Model based process design for a monoclonal antibody-producing cell line: optimisation using hybrid modelling and an agent based system

Amy Jane Green

A thesis presented for the degree of
Engineering Doctorate in Biopharmaceutical Process Development

Biopharmaceutical & Bioprocessing Technology Centre
School of Chemical Engineering and Advanced Materials
Newcastle University

November 2015

.

Acknowledgements

The author would like to acknowledge her family, friends, supervisors, and Albert for their help in completing this thesis.

Preface

This thesis describes research that was undertaken as part of an Engineering Doctorate in Biopharmaceutical Process Development which was carried out in collaboration with Fujifilm Diosynth Biotechnologies and sponsored by the Engineering and Physical Sciences Research Council (EPSRC).

The thesis takes the format of a 'thesis by portfolio' which details a number of projects that are linked by the theme of modelling and optimisation techniques to be used in the pharmaceutical industry.

Being an industrially focussed Engineering Doctorate, the projects reflect the research requirements of industry, and changed over the period of study to meet new research challenges within the company.

The concluding chapter sets out recommendations to industry that have been made based on the outcomes of the research.

Abstract

The biopharmaceutical industry has seen rapid growth over the last 10 years in the area of therapeutic medicines. These include products such as monoclonal antibodies (mAbs) produced using mammalian cell lines such as Chinese Hamster Ovary (CHO). In order to comply with the regulatory authority (FDA) Quality by Design (QbD) and Process Analytical Technology (PAT) requirements, modelling can be used in the development and operation of the bioprocess. A model can assist in both the design, scale up and control of these complex, non-linear processes. A predictive model can be used to identify optimal operating conditions, which is vital for a contract manufacturer. Traditionally industry has approached modelling through the one-unit-at-a-time method, which can fail to capture unit interactions. The research reported in this work addresses this issue by using a whole system approach, which can also capture the interactions between units. Predictive models for each of the process units are combined within an overall framework allowing for the integration of the models, predicting how changes in the output of one unit influence the performance of subsequent units. These predictions can serve as the basis for the modifications to the standard operating procedures to achieve the required performance of the whole process.

In this thesis three distinct studies are presented; the first utilises a hybridoma data set and presents a model to predict and characterise the various critical quality attributes (CQAs), such as final product glycosylation profile, and critical process parameters (CPPs) including titre and viable cell count. The second data set concerns the purification of lactoferrin using ion-exchange chromatography as a model system for developing downstream

processing models. The output of this data set varied widely, and has led to the development of a novel peak isolation methodology, which can ultimately be used to characterise the elution. The final data set contains various CQAs and CPPs for multiple units within one process. This data set has been employed within a proof of concept study to show how an agent based framework can be developed to allow for overall process optimisation.

The results showed that it is possible to link process units using a common CPP or CQA. This work shows that using a agent based system of two layers of modelling i.e. individual process unit models connected with a higher level agent model that links via a common measurement allows for the influences between units to be considered. The model presented in this work considers the use of titre, HCP, measure of heterogeneity, and molecular weight as the common measurement. It is shown that it is possible to link the units in this way with the goal of predicting and controlling the glycosylation profile of the Bulk Drug Substance (BDS).

Contents

List of Figures

List of Tables

Nomenclature

α_g	Specific production rate for growing and non growing cells (μg/10^6 cells/day)
$\hat{u_h}$	Scores for **Y**
$\hat{y_i}$	Vector of predictions
μ	Specific growth rate (h^{-1})
μ_d	Specific death rate (h^{-1})
μ_{max}	Maximum specific growth rate (h^{-1})
μ_{min}	Minimum specific growth rate (h^{-1})
ε_b	Particle voidage
ε_p	Internal particle voidage
a_{if}	PARAFAC loadings matrix one in the i direction
b_h	Unknown parameter calculated by $= u'_h t_h / t'_h t_h$
b_{jf}	PARAFAC loadings matrix two in the j direction
C_{bi}	Chromatography column concentration
$C_{fi}(t)$	Feed concentration for chromatography column
c_{kf}	PARAFAC loadings matrix three in the k direction
C_{pi}	Concentration of protein in molecule
d_{gln}	First order decomposition rate of glutamine (h)
e_h	Vector of errors for each component
e_{ijk}	PARAFAC sum of squares of residuals
E_X	Residual errors matrix of **X**
E_Y	Residual errors of **Y**
f_{glc}	Glucose feed flow rate (L/h)
f_{gr}	Initial growing fraction

F_{in} Feed flow rate (L/h)

F_{out} Outlet flow rate (L/h)

k_{ap} Specific rate of apoptotic cell formation (h^{-1})

k_d maximum cell death rate

K_{glc} Monod constant for glucose (mM)

k_{gln5} Second constant for glutamine degradation (h^{-1})

K_{gln} constant for glutamine degradation (h^{-1})

k_i Film mass transfer coefficient

K_{lysis} Lysis rate of dead cells (h^{-1})

KI_{amm} constant for cell death due to ammonia accumulation (mM)

KI_{lac} constant for cell death due to lactate for CHO cells (mM)

m_{glc} Glucose maintenance coefficient (g of glucose consumed/g of cell/hr)

m_{gln} Glutamine maintenance coefficient (g of glutamine consumed/g of cell/hr)

P^T Loadings matrix of $\mathbf{X}$

p_n Principal components analysis loadings vector for the n^{th} component

pK_a Acid dissociation constant

Q^T Loadings matrix of $\mathbf{Y}$

Q_{amm} Specific consumption/production rate of ammonia (mmol/cell/h)

Q_A Specific consumption/production rate of amino acid A (mmol/cell/h)

Q_{lac} Specific consumption/production rate of lactate (mmol/cell/h)

Q_{MAb} Monoclonal antibody specific production rate (L/cell/h)

r_{amm} Ammonia removal rate (mmol/cell/h)

T Scores matrix of $\mathbf{X}$

t_0 Column dead time

t_g Retention time under gradient

t_h Scores for $\mathbf{X}$

t_n Principal components analysis scores vector for the n^{th} component

t_R Retention time

U Scores matrix of $\mathbf{Y}$

X_g Concentration of growing cells (cell/L)

X_{ap} Apoptotic cell concentration(cell/L)

X_{ng} Concentration of non growing cells (cell/L)

X_v Viable cell count (cell/L)

Y_{A_i/A_j} yield of amino acid A_i on amino acid A_j (mmol/mmol)

$Y_{amm/gln}$ Yield of ammonia from glutamine (mmol/mmol)

y_i Vector of true values

$Y_{lac/glc}$ Yield of lactate from glucose (mmol/mmol)

$Y_{x,A}$ yield of cells on amino acid A (cell/mmol)

$Y_{x/A}$ Yield of amino acid A on cells ($mmol/cell/hr$)

Y_X/glc Yield of cells on glucose (cell/mmol)

Y_X/gln Yield of cells on glutamine (cell/mmol)

[A] Extracellular concentration of amino acid A (mM)

[AMM] Concentration of ammonia (mM)

[GLC] Concentration of glucose (mM)

[GLN] Concentration of glutamine (mM)

[LAC] Concentration of lactate (mM)

[mAb] Extracellular concentration of monoclonal antibody (mM)

$\underline{\mathbf{X}}$ 3-Dimensional input array

$\mathbf{X}$ 2-Dimensional input matrix

$\mathbf{Y}$ 2-Dimensional response matrix

B Normalised gradient ramp

F Number of components/modes in PARAFAC model

L Column length

R Resin molecule diameter

R Slope of chromatography gradient ramp

u Flow velocity

V Volume of culture (L)

List of publications

Green, A. and Glassey, J. Multivariate analysis of the effect of operating conditions on hybridoma cell metabolism and glycosylation of produced antibody. *Journal of Chemical Technology and Biotechnology*, 90(2):303–313, 2015.

Chapter 1

Introduction

Within the pharmaceutical industry there has been a shift in the focus of research and development over the last 10 years from small molecules to one which places equal emphasis on small and large molecules. This is shown in an analysis of the biotechnology industry as carried out by Evens and Kaitin (2014) who highlight the increase in commercially available biotechnology products. This trend in the pharmaceutical industry is shown in figure 1.1, with the growth of biotechnology highlighted from the increase in products in clinical trials since 2001.

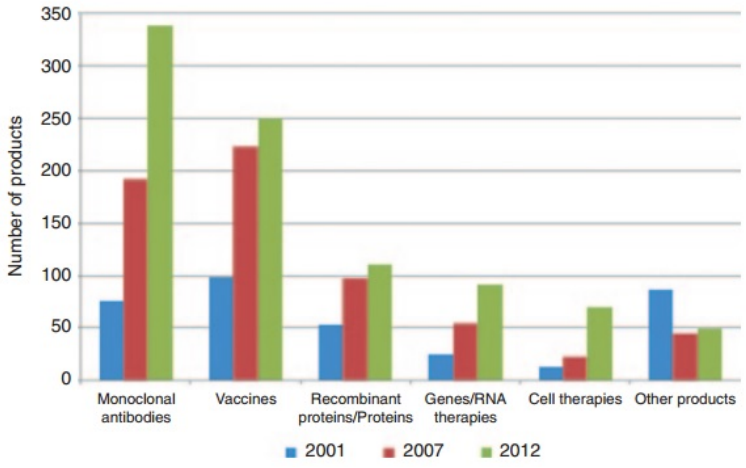


Figure 1.1: Types of biotechnology products in clinical trials during 2001, 2007, and 2012 (Evens and Kaitin, 2014)

The increased demand for biotherapeutic products over the last twenty years has prompted the development of large scale production; with emphasis

being placed on improving titre and/or yield. The primary method for doing this has been through development of the cell line and production media. However biopharmaceutical companies are now increasingly looking for innovative solutions to improve the production process through reducing the time to market, maintaining the cost effectiveness, and providing flexibility to the manufacturing process whilst maintaining critical quality attributes (CQAs). This change was driven by the introduction of an initiative by the Food and Drug Administration (FDA) covering process analytical technology (PAT) guidance and quality by design (QbD). This initiative addresses the designing, analysing, and control of critical quality attributes (CQAs) through critical process parameters (CPPs). Within the biopharmaceutical industry the identification and subsequent control of CPPs which have the greatest influence over the process allows for the establishment of a robust process with the ability to control CQAs and CPPs such as final product quality and yield. The research presented in this thesis aims to show that CPPs and CQAs of subsequent steps in a bioprocess are linked, and that overall final product CQAs can be controlled through the optimisation and control of multiple unit CPPs.

Production of complex therapeutic proteins, such as monoclonal antibodies (mAbs), is typically achieved through the use of mammalian cell lines. Mammalian cell lines which have been used extensively in the biotechnology industry include Chinese Hamster Ovary (CHO), Baby Hamster Kidney (BHK), and murine hybridoma (NS0). CHO cells are used in both industry and clinical development to produce approximately 70% of the products currently on the market (Jayapal *et al.*, 2007).

Despite their importance and popularity in biotechnology there is limited information as to the post translational activity of CHO cells, which can greatly affect the product. This lack of information includes the ability to link operational conditions to the metabolic pathways, and lack of detailed first principles models of processes such as glycosylation. Traditionally a biopharmaceutical process is developed in two stages; the first being the upstream fermentation and the second being the downstream purification.

There are reports in the literature as to the development of platform process and subsequent modelling of both of these stages (Li *et al.*, 2010; Shukla and Thömmes, 2010a; Low *et al.*, 2007). However there is currently no technique which would allow for the simultaneous development/optimisation of both upstream and downstream together.

To develop such an overall process modelling framework there are three separate factors required, these are process knowledge, process models, and data. The process knowledge is obtained through understanding the fundamental principles which govern each unit, for example the kinetics of adsorption which determine the operation of the ion exchange chromatography step. The process models can be formulated either based on fundamental understanding of the process operation or from historical data typically using multivariate data-based analysis techniques.

Multivariate modelling techniques can be used to predict future batch performance from a historical data set. With regards to process development this can aid in the optimisation of operating conditions, reduce experimental number, provide on-line monitoring, and ultimately assist in the prediction and control of critical quality attributes (CQAs) (Rathore *et al.*, 2014a; Glassey *et al.*, 2011b). The proposed whole process modelling framework will allow for the description of the whole process, the evaluation of the process and conditions, monitoring, and predicting performance. It can also serve as an evaluation tool for process scientists to use in system design. During process development, this will be used as a tool to allow the process scientists to explore the design space with minimal experimentation. To construct the overall model first each unit must be considered.

The cultivation stage of the process relies upon the growth, protein production, and harvest of the product. Cells are complex units, and as of yet their full metabolism has not been fully characterised, various product variations can be introduced at this stage which can affect the end product efficacy. Provided in this research is a data-based model which can predict the final glycosylation profile of the product. Furthermore through this research a greater understanding of the process of glycosylation is presented; showing

that glycosylation is a multi-step process with stages occurring in a particular order. Additionally this research combines first principle models with data-based models in a hybrid modelling structure, providing a characterisation of the cultivation which can be employed to monitor the process. The second unit considered is the ion exchange chromatography column.

In characterising the design space for a process and unit operation it is standard practise that a range of operating conditions are tried at various set points. In the case of chromatography this can cause significant changes in process output which can make modelling difficult. This research presents a method of peak isolation, through first principles and multivariate models. This can then allow for the modelling and prediction of the peak at various different conditions. The combination of first principles and multivariate modelling techniques have overcome the challenges of a small data set to produce accurate predictions. The final research presented in this contribution is a proof of concept study. An agent based system is used showing how a change in one process unit can influence and change the operation of another unit.

One of the main issues facing the research presented in this thesis is the lack of available data. Due to the sponsor company being a contract manufacturer it was not possible to use data which belonged to the client. Therefore the data used in this research was obtained form different studies used to establish new technology, and from other academic sources.

In summary there are three distinct modelling techniques applied in this research;

1. The application of statistical analysis and modelling tools to mammalian cell culture and purification.

2. The development of a hybrid model using MVDA modelling along side first principles models.

3. The application of an agent based model to allow for the simultaneous optimisation of process units.

1.1 Thesis layout and contribution

The chapters in this thesis present the work carried out over the course of the Engineering Doctorate study. Chapters 2 and 3 provide a summary of the key concepts which form the basis of the research, with Chapter 4 presenting the methodologies used in the development of the models. Chapters 5-7 present the results of the research, with chapters 5 and 6 being stand alone case studies and Chapter 7 being the agent model for overall process optimisation. Chapter 8 presents conclusions and recommendations for industry based upon the research carried out.

1.1.1 Chapter 2: Background information

This chapter presents background information necessary to understand the biological aspect of the work presented in this thesis. This chapter covers the structure and function of monoclonal antibodies, their production from cell cultivation, and an overview of the purification stages. Also given is a description of the bioreactor and ion exchange chromatography columns, providing the necessary information required for modelling these units.

1.1.2 Chapter 3: Literature review

This chapter presents a review fo the literature concerned with monoclonal antibody production, multivariate modelling, first principles modelling, and agent based modelling. This literature is presented in the context of biopharmaceutical development, with the focus being on techniques and tools which have been study that comply with the FDA's PAT and QbD guidelines.

1.1.3 Chapter 4: Methodologies and protocols

This chapter covers the methodologies for the experimental aspect of this research, and the protocols for the modelling research. The chapter describes

the generation of an ion exchange chromatography data set, and the origins of both the hybridoma and Chinese hamster ovary data sets. The protocols cover the techniques used to handle these data sets. Details are provided on array unfolding, data preprocessing, multivariate techniques, first principles models, hybrid model construction, and model performance assessment.

1.1.4 Chapter 5: Cell cultivation modelling

This chapter presents the models constructed for the cell cultivation process unit. This chapter uses the hybridoma data set to construct models which can then also be used, due to the similarities between the cells and products, with CHO cells. The models presented include PLS models to predict the CPPs of titre, and viable cell count, and the CQA of the final glycosylation profile. Both titre and viable cell count are also then predicted using first principles models. These two techniques are then combined to produce a hybrid model. It is this hybrid model which really develops the work shown in literature. As currently there few reports of hybrid models which predict both CPPs such as titre and viable cell count, whilst also accounting for concentrations of metabolites. This chapter presents a hybrid model which requires less data than multivariate applications whilst still being able to predict changes in operational conditions.

1.1.5 Chapter 6: Ion exchange chromatography modelling

This chapter presents a similar study for ion exchange chromatography. A lactoferrin data set was generated for use in constructing the models for this unit. Again multivariate modelling techniques were used to predict yield, with first principles models being used to generate the retention time of the experiments. The use of the first principles model to predict retention time allows for an automated model. It is this automated model, and the ability to handle widely varying data sets which is the contribution of this chapter.

1.1.6 Chapter 7: Agent based modelling

The concept presented in this chapter is the main focus of the research. Showing that an agent based model can be used to predict how changes in the operating conditions fo one process unit impact on the performance of another unit. This chapter uses the models constructed in Chapters 5 and 6 to link together the cultivation and ion exchange process steps. The chapter acts as a proof of concept showing that it is possible to construct an agent based model for a mammalian cell process. It can be seen that the cultivation and purification stages can be linked with changes to the cultivation operating conditions causing changes to the operation of the chromatography column.

1.1.7 Chapter 8: Conclusions and future work

This chapter concludes the work presented in this thesis, along with making recommendations for the implementation of the research in industry, and suggestions for future development.

Chapter 2

An Overview of Bioprocessing

Chapter 1 introduced the current position of the pharmaceutical industry and showed the increasing demand over the last decade for biotherapeutic products such as those derived from monoclonal antibodies (mAbs). This chapter aims to introduce the fundamental science around the production of biopharmaceuticals from cell culture production systems with emphasis on the impact of different process variables upon growth, production, purification, final product quality and the economic viability of the manufacturing process.

2.1 Cell culture

Cell culture is the process whereby cells are grown outside of their natural environment, under controlled conditions (*in vitro*). The term cell culture refers to cells which have been taken from a multicellular organism. These cells are eukaryotic as opposed to simpler prokaryotic cells, such as *E. coli* or *B. subtilis*; which are incapable of performing complex post-translational modifications (Wurm, 2004). Prokaryotic and eukaryotic cells are both capable of producing proteins, and both are capable of performing post-translational modification, but the complexity of these modifications is greater in eukaryotic cells. This is due to the structure of the cells; prokaryotic cells do not have internal membrane bound organelles whereas eukaryotic do. In prokaryotic cells gene expression and protein synthesis occurs in the cytoplasm whereas

for eukaryotic cells gene expression occurs in the nucleus and protein synthesis in the cytoplasm. In eukaryotic cells there are organelles (Golgi body and endoplasmic reticulum) which are the site for post translational modifications. The Golgi body enzymes catalyse these reactions which include glycosylation, phosphorylation, and ubiquitylation. This research is concerned with glycosylation, which is the addition of a carbohydrate side chain to a protein.

Disadvantages of using mammalian cell expression systems include high maintenance costs (in comparison to prokaryotic based systems), long culture duration, and high nutrient requirements. Furthermore manufacturing sites that use mammalian cells require operators that are specially trained and qualified. In contrast mammalian cells can produce proteins with a similar structure to naturally occurring proteins, as the post translational modifications are similar. This is particularly important when the product is for human use, as proteins with glycans identical to naturally produced glycoproteins are less likely to produce an immune response. Some mammalian cell lines which are commonly used expression systems for large-scale recombinant protein production are Baby Hamster Kidney (BHK-12), murine myeloma (NS0), and Chinese Hamster Ovary (CHO).

2.1.1 Baby Hamster Kidney (BHK)

Production of vaccines from cells grown in culture started in 1954 using monkey kidney cells (Vero) (Eibl *et al.*, 2008). However with increased demand for large quantities of foot and mouth disease vaccine (FMD) it was decided that the Vero cell line was too expensive (Butler, 2004). Baby hamster kidney (BHK) cells were adapted for use in producing FMD vaccine as they could easily be scaled up to large volumes (5000L) (Ozturk and Hu, 2006). As more vaccines were developed for use in humans, BHK were suggested as possible cell culture producers as they could be cultured in suspension and had been approved for vaccine production (Eibl *et al.*, 2008). Issues were raised, however, as it was discovered that BHK cells contain virus particles harmful to humans (Ozturk and Hu, 2006). Therefore BHK are now primarily used for production of veterinary medicines (Butler, 2004).

2.1.2 Hybridoma

Hybridoma cell lines were originally developed from murine myeloma cells. In the 1970s Köhler and Milstein (1975) developed the technique of producing cells which are a hybrid of B-lymphocytes and myelomas. B-lymphocytes produce the desired antibodies, but they are mortal. After fusion with myeloma cells the resulting cell can produce the antibody and become immortal, meaning they can reproduce indefinitely (Shuler and Kargi, 2010; Galfre *et al.*, 2007; Butler, 2004). Hybridomas are well adapted for large-scale production of antibodies as they can be grown suspended in culture. Furthermore the antibodies produced by hybridoma cells have a wide range of applications because of the high specificity in recognising proteins. Further information on the synthesis and production of antibodies from hybridoma cells can be found in work of Butler (2004).

Murine derived hybridomas, and the subsequent antibodies, are widely used as reagents, in purification, and for diagnostic tests. However they have had limited success in the treatment of humans. This is because antibodies secreted in mice and humans have different constant regions (F_c) and thus an antibody derived from a mouse and injected into a human could elicit an undesirable immune response (Sofer and Hagel, 1997). (More detailed information on the structure of antibodies, and the F_c region is provided in section 2.2.1.)

2.1.3 Chinese Hamster Ovary (CHO)

CHO cell lines are one of the most widely applied mammalian cell lines in industry, accounting for 70% of recombinant protein production (Li *et al.*, 2010). As previously mentioned mammalian cells are used for recombinant protein production as they have the ability to perform the required post-translation modifications. CHO cells are particularly popular as they synthesise proteins whose glycoforms are similar to native human products (Butler, 2004). Figure 2.1 shows that between 1994 and 2012 CHO cell lines

account for 40% of mAb production (Reichert, 2012).

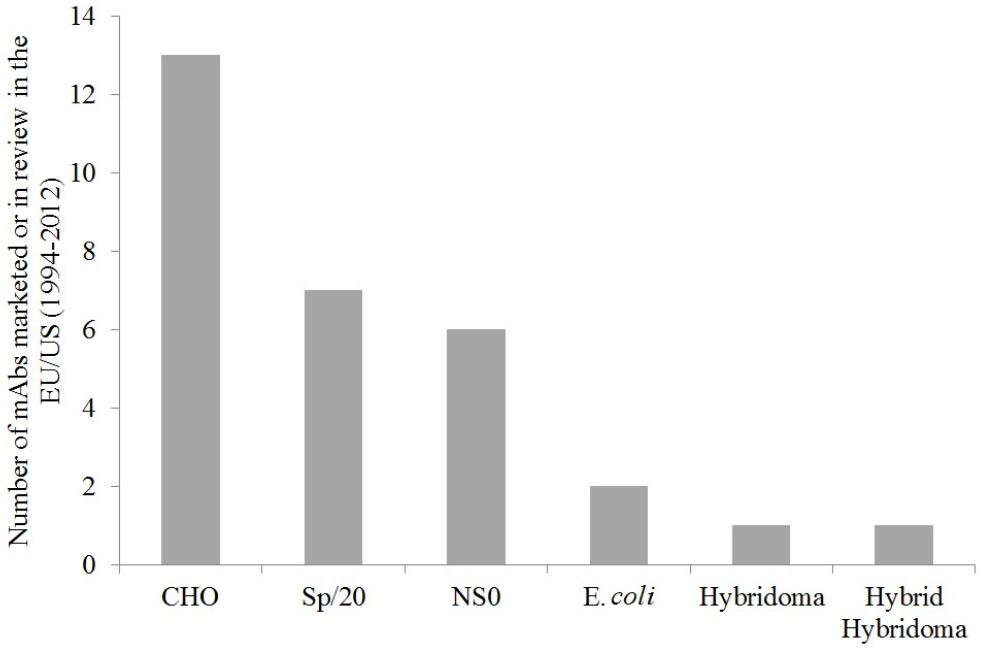


Figure 2.1: Manufacturing cell lines used for mAb production between 1994 and 2012 in the European Union and United States (Reichert, 2012)

CHO cells are popular because they work well with recombinant DNA techniques. This is when DNA is isolated and transferred from one species to another. It can also be used so that cells express high levels of a protein (over-expression) which is particularly useful in manufacture (Butler, 2004). Along with the gene of interest a marker gene is also transfected, an example would be the CHO cell line derivative known as dhfr-CHO. This is a cell line which has been altered so it is able to produce the dihydrofolate reductase enzyme (DHFR). This gives a selective advantage to the transfected cells, so that they can be grown in an environment where producing the DHFR enzyme allows them to out compete non transfected cells. Another commonly used marker gene is glutamine synthetase (GS), which was developed at Lonza. More information on both of these marker genes can be found in the work of Costa *et al.* (2010).

Table 2.1 provides a summary of common cell lines which can be obtained from cell culture collections; a brief note of the application of the cell line and references to literature concerning the cell line are provided.

Table 2.1: Examples of common cell lines obtainable from culture collections. Information on cultivation conditions obtained from Doyle and Griffiths (1998).

Cell line	Origin	Cell type	Comments	Cultivation media	References
BHK	Baby hamster kidney	Fibroblast	Used for vaccine production, cells are anchorage dependant but can be induced into a suspension.	Eagle's basal medium supplemented with 2% tryptose phosphate broth (TPB) and other supplementation, including glucose, glutamine, vitamins, lactalbumin hydrolysate and pluronic F-68	Hernandez and Brown (2010); Moreira *et al.* (1994)
CHO	Chinese hamster ovary	Epithelial	If a surface is available cells will attach to it, but they will also grow in suspension. Used extensively for genetic engineering	Suspension culture using F12 medium, with 10 % foetal bovine serum (FBS). Vessel should be gassed with 5 % CO_2. Cells require hypoxanthine, glycine, thymidine, and proline for growth.	Hacker *et al.* (2009); Li *et al.* (2010)
HeLa	Human cervical carcinoma	Epithelial	First isolated in the 1950s, it is a fast growing human cancer cell.	Cells are cultured in Eagle's MEM (EBSS) supplemented with glutamine, non- essential amino acids and 10 % FBS	Doyle and Griffiths (1998)

L	Mouse connective tissue	Fibroblast	A lot of cell development in the 1950s was based on this cell line	Cells are cultured in DMEM with 10% FBS	
L6	Rat skeletal muscle	Myoblast	Used for differentiation of muscle cells	Cells are cultured in Eagle's basal medium with 10% FBS	
MDCK	(Madine Darby) canine kidney	Epithelial	Anchorage dependant cells with favourable growth characteristics. Used in the production of veterinary vaccines.	Cells are cultured in Eagle's MEM with 10% FBS	Doyle and Griffiths (1998)
MRC-5	Human embryonic lung	Fibroblast	Has a finite life span, is used for human vaccine production.	Cells are cultured in Eagle's MEM with 10 % FBS. Different media formulations can increase lifespan of cells if desired.	Doyle and Griffiths (1998)
MPC-11	Mouse myeloma	Lymphoblast	Derived from mouse tumours, is used to produce antibodies	Cells are cultured in RPMI 1640 with glutamine and 10% FBS. The cells die in the presence of HAT medium.	
Namalwa	Human lymphoma	Lymphoblast	Derived from cells of a human suffering from Burkitt's syndrome; used for alpha interferon production.	Cells are cultured in PMI 1640 with 5-15 % FBS at 37°C and pH 6.8-7.0,	Doyle and Griffiths (1998)

NB41A3	Mouse neuroblastoma	Neuronal	Derived from tumour cells, with favourable growth characteristics. These cells have similar properties to nerve cells.	Cells are cultured in DMEM supplemented with 2mM glutmaine with 10% FBS	
3T3	Mouse connective tissue	Fibroblast	Fast growing in suspension, used for development of cell culture techniques.	Cells are cultured in DMEM supplemented with 2mM glutmaine with 10% FBS	
WI-38	Human embryonic lung	Fibroblast	Finite lifespan, used for human vaccine production.	Cells are cultured in Eagle's basal medium with 10% FBS	
Vero	African green monkey kidney	Fibroblast	Infinite lifespan, but shares characteristics with finite cells, used for human vaccine production.	Cultured in Dulbecco's modified Eagle's medium (DMEM) with 10 % FBS.	Butler (2004); Doyle and Griffiths (1998)

2.2 Monoclonal Antibodies (mAbs)

Antibodies are protein molecules, known as immunoglobulins, synthesised by the immune system of an organism in response to a foreign macromolecule, known as an antigen (Galfre *et al.*, 2007). In cell culture a population of cells derived asexually from the same parent cell are monoclonal therefore the antibodies produced are termed *monoclonal antibodies* (Galfre *et al.*, 2007). Production of mAbs in research and development was started by Köhler and Milstein (1975) who developed the used of hybridomas as an expression system. The popularity of mAbs in industrial production has grown rapidly, as shown in Figure 1.1, between 2001 and 2012 mAbs were one of the fastest growing areas of biotechnology in terms of products in clinical trials Evens and Kaitin (2014). The evidence of this growth is further supported by the four new mAb products which were approved by the Food and Drug Administration (FDA) in 2014 (Reichert, 2015) (Table 2.2).

Table 2.2: Therapeutic antibodies marketed in the EU and US in 2013/2014 (Reichert, 2015).

Product	Trade name	Target	Year
Adotrastuzuma-bemtansine	Kadcyla®	Humanised IgG1 (Breast cancer)	2013
Obinutuzumab	Gazyva®	Humanised IgG1 (Chronic lymphocytic leukemia)	2013
Siltuximab	Sylvant®	Chimeric IgG1 (Castleman disease)	2014
Vedolizumab	Entyvio®	Humanised IgG1 (Ulcerative colitis, Crohn disease)	2014
Ramucirumab	Cyramza®	Humanised IgG1 (Gastric cancer)	2014
Pembrolizumab	Keytruda®	Humanised IgG4 (Melanoma)	2014

Table 2.2 shows the mAb products brought to market in the EU and US in 2013/2014. This growth in the market is reflected in the global market value of antibody drugs, which in 2013 was $63.4 billion. It is expected that this market

will grow at a rate of 12.2 % between 2014 and 2019, with the result being that by 2019 antibody drugs will have a market value of $122.6 billion (Dewan, 2015). This development potential comes from the interest shown in not just developed, but also emerging markets (Imarcgroup, 2012). Additionally the interest in mAbs is due to the wide range of diseases that can be treated using them, as shown by the cross section of medical conditions which can be treated by the products brought to market in 2013/2014 (Table 2.2).

2.2.1 Structure

All antibodies have a similar structure containing four polypeptide chains; two light and two heavy chains as shown in Figure 2.2, named according to the molecular weight of the chains. Each of the antibody chains has a variable amino acid sequence in the F_{ab} region which is where the antigen specific binding occurs. The flexible hinge region allows for antigen binding sites to be variable distance apart. Once the antigen has bound to the F_{ab} region the F_c region (which is constant for all antibodies of the same class) interacts with other components of the immune system. For example cancer cells are not recognised by the immune system as foreign. Therefore when a mAb attaches to the surface of a cancer cell the immune system recognises the mAb as a foreign body and subsequently the cancer cell. In this way it can be said the mAbs are acting as markers.

The production of mAbs for use as human therapeutics presents many challenges. The first mAb product produced was Orthoclone, an immunosuppressent against kidney transplant rejection. The product was not successful and failed during clinical trials as patients who received the drug developed an immune response to it (Ezzell, 2001). It was discovered that the immune response was caused by the fact the mAbs were made using murine derived hybridomas. The immune system recognises mAbs produced this way as foreign and causes the human anti-murine antibody response (HAMA) (Sofer and Hagel, 1997; Butler, 2004). This is an example of how product properties and characteristics need to be considered as they can impact on the safety and/or efficacy of the product (Rathore *et al.*, 2014a).

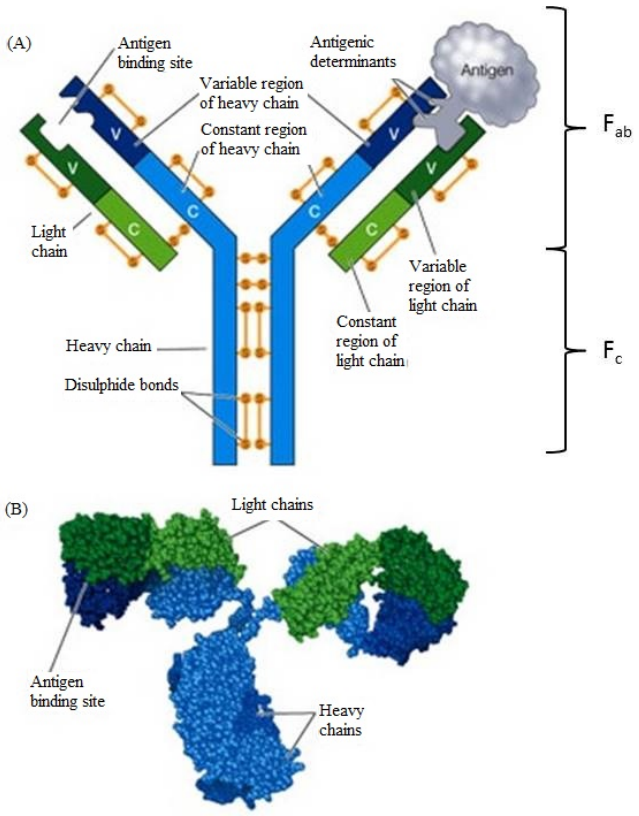


Figure 2.2: Structure of antibody (Immunoglobulin). (A) The diagram shows the four poly peptide chains (two light, two heavy), the antigen binding site, and the F_{ab} and F_c regions (B) A three dimensional model of an antibody molecule in the same orientation as (A)(Sadava *et al.*, 2007)

In the case of mAbs produced from murine hybridoma cells, CHO cells (Li *et al.*, 2010) were used as an alternative expression system as they offer control over properties and characteristics of the product as summarised by Costa *et al.* (2010):

- CHO cell produced mAbs are safe for use in humans.
- The glycan structure of the product is similar to that of naturally produced human mAbs.
- The ease of transfection i.e. the introduction of new genetic material to cell.
- They are a powerful gene amplification tool and can be used in the replication of specific genes so that more of a specific protein is produced.
- They can be easily adapted to grow in suspension and serum free media.
- They can be used to produce protein with a glycosylation profile similar to that of naturally occurring human proteins.

This research is particularly concerned with the last of these points, which is the identification and control of glycosylation of mAbs produced from CHO cell cultivation.

2.2.2 Critical Quality Attributes (CQAs)

Critical Quality Attributes (CQAs) are properties or characteristics that should be maintained within appropriate limits to ensure the desired product quality (Rathore *et al.*, 2014b). Although it is not necessary for a company to fully assess the CQAs of a biopharmaceutical product (ICH guideline Q11, 2011), it is considered the first step to a quality by design (QbD) approach. Furthermore characterising a product can help in the design and subsequent regulatory approval of the product. Some examples of mammalian cell cultivation CQAs include: the capability of attaining high product yield, the ability to perform post-translational modifications, solubility, stability, therapeutic efficiency, and time taken to be cleared from the body (Jayapal *et al.*, 2007). Defining these attributes prior to product manufacture is important, as the parameters which influence them can be identified and controlled. These are referred to as Critical Process Parameters (CPPs). Table 2.1 briefly describes some of the cultivation conditions from different cell lines. The influence of these parameters on CQAs will be discussed further in the following section.

Glycosylation

As mentioned in section 2.1 one of the benefits of mammalian cells is their ability to perform the required post translational modifications, such as for example glycosylation. There are various structures for a glycosylated proteins used for human therapeutics it is important their structure resembles a naturally occurring human one. Glycosylation is a reaction where a carbohydrate is attached to a functional group of another molecule; generally these molecules are proteins or lipids. With reference to the research presented in this thesis a glycosylation reaction produces a glycoprotein, i.e. a protein with a covalently attached carbohydrate chain. The carbohydrate chain is an

oligosaccharide formed from single monosaccharides such as glucose or fructose or others. When this chain is attached to a glycoprotein it is commonly called the glycan. The glycosylation reaction which produces the glycoprotein is a post translational modification, occurring after protein synthesis but prior to secretion from the cell.

The glycosylation of the F_c region is essential for effector functions of the antibody, such as complementary binding once in the human recipient (Butler, 2004). Additionally, approximately 20 % of human antibodies are glycosylated in the F_{ab} region, where glycans can be important for antigen binding. The level of glycosylation of antibodies is small (2-3 % weight) compared to other proteins. However, it is known that the glycan structures of antibodies can significantly impact on the immune response (an example of the HAMA response is described in section 2.2.1). Figure 2.3 shows the common glycans of IgG with 0, 1, or 2 galactose terminal residues (G0, G1, and G2)

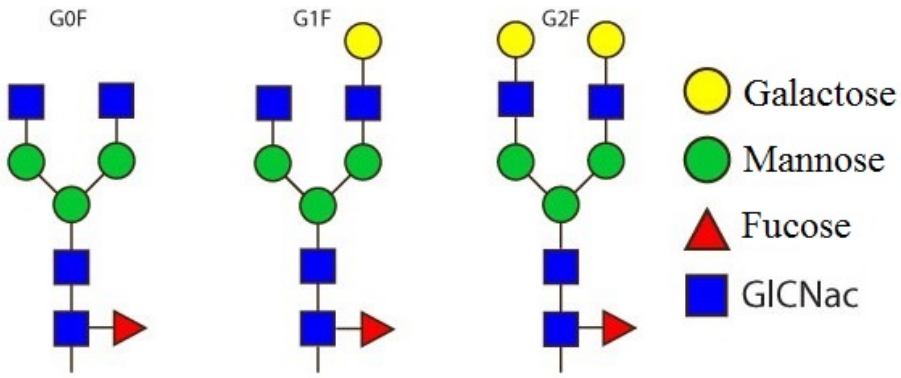


Figure 2.3: Three of the most abundant glycans in mAb biopharmaceuticals. These glycans bind to the F_c region and are terminated by 0, 1 or 2 galactoses. They are called G0, G1 or G2 respectively (Sasorith and Lefranc, 2004). GLCNac is N-Acetylglucosamine a derivative of glucose.

To avoid the undesired immune response it is important that the cultivation maximises the content of fully processed glycoproteins. There are various factors which affect the final glycoprotein, such as the metabolic profile of the cell, and the environmental conditions of the culture. Process parameters which have been shown to impact glycosylation include: concentration of ammonia, dissolved oxygen levels (DO), concentration of glucose, pH, and composition of media (Ivarsson *et al.*, 2014; Hossler *et al.*, 2009; Kunkel *et al.*, 1998; Spearman *et al.*, 2007; Pacis *et al.*, 2011a). It has

been shown that glycosylation can be controlled by temperature, with higher levels of certain glycoproteins being produced at lower temperatures (Spearman *et al.*, 2007). Furthermore Pacis *et al.* (2011a) showed that high levels of mannose glycosylation is strongly correlated with osmolarity levels and extended cultivation durations. Kunkel *et al.* (1998) investigated how varying DO levels effect glycosylation and showed that varying the DO did not impact upon quantity of mAb produced, but noticeable differences in the glycoproteins were observed. Normal levels of the 3 main glycoproteins (Figure 2.3) were observed at 50 % DO.

2.3 Bioprocessing

Techniques used for the cultivation of mammalian cells differ greatly from the techniques used to grow bacteria, yeasts, and fungi. A simplified representation of a typical mammalian cell bioprocess can be seen in Figure 2.4. There are four distinct stages of the process, namely inoculation, cultivation scale up, clarification, and purification.

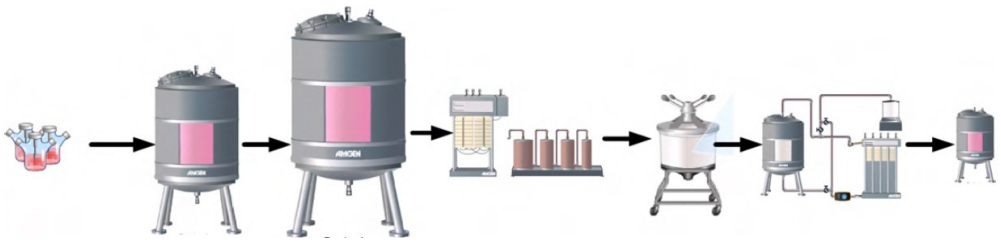


Figure 2.4: Simplified bioprocess manufacture (inoculation, scale up bioreactors, clarification and purification) (Ündey *et al.*, 2010)

The inoculation is preceded by the development of the cell line. Currently most cell line development by biopharmaceutical companies is based upon the inhibition of DHFR (described in section 2.1.3) or on Lonza's glutamine synthetase (GS) system (Costa *et al.*, 2010). The DHFR technique has been established for longer than the GS which was only developed recently, thus the DHFR technique is more widely adopted in industry. A typical cell line development scheme is shown in Figure A1 in appendix. The expression vector for the product is inserted into the host cell line, during a process known as transfection. The cells are then selected and the genes amplified, leading to

an increase in the recombinant gene copy number in the cells. The next step, single cell cloning or limiting dilution, ensures that the cells selected for processing produce the target protein. The cells are analysed by examining product titre, so that the highest producer can be chosen for expansion. Further information on cell line development can be found in the work of Lai *et al.* (2013). After the cell line has been developed, it can be transferred under aseptic conditions to a growth medium containing serum and small amounts of an antibiotic in small T-flasks. These cells form the primary culture. At this stage mammalian cells do not form aggregates but grow in the form of monolayers on a support surface (Shuler and Kargi, 2010). If a cell line is then obtained from this primary culture it is referred to as a secondary culture which can be adapted during the scale up stage to grow in a suspension rather than on a support. Scale up is important as the cells are in the growth phase and the aim is to obtain a suspension with a concentration high enough to be used in the production bioreactor and produce enough product (Shuler and Kargi, 2010).

Large scale production aims to take the cells through the growth phase into the production phase, whereby the cultivation is optimised to obtain maximum product. The production bioreactor unit is described in detail in section 2.3.1.

Antibodies can be recovered using a variety of methods depending on the purity required. For use in patients a high purity is required, which means that levels of contaminants must be reduced to acceptable levels (details can be found in ICH guideline Q3A (2008); ICH guideline Q3B (2006)). Additionally the purification process has to be validated to confirm that it removes contaminating substances regardless of whether the presence of these substances can be shown in the antibody source (Galfre *et al.*, 2007).

The first stage in antibody purification is clarification involving the lysis and subsequent removal of debris and large macroimpurities. If product is intracellular, as in the case of mAbs produced from CHO cells, the proteins need to be released. A review of techniques for large scale disruption of cells is given by Kula and Schütte (1987), with specific application to mammalian

cells described by Tait *et al.* (2009). The most common technique for clarification is through filtration and/or centrifugation (Liu *et al.*, 2010). It is important that the correct technique is used as the process of lysis will affect the viscosity (Abraham *et al.*, 2004). Typically, the use of a centrifugal separator will have a 90 % recovery of product with 99.9 % of the cell debris being removed (Sofer and Hagel, 1997). Clarification is important as removal of macroimpurities increases the performance of subsequent steps (e.g. prevents clogging of chromatography columns). The largest impurity (by volume), water, needs to be removed prior to purification, to concentrate the product. This process is generally carried out by ultra-filtration or by protein precipitation. Generally, purification is carried out using a system of chromatography techniques which are complementary to each other as this keeps the number of steps low (Liu *et al.*, 2010). The purification steps are used to remove impurities which, depending on the product source, can include viruses, endotoxins, nucleic acids, host cell proteins, mutated or modified proteins, modified oligonucleotides or peptides, cell culture additives, and processing chemicals (Shuler and Kargi, 2010). Additionally in some cases modified products are produced during purification. These can include for example, aggregates which can be highly immunogenic (Fradkin *et al.*, 2009). These impurities also have to be removed to make a safe product. A key requirement is that a purification process must be reproducible and this is achieved by demonstrating reproducibility using analytical methods, to both identify and prove the removal of impurities (Sofer and Hagel, 1997).

The research presented in this thesis is primarily concerned with the production bioreactor and the ion exchange chromatography column. However a more detailed account of the purification methods can be found in Desai (2000).

2.3.1 Production bioreactor

To obtain the required cell density and maximum product concentration the type and design of the bioreactor and the operation of it must be carefully considered. When considering the large scale production there are CQAs of

the product which influence the type of bioreactor used. Section 2.2.2 lists some of these which are important for CHO cells and in particular the glycosylation of the mAb product. These CQAs are, as discussed, influenced by CPPs which include:

- Method of air supply
- Dissolved Oxygen
- Temperature
- pH
- Culture mixing

The following section discusses the monitoring and control of these parameters, along with the different bioreactor designs.

Bioreactor design and control

The simplest and most widely used bioreactor is the stirred tank bioreactor (STR). This is the traditional design for growing cells and has been used extensively for bacterial and yeast fermentations. However mammalian cells are large (10 to 20 μm diameter), slow growing ($t_d \approx$ 10h to 50h), and often very sensitive to shear damage and therefore require a modified design (Sofer and Hagel, 1997). A typical 5L STR used in mammalian cell culture is shown in Figure 2.5. As can be seen the bioreactor is mechanically agitated using an impeller with the stirring shaft fitting through the head plate. The impeller blades are designed so that it mixes both radially and axially to minimise the shear whilst maximising the effectiveness of the mixing. The base of the vessel is also curved to promote mixing and prevent dead zones.

As can be seen in Figure 2.5 allowance is also made in the head plate for the addition of probes to control various CPPs. The first of these parameters is temperature, which is measured using a probe, and controlled by the outer jacket. The second controlled parameter is pH, which again is monitored by a probe. Optimal pH for cell cultures is approximately pH 7.4 and maximum cell growth occurs when this is maintained. One method of pH control is

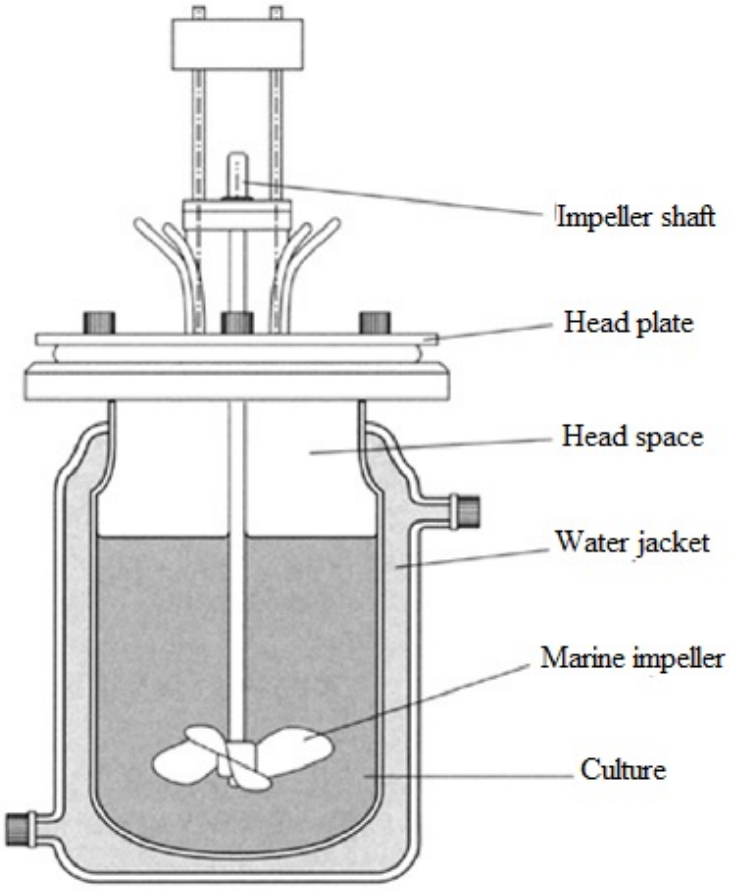


Figure 2.5: Stirred tank bioreactor (Butler, 2004).

through the addition of CO_2 however this is limited by the available head space in the bioreactor. An alternative pH control method is direct addition of acid or alkali via a pump. Another parameter which is monitored is the dissolved oxygen, as the cells need enough oxygen for the cell metabolism requirements. As culture volume increases so does the complexity of the oxygen supply system. In T-flasks the cells can generally get enough oxygen through head space diffusion, whereas for reactors larger than 1L the volume to surface ratio is too low (Shuler and Kargi, 2010). For larger volume bioreactors, they are aerated by sparging. Although this can cause issues with bubble bursting and foaming (Butler, 2004).

There are other types of bioreactors which can also be used in mammalian cell culture, the most common of these are summarised in Table 2.3.

Table 2.3: Examples of production scale bioreactors, references provide information on the structure of the bioreactor or studies carried out using the bioreactor.

Bioreactor type	Design	Comments	References
Stirred tank reactor (STR)	This is the most widely use type of bioreactor. it is a simple design consisting of a cylindrical vessel with a stirrer.	There can be issues with bubble damage, shear, and foaming.	Heath and Kiss (2007); Catapano *et al.* (2009); Butler (2004); Ivarsson *et al.* (2014)
Airlift	Column reactor with an internal draught tube. Fluid is circulated by a stream of air with passes through the inside of the draught tube	This is a simple system, with no mechanical components so it is less susceptible to breaking down. Bubble damage and foaming are minimised.	Heath and Kiss (2007); Catapano *et al.* (2009); Butler (2004)
Hollow fibre	This reactor contains bundles of hollow fibres which provide a matrix for cell growth. Liquid can flow through the fibres and the space between fibres.	Media must be pumped through the fibres to supply the cells, this can create high pressures, nutrient gradients, and uneven cell growth.	Heath and Kiss (2007); Catapano *et al.* (2009); Butler (2004); Lipman and Jackson (1998)
Packed bed	The packed element of this bioreactor provides a matrix for cell growth with a continuous flow of medium.	The aim of this reactor is to provide maximum surface area for growth. The packed element can be; glass beads, ceramic cylinder with channels, or stacked mesh plates.	Heath and Kiss (2007); Catapano *et al.* (2009); Butler (2004); Meuwly *et al.* (2007); Golmakany *et al.* (2005); Ducommun *et al.* (2002); Wang *et al.* (1992)

Fluidized bed	Cells are immobilised in beads and held in suspension by upward flow of medium.	This type of bioreactor can be difficult to operate on larger scales.	Heath and Kiss (2007); Catapano *et al.* (2009); Butler (2004)
Disposable	Has a disposable bag, which is in contact with the cell culture, it is encased in a permanent structure.	There are two types which differ in the method of agitation, one with an internal stirrer and the other uses a rocking motion (wave)	Heath and Kiss (2007); Tang *et al.* (2007); Catapano *et al.* (2009); Minow *et al.* (2012)

Cultivation characteristics

The cultivation is controlled so that the cells are provided with the right environment at each stage as shown in Figure 2.6. The first stage is the lag phase with little growth when the cells are established within the new culture. Second is the growth phase which can last between 2 - 5 days. This stage is typically characterised by the consumption of glucose, and can often be indicated through an increase in pH as toxic metabolites such as ammonia are produced. The stationary phase is relatively short and the concentration of viable cells drops significantly due to the build up of the toxic metabolites such as lactate and ammonia. However the formation of the product (i.e. mAbs) can occur during both the growth phase and after growth stops (Shuler and Kargi, 2010).

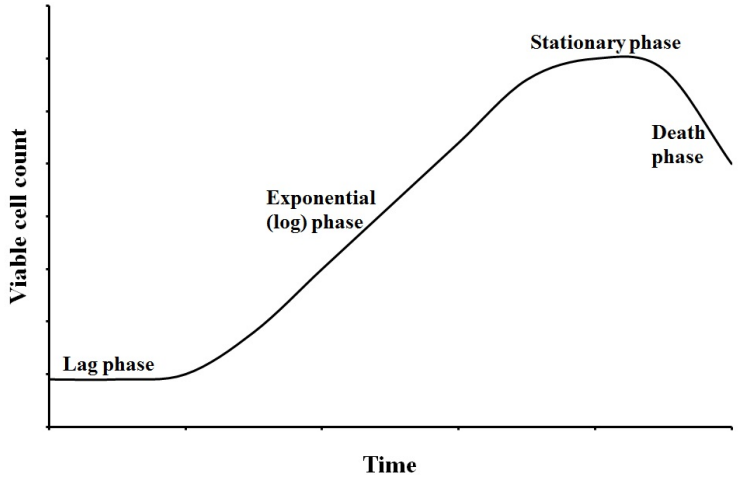


Figure 2.6: The cell growth curve: illustrating the lag phase when the bioreactor is first set up, the subsequent growth phase of cells, the production phase of the protein, and finally the death phase which occurs when the available nutrients have been consumed.

The specific growth rate (μ) of the cell population is defined as the number of new cells produced per unit of living cells present in the culture medium per unit time. It is limited by the concentration of the limiting substrate (S), which for CHO cells is typically glucose and glutamine. The specific cell growth rate is given in equation 2.1 (Xing *et al.*, 2010).

$$\mu = \mu_{max}\left(\frac{[GLC]}{K_{glc}+[GLC]}\right)\cdot\left(\frac{[GLN]}{K_{gln}+[GLN]}\right)\cdot\left(\frac{[KI_{lac}]}{KI_{lac}+[LAC]}\right)\cdot\left(\frac{[KI_{amm}]}{KI_{amm}+[AMM]}\right) \quad (2.1)$$

where [GLC] is the concentration of glucose (mmol/L), [GLN] is the concentration of glutamine (mmol/L), [LAC] is the concentration of lactate (mmol/L), and [AMM] is the concentration of ammonia (mmol/L). K_{glc} and K_{gln} are the half maximum rate concentration of glucose and glutamine for cell growth rate (mM). KI_{lac} and KI_{amm} are the half maximum rate concentration of lactate and ammonia for cell growth rate (mM). (μ_{max}) is the observed maximum specific growth rate (hr^{-1}) in the absence of any limitations from nutrients or inhibition metabolites. During culture the specific growth rate usually decreases due to the depletion of nutrients or accumulation of inhibitory metabolites. The specific rate of cell death (μ_d) is defined as the number of dying cells per unit of living cells present in the culture medium per unit of time (equation 2.2).

$$\mu_d = k_d \left(\frac{[LAC]}{KD_{lac} + [LAC]} \right) \cdot \left(\frac{[AMM]}{KD_{amm} + [AMM]} \right) \tag{2.2}$$

where k_d is the maximum cell death rate. KD_{lac} and KD_{amm} are half maximum rate concentration of lactate and ammonia for cell death rate (mM). Cell death is often low at the start of the culture and then increases due to nutrient depletion or accumulation of inhibitory metabolites. The change in the concentration of glucose and glutamine is defined as the number of millimoles of nutrient consumed per unit of living cells present in the culture medium per unit of time. For mammalian cell cultures it is generally found to have a linear relationship to cell growth, and is given in equations 2.3 and 2.4.

$$\frac{d[GLC]}{dt} = -\left(\frac{(\mu - \mu_d)}{Y_{X/glc} + m_{glc}} \right) \cdot X_v \tag{2.3}$$

$$\frac{d[GLN]}{dt} = -\left(\frac{(\mu - \mu_d)}{Y_{X/gln} + m_{gln}} \right) \cdot X_v - d_{gln}[GLN] \tag{2.4}$$

where m_{glc} and m_{gln} are the glucose and glutamine maintenance coefficients (g of glucose consumed/g of cell/hr). $Y_{X/glc}$ and $Y_{X/gln}$ are the glucose and glutamine yield coefficients (cell/mmol). d_{gln} is the decomposition rate of glutamine. The change in concentration of lactate and ammonia is defined as the number of millimoles of metabolite excreted per unit of living cells present

in the culture medium per unit time (equations 2.5 and 2.6) (Xing *et al.*, 2010).

$$\frac{d[LAC]}{dt} = Y_{lac/glc} \cdot \left(\frac{(\mu - \mu_d)}{Y_{X/glc} + m_{glc}} \right) \cdot X_v \tag{2.5}$$

$$\frac{d[AMM]}{dt} = Y_{amm/gln} \cdot \left(\frac{(\mu - \mu_d)}{Y_{X/gln}} \right) \cdot X_v - r_{amm} \cdot X_v + d_{gln}[GLN] \tag{2.6}$$

where $Y_{lac/glc}$ and $Y_{amm/gln}$ are the lactate yield from glucose and ammonia yield form glutamine respectively (mmol/mmol). r_{amm} is the ammonia removal rate (mmol/cell/hr). Equations 2.1 to 2.6 are the equations which describe the main mechanisms within the cell culture cultivation. These relationships can become more complex when further variables are included such as the production and consumption of all twenty amino acids. The inclusion of amino acids into cultivation models will be explored in this research.

2.3.2 Purification

In the process shown in Figure 2.4, a protein A chromatography column is used to actively bind the target protein, allowing host cell proteins, cell culture media, and viruses to flow through the column (Sofer and Hagel, 1997). After this, ion-exchange chromatography (IEX) is typically applied as a polishing step. When applied in this way IEX is used to reduce high molecular weight aggregates, charge-variants, residual DNA, host cell protein, leached protein A, and viral particles (Sofer and Hagel, 1997). It is often used as a technique for proving purity to regulatory bodies. The research carried out in this thesis is primarily concerned with IEX chromatography, which is discussed in more detail in subsequent sections.

Purification of mAbs

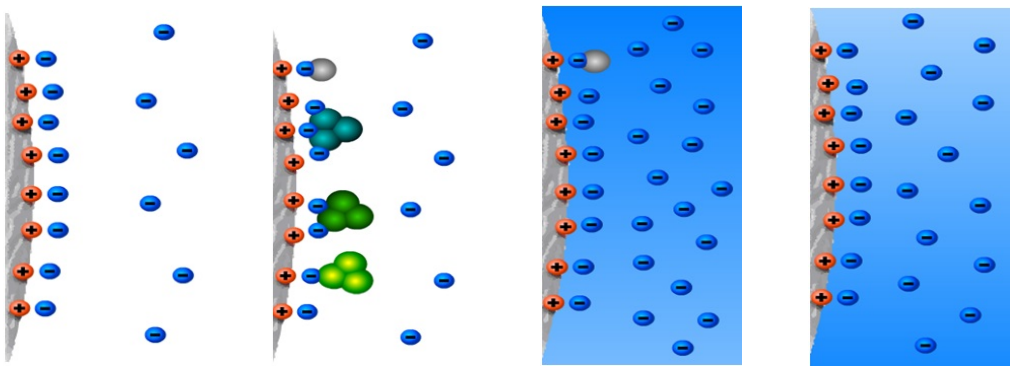
The cell line development for mammalian cell cultures over the last twenty years has significantly increased the product titre from a cultivation. However there is still a demand on industry to continue to increase final product titre, which has caused a shift in the bottleneck from cultivation to purification. The

demand for an increase in titre is an economic one, as production costs are so high the more product a process can make the more profit for the company. The purification stage of a process accounts for between 45-92 % of the total process costs (Saraswat *et al.* (2013)). One of the main purification challenges which can arise is the formation of aggregates in the process. Protein aggregation is defined as a complex, multi-stage process which involves the folding or misfolding of monomers followed by one or more assembly steps to form oligomers. As mAbs are complex molecules, understanding the aggregation process can be difficult due to the wide range of factors which can introduce stress conditions during the manufacturing process that cause aggregation. The F_c and F_{ab} regions of a mAb molecule have different properties and as such will respond differently to stress conditions (Chen *et al.*, 2010). Studies, such as that by Chen *et al.* (2010); Xu *et al.* (2012), show the successful application of IEX chromatography for the removal of aggregates.

Ion-exchange (IEX) chromatography

The basic principle of IEX chromatography is that the charged areas of the molecules in the product stream are attracted to the charged ligands on the chromatography resin. There are two types of IEX, anion and cation. In anion IEX the molecules carry a negative charge and the resin a positive, whereas in cation IEX the molecules are positively charged and the resin is negatively charged.

The five stages of IEX chromatography are shown in Figure 2.7 (GE-Healthcare, 2010). In summary the first stage is equilibration where the chromatography resin is brought to the starting state (charge and pH) that allows binding of the target molecule. The second stage is called the sample loading where the product is applied to the column. Molecules with a suitable charge will reversibly bind with the resin whilst unbound substances will wash out during the third stage (wash). The fourth stage is elution, where the conditions in the column are changed to make it unfavourable for binding. This is done through the application of a buffer of a higher ionic strength whereby the salt competes with the molecules for binding sites. This can either



(a) Equilibration (b) Sample load and Wash (c) Elution (d) Regeneration

Figure 2.7: The five main stages of ion-exchange chromatography; equilibration, wash, elution, and regeneration (GE-Healthcare, 2010).

be done gradually (gradient elution) or stepwise (step elution). The final stage includes the removal of any unbound material by application of a buffer with a high ionic strength (high salt content), and then the column is re-equilibrated to return the resin to a point where binding of target molecules can occur. To optimise the binding of molecules and to achieve the required purity various CPPs can be changed. For example to promote binding of target molecules the mobile phase buffer should be of a low conductivity (Desai, 2000). Table 2.4 summarises the main CPPs and the ideal conditions for IEX chromatography.

Table 2.4: Properties of ion-exchange (IEX) operation

Process parameter	Operation	Comments
Product concentration	The concentration of the sample loaded onto a column should not exceed the maximum binding capacity for the resin, which is specific for each resin.	When carrying out the removal of water from the product stream (as described in section 2.3) the concentration of the product should be monitored, as it can vary greatly. If the concentration is less than 1 gl^{-1} then it is recommended that the solution should be concentrated.
Viscosity	The viscosity of a sample should be determined prior too chromatography. If the viscosity is to high it can prevent diffusion of molecules and damage the resin.	If the protein and/or nucleic acid content of the solution is high then it may be quite viscous. If it is greater than 4cP is may be necessary to reduce it prior to chromatography. If the viscosity is not reduced it can increase the back pressure of the column, this in turn decreases the flow rate. The viscosity can be reduced by dilution of the feed stream.
Ionic strength (conductivity)	Elution is performed by changing the pH or ionic strength of the elution buffer. It can be difficult to control a change of pH, so a salt gradient is normally used. The most commonly used salt is NaCl.	If the solution containing the product has a high ionic strength (greater than 5 $mScm^{-1}$) at a neutral pH, there can be problems with IEX chromatography. This is a problem because the charged molecules (e.g. salt) will competitively bind to the ligands and prevent the target molecules from binding.

pH	The pH of the buffer selected for binding and elution affects the charge on weak ion exchangers but not on strong ion exchangers, this is why strong ion exchangers are used in industry.	The net charge of a molecule depends on the pH, if the pH was altered towards the isoelectric point of the substance it looses charge and therefore desorbs from the column.

