.org/10.1002/btpr.2050

Richelle, A., David, B., Demaegd, D., Dewerchin, M., Kinet, R., Morreale, A., Portela, R., Zune, Q., & von Stosch, M. (2020). Towards a widespread adoption of metabolic modeling tools in biopharmaceutical industry: a process systems biology engineering perspective. *Npj Systems Biology and Applications*, *6*(1), 6. https://doi.org/10.1038/s41540-020-0127-y

Ritacco, F. V., Wu, Y., & Khetan, A. (2018). 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. https://doi.org/10.1002/btpr.2706

Robitaille, J., Chen, J., & Jolicoeur, M. (2015). A Single Dynamic Metabolic Model Can Describe mAb Producing CHO Cell Batch and Fed-Batch Cultures on Different Culture Media. *PLOS ONE*, *10*(9), e0136815. https://doi.org/10.1371/journal.pone.0136815

Rodrigues, M. E., Costa, A. R., Henriques, M., Azeredo, J., & Oliveira, R. (2009a). Technological progresses in monoclonal antibody production systems. *Biotechnology Progress*, *26*(2), NA-NA. https://doi.org/10.1002/btpr.348

Rodrigues, M. E., Costa, A. R., Henriques, M., Azeredo, J., & Oliveira, R. (2009b). Technological progresses in monoclonal antibody production systems. *Biotechnology Progress*, *26*(2), 322–351. https://doi.org/10.1002/btpr.348

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rubingh, C. M., Bijlsma, S., Jellema, R. H., Overkamp, K. M., van der Werf, M. J., & Smilde, A. K. (2009). Analyzing Longitudinal Microbial Metabolomics Data. *Journal of Proteome Research*, *8*(9), 4319–4327. https://doi.org/10.1021/pr900126e

S. De Jong. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*, 251–263.

Saitua, F., Torres, P., Pérez-Correa, J. R., & Agosin, E. (2017). Dynamic genome-scale metabolic modeling of the yeast Pichia pastoris. *BMC Systems Biology*, *11*(1), 27. https://doi.org/10.1186/s12918-017-0408-2

Sansana, J., Joswiak, M. N., Castillo, I., Wang, Z., Rendall, R., Chiang, L. H., & Reis, M. S. (2021). Recent trends on hybrid modeling for Industry 4.0. *Computers & Chemical Engineering*, *151*, 107365. https://doi.org/10.1016/j.compchemeng.2021.107365

Schinn, S.-M., Morrison, C., Wei, W., Zhang, L., & Lewis, N. E. (2021). Systematic evaluation of parameters for genome-scale metabolic models of cultured mammalian cells. *Metabolic Engineering*, *66*(November 2020), 21–30. https://doi.org/10.1016/j.ymben.2021.03.013

Schinn, S., Morrison, C., Wei, W., Zhang, L., & Lewis, N. E. (2021). A genome-scale metabolic network model and machine learning predict amino acid concentrations in Chinese Hamster Ovary cell cultures. *Biotechnology and Bioengineering*, *118*(5), 2118–2123. https://doi.org/10.1002/bit.27714

Schofield, M. (2018). Current state of the art in continuous bioprocessing. *Biotechnology Letters*, *40*(9–10), 1303–1309. https://doi.org/10.1007/s10529-018-2593-5

Schubert, J., Simutis, R., Dors, M., Havlik, I., & Lübbert, A. (1994). Bioprocess optimization and control: Application of hybrid modelling. *Journal of Biotechnology*, *35*(1), 51–68. https://doi.org/10.1016/0168-1656(94)90189-9

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sellick, C. A., Croxford, A. S., Maqsood, A. R., Stephens, G., Westerhoff, H. V., Goodacre, R., & Dickson, A. J. (2011). Metabolite profiling of recombinant CHO cells: Designing tailored feeding regimes that enhance recombinant antibody production. *Biotechnology and Bioengineering*, *108*(12), 3025–3031. https://doi.org/10.1002/bit.23269

Selvarasu, S., Ho, Y. S., Chong, W. P. K., Wong, N. S. C., Yusufi, F. N. K., Lee, Y. Y., Yap, M. G. S., & Lee, D.-Y. (2012). Combined in silico modeling and metabolomics analysis to characterize fed-batch CHO cell culture. *Biotechnology and Bioengineering*, *109*(6), 1415–1429. https://doi.org/10.1002/bit.24445