To understand and subsequently model an IEX column it is easer to look at the bulk-fluid and particle phases separately. Figure A2 on page 292, illustrates the two phases, fluid and stationary. A mass balance constructed for the bulk-fluid phase is concerned with the concentration and movement of particles through the column. A stream entering the column is characterised by the feed concentration as a variable of time ($C_{fi}(t)$), however as this stream moves through the column, the concentration (C_{bi}) becomes dependant on time, axial position (distance travelled) in the column and radial diffusion. Additionally the space between particle (voidage (ε_b)) needs to also be considered as this is the flow path of the fluid. A mass balance for the bulk fluid phase can be found in the appendix as given by Gu (1995).

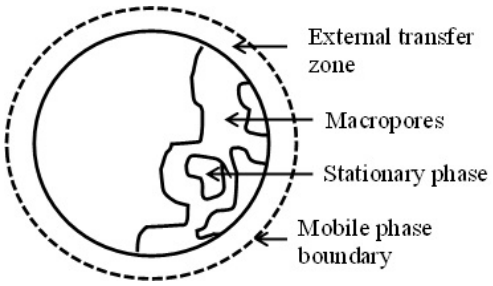


Figure 2.8: Resin bead cross sectional diagram for ion exchange chromatography, showing phase boundaries (Gu, 1995).

Figure 2.8 shows a cross section diagram of the phase boundary around a resin bead. A more detailed diagram is shown in Figure A2 on page 292. Around each resin particle there is an area which relates to the mass transfer from the bulk fluid phase to the particle phase. This is accounted for by the film mass transfer coefficient (k_i). Once a molecule has diffused through the film surrounding the bead, the diffusion into the bead is dependent upon the: molecule diameter (R), molecule concentration (C_{pi}), and internal particle voidage (ε_p). A mass balance for the particle phase can be found in the appendix as given by Gu (1995).

Chromatographic separation

The compounds which elute from the column are transported in the mobile phase to the UV detector and recorded as a Gaussian curve. The signal produced for each compound is called a peak and the trace for the whole

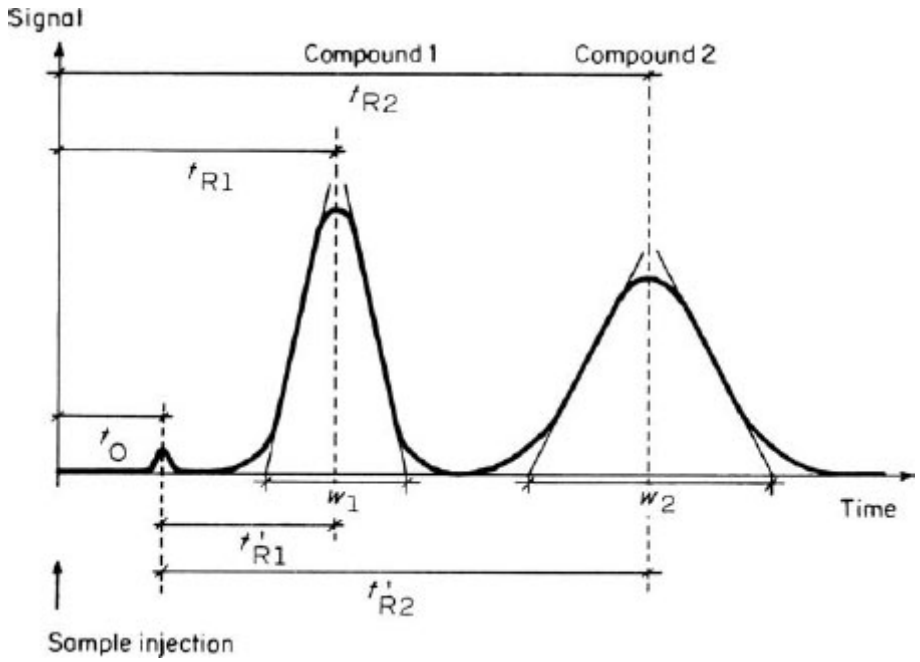


Figure 2.9: Example chromatogram showing a two component system. t_0 is the dead time, t_R is the retention time of components one and two respectively, t'_R is the net retention time, and w is the peak width (Northern Arizona Univeristy).

purification is called the chromatogram. The peaks provide information, which is both qualitative and quantitative, on the mixture being purified.

(a) Qualitative: when the chromatographic conditions are identical the retention time (t_R) of a component is constant. The retention time is the time that elapses after the sample is injected until the recording of the signal maximum by the detector. The chromatographic conditions consist of the column dimensions, the temperature, the type of resin used (stationary phase), the buffer used (mobile phase), and the flow rate.

(b) Quantitative: the peak area and peak height are both proportional to the amount of compound injected onto the column. In the case of the data presented in this research the concentration of protein is known thus the relationship with the peak area and height can be easily modelled.

Figure 2.9 shows an example of a chromatogram, where w is the peak width at the baseline, t_0 is the dead time of an un-retained compound. The dead time is the time it takes the mobile phase to pass through the column and is also referred to as the breakthrough time. The linear flow velocity (u) can be

calculated as shown in Equation 2.7 (Horvath *et al.*, 1967).

$$u = \frac{L}{t_0} \tag{2.7}$$

where L is the column length. If a compound does not bind to the column it appears on the chromatogram at t_0, additionally if the maximum binding capacity of the resin is reached there may be unbound compound of interest in the peak at t_0. t_0 is identical for all eluted substances and represents the mobile-phase residence time. The retention time is given as t_R. t'_R is the net retention time or adjusted retention time. It represents the stationary phase residence time and is different for each separated compound. The longer a compound remains in the stationary phase, the later it is eluted.

The retention time is a function of mobile phase flow velocity and column length. Meyer (2010) states that if the mobile phase is flowing slowly or if the column is long, then t_0 is large and hence so is t_R; and is therefore not suitable for characterizing a compound. In this case the retention factor (k) value is used (Equation 2.8).

$$k = \frac{t'_R}{t_0} = \frac{t_R - t_0}{t_0} \tag{2.8}$$

k is independent of the column length and mobile phase flow rate and represents the molar ratio of the compound in the stationary and the mobile phase.

2.4 Summary

In the last decade mAbs have been the fastest growing class of biotherapeutic and this trend is set to continue. It is estimated that by 2019 the market share for mAbs will be $122.6 billion (Dewan, 2015). There is a wide range of mammalian cell expression systems available including CHO, NS0, and BHK. There are a number of factors influencing which is the appropriate expression system to use; these include cost to manufacture, product yield, ability to perform post translational modifications, and complexities of the purification process. The most widely used cell expression system in industry is CHO,

accounting for 70% of recombinant protein production (Li *et al.*, 2010). This is because they are well adapted for production of proteins for use as treatments in humans due to their ability to perform post translational modifications similar to those naturally found in humans. In particular CHO cells are able to perform glycosylation and produce glycoproteins which do not produce an immune response. In addition to their ability to perform post translational modifications, CHO cells also have two well established vector expression systems (DHFR and GS) and there are platform technologies for the transfection, amplification, selection, and expansion of the cell line (Lai *et al.*, 2013). The CQAs highlighted in this chapter form the basis for the modelling work which is presented in chapter 3.

Chapter 3

Literature Review

3.1 Process development in the biopharmaceutical industry

The growth seen in the biotechnology industry over the last three decades has produced significant development in areas such as genomics, cell and protein engineering, but also in the development of the manufacturing processes. Advances include the development of large scale fermentations, optimisation of downstream processes and the development of disposable technology (Strandberg *et al.*, 1991; Wheelwright, 1991; Eibl *et al.*, 2010). Within industry the growth and discovery of new therapeutic proteins is directing the need for successful large scale production and purification to provide the economic advantages to companies as discussed by Kayser and Warzecha (2012). Furthermore there are other factors which are promoting this development of low cost faster biopharmaceutical production, namely the unsatisfied market needs, the growing competitions between companies, and the economic constraints of healthcare systems which are often turning to cheaper generic products (Gottschalk, 2003). All of these factors are placing pressure on making improvements to bioprocess development which is typically expensive and time-consuming.

Literature has shown a few of these improvements which have been

made. Farid (2007) discusses the development of scaled up bioreactors to 20,000L and Jagschies *et al.* (2006) discuss the improvements in product titre to monoclonal antibody (mAb) concentrations in excess of 5 g/L both of which have decreased the cost and time associated with bioprocess development. A recent trend has emerged which has identified downstream purification as the crucial limiting step in biopharmaceutical development (Aldington and Bonnerjea, 2007; Birch and Racher, 2006; Rito-Palomares, 2008; Farid *et al.*, 2000). It could be said that the technological advances observed in the upstream have far out stripped the advanced in the downstream, which has resulted in 50-80% of the total manufacturing cost for one biopharmaceutical product being in the purification and polishing steps (Lowe *et al.*, 2001). This means that both industry and academia are investigating new technologies which can be incorporated into platform processes to reduce costs.

It is not only the cost of the downstream processing which is an issue. In the competitive biopharmaceutical industry the time to market is also crucial. Therefore the time given to establish the process is important. Normally the process development is carried out and established during pre-clinical trials and is then subsequently scaled up, optimised, and transferred to commercial manufacturing facilities under Good Manufacturing Practise (GMP) before phase 3 clinical trial and authority inspection is implemented (Nfor *et al.*, 2009). Once the manufacturing process has been reported to a regulatory body it is very difficult and complicated to change any operation and specification later. Historically the only exception being if there is evidence to suggest that the safety, quality, or efficacy of the product are equivalent or better in the modified process (United States Food and Drug Administration (FDA), 2004a). However this has now advanced with the introduction of QbD and PAT, so that if a design space is fully characterised changes may be possible.

The current procedure for downstream process development is to investigate various conditions in a laboratory or pilot plant, then scale up to a large scale production using a general purification platform (Shukla *et al.*, 2007). This method requires large amounts of time, man power, and capital in both designing and optimising at both small and large scale. However,

ultimately it provides little understanding and few improvements to the established platform process (Low *et al.*, 2007). It is due to this that often in industry process optimisation is reliant on operator experience and often lacks a proper design approach. Therefore a systematic process design and development strategy at an early stage can greatly improve knowledge of bioprocesses and subsequently also achieve reductions in the time and capital required for manufacturing optimisation.

3.1.1 Quality considerations in development and manufacture

For therapeutic protein production it is important from a regulatory point of view to address the quality and safety of the products produced. Historically the strict quality requirements were primarily focused in manufacturing rather than the process development, and quality was controlled by specifications rather than through understanding. Now however the regulatory authorities and industry have both adopted the Quality by Design (QbD) concept which was introduced by the FDA (United States Food and Drug Administration (FDA), 2004a). QbD is a comprehensive approach to product development which includes designing and developing processes, identifying critical quality attributes, critical process parameters, and sources of variability. The aim of this approach is to improve the understanding of how the impact and interactions between process parameters can influence product quality during the process development stage (Kelley *et al.*, 2009).

As part of the FDAs initiative *'Pharmaceutical cGMPs for the 21st Century - A Risk-Based Approach'*, they also produced a further document, 'Guidance for Industry: PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance' (United States Food and Drug Administration (FDA), 2004b) designed to help with the introduction of new technologies to improve the efficiency and effectiveness of manufacturing process design, control quality and assurance (United States Food and Drug Administration (FDA), 2004a). The FDA defines Process

Analytical Technology (PAT) as being;

> ”a system for designing, analysing, and controlling manufacturing through timely measurements (i.e. during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality” (United States Food and Drug Administration (FDA), 2004b)

It is clear from this that the FDA’s PAT initiative is focused on promoting optimisation of biopharmaceutical manufacturing and quality control through companies adopting state-of-the-art methods for process control and analysis. These methods allow for the consideration of more data from the process to be able to relate this to final product quality, therefore allowing for quality to be in-built in the products.

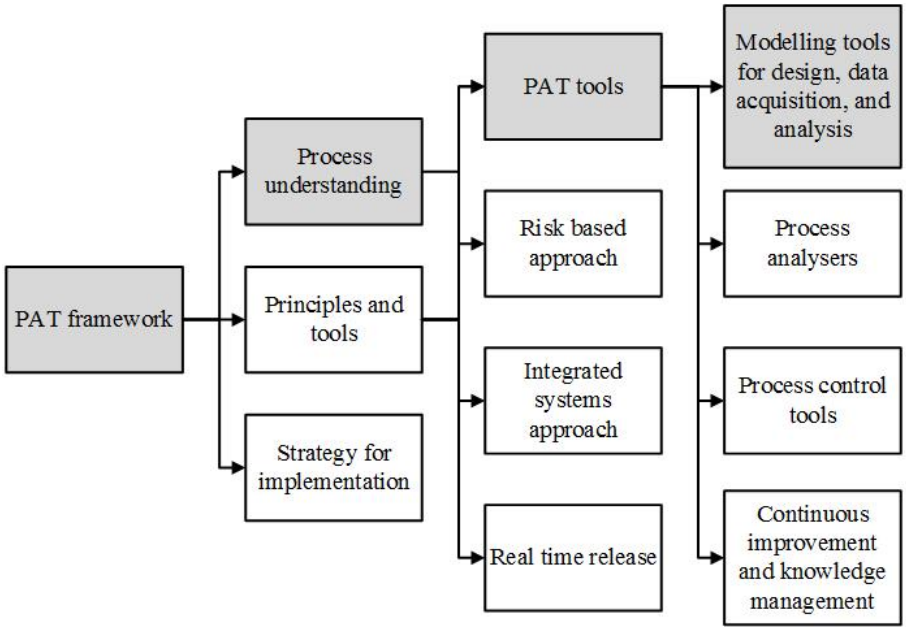


Figure 3.1: Summary of the FDA’s Process Analytical Technology (PAT) framework, highlighting where the work presented in this thesis fits within the framework.

Figure 3.1 provides a summary of the FDA’s PAT framework with the areas where this work fits in shaded in grey. It can be seen that within the framework there is a section concerning tools and a section concerning process understanding. These two areas are captured through the modelling work presented in this thesis, in that the model provides a tool for process optimisation, whilst also increasing process understanding.

In short, the biopharmaceutical industry needs a method of guaranteeing the quality, safety, and efficacy of the products as early as possible (European Medicines Agency, 2000). In order to predict, control, and analyse quality effects, it has been shown that a mathematical equation, or a model, can be highly beneficial when applied in early bioprocess development (Yu, 2008; Rathore, 2009). Nevertheless, there is currently no framework which can be applied to different bioprocesses or products. Even when the same platform process is used there are variations in cultivation and purification conditions which can vary significantly between expression systems and products. Therefore a systematic approach is needed, which can be applied to establish the relationship between critical process parameters (CPPs) and critical quality attributes (CQAs) (ICH guideline Q11, 2011). The aim in this research is to develop such a framework and models specifically for mammalian cell systems.

3.1.2 The move towards monitoring, modelling, and optimisation in the biopharmaceutical industry

Since the introduction of the FDA's QbD and PAT initiatives there have been various approaches suggested with the aim of accelerating bioprocess development. These approaches include bioprocess modelling, disposables, and high throughput technology (Carrier *et al.*, 2009; United States Food and Drug Administration (FDA), 2004a; Lee and Gilmore, 2006; Mandenius and Brundin, 2008; Milavec *et al.*, 2002; Nfor *et al.*, 2009). The focus of this research is on the development of bioprocess modelling techniques in particular.

Historically process development relied on trial and error, and one factor at a time methods i.e. empirical methods determined by intuition and experience. This relied heavily on luck being an element of success and the methods give little insight into the process (Simutis *et al.*, 1997). Up until a few years ago modelling was not considered as an important element in industrial bioprocess design as it requires a lot of effort and resources to

produce an accurate model. Recently industry has shown a greater interest in modelling (Velayudhan and Menon, 2007), with the main driving force being the requirements of the QbD and PAT incentives developed by the United States Food and Drug Administration (FDA) (2004a). These concepts and technologies use models to enhance the understanding of complicated bioprocesses in pharmaceutical manufacturing, with the long term benefit for companies being that the models can aid in the design and development of processes more quickly. They also can be used for on-line analysis, and control.

There are applications of QbD and PAT to laboratory scale fermentations of bacterial expression systems Gnoth *et al.* (2007); Carrondo *et al.* (2012); Mercier *et al.* (2013). Although these systems are not as complex as a mammalian cell expression system there are benefits in applying these approaches. Mammalian cell line development has traditionally focused on cell line and culture improvements and the PAT principles have yet to be widely adopted in the biopharmaceutical industry. However, Carrondo *et al.* (2012); Mercier *et al.* (2013). Mercier *et al.* (2013), Kirdar *et al.* (2008a), Glassey *et al.* (2011b), Teixeira *et al.* (2009a) have shown that multivariate data analysis (MVDA) can be used to extract important process information from a data set. It was shown that this reduces the complexity of the data set by eliminating co-linearity and noise (Eriksson *et al.*, 2013; Næs and Mevik, 2001). Due to these characteristics MVDA could potentially be used to identify CQAs in line with the PAT and QbD requirements. However as the biopharmaceutical industry is still in the process of adopting the PAT and QbD initiative there is no suggestion of how process/product complexity, changing cultivation environment over the cultivation trajectory, and the lack of direct signal to physiological mechanism relevance would affect the application of the PAT methods. One example in literature which focuses on the use of MVDA as a tool for PAT in the biopharmaceutical industry is by Teixeira *et al.* (2009a). In contrast to this there has recently been an increase in the literature investigating the understanding of the cellular metabolism and the impact of the changes in environmental conditions on the cell metabolism and the product quality (Ozturk and Palsson, 1991; Ozturk *et al.*, 1992; Nolan and Lee,

2011; Trummer *et al.*, 2006; Li *et al.*, 2012). This provides the motivation to develop accurate dynamic models which combine data-based models to predict process variables using process data measured during the cultivation with first principles understanding of cell metabolism and product synthesis.

3.2 Bioprocess modelling of mammalian cell systems

This thesis considers three different modelling techniques: first principles, multivariate, and hybrid. Each of these techniques has been widely reported in literature (e.g. Gadgil (2014); Huang *et al.* (2009); Psichogios and Ungar (1992)). They differ in the construction and requirements for training. First principles models use established laws of physics, with assumptions made about the system. In contrast multivariate modelling is dependent upon the collection of data. A model is then produced based upon observations of the data. Hybrid modelling combines these two techniques to utilise the best features of both. This review covers applications of all three techniques to bioprocessing with the aim of selecting the most appropriate techniques for this work.

3.2.1 First principles modelling of mammalian cell cultivation and purification

First principles modelling has been used for decades to describe the mechanisms and dynamics behind experimental observations (Tomlin and Axelrod, 2007). Within the field of biology one of the most well known models is Turing (1952) reaction-diffusion equations. This model used differential equations for morphogenesis (structure and shape) and applied it to tentacle formation of different cells. Since this work many other models have been derived which specify how concentrations of biochemical substances change with time, in and around cells (Tomlin and Axelrod, 2007). To derive

and construct these models experimental data, and knowledge or a hypothesis of the systems behaviour are required (Bailey, 1998b). First principles models have begun to find a place in industrial use in the last few decades due to the improvements in data collection. Prior to this there was not enough information on the performance and function of systems such as cells, meaning models could be hypothesised, but they could not be validated (Tomlin and Axelrod, 2007; Weis *et al.*, 2014).

Mathematical models can be classified based on the model 'architecture' and on the mathematical implementation. The term model architecture refers to whether the model is constructed 'bottom-up' or 'top-down'. The bottom-up approach uses the component parts and their interactions, with the higher level processes being determined from these parts being assembled. This approach promotes that in order to understand a system the component parts must first be understood (Milo, 2002; Hartwell *et al.*, 1999; Bhalla and Iyengar, 1999; Guido *et al.*, 2006). In contrast the top-down uses a functional model of the entire system and successively replaces each function with a model of the mechanism that implements it. As a high level system could contain many lower level mechanisms (Reid, 1985) the challenge is determining which mechanisms are involved. With both of these modelling approaches the components of the model are chosen based upon experimental data (Morris *et al.*, 2010). The mathematical implementation refers to the kind of model used to represent the dynamics of each component (Tomlin and Axelrod, 2007). Literature has shown that the most popular implementation technique is differential equation models. Differential equations can represent increase or decrease of biochemical compound concentration over time as continuously varying signals (Turing, 1952). Within these equations are parameters which include decay rates, rates of reaction, and rates of diffusion etc. values of which must be identified using experimentation.

Industrial biotechnology benefits from the use of first principles models as they can aid in understanding, predicting, and optimising the properties and behaviours of various process units (Almquist *et al.*, 2014). These benefits mean that first principles models have been used to increase yields, titre, and

productivity of a desired product (Bailey, 1998b). When modelling a biological system such as the cell metabolism of a culture during a fed batch process the response is characterised by its dependence on time. Almquist *et al.* (2014) state that to account for this time dependence the model is dependent on having accurate predictions of the rates of production and consumption each component. This is further supported by the models suggested by Kontoravdi *et al.* (2007); Naderi *et al.* (2011); Xing *et al.* (2010) who all have produced models for CHO cell metabolism and have varying methods for prediction of rates.

Developments in mathematical modelling of cell cultivations have primarily focused on two main areas: increasing understanding of cell growth and metabolism, and the prediction of cell performance in producing target products. With the development of mammalian cell culture in the 1980s this encouraged the development of mathematical expressions which could be used to monitor, predict, and optimise the culture. Furthermore the 1980s saw improvements to data generation and collection which made mathematical modelling a more viable option. It is through the work of Glacken *et al.* (1988) who began to establish models for cell growth that the more complex models can be derived.

Glacken *et al.* (1988) focused on the generation of an equation to predict cell growth rate for hybridoma cells. This equation was based on the Monod equation (Monod, 1949) but included terms for the main metabolites of hybridoma metabolism, namely glutamine, lactate, and ammonia. This work was primarily concerned with developing equations such as those which could be used in a 'bottom-up' approach to characterise a system with little experimental data. This method of characterising each aspect of the metabolism is what is referred to by Tomlin and Axelrod (2007) as an unstructured model, and from the 1980s onwards sees its use and application grow. Suzuki and Ollis (1989) took a slightly different approach, in that they hypothesised from data that the specific antibody production rate of hybridoma cells increases as the specific growth rate of cells decreases. The result is a structured model which takes into account the stages of cell division, and

related this to specific production rate, showing that the most product was produced during the G1 phase. This phase is classified as being just after mitosis during which the cells biosynthetic activity slows. G1 is thus the stage where is speeds up again.

These early models quickly established the basic equations used in almost all subsequent models, whilst also establishing the range of application. Glacken *et al.* (1988) stated that first principles models have the potential to optimise a cell culture with minimal experimentation. It is this limited experimentation aspect which has seen the popularity of this modelling technique grow. Having established both the models and the applications, the next stage in the development of first principles models for mammalian cells was the development of more sophisticated models. These include both unstructured (Bree *et al.*, 1988; Glacken *et al.*, 1989; Frame and Hu, 1991; Zeng, 1996) and structured models (Batt and Kompala, 1989; Sanderson *et al.*, 1995). Generally it is found that the structured models are more complex as they aim to characterise cell activities, which can include functions within the cell organelles. On the other hand, unstructured models view the cell as being a unit and they characterise the cell as a whole, being concerned more with the initial conditions and end result, such as metabolites and cell products (e.g. mAbs). As the unstructured models are generally simpler there has been more work conducted into the development of unstructured models as they have less computational requirements, and require less data. Jang and Barford (2000) present an unstructured model for hybridoma cells which considers not only each of the main metabolites separately but also considers both viable and non viable cells separately. This allows for a more accurate prediction of the specific antibody production rate, which is dependent on cell death. A similar approach was adopted by Xing *et al.* (2010) and Naderi *et al.* (2011), with further developments being suggested by Kontoravdi *et al.* (2007) who again produced an unstructured model but one which contained expressions for all metabolites. This increase in complexity of the models is possible due to the improvements of not only data collection methods but also of the increase in knowledge of the cell. The model presented by Kontoravdi *et al.* (2007) contains many equations for both prediction of rates and concentrations of

substances, showing how the improvements made in the predictions are at the expense of increasing computational requirements.

Although first principles models have their place as useful tools for predicting culture outcomes from small data sets, there are issues in their application. Gadgil (2015b) states that currently mathematical models for animal cell growth and metabolism cannot simulate changes to the culture pH, with this being true for all process operating conditions. It can easily be seen from literature that changing culture conditions can greatly affect the performance of a cell culture (Hwang *et al.*, 2011; Yoon *et al.*, 2005; Ahn *et al.*, 2008). It is Gerdtzen (2012) who summarises the best method currently available for overcoming this challenge in that statistical information provides a description of the system which mathematical modelling cannot, showing that for models which are both adaptive to changes in metabolism and operating conditions, a hybrid model might prove best. However, to optimise a production process it is not enough to characterise the synthesis steps, the purification steps must also be described too.

Developments in mathematical modelling of protein chromatography have focused on two main areas: intra-particle transport and interactions with the sorbent surface. The early 1990s saw new theories for intra-particle convection and diffusion. The adsorption of proteins has also been thoroughly investigated and the effects of thermodynamic aspects and protein-protein interaction investigated through the work of researchers such as Jungbauer (1996). Furthermore the design of a process was considered and aspects such as incorporating optimisation, flow rate, and column utilisation into the design stage were considered. However these advancements do not start from a blank canvas; they build upon the work carried out previously. It is through work such as that of Giddings (1960) who during the 1950s characterised the process of chromatography but did this through the method of defining the kinetics for the system.

Giddings work into kinetic processes and zone diffusion was the first attempt at characterising kinetic schemes which were representative of real chromatography processes. These include the effects of adsorption on

heterogeneous surfaces, simultaneous partition and adsorption, the adsorption of larger molecules and chemical reactions that occur which are not related to adsorption. All of these effects are important parameters to consider when modelling a column (Giddings, 1960). The use of the rate model was demonstrated by Giddings (1960), it was subsequently built upon by Gu (1995) and finally the work of Orellana *et al.* (2009). The rate model uses two partial differential equations: the first expressing the bulk fluid phase and the second for the particle phase. There is then an equation which expresses the rate of adsorption and desorption. This rate equation is chosen depending upon the system. The type of rate equation used could vary from that of second order kinetics used by Lienqueo *et al.* (2009) or using the Langmuir isotherm as shown by Shene *et al.* (2006) who show that when the dimensionless parameters are estimated correctly, it leads to a reasonably accurate prediction of the elution profile. However as hinted by Orellana the difficulty with mathematical modelling is in the estimation of the model parameters Orellana *et al.* (2009).

In other work carried out by Von Lieres and Andersson (2010) the general rate model was once again applied and a solver was created that was able to deal with complex computations and unknown parameter values. Initially work carried out by Susanto *et al.* (2008) focused on looking at a stepwise procedure which is useful when single mechanisms are considered such as the dispersion in the bulk phase. However due to current trends in separation science, where industrial mixtures containing several compounds need to be separated, this method is not applicable. Work by Von Lieres and Andersson (2010) developed the rate model to create an accurate solver to predict the concentration of protein in the column over time. Using the rate model as the basis for their model they have investigated the adsorption and desorption kinetics. From the evidence presented this method appears to have worked. However, the interesting observation is that once the main system parameters have been determined, it is the derivations which seem to be the time consuming element. Once the equations have been derived a linear solver is first explored. This linear solver considers iterative methods for solving the differential equations. It is this method of time integration that has allowed

Von Lieres and Andersson (2010) to create a simple way of elution profile prediction.

In the past few years one of the most important pieces of work that has been published is that of Shene *et al.* (2006). This explores the mathematical modelling of elution curves, however it is the first published work that draws together elements of all the previous research carried out in the twenty preceding years. This research concentrates on IEX mathematical modelling for describing a chromatographic separations classified depending on the simplifying assumptions that are considered in the derivation of the differential equations used to predict the output concentration. Shene *et al.* (2006) present an interesting yet simple solution to mathematical modelling of IEX. They targeted four key output parameters: the output concentration of desired protein, purity of desired protein, yield of desired protein and process time. They also investigated the calculation of unknown parameters which are generally determined through experimentation but the authors present a solution which allows for the calculation of these without the pre-requisite of experimentation. Furthermore their work reports the applicability of a model at different scales. This is similar to the approach taken by Li *et al.* (1998), who also showed how mathematical modelling can be applied to scale up issues. By being able to mathematically model a system and perform a few limited experiments, they have been able to predict the response of a larger column extracting the same product.

As previously discussed, the desired ability to optimise a process and control it is often achieved through a model-based approach. When considering the optimisation of a chromatographic system it is first worth considering the batch mode of operation where the elution stage is considered to be the most important step. Due to this batch operation the process is not as efficient as it could be as it would be ideal to operate a counter-current flow although this is difficult to achieve. As suggested by the work of Broughton and Gerhold (1961) a simulated moving bed (SMB) would be a working alternative. A mathematical model of this has been produced by Klatt *et al.* (2000). It is a complex model which considers two separate modules: the first

containing node balances which captures the interconnection and the switching, the second being the dynamic models. This adds another layer of complexity to the previous models investigated in literature. However this is necessary due to the set-up of an SMB system. Detail on the operation of this system can be found in Klatt *et al.* (2000). One benefit of this model is that any type of dynamic model can be incorporated into the dynamic module. Therefore the mathematical concepts in this paper may be of interest when developing a mathematical model based upon a partition design. Although this work provides some interesting ideas, they are not substantiated with evidence and would have to be developed further to prove they are valid.

3.2.2 Multivariate modelling of mammalian cell cultivation and purification

Multivariate data analysis (MVDA) and modelling is used to capture multi-linear structures in high order data sets. MVDA was first proposed by Karl Pearson in 1901 (Gorban *et al.*, 2008) and then developed by Harold Hotelling in the 1930s (Wang and Du, 2000). However it was the work of Tucker (1964) that extended common factor analysis techniques from two way data sets to higher order data sets and showed that MVDA can be used for extracting hidden structures and capturing the interactions between variables in large data sets. Referring to the United States Food and Drug Administration (FDA) (2004a) PAT framework, MVDA can be used to enhance and build quality into the process. As Glassey *et al.* (2011a) state the use of standard analytical techniques as well as the more advanced methods, such as near-infrared (NIR) spectroscopy, multi-wavelength fluorescence, and electronic nose leads to the generation of large data sets. MVDA can be used to extract useful information which results in better process understanding. Teixeira *et al.* (2009b) provide a review of various spectroscopic techniques and resulting chemometric models. They show that for data sets, such as spectra where there is a lot of information, it is necessary to first filter the data to remove redundant information, with a common technique being principal component analysis (PCA). It is also shown, how after dimension reduction is

achieved, a model is calibrated between the reduced spectra data and a target bioprocess variable with various multivariate regression methods existing for extracting process information. The most frequently used are principal component regression (PCR) and partial least squares (PLS) (Haack *et al.*, 2007; Kirdar *et al.*, 2008b; Tsang *et al.*, 2014; Rhee and Kang, 2007).

Selecting the appropriate technique to use is essential in performing MVDA correctly. As mentioned the first step in data analysis is often dimension reduction. This can be performed using PCA or PLS. PCA is a tool which can handle high dimensional, noisy, and highly correlated data (Wall *et al.*, 2003). Each data point is projected into a lower dimensional space of orthogonal components, which contain most of the variation of the original data. The orthogonal components are termed principal components (PCs) and are linear combinations of the original data. The first principal component represents the maximum variance of the original data set, with each successive PC accounting for as much of the remaining variability as possible (Teixeira *et al.*, 2009b). Other dimension reduction methods include PLS, factor analysis, projection pursuit, and independent component analysis, all of which are discussed by Fodor (2002). In the work presented in this thesis PCA is used as the dimension reduction method as it can be used to determine the important variables in the original dataset. Where PCA is used to analyse a data set, PLS can be used to obtain a regression model. These two techniques are complimentary due to the method of projecting the data into orthogonal variables (see discussion in Chapter 4), with PLS also modelling the interactions between the predicting data set and the predicted variables.

With regards to regression techniques, there are various methods reported in literature for application to pharmaceuticals. Simple regression models include multiple linear regression (MLR) and principal component regression (PCR) which correlate cause ($\mathbf{X}$) with effect ($\mathbf{Y}$) (Warnes *et al.*, 1996). However the most frequently used MVDA regression technique is PLS and its variants. The PLS algorithm projects the cause and effect data into latent variables (LVs) and then models the relationship between these new variables. Ödman *et al.* (2010) present a comparison of four different methods of

variable selection for PLS, namely: genetic algorithms, interval PLS, principal variables selection, and three-way stepwise variable elimination. They used these techniques to predict biomass and substrate concentrations, showing that the variable elimination methods resulted in the best PLS models. Issues can arise with non-linear data (Wold *et al.*, 2001b). Literature reports several methods for handling non-linear data. The first is to incorporate polynomial relationships into the PLS model (Wold *et al.*, 2001b), the second is to use ANNs (Lee *et al.*, 2006), or through the use of a hybrid structure which incorporates first principles models or mass balance equations (von Stosch *et al.*, 2011).

There are various examples in the literature of MVDA applied to mammalian cell cultivation. Recent literature has focused on the application of MVDA as a PAT solution (Mercier *et al.*, 2013; Jiang *et al.*, 2011; Teixeira *et al.*, 2009b) with Mercier *et al.* (2013) showing that PCA can be used to explore a data set, identify deviations, and consider sensitivities to scale. They also showed that data sets which are incomplete and containing gaps are difficult to use with PLS, but if the experiments are performed correctly and the MVDA is performed, it can lead to more efficient process development paths, resulting in lower development costs for new products.

As MVDA techniques can also handle multi-way data sets, they are often used to investigate batch-to-batch variation (Nomikos and MacGregor, 1995b; Albert and Kinley, 2001). This highlights one of the key points when handling multi-way data, the unfolding of the data matrix. Nomikos and MacGregor (1995b) show how 3D data may be unfolded to maintain the batch information. The 3D array is defined as $\mathbf{X}(IxJxK)$ which is unfolded into a matrix $\mathbf{X}(IxJK)$. A PLS model constructed on the unfolded matrix can be used to identify any interactions, or anomalous batches.

The application of MVDA to chromatography data can be slightly more challenging as generally data sets are of limited size and scope. If the regression model is being created for use as an on-line control or prediction tool then it is likely that one of the responses of the model will be the absorbance measurements (Laursen *et al.*, 2010b). This presents challenges in

handling the data. Zhang *et al.* (2005) state the need for a data mining system which can perform peak quantification, peak alignment, and data quality assurance. However as with other reported methods this primarily deals with data sets that contain only small batch-to-batch variations (Krishnan *et al.*, 2013b; Laursen *et al.*, 2010b).

Skov and Bro (2008) published a study which aimed to solve some of the fundamental issues with multivariate chromatography data analysis. They suggest methods of peak alignment such as correlation optimised warping (COW), with success in the application to gas chromatography mass spectrometry (GC-MS) data. Additionally they introduce the use of parallel factor analysis (PARAFAC) for the analysis of multi-way data.

3.2.3 Hybrid modelling of mammalian cell cultivation and purification

Hybrid modelling within the pharmaceutical field expanded quickly in the early 1990s, with models based on ANNs being widely reported in literature in 1992 (Johansen and Foss, 1992; Kramer *et al.*, 1992; Psichogios and Ungar, 1992). It is in this literature that the basic methodology was established, the so called parallel method (Johansen and Foss, 1992; Kramer *et al.*, 1992) with Psichogios and Ungar (1992) developing the serial approach. The main principle behind both parallel and serial methods was to use the ANN structure and to supplement these models with first principles. This produced a model, which when trained using the same process data, could produce predictions with a higher degree of accuracy. Hybrid modelling combines knowledge which is usually obtained from separate sources into one model. The advantage of this is that in combining multiple sources of information, the overall representation of the system is enhanced (Choi and Park, 2001). Hybrid modelling is also known as grey box modelling. This developed from the terms black box (e.g. multivariate) and white box (e.g. first principles). Hybrid models can be both, prior information incorporated into black box models or models where both the black box and white box elements can exist

separately (Sohlberg, 2005).

Understanding why and when to use hybrid modelling can be difficult. Standard approaches covered by black box and white box modelling techniques have different traits and thus when one cannot be applied the other generally can. Development of a white box model requires detailed knowledge of the process, with the resulting model being limited by the values used for the model parameters. Black box data driven approaches are quickly applicable and generally require less knowledge, however again experimentation is required to provide data to train the model. These black box models are based upon the underlying relationships in the data, and as such more data is required for these models that phenomenological ones. There are three possible structures for a hybrid model: one parallel (Su and McAvoy, 1993; Klimasauskas, 1998) and two serial (Martinez and Wilson, 1998) (Figure 3.2).

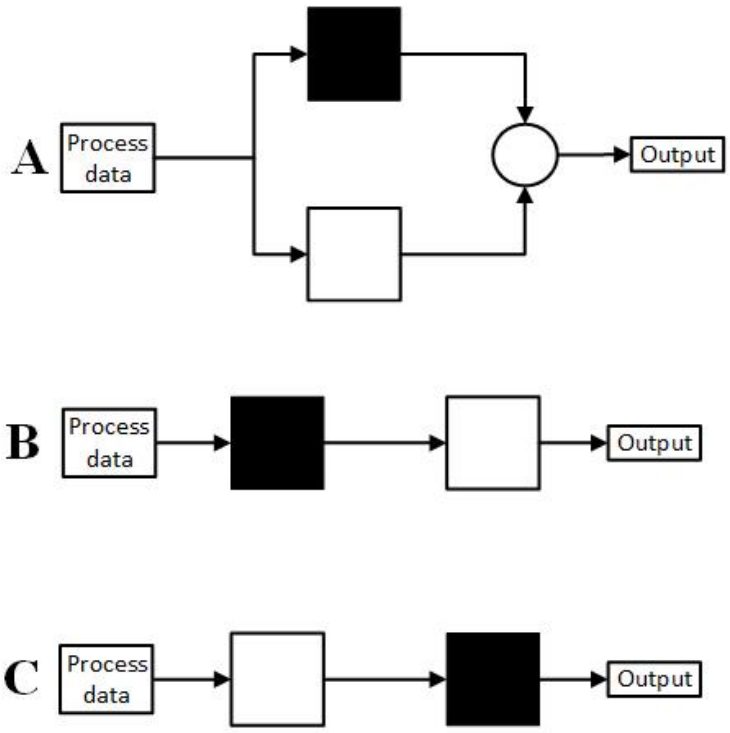


Figure 3.2: Sketch of the three ways of combining black box and white box models. A shows a parallel configuration B and C show serial configurations (von Stosch *et al.*, 2014b).

There are various instances in literature which report the parallel method (Abonyi *et al.*, 2002; Chen and Huang, 2004; Klimasauskas, 1998; Levin and Narendra, 1997; Potočnik and Grabec, 1999; Su and McAvoy, 1993). There

are various different methods which can be used to combine the outputs of the models in the structure. Two of the most commonly used methods are multiplication and superposition, as discussed by Hu *et al.* (2009). Johansen and Foss (1992) also discuss how it is possible to combine these two techniques, whereby several models can be combined using a weighting methodology to determine their contribution to the final model. This is further developed by Su and McAvoy (1993) while their presented weighing methodology following the Hammerstein model is used to determine the inclusion of the non parametric model predictions. In contrast to this the approach adopted by Fellner *et al.* (2003) focuses more on applying the weighting criteria to the mechanistic models. Having stated this however, the most commonly applied method is just the summation of the outputs (Su and McAvoy, 1993).

The serial structure (Figure 3.2 **C**) can be used as an alternative to the parallel structure (Figure 3.2 **A**), as in the serial structure C the predictions from the white box are used as inputs to the non parametric model (Aguiar and Filho, 2001; Hwang *et al.*, 2009; Schenker and Agarwal, 2000). It is only the series structure shown in Figure **C** which can be used because the parallel structure takes the predicted response from both the black and white box models and sums these based upon a weighting criteria. The structure shown in **C** essentially does the same thing but uses the response from the white box as an input to the black box model. This means that there is less of an emphasis on the response from the white box model. In contrast the model shown in **B** uses the black box model to predict the model parameters for the white box model, thus the black box and white box are not predicting the same response. Tsen *et al.* (1996) proposed an approach which was then further developed by Martinez and Wilson (1998) and uses first principles models to supplement the measured data already available and then use this expanded data matrix to train the non parametric model. Tsen *et al.* (1996) showed that in using this method predictions were obtained that were better than those using either the parallel structure (**A**) or the serial structure (**B**). The most popular serial structure used in literature is structure **B** where a non parametric model is used to predict model parameters to be used in the first principles models. Teixeira *et al.*

(2009a) and subsequently von Stosch *et al.* (2011) applied this method to large volumes of data concerned with the process operation but where very little knowledge was known about the physical properties. The non parametric models were used in both of these cases to estimate the kinetics of the system.

In determining whether to use a parallel or a serial structure the main deciding factor is the uncertainty of the white box model. Where the structural uncertainty is high, the parallel structure performs better as the non parametric model can partially account for the issues (Bhutani *et al.*, 2006; Lee *et al.*, 2002). In the case of the serial structure where the non parametric model is used to predict the parameters for the mechanistic model the extrapolation capabilities of the non parametric model are restricted by the underlying model structure and as such would not be able to predict as well (Mogk *et al.*, 2002; Fiedler and Schuppert, 2008).

There are significantly fewer examples of application of hybrid models to mammalian cell cultivation and purification compared to the literature for first principles and multivariate models. Perhaps the best examples of hybrid models developed for cell culture are those of Oliveira (2004), Anderson (2005), von Stosch *et al.* (2012), von Stosch *et al.* (2014b), and Thompson and Kramer (1994). All of these studies have focused on the development of a semi-parametric model which combines both multivariate modelling with mass balances. They showed in von Stosch *et al.* (2012) that the hybrid structure can lead to significant improvements compared to other standard modelling techniques for the prediction of cell culture biomass, protein concentration, and substrate concentration. Singhania *et al.* (2011) present a different methodology for structuring a hybrid model. The model presented in their work is concerned with cell cycle regulation, which is in contrast to the model described by von Stosch *et al.* (2012) as it is the internal process of the cell. Singhania *et al.* included Boolean variables in their model to represent activities within the cell being 'on' or 'off'. Although the model does present results which suggest it can accurately predict accumulation and degradation of proteins, the use of the Boolean variable as a 'switch' could be described as the simplest discrete system that could be incorporated into a hybrid model,

which limits its application. von Stosch *et al.* (2014a) present a case for the application of hybrid models as a PAT tool in industry. One of the main benefits is given as the flexibility of hybrid modelling. A hybrid semi-parametric model allows for the use of existing body of knowledge within a structured framework which can adapt the projected relationships and as new data in generated it can be incorporated. Additionally in this work it states the applicability of hybrid modelling to the upstream cultivation whilst further discussing the unsuitability of hybrid modelling currently to downstream units. This is an issue as the downstream units are the critical section for process development. von Stosch *et al.* (2014a) and Gao *et al.* (2009) state the need for the development of hybrid models which can be used for chromatography optimisation, product purity characterisation, and aggregation prediction.

3.2.4 Agent based modelling of cell systems

Constructing a model for a whole bioprocess is time consuming with several factors needing to be considered to achieve 'fit for purpose' models. Plant wide models which characterise every unit in a process for overall bioprocess efficiency improvements need to consider the interactions between the process units as well (Davies *et al.*, 2000; Meireles *et al.*, 2003; Zhou and Titchener-Hooker, 2003). Integrating individual models for each unit together is both time consuming and has large computational demands. Additionally when process units are investigated individually it is difficult to understand how interactions between units will affect the iterative model. Instead Sycara (1998) suggests the use of a multi agent model, which allows for the model to evolve as more information is obtained. An agent based technique is defined as a computer system which is capable of autonomous action in the environment to meet a specific objective (Wooldridge and Jennings, 1995). A multi agent system (MAS) is constructed from multiple agents that each represent a unit and work together to solve problems which are beyond the remit of any single agent (Gao *et al.*, 2009). As stated by Gao *et al.* (2009) the agents interact with one another via a computer network infrastructure shown in García-Flores and

Wang (2002); Jennings *et al.* (1995); Julka *et al.* (2002); Labrou *et al.* (1998); Wahle and Schreckenberg (2001), MAS can significantly improve the ability to model, design, and build complex systems and deal with dynamic real time data such as that in bioprocesses.

Literature shows that compared to other conventional software programs MAS has certain advantages, as summarised below.

1. MAS uses a modular approach to system development which reduces the computational demand. Additionally it is an easily adaptable system, as new agents can be included without significant modifications to the structure. (Genesereth and Ketchpel, 1994; Nwana, 1996)

2. MAS structure allows for unit agents to communicate in a network. This information exchange between agents allows for interoperation between applications. (Bradshaw *et al.*, 1997)

3. MAS can respond to time critical information. Soler *et al.* (2002) detail how changes in one unit can then be interpreted in other units in the MAS structure to create immediate responses.

4. As discussed by Sycara (1998), as the MAS structure contains several smaller structures this divides the overall problem into sub-problems. In doing so it allows for the MAS model to solve problems which would be to complex for a single centralised model to solve.

Each unit agent contains models for the particular unit, which can be developed based on first principles (Zhou and Titchener-Hooker, 2003; Boychyn *et al.*, 2004), multivariate techniques (Hiden *et al.*, 1999; Lennox *et al.*, 2001; Loukas, 2000), or hybrid modelling (Gao *et al.*, 2009). As such the development of the agent based model can be thought of as bringing together all the other modelling techniques.

3.2.5 Summary

The trend for bioprocess improvement, particularly microbial culture, over the last 15 years has focused on improvements to cell lines, and improvements in upstream technology. The literature has shown how bioprocess modelling can be applied as a successful tool for implementing the PAT initiative, which could be used to lead the way in bioprocess development in the future. The implementation of a systematic process design and development strategy at early stage development can greatly improve the knowledge of the process and subsequently can reduce the time and capital required for optimisation. The use of bioprocess modelling in this thesis expands on the work of Gnoth *et al.* (2007); Carrondo *et al.* (2012); Mercier *et al.* (2013) by exploring the application of QbD and PAT to laboratory scale fermentations of mammalian cell systems. As Teixeira *et al.* (2009a) has shown modelling can enhance and extract important process information allowing for the easy identification of CQAs in line with the PAT and QbD principles.

The literature has shown that first principles and multivariate modelling have two very different approaches and outcomes. First principles models have been shown to be beneficial in early application during the early stages of process development (Milo, 2002; Guido *et al.*, 2006; Almquist *et al.*, 2014). They generally require less experimental data to predict the variables within the models (Morris *et al.*, 2010), however their range of application is limited as they cannot be used to accurately predict how changes in the operating procedure effect the process unit (Gadgil, 2014). In contrast multivariate modelling techniques have been shown to be useful in the optimisation of process units (Teixeira *et al.*, 2009b) as they can be used to extract useful information from datasets with multivariate models generally requiring significant amounts of experimentation. Multivariate modelling can be used to optimise various process conditions, but their application is limited to the operating conditions which were varied in the process.

Hybrid modelling combines the best aspects of first principles and multivariate modelling. A hybrid model can often be used as both a predictive

and an optimisation tool. The application is dependent on the mathematical expression used and the data collected. However in general for cell cultivations it can be said that the multivariate aspect captures the dynamics of the system i.e. the rates of consumption and production (Le *et al.*, 2012), whilst the mathematical model captures the metabolism of the cell i.e. the interaction between the cell and the metabolites (Kontoravdi *et al.*, 2007).

An agent based system brings these three modelling techniques together producing a single agent for each process unit, which is then co-ordinated in a multi agent structure (MAS) (Gao *et al.*, 2009). The MAS structure brings together all the individual agents to solve problems which are beyond the remit of each individual unit. This means that MAS can significantly improve the ability to model, design, and build complex systems and handle dynamic real time data (Julka *et al.*, 2002; Wahle and Schreckenberg, 2001).

Chapter 4

Methodologies and protocols

Chapter 3 covered the literature associated with modelling of bioprocesses. It was shown that there are many different techniques which could be applied to model mammalian cell manufacture. This chapter introduces the methods used for data acquisition, along with the modelling methodology.

Mathematical and multivariate statistical techniques have been successfully applied to provide calibration, validation, and optimisation of biological processes. The application of multivariate techniques involves data pre-processing and data analysis (may include data mining, pattern extraction, identification of system). The raw process data used in the analysis generally has two issues: the quality of the data, and the applicability of the data. Pre-processing can often address the issue of data quality, whereas applicability of data is often only highlighted after the multivariate techniques have been applied. There are several decomposition methods for handling multi-way data, with multi-way principal component analysis (PCA), Tucker, and parallel factor analysis (PARAFAC) most frequently reported in the literature. All three of these techniques decompose the 3-D array into sets of scores and loadings which describe the data in a more condensed form than the original. Bro *et al.* (1999) shows that PARAFAC is a constrained version of Tucker, and Tucker is a constrained version of PCA, therefore if a data set can be modelled with PARAFAC it can be modelled by both Tucker and PCA methods. PARAFAC can be thought of as the simplest and PCA the most

complex. This chapter contains the methodologies and procedures used in this research. The techniques used are described along with reasoning behind the application.

4.1 Experimental procedure

The data used in this research was obtained from many sources including academic and industrial collaborations, and experiments. This section provides an account of the different sources of the data used and where, generated through experiments, the procedures used. For the experiments carried out by third parties, the experimental details are provided entirely to the extent provided by the third party.

4.1.1 Cultivation

Hybridoma cell line

The cultivation data in chapter 5 work was carried out in collaboration with a PhD student from ETH Zurich. The cultivation experimentation was conducted by the collaborators, details can be found in Ivarsson *et al.* (2014). In summary the experiments were conducted using a hybridoma cell line (ATCC CRL-1606) and adapted to chemically defined culture media (TurboDoma TP6, Cell Culture Technologies). The cells were cultivated in controlled parallel 1L bioreactors (DasGip) in batch mode. The culture conditions are reported in Ivarsson *et al.* (2014). Briefly, the culture environment was controlled at 37 °C, dissolved oxygen (DO) was set to 50 % air saturation and controlled by a constant gas inlet flow rate of 0.05 vvm (volume of air per unit of medium per minute), pH was controlled at 7.2 by CO_2 sparging, stirrer speed was set to 150 rpm (revolutions per minute), and osmolality was 320 $mOsm/kg^{-1}$. A parameter shift of one of the selected process variables was performed in the early exponential growth phase as described by Ivarsson *et al.* (2014) together with the sampling procedure and

Table 4.1: Experimental errors for cell measurements obtained Ivarsson *et al.* (2014).

Variable	Measurement error
Viable cell count (cells/mL)	± 0.2
Product titre (mg/ml)	± 7.2
Glycosylation (Peak area %)	± 2

analytical methods. Viable cell concentration, glucose, lactate, and ammonia concentrations, amino acid concentration, and mAb concentration were measured twice a day as off line data. The glycosylation profile was recorded at the end of the cultivation. The experimental errors for the measurements are provided in Table 4.1 and obtained from Ivarsson *et al.* (2014).

Chinese hamster ovary cell line

The cultivation data used in chapter 7 was provided from the industrial collaborators Fujifilm Diosynth Biotechnologies. This data was generated during the development of their Apollo™ mammalian expression platform. Further information on the Apollo expression platform can be found on the Apollo web page (Fujifilm Diosynth Biotechnologies, 2014). In summary the experiments were conducted using Chinese Hamster Ovary (CHO) cell line producing anti-CD20 mAb. The cells were cultivated in shake flasks, micro bioreactors (Ambr) (15ml capacity), and controlled bioreactors (2L capacity) in fed-batch mode. The culture conditions for the micro bioreactors and the 2L bioreactors are provided in Table 4.2. Data from ten shake flask cultivations was available, all ten cultivations/cell lines were taken forward and cultured in the micro bioreactors. From these ten cultivations the four cultivations with the highest yield were then cultivated in the 2L bioreactors. Measurements were collected throughout the cultivations including viable cell count, concentrations of metabolites, and the glycosylation profile of the final product was measured. Due to the confidentiality requirements detailed information cannot be provided on the techniques used to take the measurements. However

this work is concerned with the analysis of the data not the data collection, and so this should not impact on the conclusions within this research. The experimental error for viable cell count, product titre, and glycosylation peak measurements are provided in Table 4.1. The same errors were used for the measurements recorded for both hybridoma and CHO data sets. Ad the measurement error is associated with the technique for recording the data.

Table 4.2: Culture conditions for cultivations operated using the micro-bioreactors and the 2L bioreactors.

Micro-bioreactors		
Parameter	**Set point**	**Control limits**
pH	7.00	+/- 0.05
Temperature	36.5 °C	+/- 0.2
Agitation	1152 rpm	+/- 5.0
DO	30.00%	+/- 10.0
2L bioreactor		
Parameter	**Set-point**	**Control limits**
pH	7.00	+/- 0.1
Temperature	36.5 °C	+/- 0.5
Agitation	223 rpm	+/- 5.0
DO	30.00%	+/- 10.0

Also provided with the CHO cell cultivation was a data set for the downstream purification of a mAb produced from a CHO cell line. This data was not the purification data directly related to the experiments conducted for the cell cultivation, however, the sponsor company provided information supporting the fact that the mAb produced and the cell line were very similar. Hence for this research they were used as one data set. The downstream process units included:

- Cell harvest
- Protein A chromatography
- Virus removal
- Protein A chromatography
- IEX chromatography
- Viral filtrate
- Bulk drug substance (BDS)

Table 4.3: Summary of the analytical techniques used for the analysis of the downstream process units.

Measurement key	Full name	Description
UV A280	Ultraviolet absorbance at 280nm	Measured using a spectrophotmeter, used to detect protein (A280nm). Is a measure of the product titre at each stage.
Pro A	Protein A chromatography	Ligand binds to the mAb, which is then eluted and detect by HPLC. Is a measure of the product titre at each stage.
CIEX	Cation exchange chromatography	Different aggregates of the protein have different charges. CIEX is used as a measure of the heterogeneity of the protein.
SEC	Size exclusion chromatography	Size exclusion is used to determine the molecular weight of the proteins produced.
ELIZA	Enzyme-linked immunosorbent assay	The ELIZA test is used to assess the quantity of host cell protein (HCP). This is an impurity and there is a maximum allowance for the final BDS.
CE-SDS	Capillary electrophoresis gel	Gel electrophoresis is used to determine if the protein is glycosylated or not. In this application the distinct glycans are not determined.

with several analytical measurements being recorded during each of these stages. These are summarised in Table 4.3.

4.1.2 IEX chromatography

The following methodology is the experimental procedure for obtaining the lactoferrin data. This is the data used in chapter 6, the experiments were conducted at Fujifilm Diosynth Biotechnologies' Billingham site by the author of this thesis.

Materials and equipment

Recombinant human lactoferrin produced from *Aspergillus niger* var. *awamori*, a food grade organism, was used in these experiments and obtained form stock solutions at the sponsor company. The lactoferrin was produced via

a fermentation and subsequently stored in a buffer solution (15 mM sodium phosphate, 50 mM NaCl, pH 7). Lactoferrin has a molecular weight of 78,000 Daltons (78 kDa) and it is a glycosylated metalloprotein (this is a protein which contains a metal ion). The data used in chapter 6 is concerned with the purification of the lactoferrin via IEX chromatography performed using an Äkta explorer.

The IEX was performed using HiTrap™ SP (Sulphopropyl) Sepharose Fast Flow (SPFF) columns. These columns are cation exchangers which are acidic and contain negatively charged groups, termed ligands. In the case of SPFF the ligand is a straight chain called Sulphopropyl, a strong cation exchanger. The term *strong cation exchanger* refers to the pK_a value of the ligand, which for Sulphopropyl is 2-2.5. The ligands are attached to a matrix which acts as a support. In SPFF the matrix is a 6 % cross-linked agarose. A property of this matrix is that it is hydrophilic, which is common for resins used with proteins, as the interactions with the protein are weak. The columns come pre-packed from the supplier (GE Healthcare). Pre-packed columns were chosen to eliminate the variability which could have been introduced if hand packed columns had been employed. The HiTrap columns have a bed volume of 1 ml, a bed height of 25 mm, and an internal diameter of 7 mm. The average particle size of the beads is 90 μm, the range of the particle size is 45 - 165 μm. The ionic capacity of the column is 0.18-0.25 mmol H^+/ml of medium. This refers to the concentration of ligands and has been calculated when the resin is in liquid form (Jansen, 2012).

Figure 4.1 shows a schematic of an Äkta explorer (GE Healthcare), with the following basic operation. A pump is used to apply the equilibration buffer to the column, which is packed with the chromatographic medium. The sample containing the product is applied to the column using a sample pump and it is allowed to interact with the chromatographic medium. The elution is carried out using the elution buffer (containing NaCl) and the components are separated in order of increasing affinity to the chromatographic medium. The solution leaving the column is monitored to detect the composition of the stream and fractions are collected and transferred to the next step in the

purification scheme. Having removed all the target molecules the column is stripped to remove any remaining bound molecules, and the column is recharged for the next run.

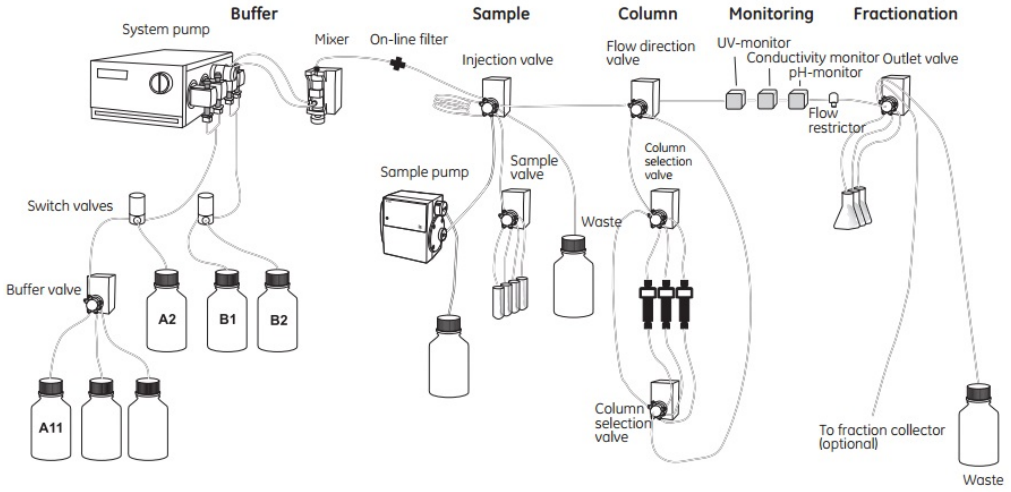


Figure 4.1: Schematic of Äkta Explorer used in the experiments conducted in this research (GE-Healthcare, 2010)

The buffers used to perform the chromatography step are summarised in table 4.4. The first buffer (A) is a phosphate buffer (NaH_2PO_4/ Na_2HPO_4) and has a low ionic strength (see Table 4.4). The second buffer (B) is again a phosphate buffer, but it has a high ionic strength due to the addition of salt (1M NaCl). The final buffer (C) is a 0.1 M solution of NaOH, and is used to store the column. In relation to the diagram shown in Figure 4.1, buffer A is placed on the line A11, buffer B is placed on B1, and buffer C is placed on A2. This set up is used so that when two buffers are required (A and B) for the gradient elution it does not require the system to switch lines on one pump.

Table 4.4: Composition and concentration of the three buffers used to perform the chromatography step

Buffer	Line	NaH_2PO_4/ Na_2HPO_4 (mM)	NaCl (M)	NaOH (M)	pH
A	A11	25	-	-	7.5
B	B1	25	1	-	7.5
C	A2	-	-	0.1	7.5

The protein sample is diluted using buffer A to concentrations between 10-30 mg/ml, and is injected into the system at the injection valve. The sample is washed over the column in a downwards motion and no sample emerges in

the effluent until on column dead volume has passed (Desai, 2000). This is the interstatial volume of the resin and is approximately 30-45%, in this instance 0.3-0.45 ml. The column effluent is then passed through different monitoring instruments and is sent to either waste or is collected. The purification scheme for lactoferrin, which was designed at Fujifilm, is shown in Table 4.5. This forms the basis for the design of experiments (DoE) used to generate the data set for further analysis described in Chapter 6.

Table 4.5: In house methodology used for IEX lactoferrin purification with SPFF resin. (* pH not specified in method ** Depends upon the sample size being loaded)

Chromatography stage	Phosphate (mM)	NaCl (M)	NaOH (M)	pH	CV	Flowrate (cm/hr)
Clean	-	-	0.1	*	2	75
Charge	25	1	-	7.5	2	75
Equilibration	25	-	-	7.5	2	75
Sample load	25	-	-	7.5	**	40
Wash (low flow)	25	-	-	7.5	1	40
Wash (high flow)	25	-	-	7.5	2	75
Elution (gradient)	25	(0-1)	-	7.5	10	75
Strip	25	1	-	7.5	1	75
Clean & store	-	-	0.1	*	2	40

4.1.3 Design of experiments (DoE)

Design of Experiments (DoE) is defined as a structured and systematic method of designing an experimental set. DoE aims to generate the most informative results whilst minimising runs, time, and resources (Mercier *et al.*, 2014). The DoE methodology was used to determine the conditions of the experimental runs in the lactoferrin data set. The factors which were chosen to be investigated were: flow rate of buffer (ml/min), pH of load buffer, pH of elution buffer, number of column volumes gradient was performed over, and concentration of protein in loaded sample (mg/ml). A minimum resolution IV design was used where the number of experimental runs is determined by Equation 4.1.

$$\text{Number of runs} = 2k + 2 \tag{4.1}$$

where k is the number of factors being investigated (Jones and Montgomery, 2010). Therefore, as 5 factors are being investigated in this experimental data set the number of runs would be 12. The minimum resolution IV design aims to estimate all the main effects independently without alias, whilst using the minimum runs. If however one or more runs were missing this would cause the design to become a resolution III design, so two extra runs are added to protect for missing data. Furthermore three centre points were added to the design. These were used to check reproducibility in the experimentation and to check for curvature in the response of a factor. So in total for the lactoferrin data set there are 15 experimental runs. The conditions of each of these runs is provided in Table B1.

4.2 Data treatment

Having performed the experiments and obtained the data as described in section 4.1 the next step is to put the data into a usable form. Data is required in this research as a tool for both training and validating models. The scope of application, prediction quality, and adaptability of the models depends upon the quantity and quality of the data.

The data for both cultivations and chromatography is multi-way, meaning the data sets have multiple samples, variables, and batches. Therefore the data is assembled in a 3D array. This presents various challenges with manipulating the data. In order to analyse all the batches simultaneously they must be time aligned. This is achieved by various techniques for example cutting the data, time shifting, interpolation. Having aligned the data other transforms can be applied. These transforms are subject to the analysis or modelling technique being used and are specific to each data set.