Setoodeh, P., Jahanmiri, A., & Eslamloueyan, R. (2012). Hybrid neural modeling framework for simulation and optimization of diauxie-involved fed-batch fermentative succinate production. *Chemical Engineering Science*, *81*, 57–76. https://doi.org/10.1016/j.ces.2012.06.031

Sha, S., Agarabi, C., Brorson, K., Lee, D. Y., & Yoon, S. (2016). N-Glycosylation Design and Control of Therapeutic Monoclonal Antibodies. *Trends in Biotechnology*, *34*(10), 835–846. https://doi.org/10.1016/j.tibtech.2016.02.013

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Sheikholeslami, Z., Jolicoeur, M., & Henry, O. (2013). Probing the metabolism of an inducible mammalian expression system using extracellular isotopomer analysis. *Journal of Biotechnology*, *164*(4), 469–478. https://doi.org/10.1016/j.jbiotec.2013.01.025

Sheikholeslami, Z., Jolicoeur, M., & Henry, O. (2014). Elucidating the effects of postinduction glutamine feeding on the growth and productivity of CHO cells. *Biotechnology Progress*, *30*(3), 535–546. https://doi.org/10.1002/btpr.1907

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Shukla, A. A. ., & Thömmes, J. (2010). Recent advances in large-scale production of monoclonal antibodies and related proteins. *Trends in Biotechnology*, *28*(5), 253–261. https://doi.org/10.1016/j.tibtech.2010.02.001

Silva, F., Resende, D., Amorim, M., & Borges, M. (2020). A Field Study on the Impacts of Implementing Concepts and Elements of Industry 4.0 in the Biopharmaceutical Sector. *Journal of Open Innovation: Technology, Market, and Complexity*, *6*(4), 175. https://doi.org/10.3390/joitmc6040175

Simutis, R., & Lübbert, A. (2017). Hybrid Approach to State Estimation for Bioprocess Control. *Bioengineering*, *4*(4), 21. https://doi.org/10.3390/bioengineering4010021

Sinner, P., Daume, S., Herwig, C., & Kager, J. (2020). Usage of Digital Twins Along a Typical Process Development Cycle. In *Advances in biochemical engineering/biotechnology* (Vol. 176, Issue December 2020, pp. 71–96). https://doi.org/10.1007/10_2020_149

Sivanandam, S. N., & Deepa, S. N. (2008). *Introduction to genetic algorithms*. Springer Berlin Heidelberg.

Smiatek, J., Clemens, C., Herrera, L. M., Arnold, S., Knapp, B., Presser, B., Jung, A., Wucherpfennig, T., & Bluhmki, E. (2021). Generic and specific recurrent neural network models: Applications for large and small scale biopharmaceutical upstream processes. *Biotechnology Reports*, *31*, e00640. https://doi.org/10.1016/j.btre.2021.e00640

Smietana, K., Siatkowski, M., & Møller, M. (2016). Trends in clinical success rates. *Nature Reviews Drug Discovery*, *15*(6), 379–380. https://doi.org/10.1038/nrd.2016.85

Smilde, A. K., Westerhuis, J. A., Hoefsloot, H. C. J., Bijlsma, S., Rubingh, C. M., Vis, D. J., Jellema, R. H., Pijl, H., Roelfsema, F., & van der Greef, J. (2010). Dynamic metabolomic data analysis: a tutorial review. *Metabolomics*, *6*(1), 3–17. https://doi.org/10.1007/s11306-009-0191-1

Sokolov, M., Morbidelli, M., Butté, A., Souquet, J., & Broly, H. (2018). Sequential Multivariate Cell Culture Modeling at Multiple Scales Supports Systematic Shaping of a Monoclonal Antibody Toward a Quality Target. *Biotechnology Journal*, *13*(4), 1700461. https://doi.org/10.1002/biot.201700461

Sokolov, M., Ritscher, J., MacKinnon, N., Bielser, J.-M., Brühlmann, D., Rothenhäusler, D., Thanei, G., Soos, M., Stettler, M., Souquet, J., Broly, H., Morbidelli, M., & Butté, A.