4.2.1 Array unfolding

Unfolding a 3 dimensional array reduces the data set down to 2 dimensions without loosing any information. The 3 dimensional data set is as shown in

figure 4.2 (a); with i batches, for measurements of j different variables available over k time points giving a three-dimensional data array $\underline{\mathbf{X}}$ of size i × j × k. There are two ways to unfold this data; the first is *batch-wise data matrix unfolding* (Nomikos and MacGregor, 1994; Nomikos and MacGregor, 1995a). The array $\underline{\mathbf{X}}$ is divided into k slices of size ixj, which are then placed side by side. This technique produces an unfolded matrix **X** of size i x jk. Using this method preserves the batch direction as every row of the unfolded matrix corresponds to a complete batch. This unfolding technique is demonstrated in Figure 4.2. The other technique for unfolding the data is *variable wise data matrix unfolding* (Wold *et al.*, 1987). In this technique the matrix **X** is divided into j slices of size ixk, when these slices are placed side by side the unfolded matrix is of size j x ik. This method preserves the variable direction. Generally for batch processes, such as cultivations, batch-wise data matrix unfolding is the more popular, since the batch-end quality is related to the complete batch history. In this way batch-wise data matrix unfolding can be used for the prediction of final product quality. The matrix unfolding in this research was performed using the batch-wise methodology.

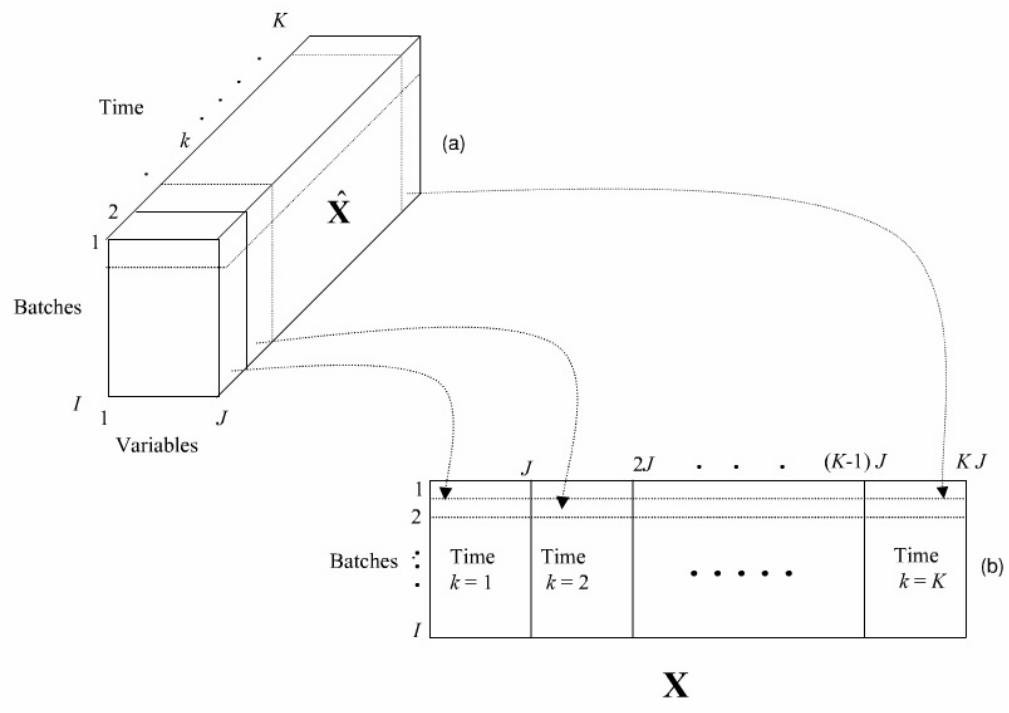
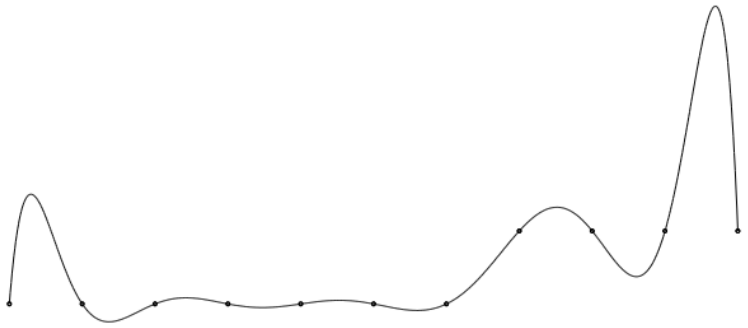


Figure 4.2: (a) Structure of a three-way data array describing input (predictor) variable measurements from a batch process; (b) unfolding of array into a large two-dimensional matrix (Nomikos and MacGregor, 1994, p. 100).

4.2.2 Cubic spline

To compare between experiments where samples have been recorded at different times a spline was applied. This is a form of interpolation whereby a polynomial function is fitted between the available data points to predict the values between measured data points. This research uses a cubic smoothing spline which is the most common form. This method was chosen as it accounts for the experimental error, allowing for the cubic curve to be placed within a specified range of measured values (Wahba, 1978).



(a) Lagrange interpolation of data points

(b) Spline interpolation of data points

Figure 4.3: (a) The increase in height of the data points in the middle causes an effect on the interpolating polynomial curve at the ends; (b) 'cubic smoothing spline' curve through the same data points, the jump in data point height does not adversely affect the fit of the polynomial. (Baker, 2014)

Figure 4.3 illustrates the difference between another commonly used form of interpolation, Lagrange interpolation (Berrut and Trefethen, 2004). The main difference can be seen in that the Lagrange interpolation uses one polynomial which can be affected by sudden changes in the data. Whereas the cubic smoothing spline uses a separate polynomial fitted between each data point, therefore the fitted curve follows the data points more closely. More information on the mathematics of spline functions can be found in the work of Wold (1974). This work uses a cubic smoothing spline function developed for MATLAB® by Henning.

4.3 Multivariate data analysis (MVDA)

To analyse and model data such as that used in this research, appropriate methods need to be used. There are various techniques which can be employed and have been summarised in chapter 3. This research explores using Parallel Factor Analysis (PARAFAC), Principal component analysis (PCA), and regression modelling using Partial Least Squares (PLS). These three techniques are appropriate because they can provide an overview of batch variation (PARAFAC) (Bro, 1997), the interactions between variables (PCA) (Wold *et al.*, 1987), and data modelling (PLS) (Wold *et al.*, 2001a).

There are other techniques which would have been equally applicable for use in this research. One common alternative when using PLS regression is to analyse the scores and loadings from the PLS model instead of performing a separate PCA analysis (Westerhuis *et al.*, 1998; Nelson *et al.*, 1996). The main reason for first applying PCA to analyse the data set is due to its size. As Kettaneh *et al.* (2005) discuss for large data sets criteria such as applicability of PCA and PLS to the larger scale, more difficulty in handling noise and non-linearities, interpretability of results, and simplicity of use are issues which need to be addressed. They showed that keeping the analysis simple made it easier to handle these challenges and interpret the results. As PCA considers the interaction between the **X**-block variables, whilst PLS considers the interactions between both the **X**-block and **Y**-block variables this makes PLS a more complex technique. The interpretation, especially for the loading values, for PLS becomes much more difficult the larger the data set. Therefore in this research PCA is applied as a pre-cursor to PLS, to determine the appropriate variables for use in the regression model.

Additionally there are also other regression techniques which could have been used. These include other variants of PLS, such as non-linear iterative partial least squares (NLPALS), orthogonal partial least squares (O-PLS), sparse partial least squares (SPLS), or other multivariate techniques such as Canonical Correlation Analysis (CCA), Tucker regression to name a few. The main choice was between the use of a linear or non-linear technique.

Relationships and measurements obtained from biological systems are not generally linear, this would naturally lead to the use of a non-linear technique for prediction (Jong, 1993). However, the data used in this research is obtained from industry and as such is not perfect. There are frequently missing values, through either instrument error or because a sample was not taken by the operator. This combined with the need for the end model to be used by an operator with limited in depth knowledge of modelling techniques means that it is necessary to have a regression model which is both easy to train and apply (Abdi, 2010). In this way linear PLS is the better option as a fluctuation in one reading of one batch will not have a significant impact on the end model. Furthermore it is also worth noting that PLS has fewer computational requirements and the results are easier to interpret, which again is a benefit for application by a process operator (Wold *et al.*, 1987; Martens, 2001).

4.3.1 Pre-processing

There are experimental and instrumental effects which are not related to the process data which can impact the profile of samples. For example, sample collection, sample preparation, and instrumental calibration. The data needs to be appropriately pre-treated prior to analysis as the type and extent of pre-treatment can greatly impact the final results. A good pre-treatment procedure enhances the process information, whilst an inappropriate pre-treatment procedure can impact on correlation (Rajalahti and Kvalheim, 2011a). Effective pre-processing is determined by how well the user knows the data, as too much pre-processing to remove noise may instead remove information.

The pre-processing steps which are carried out on the data set prior to PARAFAC are less complex than those used prior to PCA. PARAFAC is able to account for variations in batch length, whereas PCA is not. Therefore prior to PCA the data matrix was cut to ensure uniform length of batches. Furthermore, due to the nature of some of the measurements, particularly on-line measurements, it was necessary to perform data set sampling, whereby every 10th data point is included. This reduces the data set down to a size

which is more appropriate for the computational software.

As with all multivariate modelling applications it is the measurements that were taken which direct the preprocessing required. For example for cell cultivation where measurements include pH, DO, temperature etc the preprocessing steps required would be different to an application such as ion exchange chromatography which includes measurements for absorbance. Rinnan *et al.* (2009) provides a comparison of the most commonly used pre-processing techniques, although this study is for the application to Near Infra Red spectroscopy (NIR) it discusses the application of the various techniques. Considering next the literature for cell cultivation data there are various studies which discuss the different preprocessing techniques used (Selvarasu *et al.*, 2010; Dudoit *et al.*, 2002; Heyer *et al.*, 1999; Selvarasu *et al.*, 2012). These studies have shown that autoscale should be applied. This technique is a combination of mean centering and scaling the data. The application of autoscale means that any resulting analysis or regression model is a result of the variation between the samples instead of the absolute level of variation. Each variable X_i is mean centred by Equation 4.2.

$$X_{ij} = X_{ij} - \bar{X}_i \tag{4.2}$$

The literature has also shown that for applications to data such as that collected from ion exchange chromatography (IEX) the preprocessing methodology would be different. There is no study present in the literature which provides a direct comparison of different techniques for the application to IEX, hence part of this research does this (see Chapter 6, Table 6.5). However, there are similarities between IEX, spectroscopy, and liquid chromatography-mass spectrometry (LC-MS). Therefore, applications of preprocessing to these areas formed the basis of the IEX study (Rinnan *et al.*, 2009; Gromski *et al.*, 2014; Laxalde *et al.*, 2011; Wang and Kowalski, 1992; Skov and Bro, 2007; Luypaert *et al.*, 2004).

4.3.2 Parallel factor analysis (PARAFAC)

Parallel factor analysis (PARAFAC) is a tool to analyse multi-way data. A 3-D data array is used directly in the analysis with the main benefit that the models produced are simpler and therefore easier to interpret (Bro, 1997). PARAFAC decomposes the data array into new dimensions which capture the variability between batches. These new dimensions are captured in components called modes, comprised of scores and loadings, which describe the data in a condensed form. Each mode in a model consists of one score and two loadings vectors. In PARAFAC it is common that no difference is made between the scores and the loadings, and they are treated exactly the same way. The decomposition of a 3 way array is shown in Figure 4.4.

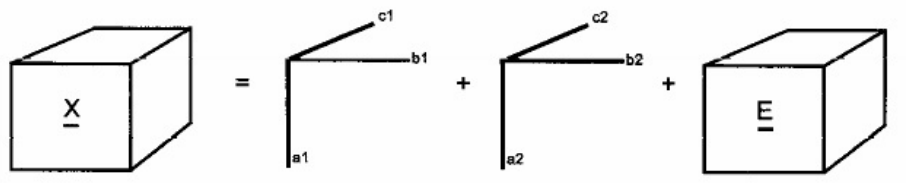


Figure 4.4: Pictorial representation of the decomposition of the array X into a two component PARAFAC model (Bro, 1997).

The array $\underline{\mathbf{X}}$ has 3 dimensions, i (time), j (variables), and k (batches). Each subsequent component produced from the PARAFAC analysis has three scores/loadings matrices, with each one corresponding to one of the original dimensions. This can be represented by the Equation 4.3.

$$\underline{\mathbf{X}}_{ijk} = \sum_{f=1}^{F} a_i b_j c_k + e_{ijk} \tag{4.3}$$

where a_i is the loading in the ith direction, b_j is the loading in the jth direction, c_k is the loading in the kth direction, and e_{ijk} is the residual error, and F is the number of components/modes in the model. The PARAFAC analysis was performed using the n-way toolbox as developed by Andersen and Bro (2000).

4.3.3 Principal component analysis (PCA)

Principal component analysis (PCA) compresses data, extracts information, and visualises observations. PCA reduces data sets with a high-dimensionality to a few dimensions which capture the main variability in the data. These new dimensions are defined in terms of principal components (PCs) that are linear combinations of the original variables. The contribution of the original variables in the PCs are called the loadings. These can be used to identify important variables in the PCs and also to show how the original variables relate to each other. The scores are the coordinates of the original data in the new space, and they show the sample/batches/experimental runs relate to each other.

A principal component is obtained by visualising the data in a matrix, **X**, the first principal component is then determined as being the 'line of best fit' which captures the maximum variation. The second principal component is a line of best fit which captures the second greatest amount of variability in the data set, and so on. The model aims to reduce the sum of squared residuals which are a measure of the discrepancy between the data and an estimation model. Eventually the principal components will reach a stage where they are no longer capturing process variability but instead are representative of noise in the data. Determining the correct amount of PCs for a model is subjective. The most simple approach would be to choose the number of PCs for the variance to of achieve a predetermined percentage, i.e. 90% (Qin and Dunia, 2000). Another method was proposed by Wold (1978b) who used the predicted error sum of squares (PRESS) which is calculated by randomly leaving out samples. This technique is quite time consuming as it requires multiple PCA models to be constructed in order to calculate the PRESS. There are other less common techniques such as that proposed by Joreskog *et al.* (1976) in which each PC must contribute at least one mth of the total variance, where m is the number of variables. Another alternative is the technique proposed by Cattell (1966) which is a scree test based on a plot of the eigenvalues of the correlation matrix. The number of PCs in the model is selected at the point which the graph drops sharply followed by a straight line with a much smaller slope. It is

this technique of using the scree plot which is applied in this research.

Figure 4.5 shows an example of a projection for a data matrix which contains two variables (x_1 and x_2), The vector shown in Figure 4.5 is the line of best fit for the data points, as it captures the maximum variance. The scores are the values of the data points projected onto the vector as shown in Figure 4.5 (a). The loadings, shown in Figure 4.5 (b), are the direction cosines of the vector, they are calculated for each variable as the distance from the variable axis to the vector. Equation 4.4 shows the formula for the scores and loadings of each PC.

$$\mathbf{X} = t_1 p_1' + t_2 p_2' + ... + t_n p_n' \tag{4.4}$$

where t_n is the scores vector and p_n is the loadings vector for the n^{th} PC. For the overall model, these vectors can be written as a matrix which contains all the scores and loadings for all the PCs in the model (Equation 4.5).

$$X = TP^T + E \tag{4.5}$$

where T represents the scores and P the loadings matrices. E is the residual errors matrix, and is not a part of the model. It is the part of the original $\mathbf{X}$ matrix which is not explained by the model (TP'). This value should be small, as a large value would suggest too much information has been removed. When constructing a PCA analysis the relationship can be described in Equations 4.6 and 4.7.

$$PC_0(\text{explained variance}) = 0\% \longrightarrow E_0(\text{residual variance}) = 100\% \tag{4.6}$$

$$PC_{max}(\text{explained variance}) = 100\% \longrightarrow E_{max}(\text{residual variance}) = 0\% \tag{4.7}$$

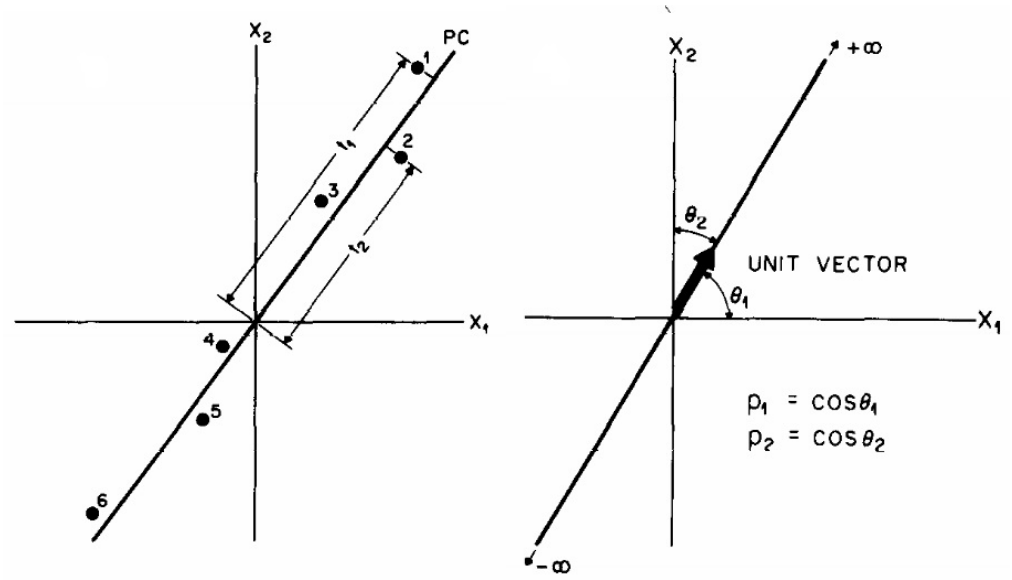


Figure 4.5: Determination of principal component for two variables (x_1 and x_2). The PC vector is shown to capture maximum variability in the data. (a) the scores projection onto the PC vector (b) the loading determination, calculated as the angle between each variable axis too the PC vector. (Geladi and Kowalski, 1986, p. 6)

Bi-plot

The analysis of the scores and loadings is carried out using a bi-plot, constructed using the bi-plot function implemented with Matlab. Bi-plots were first introduced by Gabriel (1971) as a method of graphically representing the batch and variable data on one plot. When applied in PCA the axes are principal components and in PLS they are latent variables. Bi-plots are obtained by using the singular value decomposition (SVD) to obtain a low-rank approximation of the data. Carlier and Kroonenberg (1996) provide information on the algorithm for producing bi-plots. In this research, however, it is the interpretation of the graphs that is important.

A bi-plot uses points to represent the scores of the observations on the principal components, and it uses vectors to represent the coefficients of the variables on the principal components. The location of the points can be interpreted such that those which are close together correspond to observations which have similar scores of the principal components displayed in the plot.

Additionally this corresponds to a similarity in the original observations for these batches. The vectors represent the variables, with both the length and direction of the vectors being important. The vector points in the direction which most closely resembles the variable represented by the vector, as this is the direction which has the highest squared multiple correlation with the principal components. Furthermore, the length of the vector is proportional to the squared multiple correlation between the fitted values for the variable and the variable itself. Vectors which point in the same direction correspond to variables which have similar response profiles. In the case of on-line analysis of cell cultivation data, two vectors which point in the same direction represents two variables which have a similar influence over the variation in the data.

4.3.4 Partial least squares (PLS)

Partial least squares (PLS) is a regression technique which can be applied to noisy and correlated data. Similarly to PCA, PLS reduces the dimensionality of a data set. However where PCA aimed to capture the maximum variance in one data set, PLS aims to maximise the covariance between two data sets (Geladi and Kowalski, 1986). It uses an **X** matrix (inputs) and a **Y** matrix (outputs) , where the regression model is developed so that predictions of **Y** can be made from **X**. This research uses **Y** data which is both 3 dimensional and 2 dimensional and unfolding steps for 3 dimensional arrays are performed as described in section 4.2.1.

A simplified PLS model can be said to have both inner and outer relationships. There are two outer relationships and one inner, which links the two data matrices in the model. The outer relationship for the **X** data is given by Equation 4.5 which is the main equation used in PCA. These relationships are shown in Equations 4.8 and 4.9.

$$X = TP^T + E_X \tag{4.8}$$

$$Y = UQ^T + E_Y \tag{4.9}$$

where T represents the scores and P the loadings matrices. E is the residual errors matrix, and is not a part of the model. It is the part of the original **X** matrix which is not explained by the model (TP'). Similarly to PCA the errors of both equations can be reduced to 0, however this often leads to the inclusion of noise in the model. Model order selection for PLS models is discussed in section 4.3.5.

In PLS the aim is to describe the **Y** matrix as well as possible, and hence make E_Y as small as possible. This is done through minimising the covariance of the **X** and **Y** scores as displayed graphically in Figure 4.6.

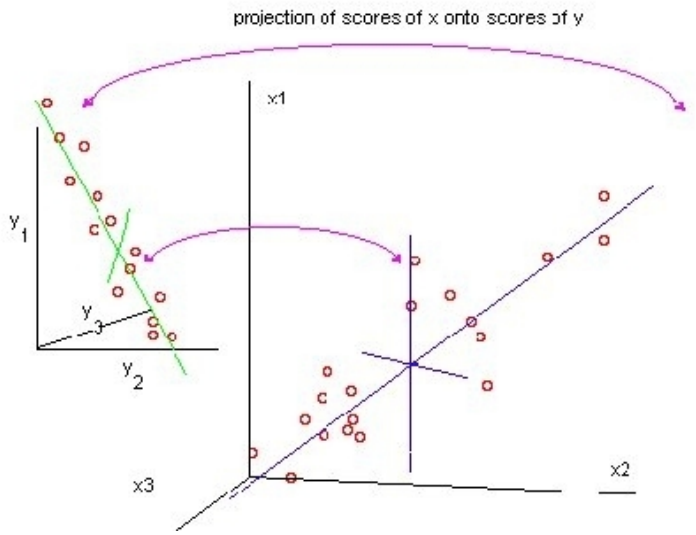


Figure 4.6: Illustration of latent variable determination; the black line represents PC of outer PLS models, the green and blue lines show latent variable which maximises the covariance of the X and Y scores (T and U) (SeparationsNow.com, 2014).

The process shown in Figure 4.6 is performed for each LV in the model, and is given in Equation 4.10.

$$\hat{u_h} = b_h t_h + e_h \tag{4.10}$$

where e_h is a vector of errors, b_h is a unknown parameter estimated by $b_h = u_h' t_h / t_h' t_h$ and t_h is the scores vector for the **X** matrix. In PLS the extraction of each pair of latent variables (t_h and u_h) is an iterative process.

4.3.5 Model structure selection

Selecting the optimum number of components in a PARAFAC model, principal components in PCA or latent variables in PLS is important to obtain the best analysis or predictions of the cultivation variables. Section 4.3.5 explains the procedure for model order selection. Additionally the selection of the most relevant model inputs will be discussed in section 4.3.5, as only the inclusion of the most relevant variables may improve the prediction performance of the model (Wold *et al.*, 2001a). For example in PLS not all the available measurements may be correlated with the data to be predicted. Including measurements which are not correlated with the data to be predicted causes a decrease in the model performance and results in poorer predictions.

Model order selection

Determining the number of components to use in a PARAFAC model can be difficult. The cross-validation method which is used in PCA cannot be applied here as the data array and the model is calculated for all components simultaneously. There are three techniques for determining the right number of components as summarised by Bro (1997). Due to the variation within the data sets used in this research the technique which utilises external knowledge was used. This technique relies on the researcher having adequate process knowledge to know when components are modelling noise. This is a subjective technique and relies upon interpretation of the loading plots.

The number of variables within each model is determined through interpretation of the scree plot. Which is a graph displaying the eigenvectors, and relates to the variation captured in each PC or LV. This means the interpretation of the scree plot is based on the assumption that important information is larger than random noise and that the magnitude of the variation of the noise levels off with the number of components (Qin and Dunia, 2000). Therefore the eigenvalues can be plotted as a function of number of components, and as discussed in section 4.3.3, when the eigenvalues start to show a linear relationship, that is the optimum number of components (Bro

and Smilde, 2014). A second option for autoscaled data is to use all components where the eigenvalue is above one (Wu and Manne, 2000). If all the variables were orthogonal to each other then every component in the model would have an eigenvalue of one, but if a component has an eigenvalue larger than one it is capturing variation from more than one variable. Another option is to look at the variance explained in each component (Joreskog *et al.*, 1976). If a two component model explains 50 % of the variation, it is likely that more components are needed. However if a 6 component model is used which explains 90 % of the variation it is likely that noise is being captured and the number of components should be reduced. The method of using cross-validation was introduced by Wold (1978a). For PCA and PLS this research uses a leave-one-out cross-validation method to select the optimum model order (R). Here R is the number of principal components or latent variables in the model. Leave-one-out cross validation systematically leaves each batch out of the training dataset once. Models are then constructed from the other batches. Next the models are validated using the left out batch and the mean Sum of Squared Errors (SSE) over all the batches that were in the training set is calculated for each model order. Generally the leave-one-out method has a larger computational requirement, and would not be appropriate for large data sets (Kearns and Ron, 1999; Kohavi, 1995). However as the data sets used in this research are relatively small, this is not an issue. The calculated root mean squared error of cross validation (RMSECV) can be plotted against the number of PCs or LVs and the point at which it is the lowest, without being linear is chosen as the number of components in the model (Bro and Smilde, 2014).

There are various other methods as described in literature to determine the number of variables in a model such as: Akaike information criterion (AIC), final prediction error criterion (FPE), bayesian information criterion (BIC), Normalised residuals sum of squares (NRSS), multiple correlation coefficient (R^2), adjusted multiple correlation coefficient (R_a^2), and Wold's R criterion. A review of these different techniques can be found in the work of Haber and Unebnhauen (1990), with a comprehensive study of the most popular technique (Wold's R criterion) given in the study carried out by Li *et al.*

(2002). However for the application to this research the RMSECV was the most appropriate as it provides relatively accurate results and is simple to use.

Input variable selection

PCA and PLS have the capability of handling noisy data, as discussed in section 4.3.1. However the quality of the analysis or predictions of a regression model can be significantly improved through eliminating measurements which are not correlated with either the investigating parameter (PCA) or the final batch quality (PLS) (Kosanovich *et al.*, 1996).

During data pre-processing techniques noise can be introduced to the system. There are techniques available in literature, such as Andersen and Bro (2010), which can be used to determine when and if a variable should be eliminated from the analysis. However due to the nature of the measurements and the size of the data matrix in this research process knowledge was used. The loadings (a measure of variable variability) can be related back to the original process data to give a clear picture of the variation and/or noise associated with each variable (Jolliffe, 2002). This technique will be used along with process knowledge to determine the applicability of retaining a variable in the models.

4.3.6 Model assessment criteria

To assess the performance of the models produced in this research the root mean squared error (RMSE) is calculated for each model (Equation 4.11).

$$RMSE = \sqrt{\left(\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}\right)} \tag{4.11}$$

where y_i is a vector containing original values, $\hat{y}_i$ is a vector containing predictions and n is the number of samples in the vectors. The RMSE value is a measure of how well the model predicted the test data set (Ramadan *et al.*, 2005). As Qi and Zhang (2001) discuss there are other techniques for

assessing model performance, such as, the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). When creating a model including more parameters than are necessary can lead to an over-fitted model. Both the AIC and the BIC assess the fit of the predictions whilst introducing an additional criteria to take into account the number of parameters in the model. This results in an AIC or BIC value which is larger for models that contain more parameters. Due to the similarity between these two methods only one is used in this research. The AIC value was chosen because the additional criteria for model parameters is smaller than for BIC (Ramadan *et al.*, 2005). The AIC is determined by Equation 4.12 (Akaike, 1987).

$$AIC = 2k + nlog(\frac{RSS}{n}) \tag{4.12}$$

where n is the number of observations, RSS is the Residual Sum of Squares and k is the number of model parameters.

4.3.7 Analytical software

The models used in this work were developed using MATLAB® (2013a, The MathWorks, Massachusetts, United States). Alongside this the PLS Toolbox chemometrics software (Eigenvector Research Inc., Washington, United States) was used for the multivariate statistical applications. For the creation of the DoE used in the chromatography experiments the software Design-Expert® (Stat-Ease Inc., Minneapolis, United States) was used.

4.4 Mechanistic modelling

This research uses mathematical models in the form of differential equations. In the last twenty years there have been various attempts to characterise mammalian cell cultures (Glacken *et al.*, 1988; Suzuki and Ollis, 1989; Fernandes *et al.*, 2013; Gadgil, 2015a). The mechanistic models used in this research are specific to CHO cells and as such CHO specific models were selected (Gadgil, 2015a; Xing *et al.*, 2010; Naderi *et al.*, 2011; Kontoravdi

et al., 2007). Within this area there is still a broad spectrum of models and two models were selected of differing complexity in terms of the number of variables (Naderi *et al.*, 2011; Kontoravdi *et al.*, 2007). The advantage of studying the model complexity is that a comparison could be drawn not only between the multivariate and mathematical models, but also between the hybrid models developed using combinations of multivariate and mathematical (Psichogios and Ungar, 1992).

4.4.1 Cultivation

Two different mechanistic models were used to characterise the system. These two models differ in complexity; with one model providing an overview of cell cultivation, and second expanding this to include all the main metabolite concentrations as well. Therefore these models can be interchanged when more is known of the process. The first model which was investigated was developed by Naderi *et al.* (2011). This is a series of equations which have been adapted from a hybridoma metabolic model for CHO cells. The model includes the main glycolysis pathways, energy related amino acid metabolism and production of biomass and mAb. The model is constructed using a mass balance for each metabolite, giving the extracellular concentrations and specific uptake/production rates during the cell cultivation. The viable cell concentration is given as a function of time, making it a dynamic model. It is assumed at the specific uptake/production rates are constant. The equations for this model are given in Appendix B part B.

The second model used is described by Kontoravdi *et al.* (2007). This is the most complex of the two models as it describes the cell growth kinetics and cell metabolism. The amino acid metabolism is based specifically on CHO cells (therefore it is not applicable to hybridoma cells), and was determined based on the work of Alberts *et al.* (2002). It is different to the model presented by Naderi *et al.* (2011) as Naderi *et al.* (2011) only include the amino acids which they have deemed to be the most important, whereas Kontoravdi *et al.* (2007) have included all twenty amino acids.

4.4.2 Purification

There will just be one mechanistic model used for the IEX chromatography unit, which is used to predict the retention time for ion exchange chromatography. This model is given in Chapter 6 section 6.5. This model was taken from literature, and is shown in the work of Shellie *et al.* (2008) and Madden *et al.* (2002).

4.5 Hybrid modelling

Hybrid modelling combines the advantages of both models, whilst simultaneously removing some of the barriers to application. Advantages include:

- Greater prediction accuracy:
    1. Lower number of model parameters when compared to non-parametric models.
    2. Model can account for changes to variables when compared to phenomenological models.
    3. Constraints can be included to avoid infeasible solutions.
- Phenomenological aspects allows for process operation space to be defined, which reduces the computational demand.
- Improved extrapolation properties when compared to non parametric models.
- Less phenomenological knowledge is required when compared to white box models.
- Hybrid models are faster to develop than white box models as less knowledge is required.

Hybrid modelling combined different sources of information in one framework. The main issue being the structure of the framework used;

literature presents two different options, these being serial or parallel and one step or multi step. The hybrid models can be structures in three different ways (Figure 3.2), the first (A) is termed a parallel structure with (B) and (C) being serial structures. Agarwal (1995) discusses these three structures and states that the black box represents non parametric models and the white box phenomenological models.

4.6 Summary

This chapter described the sources of data used in the research, including data obtained from collaborators and data generated through experimentation. Furthermore the protocol for applying the MVDA techniques used in the research was discussed, including Parallel Factor Analysis (PARAFAC), Principal Component Analysis (PCA), and Partial Least Squares (PLS). The PARAFAC analysis was conducted using the n-way toolbox (Andersen and Bro, 2000), PCA and PLS was conducted using the PLS toolbox (Eigenvector Research, Inc.). PARAFAC was selected for this research as it can highlight variation between batches, PCA was selected for this research as the variance within time series data sets with multiple variables can be analysed (Kourti and MacGregor, 1995). PARAFAC, PCA, and PLS were selected for use in this research because all three methods can handle multi-way data, they reduce both the dimensions and noise of the data, they are well established tools for use in the pharmaceutical industry, and there is evidence (in the case of PLS) of the technique being applied with hybrid modelling. The study presented in the following chapter (Chapter 5) investigates the use of PARAFAC, PCA, and PLS to allow for an informed decision as to which techniques to use in the chromatography modelling and the agent based model. Similarly to the array of multivariate technique available, there are numerous first principles models which can be used to describe the cultivation, this research investigates the incorporation of two cultivation models of varying complexity into a hybrid structure. One mechanistic model has been chosen for the IEX chromatography unit, as it is possible to specify the process parameters specific to the system within the model. The next chapter presents a study of

hybridoma glycosylation, with the aim of selecting the most appropriate MVDA techniques.

Chapter 5

An investigation of the effects of operating conditions on hybridoma cell metabolism and glycosylation of produced antibody

The previous three chapters have provided an overview of bioprocessing, introduced some of the literature associated with modelling bioprocesses, and discussed the techniques which this research uses. This chapter covers the research carried out to investigate the ability to model and predict the metabolism and final glycosylation profile of Hybridoma cells using multivariate data analysis (MVDA) and first principles modelling. The MVDA uses both on-line and off-line measurements with the aim of predictions from on-line measurements being used for real-time control of a cultivation.

The study presented in this chapter aims to establish the link between process operating conditions, cell metabolism and the production of different glycosylated forms of a monoclonal antibody (mAb) produced by a murine Hybridoma cell line. This is the first step towards the development of a hybrid model, and provides the link between on line monitoring and control of the final glycosylated product. Glycosylation is important as it can have a major impact on the safety and efficacy of the product. When compared to a naturally

occurring protein, a protein that is missing carbohydrate side chains has the potential to aggregate, which may enhance a patients immune response (Sofer and Hagel, 1997). The literature has shown that no strategy currently exists for on-line control of glycosylation. This is primarily because glycosylation is a non-template driven cellular process. However recent work by Amand *et al.* (2014) suggests that it would be possible to control glycosylation through cultivation conditions. Therefore as control of glycosylation is possible this study will aim to investigate the routes by which it can be achieved.

5.1 Methodology

The data used in this study was collected as summarised section 4.1.1. This work was carried out in collaboration with Marija Ivarsson, a PhD student at ETH Zürich, and the cultivation experimentation was conducted by the collaborators. More detailed information on the experimental procedure is given by Ivarsson *et al.* (2014).

The data set contains 13 separate cultivations, these experiments were conducted using the *one factor at a time* principle. All cultivations were operated at standard conditions until 30 hours, when a process shift was introduced. This changed the level of one of the investigated variables in each batch. The standard cultivation settings were;

- 50 % dissolved oxygen
- 380 $mOsm/kg^{-1}$ osmolality
- 7.2 pH
- 0.05 vvm aeration rate

The conditions of the 13 batches are shown in Table 5.1. As can be seen batches 1-3 investigated three different levels of dissolved oxygen, a low level (10%), a high level (90%) and a centre point (50%). Batches 4-6 and 13 investigated four different levels of osmolarity ($mOsm/kg^{-1}$), a low level (350 $mOsm/kg^{-1}$), a high level (420 $mOsm/kg^{-1}$), a very high level (450

$mOsm/kg^{-1}$), and a centre point (380 $mOsm/kg^{-1}$). Batches 7-9 investigated changes to pH, with a low level (7.0) a high level (8.0) and a centre point (7.2). The final three batches, 10-12 investigate changes to aeration rate, with a low level (0.05 vvm), high level (0.20 vvm), and a centre point (0.10 vvm). It can be seen that there is a total of 3 batches (1,5, and 10) which are all operated at the same conditions.

Table 5.1: Experimental conditions of each of the 13 experimental batches used to construct and validate the PCA and PLS models. There was a condition change at 30 hours into the cultivation. Prior to the parameter shift the cultivations were operated at a standard setting (50 % dissolved oxygen; 380 $mOsm/kg^{-1}$ osmolality; 7.2 pH; 0.05 vvm sparging). Highlighted in bold is the parameter that was being investigated in each batch.

Batch identifier	Dissolved oxygen (%)	Osmolality ($mOsm/kg^{-1}$)	pH	Sparging (vvm)
1	**50**	380	7.2	0.05
2	**10**	380	7.2	0.05
3	**90**	380	7.2	0.05
4	50	**350**	7.2	0.05
5	50	**380**	7.2	0.05
6	50	**450**	7.2	0.05
7	50	380	**7.0**	0.05
8	50	380	**7.5**	0.05
9	50	380	**8.0**	0.05
10	50	380	7.2	**0.05**
11	50	380	7.2	**0.10**
12	50	380	7.2	**0.20**
13	50	**420**	7.2	0.05

Initial treatment

The techniques used for the initial treatment of the data are described in chapter 4 section 4.2. The matrices were unfolded in a batch wise way (Nomikos and MacGregor, 1994). Batch wise data unfolding was chosen as it preserves the variability between batches, which for this data set and application were important as literature has shown small changes to cultivation can greatly impact performance (Gomez *et al.*, 2010). Having unfolded the data the pre-processing steps can then be applied. An example of the raw data

prior to pre-processing is shown in Figure C1, which corresponds to batch 9 in Table 5.1. The data set presents various challenges, including:

- variation in the cultivation duration
- random sampling time of the off-line data
- incomplete recordings

The duration of the cultivations varied considerably, ranging from 74.2-94.8 hours. Of the 13 cultivations, 11 were between 72-78 hours and only two extended to >90 hours. Figure C2 shows a comparison of the raw off-line data for batches 9 (78 hours) and 11 (90 hours). As can be seen in Figure C2 (a) the cells have achieved the four stages of cell curve characterised by the cell growth curve (chapter 2,section 2.3.1, Figure 2.6). Similarly for figures C2 (b-e) the consumption/production profiles which would be expected are shown for both batches. In particular it is noted that for all five off-line measurements the expected profile is achieved by 70 hours and as shown by the titre (b) the maximum production of the product has been achieved. Therefore it was not detrimental to the modelling of the cultivation to cut the 15 batches to the length of the shortest run (74.2 hours). It is necessary to cut the data because when applying MVDA techniques the matrices used must be of equal length as if there are too many missing data points it becomes difficult to apply.

To account for the random sampling time of the off-line data a cubic spline was used. Figure C3 shows the data for batch 9 after the cubic spline interpolation has been applied and after the data cut (to 74.2 hours). The cubic spline was applied to the off-line data using the time the on-line data measurements were recorded. This was done so that a comparison could be achieved at a specific point of time in the cultivation duration. Additionally the data was sampled, this is to reduce the size of the matrices. The on-line data consisted of ∼9000 measurements for each variable. This is a very large data set to work with; therefore the data was sampled whereby every 10th measurement is used.

The final issue encountered in the data set was incomplete recordings. The main problem arises with the ammonia measurements as it was only

successfully recorded for batches (8-11 and 13-15). Models were constructed which included and excluded the ammonia data to determine if the lack of available data was influencing the model. This is discussed further in the following sections.

Pre-processing

The methodology for pre-processing the data is given in Chapter 4. There are various pre-processing techniques which can be applied to multivariate data, which can account for scattering (spectroscopy (Næs, 1989)), to show hidden peaks (spectroscopy (Savitzky and Golay, 1964)), or to smooth noisy data (Press *et al.*, 2007). In this application the only pre-processing technique which was used is autoscale, which combines mean centering with division of each column by the standard deviation. Autoscale was used because it allows for direct comparison between the variables.

Multivariate data analysis (MVDA)

The protocol for constructing and applying MVDA models is given in chapter 4, section 4.3. The application of multivariate data analysis techniques (PARAFAC, PCA, and PLS) were carried out using MATLAB® (R2014a, the MathWorks Inc.) and the PLS Toolbox™ (Eigenvector Research Inc.). PCA was applied to get an overview of the process data. Combinations of on-line and off-line data were explored with the aim of establishing the CPPs so that a predictive model can then be constructed which focuses on using these to predict CQAs. For PCA all 13 experimental batches were used with a leave one out cross validation method applied. Having established trends and clusters in the data, PLS modelling was used to develop regression models.

The on-line measurements collected for each cultivation included the % of Dissolved Oxygen (DO) in the cultivation medium, the % of Oxygen (O_2) in the air feed to the cultivation, the % of Carbon Dioxide (CO_2) in the air feed to the cultivation, the pH of the cultivation medium (pH), the flow of base to the cultivation (base), the stirrer speed of the impeller, and the temperature of

the cultivation.

5.2 Results and discussion

This section presents three analysis/modelling techniques. The first of which was Parallel Factor Analysis (PARAFAC), which was explored as it is a multi-way method where there is no rotation issue as there is for PCA or PLS. Furthermore the resulting models are simple typically and therefore easy to interpret. This would be an advantage for this research as it is intended to be adopted in industry. The second technique was principal component analysis (PCA), and the final technique is partial least squares (PLS). Description of the application of these tools is given in chapter 4 section 4.3. The results of the application to a hybridoma cell line regarding the analysis and prediction of the cell metabolism and product production, but also the applicability of the technique to hybrid modelling and this research are discussed in the following sections.

5.2.1 Parallel factor analysis

The first multivariate technique used on the data set was parallel factor analysis (PARAFAC), section 4.3.2 in chapter 4 provides details on the application of PARAFAC. As described in section 4.3.2 PARAFAC analysis uses data contained within a 3-D array, therefore data unfolding was unnecessary. The aim of this case study was to determine whether PARAFAC can be used as a tool in industry for analysing mammalian cell cultivation. It was compared with principal component analysis in section 5.2.2 to determine which is the most appropriate technique to take forward in this research. With PARAFAC there is no rotation problem, which is beneficial for data sets such as that being investigated in this study where batch length and sampling intervals are irregular. PARAFAC allows for the visualization of the data in a way that is relative to the raw information, therefore making it useful as a precursor to principal component analysis (Bro, 1997). The data was arranged in a

three-way array with time, batches and variables representing the dimensions, respectively. Where batch lengths were unequal, 'not a number' was entered. To construct the PARAFAC analysis, the n-way toolbox was used.

On-line data

The PARAFAC model constructed for the on-line data contained three components (Model 1). This was selected as the analysis was shown to explain the effects of interest, i.e. the constructed model was shown to explain variability in cultivation duration (Figure C4 (a)), variables (Figure C4 (b)) and batches (Figure C4 (c)). Using more components did not yield any further useful observations. The analysis indicated that PARAFAC can capture the variation in the system and the controlled parameter shift at 30 hours. Figure C4 (a) demonstrates this shift in Mode 1 of the model for each of the three components, but primarily in component 3. PARAFAC was especially beneficial in assessing the quality of the experimental runs as it highlighted specific runs deviating from the typical observed behaviour (as illustrated by Figure C4 (c)). However this deviation was not excessive and could be explained by the different operational conditions each experiment operated under. Figure C4 (c) shows that of the 13 batches used in the analysis (Table 5.1), batch 3 was highlighted as exhibiting different behaviour. This batch was carried out with a DO shift to a high value at 30 hours. Subsequent analyses also highlighted the different behaviour of this batch (see results of PCA in section 5.2.2).

Off-line data

The PARAFAC model constructed on the off-line data also contained three components (Model 2). Again additional components did not yield further results. The model shows variability in cultivation duration (Figure C5 (a)), variables (Figure C5 (b)) and batches (Figure C5 (c)). Figure C5 (a) shows the variation during the duration of the cultivation. As can be seen component one highlights the variation at the beginning of the cultivation, component two

during the middle and component three at the end of the cultivation. This reflects the cell growth curve, and suggests that the dominant variation in mode one is to do with growth and production. Figure C5 (b) highlights that the most important variables were viable cell count, percentage viability, glucose, lactate and titre, with the amino acids not having much influence over any of the three components in the model. Figure C5 (c) shows the analysis of the batch variation. As can be seen there is a great deal of variation and influence over the model in all 13 batches. Component one highlights that all 13 batches are important to the model, component two illustrates the difference between the pH runs (7-9) and the other runs. Finally component 3 highlights a big difference between runs (1-3 and 13) and the rest of the runs used to build the model. The reason for this is not clear as run 2 was operated at the same conditions as runs 5 and 10, and run 13 was investigating osmolality whereas runs 1-3 investigated DO. PARAFAC has shown to be good at identifying sources of variation within a data set, but not the reasons for the variation.

5.2.2 Principal component analysis

Principal components analysis (PCA) was used to increase understanding of the cell cultivation data, with the aim of establishing the influence of various operating parameters on the final CPPs and CQAs which include titre and glycosylation profile. The PCA analysis performed on the on-line data was used as a variable reduction method, whereby the data was analysed to determine which variables account for variation and which are sources of noise in the data. The PCA analysis on the off-line data was used to understand the influence of changes to operating parameters on various measured parameters and to establish whether it would be possible to predict or control these measured CQAs.

On-line data

The first PCA carried out on the data was conducted on a matrix which included all 13 experimental batches with the 7 on-line variables. The aim of

this analysis was to determine if all variables had an equal effect upon the variation in the batches, and if the conditions of certain batches impacted the analysis more. The resulting model contained 5 principal components (PCs) and accounted for 84.41% of the variation in the data. Figure 5.1 is a bivariate scores plot for PC1 and PC2, which are the two PCs that captured the most variance. It can be seen that for PC1 there is a wide spread in the batches which investigated pH (blue), with batch 9 lying close to the 95% confidence limit, this is due to the fact batch 9 was operated at pH8 which is outside the normal operating pH for mammalian cells (Table 2.1, 13). PC2 shows a wide spread in the dissolved oxygen (DO) experiments, suggesting that the variation in PC2 is accounted for by DO. Batch 3 was close to the confidence limit and as this batch was operated at 90% DO this was not surprising.

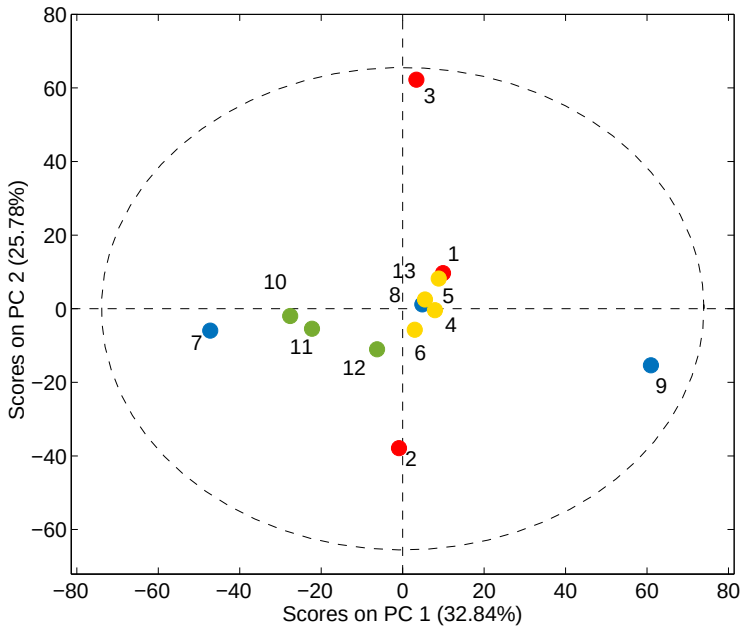


Figure 5.1: Bivariate scores plot showing PC1 and PC2 for the PCA analysis containing the on-line data for all 7 measured variables. The scores are grouped to show dissolved oxygen (red), osmolality (yellow), pH (blue), and sparger (green).

Although Figure 5.1 suggests a relationship between PC1 and pH, and PC2 and DO, the random scatter of most of the points shows that there are other influences which are causing variation. To investigate these the loadings for the model can be used to see how each of the variables influence the variation. Figure 5.2 shows the loadings for all 7 variables for PC1, the figure

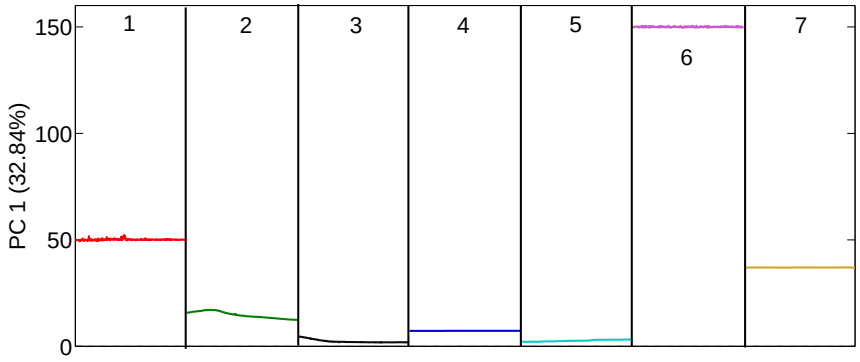


Figure 5.2: Loadings plot for PC1 showing all 7 variables; (1) DO, (2) O_2, (3) CO_2, (4) pH, (5) base, (6) temperature, and (7) stirrer speed.

suggests that temperature and stirrer speed have a significant impact on the variation in the data. However when observing the raw data it can be seen that these two variables change very little. Therefore it is more likely that the effect seen in Figure 5.2 is due to the scaling of the data. During the scaling process each data measurement within a variable was weighted and scaled down, therefore if a variable only changes by a small amount this small change will be magnified. Due to this process and technique knowledge a PCA was subsequently performed with stirrer speed and temperature removed from the analysis.

Figure 5.3 shows the weightings for the analysis with stirrer speed and temperature removed. The new model was constructed using 3 PCs and accounted for 89.04% of the variation in the data. As can be seen from the weightings the variable dissolved oxygen is now the most significant. It can be seen that the weightings for the variables O_2, CO_2, and base vary over time with the importance of O_2 and CO_2 decreasing with time whilst the importance of the base variable increases over time. This would suggest that O_2 and CO_2 are important during the cell growth stage of the cultivation, with the base being important during the production stage. The weightings for PC2 and PC3 are given in the appendix in Figures C6 and C7. As can be see dissolved oxygen is the mots important variable when considering the variation captured in each of the three principal components.

The weightings have then been used along with the scores values for each

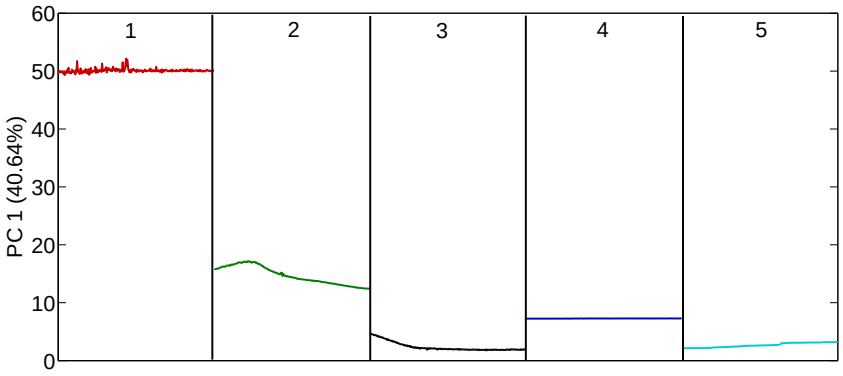


Figure 5.3: Loadings plot for PC1 showing 5 on-line variables; (1) DO, (2) O_2, (3) CO_2, (4) pH, and (5) base.

batch to produce the bi-plot shown in Figure 5.4.

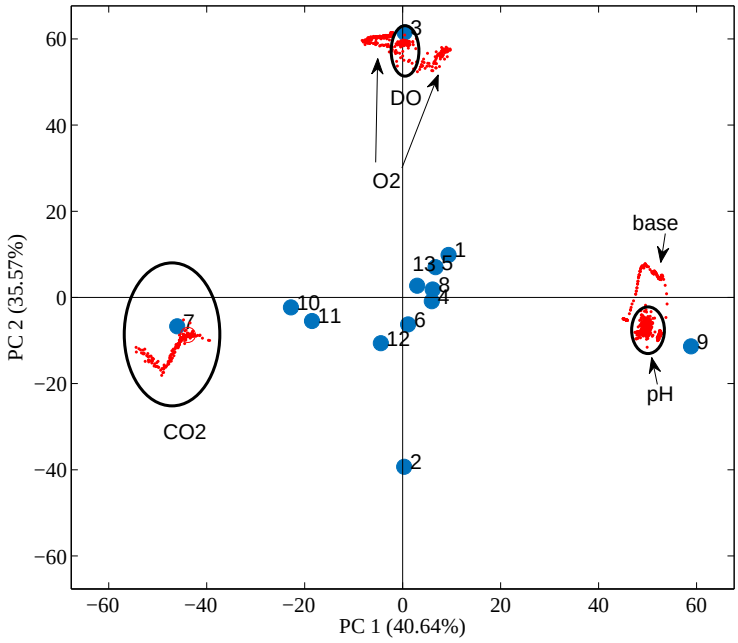


Figure 5.4: Bi-plot for PC1 and PC2 for the PCA model constructed using 5 on-line variables. The scores are label with their batch number, and the variables are highlighted in ellipses.

It can be seen that there is a direct correlation between the variable DO and the batch which was operated at the highest concentration of DO (batch 3). Additionally it can be seen that there is a relationship between the DO concentration and the O_2 in the cultivation which is as would be expected. Similarly the relationship between the base concentration and the pH is as expected, along with the importance on batch 9 (the high pH batch). What is

perhaps a less obvious relationship is that between the low pH batch (batch 7) and the level of CO_2 in the cultivation. From analysis of the loadings values it is known that the group of carbon dioxide points, batch 7 is close to, relate to near the end of the cultivation, i.e. during the end of the growth phase and through the production phase. This strong correlation to the carbon dioxide concentration suggests that the cells are growing more. This links with the low pH level this batch was operated at because as shown in Trummer *et al.* (2006) high pH levels inhibit cell growth.

In summary the PCA analysis of the on-line data has shown that for any subsequent models the temperature and stirrer speed data should be discarded. Furthermore the analysis has shown a correlation between the operating conditions and the on-line data showing that there is a link between the CPPs and the CQAs.

Off-line data

There were various off-line measurements taken from the 13 cultivations these include; glucose concentration, lactate concentration, titre, viable cell count, amino acid concentrations, and glycosylation of final product. These measurements can provide a lot of information on whether the variation in the data set is ordered or random. The first PCA analysis performed combined the DoE operational set point data given in Table 5.1 with the final product titre. The resulting model contained 3 PCs which explained 72.75% of the variation in the data. Figure **??** shows the bivariate scores plot for PC1 and PC2. It is shown that there is a strong link between the variation captured in PC1 and pH, and the variation captured in PC2 and the concentration of gases dispersed in the media. To obtain a more thorough understanding of how the variables contribute to this variation a bi-plot has been produced (Figure 5.5).

As can be seen there is a strong relationship for all batches which were operated under the extreme condition and the variable they were testing. For example batch 9 operated at the high pH value of pH8, and this batch is shown to be strongly related to the pH variable, similarly batch 12 which operated at

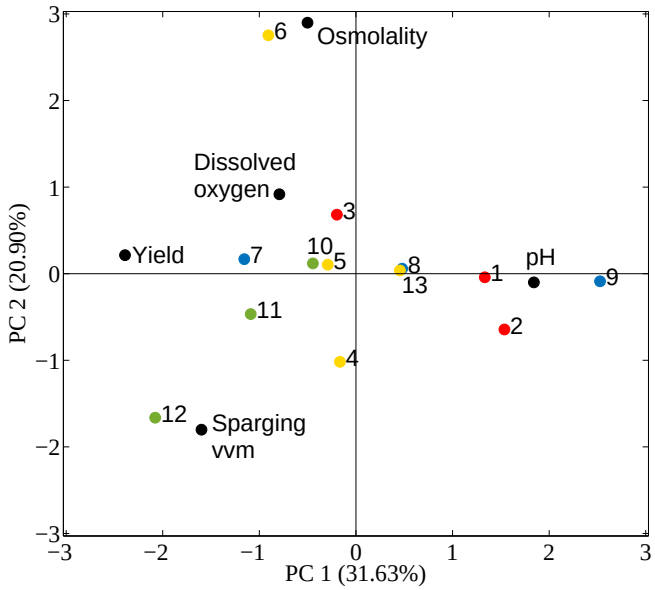


Figure 5.5: Biplot showing PC1 and PC2 for the PCA analysis containing the off-line data for operating parameters and titre. The scores are grouped to show dissolved oxygen (red), osmolality (yellow), pH (blue), and sparger (green). The variable loadings are shown in black and labelled as to the variable they relate to.

0.2 vvm is shown to be closely correlated to the sparging variable. What is perhaps not as expected are the results for the DO batches (1, 2, and 3). Batches 1 and 2 are shown to be more closely related to the variable pH than to the DO variable. A possible reason for this is that the affect of the pH of the cultivation media and the affect of the level of dissolved oxygen on the cell are closely related. As discussed by Naciri *et al.* (2008) at lower concentrations of dissolved oxygen the cells are more susceptible to changes in pH. Whereas when the DO is in excess small changes to the pH do not influence the cells as much.

The variable 'yield' is shown to be positively correlated to the dissolved oxygen levels and the osmolality, suggesting a higher level of dissolved oxygen promotes a higher titre. Additionally there is a negative correlation between titre and pH suggesting pH can inversely affect the production of protein. This is supported by batch 7, which operated at pH 7 and is shown to be closest to the yield variable. With regards to using this data in a PLS model some difficulties may arise around batches which are far from the centre cluster i.e. batches 6, 9, and 12. This is because the data set is small in size and

with some of the batches being extreme values it means these batches influence the trained model more. Additionally if these batches are used for validation they are so different to the batches which would be used to train the model that the prediction would be expected to be poor.

The second PCA model constructed from the off-line data included in the matrix the variables; glucose concentration, lactate concentration, titre, and viable cell count. The resulting model contained 4 PCs and explained 89.15% of the variation in the data. Figure 5.6 shows a bi-plot of the resulting scores and loadings for the model. It can be seen that there is a relationship between the batches testing DO level and the viable cell count. This is as expected as the cells require oxygen to grow and thus changes to DO levels would result in higher or lower growth. What is perhaps not expect is the position of batch 1, as although this batch is considered to be one of the DO test group, it has similar conditions to batches 5 and 10. These three batches are relatively spread out across PC1, indicating that PC1 captures a variable which is not measured and varies between these three batches. The glucose concentration is indicated to be important in batches 7, 8, and 9, with batch 7 especially the glucose is important during the end of the cultivation. As the pH for batch 7 is the lowest for all the pH test set (pH 7) it suggests that higher pH affects the cultivation more (Naciri *et al.*, 2008). All the batches relating to the sparger vvm and the high osmolality batch are related to the lactate concentration and the product titre. This is an interesting grouping, as all the batches within the group are concerned with mass transfer parameters and it could indicate that it is the distribution of the different components within the media that affect the cell production.

With regards to the variables it can be seen that the product titre is closely linked to the lactate concentration, and the viable cell count is inversely related to the glucose concentration. This follows what is known of the cell cycle in that glucose is used for cell growth, and lactate is a by-product of growth. With respect to using this data in a PLS model it can be seen that all 4 variables are important to the data variation and should be included. The batches have again shown that there is some variation in this case with batches

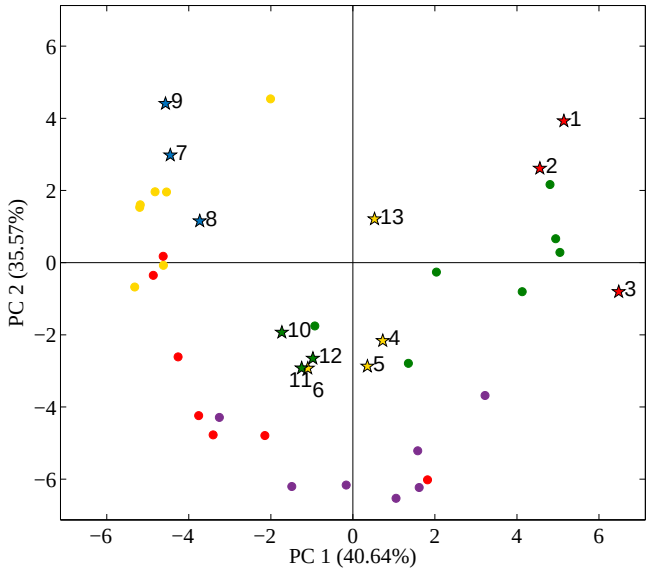


Figure 5.6: Bi-plot showing PC1 and PC2 for the PCA analysis containing the off-line data for glucose (yellow circles), lactate (red circles), titre (purple circles), and viable cell (green circles). The scores are grouped to show dissolved oxygen (red stars), osmolality (yellow stars), pH (blue stars), and sparger (green stars).

1, 2, 3, and 13 which may prove challenging in a predictive model. The bi-plot for PC3 and PC4 is given in Figure C8 in the appendix. As can be seen there are no strong links between the operating conditions of the batches and the profiles of the different off-line variables.

So far the PCA analyses presented in this chapter have focused on the CPPs given in the experimental plan, and the CPPs of viable cell count and product titre. These are important parameters and will be included in the final model network, however the main focus of this chapter is to investigate the CQA of the final product glycosylation profile. It is known that there is a link between the amino acids in the media and the glycosylation of the final product (Ivarsson *et al.*, 2014), thus the next PCA analysis considered the operating parameters, titre, and glycosylation profile. The resulting model contains 5 PCs and captures 89.18% of the variation in the data.

Figure 5.7 shows the bi-plot of the model for PC1 and PC2, with a bi-plot for PC3 and PC4 given in Figure C9. The figure suggests a relationship between the operating parameters and the final glycosylation profile. The variable pH seems to be strongly related to the concentration of G0F, further

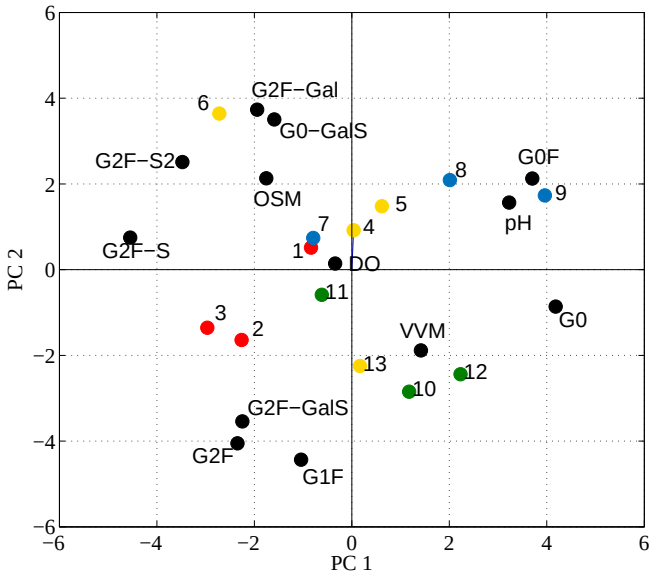


Figure 5.7: Bi-plot showing PC1 and PC2 for the PCA analysis containing the off-line data for the operating parameter set points and final product glycosylation profile. The scores are coloured to show dissolved oxygen (red), osmolality (yellow), pH (blue), and sparger (green). The glycans and the off-line operating parameters (black) are labelled to identify which item they are.

supported by the scores for batches 8 and 9, which shows that an increase in pH causes a stronger correlation to G0F. This is the only glycan which is positively correlated to pH. In the opposite quadrant (bottom left) it can be seen that there are 3 glycans (G1F, G2F, and G2F-GalS). The main difference here is that G0F does not contain any galactose, whereas G1F (one galactose), G2F (two galactose) and G2F-GalS (two galactose) all do, with all four glycans being fucosylated. This indicates that pH can influence the ability of galactose to bind to the protein. This is supported by the work of Borys *et al.* (1993), who showed that cell external pH can influence glycosylation. This study focuses on glycosylation as a whole and not the individual glycans, therefore there is scope for a study investigating the effect of pH on the individual glycosylation mechanisms. Considering now the fucosylation of the glycans; returning to G0F in the top right quadrant, if we compare this to G0 in the bottom right quadrant it can be seen that both of these are influenced by the pH, which affects the binding of galactose. However G0, which is not fucosylated is shown to be correlated to the gas flow rate (VVM) and is confirmed by the placement of batch 12 (0.2 vvm) indicating that higher the

flow rate of air the more the glycosylation pathway is influenced to produce G0. This is suggested to be further supported by the glycans G2F-S, G2F-S2, G2F-Gal found in the top left quadrant and which are all negatively correlated to the VVM and found to be fucosylated. The issue however arises with the glycan G0-GalS, also found in the top left quadrant, which is not fucosylated. With this in mind, considering the placement of the variable OSM (osmolality) it suggests that the processes of fucosylation and sialylation (addition of sialic acid) are negatively correlated to osmolality. This is confirmed in the work of Konno *et al.* (2012) and Ivarsson *et al.* (2014) who showed that the fucose content of monoclonal antibodies could be controlled by culture medium osmolality. The placement of the glycan G0-GalS does indicate that sialylation is the more dominant process step, with the placement of the glycan G2F-GalS (in the top right quadrant) suggesting that the addition of galactose must occur before sialylation occurs. This is supported by Raju *et al.* (2001) and Kaneko *et al.* (2006) who both discuss how sialic acid commonly binds to galactose.

The placement of the DO variable near the origin shows that, when operated at 50%, the DO does not influence the glycosylation. It is only when the DO is operated at extremes that it becomes an influencing factor (batches 2 and 3). Meuwly *et al.* (2006) conducted a study to increase production and scale up of a cell culture process. They successfully increased product titre whilst maintaining a similar glycosylation profile. To summarise the findings of the effects of operating parameters a list is given below.

1. pH is correlated to the addition of galactose.
2. A higher pH means galactose does not bind.
3. The processes of fucosylation and sialylation are negatively correlated.
4. A higher osmolality promotes sialylation.
5. Higher gas sparing (vvm) appears to have some inhibitory effects on the sialylation and fucosylation.

Having established that there are relationships between the operating parameters and the final glycoyslation profile, this indicates that it should be

possible to produce a PLS model that can predict glycosylation.

Although the last PCA model established that there is a link between the operating parameters and the final glycosylation profile, there are references in literature which suggest the metabolism of the cell is an indicator of glycosylation (Yi *et al.*, 2012; Nyberg *et al.*, 1999; Butler, 2006; Wong *et al.*, 2005a). The amino acids are a measurement of the metabolism, therefore they would be used as an intermediate in the modelling scheme. Using the operating parameters (DO, pH, osmolality, and gas flow rate) to predict amino acid concentrations and then use the amino acid concentrations to predict glycosylation. Additionally this is perhaps the point at which multivariate modelling would be most useful as there are currently no mathematical models reported in literature which fully characterise the metabolism of hybridoma or CHO cells. Therefore a final PCA model was constructed using the amino acid concentrations at the time the glycosylation profile was recorded, and the glycosylation profile for the 9 glycans. The resulting model contained 4 PCs and captured 93.35% of the variation in the data with the majority of the variation being captured in PC1 (63.92%).

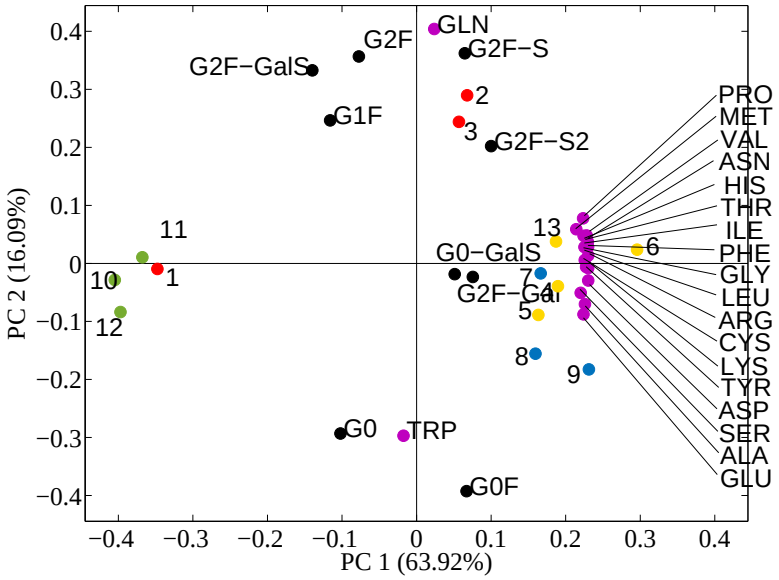


Figure 5.8: Bi-plot showing PC1 and PC2 for the PCA analysis containing the off-line data for the amino acid concentrations and final product glycosylation profile. The scores are coloured to show dissolved oxygen (red), osmolality (yellow), pH (blue), and sparger (green). The loadings are colour coded to show glycans (black) and amino acids (pink).

Figure 5.8 gives the bi-plot for the model for PC1 and PC2. The variables have been colour coded to shown glycans (black) and amino acids (purple). The scores have also been colour coded to show the batch groupings i.e. batches 1-3 relating to DO (red), osmolality (yellow), pH (blue), and sparger vvm (green). Initial observations suggest that there is not a strong correlation between the end amino acid concentration and the end glycosylation profile. This observation comes from the close grouping of 18 of the amino acids in the two right hand quadrants. There are two positive observations: firstly that concentration of tryptophan (TRP) seems to be correlated to the glycans which contain no galactose. Literature can aid in understanding this relationship: Lécorché *et al.* (2012) conducted a study showing that the use of both tryptophan and galactose in glycosylation is related. This has also been shown by Vliegenthart and Casset (1998) who showed that the addition of tryptophan occurs prior to the addition of galactose. This can be related back to Figure 5.8 suggesting that the higher concentration of tryptophan occurs when more of G0 and G0F are formed as these two do not consume the tryptophan. The second observation is that glutamine (GLN) appears to be important in the formation of the majority of the glycans. This disagrees with the findings of Taschwer *et al.* (2012) who showed that CHO cells produced in normal media and glutamine free media contained no significant differences in the final glycosylation profile. Therefore considering again Figure 5.8 the position of the GLN and TRP loadings relative to the other amino acid loadings suggests that the cell metabolism is being captured with glutamine being important to cell growth and the other amino acids being subsequently used in the production of protein. On this basis it suggests the variation in PC1 is related to cell production of protein, and the variation in PC2 is related to cell growth. Subsequently the clustering of the amino acids on the right hand side of the figure suggests that the individual concentrations of the amino acids have less of an impact on the glycosylation of the product. Figure C10 shows the bi-plot for PC3 and PC4, some of the labels have been removed for the variables which are close to the origin for ease of interpretation. The proximity of these variables to the origin shows that they have little influence on the variation captured by either PC3 or PC4. This figure shows little influence of the operating set points or the amino acids on the final glycosylation profile.

However it does again highlight that glutamine and tryptophan are important to the variation captured. In summary this analysis indicates that the off-line variables from the previous analysis (DO, VVM, pH, and OSM) would be better to use as predictors for final product glycosylation. This can be concluded from the fact that there appears to be two relationships apparent in Figure 5.8, the first being the importance of the amino acids in cell growth and the production of protein. The second being the relationship between operating conditions and glycans.

In summary the PCA analyses presented in this section have shown that it was necessary to remove two of the on-line variables (temperature and stirrer speed) as they introduced noise in to the system. From the off-line data it has been shown that there are strong correlations between the operating parameters and three of the target CPPs and CQAs (titre, viable cell count, and final product glycosylation profile). Additionally the analyses have shown how changes to these parameters made through the experiments have influenced the final CPPs/CQAs. The next section will consider using these data matrices to develop predictive models using partial least squares (PLS).

5.2.3 Partial Least Squares

Partial least squares (PLS) is a regression technique used to predict a **Y** matrix from an **X** matrix (see Chapter 4 section 4.3.4 for details). PLS was applied in this research to predict various CPPs/CQAs, namely titre, viable cell count, and glycosylation profile. The aim was to assess how well various input data correlates with these to enable their accurate prediction.

From the PCA analyses performed in the previous section it was known that there are relationships between the operating parameters, the on-line measurements, and the off-line measurements (glucose and lactate) and the CPPs/CQAs of interest (titre, viable cell, and glycans). Therefore for each of the CPPs/CQAs three models were constructed to predict them using each of the three identified sets of input variables. For all the PLS models constructed, the training data consisted of 11 batches with 2 batches being used as

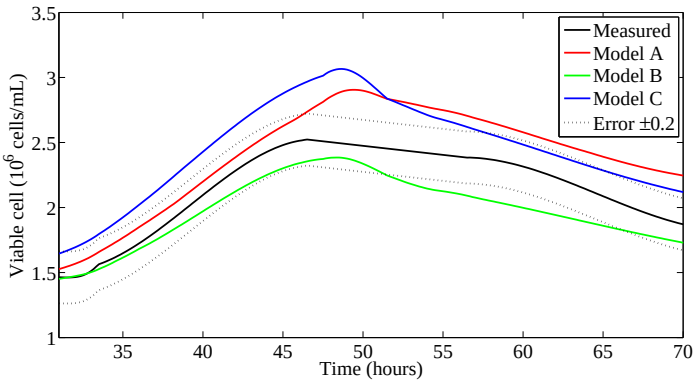
validation data. The batches used for validation were batch 5 and batch 13. These batches were chosen because they were not extremes of any of the variables, hence the model was not predicting outside of the range it was trained. Additionally both batches were part of the osmolality group of which there were more experimental batches. Therefore, selecting these two would not decrease the models ability to predict osmolality changes.

Three models are used to predict each of the viable cell, product titre, and glycosylation profile. This produces nine models in total. Table 5.2 reports the number of latent variables, the variance captured in the **X** and **Y** blocks, the RMSE, and AIC values for these models.

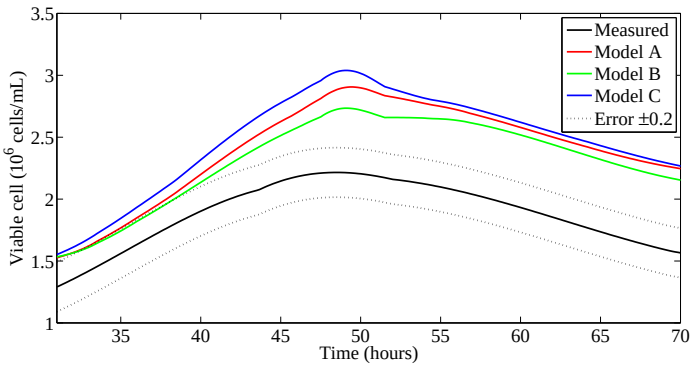
Table 5.2: Model information for the three models constructed to predict the viable cell count, product titre, and final glycosylation. The RMSE and AIC values are reported as mean values for the training and validation batches respectively.

CPP/CQA	Model	X-Block data	LVs	X-block variance captured	Y-block variance captured	RMSE		AIC	
						train mean	val mean	train mean	val mean
Viable cell count	**A**	Operating parameters	3	74.27	48.11	0.28	0.43	11.57	4.73
	B	Glucose + Lactate	3	80.00	73.44	0.24	0.21	14.36	12.12
	C	On-line data	4	81.08	74.56	0.23	0.36	15.20	17.04
Product titre	**D**	Operating parameters	4	80.18	89.50	7.29	6.12	27.30	29.46
	E	Glucose + Lactate	3	84.52	81.66	6.84	9.88	29.77	35.76
	F	On-line data	4	69.47	87.79	6.57	9.57	28.42	35.59
Glycosylation	**G**	Operating parameters	3	76.66	46.16	0.83	1.37	0.87	1.10
	H	Glucose + Lactate	4	93.91	65.36	0.63	1.40	1.90	3.51
	I	On-line data	4	80.16	70.18	0.62	1.31	0.97	3.50

The first model (model A) to predict the viable cell count used the operating parameter set points as the **X**-block, the second model (model B) used the off-line glucose and lactate measurements as the **X**-block, and the third model (model C) used the on-line measurements (DO, O_2, CO_2, pH, and base) as the **X**-block. Table 5.2 shows that model B offers the lowest RMSE value for both the training and the validation batches. Models A and C show relatively low RMSE values for the training batches but higher values for the validation batch. However, the AIC values for the validation batch of model B is significantly higher than for model A. As the AIC is a measure of model complexity it suggests that whilst there is a slight improvement on the prediction made with model B the resulting increase in model complexity would mean model A would be a better choice. The predictions of the three models for the validation batches are shown in Figure 5.9.



(a) Batch 5



(b) Batch 13

Figure 5.9: Measured and predicted values for viable cell count for the two validation batches using three models. Model A (X-block: operating parameters), Model B (X-block: glucose and lactate), Model C (X-block: on-line data). Measurement error (± 0.2) is included for the measured data (black dotted lines).

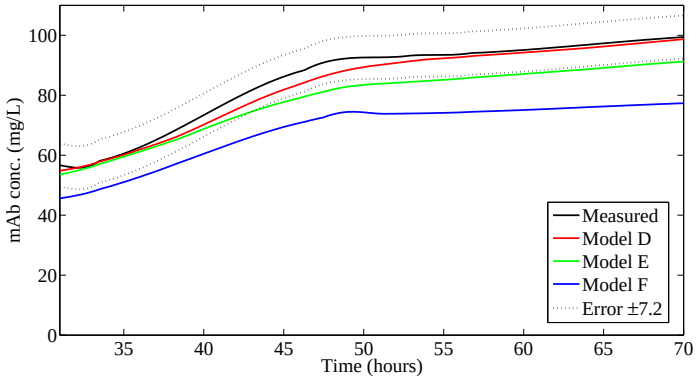
Figure 5.9 shows that model B does indeed predict closest to the measured data. When the measurement error for the viable cell count is included (Ivarsson *et al.*, 2014) it can be seen that only model B remains within the limits for the prediction of batch 5, but is outside the limit for batch 13. The main variation between the measured and predicted data for model B occurs at the cell count maxima, where the prediction is shown to achieve a higher maximum than the measured data. This is likely to be due to variations in the batch data used to train the model.

Figures C11-C22 in the appendix show the predictions for the training runs using each of the models. For ease of interpretation the batches have been split into groups, for example Figure C11 shows the predictions for the DO batches (1-3) for model A. It can be seen for model A for the predictions the greatest error is associated with the maximum growth phase for all batches, whereas for model B the issue is not with a particular phase of the growth curve, but more with individual batches, in particular the training predictions for batch 6 and batch 8 are poor. For model C the issue is again with individual batches, but in this case it is batches 4 and 6 which show a large error.

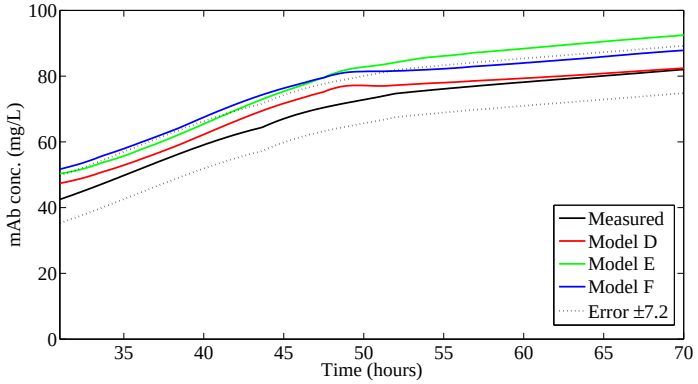
To understand these errors the data used to train the models should again be considered. For model A, the **X**-block contained the set points of the cultivation operating parameters post shift (30 hours). Figure C23 shows the correlation between batch 3 and DO, batch 6 and osmolality, batch 9 and pH, and batch 12 and sparger vvm. These batches were the highest setting for each of the variables i.e. batch 3 is the 90% DO oxygen batch. As the predictions in the appendix for the training runs have shown these four batches to be the best predicted ones, it suggests that the model predicts better for the extremes of the operational parameter settings. For model B the **X**-block contained off-line glucose and lactate concentrations. Figure C25 shows this bi-plot for the model. As can be seen, batches 6 and 8 are closely correlated to the lactate concentration during the shift in cultivation conditions. Observing again the predictions for these graphs it can be seen that the predictions for batches 6 and 8 are very similar. This suggests that the glucose and lactate measurements for these batches were similar. This indicates that for high osmolality (batch 6)

perhaps lactate and glucose are not good indicators of viable cell count as they do not capture the effect of high osmolality (Ozturk *et al.*, 1992). Finally the training predictions of model C are shown in Figures C19-C22. For all batches except 4 and 6 the predictions were good. Batches 4 and 6 are both part of the osmolality group, Figure C27 shows no clear correlation between these two batches and any of the variables used to train the model.

Similarly to the models constructed for viable cell count, three PLS models were constructed using the product titre as the **Y**-block. The first model (model D) used the operating parameter set points as the **X**-block, the second model (model E) used the off-line glucose and lactate measurements as the **X**-block, and the third model (model F) used the on-line measurements (DO, O_2, CO_2, pH, and base) as the **X**-block. The number of latent variables, the variance captured in the **X** and **Y** blocks, the RMSE, and AIC values are reported in Table 5.2 which shows that when considering both the training and predicted RMSE and AIC values, model D provides the lowest values for all of these. Models E and F offer comparatively low RMSE and AIC values for the training batches, but the values associated with the validation batches are much higher. The predictions of the three models are shown in Figure 5.10 demonstrating that model D predicts within the measurement error for both validation batches, and model E predicts within the limits for batch 5 but not for batch 13. Additionally it can be seen that model F does not predict within the limits for either validation batch. This suggests that the input data used for models E and D might be very similar between the batches resulting in the prediction for the validation batch resembling one of the training batches. This effect can also be seen in model D, where the input data was the operational parameter set points which were distinctly different for each batch and the best predictions were obtained. Figures C29-C40 show the predictions made for the training runs for each model. It can be seen that the trend observed in the validation data, of model D predicting best, is also shown here. The training predictions for all three models do not highlights any outliers, i.e. there are no batches which heavily influence the models.



(a) Batch 5



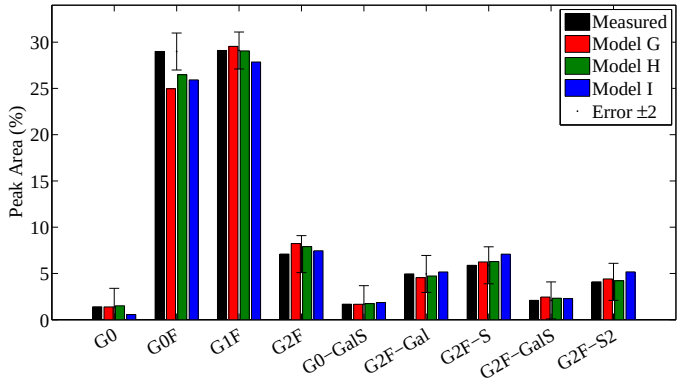
(b) Batch 13

Figure 5.10: Measured and predicted values for titre for three models. Model D (X-block: operating parameters), Model E (X-block: glucose and lactate), Model F (X-block: on-line data).

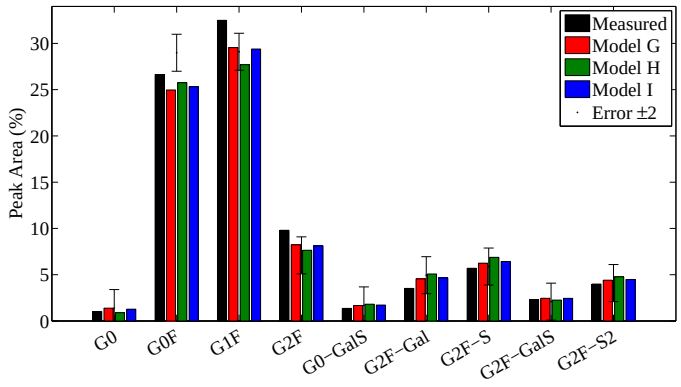
Glycosylation profile of the final product was also used as the **Y**-block. Again three PLS models were constructed, with the first model (model G) using the operating parameter set points as the **X**-block, the second model (model H) using the off-line glucose and lactate measurements as the **X**-block, and the third model (model I) using the on-line measurements (DO, O_2, CO_2, pH, and base) as the **X**-block. The number of latent variables, the variance captured in the **X** and **Y** blocks, the RMSE, and AIC values are reported in Table 5.2 which shows that for all three models the RMSE and AIC values are very similar, with models G and I offering the lowest errors. The predictions of the three models for the validation batches are shown in Figure 5.11.

It is shown that the predictions for both validation batches for 7 of the glycans have a similar accuracy across all three models. Issues predominantly arise in the predictions of two of the glycans G0F and G1F. The predictions for the training data (Figures C41-C52) show the same trend, where the predictions for G0F and G1F are consistently the poorest predictions for all three models. Figures C53-C55 show bi-plots of the **Y**-block data for each model (A-C). In all three models three groups of glycans can be distinguished. The first group contained G0 and G0F, of which G0F is one the most abundantly present glycans. These two glycans represent core-fucosylated, bi-antennary structures with no galactose. The next group contained G1F (most abundant glycan), G2F, and G2F-GalS all of which which were fucosylated and either had one galactose, or were fully galactosylated. The final group contained G2F-Gal, G2F-S2, G2F-S, and G0-GalS which are of a higher complexity to the rest (Ivarsson *et al.*, 2014). In all three models there is shown to be a correlation between the fucosylation of the initial protein (G0F) and the pH. This is supported by Gawlitzek *et al.* (2000) and Pacis *et al.* (2011b) who showed that pH affect the kinetic rates of glycosyltranserases within the Golgi apparatus thereby affecting the glycosylation. These three groups, which appear in all three models, suggest that glycosylation is a step-by-step process, with one step having to occur before another can (Hooker *et al.*, 1995). In all three models batches 8 and 9 (high pH) show a correlation with the glycan G0F. This can be explained through the study presented by Ivarsson *et al.* (2014) who showed that higher pH caused a significant decrease

in terminal galactosylation (glycans G1F and G2F) of which G0F is the starting point. Figures C54 and C55 show that there is a correlation between G1F and G2F with batches 1 and 2. This suggests that the terminal glycosylation is affected not only by variations in pH but also by variations in other operating parameters, suggesting that the operating parameters can be used to control the glycosylation profile (McCracken *et al.*, 2014).



(a) Batch 5



(b) Batch 13

Figure 5.11: Measured and predicted values for glycosylation for three models. Model G (X-block: operating parameters), Model H (X-block: glucose and lactate), Model I (X-block: on-line data).

5.2.4 First principles modelling of cell cultivation

Mathematical modelling has been applied for the characterisation of biological phenomena (Bailey, 1998a), to organise high throughput data (Schilling *et al.*,

1999), and to guide experimentation for applications such as improving media (Gadgil, 2014) or, as in the case of this research, for optimising process operation(Fernandes *et al.*, 2012). The models presented in this section are unstructured in that they do not take into account the inner structure of the cells. The models were primarily used to provide a comparison to the results obtained from the PLS modelling. Two models are described: the first was presented by Naderi *et al.* (2011), and the second by Kontoravdi *et al.* (2007). These two models were used, with model parameter values from literature, to predict viable cell count and product titre. With regards to the glycosylation, there are two reports in literature of mathematical models which can be used to predict glycosylation, as presented by Krambeck and Betenbaugh (2005) and del Val *et al.* (2011). However these models rely upon having accurate values of components inside the cell, and the calculation of various model parameters. With the data sets available for this research it is not possible to apply either of these models.

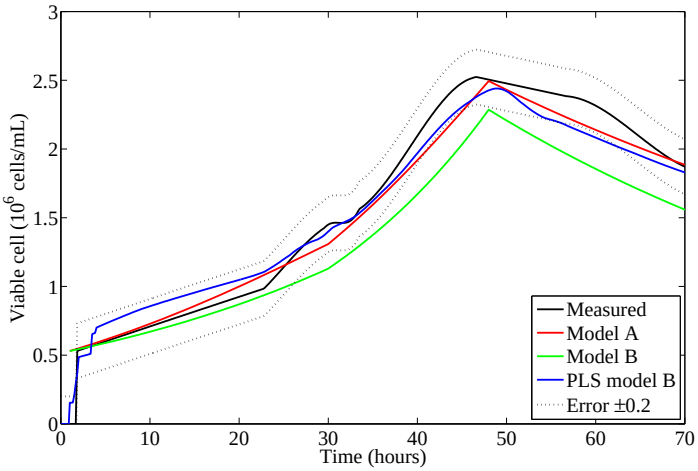
The main limitation encountered during the first principles (FP) model development was the lack of ability to account for varying process conditions. Within the hybridoma data set used in this chapter, the process conditions were varied from batch to batch and it has been shown that these variations have caused a significant change to the outputs of the cultivation. Thus the FP models were used primarily to capture changes to variables such as the metabolites. This presents a situation whereby the data reflects operating changes but the FP models capture variation in media composition. Matlab was used to run the models, with the differential equations being run as functions. An average batch length of 79 hours was used for both models, with the integration for each loop being over a one hour period. The values of model parameters used are provided in Tables B2 and B3, with the starting conditions obtained from the experimental hybridoma data set. As the levels of starting cell count and basic metabolites were uniform for all 13 batches the average for each of these values was used as the starting parameter. The average was taken to account for any minor variations in levels between batches. This resulted in one prediction for each FP model.

Similarly to the approach adopted for the PLS models two FP models were used. Both models predicted cell growth, cell death, consumption and production of the main metabolites, and product titre. The main difference between the two models was the inclusion of all 20 amino acids in the second model. The model prediction accuracy is summarised in Table 5.3.

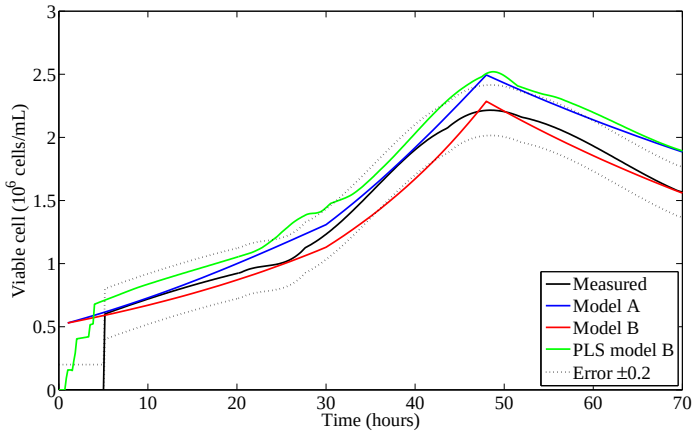
Table 5.3: Model information for two first principles models and the best performing PLS model for viable cell count. The RMSE values are reported; to enable direct comparison with PLS models the values have been listed with regards to the training and validation batches used in section 5.2.3.

CPP	Model	Reference	RMSE	
			Training batches	Validation batches
Viable cell count	A	Naderi *et al.* (2011)	0.49	0.18
	B	Kontoravdi *et al.* (2007)	0.58	0.22
	Best PLS (model B)	Section 5.2.3 Subsection:Viable cell	0.24	0.21
Product titre	A	Naderi *et al.* (2011)	17.56	14.04
	B	Kontoravdi *et al.* (2007)	14.43	10.13
	Best PLS (model D)	Section 5.2.3 Subsection:Titre	7.29	6.12

The first noticeable difference is that the predictions of the FP models vary significantly. The predictions of model A are significantly better and are comparable to the predictions of the best PLS model. This is shown in Figure 5.12, which shows these predictions along with the original measured data for both validation batches (5 and 13). To understand why one FP model is better than the other, the structure of the two models should be considered. Model A uses equations to determine cell count, product titre, and only the basic metabolites (glucose, lactate, glutamine, ammonia, glutamate, asparagine, aspartate, and alanine). Only the basic metabolites are included as the authors stated that these are the main components for the cell metabolism. On the other hand model B included equations for cell count, product titre, glucose, lactate, ammonia, and all 20 amino acids. This makes the second model significantly more complex.



(a) Batch 5

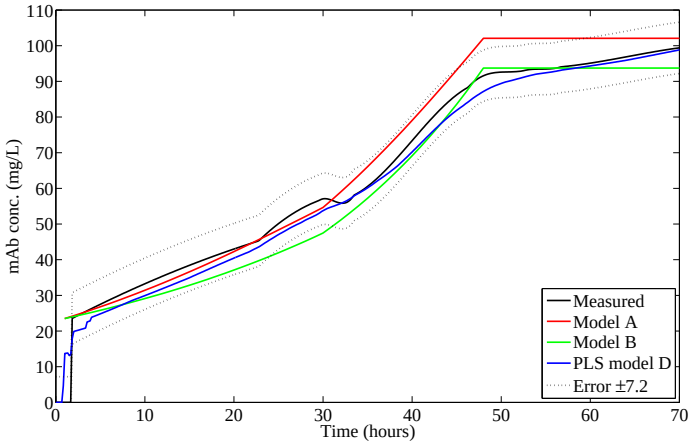


(b) Batch 13

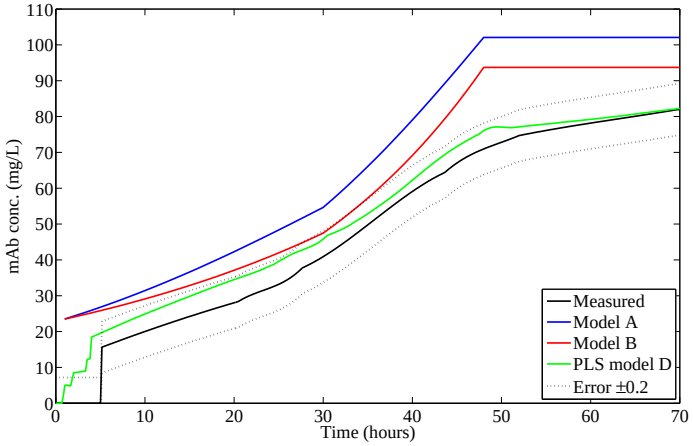
Figure 5.12: Measured and prediction data for viable cell count for validation batches. The predictions shown are the first principles model derived from the Naderi equations (Model A), the first principles model derived from the Kontoravdi equations (Model B), and the best performing PLS model from the previous section (PLS model B).

Considering the predictions shown in Figure 5.12 it can be seen that the predictions of both FP models are quite simplistic. The three main stages of cell growth can be observed (lag, growth, and production), with the peak viable cell count accurately predicted. To explore the large RMSE values the entire data set batches are shown in Figure C56 (showing the raw viable cell count for all 13 batches and model A and B predictions). As can be seen there is a lot of variation between the 13 batches, which results in larger RMSE values.

The models presented by Naderi *et al.* (2011); Kontoravdi *et al.* (2007) were also used to predict the final cell titre, with the results being summarised in Table 5.3. It can be seen that there is a noticeable difference between the RMSE values for both FP models, and that the prediction from the best PLS model is significantly better. This is supported in Figure 5.13, which shows the measured and predicted data for the two validation batches. It can be seen that the PLS prediction made for batch 5 is comparable to the FP model predictions, whereas for batch 13 the PLS prediction is significantly better. This is again showing the influence of the varying operating conditions on the ability of the FP model to predict the batches.



(a) Batch 5



(b) Batch 13

Figure 5.13: Measured and prediction data for product titre for validation batches. The predictions shown are the first principles model derived from the Naderi equations (Model A), the first principles model derived from the Kontoravdi equations (Model B), and the best performing PLS model from the previous section (PLS model D).

The higher RMSE values obtained for the titre predictions over the viable cell predictions can be explained through the comparison of all batches given in Figure C57. As can be seen the output of the product titre varies significantly with the end titre value ranging from $\approx 60 - 100 mg/ml$.

It is easy to understand the benefits of FP mathematical models for the biopharmaceutical industry. Being able to use a model which requires only a small amount of experimentation to make predictions as to the performance of a process unit is advantageous. However this rarely happens in practise. The models presented in this chapter used model parameters from literature. These values were specific for hybridoma cells but not for the specific cell line. To be able to use these models in industry these model parameters would have to be determined and a certain amount of experimentation is required. Once the system specific parameters have been determined the first principals models can be used to predict how changes to the media would impact the cultivation. However this presents an issue in that if the operating conditions are changed then this can significantly impact the ability of the FP model to predict the outcome.

For this data set, where the variation between batches is introduced through operational changes, the FP models fail to capture the variation. The resulting predictions of the FP models can be viewed as being 'blocks' with the first block being the cell lag phase, the second being cell growth, and the final being production. Figure 5.13 has shown that neither of the FP models can handle variation in the operating conditions well. This is because the rates of production and consumption of metabolites and the cell growth rate are assumed to be constant or are calculated. The first model (Naderi *et al.*, 2011) used literature values for the rate of production/consumption of the metabolites and the rate of cell growth which are constant. For the second model (Kontoravdi *et al.*, 2007) the rates of cell growth and production/consumption of metabolites were determined by an equation for each individual rate. This means that the rates vary with every iteration of the model, but are again constrained by the fact that the model cannot handle operational changes. The next section introduces a hybrid model whereby a training data set is used to

predict the rates of production and consumption of the various metabolites as well as the cell growth rate, which are subsequently used in the FP models.

5.2.5 Hybrid modelling

As has been shown in previous sections, there are benefits in applying both the multivariate and the first principles techniques. The multivariate data set captured the dynamic changes in the process conditions showing how varying the conditions impacted on both the CPPs and CQAs. This shows how changes to the final product titre and glycosylation profile can be captured. In contrast, the first principles models allowed for predictions with significantly less input data required. Additionally the first principles models can predict for parameters which are not necessarily measured in the cultivation such as amino acid concentrations. The models presented in this section are a combination of both of these techniques with the aim of combining the benefits of both. The model structure used both the first principles models was presented in the previous section (as given by Naderi *et al.* (2011) and Kontoravdi *et al.* (2007)), with multivariate modelling predicting model parameters required for these equations.

In the previous section the first principles models relied heavily upon values from literature for the rates of production/consumption of metabolites, and the cell growth rate constant. In this section PLS was used to predict these values, to take into account the variation between batches introduced by different operating conditions. Additionally as the rates were predicted using PLS, this should allow for the variation over time to be accounted for. There are hybrid models presented in literature which aim to achieve the same purpose, such as Galvanauskas *et al.* (2004) who used artificial neural networks for the prediction of parameters in first principles models, or von Stosch *et al.* (2011) and von Stosch *et al.* (2014a) who used PLS models for the same purpose. The work presented in this section aims to build upon the knowledge of hybrid modelling by using a more complex first principles model, and predicting not only titre, viable cell count, and main metabolites, but also other metabolites such as all amino acids. This additional

characterisation of the cultivation can lead to better predictions of the glycosylation profile, as the predictions of the profile can be made from multiple sources and combined.

Methodology

The basic structure of the hybrid models used in this section is shown in Figure 5.14, in that PLS models are used to predict the rates and these are then fed into a system of ordinary differential equations (as used in section 5.2.4). The PLS models were constructed for each individual metabolite, and predicted independently. Additionally the cell growth rate was also predicted in a similar manner. The rates were predicted at 445 time points during the cultivations as this was the sampled data matrix produced in section 5.1. Having trained the PLS models each validation batch was predicted separately, and these predicted values were then used in the differential equations to obtain a prediction for the cultivation.

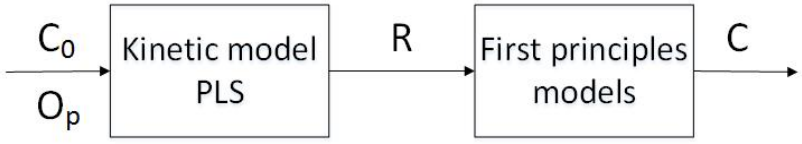


Figure 5.14: Basic hybrid model construction: the backbone of the model is a mass balance, which is represented by ordinary differential equations and is given by Naderi *et al.* (2011) and Kontoravdi *et al.* (2007). PLS models are used to predict the rates. Where C_0 is the initial concentration of metabolites, cell count, and product titre; O_p is the operating parameter set points; R is the rates of production and consumption of metabolites and the cell growth rate; and C is the time series data for metabolites, viable cell count, and product titre.

Two PLS models were constructed, the first using the on-line data as the **X**-block. The second used the operational parameter set points as the **X**-block. The operational parameter set points were chosen as the inputs to the model because they are factors which the operator can directly influence. Thus as this is ultimately a tool for use in industry it has to meet the needs of the person using the model. Initially the results of the hybrid models were compared against each other, with comparisons to the models produced using just PLS

Table 5.4: Hybrid model key, specifying how each model was constructed.

Model key	Model construction
Hybrid model A	Model constructed using the on-line data as the PLS X-block, and using the Kontoravdi ODE system.
Hybrid model B	Model constructed using the operational parameter set point as the PLS X-block, and using the Kontoravdi ODE system.
Hybrid model C	Model constructed using the on-line data as the PLS X-block, and using the Naderi ODE system.
Hybrid model D	Model constructed using the operational parameter set point as the PLS X-block, and using the Naderi ODE system.

and first principles presented in the discussion section (section 5.2.5). Matlab was used to run the models, with the differential equations being run as functions. An average batch length of 79 hours was used for both models, with the integration for each loop being in steps of the 445 sample points. A more detailed flow chart of the hybrid model is given in Figure 5.15, showing the flow of work within the model network. The following sections present the predictions, with the key for the models given in Table 5.4.

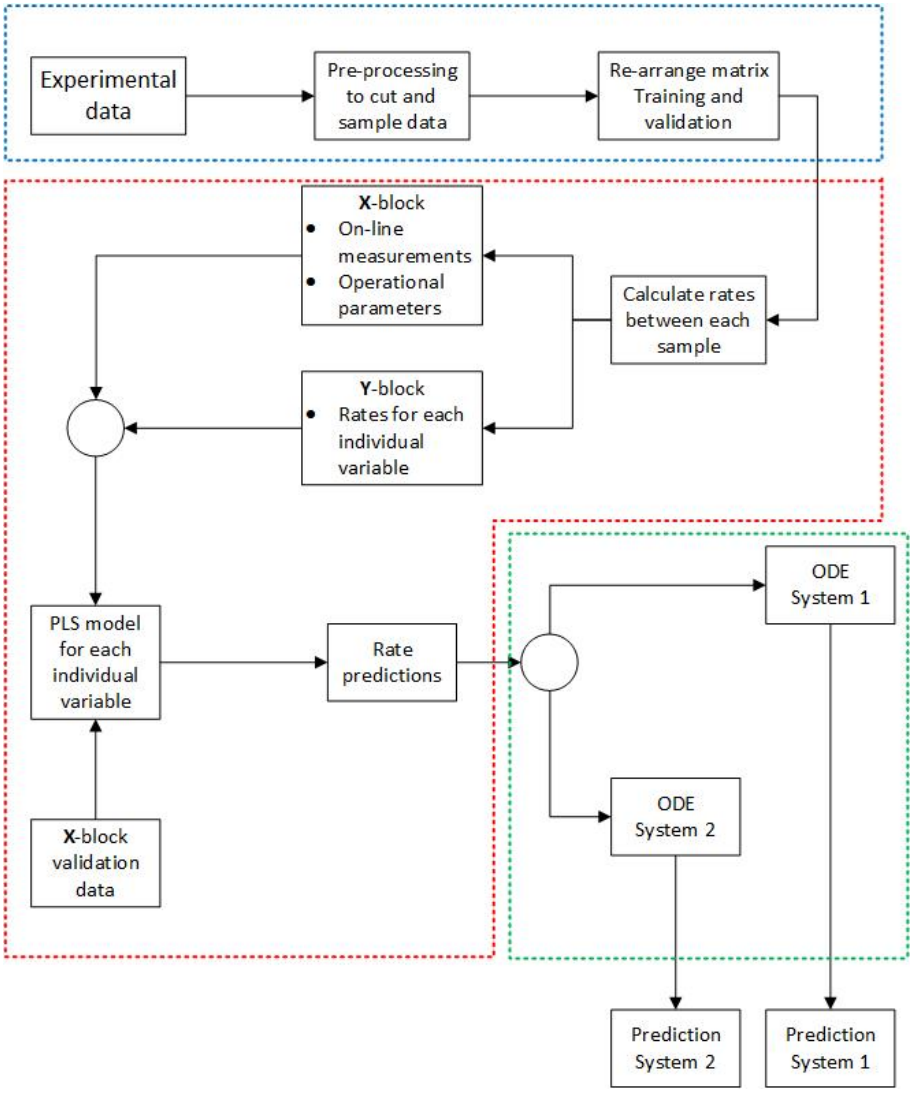


Figure 5.15: Model network, showing the flow of information. The blue box shows the data pre-processing and manipulation, the red box shows the PLS model, and the green box shows the first principles ODE equations. Four predictions were made from the system, relating to the two first principles models used with two different multivariate models predicting the input data.

Four predictions were made from different models. The aim of the hybrid model was to investigate whether it could more accurately predict not only the start and end cell count but also the inflections seen during the cultivation. This includes the maximum cell count and the points at which cell growth and cell death accelerate. The models produced using only PLS or first principles struggled with predicting these changes accurately.

The predictions made for the validation batches (batches 5 and 13) are shown in Figure 5.16 with the model assessment criteria reported in Table 5.5.

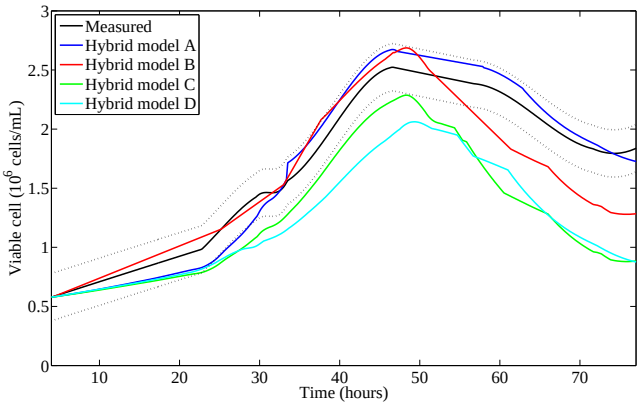
Table 5.5: RMSE values for four hybrid models, as given in Table 5.4, for viable cell predictions shown in Figure 5.16.

CPP	Model	Details and reference	RMSE	
			Batch 5	Batch 13
Viable cell count	Hybrid model A	On-line **X**-block Kontoravdi *et al.* (2007)	0.33	0.32
	Hybrid model B	Operational parameter **X**-block Kontoravdi *et al.* (2007)	0.26	0.17
	Hybrid model C	On-line **X**-block Naderi *et al.* (2011)	0.49	0.51
	Hybrid model D	Operational parameter **X**-block Naderi *et al.* (2011)	0.51	0.28
Titre	Hybrid model A	On-line **X**-block Kontoravdi *et al.* (2007)	15.72	19.84
	Hybrid model B	Operational parameter **X**-block Kontoravdi *et al.* (2007)	11.56	9.96
	Hybrid model C	On-line **X**-block Naderi *et al.* (2011)	11.56	10.55
	Hybrid model D	Operational parameter **X**-block Naderi *et al.* (2011)	12.56	9.48

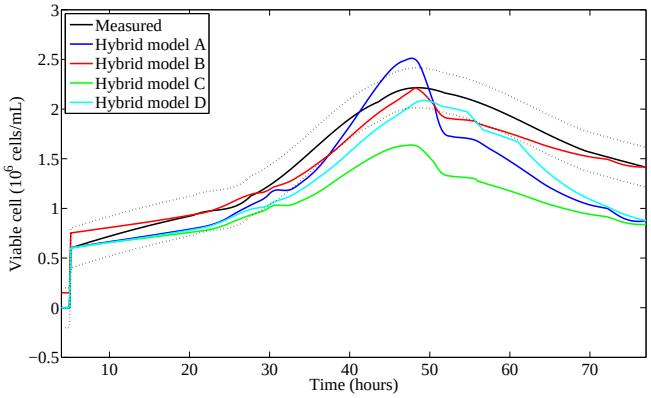
As with the first principles only models in section 5.2.4, there is a distinct difference in the predictions from both ODE systems. The predictions using the Kontoravdi *et al.* (2007) ODEs are marginally better than those made using the Naderi *et al.* (2011) ODEs. The results of the predictions suggest that as the more complex model which includes all 20 amino acids (Kontoravdi *et al.*, 2007) predicts the viable cell count better some of the amino acids included in the model have more significance on the system than is suggested by Naderi *et al.* (2011). Additionally it can be seen that the operational parameter set point **X**-block has also produced lower RMSE values for both validation batches. This suggests that using the on-line data (which contains significantly

more information) does not produce models which are more accurate and therefore to simplify the hybrid model the rate predictions made using the operational parameter set points is best.

Figure 5.16 shows the three main stages of cell growth can be observed (lag, growth, and production), for all four models. It can be seen that the greatest error is introduced after the maximum cell count has been reached. This is most likely due to the variation introduced in the data used to train the PLS models, as certain batches vary greatly (batch 9, pH8). Additionally, hybrid models 1 and 3 show a profile which is not smooth, which can be attributed to the calculation of the rates. The rate was calculated between each sample point from the on-line data. On the other hand the rates calculated using the operational parameters produce a much smoother profile because only one value is used to predict the rate at all sample points. In summary for the viable cell count the model which produced the most accurate predictions was *hybrid model 2*.



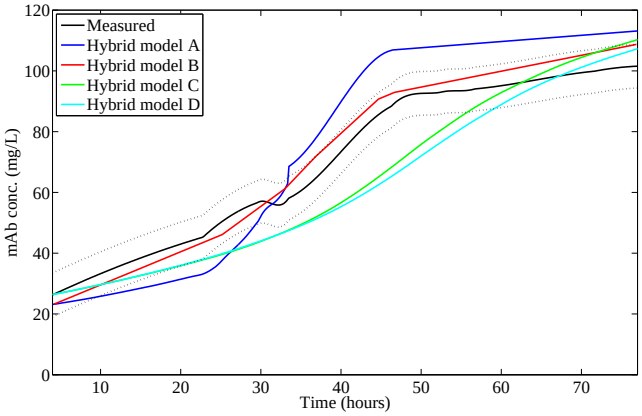
(a) Batch 5



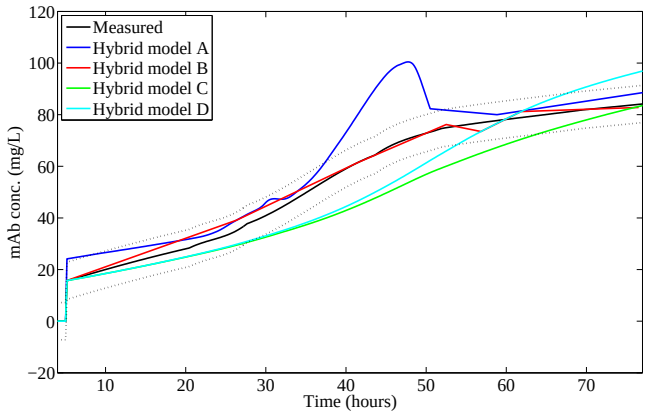
(b) Batch 13

Figure 5.16: Predictions for viable cell count from four hybrid models. Hybrid model one uses on-line X-block data and Kontoravdi *et al.* (2007) ODEs, hybrid model two uses operational parameter X-block data and Kontoravdi *et al.* (2007) ODEs, hybrid model three uses on-line X-block data and Naderi *et al.* (2011) ODEs, and hybrid model four uses operational parameter X-block data and Naderi *et al.* (2011) ODEs. The corresponding model assessment values are reported in Table 5.5.

The models given by Naderi *et al.* (2011); Kontoravdi *et al.* (2007) were also used to predict the product titre, with the results being summarised in Table 5.5. It can be seen (Figure 5.17) that there is a distinct difference between the models constructed using the two ODE systems. For both validation batches it can be seen that hybrid models 1 and 2 over-predict the titre, whilst hybrid models 3 and 4 under-predict the titre. This is because both models directly link viable cell count to product titre; models 1 and 2 predicted higher values for the viable cell count and models 3 and 4 predicted lower values. The viable cell count prediction based on the Naderi *et al.* (2011) ODE system showed a profile which was not smooth, however the titre predictions are smooth. This suggests that although the product titre is linked to the viable cell count it is the predicted rate of production of protein which is the critical component. The predictions of titre based on the Kontoravdi *et al.* (2007) ODE system are similar to those made for the viable cell count in that the profile is not smooth. This is particularly evident for model A, where an anomalous peak shown in the viable cell count is reflected in the product titre. In summary the RMSE values and Figure 5.17 suggest that the best model is *hybrid model 2*.



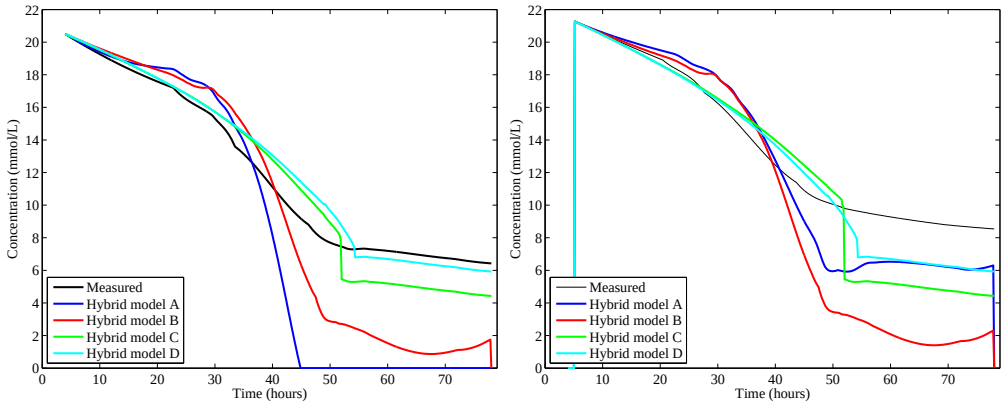
(a) Batch 5



(b) Batch 13

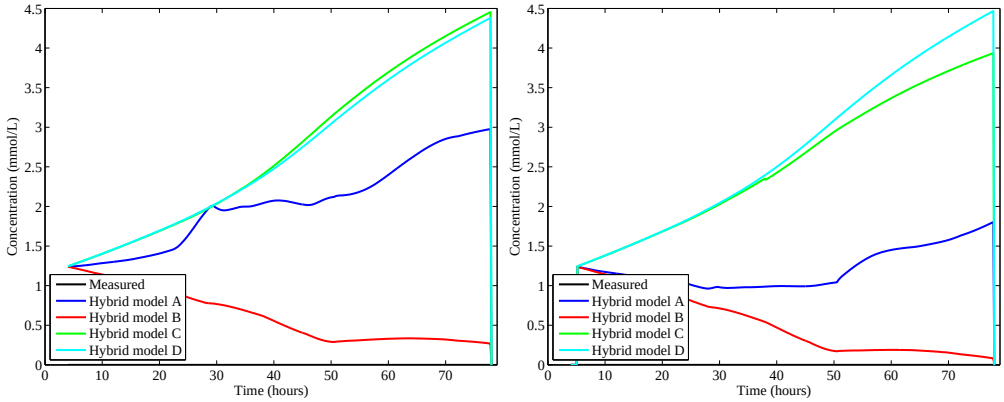
Figure 5.17: Predictions for product titre using four hybrid models. Hybrid model one uses on-line X-block data and Kontoravdi *et al.* (2007) ODEs, hybrid model two uses operational parameter X-block data and Kontoravdi *et al.* (2007) ODEs, hybrid model three uses on-line X-block data and Naderi *et al.* (2011) ODEs, and hybrid model four uses operational parameter X-block data and Naderi *et al.* (2011) ODEs. The corresponding model assessment values are reported in Table 5.5.

The hybrid model also allows for the prediction of metabolites. Using the hybridoma data set, which contains concentrations of amino acids and base metabolites (glucose, lactate, and ammonia) predictions can be made for the validation batches. These predictions are shown in Figure 5.18 with the remaining predictions shown in Figures C56 and C55 in appendix C.



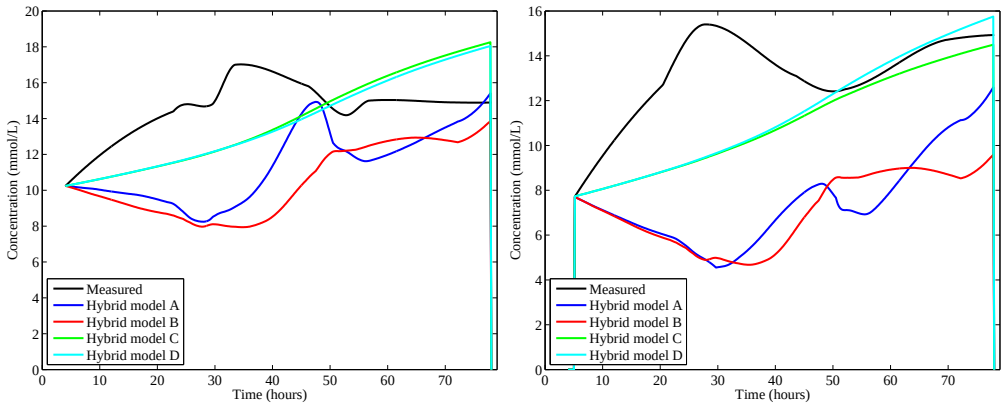
(a) Glucose - Batch 5

(b) Glucose - Batch 13



(c) Ammonia - Batch 5

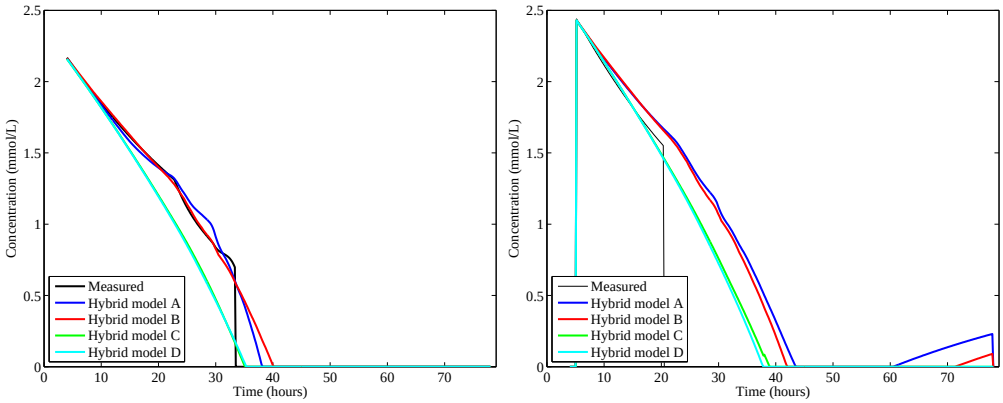
(d) Ammonia - Batch 13



(e) Lactate - Batch 5

(f) Lactate - Batch 13

Considering first the predictions for glucose (Figure 5.18 (a) and (b)) it can be seen that a general trend is observed in that there was a sharp initial decline which levels out. This can be seen for all four models, with hybrid



(g) Glutamine - Batch 5 **(h) Glutamine - Batch 13**

Figure 5.18: Predictions for glucose, lactate, and glutamine from four hybrid models for batches 5 and 13. Hybrid model one uses on-line X-block data and Kontoravdi *et al.* (2007) ODEs, hybrid model two uses operational parameter X-block data and Kontoravdi *et al.* (2007) ODEs, hybrid model three uses on-line X-block data and Naderi *et al.* (2011) ODEs, and hybrid model four uses operational parameter X-block data and Naderi *et al.* (2011) ODEs. The corresponding model assessment values are reported in Table 5.5.

model 4 providing the best prediction. For ammonia predictions (Figure 5.18 (c) and (d)), there was no measured data with which to compare the validation batches. The ammonia model was trained from the two batches which contained measurements. As the training set was so small this is likely the cause for variation in the predictions of the four hybrid models. As can be seen hybrid model 2 did not predict ammonia concentration accurately, even though there was no measured data to compare to this, it can be said with confidence because ammonia is produced as a by product in the cultivation so the concentration should rise with time. The final lactate concentration (Figure 5.18 (e) and (f)) was most accurately predicted with hybrid models 1 and 2, however it can be seen that the time series profile of the lactate concentration did not follow the same trend as the measured data. This difference is likely caused by the models used, which are not specific to this cell line. The work presented in this chapter is a precursor to the agent based model for CHO cell products, and as such the equations used were for CHO cells. It is possible to use these models for hybridoma production as they were adapted from equations for hybridoma cells. Therefore the issue observed with the lactate predictions is likely due to a variation in this equation which adapts it for use

with CHO cell metabolism. Therefore, to be able to comment on the applicability of the models it is other factors which are important, such as accurate end predictions, and realistic cultivation profile. By this it is meant that the prediction reaches an end value which is approximately the measured end value, the profile is approximately the same as the measured. The model applicability was better tested with the CHO cell data (Chapter 7). The final main metabolite is glutamine (Figure 5.18 (g) and (h)), for which all four hybrid models provided accurate predictions, showing all glutamine to be consumed by $\approx 30 hours$. As Jeong and Wang (1995) discuss, glutamine plays an important role in stimulating the growth of the cells. They showed that the specific growth rate is a strong function of glutamine. Additionally glutamine is important in mAb production, hence why the glutamine is consumed at the beginning (cell growth) and the consumption rate increases (cell growth and mAb production).

The FP model presented by Naderi *et al.* (2011) takes into account asparagine, aspartic acid, and glutamic acid and the model presented by Kontoravdi *et al.* (2007) takes into account all 20 amino acids. Focusing first on the predictions made for asparagine, aspartic acid, and glutamic acid it can be seen that hybrid models 3 and 4 provide a much smoother profile and for all three provide the most accurate prediction. For the other amino acids, which are only predicted using hybrid models A and B, it can be seen that the predictions are most accurate, and show the greatest accord between models prior to the shift change at 30 hours, with the greatest variation from the measured data being after the 30 hour point. This shows that the models are very easily influenced by the variation in the operational parameters captured by the rates. For all amino acids the concentration is very low, for example histidine (Figure C56 (h)) concentration is approximately 1.1 mmol/L for the duration of the cultivation. The predictions of both hybrid models A and B give the concentration as being approximately 0.7 mmol/L for the cultivation duration. The difference between these two is relatively small, hence the predictions can be classed as good.

The RMSE values for all metabolites are reported in Table 5.6. To

Table 5.6: Model assessment values for four hybrid models, as given in Table 5.4, for viable cell predictions shown in Figures C56 and C55 (beginning on page 333. – denotes that RMSE was not calculated as model does not include this metabolite.

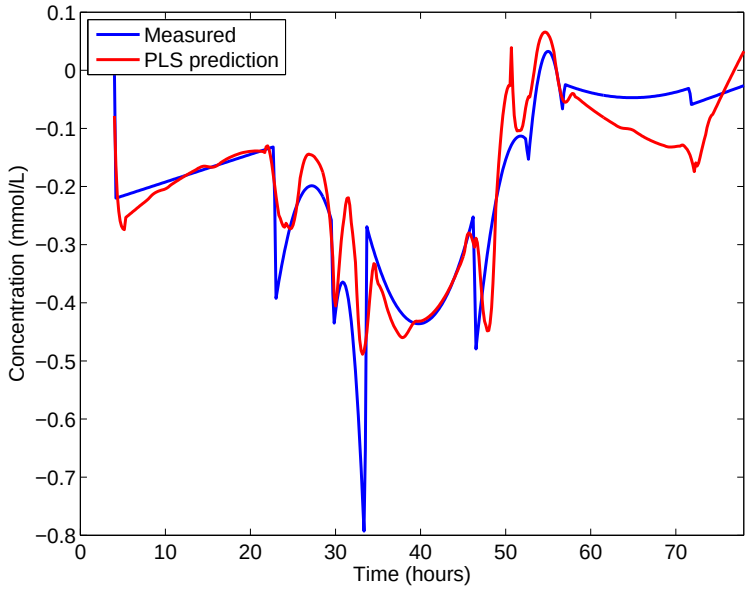
	RMSE							
	Batch 5				Batch 13			
Hybrid model number	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Glucose	5.15	3.62	3.25	2.95	2.12	4.64	2.52	1.55
Ammonia	**	**	**	**	**	**	**	**
Lactate	4.02	4.92	2.55	2.51	6.22	6.58	2.75	2.64
Alanine	0.20	0.20	–	–	0.26	0.18	–	–
Arginine	0.10	0.07	–	–	0.09	0.06	–	–
Asparagine	0.08	0.04	0.04	0.04	0.02	0.03	0.05	0.06
Aspartic acid	0.06	0.04	0.02	0.02	0.02	0.03	0.06	0.06
Cysteine	5.15	3.62	–	–	2.12	4.64	–	–
Glutamine	0.09	0.10	0.14	0.15	0.59	0.56	0.44	0.43
Glutamic acid	0.10	0.08	0.07	0.07	0.08	0.10	0.11	0.10
Glycine	0.51	0.51	–	–	0.73	0.70	–	–
Histidine	0.53	0.49	–	–	0.59	0.60	–	–
Isoleucine	1.08	1.13	–	–	1.61	1.43	–	–
Leucine	0.48	0.34	–	–	0.16	0.24	–	–
Lysine	1.37	1.31	–	–	1.45	1.47	–	–
Methionine	0.12	0.14	–	–	0.18	0.16	–	–
Phenylalanine	0.07	0.07	–	–	0.04	0.05	–	–
Proline	0.502	0.43	–	–	0.25	0.17	–	–
Serine	0.40	0.25	–	–	0.38	0.40	–	–
Threonine	0.27	0.26	–	–	0.29	0.30	–	–
Tryptophan	0.66	0.63	–	–	0.71	0.72	–	–
Tyrosine	0.50	0.41	–	–	0.22	0.41	–	–
Valine	0.47	0.41	–	–	0.41	0.35	–	–

**** RMSE not calculated as there was no measured data to compare too.**

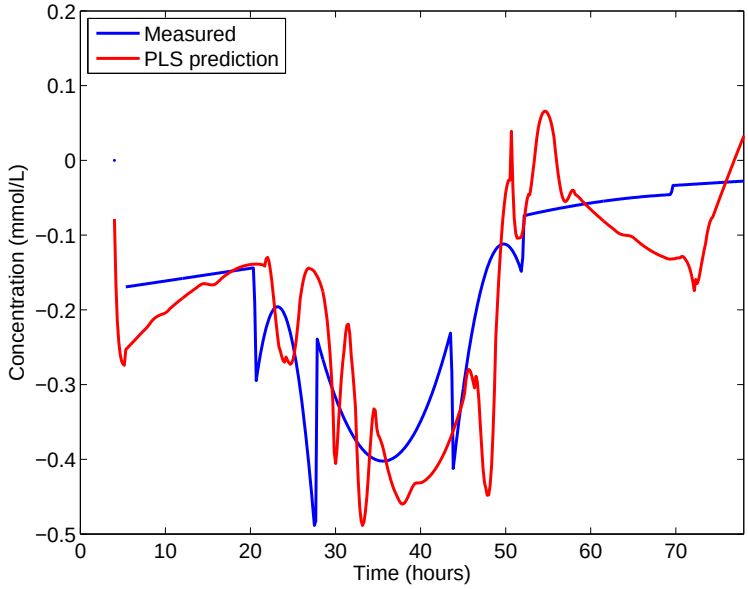
determine which model is best is difficult as the values are comparable for most of the metabolites, therefore the best model is determined as the one which provides the best prediction of the main metabolites. The model which predicts glucose, lactate, ammonia, and glutamine best is *hybrid model 4*. This model used the operational parameter set points to predict the rates and the FP model as presented by Naderi *et al.* (2011).