(2017). Robust factor selection in early cell culture process development for the production of a biosimilar monoclonal antibody. *Biotechnology Progress*, *33*(1), 181–191. https://doi.org/10.1002/btpr.2374

Sokolov, M., Soos, M., Neunstoecklin, B., Morbidelli, M., Butté, A., Leardi, R., Solacroup, T., Stettler, M., & Broly, H. (2015). Fingerprint detection and process prediction by multivariate analysis of fed-batch monoclonal antibody cell culture data. *Biotechnology Progress*, *31*(6), 1633–1644. https://doi.org/10.1002/btpr.2174

Sonnenschein, N., Golib Dzib, J., Lesne, A., Eilebrecht, S., Boulkroun, S., Zennaro, M.-C., Benecke, A., & Hütt, M.-T. (2012). A network perspective on metabolic inconsistency. *BMC Systems Biology*, *6*(1), 41. https://doi.org/10.1186/1752-0509-6-41

Sridhara, V., Meyer, A. G., Rai, P., Barrick, J. E., Ravikumar, P., Segrè, D., & Wilke, C. O. (2014). Predicting Growth Conditions from Internal Metabolic Fluxes in an In-Silico Model of E. coli. *PLoS ONE*, *9*(12), e114608. https://doi.org/10.1371/journal.pone.0114608

Strasser, L., Farrell, A., Ho, J. T. C., Scheffler, K., Cook, K., Pankert, P., Mowlds, P., Viner, R., Karger, B. L., & Bones, J. (2021). Proteomic Profiling of IgG1 Producing CHO Cells Using LC/LC-SPS-MS3: The Effects of Bioprocessing Conditions on Productivity and Product Quality. *Frontiers in Bioengineering and Biotechnology*, *9*(April), 1–16. https://doi.org/10.3389/fbioe.2021.569045

Suástegui, M., Yu Ng, C., Chowdhury, A., Sun, W., Cao, M., House, E., Maranas, C. D., & Shao, Z. (2017). Multilevel engineering of the upstream module of aromatic amino acid biosynthesis in Saccharomyces cerevisiae for high production of polymer and drug precursors. *Metabolic Engineering*, *42*(May), 134–144. https://doi.org/10.1016/j.ymben.2017.06.008

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sumit, M., Dolatshahi, S., Chu, A.-H. A., Cote, K., Scarcelli, J. J., Marshall, J. K., Cornell, R. J., Weiss, R., Lauffenburger, D. A., Mulukutla, B. C., & Figueroa, B. (2019). Dissecting N-Glycosylation Dynamics in Chinese Hamster Ovary Cells Fed-batch Cultures using Time Course Omics Analyses. *IScience*, *12*, 102–120. https://doi.org/10.1016/j.isci.2019.01.006

Tan, Z., Yoon, J. M., Chowdhury, A., Burdick, K., Jarboe, L. R., Maranas, C. D., & Shanks, J. V. (2018). Engineering of E. coli inherent fatty acid biosynthesis capacity to increase octanoic acid production. *Biotechnology for Biofuels*, *11*(1), 87. https://doi.org/10.1186/s13068-018-1078-z

Teixeira, A. P., Alves, C., Alves, P. M., Carrondo, M. J. T., & Oliveira, R. (2007). Hybrid elementary flux analysis/nonparametric modeling: application for bioprocess control. *BMC Bioinformatics*, *8*(1), 30. https://doi.org/10.1186/1471-2105-8-30

Teixeira, A. P., Carinhas, N., Dias, J. M. L., Cruz, P., Alves, P. M., Carrondo, M. J. T., & Oliveira, R. (2007). Hybrid semi-parametric mathematical systems: Bridging the gap between systems biology and process engineering. *Journal of Biotechnology*, *132*(4), 418–425. https://doi.org/10.1016/j.jbiotec.2007.08.020

Teixeira, A. P., Clemente, J. J., Cunha, A. E., Carrondo, M. J. T., & Oliveira, R. (2006). Bioprocess Iterative Batch-to-Batch Optimization Based on Hybrid Parametric/Nonparametric Models. *Biotechnology Progress*, *22*(1), 247–258. https://doi.org/10.1021/bp0502328

Teixeira, A. P., Cunha, A. E., Clemente, J. J., Moreira, J. L., Cruz, H. J., Alves, P. M., Carrondo, M. J. T., & Oliveira, R. (2005). Modelling and optimization of a recombinant BHK-21 cultivation process using hybrid grey-box systems. *Journal of Biotechnology*, *118*(3), 290–303. https://doi.org/10.1016/j.jbiotec.2005.04.024

Templeton, N., Dean, J., Reddy, P., & Young, J. D. (2013). Peak antibody production is associated with increased oxidative metabolism in an industrially relevant fed-batch CHO cell culture. *Biotechnology and Bioengineering*, *110*(7), 2013–2024. https://doi.org/10.1002/bit.24858

Templeton, N., Lewis, A., Dorai, H., Qian, E. A., Campbell, M. P., Smith, K. D., Lang, S. E., Betenbaugh, M. J., & Young, J. D. (2014). The impact of anti-apoptotic gene Bcl-2∆ expression on CHO central metabolism. *Metabolic Engineering*, *25*, 92–102. https://doi.org/10.1016/j.ymben.2014.06.010

Templeton, N., Smith, K. D., McAtee-Pereira, A. G., Dorai, H., Betenbaugh, M. J., Lang, S. E., & Young, J. D. (2017). Application of 13C flux analysis to identify high-productivity CHO metabolic phenotypes. *Metabolic Engineering*, *43*(December 2016), 218–225. https://doi.org/10.1016/j.ymben.2017.01.008

Templeton, N., Xu, S., Roush, D. J., & Chen, H. (2017). 13C metabolic flux analysis identifies limitations to increasing specific productivity in fed-batch and perfusion. *Metabolic Engineering*, *44*(August), 126–133. https://doi.org/10.1016/j.ymben.2017.09.010

Thakur, G., Bansode, V., & Rathore, A. S. (2022). Continuous manufacturing of monoclonal antibodies: Automated downstream control strategy for dynamic handling of titer variations. *Journal of Chromatography A*, *1682*, 463496. https://doi.org/10.1016/j.chroma.2022.463496

Tomba, E., Barolo, M., & García-Muñoz, S. (2012). General Framework for Latent Variable Model Inversion for the Design and Manufacturing of New Products. *Industrial & Engineering Chemistry Research*, *51*(39), 12886–12900. https://doi.org/10.1021/ie301214c

Tripathi, N. K., & Shrivastava, A. (2019). Recent Developments in Bioprocessing of Recombinant Proteins: Expression Hosts and Process Development. *Frontiers in Bioengineering and Biotechnology*, *7*(December), 1–35. https://doi.org/10.3389/fbioe.2019.00420

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Trygg, J., Holmes, E., & Lundstedt, T. (2007). Chemometrics in Metabonomics. *Journal of Proteome Research*, *6*(2), 469–479. https://doi.org/10.1021/pr060594q

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, *16*(3), 119–128. https://doi.org/10.1002/cem.695

Tulsyan, A., Garvin, C., & Undey, C. (2019). Industrial batch process monitoring with limited data. *Journal of Process Control*, *77*, 114–133. https://doi.org/10.1016/j.jprocont.2019.03.002

Tulsyan, A., Garvin, C., & Ündey, C. (2018). Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems. *Biotechnology and Bioengineering*, *115*(8), 1915–1924. https://doi.org/10.1002/bit.26605

Valle, S., Li, W., & Qin, S. J. (1999). Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. *Ind. Eng. Chem. Res.*, *38*, 4389–4401.

Vande Wouwer, A., Renotte, C., & Bogaerts, P. (2004). Biological reaction modeling using radial basis function networks. *Computers & Chemical Engineering*, *28*(11), 2157–2164. https://doi.org/10.1016/j.compchemeng.2004.03.003

Vernardis, S. I., Goudar, C. T., & Klapa, M. I. (2013). Metabolic profiling reveals that time related physiological changes in mammalian cell perfusion cultures are bioreactor scale independent. *Metabolic Engineering*, *19*, 1–9. https://doi.org/10.1016/j.ymben.2013.04.005

Vodopivec, M., Lah, L., Narat, M., & Curk, T. (2019). Metabolomic profiling of CHO fed-batch growth phases at 10, 100, and 1,000 L. *Biotechnology and Bioengineering*, *116*(10), 2720–2729. https://doi.org/10.1002/bit.27087

von Stosch, M., Hamelink, J.-M., & Oliveira, R. (2016). Hybrid modeling as a QbD/PAT tool in process development: an industrial E. coli case study. *Bioprocess and Biosystems Engineering*, *39*(5), 773–784. https://doi.org/10.1007/s00449-016-1557-1

von Stosch, M., Oliveira, R., Peres, J., & Feyo de Azevedo, S. (2014). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers & Chemical Engineering*, *60*, 86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008

von Stosch, M., & Willis, M. J. (2017). Intensified design of experiments for upstream bioreactors. *Engineering in Life Sciences*, *17*(11), 1173–1184. https://doi.org/10.1002/elsc.201600037