For all the PLS models used to predict the rates 5 latent variables were used. This value was chosen as it captured the maximum variation in the data without including noise. The eigenvalues and RMSECV values were checked

for all models and for this data set 5 LVs was determined to be the correct amount. As the data used in the **X** and **Y** blocks in each model was very similar i.e. the **X**-block were the same, and the **Y**-block was always a rate, it means the model structure is similar. The data for the rates was preprocessed using a Savgol smoothing (Savitzky and Golay, 1964) filter with a window of 15, and autoscale. The predictions of the rate of glucose consumption are shown in Figure 5.19. It can be seen for batch 5 that the main features were preserved. For batch 13 it can be seen there was a slight delay in the predictions.



(a) Batch 5



(b) Batch 13

Figure 5.19: Predictions for glucose consumption rate for batches 5 and 13. For the model which used on-line measurements as the X-block data.

The best predictions from the four hybrid models were shown to be using the models which use the operation parameter set points as the **X**-block for the rate predictions. As can be seen from Figure 5.19 there are features within the rate data, such as the sharp peaks, and the sudden changes in the rate (e.g. at $\approx 32 hours$). When the on-line data was used as the **X**-block, the PLS model captured the variation within both blocks of data and tended to over-predict the features in the validation batches predictions. However when the operational parameter set points were used as the **X**-block data, this relied only upon the features contained in the **Y**-block rates used to train the models. This produced smoother predictions and did not introduce as much noise into the data set.

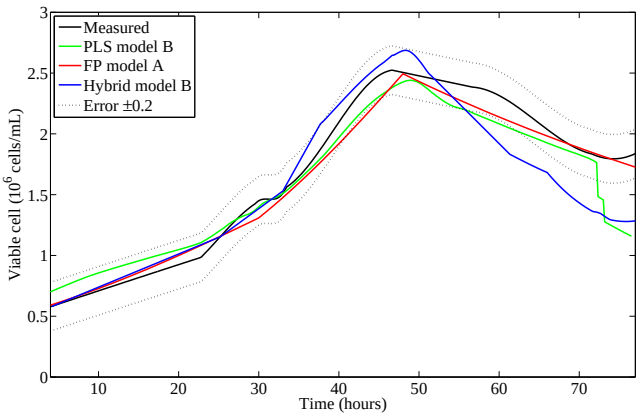
The predictions of the best PLS, first principles, and hybrid models are shown in Figures 5.20 and 5.21 with Table 5.7 summarising the RMSE values for each model.

Table 5.7: Model information for two first principles models and the best performing PLS model for product titre. The RMSE values are reported; to enable direct comparison with PLS models the values have been listed with regards to the training and validation batches used in section 5.2.3.

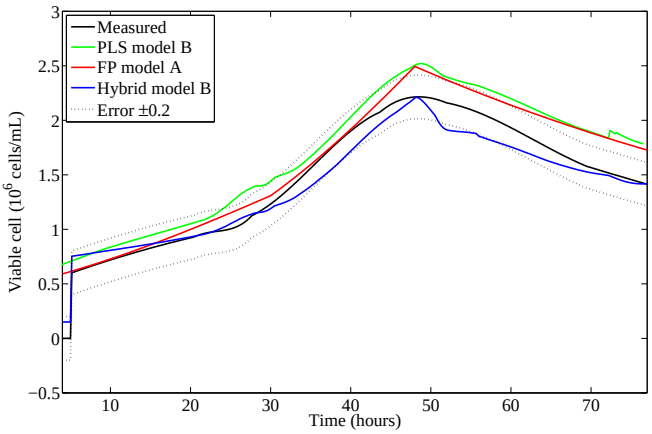
CPP	Model	Reference	RMSE	
			Batch 5	Batch 13
Viable cell count	PLS model B	Section 5.2.3 **X**-block: glucose and lactate	0.24	0.21
	FP model A	Naderi *et al.* (2011)	0.49	0.18
	Hybrid model B	PLS: operational parameters Kontoravdi *et al.* (2007)	0.26	0.17
Product titre	PLS model D	Section 5.2.3 **X**-block: operational parameter set points	7.29	6.12
	FP model A	Naderi *et al.* (2011)	17.56	14.04
	Hybrid model B	PLS: operational parameters Kontoravdi *et al.* (2007)	11.56	9.96

It can be seen that for both the viable cell count and the titre the worst predictions were provided using the first principles models. For the viable cell count the hybrid model was the best (clearly illustrated in Figure 5.20 (b)). However for product titre, the PLS only model produced a more accurate prediction. The ODEs used to determine both the viable cell count and the product titre are linked in that the product titre is a function of the viable cell

count. To determine the titre, the viable cell count was multiplied by the production factor. This is a simple relationship and states that every viable cell will produce the same amount of product. In the case shown in Figures 5.20 and 5.21, the relatively accurate viable cell count but the poorer titre prediction for the hybrid model suggests that this relationship is too simple for the model. It suggests that instead the relationship is dynamic, and as with the rates of metabolites, the rates of production should be calculated at various time points. This is supported by Figure 5.21(a) which shows the titre is initially under-predicted and then over-predicted.

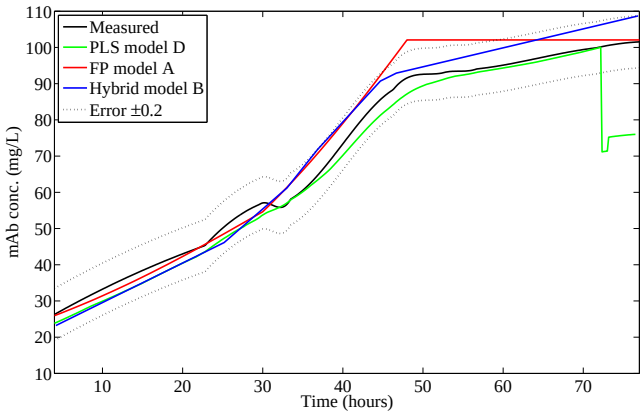


(a) Batch 5

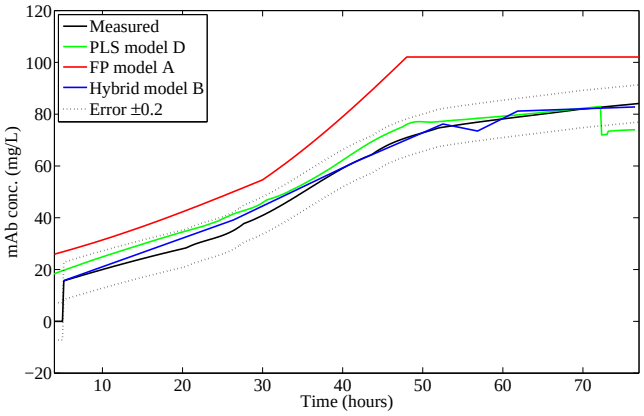


(b) Batch 13

Figure 5.20: Predictions for the best PLS (model B), first principles (Naderi *et al.*, 2011), and hybrid models (operational parameter set points PLS, (Kontoravdi *et al.*, 2007) FP model).



(a) Batch 5



(b) Batch 13

Figure 5.21: Predictions for the best PLS (model D), first principles (Naderi *et al.*, 2011), and hybrid models (operational parameter set points PLS, (Kontoravdi *et al.*, 2007) FP model.)

5.3 Conclusions

There were three main aims of this study. The first was to analyse the interactions in the hybridoma data set, the second was to study glycosylation of the final product mAb to determine if any conclusions could be drawn about the formation of glycans, and to explore if it was possible to model final glycosylation profile. The final aim was to assess the different modelling and analysis techniques to ensure the appropriate methods were chosen to meet the needs of this research and the industrial partner.

The PLS models presented in this study have shown that for this data set the best models are achieved with the removal of stirrer speed and temperature. This would have to be reassessed for subsequent data sets. The cultivation is highly dependent on the conditions it is operated at, with small changes to operating variables producing large changes in the cultivation variables profile (illustrated in Figure C13 in appendix C). This can be clearly seen with run 9, which increased to pH9, and resulted in cell death. PLS models were used to predict viable cell count, product titre, and glycosylation profile relatively accurately. The first principles models presented in this chapter have shown that when there is little available data on a system, it is possible to model it and obtain fairly accurate predictions. In this way it could be said that first principles models would be appropriate to use at the beginning of a new study to focus the experimentation subsequently performed. However, this study has also shown that statistical modelling may be used with first principles modelling to achieve a better prediction through a hybrid network. The statistical models can be used to predict rate of consumption/production of amino acids and take into account cultivation changes, and the first principles models can be used to predict the concentration levels. The hybrid models are able to predict both final product titre and to account for changes throughout the production (where rate of production changes). This shows that the PLS aspect of the hybrid model is able to incorporate changes from conditions (which cause the rate changes) to produce more accurate responses with the first principles models accounting

for the relationship between metabolites in the cultivation.

The PCA analysis of the final product glycosylation profile suggests that for the glycans measured, the mechanism of fucosylation is very important. This is the initial glycosylation mechanism which attaches a fucose molecule to the protein. As has been shown in both the analysis and the regression models, the greatest variation is with G0F, G1F, and G2F. This variation suggests fucosylation is sensitive to changes in culture conditions. The other glycans which are formed through the addition of more molecules to one of G0, G0F, G1F or G2F have a much lower error associated with the prediction. This indicates that the mechanism for attaching these further sugar molecules is independent of the conditions. Due to the lack of appropriate first principles models for glycosylation, as the main mechanisms behind glycosylation control are not yet fully understood, the hybrid model did not take into account prediction of the glycosylation profile. The ability of the PLS model to predict the glycosylation profile is satisfactory for further use in this research.

The final important aspect of the study presented in this chapter was to assess the applicability of the different techniques for use in subsequent work. PARAFAC was investigated as literature showed that it is both easy to apply and interpret (Bro, 1997). This is an advantage in the work presented in this thesis as it is to be used by the industrial sponsor and creating a model which can be applied by any operator without specific training would be beneficial. However to achieve the goals of the project PARAFAC was not adequate. It has been shown that it can capture variation in the data, however it is difficult to determine the causes of variation without prior knowledge of the system. Furthermore, interactions between the different variables were not captured, PARAFAC places greater importance on the variables which cause variation in the data set as a whole. In contrast to this PCA has been shown to be an effective analysis technique. Some of the issues highlighted by literature (Rajalahti and Kvalheim, 2011a), such as ensuring correct data pre-processing is carried out in industry, can be overcome by script formation and development of functions which carry out the pre-processing for the operator. PCA can clearly identify which variables are important for the analysis and

which are just generating model noise. Additionally the influence of variables on batch variation can be easily determined. The similarities between PCA and PLS (as described in chapter 4) mean that they can be easily used together, with conclusions from PCA informing the construction of PLS models. Furthermore the wide application and adoption of PLS to all areas of industry make it the obvious choice for this research as it can easily be used within the industrial sponsor company. The use of the first principles models in the hybrid network is beneficial for industry where one cell line is used as the main producer. This means that the same model structure can theoretically be applied to different systems where CHO cells are the expression system. This then just requires the inclusion of a data set relevant to the specific mAb to train the PLS models for the estimation of the relevant rate values.

5.4 Summary

Over the last ten years since the introduction of the PAT guidelines by the FDA, the biopharmaceutical industry has placed a greater importance on the identification of critical quality attributes and critical process parameters. There have been applications of PAT and QbD to bacterial expression system cultivations (Gnoth *et al.*, 2007; Carrondo *et al.*, 2012; Mercier *et al.*, 2013), and developments to mammalian cell lines and culture media. The research presented in this study has shown that the MVDA techniques used by Mercier *et al.* (2013), Kirdar *et al.* (2008a), Glassey *et al.* (2011b), and Teixeira *et al.* (2009a) can be successfully applied to hybridoma cell data sets, which include glycosylation information. Of the various MVDA techniques available, this study has focused on PCA and PLS as these are already widely used in industry. There are other analysis and regression techniques which could be used in this scenario such as O-PLS (orthogonal PLS) or NL-PLS (non-linear PLS) which could also be applied but as this work is for transfer to the industrial partner PLS was used. PCA has been highlighted as an appropriate tool for mammalian cells as it not only assesses the variables but also the interactions between the variables. This captures information which it is not possible to capture with standard measurement techniques. For example cell

metabolism, is not easily measured on-line as it is a system of processes within the cell and is dependent on many factors. PCA has been shown to feed into PLS, with the identification and removal of unimportant factors in PCA yielding better results in the PLS regression models. This study has also shown how first principles models can be applied to model systems where there is little information available. This is useful for early stage development of a process, meaning that the models can guide subsequent experimentation thereby reducing development costs. These PLS and first principles models can be applied through a hybrid network to improve predictions of the system. The PLS models effectively capture variation in process conditions whilst the first principles models go some way towards capturing aspects of the cell metabolism such as the production and consumption of metabolites. The conclusions in this chapter have highlighted the possibility of using a hybrid modelling technique to improve the performance of the predictions for the cultivation stage, chapter 6 investigates the application of a similar model network for ion exchange chromatography.

Chapter 6

Modelling of ion exchange chromatography

As discussed in Chapter 3 optimising downstream processes can have a significant impact on manufacturing costs. This can be achieved in a number of ways, such as decreasing the number of process steps, avoiding complex steps, and reducing the raw materials costs. Shukla and Thömmes (2010b) show that many companies when faced with a new impurity challenge just select from a set of standard chromatographic platforms. In the purification process many criteria have to be met not only for cost effectiveness but also to meet stringent purity specifications. Particularly challenging in this context can be optimising a process within the design space allowed by the regulating bodies. Ideally ion exchange (IEX) modelling of mAbs would be considered in this chapter, however, due to limitations of available data the purification of a single component system containing the protein lactoferrin (LF) is used. This chapter therefore deals with downstream processing (DSP) modelling.

6.1 Research aim

Various methods for modelling of IEX chromatography, such as the multi-component first principles model presented by Gu (1995) or the multivariate control model presented by Laursen *et al.* (2010a, 2011), are

described in literature. All of these techniques have in common a need for data. In the case of first principles models data is required to determine the constants such as the equilibrium constant, and for multivariate models data is required to train the models. This chapter presents a structured approach to modelling chromatographic data sets which can be used to establish a design space.

Starting point

- data set of chromatographic data: for the purification of lactoferrin on an SP Sepharose Fast Flow resin
- single component system with known physiochemical properties (charge and hydrophobicity)
- specifications for desired protein in terms of recovery level

Research aims

- to determine operating conditions from a predetermined final yield
- to develop a model which can be used within the ABM framework
- to produce predicted chromatogram of a system based upon changes to the inputs of the model

The main challenge in the modelling will be handling a data set which varies greatly. The following sections discus background information, raw data collection, interpretation and the methodology employed to model the system.

6.2 Background information

This section aims to provide additional information not covered in Chapter 2 which is necessary for the understanding of the research presented in this Chapter.

6.2.1 Protein

Lactoferrin (LF) belongs to the group of transferrin proteins. These are proteins which are involved in the transportation of iron in the body. LF is a

globular protein (as opposed to being fibrous, disordered or membrane), and has a molecular mass of approximately 80 kDa. LF is naturally occurring in various secretary fluids such as milk, saliva, and tears but it can also be produced using recombinant technology (Van Berkel *et al.*, 2002; Adrio and Demain, 2003). LF consists of one polypeptide chain containing approximately 700 amino acids. Figure 6.1 shows that it contains two homologous domains, called the N- and C- lobes, with the two domains being connected by a short section of α-helix (Spik *et al.*, 1994; Baker and Baker, 2005). As can be seen in Figure 6.1 both of the lobes contain two sub-domains known as the N1, N2 and C1, C2 respectively. Additionally each of the lobes contains one iron binding site and one glycosylation site. The degree of glycosylation of the protein can be different which is why the molecular weight of the molecule can vary from 76 to 80 kDa (Hakansson *et al.*, 1995). LF belongs to the group of proteins known as basic proteins, which have a high isoelectric point (pI) and therefore they tend to be positively charged at physiological pH (7.4) with the isoelectric point of LF reported as 8.7 (Preedy *et al.*, 2013).

LF was selected as the target protein for this chapter from the stock proteins at the sponsor company because it is a glycoprotein. In the previous chapter the glycosylation profile of the final product was as discussed a key CQA. As such models developed for the IEX purification of LF should be transferable to other glycoproteins more readily than non-glycosylated proteins.

In order to purify LF by IEX (see section 2.3.2 for details of IEX principles) it is necessary to modify the protein charge. The adsorption of the protein onto the resin is driven by the ionic interaction between the oppositely charged groups on the molecule and the functional ligand on the resin. Increasing the salt concentration (by a gradient) causes the molecules with the weakest ionic interactions to elute from the column first, and the molecules with a stronger ionic interaction to elute later in the gradient. The optimum pH of column operation depends on the pI of the protein and the pKa of the ligand on the resin. LF has a pI of 8.4 and with most cation exchangers having a pKa

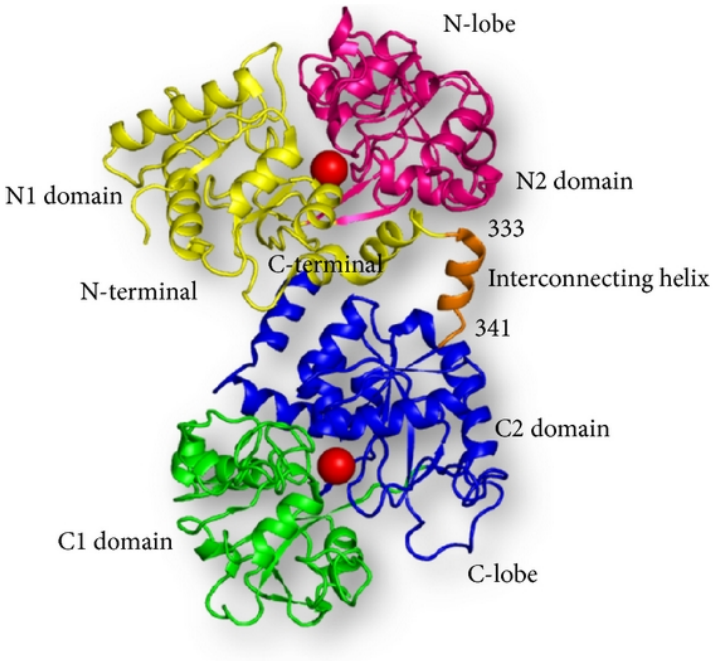


Figure 6.1: Image of lactoferrin showing the N-lobe with the N1 and N2 sub-domains, and the C-lobe with the C1 and C2 sub-domains. Also shown is the alpha helix which connects the two domains, and the iron binding site (highlighted as red balls) (Frank, 2014).

of 1.2. The buffer should be somewhere between the two at pH 6.0. If the pH of the mobile phase is increased it causes the molecule to become less positively charged, meaning the protein can no longer form an ionic interaction with the ligand causing the protein to elute.

6.3 Experimental data

This section presents the data collection method, and a preliminary analysis of the raw data used for model development.

The experiments conducted for this case study were performed at Fujifilm Diosynth Biotechnologies (Billingham) by the author of this thesis. The experiments were carried out using the framework of the company's platform process, that was developed in-house and is an optimised process. A design of experiments methodology was applied that investigated changes to key process parameters to see how these changes would affect operational performance,

with the aim of producing a set of experiments which characterise the design space. A full summary of the experimental procedure is given in Chapter 4.

The DoE used in this work was a minimum resolution IV design. This design is routinely used at the sponsor company as a way of performing scouting experiments (Tamhane, 2009). Scouting experiments are used to characterise a design space and Monks *et al.* (2012) summarises how this aids in the development of a design space. Figure 6.2 shows how the use of a design of experiments methodology to establish the design space is a key aspect of designing a new process.

Within the DoE the variation in buffer pH was obtained through variation of the concentration of sodium phosphate added during buffer preparation. The initial concentration of protein in the load samples was obtained through measuring the optical density of the sample. With the same methodology employed to determine the yield at each stage of all the experiments.

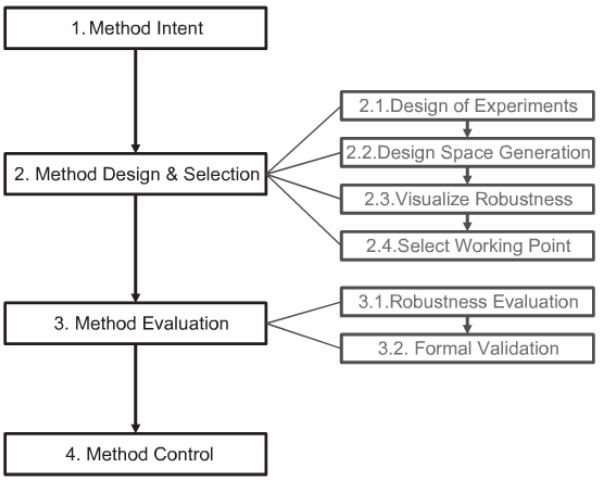


Figure 6.2: Quality by design work flow as presented by Monks *et al.* (2012)

Table 6.1 provides a summary of the conditions of the 13 batches (see section 4.1.2 for full details on experimental procedure). For all 13 batches the same column was used, having been cleaned and regenerated at the end of each batch.

Table 6.1: Operating conditions for all 13 batches, showing the values which were varied within the DoE.

Batch	Flow rate (ml/min)	Load pH	Gradient CV	Load concentration (mg/ml)	Elution pH
1	1.0	7	12	20	7
2	1.5	8	8	10	8
3	1.5	8	8	30	8
4	0.5	8	16	30	8
5	0.5	8	8	10	8
6	1.0	7	12	20	7
7	1.5	6	16	30	6
8	1.5	6	8	10	6
9	0.5	6	16	10	6
10	0.5	6	16	10	6
11	1.5	6	16	10	8
12	0.5	8	8	30	8
13	0.5	8	8	30	6
14	1.5	8	16	10	6
15	1.0	7	12	20	7

6.4 Data preprocessing

The MVDA techniques applied in this chapter are used to establish relationships between the inputs and outputs of the process. Therefore the data is time aligned between batches and variables (for information on the time alignment procedure see Chapter 4).

The elution can be used to show the efficiency of the separation, the resolution, and the yield of each component. This research is focused on the prediction of the elution peak for LF. As the conditions of each batch (Table 6.1) vary greatly it causes similarly high levels of variation in the elution. One such variation in the elution peak is the retention time, which can cause issues with batch alignment. The literature offers various techniques for dealing with peaks which are not aligned between batches of similar conditions. These include dynamic time warping (Sakoe and Chiba, 1978), correlation optimised warping (Nielsen *et al.*, 1998), and icoshift (Tomasi *et al.*, 2011). However, these techniques are more commonly used in multicomponent systems, where there are patterns and interactions between peaks, and in data where there is

not as much batch to batch variability. The batches used in this study have retention times ranging from 14 minutes to 51 minutes, thus the method of cutting the batches to the length of the shortest batch, used in Chapter 5, can not be applied here. Cutting the batch data would in many cases loose all the data relating the elution, whereas extrapolating the shortest batches to fit the length of the longest ones would not work either as this distorts the nature of the data.

As each batch contains a pure sample of lactoferrin it is assumed that the elution peak corresponding to the protein will be Gaussian (Giddings, 2002). The retention time and peak width values were used to determine the start and end point of the elution peak as given in Equations 6.1 and 6.2.

$$t_{e,start} = t_R - \frac{1}{2}\sigma \tag{6.1}$$

$$t_{e,end} = t_R + \frac{1}{2}\sigma \tag{6.2}$$

where $t_{e,start}$ and $t_{e,end}$ represent the start and end point of the elution peak respectively, t_R is the retention time, and σ is the peak width. Having determined the values of $t_{e,start}$ and $t_{e,end}$ the elution peak from each batch and the on-line variable measurements corresponding to the same time points can be isolated. As all the on-line measurements were recorded at slightly different times the data was aligned to the times of the absorbance readings so that a direct comparison can be made between the sample points of all the variables. Subsequently the 'maxcellsize' function in Matlab was used to assess which batch had the maximum number of sample points for the peak data. This value was then used to interpolate all the batches so that each batch had the same number of sample points, the resulting batch information is shown in Figure 6.3. Information on the interpolation technique employed can be found in Chapter 4.

The resulting matrix was formed from only the peak data, aligned between variables and between batches.

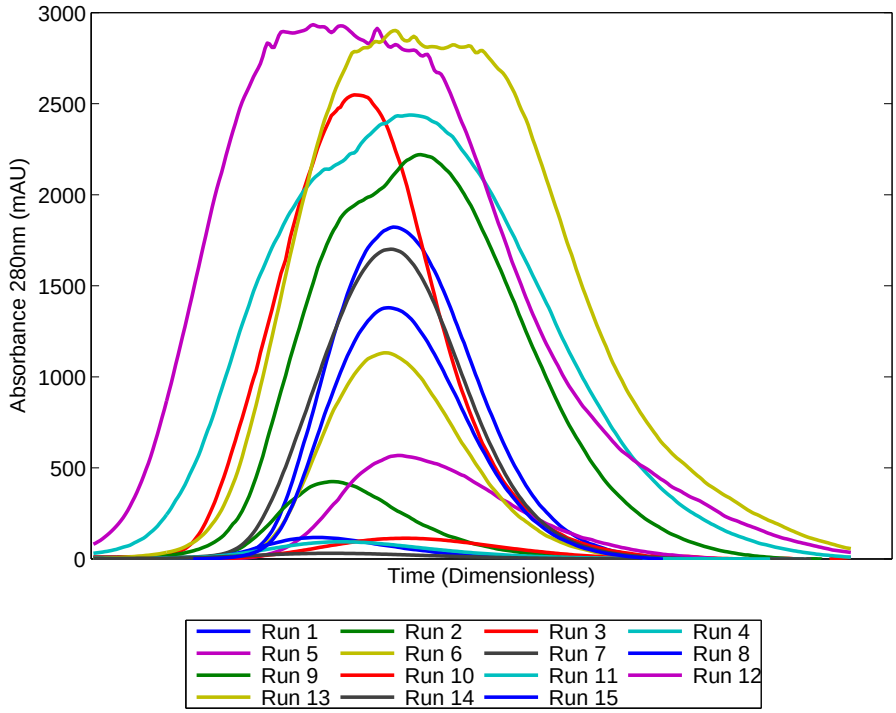


Figure 6.3: lactoferrin elution peaks for 15 experimental batches (Table B1) the figure is not representative of the retention time of each batch, as the peaks have been over layed to demonstrate variations in peak height and width.

6.5 First principles prediction of retention time

Initial attempts were made to predict the retention time using PLS modelling (Figure D4 in appendix D shows the predictions), however the predictions made from the resulting models were very poor highlighting the need for an alternative modelling technique. First principles models, based on the fundamental concepts of chromatography offer an alternative solution. In the case of IEX chromatography these models include aspects associated with mass transfer and adsorption. As highlighted in the methodology the prediction of the correct retention time is very important, this section presents a model based upon one from literature used to predict the retention time (Shellie *et al.*, 2008; Madden *et al.*, 2002).

6.5.1 Methodology

The LF purification was performed under a gradient elution, therefore the retention model used was the *'Gradient Elution Retention Model'* as described by Shellie *et al.* (2008). In this model the retention factor (k_g) under gradient conditions is given by Equation 6.3.

$$log(k_g) = a_g + b_g logR \tag{6.3}$$

where the subscript g is used to denote that it is under gradient conditions and R is the slope of the gradient ramp. A plot of $logk_g$ versus $logR$ can be used to determine the retention behaviour where a_g is the y-intercept and b_g is the slope of the relationship.

Having determined the constants of the system, equation 6.4 first presented in Jandera and Churáček (1974) and later applied by both Baba *et al.* (1985) and Shellie *et al.* (2008), predicting retention time, can be used.

$$t_g = \left(\frac{1}{u}\right)\left(\frac{1}{B}\right)\left(\left[(zb_i+1)Ba_i t_0 u + C_s^{(zb_i+1)}\right]^{\frac{1}{zb_i+1}} - \frac{C_s^{\frac{1}{z}}}{B}\right) + t_0 \tag{6.4}$$

where t_g is the retention time under gradient conditions, u is the mobile phase flow rate, B is the normalised gradient ramp ($B = R/u$, mM/mL of column), z is a parameter used to describe the shape of the gradient profile (z=1 for linear gradients), C_s (mM) is the starting concentration for the gradient, t_0 is the void time, and a_g and b_g are as determined from Equation 6.3.

6.5.2 Results and discussion

To be able to produce a model which can predict the retention time, a data set must first be used to generate the constants. In the case of the model presented in Equation 6.4, the slope of the gradient ramp (R), the value used to describe the gradient (z), the dead time (t_0), and values of a_i and b_i need to be determined from the behaviour of the system. A summary of the Matlab files constructed for the retention time prediction model are listed as a flow chart in

Figure 6.4.

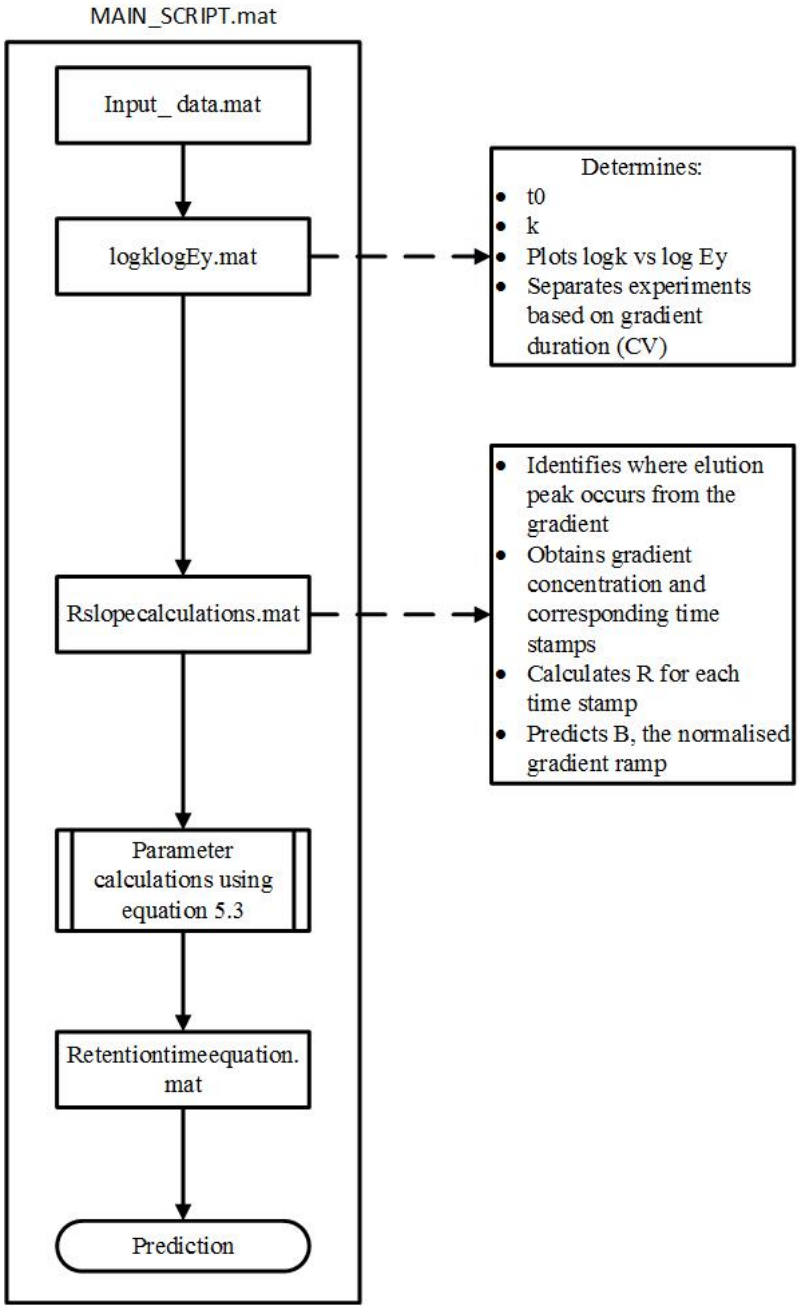


Figure 6.4: Flow chart showing the progression of files within the Matlab code used to predict retention time

Table 6.2 shows the values generated using the Matlab script for the 14 batches used to train the model. Issues arose during the application of the model due to the wide scouting nature of the data set. Some of the factors tested in the DoE, such as elution duration (CVs) and concentration of protein, can greatly affect the behaviour of the system. Therefore within the model code the influence of these factors could be considered separately. This allows the user to specify how long the gradient occurs for and predicts the retention

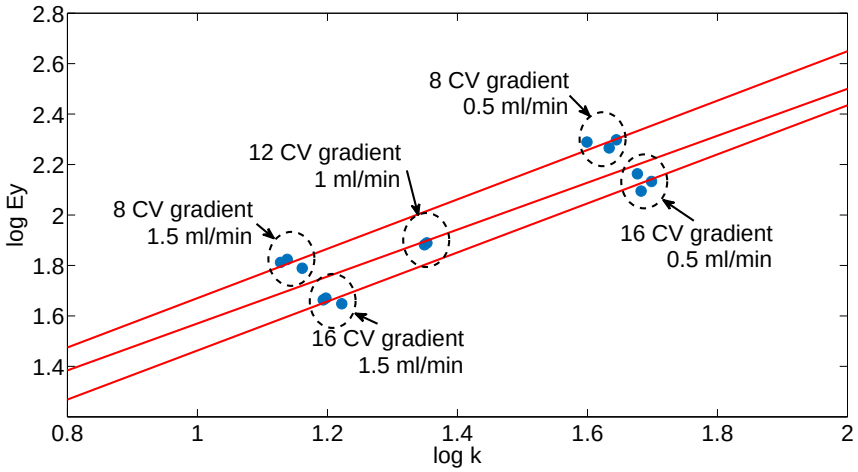
Table 6.2: Values generated from lactoferrin data set for the constants in Equation 6.4 and used to estimate the constants for the validation batch.

Batch	t_0 (min)	k	E_y (mM)	R (mM/min)	B (mM/mL)
1	1.00	22.35	76.20	8.34	8.34
2	0.67	10.17	64.87	18.75	12.50
3	0.67	11.82	61.53	18.75	12.50
4	2.00	71.81	124.40	3.12	6.25
5	2.00	51.46	194.60	6.25	12.50
6	1.00	14.85	77.50	8.34	8.34
7	0.67	7.49	44.47	9.38	6.25
8	0.67	6.40	66.67	18.75	12.50
9	2.00	68.24	135.80	3.12	6.25
10	2.00	61.42	145.60	3.12	6.25
11	0.67	6.56	46.80	9.38	6.25
12	2.00	56.38	198.60	6.25	12.50
13	2.00	53.26	184.40	6.25	12.50
14	0.67	6.15	46.07	9.38	6.25

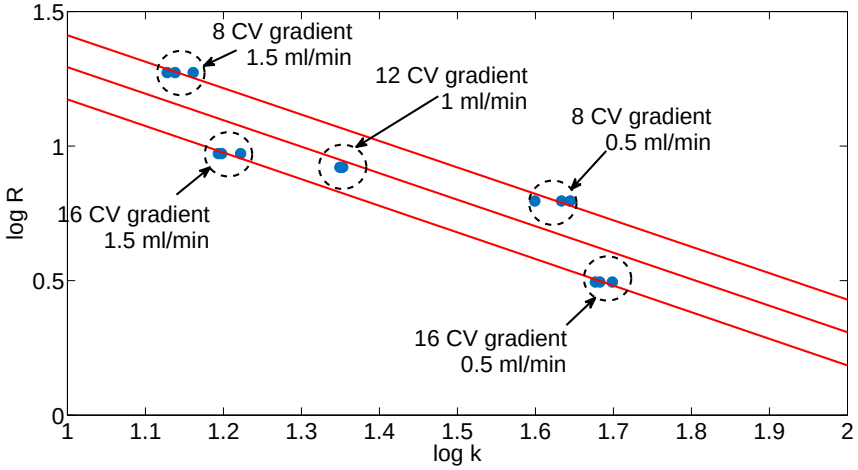
time based upon this. A similar approach was adopted by Madden *et al.* (2002) who used the differentiation based on R.

The impact of this can be seen in Figure 6.5, where the data points can be seen to separate based upon the gradient CV and the velocity. It can be seen that k is influenced by the velocity and Ey is influenced by the gradient CVs. This is as would be expected as k is the retention factor and it is a function of time and Ey is the concentration of eluent species. The gradient CVs are related to the eluent species because if a gradient is to achieve 100% concentration, then a shorter gradient duration would mean a higher concentration.

It is from the slopes shown in Figure 6.5 (b) that the constants given in Equation 6.3 are determined and subsequently the normalised gradient ramp (B). These values are then used in Equation 6.4 to predict the retention time. This example highlights how even first principles models are reliant upon the data collected and used to construct them. To predict the retention time for the system presented in this section outside of the operating parameters used (i.e. flow rate outside of 0.5-1.5 ml/min or CV gradient 8-16) extrapolation could be used but accurate predictions would only come from further experimentation.



(a) logk vs. logEy



(b) logk vs. logR

Figure 6.5: (A) Plot of log k versus log Ey, where k is given in Equation 2.8 and Ey is the concentration of the eluent species. (B) Plot of log k versus log R, where k is given in Equation 2.8 and R is the slope of the gradient, this graph is used to calculate the constants a_i and b_i as given in Equation 6.3. The ellipses represent the groups of data points, which are separated based upon gradient column volumes and flow rate.

For the validation batch (batch 15) the values generated using the Matlab code, together with the predicted and measured retention time, are shown in Table 6.3.

Table 6.3: Values generated from lactoferrin data set for the validation batch, and the subsequent predicted retention time using equation 6.4.

t_0 (min)	R (mM/min)	B (mM/mL)	C_s (mM)	z	a_i	b_i	Predicted retention time	Measured retention time
1.00	8.34	0.92	3.12	1	2.27	-0.98	23.44	23.48

This model is limited in the sense that the maximum and minimum CV gradient, flow rate are specified by the limits of the experimental DoE. To use the current data set to be able to predict within these limits would be possible, as the values for equation 6.4 could be interpolated.

In summary the application of the first principles model to the data has produced an accurate prediction of the retention time. This study has shown that the prediction method is reliable, however issues may arise from the fact that a dataset is required to derive the constants needed for the equation. Therefore this equation cannot be applied to a completely new process, without exploratory experiments first being conducted. Within industry this is generally not an issue as standard practise requires preliminary experiments to be conducted which, as was shown here, can be adapted to provide the necessary information.

6.6 Principal component analysis of IEX data

This section presents two principal component analyses; the first using the off-line data (Table 6.1) and the second using the on-line data (conductivity, pH, temperature, concentration, pressure). The aim is to identify the variables which have the greatest impact on the CPP, yield. PCA analysis is used in this research as a data reduction technique to determine significant input variables to include in the PLS models. Inclusion of variables having a large influence

on the model due to pre-processing as opposed to the amount of variation in data they explain can be very detrimental to the model accuracy (Li *et al.*, 2014). The removal of variables should be done in such a way that the main sources of variation in the data are retained. Therefore this study uses PCA as it allows for the variables' impact on the variation to be seen (through loadings values). Through process knowledge and comparison of the loadings values and raw data it can be assessed as to whether each individual variable is a source of data variation. Traditionally PCA is also used to determine batches which are termed outliers, from the data set (Bro and Smilde, 2014). This is done by unfolding the data batch-wise and analysing the scores. In this analysis the data was unfolded in this manner, however due to the scope of the DoE and the lack of repeats of experiments no batch removal occurred. It would be expected that there would be a few batches highlighted as being outside of the 95% confidence limit, but as the experiments were conducted using a scouting DoE design (Tamhane, 2009), it will be assumed that these batch variations can be accounted for due to the operating conditions.

6.6.1 Methodology

The protocols to perform the multivariate analysis and modelling are described in Chapter 4 for principal component analysis (PCA). The first PCA analysis was performed using the operational parameters from the DoE, namely flow rate, load pH, gradient CV, elution pH, and loading concentration. As yield is a CPP of interest, the batches were categorised as high yield, medium yield, and low yield as this allowed for easy determination of the variables that influence the yield. The PCA analysis of the on-line data included the variables measured on-line: conductivity, buffer concentration, pH, pressure, flow rate, and temperature.

For both the off-line and on-line analyses the data matrices were unfolded in a batch wise manner to preserve the variation between batches. The cross validation technique 'leave-one-out' was applied to both the off-line and on-line analyses. The number of PCs selected for each analysis was determined from the eigenvalues and root mean squared error of cross

validation (RMSECV) (Bro *et al.*, 2008). Matlab software along with the Eigenvector program PLS-Toolbox were used to perform the data analysis.

Pre-processing

Section 6.7 provides a detailed case study of the pre-processing techniques which could be applied to this data set. However, the PCA analyses covered in this section are used as a precursor to determine the underlying variable interactions and whether any variables need to be removed prior to the pre-processing study. Therefore for the PCA analyses the only pre-processing technique which was applied was autoscale as this will allow for direct comparison between variables. For a detailed explanation of autoscaling refer to Table 6.5 on page 180.

6.6.2 Results and discussion

Off-line data analysis

Process knowledge can be expanded by obtaining a better understanding of the interactions between the variables in the process. With respect to this research this means determining the variables which impact on the CQAs and CPPs (Huang *et al.*, 2009). A PCA analysis was performed to identify the main sources of variation in the data and to identify the variables that cause this variation and thus have the greatest impact on the CQAs and CPPs. In terms of this study the main CPP can be said to be the yield of the product obtained from the process. The data collected for this study was obtained as part of a scouting DoE and as such the recorded yield for each batch varied greatly. Therefore separate analyses were performed; the first being on all of the data in one matrix, then the next three considered the batches as groups of high, medium, and low yield.

The conditions for the DoE are summarised in Table 6.4 along with the corresponding off-line measurement of the yield. The yield measurements

Table 6.4: DoE conditions and yield obtained during elution for all batches. The batches were categorised as high yield ($y > 70\%$) in red, medium yield ($30\% < y < 70\%$) in yellow, and low yield ($y < 30\%$) in blue

Batch	Flow rate (ml/min)	Load pH	Gradient CV	Load concentration	Elution pH	Elution yield (%)
1	1.0	7	12	20	7	67.64
2	1.5	8	8	10	8	6.59
3	1.5	8	8	30	8	83.65
4	0.5	8	16	30	8	81.06
5	0.5	8	8	10	8	6.37
6	1.0	7	12	20	7	35.41
7	1.5	6	16	30	6	64.57
8	1.5	6	8	10	6	1.11
9	0.5	6	16	10	6	63.15
10	0.5	6	16	10	6	7.85
11	1.5	6	16	10	8	7.19
12	0.5	8	8	30	8	95.29
13	0.5	8	8	30	6	94.06
14	1.5	8	16	10	6	1.84
15	1.0	7	12	20	7	44.17

were categorised as high yield ($y > 70\%$), medium yield ($30\% < y < 70\%$), and low yield ($y < 30\%$). There is a large range in the achieved yields from the batches, ranging from 1.12% to 95.29% of the original sample loaded on the column. Within the biopharmaceutical industry the aim is to achieve as high a yield as possible whilst minimising the cost of the process, therefore the aim of the PCA is to identify the conditions by which the highest yields were achieved and then in subsequent work determine if it is possible to economically achieve these levels.

The constructed PCA model accounts for 73.33% of the cumulative variance in the first 3 PCs. It was decided further PCs should not be included in the model through studying the eigenvalues and RMSECV values. The RMSECV values showed a considerable increase after 3 PCs, which corresponded to a levelling in the eigenvalue at the same point (see Figure 6.6).

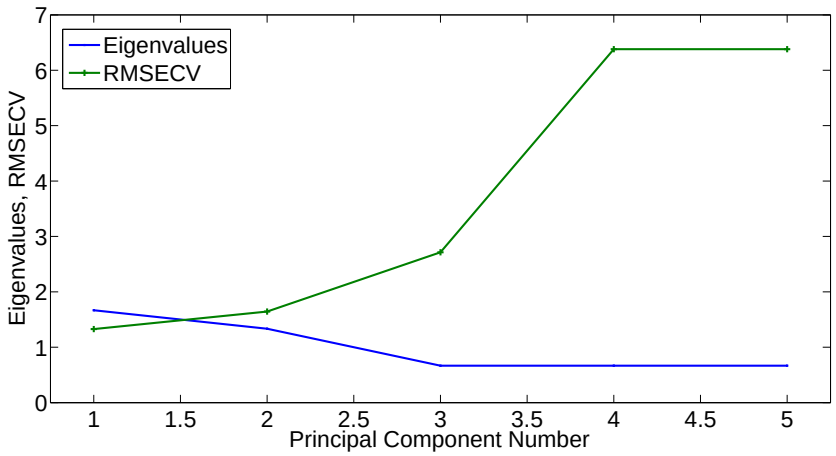


Figure 6.6: RMSECV and eigenvalues for off-line (DoE conditions) measurements PCA analysis. The eigenvalues show an significant increase after 3 PCs, which corresponds to a levelling off of the RMSECV.

To assess the variation between batches the bivariate scores plots were used. For this particular model PC 1 and PC2 capture 33.33% and 26.67% of the variance respectively (Figure 6.7). Figure 6.7 shows that there is a relationship between the DoE operating parameters and the yield. With regards to batches 7 and 10 shown on the left of the plot and batches 3 and 5 shown on the right hand side of the plot the distinction should be made that although the same score value was obtained for PC 1 there was a slight difference in PC2. Batches 5 and 10 both had positive scores and batches 3 and 7 were both negative. To determine why these batches have the same score for PC1 and similar scores for PC2, the values of the initial conditions (Table 6.4) are considered. Batches 3 and 5 both have load and elution pHs of 8 and gradient CVs of 8, and batches 7 and 10 both have load and elution pH of 6 with gradient CV of 16. Looking now at batch 2, it can be seen that for these three variables the conditions are the same as for batches 3 and 5. Also for batch 9 the conditions are the same as for batches 7 and 10. It is known that for cation chromatography a higher pH means there are more positively charged ions to compete with the binding sites on the resin. Thus a higher pH causes the elution to occur faster. Similarly if the gradient is operated from 0% - 100% over a shorter time frame then the protein will elute quicker, as the concentration is increased more rapidly. This suggests that PC 1 is capturing the retention time of the protein with the positive scores being a shorter

retention time and the negative scores being a longer retention time. As the retention time is independent of the final yield this is why there is no grouping of the batches by yield in the first PC.

Considering now the variation in PC 2, there appears to be two variables influencing the variation captured in PC2. Considering first the load concentration there is an implied relationship between the positive score values and the high concentration, and between the negative score values and the low concentration. However there is an exception to this in batch 9. This shows that the more product loaded onto the column the more product there is eluted. However considering now the batches with the positive PC2 scores (batches 4, 5, 9, 10, 12, and 13) all have a flow rate of 0.5 ml/min, and the batches with negative scores (batches 2, 3, 7, 8, 10, 11, and 14) all have a flow rate of 1.5 ml/min. This shows that there is a degree of independence between the yield and the flow rate, with the flow rate determining instead the percentage of original sample that is eluted. As Shoji *et al.* (2009) mentions a high flow rate gives less time for the product to bind to the resin within the column therefore suggesting the performance of the adsorption as a possibility for having impact on the yield. In summary the scores for PC2 suggest that load concentration influences yield whilst flow rate can be used to control it. To further assess the influence of the different variables on the variation, the loadings plots for both PC1 and PC2 can be used (Figure 6.8).

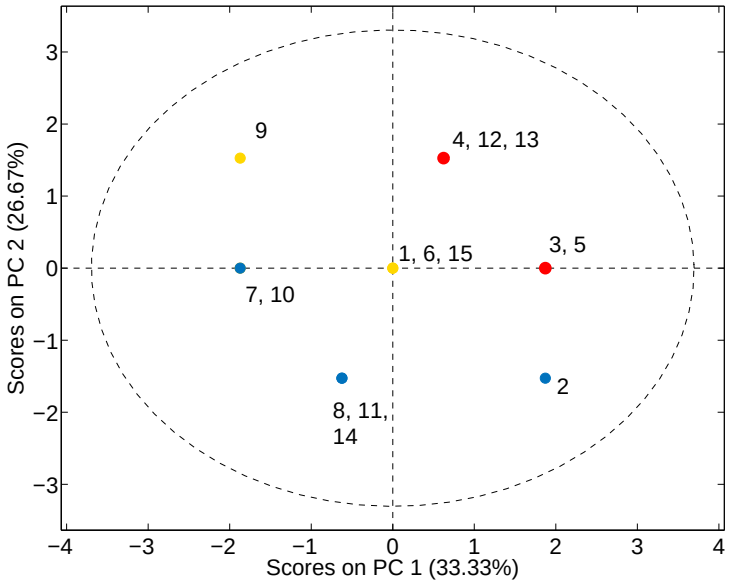
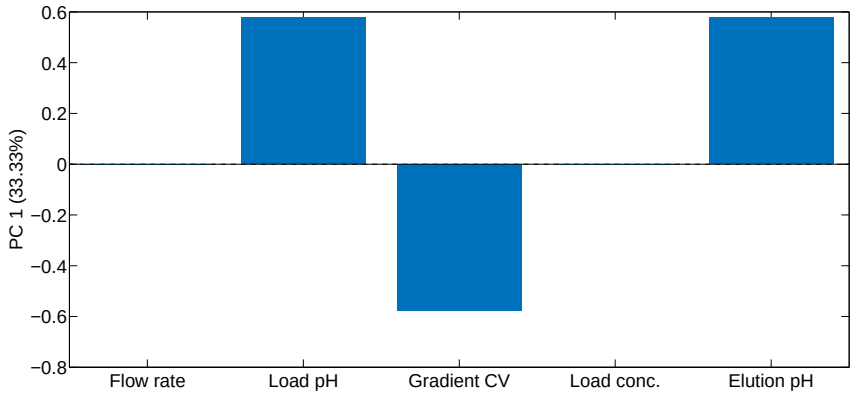
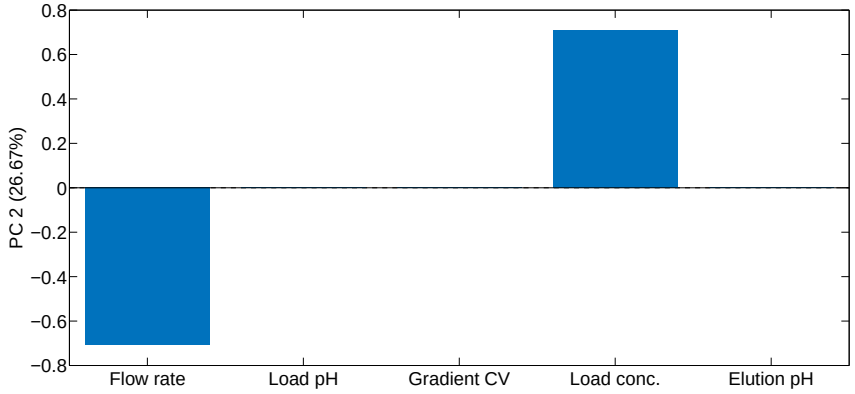


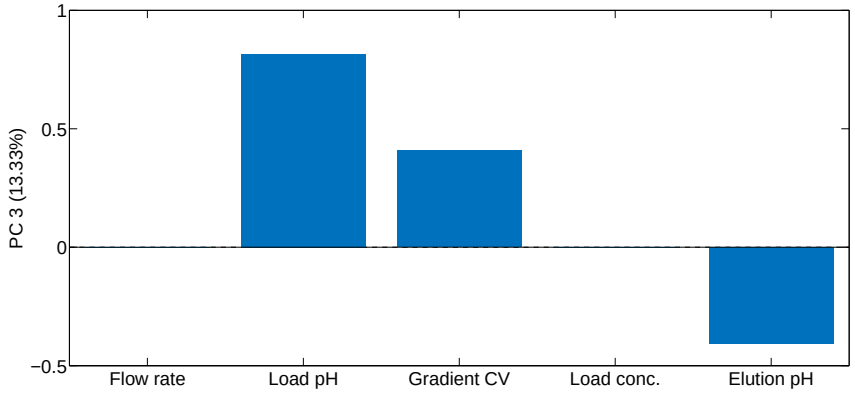
Figure 6.7: Bivariate scores plot of PC1 and PC2 for the off-line data analysis, showing the high yield (red), medium yield (yellow), and low yield (blue) batches. Batch numbers refer to Table 6.4.



(a) PC 1



(b) PC 2



(c) PC 3

Figure 6.8: Loadings of PC1 (a), PC2 (b) and PC3 (c) for the model constructed using off-line data (X-block is DoE operational parameters).

The loadings plot for PC 1 confirms that the variables contributing most to the variation in PC 1 are load pH, elution pH, and gradient CV with weightings of approximately 0.6. As highlighted from the analysis of the scores plot all three of these variables contribute equally to the variation in PC 1. The variables of flow rate and load concentration do not influence the variation in PC 1, instead they can be seen to be responsible for all the variation captured in PC 2, where again these two variables have equal loadings weights of ≈ 0.6 and equally influence the variation in PC 2. Plot (c) shows the loadings for PC3, the scores did not show a distinct relationship, thus a plot of them is not included. However, the loadings show that for PC 3 the load pH accounts for the variation captured in this PC. The high weighting for load pH, and the subsequent smaller weightings for gradient CV and elution pH suggest that this PC captures the absorption of the protein on to the resin. As discussed previously the pH of the mobile phase can significantly affect the ability of the protein to bind. If the pH is closer to the isoelectric point of the protein then there is more competition with the protein for available binding sites, and vice versa, if the pH is (in the case of cation exchange) lower than the isoelectric point, it means less competition for binding sites.

On-line data analysis

The on-line data analysis was performed on the data after the process of peak identification and isolation as described in section 6.4. This is in line with the method developed by Lu *et al.* (2004), who state that a process may occur in stages and it would be advantageous to analyse and model these stages separately. Therefore for this PCA analysis the matrix contains only the on-line variables after they have been cut to the duration of the elution peak. The purpose of performing a PCA analysis on the on-line data is to reduce the data set. There are several on-line variables measured but not all of them account for the maximum variation in the system, some may only have a small influence. Therefore it would be unwise to build a model using variables which do not characterise the outputs. The PCA analysis identifies variables for removal and variables for inclusion in subsequent models. Furthermore it

is important to ensure that the variables included in the models are representative of process variation and not noise, which is why process understanding is a key factor in variable removal. Process understanding can cover not just the operation of the units but also the process of modelling, for example as Fransson *et al.* (2001) explain certain data pre-processing techniques, such as autoscale, can manipulate the variables, making them appear to contribute more to the variation when in reality they have little influence on the process. PLS regression can be similarly used for variable reduction and it has the added advantage of considering not only the variation in **X**-block data but also **Y**-block data, and how they interact. PCA was chosen as the variable reduction technique as one potential application from this research is whether the process can be monitored on-line whereby it might be possible to use the changes in the correlation of the variables to control them. This technique has been successfully developed and applied for batch processes by Lu *et al.* (2004). Additionally Maitra and Yan (2008) discuss the relative methods of both techniques and show that even though PCA is used here as the primary variable reduction technique it is also worthwhile checking any subsequent PLS models to ensure the correct variables are used.

An initial PCA analysis was performed using all the on-line variables. The model was constructed using 5 PCs after observation of the RMSECV and eigenvalues, capturing 89.60% of the variance in the data. As can be seen in Figure 6.9 after 5 PCs the value of the RMSECV increases with the inclusion of PC6 suggesting that by this PC the model is beginning to capture noise.

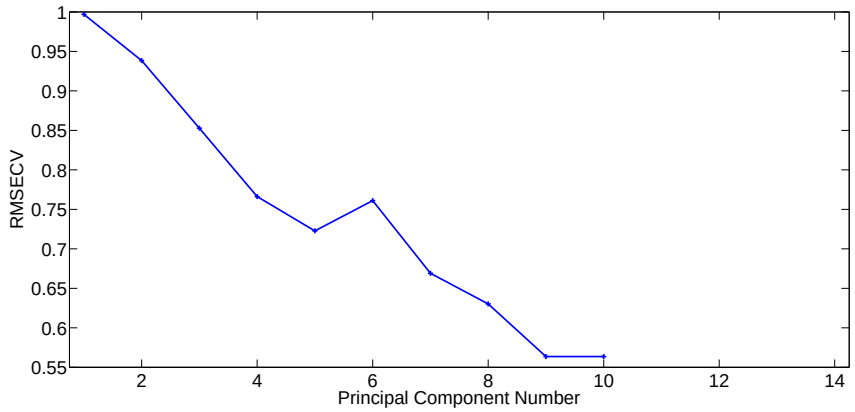


Figure 6.9: RMSECV plot for on-line data (conductivity, concentration, pH, pressure, flow, and temperature) measurements for PCA analysis. Plot suggests that the increase in RMSECV for PC6 could indicate that PC6 captures noise.

To assess whether any variables have a negligible impact on the analysis the first principal component, which captures the maximum variation, is examined (Figure 6.10). The first observation is that the significance of all the variables is low, with the highest weighting being ≈ 0.03. However relative to each other the most significant variables are shown to be pressure, flow, and temperature.

The raw data shows that these variables are controlled to relatively narrow set points with only small fluctuations of ± 0.5. Therefore it can be assumed that the larger weighting given by this analysis to these variables is due to the aforementioned effects from pre-processing techniques. As a result of this, if these variables were included in a predictive model they would place a greater emphasis on noise which does not influence the absorbance.

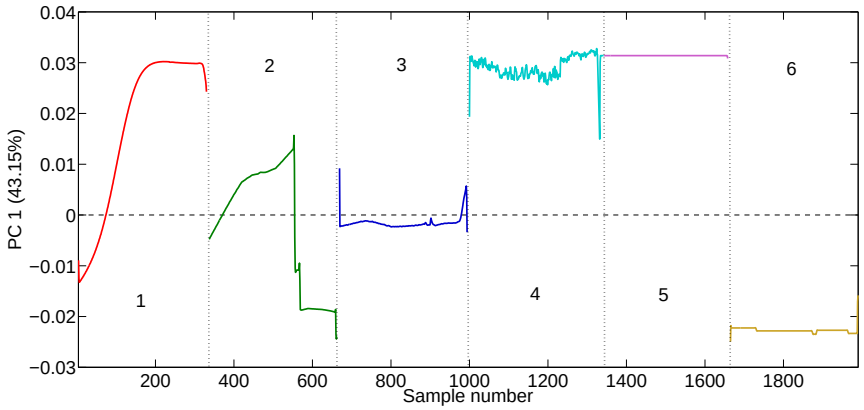
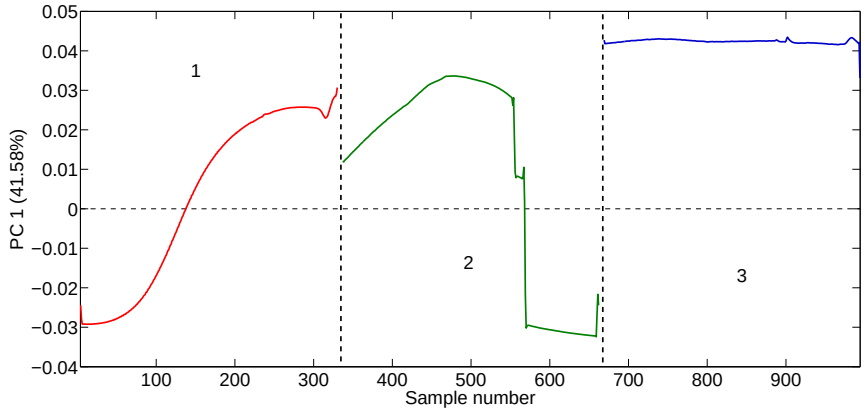


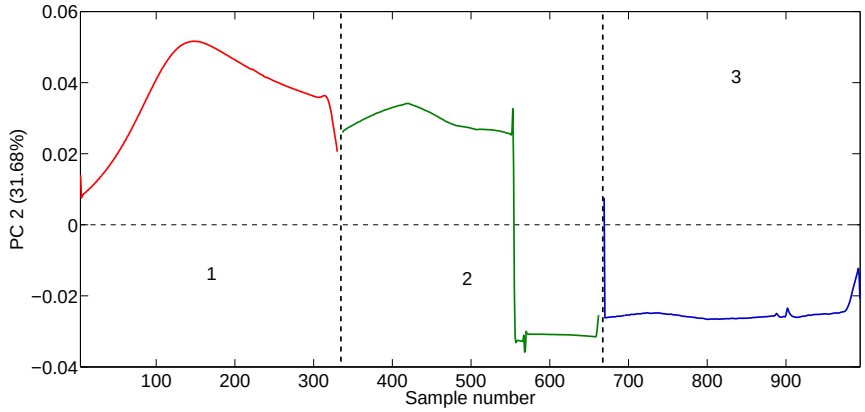
Figure 6.10: Loadings plot for PC1, showing the six on-line variables; (1) conductivity (2) concentration (3) pH (4) pressure (5) flow (6) temperature. The figure shows that relative to each other the most significant variables are pressure, flow, and temperature. The vertical dashed black lines are used to distinguish between the individual variables.

McGuffin and Chen (1997) showed that the small compressibility of the mobile and stationary phases means that in liquid chromatography pressure has a negligible influence on solute retention in liquid chromatography. Thus the relative importance of pressure shown in Figure 6.10 is more likely due to issues in scaling the data. Therefore pressure is removed from the analysis. As shown in the analysis of the off-line (operating parameter) data the flow rate of the mobile phase is an important variable in the process. However with the time series data used here in the on-line analysis, the flow rate was held at a constant value throughout the process. Therefore, a similar effect to that seen with the pressure variable is observed, in that when all 6 variables were scaled, the small variation in the fluctuations of the flow rate were given a greater importance. So flow rate was removed from the on-line analysis and only considered through the off-line data. Rubinson and Rubinson (2000) stated the importance of temperature in IEX separations and show that temperature can be changed to improve liquid chromatography separations. However if reproducibility is critical then the column should be held at a constant temperature. In the case of this data set temperature was not an investigated parameter, and for column reproducibility it was held at a constant. Slight fluctuations of ± 0.3°C were observed around the set point of 25.5°C. Similarly to the pressure and flow rate the supposed importance of this variable is more likely to be due to scaling issues.

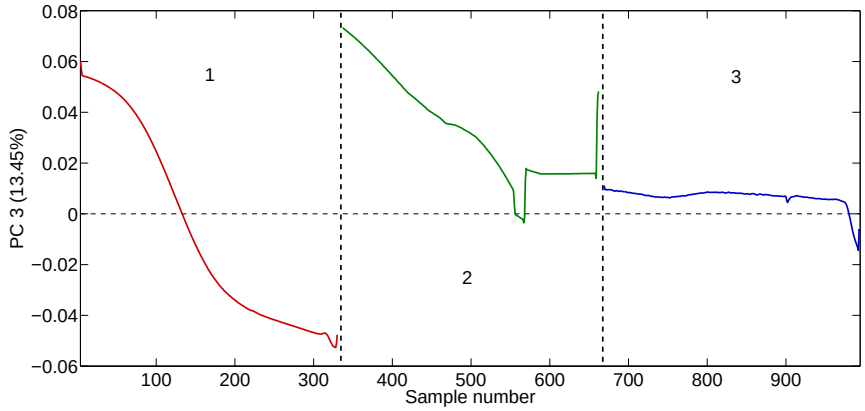
A second PCA analysis was performed with these three variables removed, and containing conductivity, concentration, and pH only. Upon reviewing the RMSECV and eigenvalues, 3 PCs were included in the model, explaining 86.72% of the variation. The loadings plot for PC 1 is shown in Figure 6.11.



(a) PC 1



(b) PC 2



(c) PC 3

Figure 6.11: Loadings of PC1 (a), PC2 (b) and PC3 (c) for the on-line data analysis; (1) conductivity (2) concentration (3) pH. All three PCs show relatively small weightings for all three variables. The vertical dashed black lines are used to distinguish between the individual variables.

Similarly to the first analysis the weightings of the variables in all 3 PCs were relatively small with the largest weighting being only 0.08. However, considering the variable loadings relative to each other it can be seen for PC1 the most important variable for the variation was the pH (blue), with pH remaining important throughout the entire duration of the elution peak. This confirms the finding of the off-line analysis that the pH affects not just the ability of the protein to bind to the resin but also the subsequent desorption of the protein during the elution. Observing the weightings given to the conductivity (red) over the duration of the elution peak, it can be seen that initially the weightings were negative, with a gradual change to being positive. This corresponds to the gradient performed over the elution peak, whereby the conductivity is increased promoting desorption of the protein. Finally, in the case of the concentration (green) it can again be seen that there was a change from the weightings from positive to negative. This is likely to be representative of the concentration gradient achieving 100%. For PC 2 it can again be seen that the weightings for pH remained constant, which is as expected as the pH was held constant for the duration of the elution. Unlike for the removed variables (temperature, pressure, and flow) this does not warrant the removal of pH as there is still batch to batch variation. A similar trend to that which is shown in PC1, is shown for concentration. With a sudden change in the weightings from positive to negative. This highlights the concentration reaching 100% of buffer B. The variable with the largest weightings value in PC 2 is conductivity, with the largest value occurring at approximately sample number 150 indicating that this is where the elution peak maximum was attained. PC 3 again shows similar findings for all 3 variables as was shown in PCs 1 and 2.

To summarise, the PCA analysis performed on the on-line data showed that the variables within the measured data which are not directly correlated to the variation in this data set are temperature, pressure, and flow rate. Having said this, however, the analysis also showed that for all the on-line variables, the weightings shown in the loadings plots were very small. As PCA is a method for projecting data onto a subspace, it is based on the principle of maximising, for each component, the sum of squares of the correlation

coefficients. This results in each component being strongly correlated with some variables and weakly correlated with the others. This means the loadings values are a measure of the correlation between the original variables and the components used to make the model. Low values show there is little correlation between the original data and any of the variables. This suggests that the variation in the data set is caused by variables which are not included in this analysis (Suhr, 2005).

6.7 Pre-processing of chromatographic data

A review of the literature showed that there was no comprehensive study conducted for pre-processing of IEX chromatography data, therefore this section presents such a study. The basis for this work was determined from similar applications to chemometrics and spectroscopy.

The previous section presented a principal component analysis of the data, to perform the analysis a pre-processing method outlined in literature was used. Engel *et al.* (2013) showed that for the type and form of data analysed in the on-line PCA, autoscale was an appropriate technique to use, as there was not a great deal of variation between batches for each variable. The PCA analysis served the purpose of understanding the data better. Having performed the analysis and considered the additional data, which was used as the **Y**-block in PLS, it was determined that autoscale alone might not be sufficient pre-processing. This conclusion was drawn primarily through the variation shown in the absorbance readings. There is plenty of literature concerned with spectroscopy where a similar range of variation is shown, specifying the need for other pre-processing techniques such as normalisation and smoothing (Rinnan *et al.*, 2009; Gromski *et al.*, 2014; Laxalde *et al.*, 2011; Wang and Kowalski, 1992; Cervera *et al.*, 2009). However, there exists no comprehensive list of the pre-processing methods which should be applied to liquid, and in particular ion-exchange, chromatography. Therefore in order to perform the next analysis using PLS, to continue the variable investigation from the previous section and to produce predictive models, this section

investigates pre-processing techniques.

Pre-processing a data set is an integral part of being able to perform data modelling. Various effects can be introduced into collected data through instrumentation and experimental differences, which are not related to the differences between the outputs of the batches. Data pre-processing is used to remove these effects and to ensure that all experimental batches can be analysed. The methods of data pre-processing used can heavily influence the final results of the model and should therefore be carefully considered. A good pre-processing procedure enhances the chemical/compositional information content of the data, whilst an inappropriate procedure affects the correlation structure. Rajalahti and Kvalheim (2011b) state that a crucial factor of data analysis is the analytical technique used to collect the data, as there is no 'catch all' method which can be used for all data sets.

This study is concerned with the pre-processing of on-line ion exchange (IEX) chromatographic data, in particular measurements associated with conductivity, pH, concentration, and UV absorbance. One example of a similar application in the literature is that of Skov and Bro (2007), which presented a solution to the problem of peak shift. The effects introduced through experimental issues like column bleed are accounted for through the use of alignment and warping of the peaks. Skov and Bro (2007) showed that it is possible to produce a predicted chromatogram which is easier to interpret and analyse. However the data set used in the research presented in this case study has peak shifts which are dependent on changes introduced to the system deliberately. Skov and Bro (2007) detailed a methodology for analysis of unintentional peak shifts, this research aims to present a methodology for the optimisation of a design space where peak shifts are viewed as a consequence of operating conditions.

Although there is no literature directly relating to the pre-processing of a data set used to characterise a design space for IEX, comparisons can be made to literature concerning spectroscopy through the similarities in the measurements recorded. Stordrange *et al.* (2002) present a comparison of different pre-processing procedure applied to NIR data, using standard

techniques such as normalisation, differentiation and multiplicative scatter correction (MSC). There are other techniques presented in the literature such as orthogonal signal correction (OSC) (Wold and Sjöström, 1998), optimised scaling (OS) (Karstang and Manne, 1992), or standard normal variate (SNV) which was investigated by Barnes *et al.* (1989). Luypaert *et al.* (2004) presents a methodology for NIR pre-processing which applies SNV, detrend correction, offset correction, and derivation. Another methodology presented by Artursson *et al.* (2000) to X-ray powder diffraction uses wavelength transforms, Fourier transforms, and Savitzky-Golay derivative (Savitzky and Golay, 1964), to improve PLS predictions. These examples all show that the type of pre-processing applied is very dependent on the characteristics of the data set.

The literature was investigated to find applications of pre-processing to similar types of data. The aim of this was to identify pre-processing methods to use in this research, and subsequently eliminate techniques which would not be appropriate. These applicable techniques were then investigated through a study which built PLS models and used the fit of the model to determine the applicability of the pre-processing techniques. Factors which were considered include time alignment of data, normalisation to account for differences in signal intensities (Arneberg *et al.*, 2007), smoothing of noise data (Savitzky and Golay, 1964), use of 1st or 2nd derivative to remove background noise (Savitzky and Golay, 1964), and scaling of variables (Karstang and Manne, 1992). The root mean squared error (RMSE), normalised root mean squared error (NRMSE) and the Akaike information criterion (AIC) were used to determine the models which provide the best fit.

6.7.1 Review of techniques

This section provides a review of the literature and presents the techniques which are applicable to the data used in this research.

Time alignment

Time alignment is an important step in preparing a data set for multivariate analysis, and is one of the more challenging aspects. Skov *et al.* (2006) state that the reason for time aligning data is to bring the data to a form where the elements in the matrix for each sample describe the same phenomena. Not performing time analysis means that the assumption made in multivariate analysis of bi-linearity (2-way analysis) or tri-linearity (3-way analysis) is no longer valid. The literature shows various methods for time aligning chromatographic data. One method is correlation optimised warping (COW) which works by using assumed peak shape and area properties (Nielsen *et al.*, 1998; Bylund *et al.*, 2002; Pravdova *et al.*, 2002; Skov and Bro, 2007). COW is useful in applications where the goal is to determine the difference between a experiment and a control experiment. Where alignment of corresponding features in the measurements is important, COW uses a reference experiment (control) and a user defined *shift window* which determines by how much a peak can shift, and in this way is useful for minor peak corrections (Skov *et al.*, 2006). Fourier transforms are also regularly used in the literature to correct for time alignment issues (Grung and Kvalheim, 1995; Wong *et al.*, 2005b; Zheng *et al.*, 2013). Methods which use Fourier transforms differ from those of warping methods in that the whole chromatogram is used, and multiple peaks are detected using a transform which then aligns these based upon a reference chromatogram. Literature has shown both warping and Fourier transform methods to be effective, however, the use of reference chromatograms in all the methods means they are not appropriate for this study.

The techniques described are generally applied to data sets where there are only small variations from the control experiment. However the data set used in this research was collected via use of a DoE to characterise the design space for the purification of the protein. This presents a challenge in the fact that each of the experiments varied greatly and no one control experiment could be used as a reference. Therefore a method which isolates each peak was developed. Hoffmann *et al.* (2012) present a solution which uses peak group identification instead of retention time to determine peaks to align, this is then

combines with the time warping methods shown in (Nielsen *et al.*, 1998; Bylund *et al.*, 2002; Pravdova *et al.*, 2002; Skov and Bro, 2007). Although this method goes some way towards the peak isolation needed in this work it still relies on having one control experiment and peak shape similarity between experiments.

This research presents a new methodology for time alignment of single component chromatograms, through peak isolation. A methodology is provided in section 6.4 detailing how the peaks are aligned in this research. Once the peaks were aligned this allowed for the other pre-processing techniques to be investigated.

6.7.2 Methodology

When considering pre-processing techniques there are four distinct categories which the different methods can fall into. These are transformations, normalisation, filtering, scaling, and variable alignment. When processing data, the first method to consider is transformations, followed by normalisation methods. After this the filtering and finally the scaling methods can be applied. Table 6.5 summarises the results of a review of the literature, where the techniques which might be appropriate to this application are outlined. The table lists the category the technique falls under, the name of the technique, a brief description of the basic principles, the advantages and disadvantages, and finally references which provide either theory or application of the technique.

To perform the study each of the technique groups was investigated individually. Each of the normalisation methods being tested was applied to the data set and a PLS model was constructed. The RMSE and NRMSE values were calculated for each training batch and the validation batch to show how well the pre-processed data can be modelled. The model with the lowest RMSE and NRMSE values was selected as the appropriate technique. This was also compared against the values obtained from a model constructed with no pre-processing applied to ensure that it was necessary to apply the technique group, i.e. to check whether the data needed to be normalised at all.

The Akaike Information Criterion (AIC) was also used as it accounts for the model complexity as well as the model fit. Further information on AIC is provided in Chapter 4. In order to compare between the models constructed using different pre-processing techniques all the models contained 3 latent variables. The models all used a cross validation (leave one out) method.

Table 6.5: Summary table of the pre-processing techniques selected to be applied to the on-line measurements. The techniques are categorised into; normalisation, filtering, scaling, and variable alignment. For each section the technique, basic principles, advantages, limitations, and references are provided.

Pre-processing technique sub-heading	Technique	Basic principles	Advantages/Limitations	References
Normalisation	Normalise (1-Norm)	Divides each variable by the sum of the absolute value of all variables. $\mathbf{x_{corr}} = \frac{\mathbf{x_{org(i)}}}{\sum_{j=1}^{n} \lvert \mathbf{x_{org(i,j)}} \rvert}$	**Adv:** Applied to all variables as one **Lim:** Outliers can cause other variables to be incorrectly normalised.	Listgarten and Emili (2005) Rinnan *et al.* (2009) Eigenvector Inc. (2015)
	Normalise (2-Norm)	Divides each variable by the sum of the squared value of all variables. $\mathbf{x_{corr}} = \frac{\mathbf{x_{org(i)}}}{\sum \mathbf{x_{org(i,j)}}^2}$	See 1-Norm	Listgarten and Emili (2005) Rinnan *et al.* (2009) Eigenvector Inc. (2015)
	Normalise (Inf-Norm)	Divides each variable by the maximum value observed for all variables. $\mathbf{x_{corr}} = \frac{\mathbf{x_{org(i)}}}{max(\mathbf{x_{org(i)}})}$	See 1-Norm	Listgarten and Emili (2005) Rinnan *et al.* (2009) Eigenvector Inc. (2015)

Normalisation	Multiplicative Scatter Correction (MSC)-Mean	Sample chromatogram, $\mathbf{x_{org}}$ is regressed against a reference, $\mathbf{x_{ref}}$ and the fit used as a correction. $\mathbf{x_{org}} = b_0 + b_{ref,1} \cdot \mathbf{x_{ref}} + \mathbf{e}$ The data is then corrected using this reference $\mathbf{x_{corr}} = \frac{\mathbf{x_{org}} - b_0}{b_{ref,1}} = \mathbf{x_{ref}} + \frac{\mathbf{e}}{b_{ref,1}}$ In this instance $\mathbf{x_{ref}}$ is the mean sample.	**Adv:** Spectral features are preserved, whilst background offsets and slopes are largely removed. **Lim:** Selection of mean maybe inappropriate for data set. The corrected data is not representative of the relationship between the **X** and **Y** blocks, so MSC can remove information from **X** that is related to **Y** and vice versa.	Wold *et al.* (1998) Geladi *et al.* (1985) Rinnan *et al.* (2009) Dhanoa *et al.* (1994)
	Multiplicative Scatter Correction (MSC)-Median	See (MSC) mean. In this instance $\mathbf{x_{ref}}$ is the median sample.	**Adv:** See (MSC)-mean **Lim:** Selection of median chromatogram may not be representative of the whole data set. See (MSC)-mean for corrected data issues. **NOTE:** Both (MSC) mean and median were tested to determine which reference sample was most appropriate.	Wold *et al.* (1998) Geladi *et al.* (1985) Rinnan *et al.* (2009)

Normalisation	Standard Normal Variate (SNV)	Calculates the mean and std dev for a chromatogram, assigning weightings to the samples deviating the most from the mean. Entire sample is normalised by the std dev and a user definable offset, δ. $\mathbf{x_{corr}} = \frac{\mathbf{x_{org}} - a_0}{a_1}$ $a_1 = \sqrt{\frac{\sum_{j=1}^{n}(X_{i,j} - \bar{x_i})^2}{(n-1)}} + \delta^{-1}$	**Adv:** Known to work well for processes where similar signals are obtained for each sample. No common reference signal is required. **Lim:** Can be sensitive to noisy samples (outliers) **NOTE:** SNV and (MSC)-mean are similar but with SNV a common reference signal is not required. Instead, each sample is processed on its own, isolated from the remainder of the set.	Rinnan *et al.* (2009) Dhanoa *et al.* (1994) Li *et al.* (2004) Barnes *et al.* (1989)
Filtering	Smoothing (SavGol)	The SavGol smoothing filter is a 2nd order polynomial which is fitted within a window of defined size to smooth the data. $\mathbf{x_{smooth}} = \sum_{i=-n}^{i=n} C_i X_{i+n}$	**Adv:** Removes high frequency noise. **Lim:** If window is two large, it can over fit the data. Causing loss of important information.	Savitzky and Golay (1964)

Filtering	Derivative (SavGol)	To find the derivative at centre point, a polynomial is fitted within a window of the raw data. The parameters from this polynomial can then be used to calculate the derivative of any order for this function. This is applied to all points in the batch sequentially. The size of the window and the degree of the derivative are user defined.	**Adv:** Accentuates high frequency characteristics of the data, such as hidden peaks and removes low frequency features such as baselines **Lim:** Accentuates high frequency noise.	Savitzky and Golay (1964) Rinnan *et al.* (2009) Laxalde *et al.* (2011)
Scaling	Autoscale	First the data is mean centred, then each column is divided by the standard deviation of that column.	**Adv:** Each variable has equal influence of the output signal. **Lim:** The noise in signals is also given equal influence, can accentuate noise.	Detroyer *et al.* (2000) Hendriks *et al.* (2005) Eigenvector Inc. (2015)

	Group scale	Similar to autoscale, the main difference being that the variables are split into equally sized blocks and each block is scaled by the grand mean of their standard deviations. Within each block, the mean of the standard deviations for each variable included in the block is used to scale all columns.	See Autoscale **NOTE:** Group scale is typically applied when the variables are within equally sized blocks, and within these blocks they are for a given unit of measure.	Eigenvector Inc. (2015) Kukharev and Kuźmiński (2005) Hua *et al.* (2008)
Scaling	Mean centre (MC)	The mean of each variable is subtracted from the value recorded for the variable at each sample.	**Adv:**Reduces all data relevant to the mean. **Lim:** Not always sufficient for heteroscedastic data.	Moco *et al.* (2008) Shaw *et al.* (1993) Liu *et al.* (2003) Wang and Kowalski (1992)
	Median centre (MedC)	The median of each variable is subtracted from the value recorded for the variable at each sample.	**Adv:** Reduces all data so variation is relative to the median. **Lim:** Variable of interest can often be independent of the median value.	Laxalde *et al.* (2011) Eigenvector Inc. (2015)

	Pareto*	Measurements for each variable are divided by the square root of the variable standard deviation.	**Adv:** Scales variables so that noise level is equal for all, providing it is the magnitude of the scaling factor. **Lim:** Sensitive to large fold changes.	Ma *et al.* (2008) Gromski *et al.* (2014) Eigenvector Inc. (2015) Kim *et al.* (2015)
	Poisson*	Measurements for each variable are divided by the square root of the mean of the variable.	See Pareto	Keenan and Kotula (2004) Engel *et al.* (2013) Kim *et al.* (2015)
Scaling	Variance (std)*	Measurements for each variable are divided by the standard deviation of the variable.	See Pareto	Engel *et al.* (2013) Eigenvector Inc. (2015)
	Log Decay*	Each measurement is scaled by a continuously decreasing log function of the form: $\mathbf{x_{scaled,i}} = e^{\frac{-x_i}{n\tau}}$	**Adv:** Takes into account sensitivity of instrumentation. **Lim:** Typically used in mass spec. as opposed to LC.	Engel *et al.* (2013) Kim *et al.* (2015)
Variable alignment	Correlation Optimised Warping (COW)	Performs a piece-wise transformation of each sample adjusting the segments to best correlate to a reference sample.	**Adv:** Can align multiple peaks. **Lim:** Can compress and distort chromatographic data.	Engel *et al.* (2013) Eigenvector Inc. (2015)

***Techniques are also applied in series with mean-centering applied first**

6.7.3 Results

Table 6.6 shows the results of the pre-processing study. The first column of the table details the technique being applied in each model, the highlighted rows indicate the technique chosen as the best option for the pre-processing method (i.e. normalisation, filtering, and scaling). Also reported are the RMSE, NRMSE, and AIC values for both the training and validation batches; for the training batches the reported values are mean averages. The final column in the table is the model identifier which is used in the discussion to refer to specific models.

Although the table shows the mean values for the three assessment criteria for the training batches, there can be large variations within this. This is shown in Figure D5 (in Appendix D), where the AIC values are reported for models 2 and 3 (Table 6.6). It can be seen that there is large variation between the error for each batch, taking into account only the AIC values for the validation batch (15) then Figure D5 would suggest that the most appropriate techniques would be MSC median as this is the lower value. However Figure D5 also shows that the AIC values for the training batches (1-14) are larger for each individual batch and also overall (average). Therefore when performing model selection the values for both the average of all the batches and the value for the validation was used. In the example shown in Figure D5 (Appendix D) the training batch average was first be considered which for MSC mean was 65.4 and for MSC median was 92. The AIC values for the validation batch were subsequently considered; for MSC mean it was 85.8 and for MSC median it was 73.5. The difference between these two values was lower and therefore determining the best model involves considering both the best validation prediction but also the best cross validation of the training batches.

Table 6.6: Summary of the average training (denoted T) and validation (denoted V) RMSE, NRMSE, and AIC values calculated for each model. For each group of pre-processing options the optimum technique is highlighted in blue to show which techniques were used in the final model.

Model	Technique applied	RMSE		NRMSE		AIC	
		T	V	T	V	T	V
1	No pre-processing	236.7	206.6	4.690	0.450	95.2	85.96
	Normalisation						
2	MSC (mean)	75.2	204.8	0.138	0.440	65.43	85.83
3	MSC (median)	518.7	90.3	4.708	0.191	92.01	73.56
4	Normalise (1-Norm)	133.5	542.2	0.275	1.182	69.09	100.43
5	Normalise (2-Norm)	126.3	542.1	0.260	1.182	68.45	100.43
6	Normalise (inf-Norm)	143.0	541.7	0.301	1.182	71.05	100.42
7	SNV	99.2	541.4	0.192	1.182	66.39	100.41
	Normalisation + filtering						
8	Derivative 1st order	665.0	544.4	1.361	1.191	92.98	100.49
9	Derivative 2nd order	664.8	542.4	1.361	1.189	92.97	100.44
10	Smoothing (Savgol)	69.2	179.8	0.125	0.390	65.89	83.88
	Normalisation, Filtering, Scaling						
11	Autoscale	62.0	176.6	0.119	0.38	61.32	79.21
12	Group scale	62.0	176.6	0.119	0.38	61.32	79.21
13	Mean centre (MC)	69.2	177.1	0.125	0.38	64.04	83.65
14	Median centre (MedC)	72.7	161.1	0.136	0.35	65.33	82.23
15	Pareto	71.6	176.5	0.136	0.38	65.12	83.6
16	Poisson	97.1	176.9	0.174	0.38	65.58	83.63
17	Variance (std)	79.1	176.0	0.158	0.38	65.39	83.55
18	Log decay	105.1	185.2	0.200	0.4	66.95	84.32
19	Pareto + MC	71.7	176.7	0.136	0.38	64.09	83.61
20	Poisson + MC	71.9	176.7	0.136	0.38	64.09	83.62
21	Variance (std) + MC	62.0	176.6	0.119	0.38	61.32	79.21
22	Log decay + MC	64.0	177.0	0.136	0.38	64.48	83.64
	Variable Alignment						
23	COW	113.4	192.6	0.218	0.41	72.2	99.63

It was determined that the best normalisation technique to apply was MSC mean. A distinct improvement can be seen from model 1 where no pre-processing was applied. Similar improvements to the predictive ability of a model after the application of MSC were reported in the study conducted by Isaksson and Næs (1988) or Helland *et al.* (1995) with NIR data. Maleki *et al.* (2007) also used this technique for NIR, but with the application of MSC to on-line measurements for control. Their study is interesting as normally a model constructed using MSC is specific to the data set, whereas their study analyses each spectra as it is produced and then adds it to the data set. Thus it could be said it adapts as it improves the performance of the model with each batch.

The next class of pre-processing techniques investigated were filtering methods, see models 8-10 in Table 6.6. MSC (mean) was applied to both data matrices before the techniques detailed in models 8-10 were tested. It can be seen from Table 6.6 that it was not necessary to take the derivative of the data. The application of the derivative makes the predictions worse therefore it was not applied in further models. The smoothing filter, on the other hand, showed a slight improvement to the model (model 10). The application of smoothing filters in literature was shown to also provide improvements in model predictions (Reed *et al.*, 2011). Luna *et al.* (2013) conducted a study of pre-processing techniques to determine the most appropriate method for their application to NIR spectroscopy, and applied these techniques in further work (Luna *et al.*, 2015). Although these studies were not just focused on smoothing techniques, it can clearly be seen that Savgol smoothing has a beneficial effect on the final model, hence the smoothing filter was included in further models.

The final class of techniques considered was filtering methods, see models 11-22 (See Table 6.6). These models were constructed with MSC (mean) and smoothing functions having been applied. The applied scaling techniques only offered relatively small improvements to the data. Dalal and Zickar (2011) made an interesting point that quite often in MVDA scaling techniques such as mean centering are applied to data without really understanding the need. Often mean centering is used to reduce collinearity

between variables (Kim *et al.*, 2015), and in the application to chromatography this is essential. As previously discussed the models constructed in this study are comprised of components which represent a measure of the correlation between the original variable data. As was shown in the PCA analysis, the on-line measured variables do not capture the variation in the data, therefore it is likely that the variation is correlated to a variable which is not measured such as the binding kinetics of the protein and the ligand. As Craig *et al.* (2006) stated scaling data gives equal weight to each data value in a time series. This means that systematic changes with small variance can be more easily detected. This is beneficial in this application as some elution peaks may be broad and shallow whilst others are sharp and intense. Scaling gives equal weight to each data point and thus the small changes can be seen as well as large ones. It can be seen that the lowest AIC values are for the models which used autoscale, group scale, and variance (std) with mean centre. This is expected as these three methods are all based on the same principle. Autoscale mean centres the data and then scales using the standard deviation, group scale does this for blocks of variables, and variance (std) with mean centre is just applying the techniques separately. This explains why the AIC values obtained are so similar. Additionally group scale is designed for multi-way models where the variables are in equally sized blocks. The user defines the sizes of these blocks and group scale then treats each variable individually within the assigned block using mean centre and variance scaling techniques to counteract the effect the different measurement techniques used for the variables. In the data set used here the data was manually decomposed into a 2D matrix prior to the application of the techniques. This removes the need for group scaling, as a result the group scale technique is performing in this study in the same way as autoscale.

Variable alignment is used to align the elution peaks with in the **Y**-block and align similar data patterns in the **X**-block. Although this technique was explored in this study the peak isolation process discussed in section 6.4 aligns the peaks over 332 sample points. The additional variable alignment over-corrects the data by aligning small inconsistencies between the batches. An example of this would be the distortion in the peaks observed in batches 4

and 9 as shown in Figure 6.12. The variable alignment technique COW was applied, which has a narrow window size of 10 sample points. It is assumed that the poor fit observed in this model (23) is due to the algorithm attempting to align small inconsistencies in the shape of the elution peaks across the small sample size, for example the highlighted bump in batches 4 and 9. However these peaks shown in the data are more likely to be cause by extra resolution, showing a second peak starting to separate. This is shown as a shoulder in the main peak.

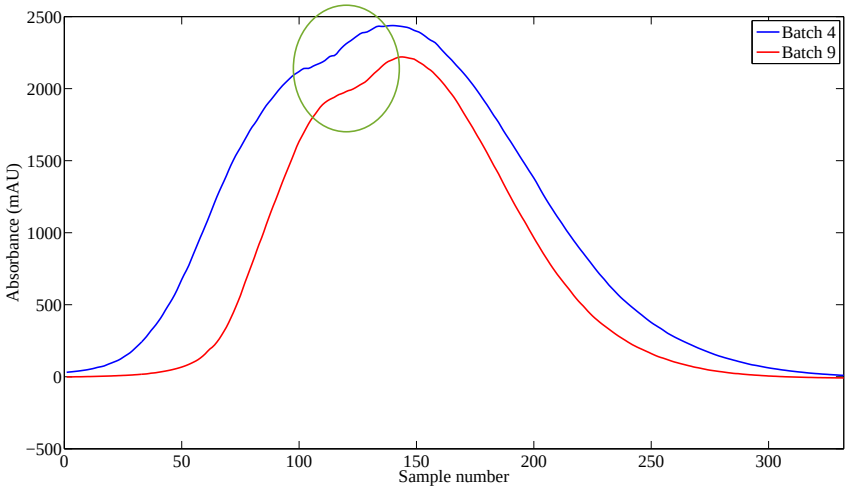


Figure 6.12: Inconsistent peak shape shown between batches 4 (blue) and 9 (red).

When the measured and predicted values are then observed for model 23 in Figure 6.13 it can be seen that the predicted values now attempt to align to the points of change located at the bump. This effect is particularly noted in the low yield batches where small changes to he peak shape are amplified as the elution peak occurs over a small absorbance range.

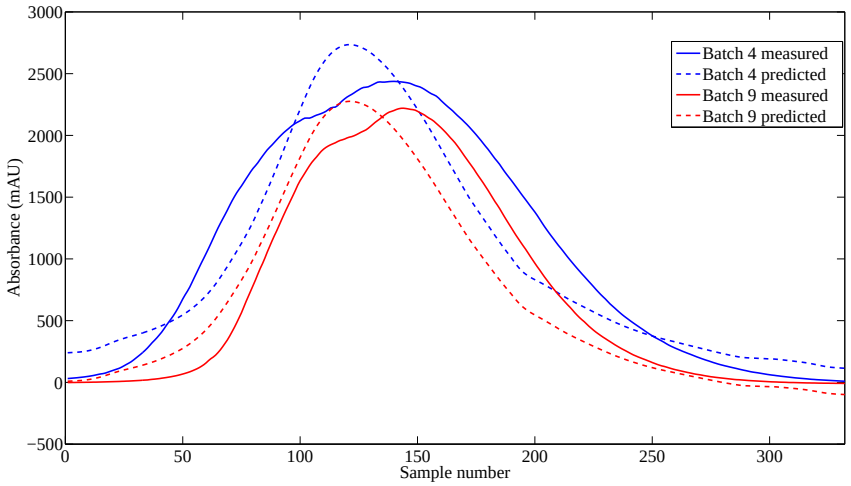


Figure 6.13: Inconsistent peak shape shown between batches 4 (blue) and 9 (red).

The techniques identified as optimum are replicates of each other. With the best pre-processing techniques to use for this application being identified as those in model 11.

1. Normalisation - MSC (mean)
2. Filtering - Smoothing (SavGol)
3. Scaling - Autoscale
4. Variable alignment - N/A

The average RMSE, NRMSE, and AIC values for model 11 across all batches were 62.0, 0.119, and 61.32, respectively. This combination of pre-processing techniques showed the smallest variation between the predictions for the training batches and the prediction for the validation batch. The measured data and the prediction for the training batches and validation batch are shown in Figures 6.14-6.17. The training data has been split into three figures showing the high, medium, and low yield batches for ease of interpretation. Batches 1 and 6 on Figures 6.15 and 6.16 respectively show that the ability of the model to predict the centre points is poor. One possible reason maybe the wide range of outputs recorded for the centre points, although it would be expected that they would of been roughly the same. This could possibly account for the poorer fit of the model to the validation batch (batch 15) Figure 6.14.

Discussion of the variation observed between the centre point batches is provided in section 6.10.

Additionally it can be seen that the model struggles to predict batches 4, 9, 12, and 13. These batches do not display the expected Gaussian peak, which the rest of the batches do. The model is fitting a Gaussian peak to these four batches as well, and is a limitation of the PLS algorithm which is discussed further in the next section.

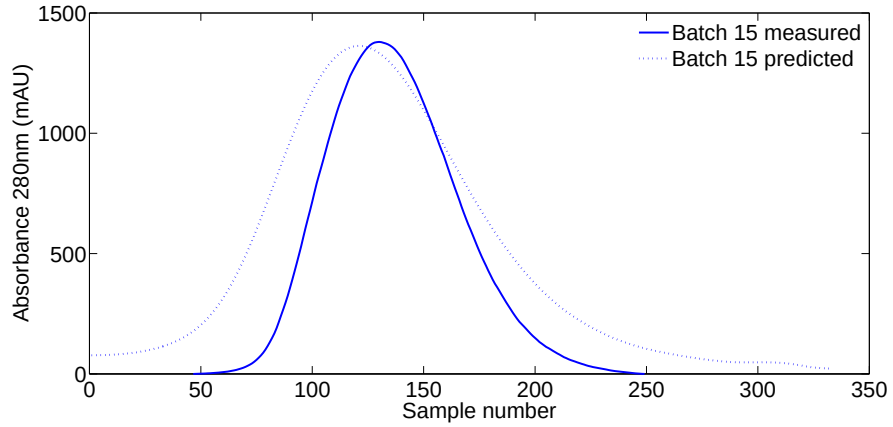


Figure 6.14: Measured and predicted absorbance for validation batch (15) using the pre-processing techniques described in model 11 (Table 6.6)

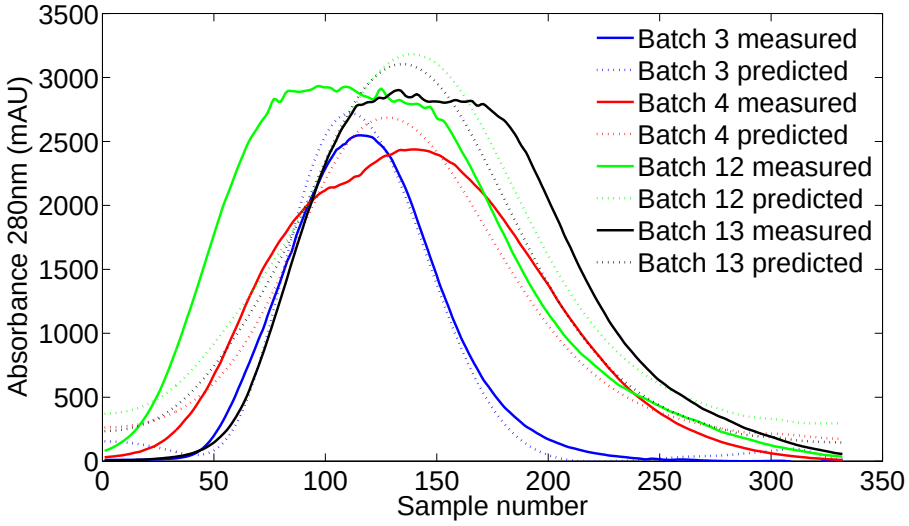


Figure 6.15: Measured and predicted absorbance for the high yield batches (3, 4, 12, and 13) using the pre-processing techniques described in model 11 (Table 6.6)

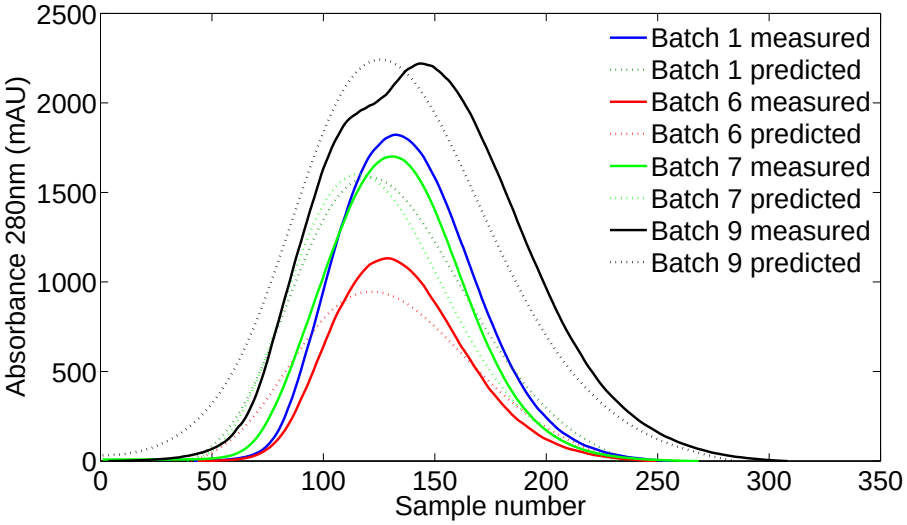


Figure 6.16: Measured and predicted absorbance for the medium yield batches (1, 6, 7, and 9) using the pre-processing techniques described in model 11 (Table 6.6)

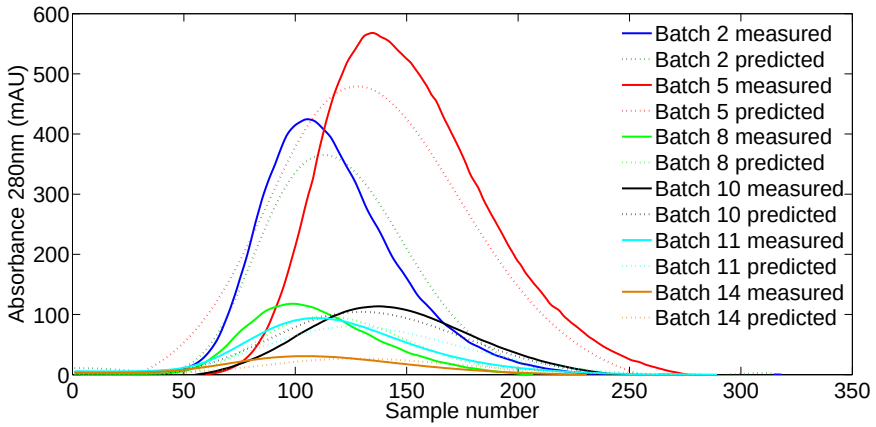


Figure 6.17: Measured and predicted absorbance for the low yield batches (2, 5, 8, 10, 11, and 14) using the pre-processing techniques described in model 11 (Table 6.6)

In summary this study has investigated 23 separate pre-processing methods which are commonly applied in chemometric and chromatographic applications to the lactoferrin data set. The data set was split into training and validation, due to the small size of the set 14 batches were used as training and 1 as validation batch. The methodology of applying each technique and assessing it using a PLS model and model assessment criteria was taken from literature, with the intent of evaluating which pre-processing techniques were best to apply to a scouting data set. For this application the best results were shown to be normalisation using multiplicative scatter correction (mean), Savitzky-Golay filtering (smoothing), and centering and scaling the data using

autoscale. The information used in this study was taken forward and applied in the next section, which is concerned with the multivariate modelling of the chromatography data.

6.8 Multivariate modelling of chromatographic data

This section presents the models constructed to predict the performance of the purification using the CPP of yield. The models presented are described as a proof of concept to show whether it is possible to both predict CPPs from the operating conditions, and subsequently if the yield is known predict the operating conditions which would achieve that yield. Of the three most commonly used latent variable methods, as given by Rajalahti and Kvalheim (2011b) the first, principal component analysis (PCA), was presented in section 6.6. The second, partial least squares (PLS) is presented in this section, with the final method being principal component regression (PCR). As shown by Hemmateenejad *et al.* (2007) PCR and PLS are comparative methods, with both providing similar predictions. However as Hemmateenejad *et al.* (2007) showed in their study PLS is a better tool to use with absorbance data. Additionally they showed that to obtain similar results from both PLS and PCR then generally more factors are required in the PCR model making it more complex. This section aims to describe the use of PLS to extract relevant information from the measured data. As Rajalahti and Kvalheim (2011b) states, multivariate methods can be used to simplify complex pharmaceutical data and thus make visualisation easier. This fits in with the PAT initiative as it aims to increase process understanding and control whilst at the same time reducing the uncertainty and variation in the end product (United States Food and Drug Administration (FDA), 2004a). With the objective to build quality throughout the manufacturing process, also referred to as Quality by Design (QbD), multivariate data analysis has a central role in the PAT initiative (von Stosch *et al.*, 2014a).

The current literature on MVDA applications to chromatography appears to be very much dominated by liquid chromatography/high-resolution mass

spectrometry (LC/HRMS) and gas chromatography (GC) (Krishnan *et al.*, 2013a; Skov *et al.*, 2006). Only limited references to IEX chromatography are available currently, with the most relevant to the work presented in this section being Bro *et al.* (1999) who used PARAFAC to model chromatographic data which contained retention time shifts. This is a similar issue as shown in this data set, where the retention times change. However PARAFAC as a modelling tool was explored in Chapter 5 but it was found that PLS was a better predictive tool, further confirmed by Bro (1997) and as such was not explored further in this work.

A study of the currently available modelling literature for IEX chromatography would not be complete without also referencing the work produced around the ChromX project from the Karlsruhe Institute of Technology (KIT). This project aims to produce a toolbox that solves mechanistic chromatography models and produce a better understanding of what occurs in the column. There are various publications such as Huuk *et al.* (2014) that are related to IEX chromatography.

The research presented in this section differs from the currently published models in that it aims to use a database of chromatographic data to characterise the design space and from this predict both how the column will operate, but also based upon final product measurements the optimum conditions for the column to operate at. This section presents the first stage, as it looks at assessing how well PLS models can predict these aspects.

6.8.1 Results and discussion

Off-line data

The PCA analysis of the off-line data presented in section 6.6 highlighted how the differences in batch yield are influenced by the operating conditions specified in the DoE. From this a PLS model was developed using the DoE (operational set points for flow rate, load pH, gradient CV, load concentration, and elution pH) data as the input and the elution yield as the output. The result

of this model is shown in Figure 6.18, for a model constructed using 3 latent variables (LVs) which accounted for 73.33% of the cumulative **X**-block variance and 93.69% of the cumulative **Y**-block variance with the number of LVs included determined by cross validation of the root mean squared error of calibration (RMSEC).

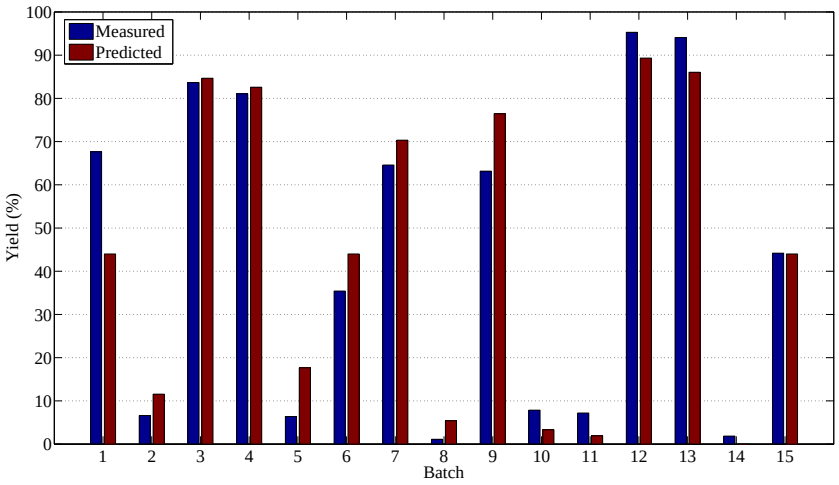


Figure 6.18: Elution yield percentages determined using a PLS model containing 3 LVs. Batch 15 is the validation batch with batches 1-14 being used to train the model, measured data is shown in blue and predicted data is shown in red.

Figure 6.18 shows that for 10 of the training batches the predictions were accurate to $\pm 5\%$. One interesting observation is the predictions made for the three centre points (replicates) which all predicted a value of 43.98% yield, which is as would be expected as the data in the input (**X**-block) for these runs is identical. This further highlights that there were influences on the system which have not been measured, (see discussion in section 6.10). Putting aside the centre point batches (1, 6, and 15) it can be seen that poor predictions were made for batches 11, 12, 13, and 14.

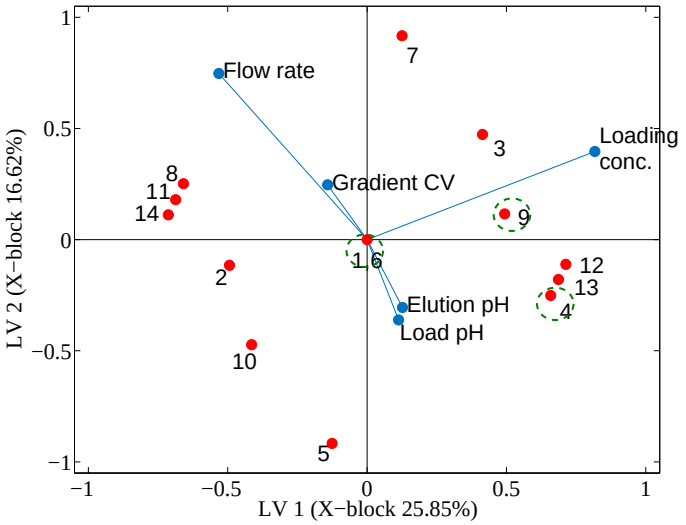


Figure 6.19: Bi-plot for off-line yield prediction showing the scores (red dots) and loadings (blue dots) values for LV1 and LV2 for the X-block data. Highlighted in the green ellipses are the batches which showed poor predictions.

Figure 6.19 shows a bi-plot (information on construction of bi-plot provided in Chapter 4) of the scores and loadings for the **X**-block data. There is a symmetry to the scores with batches that have the opposite settings for all DoE variables being mirror images of each other. With regards to being able to determine why the predictions for batches 1, 4, 6, and 9 are slightly poorer than for other batches the **X**-block data does not provide any answers. However Figure 6.20 which shows a bi-plot for the **Y**-block data does show a clear pattern. LV1 (explaining 87.88% of the variation) is clearly related to the yield percentage, this is shown in Figure 6.20 through the colour coding of the batches; high yield (red), medium yield (yellow), and low yield (blue).

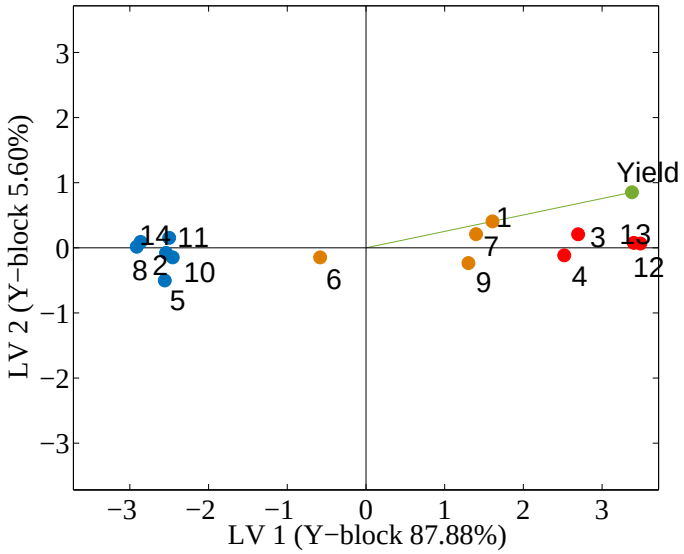


Figure 6.20: Bi-plot for off-line yield prediction showing the scores and loadings values for LV1 and LV2 for the Y-block data. Shown in red are the batches with a high yield, in orange are the medium yield batches, and in blue are the low yield batches.

From the PCA analysis it is known that the yield is highly correlated to the loading concentration of protein. Referring to the loadings concentration and elution yield given in Table 6.7, it can be seen that in general a loading concentration of 30 mg/ml produces a high or medium yield batch, and a load concentration of 10 mg/ml produces a low yield batch. This suggests a linear relationship between load concentration and total elution yield. This would imply that the other characteristics of the elution peak i.e. retention time, peak width (which for multi-component systems is a reflection of the resolution), and peak shape (is it symmetrical) are influenced more by the other operating parameters (load pH, elution pH, flow rate, and gradient CVs). There is further evidence to suggest this in Figures 6.3 and 6.19; Figure 6.3 shows that batches 4, 12, and 13 all have distinctive shapes, especially at the peak maxima as they do not follow the Gaussian curve. With this in mind, considering Figure 6.19 these three batches (4, 12, and 13) are all grouped together with influences shown from the load and elution pH.

For batches 2, 5, 11, 10, and 14 in Figure 6.3 the peaks are all asymmetrical, which is again reflected in Figure 6.19. The distinction as to one variable being responsible is not clear, instead it is suggested that flow rate, gradient CV, load pH, and elution pH all equally influence the asymmetry of

Table 6.7: Load concentrations and yield obtained during elution for all batches. The batches have been categorised as high yield ($y < 50\%$) in red, medium yield ($10 < y < 50\%$) in yellow, and low yield ($y < 10\%$) in blue

Batch	Load concentration	Elution yield (%)	Discrete value
1	20	67.64	0
2	10	6.59	-1
3	30	83.65	1
4	30	81.06	1
5	10	6.37	-1
6	20	35.41	0
7	30	64.57	0
8	10	1.11	-1
9	30	63.15	0
10	10	7.85	-1
11	10	7.19	-1
12	30	95.29	1
13	30	94.06	1
14	10	1.84	-1
15	20	44.17	0

the peaks.

A second PLS model was constructed with the aim of determining whether a model could predict if the batch has a high, medium, or low yield. To achieve this the batches were assigned a discrete value of either 1, 0, or -1 depending on if they are classed as high, medium, or low yield. These values are shown in column 4 of Table 6.7. The model was developed using these discrete values as the outputs to determine whether the operating conditions would give a high, medium, or low yield. This analysis could often be carried out using a clustering method, however, as Rokach and Maimon (2005) state the goal of clustering is to discover a new set of categories. Whilst classification methods (as used here) are predictive tool where the classes are predefined.

The first PLS model used 14 of the batches as training batches, and one as a validation batch. As the aim is now to predict one of three categories, the data was split into 12 training batches and 3 validation batches, representing the high (batch 3), medium (batch 15) and low (batch 5) yield categories. The **X**-block and **Y**-block of the resulting model captured 66.9% and 94.7% of the

resulting variation respectively. Figure 6.21 shows the resulting predictions that were made using the model. For ease of interpretation of the graph the measured data has been offset by +0.05 as the lines were overlying. As can be seen the model manages to predict both the training and validation batches with a high degree of accuracy. The only anomalous result was batch 5, which was measured at a low yield but was predicted to have a medium yield. Referring back to the DoE batch operating conditions shown in Table 6.4 there are no obvious causes for why the prediction for batch 5 should be poor.

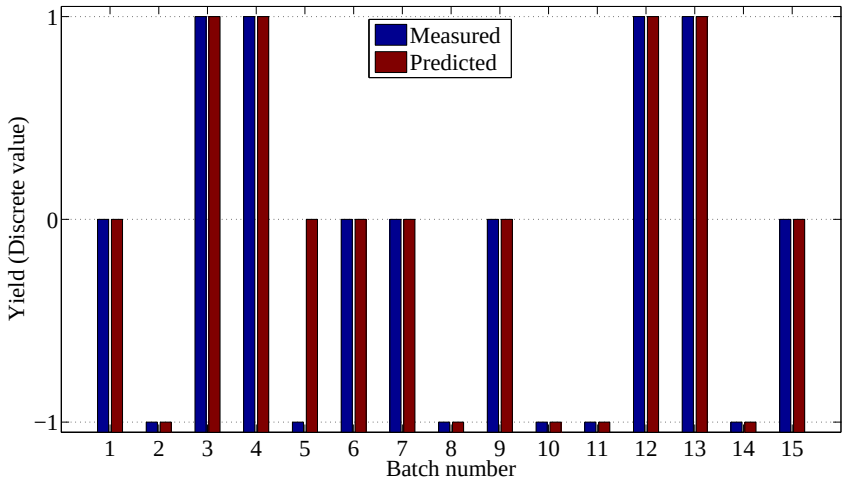


Figure 6.21: Yield predictions for discrete yield values for off-line data prediction. Using 12 batches as training and 3 as validation (3, 5, and 15). Measured values are shown in blue and predicted values are shown in red.

On-line data

The literature review presented in section 6.7 detailed the selection method for choosing the pre-processing steps to use with the on-line data. In that study the PLS model used to assess the different techniques was constructed from an **X**-block containing the on-line variables conductivity, concentration, and pH and the **Y**-block contained the on-line absorbance data. From the predictions presented in Figures 6.14 - 6.17 it can be seen that they are not ideal. Issues present across the predictions made for all batches include; inaccurate peak height, width, and baseline. It is important to obtain a good prediction for the elution peak as it provides information on yield, resolution and behaviour of

the purification. Therefore another model was constructed to predict this.

From conducting the study in section 6.7 it is apparent that PLS as a tool can not model data where the general pattern of the **X** and the **Y** block data is different. To illustrate this point, the conductivity had a linear slope, whereas the absorbance data had a Gaussian slope. Therefore two methods were be used to try to improve the on-line PLS predictions: the prediction of the cumulative area under the curve, and the prediction of the slope of the curve.

Area under curve

The **X** and **Y** block data was taken from the start of the elution peak till the end of the elution peak as described in section 6.4. Subsequently trapezoidal numerical integration was performed on each successive sample point to obtain the area under the curve, the values were recorded as a cumulative area. To perform the integration the 'trapz' function in matlab was used, which takes the specified sample range and breaks down the area in trapezoids and calculates this area. Information on the application of the trapz function within matlab can be found at MathWorks (2015), with explanation of the background being given by Dautray and Lions (2000).

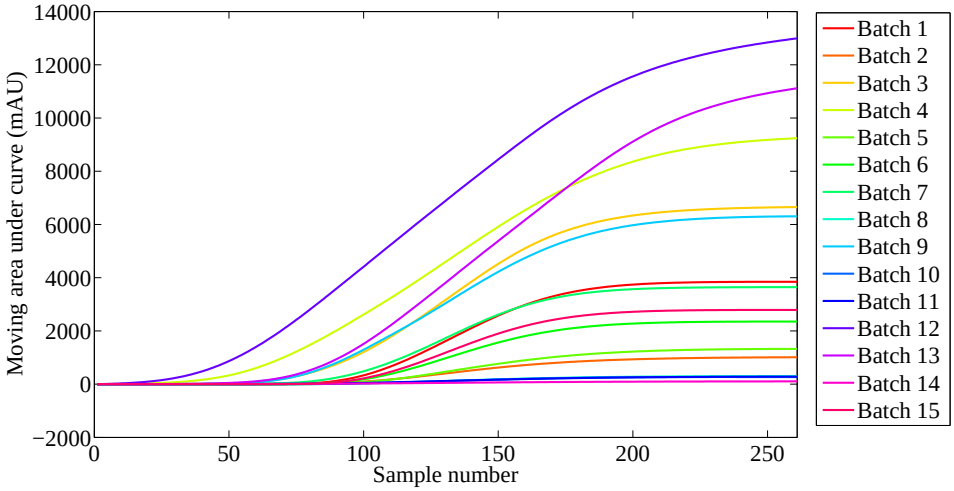


Figure 6.22: The cumulative area under the curve calculated using the trapezoidal numerical integration function within matlab for each of the 15 batches.

Figure 6.22 shows the cumulative area under the curve, with the data now

more closely resembling a sigmoid curve. This pattern now more closely resembles the pattern shown in the **X**-block data, hence a PLS model was constructed. The resulting model contained 5 LVs which accounted for 99.02% of the variation in the **X**-block data and 61.11% of the variation in the **Y**-block data. Figures D6 - D8 show the predictions for the training data set and Figure 6.23 shows the measured and predicted data for the validation batch.

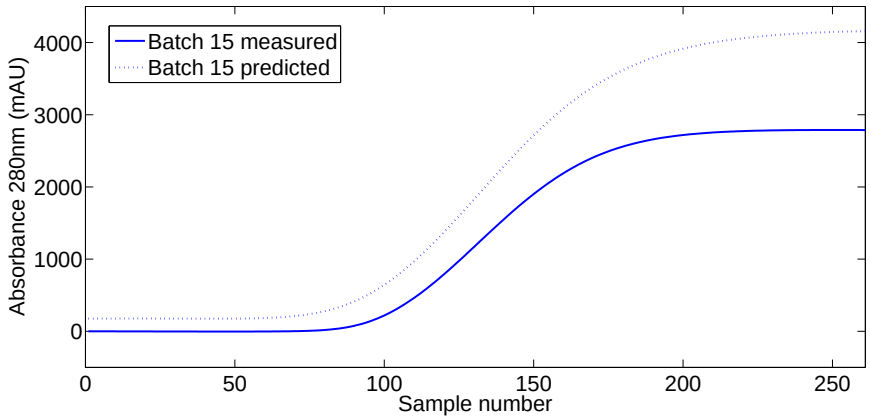


Figure 6.23: Measured and predicted values for the validation batch (batch 15) for the PLS model constructed to predict the cumulative area under the curve.

As can be seen the prediction for the validation batch over estimates the peak area quite considerably. This suggests that the high yield batches (3, 4, 12, and 13) might be exerting a greater influence over the system. To investigate this, a bi-plot of the scores and loadings for LV1 and LV2 in the **Y**-block data was constructed (Figure 6.24).

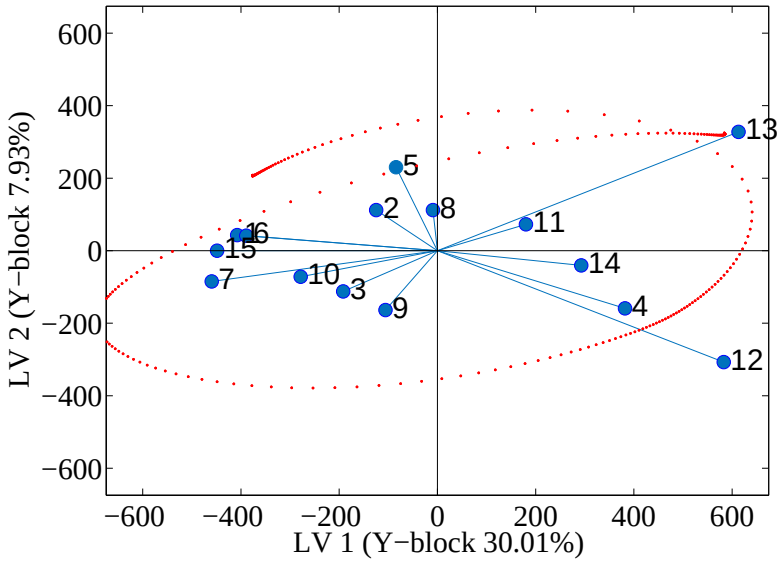


Figure 6.24: Bi-plot for the Y-block data for LV1 and LV2, showing the loadings for the Y-variable in red and the scores values for all 15 batches in blue.

The bi-plot shows that batches 12 and 13, which are the two highest yield batches, have a strong influence on the predicted area under the curve. Referring back to Figure 6.22 showing the original measured data for the area under the curve, it can be seen that batches 12 and 13 have a different shape to the other 13 batches. This is due to both the high amount of protein loaded onto the column and high amount of this binding to the resin. This produces a much wider peak with a broader and flatter inflection point. A second model was then constructed with these two batches removed, the model contained 5 LVs and accounted for 98.97% of the **X**-block variation and 63.98% of the **Y**-block variation.

Figures D9 - D11 show the predictions for the training runs and Figure 6.25 shows the measured and predicted values for the validation run. Which are much better than the predictions presented in section 6.7 (Figures 6.14 - 6.17). Figure 6.26 shows the area under the curve transposed back into absorbance values. Additionally values for RMSE, NRMSE, and AIC given in Table 6.8. The predicted data for the validation batch is lower than the measured data, which is most likely due to the fact the validation batch was one of the centre point batches and the three centre point batches varied considerably in their outputs. When comparing the RMSE, NRMSE, and AIC

values obtained for the mean of the training predictions and the validation batch, it can be seen that there is not much difference between them. Therefore it was determined that the model is adequate for the data set.

Table 6.8: RMSE, NRMSE, and AIC mean values for training runs and final values for validation batch for PLS model constructed to predict the area under the curve.

	RMSE	NRMSE	AIC
Mean for training batches	1616	4.06	91.82
Value for validation batch	1642	3.40	102.25

This model has been shown to give good predictions for batches 1-12 and 15. However, the model cannot predict accurately the batches which have a significantly higher yield. This suggests that a 'one model fits all' approach is not applicable in this scenario. One possibility would be to conduct further experiments particularly focused on the operating conditions which produce the high yield and use these to model for yields $> 90\%$. However for industry the aim would be to reduce the number of experiments conducted as the biopharmaceutical industry is highly competitive, and improved time-to-market can mean a significant business advantage over competitors. Therefore other modelling techniques were investigated to explore whether they can handle the higher yield batches.

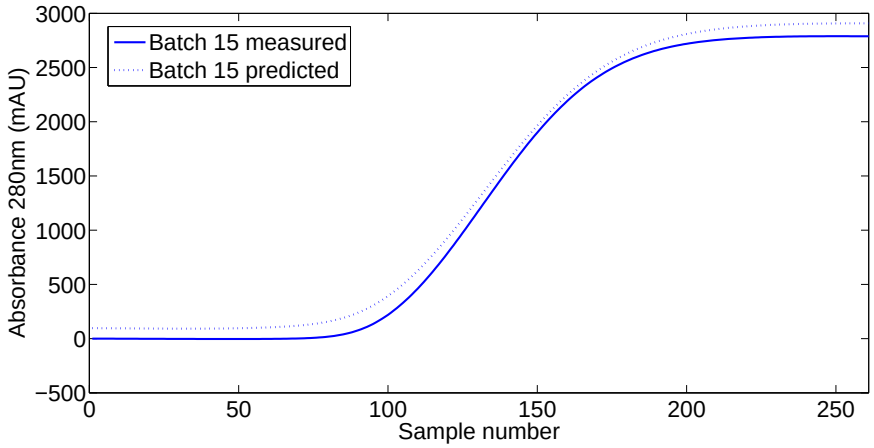


Figure 6.25: Measured and predicted values for the validation batch (batch 15) for the PLS model constructed to predict the cumulative area under the curve.