Walsh, G. (2018). Biopharmaceutical benchmarks 2018. *Nature Biotechnology*, *36*(12), 1136–1145. https://doi.org/10.1038/nbt.4305

Wang, Z., & Georgakis, C. (2017). An in silico evaluation of data-driven optimization of biopharmaceutical processes. *AIChE Journal*, *63*(7), 2796–2805. https://doi.org/10.1002/aic.15659

Wang, Z., & Georgakis, C. (2019). A Dynamic Response Surface Model for Polymer Grade Transitions in Industrial Plants [Research-article]. *Industrial & Engineering Chemistry Research*, *58*(26), 11187–11198. https://doi.org/10.1021/acs.iecr.8b04491

Wayman, J. A., Glasscock, C., Mansell, T. J., DeLisa, M. P., & Varner, J. D. (2019). Improving designer glycan production in Escherichia coli through model-guided metabolic engineering. *Metabolic Engineering Communications*, *9*(March), e00088. https://doi.org/10.1016/j.mec.2019.e00088

Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, *12*(5), 301–321. https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S

Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E. J., Edlund, U., Shockcor, J. P., Gottfries, J., Moritz, T., & Trygg, J. (2008). Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry*, *80*(1), 115–122. https://doi.org/10.1021/ac0713510

Winter, G., & Krömer, J. O. (2013). Fluxomics - connecting ‘omics analysis and phenotypes. *Environmental Microbiology*, *15*(7), 1901–1916. https://doi.org/10.1111/1462-2920.12064

Wold, S, Johansson, E., & Cocchi, M. (1993). PLS: Partial Least Squares Projections to Latent Structures. In *3D QSAR in Drug Design: Theory, Methods and Applications* (pp. 523–550). ESCOM Science Publisher.

Wold, Svante. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, *20*(4), 397–405. https://doi.org/10.1080/00401706.1978.10489693

Wold, Svante, Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

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

Wongkittichote, P., Ah Mew, N., & Chapman, K. A. (2017). Propionyl-CoA carboxylase – A review. *Molecular Genetics and Metabolism*, *122*(4), 145–152. https://doi.org/10.1016/j.ymgme.2017.10.002

Worley, B., & Powers, R. (2013). Multivariate Analysis in Metabolomics. *Current Metabolomics*, *1*(1), 92–107. https://doi.org/10.2174/2213235X130108

Wu, S. G., Wang, Y., Jiang, W., Oyetunde, T., Yao, R., Zhang, X., Shimizu, K., Tang, Y. J., & Bao, F. S. (2016). Rapid Prediction of Bacterial Heterotrophic Fluxomics Using Machine Learning and Constraint Programming. *PLOS Computational Biology*, *12*(4), e1004838. https://doi.org/10.1371/journal.pcbi.1004838

Wurm, F. M. (2004). Production of recombinant protein therapeutics in cultivated mammalian cells. *Nature Biotechnology*, *22*(11), 1393–1398. https://doi.org/10.1038/nbt1026

Xie, Q., Dai, Z., Hovy, E., Luong, M., & Le, Q. V. (2020). Unsupervised Data Augmentation for Consistency Training. *Advances in Neural Information Processing Systems*, *33*, 6256–6268.

Xiong, K., la Cour Karottki, K. J., Hefzi, H., Li, S., Grav, L. M., Li, S., Spahn, P., Lee, J. S., Ventina, I., Lee, G. M., Lewis, N. E., Kildegaard, H. F., & Pedersen, L. E. (2021). An optimized genome-wide, virus-free CRISPR screen for mammalian cells. *Cell Reports Methods*, *1*(4), 100062. https://doi.org/10.1016/j.crmeth.2021.100062

Yang, A., Martin, E., & Morris, J. (2011). Identification of semi-parametric hybrid process models. *Computers & Chemical Engineering*, *35*(1), 63–70. https://doi.org/10.1016/j.compchemeng.2010.05.002

Yang, O., Qadan, M., & Ierapetritou, M. (2020). Economic Analysis of Batch and Continuous Biopharmaceutical Antibody Production: a Review. *Journal of Pharmaceutical Innovation*, *15*(1), 182–200. https://doi.org/10.1007/s12247-018-09370-4