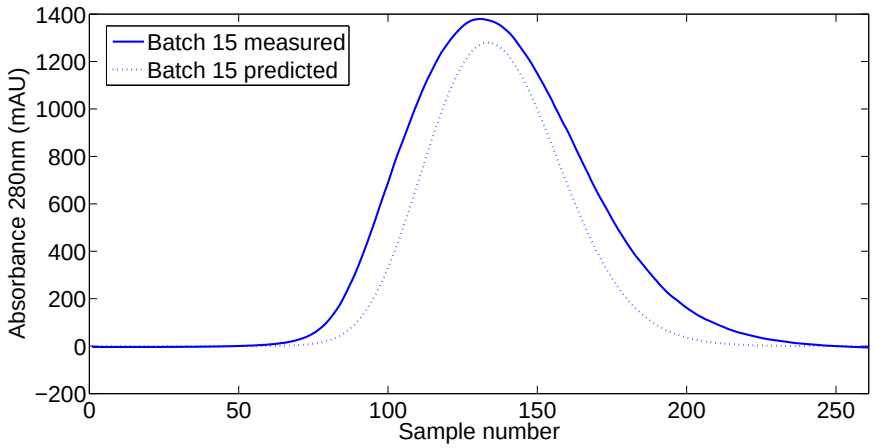


Figure 6.26: Measured and predicted values for the validation batch (batch 15) for the PLS model constructed to predict the cumulative area under the curve.

Slope of curve

The second method investigated to transform the curve data into a usable form was to calculate the slope of the curve at set sample points. This technique was explored to establish whether it could account for both the lower yield and the high yield batches which the area under the curve method had difficulties with. The on-line variables of conductivity, concentration, and pH formed the **X**-block, with the slope of the absorbance curve forming the **Y**-block. Initially a model was constructed using all of the **X** and **Y** block data. The gradient was determined using Equation 6.5.

$$slope = \frac{y_2 - y_1}{x_2 - x_1} \tag{6.5}$$

where y_1 and y_2 are the y-axis co-ordinates for the first and second points, respectively, and x_1 and x_2 are the x-axis co-ordinates for the first and second points, respectively. Figure 6.27 shows the data for the gradient of the absorbance (mAU/min) prior to pre-processing.

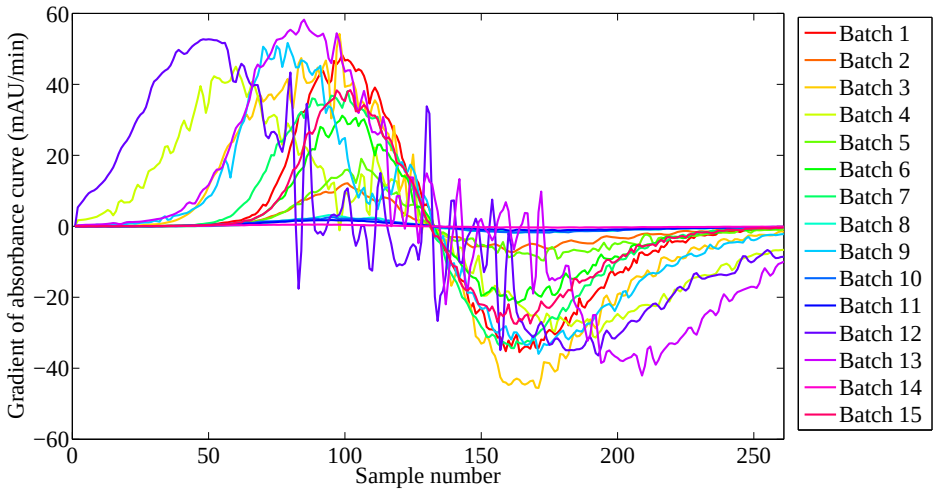


Figure 6.27: Gradient of the absorbance slope for each batch, shown prior to pre-processing. Calculated using the relationship given in Equation 6.5.

The model constructed for this data contained 5 LVs and accounted for 99.27% of the variation in the **X**-block and 64.78% of the variation in the **Y**-block. Figure 6.28 presents the measured and predicted data for the validation run, with the measured and predicted data for the training runs given in the appendix in Figures D12 - D14. The measured data is presented after the savgol smoothing filter was applied for ease of interpretation. As can be seen from Figures D12 - D14 this model struggled to predict for data which deviated from the Gaussian curve, for example batch 9 on Figure D13. The distortion which was identified in the initial half of the curve in Figure 6.12 again caused issues here. In total, five of the training batches (4, 6, 9, 12, and 13) were not predicted with sufficient accuracy.

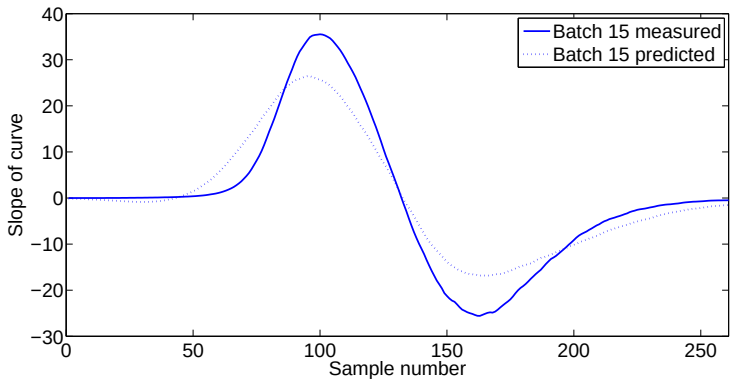


Figure 6.28: Measured and predicted values for batch 15 (validation batch) The plotted measured data is shown after a Savgol smoothing filter has been applied for ease of interpretation.

Using the bi-plot shown in Figure 6.29 for the interaction between batches and output variables, it can be seen again that batch 12, and to some extent batch 13, influence the model, similarly to the effect observed when modelling the area under the curve.

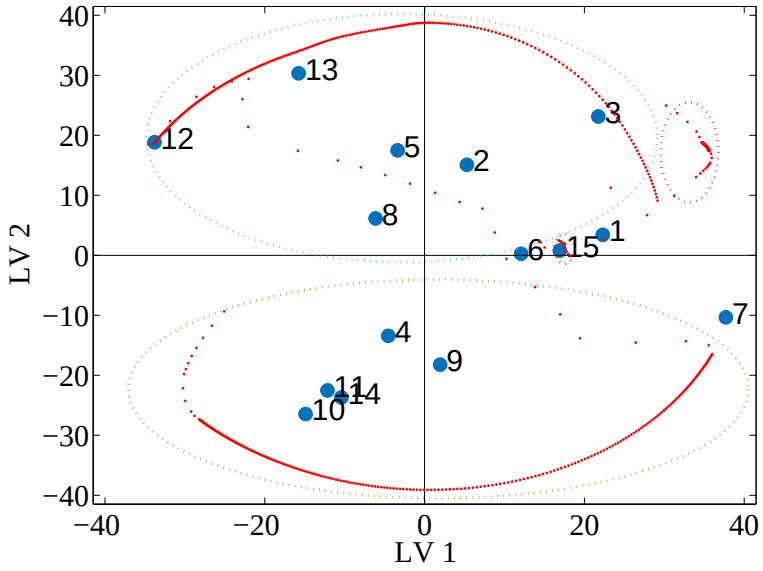


Figure 6.29: Bi-plot for the Y-block data for LV1 and LV2, showing the loadings for the Y-variable in red and the scores values for all 15 batches in blue.

A new model was constructed with batches 12 and 13 removed from the training data set. The resulting model contained 5 LVs and explained 99.01% of the variation in the **X**-block and 56.33% of the variation in the **Y**-block. As can be seen from Figure 6.30, the prediction of this model for the validation batch was a much better fit. The model managed to predict the initial and final inflection points accurately, these points are representative of when the protein begins and finishes eluting. Which would theoretically mean that for multi-component systems the resolution would be better predicted. Figures D15 - D17 show the predictions for the training runs using this model. Although these are accurate for most batches, it appears there is still an issue in predicting the unusual curve shape seen in batches 4 and 9.

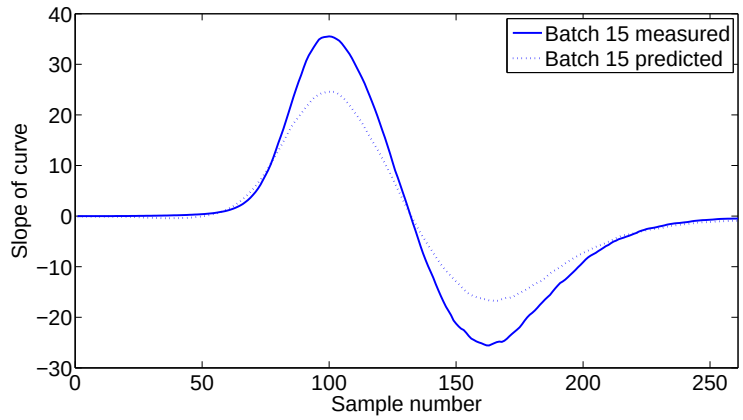


Figure 6.30: Measured and predicted values for batch 15 (validation batch). The plotted measured data is shown after a Savgol smoothing filter has been applied for ease of interpretation.

Figure 6.31 shows the predicted gradient for batch 15 transformed back into the absorbance curve of the measured data, with the RMSE, NRMSE, and AIC values for this model reported in Table 6.9. As can be seen the model accurately predicts the peak width but struggles with predicting the peak height. This is in contrast to the model built using the area under the curve (Figure 6.26) which showed a more accurate prediction of peak height with a worse prediction for the peak width. This suggests that a possible step forward could be combining both of these models. Revisiting the issue of the

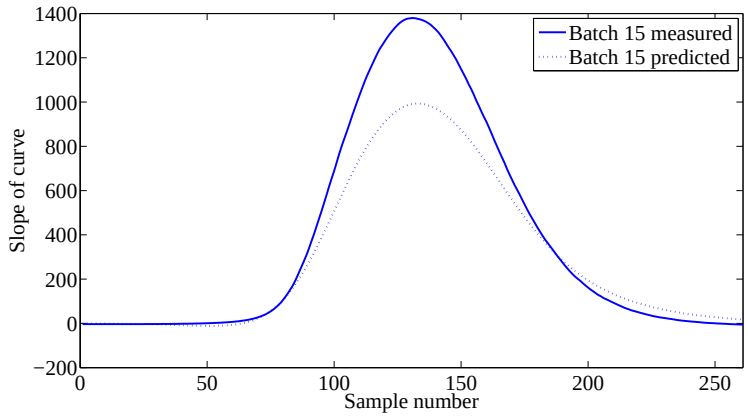


Figure 6.31: Measured and predicted values for batch 15 (validation batch), showing the predictions made from the PLS curve gradient transformed back into the absorbance profile.

predictions for the training batches 4 and 9, it can be seen from Figure 6.32 that the model is attempting to fit a Gaussian curve to the data. The distortion from this curve at the start of the elution peak is not accounted for, as it only appears in two batches. It is assumed that this is not a key part of the elution

Table 6.9: RMSE, NRMSE, and AIC mean values for training runs and final values for validation batch for PLS model constructed to predict the gradient of the slope.

	RMSE	NRMSE	AIC
Mean for training batches	246.82	0.35	57.61
Value for validation batch	159.10	0.32	71.90

and were instead a result of experimental error, therefore further models to predict this were not produced.

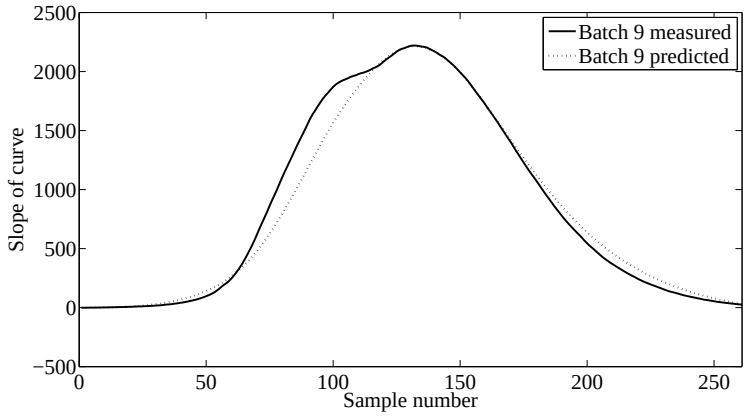


Figure 6.32: Measured and predicted values for batch 9, showing the predictions made from the PLS curve gradient transformed back into the absorbance profile.

In summary it is possible to predict the final elution yield using a PLS model where the operating conditions are the **X**-block data with high accuracy. The on-line elution profile presents more significant modelling challenges, however two possible solutions were shown to predict different aspects of the curve well. The next step was the evaluation of whether these models can be combined.

6.9 Combining multivariate models for prediction of elution peak

This section combines the PLS models presented in the previous section. The aim was to produce a prediction for the elution peak which was more accurate

than the predictions for the two models separately. The simplest method as described by Hastie *et al.* (2009) is to take the mean of the model predictions. This could present issues in this case, where the two models predict different parts of the time series data with higher accuracy. Therefore a more common approach to use is to have a measurement of the fit of the predictions at each time point. This has been done by Neuman (2003) who used the Bayesian information criterion as the measure and from this selected the prediction at each sample point which was most accurate. As similar method was adopted by Pan *et al.* (2006) who used the Akaike information criterion as the measure of accuracy.

6.9.1 Methodology

As discussed in section 6.8 the predictions of the two models constructed using the area under the curve and the gradient of the slope showed significant improvements on the model constructed in the pre-processing study which used the on-line absorbance measurements as the model output. This section reports attempts of improving the area under the curve and gradient models further by combining them into one prediction.

Two different methods were used: the first used a combined average of the predictions, with the second using an AIC weighted method. The basic premise for both of these methods is shown in Figure 6.33.

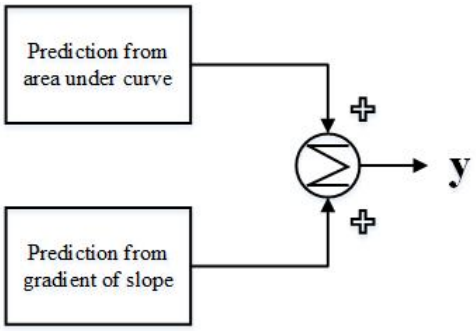


Figure 6.33: Basic premise for combining multiple PLS models into one prediction.

To produce the first of these models, the average prediction of each of the 261 sample numbers were taken and used as the final value. For the second

model the AIC value was calculated for both models at each sample point, the lowest AIC was identified and the value corresponding to this was taken as the final value. This method is similar to that described by Hastie *et al.* (2009), which was applied for example in the work of Pan *et al.* (2006).

6.9.2 Results and discussion

As the area under the curve PLS model (henceforth referred to as area model) predicted the peak height accurately, and the gradient of the slope PLS model (henceforth referred to as gradient model) predicted the peak width accurately, the combined models were used to produce improvements to these. The resulting prediction for the average method (henceforth referred to as average model) showed a significant improvement on the two previous models, with the AIC weighted method (henceforth referred to as AIC model) providing further improvements. Figure 6.34 shows the final predictions for the validation batch, and Table 6.10 provides the RMSE, NRMSE, and AIC values for the four models.

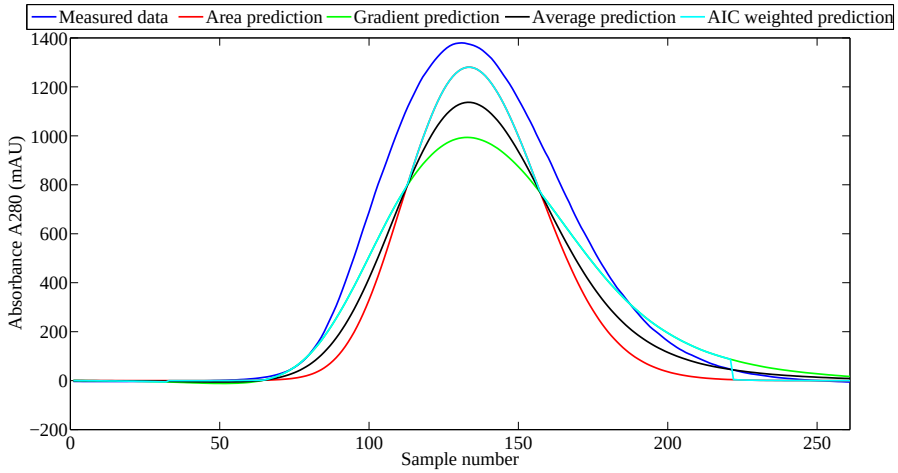


Figure 6.34: Predictions for batch 15 (validation batch), blue line shows original measured data, red line shows the prediction made using the PLS area model, the green line show the prediction made using the gradient PLS model, the black line shows the predictions from the average model, and the cyan line shows the predictions from the AIC weighted model.

As can be seen from Figure 6.34 the average model has improved the

Table 6.10: RMSE, NRMSE, and AIC values for the area under the curve PLS model, the gradient of the slope PLS model, the average hybrid model, and the AIC weighted hybrid model for the validation batch (batch 15).

	RMSE	NRMSE	AIC
Area PLS model	155.50	0.32	71.60
Gradient PLS model	159.10	0.32	71.90
Average hybrid model	144.26	0.29	70.63
AIC weighted hybrid model	101.10	0.20	66.01

prediction for the validation batch. The inflection points marking the start and end of the peak were more accurately predicted. However, it can be seen that although the peak height prediction improved, there is still an error between the predicted and the measured data. In comparison, the prediction from the AIC model can be seen to very closely predict the inflection points and achieves the peak height to within $\pm 10\%$. It can also be observed in Table 6.10 that for all 3 model assessment criteria the lowest value was achieved using the AIC weighted method. Therefore this method was identified as the best fit of the four tested approaches. However, there appears to be an issue occurring when one model has the lowest AIC value but then for the next sample point it is the other model which is lowest. This is illustrated in Figure 6.34 at sample point ≈ 225. If these curves were being used as the only method of predicting the yield then this would possibly be more of an issue. However as these models were constructed with the aim of being used to predict the resolution in a multi-component system and as a measure of column performance, these small inconsistencies are not as critical, as it is the width, height, and retention time that are critical factors.

In summary, both the averaging and AIC weighted methods showed improvements on the predictions of the area and gradient models, with the AIC weighting method showing the most significant improvements.

6.10 Other factors influencing column and model performance

As mentioned throughout this chapter there are factors which are considered external influences which may have had an effect on the process. These factors are ones which were not measured hence their actual impact cannot be quantified, although they may have contributed to some of the modelling inconsistencies observed.

Protein

The protein stock in pH 6, 7, and 8 solutions were made at the start of the working week. From these three stock solutions the 15 experimental batches were performed. Although lactoferrin is a stable protein, observing the centre point batches (batches 1, 6, and 15) it can be seen that the measured outputs are significantly different. Between batches 1 and 6 there is significant drop in both the on-line and off-line recorded yield. Due to unforeseen experimental issues an additional stock solution of pH 7 had to be made to run the final centre point batch (batch 15) and it can be seen that there is an increase of the yield. This suggests that the protein in some way degraded in the stock solution. Figure 6.35 shows the measured on-line absorbance for the three centre point batches. As can be seen the batch run in the middle of the week, batch 6, has the lowest absorbance, with a similar increase being shown in batch 15, possibly due to the need to remake the stock solution. This decrease in protein between batches 1 and 6 may indicate a loss of protein, as the solution was well mixed before taking a sample suggesting that the protein degraded. The only possible cause may be attributed to the exceptionally hot weather the week of experimentation (Abe *et al.*, 1991; Steijns and Van Hooijdonk, 2000), suggesting possible thermal degradation of the protein while the solution was out of the cool room during the experiment. Additionally the work of Sreedhara *et al.* (2010) suggests that the varying pH of both the stock solutions and the buffer used to elute the protein may have

effected the thermal stability of the protein, with a decrease in the pH causing a decrease in the denaturation temperature.

If thermal stability of the protein was the only issue affecting protein denaturation then it would be expected in Figure 6.35 that batches 1 and 15 would be very similar. However, it can be seen that there is less protein present in batch 15, inferred from the smaller elution peak. As the protein concentration of the stock solutions was checked after they were made this suggests there is another influence causing the smaller amount of protein in batch 15. One possible cause could be issues arising from the buffers.

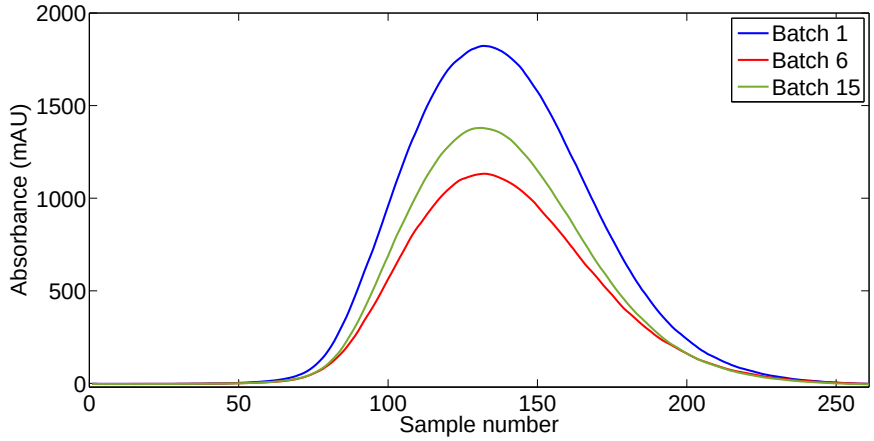


Figure 6.35: Measured on-line absorbance for centre point batches (batches 1, 6, and 15).

Buffers

The buffers used throughout the experiments were prepared at the start of the week and stored in a refrigerator. There is some suggestion that the buffers may have altered over the course of the experiments. As the centre points, batch 1 and 15, were both performed using newly made stock solutions where the protein was stored in the freezer prior to being used suggesting the obtained results would be very similar. However, it can be seen in Figure 6.35 that batch 15 had a significantly lower yield. This suggests that another factor, possibly the buffers, was impacting the process. The pH and conductivity of the buffers was checked prior to each batch, however other not measured issues can problems can arise, such as flocculation (Piazza and Garcia, 2010).

Resin

The performance of a column can decrease over time due to excessive use as shown by Hou *et al.* (2010). With excessive use it is assumed that the binding capacity of the resin decreases and hence less protein would absorb onto the matrix. As with the protein degradation it would be expected that issues with the resin would be observed in the centre point batches. Generally, excessive use of a resin refers to hundreds of batches over many months as opposed to the 15 batches run over a week for this data set. However there are reports in the literature which show that matrix degradation can occur over as few as 50 batches, as shown by Yang *et al.* (2015). Although these effects are unlikely to have influenced the data set used in this chapter they should be considered for application of the models produced for other data sets.

6.11 Recommendations

From the work presented in this chapter there are a few recommendations which can be made to further improve both the accuracy of the models but also the scope of application. This section summarises the main recommendations with possibilities for future development.

Dataset

A main factor impacting the accuracy of the models is the size of the data set used to both train and validate them. Generally applications of MVDA in literature use data sets which are both large and include more replicates. The variation seen in the three centre point batches suggests that similar effects would be observed in the other 12 batches had replicates been performed.

The small number of experiments used to construct the models in this chapter was deliberately chosen, as a key driver for industry is the reduction of both time-to-market and resources. Having said this, the experiments in this

research have shown that obtaining 'good' data for use in model development and design space characterisation is often very dependent on external factors. As a result of this data generated in this way could not be applied to produce an on-line monitoring tool. Hence it is the recommendation of this work that although DoE designs such as that used here can be applied, the need for replicates of experiments cannot be avoided.

Model scope

The models produced in this chapter focused primarily on off-line predictions of yield, which were accurate, and suggest the models could be similarly applied to predict other CPPs and CQAs. The second focus of this research was on-line predictions. This chapter considered only the prediction of the elution peak of a single component system and further work could be carried out to predict other aspects of the chromatogram such as the flow-through and wash, the strip, and the caustic stages.

Multi-component systems

Another limitation of the models developed in this chapter is that they only apply to single component systems. The techniques used for the predictions made from and for the off-line data could be readily applied to multi-component systems. However, for the models generated for on-line data significant changes would need to be accounted for. Figure 6.36 shows a simple drawing of two multi-component systems, the bottom one shows a system where good resolution is achieved with all three components (highlighted in red squares) distinctly separated. The problems arise when applying the models to systems such as that shown in the top chromatogram. The blue box highlights three components where the resolution is worse, and the peaks are not distinct.

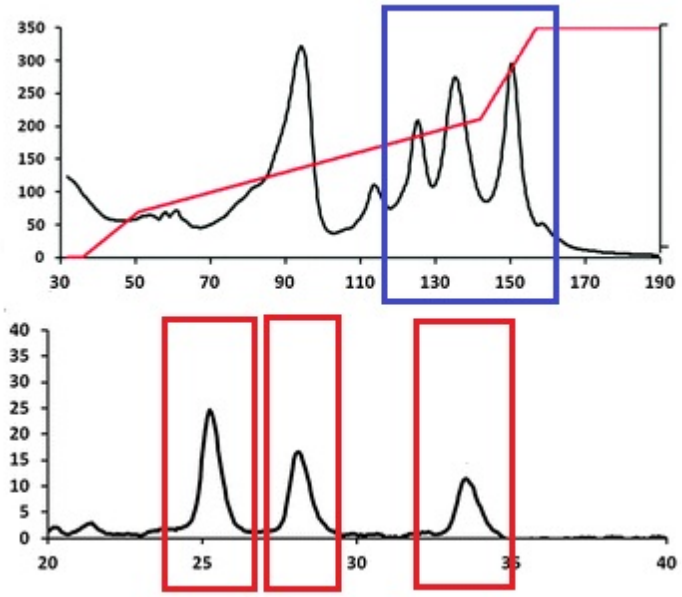


Figure 6.36: Example multi-component chromatograms, showing poor resolution (top) and good resolution (bottom).

The on-line models produced in this chapter rely on being able to calculate both the gradient of the slope and the area under the curve for the complete elution peak. Where the resolution is poorer and the peaks over lap, this is not possible. To be able to deal with this type of a system, further modelling techniques should be explored, such as first principles, which would give an indication as to the character of each individual component.

6.12 Conclusions

There were two main aims of this study; the first was to be able to predict CQAs from process operating conditions. The second was to see whether it is possible to predict the chromatogram of the system with the aim of being used for on-line control with the goal of using multivariate data analysis to achieve both of these aims and where necessary supplement with additional modelling techniques.

The work presented in this study addressed firstly the need to isolate the elution peaks when there is a lot of variation in the retention time. The methodology suggested within the work used a first principles model to predict the retention time, and then isolated the single component peak. The on-line data measured over the duration of the elution peak was then aligned between batches to the length of the longest batch, and the variables were then aligned within the batches. This allowed for a direct comparison of measurements at a given time point.

The results of the PCA analysis on the off-line data showed that the yield of lactoferrin was directly related to the load concentration, with the flow rate also being an important factor. The PCA analysis of the on-line data, which was used for variable removal, showed that when building a predictive model the variables which accounted for the variation in the data set were conductivity, pH, and the concentration of buffers over the gradient. The PCA analysis also showed that although no batches were considered as falling outside the 95% confidence limit, the analysis was influenced more by batches 12 and 13. These two batches were the highest yield batches, and were significantly higher than the rest.

Having performed the PCA analysis the results were then used to construct a PLS model to predict absorbance. An initial pre-processing study was carried out to determine the pre-processing which was most appropriate for the data set. To ensure that the only the batch to batch variation was captured and not noise, whilst also making sure the data was not over

pre-processed. It was determined that the necessary techniques were MSC mean, Savgol smoothing, and autoscale. The resulting PLS model showed a prediction which was Gaussian in shape, however the correct peak width, height, and baseline were not achieved. It was determined that the reason for this was that PLS can only operate when the shape of the data is similar in both the **X**-block and the **Y**-block.

Therefore two subsequent PLS models were constructed, the first which used the cumulative area under the curve and the second used the slope of the gradient. Both of these techniques improved the baseline prediction, with the area under the curve improving the peak height, and the gradient improving the prediction for peak width. This lead to two final models which looked at combining the outputs from the area under the curve and the gradient models. The first of these combined models used an average, and again gave improvements to the prediction. The second used an AIC weighted response, which gave the most accurate prediction of all the models.

Additionally PLS was used to predict the yield from the operating conditions. Two models were constructed, the first to predict the actual yield percentages attained and the second to predict the discrete yield attained (i.e. predicting if the batch was high, medium or low yield). Both of these models performed well, leading to the conclusion that a similar model could be applied to other final CPPs allowing them to be predicted in a similar manner. Having developed the tools for the prediction of both on-line and off-line chromatography measurements on the single component lactoferrin process, these techniques could be transferred and applied to a more complex system.

6.13 Summary

Over the last twenty years there have been significant developments to the modelling of chromatography, the most notable examples being Gu (1995) and Skov and Bro (2007). The examples presented in literature have predominately focused on modelling the chromatogram with less emphasis to prediction of off-line measurements. The research presented in this chapter has shown that the MVDA and first principles techniques described by Mercier *et al.* (2013), Bro and Smilde (2014), Shellie *et al.* (2008), and Madden *et al.* (2002), to name a few, can be successfully applied to be used for the prediction of off-line CPPs. Additionally, it has been shown that a data set used to collect the maximum information through the minimum number of experiments can successfully be applied to predict within the limits of the design. This works with the PAT guidelines given by the FDA which places a greater importance on the identification and control of CQAs and CPPs. Most importantly this chapter has succeeded in demonstrating a proof of concept study, showing that MVDA and first principles techniques can successfully be used to characterise various aspects of IEX chromatography. The conclusions highlighted the fact that these models could now be transferred to a mAb system, and can be used in conjunction with an agent based model, which is presented in Chapter 7.

6.14 Acknowledgements

The author would like to acknowledge Graham McCreath, John Liddell, and Emma McEwan at Fujifilm Diosynth Biotechnologies for their assistance in the generation of the data used in this chapter.

Chapter 7

Agent based model for Chinese Hamster Ovary (CHO) cell cultivation and purification

The work presented in the previous two chapters presented the initial modelling and characterisation of the bioreactor and Ion Exchange Chromatography (IEX) process units. This chapter covers the research performed to investigate the applicability of agent based modelling (ABM) to multiple bioprocess units. The work utilises the models constructed in the previous two chapters and applies them to a Chinese Hamster Ovary (CHO) cell line producing a monoclonal antibody (mAb).

The growth in the mAb market over the last decade is set to continue, and over the next 3-6 years the antibody market is projected to grow to $30bn or more (Ziegelbauer and Light, 2008). These and other therapeutic proteins are produced on a large scale in industry using various recombinant cell lines as the expression system. These cell lines can be bacterial (e.g. *E. coli*), yeasts, or mammalian cells (e.g. Chinese Hamster Ovary (CHO)), with CHO cells being the most popular expression system as they offer various benefits. The main one being that produced products bear a similarity to human produce proteins. This similarity it mainly in the post translational modifications, such as the glycosylation of the proteins, which is incorrect can produce an immune

response in the human patient.

Biopharmaceutical products must have a very high purity, and the concentration of host cell proteins (HCP) and DNA should be in the range of parts per million relative to the target product. Regulations also state that the final product should contain no micro-organisms, should contain less than 10 ng of DNA per dose and less than one virus per million doses (Low *et al.*, 2007). This means a very stringent purification process is needed for mAbs. Typically this purification process includs a three-column chromatography process consisting of a protein A affinity column (initial capture), followed by cation exchange (CIEX) and anion exchange (AIEX) columns as polishing steps and finally a virus filtration (VF) step. This process is summarised in Figure 7.1, which gives details on the substances removed in each stage.

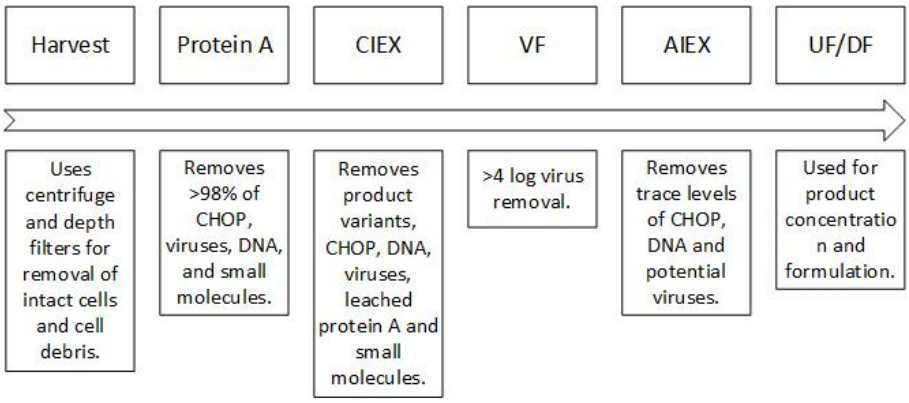


Figure 7.1: Purification process for mAbs. Showing a three column chromatography stage with ultrafiltration/diafiltration (UF/DF). The three column chromatography included an affinity step, a cation exchange step, and an anion exchange step. *CHOP refers to Chinese Hamster Ovary Proteins (Host cell proteins produced by CHO cells) (Mehta *et al.*, 2008)

The three stage chromatography purification process is used because it is able to meet the stringent purification requirements. However this process is very expensive, especially the protein A affinity step which accounts for almost 35% of the total raw materials costs for downstream purification (Kelley, 2009). The growing demand for mAbs and the increasing market competition has lead to a focus on reducing manufacturing costs and improving the efficiency of processes on the industrial scale.

There are various methods which can be employed to reduce the total cost

of production. Low *et al.* (2007) discuss the benefits of changing the current standard process units to include other units such as high-flow high capacity chromatography resins, non-chromatographic processes such as membrane absorbers, precipitation, flocculation or crystallization. Suggestions have been made for an alternative capture step as protein A does have several major drawbacks. Apart from the high resin cost, protein A also has a limited binding capacity and issues with clean in place (CIP) due to the sodium hydroxide damaging the ligands. Various authors have suggested the use of cation exchange chromatography as an alternative capture step in mAb purification (Arunakumari *et al.*, 2007; Stein and Kiesewetter, 2007; Urmann *et al.*, 2010). The consensus is that although CIEX is not as efficient as protein A in terms of selectivity of HCP it can be used with a HCP precipitation step to get similar HCP removal levels as protein A. In cases such as this the deciding factor is whether cost of process complexity is the main issue.

This chapter presents agent based modelling (ABM) as a tool for reduction in cost of production. Providing a flexible framework for the use of multiple bioprocess models each characterising a separate process unit. This method is able to handle process interactions, and can provide a plant wide process characterisation. As Gao *et al.* (2009) state, an agent based framework is constructed using process knowledge, process models, and a group of functional agents. Initially the process knowledge and process models are applied for each process unit individually. Then the functional agents act as the connection between the units. This form of modelling is a move away from the more traditional form of bioprocess modelling which considers each process unit separately. The literature has shown that ABMs are commonly used in applications such as social sciences where it is the interactions between various variables that are of interest. They have more recently started to be applied to bio applications such as Segovia-Juarez *et al.* (2004) who applied an ABM to characterise tuberculosis in the lung. This work shows the scope that ABMs have for applications to biosystems which have a higher level of unpredictability than say chemical systems. Whilst Gao *et al.* (2009) developed an ABM for the production and purification of alcohol de-hydrogenase (ADH) from a *Saccharomyces cerevisiae*, this work shows the

Table 7.1: Summary of available data for mAb process. 'Measurement' refers to the technique used and the comments give details on what the technique is measuring. The experiment section shows whether the measurement was recorded for that particular experiment.

		Experiment			
Measurement	**Comments**	**Lab scale 1**	**Lab scale 2**	**Demo**	**MCC**
UV A280	Titre	X	X	X	X
Protein A	Titre	X	X	X	X
CIEX	Measure of heterogeneity	X	X	–	–
SEC	Measure of molecular weight	X	X	X	X
HCP	Host cell protein	X	X	X	X
CE-SDS	Measure of whether protein is glycosylated	X	X	X	X

applicability of ABMs to optimisation of mammalian cell based production and the subsequent purification.

7.1 Process data

Data set one

Two data sets were used for the construction of the models in this chapter. The first data set was generated by the sponsor company Fujifilm Diosynth Biotechnologies for the characterisation of a new process. Fujifilm have two sites for contract manufacturing, one based in the UK at Billingham, the second in the United States of America in North Carolina. The primary aim of the experiments was to determine the repeatability of the culture and purification process at the two sites, to check that the end product had the same purity, glycosylation profile, level of aggregation, and molecular weight. This data set included off-line measurements which were obtained using different analysis techniques Table 7.1 summaries the different measurements that were recorded.

As can be seen form Table 7.1 the data set is relatively small, containing only 4 experimental runs. This is standard procedure in that the main development for each process unit is carried out individually, so when the entire process is checked, less experimentation is required. This method is beneficial for the company in that it reduces costs, however it is not ideal for constructing a model as more experiments and replicates would produce a more robust model. However, the data set is sufficient to be able to use as proof of concept in this chapter.

The analytical measurements were taken after each process unit (as summarised in Figure 7.1). However for the virus inactivation and virus filtrate stages there is a high proportion of missing data. This is not an issue for this research as the main area of focus is the cultivation and ion exchange stages.

Data set two

The second data set used in this research was again supplied by the sponsor company Fujifilm Diosynth Biotechnologies. This data set was produced as part of their work to establish their mammalian cell systems platform process. The data was from the upstream cultivation of CHO cells and the production of a mAb product. Due to confidentiality the exact mAb product has not been disclosed however it has a similar structure, glycosylation profile, molecular weight, and downstream purification process to the mAb used in the first data set. Therefore for the purpose of this research they will be assumed to be sufficiently close to demonstrate the proposed concepts of ABM. However any subsequent development work on this tool should be carried out using the upstream and downstream data from a single production system.

The data set included fed batch shake flask, lab scale, and 2L cultivations. The work was originally used as a cell line and scale up study, and thus included data on 10 different cell lines. Furthermore the data included both off-line measurements for viable cell count, titre, glutamine, glutamate, glucose, lactate, ammonia, and the glycosylation profile for the end of cultivation product. Additionally there were on-line measurements for the 2L

bioreactors, providing measurements for dissolved oxygen (DO), CO_2, air flow, pH, stirrer speed, and temperature. Finally provided for the fed batch shake flasks and lab scale reactors were the set points for pH, temperature, agitation, and DO (Table 4.2 on page 66).

7.2 Methodology

Unit models

The individual process units were characterised using the models determined in Chapters 5 and 6. These chapters used a hybridoma data set and a lactoferrin data set to determine the modelling methodology which provided the most accurate predictions. The identified best models were then used in this chapter and applied to the CHO cell data. Table 7.2 summarises the models, providing the target CPP or CQA and the model used to predict it.

Table 7.2: Summary of the best models from Chapters 5 and 6. The table lists the target CPP or CQA that the model is predicting, the type of model used, and the data required as an input to the model.

Process unit	CPP/CQA	Model type	Data input / X-block	Comments
Bioreactor	Viable cell count	Hybrid Kontoravdi *et al.* (2007)	Operational set points	Uses PLS to determine rates of consumption/production of metabolites. First principles models then used to provide concentrations of metabolites.
	Titre	PLS	Operational set points	The hybrid model used for viable cell count also contained a prediction for titre. However PLS only model was less complex a produced a better prediction.
	Glycosylation profile	PLS	On-line measurements	No hybrid or first principles model available. The PLS only model produced accurate results.
Ion exchange chromatography	Retention time	First principles	Data needed to calculate constants	A potential issue is that data is required to predict the constants for the model. This data is not available for the mAb purification data set.
	Yield	PLS	Operational set points	The PLS only model produced an accurate prediction for yield, and the operational set points were available for the mAb data set.
	Elution profile	Hybrid	AIC weighted (Area under curve and gradient of slope)	Similarly to the retention time model, data was required for the mAb data sets to be able to train this model. Therefore it can not be applied to this data set, but could be used in further research where more data was available.

Not all of the models listed in Table 7.2 can be used in this chapter as the CHO data set does not contain the required data to train the models. For example the downstream IEX data only includes the analytical measurements and the operational set points of the platform process used within the sponsor company. Therefore the CPPs of retention time and elution profile could not be determined as these require on-line data sets. However, the model used to predict the CPP yield could be used as the model only required the operational parameters.

Agent framework

The agent framework is used to support the integration of the unit operation models and to simulate the interactions between them. Figure 7.2 shows the hierarchical nature of the ABM framework.

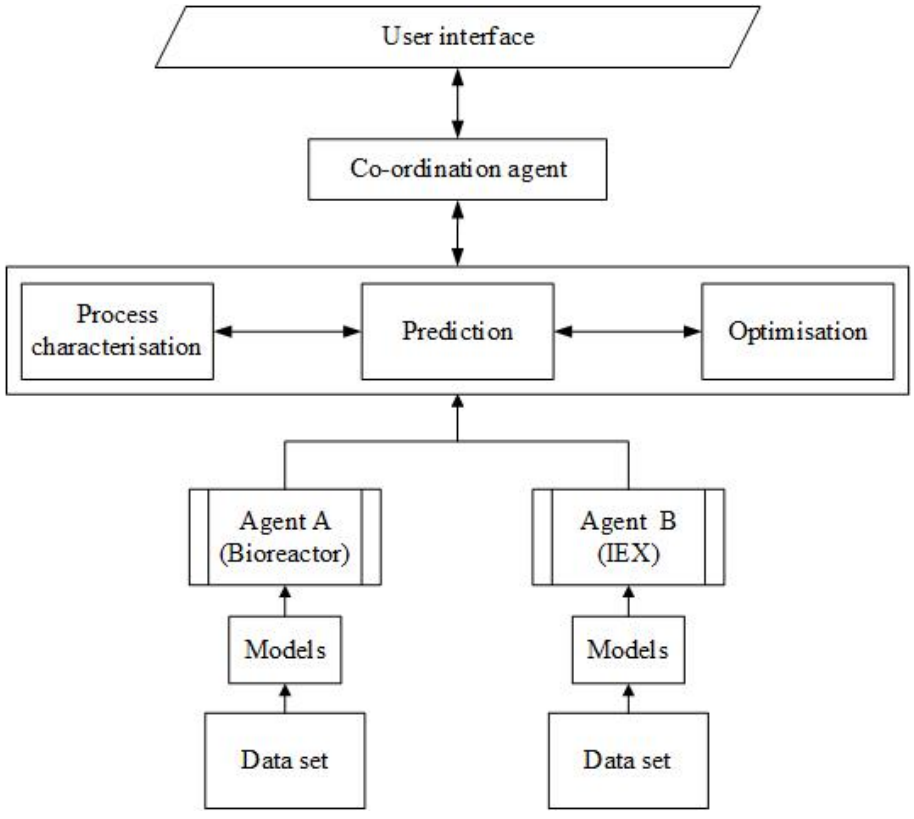


Figure 7.2: Agent based model (ABM) framework, showing the hierarchical nature of the model. The model can be adapted and improved by the use of different or additional data to train the process unit models.

This chapter is concerned with the higher level agent which co-ordinates

the process unit models. This higher level agent is able to integrate the unit operations and communicate between them to simulate the interactions thereby giving the desired whole process description. This ability to communicate between process units allows for the investigation of how changing process parameters affect plant performance, and thus determine the optimum operating conditions. The 'models' (see Figure 7.2) called by the unit agent, can contain multiple models such as first principles or multivariate (such as those in Table 7.2). This allows for different CPPs or CQAs to be modelled. Furthermore there is no limit to the number of process units that can be investigated so long as the data and models are available.

The co-ordination agent acts as a link between the unit operation agents. This allows the unit agents to operate as one when evaluating the process conditions to improve the whole process performance. Communication between the unit agents is through a common parameter, which is modelled, predicted, or optimised. This common parameter can be any measurement which is recorded for each unit, for example in data set one the level of host cell protein could be used. The unit agent then builds various models using the common parameter as a reference. The ABM presented in this chapter was constructed using matlab software.

7.3 Results and discussion

The models and results presented in this section act as a proof of concept for applying an agent based model to a mammalian cell system. The models presented in this section show how to apply this technique to achieve improvements in process performance. Additionally suggestions are made for further work to further develop the overall model.

Unit CPP and CQA prediction

Initially to develop the agent model the first data set, which contained measurements of key parameters after each process unit, was used. This data

set included measurements on titre (measured using both UV and protein A), remaining HCP (host cell protein), heterogeneity (CIEX (cation exchange chromatography): reported as main peak, acidic variants and basic variants), molecular weight (SEC (size exclusion chromatography): reported as main peak, high molecular weight, and low molecular weight), and whether the protein is glycosylated or not (CE-SDS (capillary electrophoresis-Sodium Dodecyl Sulfate gel): reported as main peak or none glycosylated heavy chains). The data for most of these measurements was incomplete, with values for the harvest only being recorded for titre. The value of the harvest is the point which can be related to the end of cultivation data (linking to the models in Chapter 5). Therefore this proof of concept uses the titre as the main predictor matrix.

The first models constructed focused on being able to predict the CPPs/CQAs listed in Table 7.1 from the titre values. Six PLS models were constructed all using the product titre (UV) as the **X**-block data, the other CPPs/CQAs listed in Table 7.1 were the **Y**-block for each successive model. The titre was used as the **X**-block data because it was the only CPP in the data set which contained values for the end of the cultivation. This was important because the unit agents for the cultivation and ion exchange chromatography could be used to determine how changes to one unit would impact the operating parameters of the other unit. As the overall agent model is designed so that it can be used by anyone regardless of modelling knowledge, the PLS models would ultimately use automated selection for the number of latent variables. However for demonstration in this proof of concept all models contain 3 LVs. This value was checked for all six model shown here to ensure that significant amounts of data were not being missed or noise included. Table 7.3 details the Y-block used in the models and the RMSE value for predicted measurement after each process unit.

The predictions for HCP at each stage of the DSP are shown in Figure 7.3, as can be seen the predictions made for virus removal and the subsequent units is relatively accurate when compared with the prediction for the protein A eluate. This is reflected in the RMSE values (see Table 7.3) which are

significantly higher for the protein A eluate. This variation is most likely a reflection of the small data set size. Three training batches are not sufficient to construct an accurate model, especially as the variation between these three batches is significant. For the HCP values recorded after the protein A the range is 1364 - 5234. Additionally, as Figure 7.1 shows, the protein A and IEX steps are used to remove DNA. As the levels of HCP vary greatly prior to these stages it confirms that HCP levels cannot be controlled.

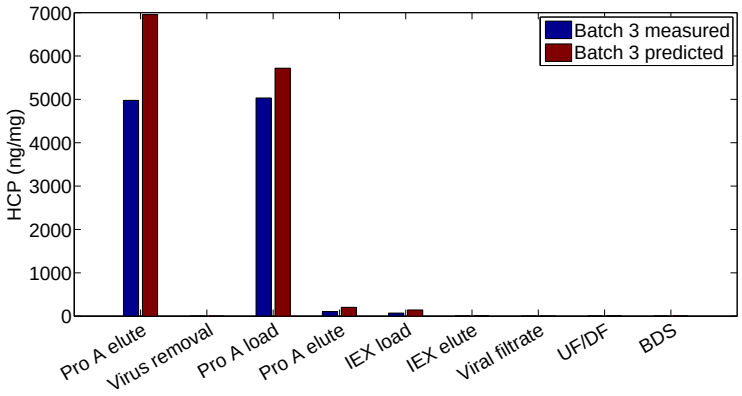


Figure 7.3: Measured and predicted values for HCP (host cell protein) for batch 3. Prediction based on the product titre values after each process unit. Model contained 3 training batches, 1 validation batch, and 3 latent variables.

Table 7.3: Table showing the data used to train the model (Y-block) and the RMSE value for the predicted measurement of each process unit. *- demotes that there was not measured data so the model could not predict the unit value.

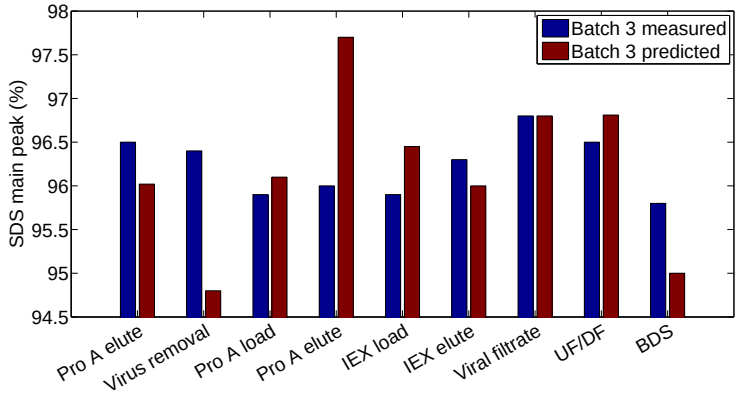
		RMSE									
Model	**Y-block CPP/CQA**	*Harvest*	*Pro A elute*	*Virus removal*	*Pro A load*	*Pro A elute*	*IEX load*	*IEX elute*	*Viral filtrate*	*UF/DF*	*BDS*
1	*HCP*	-	1977.20	0.00	685.83	97.79	72.87	3.99	0.00	3.84	3.57
2	*SDS main peak*	-	0.43	1.55	0.22	1.70	0.50	0.24	0.01	0.30	0.78
3	*SDS NGHC peak*	-	0.28	0.33	0.23	0.33	0.21	0.13	0.13	0.31	0.21
4	*SEC main peak*	-	1.78	0.00	0.59	0.38	0.42	0.40	0.00	0.29	0.69
5	*SEC HMW peak*	-	1.57	0.00	6.86	0.23	0.25	0.16	0.00	0.38	0.49
6	*SEC LMW peak*	-	0.23	0.00	0.06	0.06	0.06	0.16	0.00	0.03	0.02

As HCP is an uncontrolled variable it means that there is a lot of variation in the recorded levels, this makes it difficult to predict. Therefore only the end process units (i.e. virus removal onwards) can be said to be predicted with any accuracy, as the level of removal is known (i.e. 98% removed by protein A).

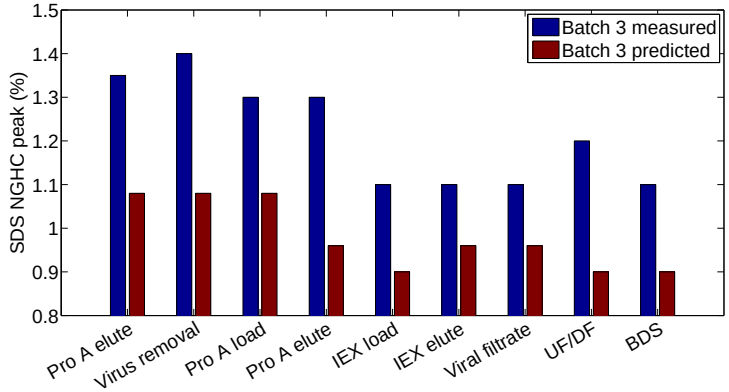
Figure 7.4 shows the predictions of product glycosylation. As can be seen both visually using Figure 7.4 and from the RMSE values in Table 7.3 the predictions initially appear to be accurate. However, considering the relatively narrow range of the measurements recorded, the predictions do not appear to be as good. For example in Figure 7.4(a) the prediction is worse for the virus removal and second protein A eluate. This suggests that the most variation between batches occurred for these units. Taking for example the prediction made for the virus removal Table 7.4 shows the values used to train the model (batches 1, 2, and 4) and the values for the validation batch (batch 3). Comparing batches 3 and 4, it can be seen that the glycosylated main peak values are approximately the same, however the titre value for batch 3 is much lower than the rest. Therefore when the model was trained with batches 1, 2, and 4 and subsequently the titre for batch 3 was used to predict the glycosylated main peak the value predicted was much lower. This shows two things; the first that a larger data set would be better. Secondly, it suggests that there is not a direct relationship between the recorded titre and the glycosylated product, and thus titre cannot be used to predict it. The agent model would also be able to predict the glycosylation profile after each unit, as shown for the fermenter (Chapter 5) and this prediction could be used as an additional input to the DSP models. Due to lack of data this could not be implemented in this research, however the models in Chapter 5 show that it is possible to accurately predict.

Table 7.4: Comparison for the viral unit of the X-block data, titre, used to train the model, and the Y-block data, SDS main peak, predicted by the model.

	Virus removal unit	
	Titre (mg/ml)	Glycosylated form SDS main peak (%)
Batch 1	12.10	97.70
Batch 2	14.40	97.80
Batch 3	11.70	96.40
Batch 4	12.20	96.20



(a) Main peak



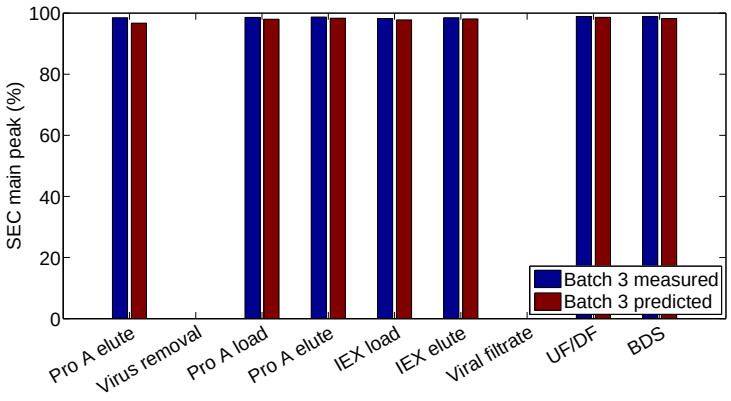
(b) NGHC peak

Figure 7.4: Measured and predicted values for glycosylation ((a) main peak (b) NGHC (non-glycosylated heavy chains)) for batch 3. Prediction from the product titre values after each process unit. Model contained 3 training batches, 1 validation batch, and 3 latent variables.

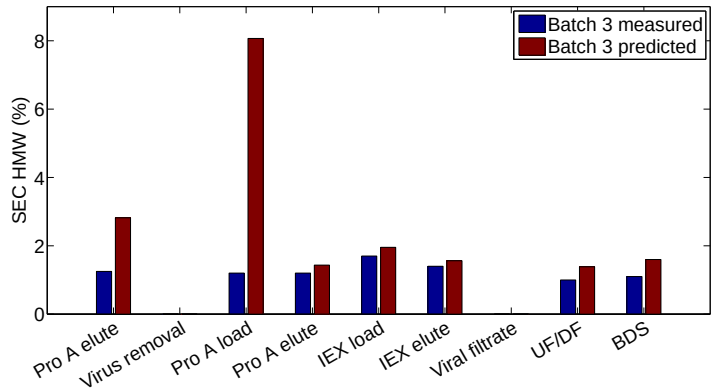
Figure 7.5 shows the predictions of the size of the product, with (a) showing the main peak, (b) showing the high molecular weight (HMW) peak, and (c) showing the low molecular weight (LMW) peak. For the main peak the predictions are good, however there is little variation between the measurements for each batch, therefore no firm conclusion can be drawn as to the appropriateness of using titre as a predictor. In contrast for the HMW and LMW peaks there was variation in the measurements used to both train and validate the model. The variation in the HMW and LMW peaks is not directly linked to the titre measurements, and thus the titre can not be used to accurately predict the SDS main, HMW and LMW peaks. Similarly to the glycosylation profile, the molecular weight could be predicted after the cultivation and used as an input to the DSP models.

The final measurement recorded in data set 1 was the measure of heterogeneity using CIEX. These values were only recorded for two of the four batches, and as such no model could be constructed as there would only be one training and one validation batch. Figure 7.6 shows the measured data for the main, acidic, and basic peaks which further demonstrates why a model could not be constructed, as there is significant variation between the two batches.

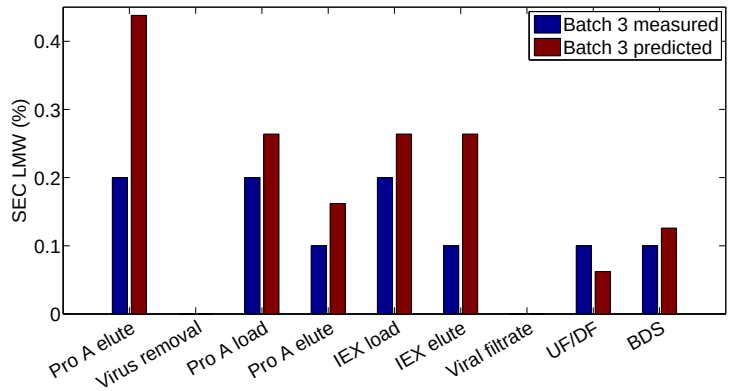
Having considered using titre as a way of predicting the other CPPs/CQAs it can be seen that this is not a viable method for any of the variables. This presents a challenge, as for this data set there are no other measurements which can be used.



(a) Main peak

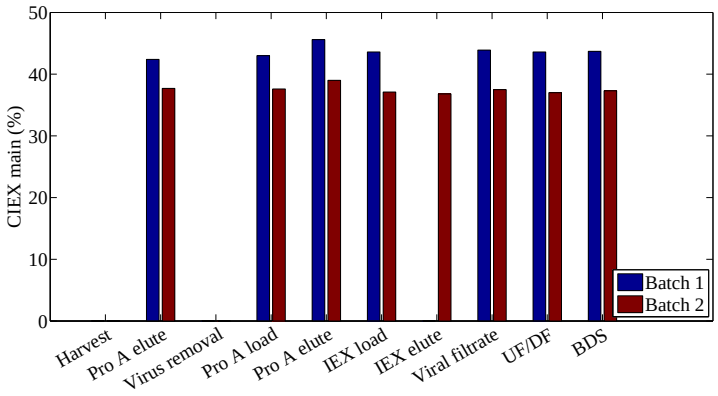


(b) HMW peak

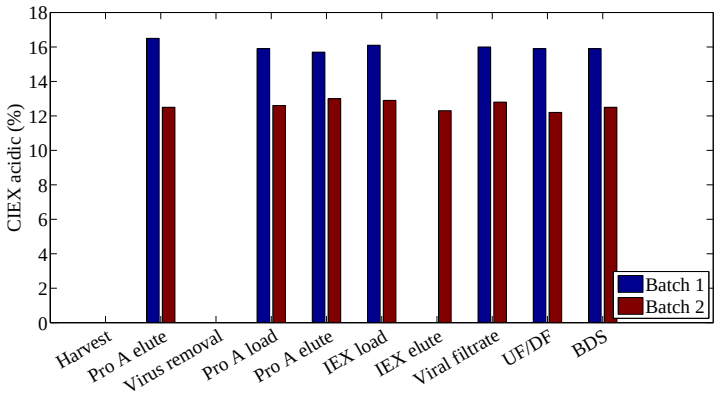


(c) LMW peak

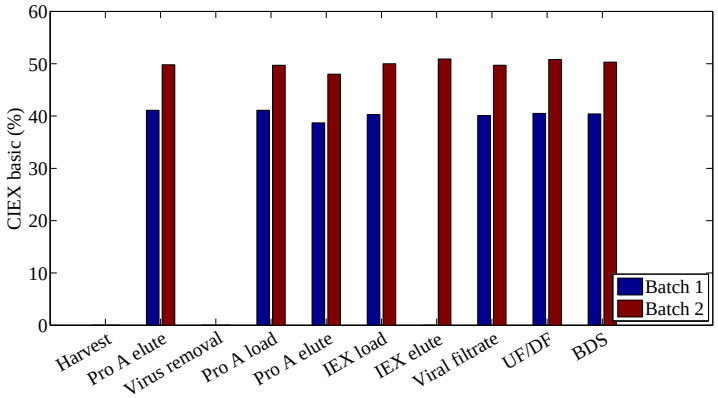
Figure 7.5: Measured and predicted values for protein size ((a) main peak (b) HMW (High molecular weight) (c) LMW (low molecular weight)) for batch 3. Predicted using the product titre values after each process unit. Model contained 3 training batches, 1 validation batch, and 3 latent variables.



(a) Main peak



(b) HMW peak



(c) LMW peak

Figure 7.6: Measured heterogeneity values for batches 1 and 2 ((a) main peak (b) acidic variants (c) basic variants).

Titre as agent model parameter

As previously mentioned to co-ordinate between the unit agents a common parameter is needed. From the data sets used in this research the only measurement which has been recorded for both the cultivation and the DSP steps is the product titre. Considering the models constructed in Chapters 5 and 6 which showed that the operating conditions can be directly related to titre, the titre was used in the following models as the communicating parameter. Obviously other process parameters relevant to the task would be used in a real application, titre is used here to demonstrate the concept of ABM.

Data set one was used to construct a model where in the **Y**-block data is the titre recorded after the 9 process units of harvest till UF/DF. The **X**-block data is the titre obtained at the end of the whole process i.e. the BDS (bulk drug substance). The model structured in this way should allow for the operator to specify the desired attributes of the end product and from this be able to predict the operating conditions for the various process units. Additionally the model could be applied by adjusting the set points of the cultivation to determine the impact downstream.

As both data set one and data set two are for mAb products and the downstream processing units were operated according to in house operating procedures at the sponsor company, it is assumed that the relationship characterised in the downstream processing of data set one also apply to data set two. Therefore the model is trained using all four batches within data set one, and tested using data set two. Two consequences of this are that there are more batches included in the training data set, and there is information provided on the operating parameters for the batches in data set one. This means that from variations to the titre implemented by the co-ordination agent, the operating parameter set points can be predicted and these can be compared to actual values. Figure 7.7 shows the measured data from data set one, and the prediction made using batch 3 of data set two.

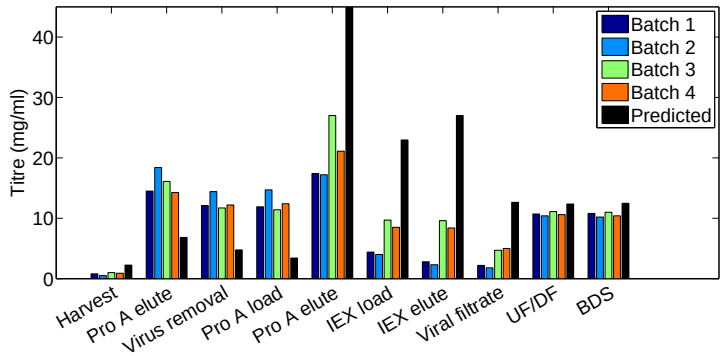


Figure 7.7: Measured data for batches 1 to 4 from data set one. These four batches were used to train the model which predicts the titre for each process unit from the titre of the BDS (bulk drug substance). The measured data (batches 1-4) were as the X-block to train a model for predicting unit titre. The black columns show the prediction made for batch 3 from data set 2.

As can be seen from the predicted batch, a higher titre achieved at the BDS stage is reflected in the a higher titre obtained at the end of the cultivation. It is shown across all the training and validation batches that a higher titre at the end of the cultivation means a lower titre during the first and second protein A stages with the protein becoming more concentrated in the elution of the second protein A column with all subsequent stages having a higher titre. It can not be confirmed if the predicted titre for the downstream processing units from the first protein A eluate are correct. However the predicted value for the harvest, was 2.3 mg/ml and the measured titre was 2.25 mg/ml, showing that for this unit at least the prediction is good.