Yang, S., Navarathna, P., Ghosh, S., & Bequette, B. W. (2020). Hybrid Modeling in the Era of Smart Manufacturing. *Computers & Chemical Engineering*, *140*, 106874. https://doi.org/10.1016/j.compchemeng.2020.106874

Yeo, H. C., Hong, J., Lakshmanan, M., & Lee, D.-Y. (2020). Enzyme capacity-based genome scale modelling of CHO cells. *Metabolic Engineering*, *60*(April), 138–147. https://doi.org/10.1016/j.ymben.2020.04.005

Yu, L. X., Amidon, G., Khan, M. A., Hoag, S. W., Polli, J., Raju, G. K., & Woodcock, J. (2014). Understanding pharmaceutical quality by design. *AAPS Journal*, *16*(4), 771–783. https://doi.org/10.1208/s12248-014-9598-3

Zampieri, G., Vijayakumar, S., Yaneske, E., & Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLOS Computational Biology*, *15*(7), e1007084. https://doi.org/10.1371/journal.pcbi.1007084

Zaykin, D. V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology*, *24*(8), 1836–1841. https://doi.org/10.1111/j.1420-9101.2011.02297.x

Zhou, B., Xiao, J. F., Tuli, L., & Ressom, H. W. (2012). LC-MS-based metabolomics. *Molecular BioSystems*, *8*(2), 470–481. https://doi.org/10.1039/C1MB05350G

Zhou, W., Rehm, J., Europa, A., & Hu, W. (1997). Alteration of mammalian cell metabolism by dynamic nutrient feeding. *Cytotechnology*, *24*, 99–108.

Zürcher, P., Sokolov, M., Brühlmann, D., Ducommun, R., Stettler, M., Souquet, J., Jordan, M., Broly, H., Morbidelli, M., & Butté, A. (2020). Cell culture process metabolomics together with multivariate data analysis tools opens new routes for bioprocess development and glycosylation prediction. *Biotechnology Progress*, *36*(5), 1–11. https://doi.org/10.1002/btpr.3012

Acknowledgments

After this long Dissertation, it is the right time to finally acknowledge all the ones that gave an important contribution along this journey.

First of all, I would like to thank my supervisor, Prof. Pierantonio Facco, for all the technical and personal support provided since my first day at CAPE-Lab. Thanks for all the discussions that deeply contributed to my PhD project and my growth, as a researcher and person. I will remember these teachings for my entire future career. I will also remember all the food advice given along these years; it has been really appreciated.

I would also like to thank Prof. Massimiliano Barolo and Prof. Fabrizio Bezzo for the help and guidance throughout this journey.

Thanks to the industrial collaborators at GSK, Mr. Antonio Benedetti, Mrs. Paloma Diaz-Fernandez, and Mr. Gary Finka, for the possibility to work on stimulating, challenging, and impactful problems.

Thanks to Prof. Cleo Kontoravdi for hosting me with such short notice and for the invaluable opportunity to work on cutting edge problems; as well as to all colleges I had the pleasure to meet at Imperial College, especially Ben, James, Thanasis, and Kostis.

Un ringraziamento davvero speciale alla mia Famiglia, Marilena, Andrea e Clara, per avermi sempre spronato, supportato, appoggiato, e per avermi cresciuto come la persona che sono oggi. Grazie per avermi sempre dato la possibilità di perseguire i miei sogni, mi avete permesso di essere qui oggi.

Un ringraziamento a Raf, che ha creduto in me da sempre e mia ha accompagnato in tantissime avventure, a Dalmo, che mi ricorda sempre di vivere la vita con tranquillità, a Francesco, che ha condiviso con me le gioie ed i dolori di questo percorso, ed a Fabri, che ha sempre accompagnato in tutti i miei cambi di rotta con preziosi consigli.

Un ringraziamento a tutti gli amici, a casa ed in giro per il mondo, ai coinquilini vari, e agli amici del CAPE con cui ho avuto il piacere di condividere momenti in questi ultimi anni.

Enfim, mas não menos importante, um agradecimento muito especial a Ingrid. Obrigado por me levar nessa viagem, pelo apoio e amor, por sempre me aturar nos momentos mais difíceis. Obrigado por me dar a força para continuar todos os dias, sem você este objetivo nunca teria chegado. Foi uma longa viagem cheia de aventuras, mas com você todo desafio era mais leve. Obrigado pela vida que temos!