Having predicted the end of cultivation titre this allowed for the culture conditions to be predicted. A new model was constructed using the titre as the **X**-block with the **Y**-block consisting of the initial conditions and set points for glucose, lactate, ammonia, glutamine, glutamate, dissolved oxygen, agitation, temperature, and pH. The model was trained using data set two, with batch 3 being used as the validation batch. The inclusion of dissolved oxygen, agitation, temperature, and pH was redundant for this data set as for all batches the set points were the same. However as they did not negatively impact the model they were retained, so that the model is established in its final form for use within the sponsor company at a later date. Having established the initial

set points for the cultivation a model was constructed using data set two. The model used was hybrid model D from Chapter 5 as this model only required the initial concentrations and rates of the main metabolites. Therefore, the glucose, lactate, ammonia, glutamine, and glutamate concentration profiles were predicted throughout the cultivation duration.

Figure 7.8 shows the prediction for glucose for batch 3. As can be seen the general profile for the concentration is shown over the entire duration. However the model does not seem to have predicted the high extremes as accurately. In particular the recording at day 11, which is representative of when the bioreactor was batch fed. This suggests that the model may provide better predictions for instances where the cultivation is operated without subsequent human interferences such as feeding. However to be able to draw firm conclusions as to the ability of the model to predict instances such as batch fed cultivations then more data is required. Additionally a parameter should be included in the model which allows the operator to specify the day the feed will occur on, to see how this affects the cultivation. This would require significant amounts of experimentation to fully characterise.

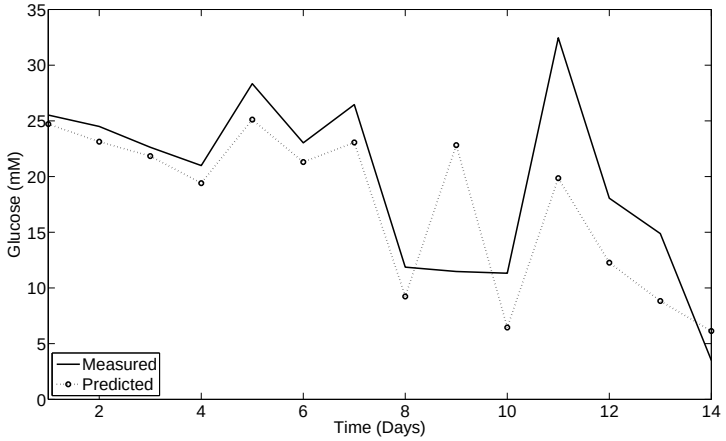


Figure 7.8: Measured and predicted values for glucose concentration for batch 3 of data set two. Model was constructed using end titre to predict initial conditions. Which were then used with the operating parameter set points in a hybrid model which predicts the rates for the equations presented by Naderi *et al.* (2011).

Figure 7.9 shows the measured and predicted lactate concentration for batch 3. The general trend is well represented however there appears to be an

offset, with the predicted concentrations being higher for each day. There are a few possible reasons as to why this might have occurred. The first is that the **X**-block data used to predict batch 3 may have been very similar values to one of the training batches and thus the predicted result is similar to that training batch. This would suggest then that there is variation between the batches which is not captured in the variables used in the models, such as the variation between cell line. The variation in cell lines is difficult to capture as it is not a quantifiable measurement.

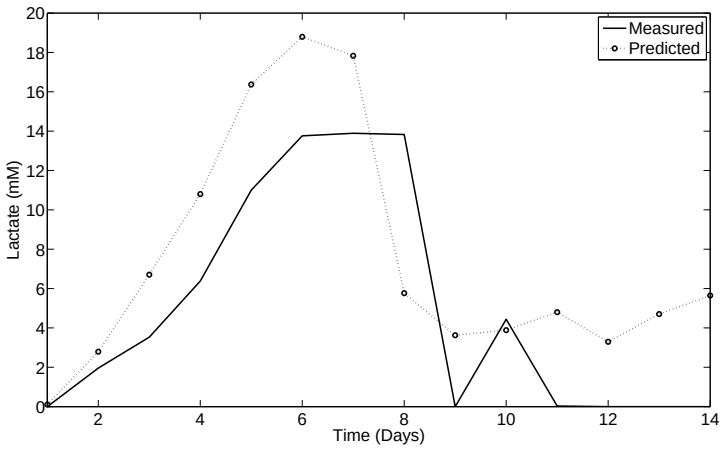


Figure 7.9: Measured and predicted values for lactate concentration for batch 3 of data set two. Model was constructed using end titre to predict initial conditions. Which were then used with the operating parameter set points in a hybrid model which predicts the rates for the equations presented by Naderi *et al.* (2011).

Figure 7.10 shows the measured and predicted ammonia concentration for batch 3. The model predicts the concentration accurately for the first 6 days, and fairly accurately after. Between days 6 and 12 it could be said that the model is not as accurate, however when it is considered that the only actual measurement that was provided was the BDS titre, and this is the prediction made after 3 models have been applied it can be seen that relatively speaking the prediction is good.

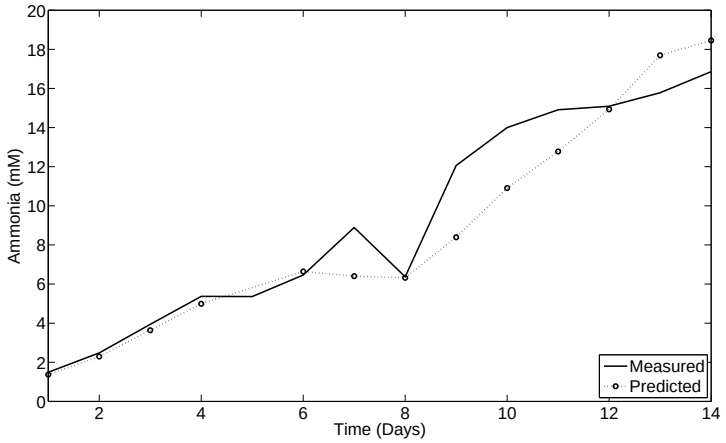


Figure 7.10: Measured and predicted values for ammonia concentration for batch 3 of data set two. Model was constructed using end titre to predict initial conditions. Which were then used with the operating parameter set points in a hybrid model which predicts the rates for the equations presented by Naderi *et al.* (2011).

Figures 7.11 and 7.12 show the glutamine and glutamate concentration profiles respectively. The predictions made for both of these are accurate throughout the cultivation. This suggests that they are independent of the fed batch process, as the glucose which is dependant on it showed issues with predictions the concentration after the culture has been fed. Considering the growth curve of the cells the glutamine can be seen to be consumed during the cell production stage, and produced during the cell death/product production stage. In contrast the glutamate can be seen to be produced during cell production and consumed during the cell death/product production stage. This is in agreement with the study conducted by Altamirano *et al.* (2001) who investigated CHO cell metabolism. They also showed that glutamate could be used as a limiting factor similar to glucose.

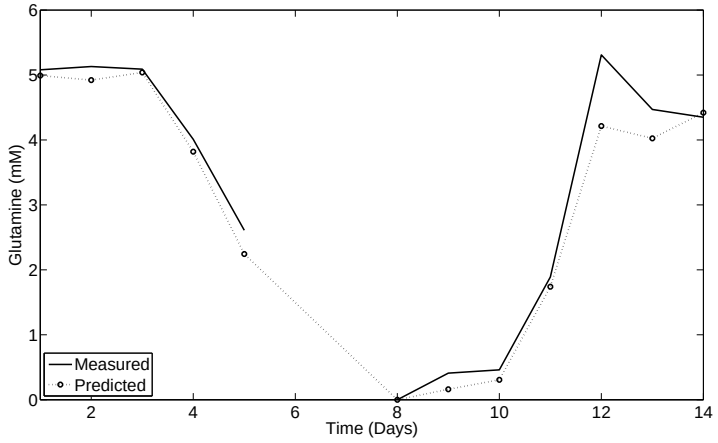


Figure 7.11: Measured and predicted values for glutamine concentration for batch 3 of data set two. Model was constructed using end titre to predict initial conditions. Which were then used with the operating parameter set points in a hybrid model which predicts the rates for the equations presented by Naderi *et al.* (2011).

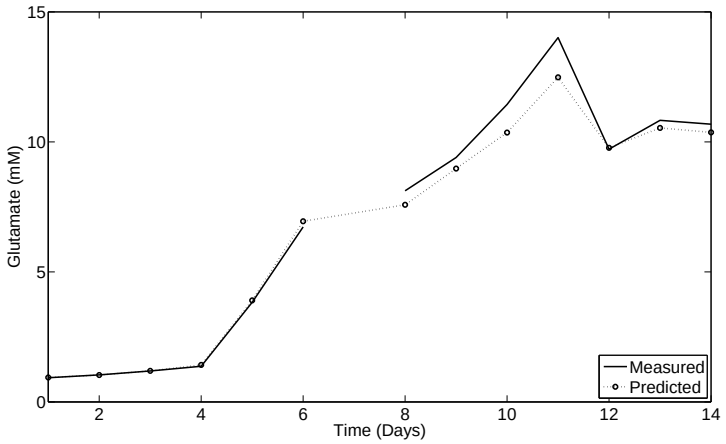


Figure 7.12: Measured and predicted values for glutamate concentration for batch 3 of data set two. Model was constructed using end titre to predict initial conditions. Which were then used with the operating parameter set points in a hybrid model which predicts the rates for the equations presented by Naderi *et al.* (2011).

The final CPP predicted for the cultivation was the viable cell count (Figure 7.13). As can be seen the lag, and initial growth phases are predicted well, as is the final viable cell count. The prediction for the maxima is not as good. However as this model is used as a guide, it can provide a lot of information such as the growth rate, the time point of the cell count maxima, and the final cultivation condition.

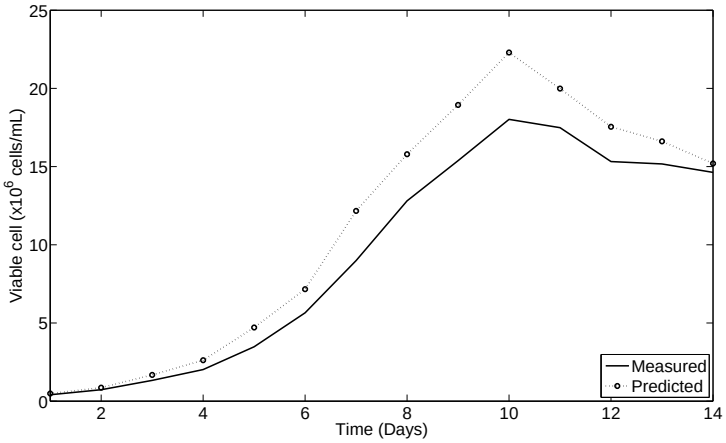


Figure 7.13: Measured and predicted values for viable cell count for batch 3 of data set two. Model was constructed using end titre to predict initial conditions. Which were then used with the operating parameter set points in a hybrid model which predicts the rates for the equations presented by Naderi *et al.* (2011).

In summary it has been shown that using the titre as the communicator between the process units is possible, and does yield good results. Subsequently it has also been shown that if the final product titre is known for the cultivation, it is possible to predict the initial conditions and operating set points with a relatively good degree of accuracy. The limitations of the data set mean that it is not possible to test the ability of using the end unit titre to test other process units. Initially the aim was to predict the operating conditions and subsequent elution of the IEX unit, but the information is not available. The following illustrates the method that would be used when a complete data set becomes available.

Using the titre predicted as the end value for the IEX unit as the **X**-block of the model, the operating parameters would be predicted. This is similar to the development shown in Chapter 6. To illustrate the point, a model was shown here which used the lactoferrin yield and operational parameter information to train the model, and the CHO cell data titre to predict the operating conditions. This model (as presented in Chapter 6 and trained using the lactoferrin data) is not a true representation of the operating conditions for the CHO cell, as the lactoferrin is a completely different protein and thus would be operated differently. The aim was just to show the application of the modelling technique. The resulting predictions for the operating conditions are

Table 7.5: Predictions for the operation of the IEX process unit. Predictions made using a PLS model which was trained with the lactoferrin data set from Chapter 6. Model and predictions aim to illustrate how they should be applied for this unit when the appropriate is obtained.

	Operating parameter				
	Flow rate (ml/min)	**Load pH**	**Gradient (CV)**	**Load concentration (mg/ml)**	**Elution pH**
Batch 3	0.62	6.98	17.02	44.84	6.96

given in Table 7.5. The model predicts the optimum load and elution pH values, the optimum flow rate, the column volumes the gradient should be performed over, and the concentration of the load sample which would achieve an end titre as specified.

The models shown here for the cultivation and IEX are relatively simple. However they allow for simultaneous optimisation of process units to achieve one desired end goal. The next step in the process would be to predict and optimise for multiple end CPPs and CQAs. It would be possible to predict these by incorporating further models into the unit agents. The main limitations are lack of available data. For many CPPs and CQAs, such as glycosylation profile, it is not common practise to measure them after every process unit. This limits the applicability of the agent based model. Another limitation is the ability to model, it has been demonstrated in Chapter 5 that it is possible to predict the glycosylation profile at the end of the cultivation. However, as no data is currently available it is only speculation that other measurements, such as aggregates, could be predicted.

Perhaps one of the main benefits of adopting an ABM is the ability to predict how changes to the operation of one unit affect other units. An ABM could be used to determine how changes to the operating parameters of the cultivation would impact on the down stream processing. This then allows for predictions to the operating conditions of the DSP units to maintain the same product quality attributes. For example if a change in the temperature or pH of the cultivation causes a change in the aggregation of the product (Hwang *et al.*, 2011; Bollin *et al.*, 2011) this can be predicted and the subsequent DSP unit

operating conditions altered to account for this. This results in a BDS which is uniform across all batches.

7.4 Suggestions for further development

The work presented in this chapter was carried out to provide a framework for process development. The main suggestion for further development would be to test the model using more data. In which the upstream and downstream units have been characterised, and preferably with more batches and repeats. This would allow for testing of the IEX models developed in Chapter 6.

One of the main issues encountered within the research presented in this chapter was that the main process unit characterisation was for the bioreactor, however the data supplied for process units (data set one) was predominantly for downstream and measurements of final cultivation output were not included. The only CPP which was measured was titre. In further development it would be beneficial to see if any other measured factor could be used as the communicator between process units, i.e. the glycosylation profile.

7.5 Conclusions

The main aim of the study presented in this chapter was to determine if it was possible to predict the operating conditions of process units having specified the end product. A second aim was to determine whether the models presented in previous chapters could be used to characterise the profile of the operation of the individual units. The work in this chapter mainly focused on the bioreactor cultivation due to insufficient data for downstream units.

The work presented firstly addressed whether a CPP such as titre, measured after each process unit, could be used to predict other CPPs and CQAs. The results showed that titre was not an accurate predictor of other measurements. This is likely because titre is a measure of quantity whereas the other CPPs and CQAs are measured of the quality of the product. The amount of the product does not reflect the form the individual proteins take. This was challenging for this data as it meant to be able to predict the unit operation of the only unit which was fully characterised and data available for (the bioreactor) the product titre had to be used as the communicator between the BDS and the unit. This however does leave the potential for development at a further stage, in that other measurements such as host cell protein or glycosylation profile could be used as the communicator to determine how the unit should be operated to achieve specific end results for these measurements. Additionally it is worth noting that a better prediction could of been made through the use of total titre as this would remove issues due to dilution.

The results for the agent model constructed using titre as the communicator were much more encouraging. The model showed that when a specific BDS titre was chosen the initial operating conditions, and the subsequent metabolite profiles could be predicted relatively well. Slight issues arose around the prediction of the fed batch element, in that the training data used for the metabolite profiles were fed batch a slightly different times, and combined with the small data set meant errors were introduced. This is particularly evident in the glucose profile, however this could be easily accounted for if more batches were used to train and the time of the fed batch

was used as a model input.

The validity of the IEX models could not be tested as there was no available data for the unit. However the application of the operational parameter prediction shown in this chapter demonstrates the simplicity of the model. This model should be tested as soon as possible and if it is determined that the results are valid then it shows how simple it would be to make changes to the downstream operation. Additionally the simplicity of this model for the IEX suggests that similar models could be generated for the other chromatography columns making it easy to provide a quick simple optimisation of those units as well.

In conclusion the work presented in this chapter has proved the concept that agent based model can be used for entire bioprocess optimisation of CHO cells. So far the concept has only been applied using the product titre as the communicating measurement. However there is evidence to suggest other parameters could be used instead. This is a critical point as it is this communicator which is the optimised element.

7.6 Summary

In summary the work presented in this chapter was carried out as a proof of concept with the aim being to determine whether agent based model (ABM) could be adapted to be used for overall optimisation of a CHO cell cultivation and purification. The results have shown, through the development of the model for the first CPP, titre, that the ABM is applicable. It is possible to specify the desired end titre and to predict the titre required to achieve this after the cultivation, and subsequently the conditions the cultivation needs to be operated at to achieve this. Due to lack of data the IEX unit models could not be validated. However this has identified one of the key areas for further development which has been touched upon in this research, the use of first principles models. If first principles models were used in the characterisation of the process units then issues encountered here with the IEX column would not of had such a big impact. Work was presented in Chapter 5 which characterised the metabolism of the cells and could predict titre, viable cell count, and metabolite concentrations. Chapter 5 showed that the predictions from these models were not as accurate as for multivariate or hybrid models but the use of them would significantly reduce the experimental data required. Additionally issues may arise in the mathematical modelling of CQAs such as the glycosylation profile. Work presented in this thesis managed to predict glycosylation profile only through the use of multivariate PLS modelling. This may present issues in further development.

Chapter 8

Conclusions and future work

This project has covered three areas of bioprocess modelling. The first, modelling of cell cultivation, the second, modelling of ion exchange chromatography, and the third, agent based modelling of multiple process units. The diversity in the units considered is a reflection of the need identified by Fujifilm Diosynth Biotechnologies, for development of both upstream and downstream. The research has shown the potential of agent based modelling for bioprocesses. With the potential for the agent based model to be used with various CPPs and CQAs to contribute towards the optimisation of a process within the design space, as defined in the QbD initiative. For a pharmaceutical product this can improve the efficiency, maintain product quality, and provide more information of the operation of the system.

8.1 Modelling of cell cultivation

The research presented in this thesis concerning the modelling of cell cultivation showed that multivariate, first principles, and hybrid models could be used to predict the final glycosylation profile, the profile of the metabolites in the cultivation, and the product titre. Of the three modelling techniques only the multivariate modelling was successfully applied to predict the glycosylation profile. This is because the mechanism behind the production of the various glycans is not yet fully understood and characterised. Whereas the

multivariate models are produced from data. This meant the data fro the operational set points of each batch could be used with the final glycosylation measurements. THis produced accurate models which could predict the final glycosylation profile.

Similarly for product titre the multivariate PLS models produced which used the operational set points, and concentration of glucose and lactate to predict the product titre which resulted in satisfactory predictions. Additionally the product titre could be predicted using first principles models and through hybrid modelling with the best predictions being with the hybrid model. A similar situation occurred for the modelling of the metabolites during the cultivation, with the hybrid model producing the best results. The hybrid model had the additional benefit of providing predictions for metabolites which are not regularly measured during standard operating procedures.

With regards to the implementation of these models in industry the research has shown that the hybrid approach has the most benefits. As the hybrid models combine the best of both multivariate and first principles. Considering these two separately it has been shown that the multivariate models are generally more accurate than the first principles models. However, the first principles models have the benefit of requiring less data, and thus can be applied at an earlier stage in the process development. The hybrid model combines these two approaches and allows for the flexibility of the first principles model in accounting for changes to the cultivation conditions (i.e. different starting glucose concentrations) whilst combining the ability of the multivariate models to account for changes in the operational set points. To further develop the models presented in this thesis the main recommendation would be to train the models using a data set specific to CHO cells.

8.2 Modelling of ion exchange chromatography

A similar approach was taken in the modelling of the ion exchange chromatography column. In that it was considered as a separate unit prior to consideration for use in the agent based model. Again multivariate PLS

modelling was used to predict the yield, and the UV absorbance output for the elution peak. The yield was predicted well. However, problems were encountered with th prediction of the elution peak. This was due to the non-linear nature fo the peak. This problem was solved by using the gradient of the elution peak curve, and the cumulative area under the curve as the output of the PLS models.

Additionally a problem was encountered in using PLS modelling to predict the retention time. This problem was solved by using a first principles model. The model used the lactoferrin data set to determine the constants of the model and from this predicted the retention time. From the validation batch used to test the model it suggested that the model performed well. However, it is uncertain how this model would perform when predicting batches in which the operating conditions are vastly different to the conditions of the batches used to train the model.

There is a need for accurate IEX models to be used in industry as they can be implemented during early stages of development to help direct and establish the platform process. The benefit of employing the models lies both in reduction of costs, as fewer experiments would be required, but also that less time would be required. In order to develop the models presented in this thesis further, a data set specific to monoclonal antibodies should be used. Additionally other modelling techniques should be explored, with one suggestion being the use of non-linear PLS.

8.3 Agent based modelling

This thesis also presented a proof of concept study for the application of agent based modelling to bioprocessing. The main aim of this research was to determine whether agent based modelling could successfully be used to predict how changes to the operational parameters of different units would effect certain CPPs and CQAs, and what subsequent changes would have to be made to the operation of other process units in order to maintain the same product quality, and efficacy of the bulk drug substance.

The model was limited by the lack of data, in that two distinct data sets were used, for upstream and one for downstream. Both of these data sets were for CHO cells but they were not for the same process. This problem was further exacerbated by the lack of available data in the downstream processing relating to measurements taken post cultivation. This was a challenge as these measurements were required to relate the two data sets. This problem was handled by using the product titre as the communicating CPP between the process units.

It was shown that the product titre did not predict well the other CPPs and CQAs (HCP, SEC, CIEX etc). One possible reason for this may be the nature of the measurements. For example the CIEX measurement after each process unit was either 'main peak', 'acidic peak', or 'basic peak' which is not a definitive value. However, the model did show that the use of the product titre in combination with the hybrid model allowed for changes to be made the BDS and predictions as to the operational condition set points and cultivation profile to be made.

Additionally it can be seen for all aspects of this thesis, especially the agent based work, that high throughput technology would be beneficial in generating the data required to train and test the models. High through put technology would allow for the quick generation of data sets with minimal time, money, and effort required to produce them.

8.4 Summary

In summary the work presented in this thesis has shown that agent based models can be applied to bioprocesses. To further develop the study, a more robust data set is required, which is specifically for a monoclonal antibody producing Chinese hamster ovary cell line. Additionally the agent based model requires further testing with a data set which contains upstream and downstream data which is for the same process. However the positive potential impact of the development of the agent based model would more than justify this. As the agent based model has the potential to simplify the optimisation of

the cultivation and purification of mammalian cell products.

References

Abdi, H. Partial least squares regression and projection on latent structure regression (pls regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):97–106, 2010.

Abe, H., Saito, H., Miyakawa, H., Tamura, Y., Shimamura, S., Nagao, E., and Tomita, M. Heat stability of bovine lactoferrin at acidic ph. *Journal of Dairy Science*, 74(1):65–71, 1991.

Abonyi, J., Madar, J., and Szeifert, F. Combining first principles models and neural networks for generic model control. In *Soft Computing and Industry*, pages 111–122. Springer Science Business Media, 2002. doi: 10.1007/978-1-4471-0123-9_10. URL http://dx.doi.org/10.1007/978-1-4471-0123-9_10.

Abraham, V., Taylor, D., and Haskins, J. High content screening applied to large-scale cell biology. *Trends in Biotechnology*, 22(1):15–22, 2004. ISSN 0167-7799. doi: http://dx.doi.org/10.1016/j.tibtech.2003.10.012. URL http://www.sciencedirect.com/science/article/pii/S0167779903003068.

Adrio, J. and Demain, A. Fungal biotechnology. *International Microbiology*, 6 (3):191–199, sep 2003. doi: 10.1007/s10123-003-0133-0. URL http://dx.doi.org/10.1007/s10123-003-0133-0.

Agarwal, M. Combining neural and conventional paradigms for modeling, prediction, and control. In *Proceedings of International Conference on Control Applications*. Institute of Electrical & Electronics Engineers (IEEE), 1995. doi: 10.1109/cca.1995.555789. URL http://dx.doi.org/10.1109/cca.1995.555789.

Aguiar, H. and Filho, R. Neural network and hybrid model: a discussion about different modeling techniques to predict pulping degree with industrial data. *Chemical Engineering Science*, 56(2):565–570, jan 2001. doi: 10.1016/s0009-2509(00)00261-x. URL http://dx.doi.org/10.1016/s0009-2509(00)00261-x.

Ahn, W., Jeon, J., Jeong, Y., Lee, S., and Yoon, S. Effect of culture temperature on erythropoietin production and glycosylation in a perfusion culture of recombinant cho cells. *Biotechnology and bioengineering*, 101 (6):1234–1244, 2008.

Akaike, H. Factor analysis and aic. *Psychometrika*, 52(3):317–332, 1987.

Albert, S. and Kinley, R. Multivariate statistical monitoring of batch processes: an industrial case study of fermentation supervision. *TRENDS in Biotechnology*, 19(2):53–62, 2001.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. *Molecular Biology of the Cell*. Garland Science, 2002.

Aldington, S. and Bonnerjea, J. Scale-up of monoclonal antibody purification processes. *Journal of Chromatography B*, 848(1):64–78, mar 2007. doi: 10.1016/j.jchromb.2006.11.032. URL http://dx.doi.org/10.1016/j.jchromb.2006.11.032.

Almquist, J., Cvijovic, M., Hatzimanikatis, V., Nielsen, J., and Jirstrand, M. Kinetic models in industrial biotechnology – improving cell factory performance. *Metabolic Engineering*, 24:38–60, jul 2014. doi: 10.1016/j.ymben.2014.03.007. URL http://dx.doi.org/10.1016/j.ymben.2014.03.007.

Altamirano, C., Illanes, A., Casablancas, A., Gamez, X., Cairo, J., and Godia, C. Analysis of cho cells metabolic redistribution in a glutamate-based defined medium in continuous culture. *Biotechnology Progress*, 17(6): 1032–1041, 2001.

Amand, M., Tran, K., Radhakrishnan, D., Robinson, A., and Ogunnaike, B. Controllability analysis of protein glycosylation in cho cells. *PloS one*, 9(2): 1–16, 2014. doi: 10.1371/journal.pone.0087973. URL http://dx.doi.org/10.1371/journal.pone.0087973.

Andersen, C. and Bro, R. The n-way toolbox for matlab. *Chemometrics and Intelligent Laboratory Systems*, 52:1–4, 2000.

Andersen, C. and Bro, R. Variable selection in regression — a tutorial. *Journal of Chemometrics*, 24(1112):728–737, 2010. ISSN 1099-128X. doi: 10.1002/cem.1360.

Anderson, A. A hybrid mathematical model of solid tumour invasion: the importance of cell adhesion. *Math. Med. Biol*, 22(2):163–186, 2005.

Arneberg, R., Rajalahti, T., Flikka, K., Berven, F. S., Kroksveen, A. C., Berle, M., Myhr, K.-M., Vedeler, C. A., Ulvik, R. J., and Kvalheim, O. M. Pretreatment of mass spectral profiles: Application to proteomic data. *Analytical Chemistry*, 79(18):7014–7026, sep 2007. doi: 10.1021/ac070946s. URL http://dx.doi.org/10.1021/ac070946s.

Artursson, T., Hagman, A., Björk, S., Trygg, J., Wold, S., and Jacobsson, S. P. Study of preprocessing methods for the determination of crystalline phases in binary mixtures of drug substances by x-ray powder diffraction and multivariate calibration. *Applied Spectroscopy*, 54(8):1222–1230, aug 2000. doi: 10.1366/0003702001950805. URL http://dx.doi.org/10.1366/0003702001950805.

Arunakumari, A., Dembecki, J., Ferreira, G., and Patel, K. Cchromatography: A two-column process to purify antibodies without protein a. *BioPharm International*, 20(5), may 2007.

Baba, Y., Yoza, N., and Ohashi, S. Computer-assisted prediction of retention times for inorganic polyphosphates in gradient ion-exchange chromatography. *Journal of Chromatography A*, 350:461–467, jan 1985. doi: 10.1016/s0021-9673(01)93552-6. URL http://dx.doi.org/10.1016/s0021-9673(01)93552-6.

Bailey, J. Mathematical modeling and analysis in biochemical engineering: Past accomplishments and future opportunities. *Biotechnol. Prog.*, 14(1): 8–20, feb 1998a. doi: 10.1021/bp9701269. URL http://dx.doi.org/10.1021/bp9701269.

Bailey, J. Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. *Biotechnology progress*, 14 (1):8–20, 1998b.

Baker, E. N. and Baker, H. M. Lactoferrin. *Cellular and Molecular Life Sciences*, 62(22):2531–2539, nov 2005. doi: 10.1007/s00018-005-5368-9. URL http://dx.doi.org/10.1007/s00018-005-5368-9.

Baker, K. UCLA department of mathematics online resource: Cubic spline curves. http://www.math.ucla.edu/~baker/149.1.02w/handouts/dd_splines.pdf, 2014. [Accessed on: 4th October 2014].

Barnes, R. J., Dhanoa, M. S., and Lister, S. J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5):772–777, jul 1989. doi: 10.1366/0003702894202201. URL http://dx.doi.org/10.1366/0003702894202201.

Batt, B. and Kompala, D. A structured kinetic modeling framework for the dynamics of hybridoma growth and monoclonal antibody production in continuous suspension cultures. *Biotechnol. Bioeng.*, 34(4):515–531, aug 1989. doi: 10.1002/bit.260340412. URL http://dx.doi.org/10.1002/bit.260340412.

Berrut, J. and Trefethen, L. Barycentric lagrange interpolation. *Siam Review*, 46(3):501–517, 2004.

Bhalla, U. and Iyengar, R. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387, jan 1999. doi: 10.1126/science.283.5400.381. URL http://dx.doi.org/10.1126/science.283.5400.381.

Bhutani, N., Rangaiah, G., and Ray, A. First-principles, data-based, and hybrid modeling and optimization of an industrial hydrocracking unit. *Industrial & Engineering Chemistry Research*, 45(23):7807–7816, nov 2006. doi: 10.1021/ie060247q. URL http://dx.doi.org/10.1021/ie060247q.

Birch, J. and Racher, A. Antibody production. *Advanced Drug Delivery Reviews*, 58(5-6):671–685, aug 2006. doi: 10.1016/j.addr.2005.12.006. URL http://dx.doi.org/10.1016/j.addr.2005.12.006.

Bollin, F., Dechavanne, V., and Chevalet, L. Design of experiment in cho and hek transient transfection condition optimization. *Protein expression and purification*, 78(1):61–68, 2011.

Borys, M., Linzer, D., and Papoutsakis, E. Culture ph affects expression rates and glycosylation of recombinant mouse placental lactogen proteins by chinese hamster ovary (cho) cells. *Nature Biotechnology*, 11(6):720–724, 1993.

Boychyn, M., Yim, S., Bulmer, M., More, J., Bracewell, D., and Hoare, M. Performance prediction of industrial centrifuges using scale-down models. *Bioprocess and biosystems engineering*, 26(6):385–391, 2004.

Bradshaw, J., Dutfield, S., Benoit, P., and Woolley, J. Kaos: Toward an industrial-strength open agent architecture. *Software agents*, pages 375–418, 1997.

Bree, M., Dhurjati, P., Geoghegan, R., and Robnett, B. Kinetic modelling of hybridoma cell growth and immunoglobulin production in a large-scale suspension culture. *Biotechnol. Bioeng.*, 32(8):1067–1072, oct 1988. doi: 10.1002/bit.260320814. URL `http://dx.doi.org/10.1002/bit.260320814`.

Bro, R. Parafac. tutorial and applications. *Chemometrics and Intelligent Laboratory systems*, 38:149–171, 1997.

Bro, R. and Smilde, A. Principal component analysis. *Analytical Methods*, 6 (9):2812–2831, 2014. ISSN 1759-9660. doi: 10.1039/C3AY41907J.

Bro, R., Andersson, C., and Kiers, H. Parafac2—part ii. modeling chromatographic data with retention time shifts. 13:295—-309, 1999. ISSN 1099-128X. doi: 10.1002/(SICI)1099-128X(199905/08)13:3/4⟨295::AID-CEM547⟩3.0.CO;2-Y.

Bro, R., Kjeldahl, K., Smilde, A., and Kiers, H. Cross-validation of component models: a critical look at current methods. *Analytical and bioanalytical chemistry*, 390(5):1241–51, 2008. ISSN 1618-2642. doi: 10.1007/s00216-007-1790-1.

Broughton, D. and Gerhold, C. Continuous sorption process employing fixed bed of sorbent and moving inlets and outlets, May 23 1961. US Patent 2,985,589.

Butler, M. *Animal Cell Culture and Technology*. BIOS Scientific Publishers, 2nd edition, 2004. ISBN 1-859960499.

Butler, M. Optimisation of the cellular metabolism of glycosylation for recombinant proteins produced by mammalian cell systems. *Cytotechnology*, 50(1-3):57–76, 2006.

Bylund, D., Danielsson, R., Malmquist, G., and Markides, K. E. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography–mass spectrometry data. *Journal of Chromatography A*, 961(2):237–244, jul 2002. doi: 10.1016/s0021-9673(02)00588-5. URL `http://dx.doi.org/10.1016/s0021-9673(02)00588-5`.

Carlier, A. and Kroonenberg, P. Decompositions and biplots in three-way correspondence analysis. *Psychometrika*, 61(2):355–373, 1996.

Carrier, T., Heldin, E., Ahnfelt, M., Brekkan, E., Hassett, R., Peppers, S., Rodrigo, G., Van Slyke, G., and Zhao, D. High-throughput technologies in bioprocess development. *Bioprocess, Bioseparation, and Cell Technology*, Oct 2009. doi: 10.1002/9780470054581.eib134.

Carrondo, M., Alves, P., Carinhas, N., Glassey, J., Hesse, F., Merten, O., Micheletti, M., Noll, T., Oliveira, R., Reichl, U., Staby, A., Teixeira, A., Weichert, H., and Mandenius, C. How can measurement, monitoring, modeling and control advance cell culture in industrial biotechnology? *Biotechnology Journal*, 7(12):1522–1529, 2012. doi: 10.1002/biot.201200226. URL `http://dx.doi.org/10.1002/biot.201200226`.

Catapano, G., Czermak, P., Eibl, R., and Eibl, D. *Cell and Tissue Reaction Engineering*. 2009.

Cattell, R. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

Cervera, A., Petersen, N., Lantz, A., Larsen, A., and Gernaey, K. Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation. *Biotechnology progress*, 25(6):1561–1581, 2009.

Chen, J. and Huang, T. Applying neural networks to on-line updated PID controllers for nonlinear process control. *Journal of Process Control*, 14(2): 211–230, mar 2004. doi: 10.1016/s0959-1524(03)00039-8. URL `http://dx.doi.org/10.1016/s0959-1524(03)00039-8`.

Chen, S., Lau, H., Brodsky, Y., Kleemann, G., and Latypov, R. The use of native cationexchange chromatography to study aggregation and phase separation of monoclonal antibodies. 19:1191–1204, 2010. ISSN 1469-896X. doi: 10.1002/pro.396.

Choi, D. and Park, H. A hybrid artificial neural network as a software sensor for optimal control of a wastewater treatment process. *Water research*, 35 (16):3959–3967, 2001.

Costa, A., Rodrigues, M., Henriques, M., Azeredo, J., and Oliveira, R. Guidelines to cell engineering for monoclonal antibody production. *European Journal of Pharmaceutics and Biopharmaceutics*, 74(2):127–38, 2010. ISSN 0939-6411. doi: 10.1016/j.ejpb.2009.10.002.

Craig, A., Cloarec, O., Holmes, E., Nicholson, J., and Lindon, J. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7):2262–2267, apr 2006. doi: 10.1021/ac0519312. URL `http://dx.doi.org/10.1021/ac0519312`.

Dalal, D. K. and Zickar, M. J. Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods*, 15(3):339–362, dec 2011. doi: 10.1177/1094428111430540. URL `http://dx.doi.org/10.1177/1094428111430540`.

Dautray, R. and Lions, J.-L. *Mathematical Analysis and Numerical Methods for Science and Technology*. Springer Berlin Heidelberg, 2000. doi: 10.1007/978-3-642-58090-1. URL http://dx.doi.org/10.1007/978-3-642-58090-1.

Davies, J., Baganz, F., Ison, A., and Lye, G. Studies on the interaction of fermentation and microfiltration operations: erythromycin recovery from saccharopolyspora erythraea fermentation broths. *Biotechnology and bioengineering*, 69(4):429–439, 2000.

del Val, I. J., Nagy, J. M., and Kontoravdi, C. A dynamic mathematical model for monoclonal antibody n-linked glycosylation and nucleotide sugar donor transport within a maturing golgi apparatus. *Biotechnol Progress*, 27(6): 1730–1743, sep 2011. doi: 10.1002/btpr.688. URL http://dx.doi.org/10.1002/btpr.688.

Desai, M. *Downstream Processing of Proteins*. Humana Press, 2000. ISBN 0-89603-564-6.

Detroyer, A., Schoonjans, V., Questier, F., Heyden, Y. V., Borosy, A., Guo, Q., and Massart, D. Exploratory chemometric analysis of the classification of pharmaceutical substances based on chromatographic data. *Journal of Chromatography A*, 897(1-2):23–36, nov 2000. doi: 10.1016/s0021-9673(00)00803-7. URL http://dx.doi.org/10.1016/s0021-9673(00)00803-7.

Dewan, S. Antibody drugs: Technologies and global markets, 2015. URL http://www.bccresearch.com/market-research/biotechnology/antibody-drugs-market-bio016j.html.

Dhanoa, M., Lister, S., Sanderson, R., and Barnes, R. The link between multiplicative scatter correction (msc) and standard normal variate (snv) transformations of nir spectra. *Journal of Near Infrared Spectroscopy*, 2: 43–47, 1994. doi: 10.1255/jnirs.30.

Doyle, A. and Griffiths, J. *Cell and tissue Culture: Laboratory Procedures in Biotechnology*. Wiley, 1998.

Ducommun, P., Ruffieux, P., Kadouri, A., Stockar, U., and Marison, I. Monitoring of temperature effects on animal cell metabolism in a packed bed process. *Biotechnology and Bioengineering*, 2002. ISSN 1097-0290. doi: 10.1002/bit.10185.

Dudoit, S., Fridlyand, J., and Speed, T. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87, 2002.

Eibl, R., Eibl, D., and Pörtner, R. *Cell and tissue Reaction Engineering: Principles and Practise*. Springer-Verlag, 2008.

Eibl, R., Kaiser, S., Lombriser, R., and Eibl, D. Disposable bioreactors: the current state-of-the-art and recommended applications in biotechnology. *Applied microbiology and biotechnology*, 86(1):41–49, 2010.

Eigenvector Inc. Advanced preprocessing: Sample normalization, 2015. URL http://wiki.eigenvector.com/index.php?title=Advanced_Preprocessing:_Sample_Normalization.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J., Downey, G., Blanchet, L., and Buydens, L. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, 50:96–106, oct 2013. doi: 10.1016/j.trac.2013.04.015. URL http://dx.doi.org/10.1016/j.trac.2013.04.015.

Eriksson, L., Byrne, T., Johansson, E., Trygg, J., and Vikström, C. *Multi-and megavariate data analysis basic principles and applications*. Umetrics Academy, 2013.

European Medicines Agency. Development pharmaceutics for biotechnological and biological products, 2000.

Evens, R. and Kaitin, K. The biotechnology innovation machine: A source of intelligent biopharmaceuticals for the pharma industry—mapping biotechnology's success. *Clinical Pharmacology and Therapeutics*, 95: 528–532, 2014. URL http://www.ncbi.nlm.nih.gov/pubmed/24448474.

Ezzell, C. Magic bullets fly again, 2001.

Farid, S., Washbrook, J., Birch, J., and Titchener-Hooker, N. A hierarchical framework for modelling biopharmaceutical manufacture to address process and business needs. In *Computer Aided Chemical Engineering*, pages 673–678. Elsevier BV, 2000. doi: 10.1016/s1570-7946(00)80114-5. URL http://dx.doi.org/10.1016/s1570-7946(00)80114-5.

Farid, S. Process economics of industrial monoclonal antibody manufacture. *Journal of Chromatography B*, 848(1):8–18, mar 2007. doi: 10.1016/j.jchromb.2006.07.037. URL http://dx.doi.org/10.1016/j.jchromb.2006.07.037.

Fellner, M., Delgado, A., and Becker, T. Functional nodes in dynamic neural networks for bioprocess modelling. *Bioprocess Biosyst Eng*, 25(5):263–270, mar 2003. doi: 10.1007/s00449-002-0297-6. URL http://dx.doi.org/10.1007/s00449-002-0297-6.

Fernandes, R., Bodla, V., Carlquist, M., Heins, A., Lantz, A., Sin, G., and Gernaey, K. Applying mechanistic models in bioprocess development. *Advances in biochemical engineering/biotechnology*, 132:137–166, 2013. ISSN 0724-6145. doi: 10.1007/10_2012_166.

Fernandes, R. L., Bodla, V. K., Carlquist, M., Heins, A.-L., Lantz, A. E., Sin, G., and Gernaey, K. V. Applying mechanistic models in bioprocess development. In *Measurement, Monitoring, Modelling and Control of Bioprocesses*, pages 137–166. Springer Science Business Media, 2012. doi: 10.1007/10_2012_166. URL http://dx.doi.org/10.1007/10_2012_166.

Fiedler, B. and Schuppert, A. Local identification of scalar hybrid models with tree structure. *IMA Journal of Applied Mathematics*, 73(3):449–476, jan 2008. doi: 10.1093/imamat/hxn011. URL http://dx.doi.org/10.1093/imamat/hxn011.

Fodor, I. A survey of dimension reduction techniques, 2002.

Fradkin, A., Carpenter, J., and Randolph, T. Immunogenicity of aggregates of recombinant human growth hormone in mouse models. *Journal of pharmaceutical sciences*, 98(9):3247–3264, 2009.

Frame, K. and Hu, W. Kinetic study of hybridoma cell growth in continuous culture. i. a model for non-producing cells. *Biotechnol. Bioeng.*, 37(1): 55–64, jan 1991. doi: 10.1002/bit.260370109. URL http://dx.doi.org/10.1002/bit.260370109.

Frank, A. Lactoferrin molecule rendering, 2014. URL https://commons.wikimedia.org/wiki/File:Lactoferrin_molecule_rendering.png#/media/File:Lactoferrin_molecule_rendering.png.

Fransson, M., Sparén, A., Lagerholm, B., and Karlsson, L. On-line process control of liquid chromatography. *Anal. Chem.*, 73(7):1502–1508, apr 2001. doi: 10.1021/ac001149w. URL http://dx.doi.org/10.1021/ac001149w.

Fujifilm Diosynth Biotechnologies. Apollo™mammalian expression platform, 2014. URL {http://www.fujifilmdiosynth.com/mcc/apollo.htm}. Brochure.

Gabriel, K. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

Gadgil, M. Development of a mathematical model for animal cell culture without pH control and its application for evaluation of clone screening outcomes in shake flask culture. *Journal of Chemical Technology & Biotechnology*, 90(1):166–175, feb 2014. doi: 10.1002/jctb.4302. URL http://dx.doi.org/10.1002/jctb.4302.

Gadgil, M. Development of a mathematical model for animal cell culture without ph control and its application for evaluation of clone screening outcomes in shake flask culture. *Journal of Chemical Technology Biotechnology*, 90:166–175, 2015a. ISSN 0268-2575. doi: 10.1002/jctb.4302.

Gadgil, M. Development of a mathematical model for animal cell culture without ph control and its application for evaluation of clone screening outcomes in shake flask culture. *Journal of Chemical Technology and Biotechnology*, 90(1):166–175, 2015b.

Galfre, G., Secher, D., and Crawley, P. *Ullman's Biotechnology and Biochemical Engineering*, volume 1. Wiley-VCH, 2007. ISBN 978-3-527-31603-8.

Galvanauskas, V., Simutis, R., and Lübbert, A. Hybrid process models for process optimisation, monitoring and control. *Bioprocess Biosyst Eng*, 26 (6):393–400, nov 2004. doi: 10.1007/s00449-004-0385-x. URL http://dx.doi.org/10.1007/s00449-004-0385-x.

Gao, Y., Kipling, K., Glassey, J., Willis, M., Montague, G., Zhou, Y., and Titchener-Hooker, N. Application of agent-based system for bioprocess description and process improvement. *Biotechnol Progress*, 26(3):706–716, dec 2009. doi: 10.1002/btpr.361. URL http://dx.doi.org/10.1002/btpr.361.

García-Flores, R. and Wang, X. A multi-agent system for chemical supply chain simulation and management support. *Or Spectrum*, 24(3):343–370, 2002.

Gawlitzek, M., Ryll, T., Lofgren, J., and Sliwkowski, M. Ammonium alters n-glycan structures of recombinant tnfr-igg: Degradative versus biosynthetic mechanisms. *Biotechnology and Bioengineering*, 68(6):637–646, 2000.

GE-Healthcare. *Getting Started with Äkta Explorer*, 2010. URL https://www.gelifesciences.com/gehcls_images/GELS/Related%20Content/Files/1314807262343/litdoc28953706_20131222225146.pdf.

Geladi, P. and Kowalski, B. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 1986. URL http://www.sciencedirect.com/science/article/pii/0003267086800289.

Geladi, P., MacDougall, D., and Martens, H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, 39(3):491–500, May 1985. URL http://as.osa.org/abstract.cfm?URI=as-39-3-491.

Genesereth, M. and Ketchpel, S. Software agents. *Commun. ACM*, 37(7): 48–53, 1994.

Gerdtzen, Z. Modeling metabolic networks for mammalian cell systems: General considerations, modeling strategies, and available tools. In *Genomics and Systems Biology of Mammalian Cell Culture*, pages 71–108. Springer, 2012.

Giddings, J. C. *Dynamics of chromatography: Principles and theory*. CRC Press, March 2002.

Giddings, J. Kinetic processes and zone diffusion in chromatography. *Journal of Chromatography A*, 3:443–453, 1960.

Glacken, M., Adema, E., and Sinskey, A. Mathematical descriptions of hybridoma culture kinetics: I. initial metabolic rates. *Biotechnology and bioengineering*, 32(4):491–506, 1988. ISSN 1097-0290. doi: 10.1002/bit.260320412.

Glacken, M., Huang, C., and Sinskey, A. Mathematical descriptions of hybridoma culture kinetics. III. simulation of fed-batch bioreactors. *Journal of Biotechnology*, 10(1):39–65, apr 1989. doi: 10.1016/0168-1656(89)90091-6. URL http://dx.doi.org/10.1016/0168-1656(89)90091-6.

Glassey, J., Gernaey, K., Clemens, C., Schulz, T., Oliveira, R., Striedner, G., and Mandenius, C. Process analytical technology (pat) for biopharmaceuticals. *Biotechnology journal*, 6(4):369–377, 2011a.

Glassey, J., Gernaey, K., Clemens, C., Schulz, T., Oliveira, R., Striedner, G., and Mandenius, C. Process analytical technology (PAT) for biopharmaceuticals. *Biotechnology Journal*, 6(4):369–377, 2011b. doi: 10.1002/biot.201000356. URL http://dx.doi.org/10.1002/biot.201000356.

Gnoth, S., Jenzsch, M., Simutis, R., and ű bbert, A. Process analytical technology (pat): Batch-to-batch reproducibility of fermentation processes by robust process operational design and control. *Journal of Biotechnology*, 132(2):180–186, 2007. doi: 10.1016/j.jbiotec.2007.03.020. URL http://dx.doi.org/10.1016/j.jbiotec.2007.03.020.

Golmakany, N., Rasaee, M., Furouzandeh, M., Shojaosadati, S., Kashanian, S., and Omidfar, K. Continuous production of monoclonal antibody in a packedbed bioreactor. *Biotechnology and Applied Biochemistry*, 41, 2005. ISSN 1470-8744. doi: 10.1042/BA20040121.

Gomez, N., Ouyang, J., Nguyen, M., Vinson, A., Lin, A., and Yuk, I. Effect of temperature, ph, dissolved oxygen, and hydrolysate on the formation of triple light chain antibodies in cell culture. *Biotechnology Progress*, pages 1438–1445, 2010. ISSN 1520-6033. doi: 10.1002/btpr.465.

Gorban, A., Kegl, B., Wunsch, D., and Zinovyev, A. *Principal manifolds for data visualization and dimension reduction*, volume 1. Springer, 2008.

Gottschalk, U. Biotech manufacturing is coming of age. *BioProcess International*, pages 1–7, 2003.

Gromski, P., Xu, Y., Hollywood, K., Turner, M., and Goodacre, R. The influence of scaling metabolomics data on model classification accuracy. *Metabolomics*, oct 2014. doi: 10.1007/s11306-014-0738-7. URL http://dx.doi.org/10.1007/s11306-014-0738-7.

Grung, B. and Kvalheim, O. M. Retention time shift adjustments of two-way chromatograms using bessel's inequality. *Analytica Chimica Acta*, 304(1): 57 – 66, 1995. ISSN 0003-2670. doi: http://dx.doi.org/10.1016/0003-2670(94)00587-C. URL http://www.sciencedirect.com/science/article/pii/000326709400587C.

Gu, T. *Mathematical Modeling and Scale-up of Liquid Chromatography*. Springer, 1st edition, 1995. ISBN 3-540-58884-1.

Guido, N., Wang, X., Adalsteinsson, D., McMillen, D., Hasty, J., Cantor, C., Elston, T., and Collins, J. A bottom-up approach to gene regulation. *Nature*, 439(7078):856–860, feb 2006. doi: 10.1038/nature04473. URL http://dx.doi.org/10.1038/nature04473.

Haack, M., Lantz, A., Mortensen, P., and Olsson, L. Chemometric analysis of in-line multi-wavelength fluorescence measurements obtained during cultivations with a lipase producing aspergillus oryzae strain. *Biotechnology and bioengineering*, 96(5):904–913, 2007.

Haber, R. and Unebnhauen, H. Structure identification of non-linear dynamic systems a survey on input output approaches. *Automatica*, 26:651–677, 1990.

Hacker, D., De Jesus, M., and Wurm, F. 25 years of recombinant proteins from reactor-grown cells - where do we go from here? *Biotechnology advances*, 27:1023–1027, 2009.

Hakansson, A., Zhivotovsky, B., Orrenius, S., Sabharwal, H., and Svanborg, C. Apoptosis induced by a human milk protein. *Proceedings of the National Academy of Sciences*, 92(17):8064–8068, aug 1995. doi: 10.1073/pnas.92.17.8064. URL http://dx.doi.org/10.1073/pnas.92.17.8064.

Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. From molecular to modular cell biology. *Nature*, 402(supp):C47–C52, dec 1999. doi: 10.1038/35011540. URL http://dx.doi.org/10.1038/35011540.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer New York, 2009. doi: 10.1007/978-0-387-84858-7. URL http://dx.doi.org/10.1007/978-0-387-84858-7.

Heath, C. and Kiss, R. Cell culture process development: Advances in process engineering. *Biotechnology Progress*, 2007. ISSN 1520-6033. doi: 10.1021/bp060344e.

Helland, I. S., Næs, T., and Isaksson, T. Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 29(2):233–241, oct 1995. doi: 10.1016/0169-7439(95)80098-t. URL http://dx.doi.org/10.1016/0169-7439(95)80098-t.

Hemmateenejad, B., Akhond, M., and Samari, F. A comparative study between pcr and pls in simultaneous spectrophotometric determination of diphenylamine, aniline, and phenol: effect of wavelength selection. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 67 (3):958–965, 2007.

Hendriks, M., Cruz-Juarez, L., Bont, D. D., and Hall, R. Preprocessing and exploratory analysis of chromatographic profiles of plant extracts. *Analytica Chimica Acta*, 545(1):53–64, jul 2005. doi: 10.1016/j.aca.2005.04.026. URL http://dx.doi.org/10.1016/j.aca.2005.04.026.

Henning, J. Interpolation utilities. MATLAB Central. http://www.mathworks.co.uk/matlabcentral/fileexchange/36800-interpolation-utilities/content/cubiconv.m. [Accessed on: 2nd February 2014].

Hernandez, R. and Brown, D. Growth and maintenance of baby hamster kidney (BHK) cells. *Current protocols in microbiology*, 2010. doi: 10.1002/9780471729259.mca04hs17.

Heyer, L., Kruglyak, S., and Yooseph, S. Exploring expression data: identification and analysis of coexpressed genes. *Genome research*, 9(11): 1106–1115, 1999.

Hiden, H., Willis, M., Tham, M., and Montague, G. Non-linear principal components analysis using genetic programming. *Computers & chemical engineering*, 23(3):413–425, 1999.

Hoffmann, N., Keck, M., Neuweger, H., Wilhelm, M., Högy, P., Niehaus, K., and Stoye, J. Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets. *BMC Bioinformatics*, 13(1):214, 2012. doi: 10.1186/1471-2105-13-214.

Hooker, A., Goldman, M., Markham, N., James, D., Ison, A., Bull, A., Strange, P., Salmon, I., Baines, A., and Jenkins, N. N-glycans of recombinant human interferon-γ change during batch culture of chinese hamster ovary cells. *Biotechnology and bioengineering*, 48(6):639–648, 1995.

Horvath, C., Preiss, B., and Lipsky, S. Fast liquid chromatography. investigation of operating parameters and the separation of nucleotides on pellicular ion exchangers. *Analytical chemistry*, 39(12):1422–1428, 1967.

Hossler, P., Khattak, S., and Li, Z. Optimal and consistent protein glycosylation in mammalian cell culture. *Glycobiology*, 19(9):936–49, 2009. ISSN 0959-6658. doi: 10.1093/glycob/cwp079.

Hou, Y., Jiang, C., Shukla, A. A., and Cramer, S. M. Improved process analytical technology for protein a chromatography using predictive principal component analysis tools. *Biotechnology and Bioengineering*, 108 (1):59–68, jul 2010. doi: 10.1002/bit.22886. URL http://dx.doi.org/10.1002/bit.22886.

Hu, B., Qu, H., Wang, Y., and Yang, S. A generalized-constraint neural network model: Associating partially known relationships for nonlinear regressions73. *Information Sciences*, 179(12):1929–1943, may 2009. doi: 10.1016/j.ins.2009.02.006. URL http://dx.doi.org/10.1016/j.ins.2009.02.006.

Hua, Y., Tu, K., Tang, Z., Li, Y., and Xiao, H. Comparison of normalization methods with microRNA microarray. *Genomics*, 92(2):122–128, aug 2008. doi: 10.1016/j.ygeno.2008.04.002. URL http://dx.doi.org/10.1016/j.ygeno.2008.04.002.

Huang, J., Kaul, G., Cai, C., Chatlapalli, R., Hernandez-Abad, P., Ghosh, K., and Nagi, A. Quality by design case study: an integrated multivariate approach to drug product and process development. *International journal of pharmaceutics*, 382(1-2):23–32, 2009. ISSN 0378-5173. doi: 10.1016/j.ijpharm.2009.07.031.

Huuk, T. C., Hahn, T., Osberghaus, A., and Hubbuch, J. Model-based integrated optimization and evaluation of a multi-step ion exchange chromatography. *Separation and Purification Technology*, 136:207–222, nov 2014. doi: 10.1016/j.seppur.2014.09.012. URL http://dx.doi.org/10.1016/j.seppur.2014.09.012.

Hwang, S., Yoon, S., Koh, G., and Lee, G. Effects of culture temperature and ph on flag-tagged comp angiopoietin-1 (fca1) production from recombinant cho cells: Fca1 aggregation. *Applied microbiology and biotechnology*, 91 (2):305–315, 2011.

Hwang, T., Oh, H., Choi, Y., Nam, S., Lee, S., and Choung, Y. Development of a statistical and mathematical hybrid model to predict membrane fouling and performance. *Desalination*, 247(1-3):210–221, oct 2009. doi: 10.1016/j.desal.2008.12.025. URL http://dx.doi.org/10.1016/j.desal.2008.12.025.

ICH guideline Q11. Development and manufacture of drug substances (chemical entities and biotechnological/biological entities), 2011.

ICH guideline Q3A. Impurities in new drug substances, 2008.

ICH guideline Q3B. Impurities in new drug products, 2006.

Imarcgroup. Global biopharmaceutical market report forecast (2012-2017), 2012. URL http://www.imarcgroup.com/biotechnology-industry/#sthash.BDqBLY4Y.dpuf.

Isaksson, T. and Næs, T. The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. *appl spectrosc*, 42(7): 1273–1284, sep 1988. doi: 10.1366/0003702884429869. URL http://dx.doi.org/10.1366/0003702884429869.

Ivarsson, M., Villiger, T., Morbidelli, M., and Soos, M. Evaluating the impact of cell culture process parameters on monoclonal antibody n-glycosylation. *Journal of biotechnology*, 2014.

Jagschies, G., Gronberg, A., Bjorkman, T., Lacki, K., and Johansson, H. Technical and economical evaluation of downstream processing options for monoclonal antibody (mab) production. 2006.

Jandera, P. and Churáček, J. Gradient elution in liquid chromatography. *Journal of Chromatography A*, 91:223–235, apr 1974. doi: 10.1016/s0021-9673(01)97902-6. URL http://dx.doi.org/10.1016/s0021-9673(01)97902-6.

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

Jansen, J., editor. *Protein Purification*. 2012.

Jayapal, K., Wlaschin, K., Hu, W.-S., and Yap, M. Recombinant protein therapeuticics from cho cells - 20 years and counting. *AIChE Chemical Engineering Progress*, 103:40–47, 2007. URL http://www.aiche.org/sites/default/files/docs/pages/CHO.pdf.

Jennings, N., Corera, J., and Laresgoiti, I. Developing industrial multi-agent systems. In *ICMAS*, pages 423–430, 1995.

Jeong, Y. and Wang, S. Role of glutamine in hybridoma cell culture: Effects on cell growth, antibody production, and cell metabolism. *Enzyme and microbial technology*, 17(1):47–55, 1995.

Jiang, W., Kim, S., Zhang, X., Lionberger, R., Davit, B., Conner, D., and Lawrence, X. The role of predictive biopharmaceutical modeling and simulation in drug development and regulatory evaluation. *International journal of pharmaceutics*, 418(2):151–160, 2011.

Johansen, T. and Foss, B. Nonlinear local model representation for adaptive systems. In *Singapore International Conference on Intelligent Control and Instrumentation [Proceedings 1992]*. Institute of Electrical & Electronics Engineers (IEEE), 1992. doi: 10.1109/sicici.1992.637617. URL http://dx.doi.org/10.1109/sicici.1992.637617.

Jolliffe, I. *Principal component analysis*. Wiley Online Library, 2002.

Jones, B. and Montgomery, D. Alternatives to resolution iv screening designs in 16 runs. *International Journal of Experimental Design and Process Optimisation*, 1(4):285–295, 2010.

Jong, S. D. Simpls: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*, 18(3):251–263, 1993.

Joreskog, K., Klovan, J., and Reyment, R. *Geological factor analysis*. Elsevier Scientific Pub. Co., 1976.

Julka, N., Srinivasan, R., and Karimi, I. Agent-based supply chain management—1: framework. *Computers & Chemical Engineering*, 26(12): 1755–1769, 2002.

Jungbauer, A. Insights into the chromatography of proteins provided by mathematical modeling. *Current Opinion in Biotechnology*, 7(2):210–218, 1996.

Kaneko, Y., Nimmerjahn, F., and Ravetch, F. Anti-inflammatory activity of immunoglobulin g resulting from fc sialylation. *Science*, 313(5787): 670–673, 2006.

Karstang, T. V. and Manne, R. Optimized scaling. *Chemometrics and Intelligent Laboratory Systems*, 14(1-3):165–173, apr 1992. doi: 10.1016/0169-7439(92)80101-9. URL http://dx.doi.org/10.1016/0169-7439(92)80101-9.

Kayser, O. and Warzecha, H. *Pharmaceutical biotechnology: drug discovery and clinical applications*. John Wiley & Sons, 2012.

Kearns, M. and Ron, D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

Keenan, M. R. and Kotula, P. G. Accounting for poisson noise in the multivariate analysis of ToF-SIMS spectrum images. *Surf. Interface Anal.*, 36(3):203–212, mar 2004. doi: 10.1002/sia.1657. URL http://dx.doi.org/10.1002/sia.1657.

Kelley, B. Industrialization of mab production technology: the bioprocessing industry at a crossroads. In *MAbs*, volume 1, pages 443–452. Taylor & Francis, 2009.

Kelley, B., Blank, G., and Lee, A. Downstream processing of monoclonal antibodies: Current practices and future opportunities. In *Process Scale Purification of Antibodies*, pages 1–23. Wiley-Blackwell, jan 2009. doi: 10.1002/9780470444894.ch1. URL http://dx.doi.org/10.1002/9780470444894.ch1.

Kettaneh, N., Berglund, A., and Wold, S. Pca and pls with very large data sets. *Computational Statistics & Data Analysis*, 48(1):69–85, 2005.

Kim, S., Kano, M., Nakagawa, H., and Hasebe, S. Input variable scaling for statistical modeling. *Computers & Chemical Engineering*, 74:59–65, mar 2015. doi: 10.1016/j.compchemeng.2014.12.016. URL http://dx.doi.org/10.1016/j.compchemeng.2014.12.016.

Kirdar, A., Green, K., and Rathore, A. Application of multivariate data analysis for identification and successful resolution of a root cause for a bioprocessing application. *Biotechnology Progress*, 24(3):720–726, 2008a. doi: 10.1021/bp0704384. URL http://dx.doi.org/10.1021/bp0704384.

Kirdar, A., Green, K., and Rathore, A. Application of multivariate data analysis for identification and successful resolution of a root cause for a bioprocessing application. *Biotechnology progress*, 24(3):720–726, 2008b.

Klatt, K., Hanisch, F., Dünnebier, G., and Engell, S. Model-based optimization and control of chromatographic processes. *Computers & Chemical Engineering*, 24(2):1119–1126, 2000.

Klimasauskas, C. Hybrid modeling for robust nonlinear multivariable control. *ISA Transactions*, 37(4):291–297, sep 1998. doi: 10.1016/s0019-0578(98)00030-5. URL http://dx.doi.org/10.1016/s0019-0578(98)00030-5.

Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

Köhler, G. and Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256:495–497, 1975. doi: 10.1038/256495a0. URL http://www.nature.com/nature/journal/v256/n5517/abs/256495a0.html.

Konno, Y., Kobayashi, Y., Takahashi, K., Takahashi, E., Sakae, S., Wakitani, M., Yamano, K., Suzawa, T., Yano, K., Ohta, T., *et al.* Fucose content of monoclonal antibodies can be controlled by culture medium osmolality for high antibody-dependent cellular cytotoxicity. *Cytotechnology*, 64(3): 249–265, 2012.

Kontoravdi, C., Wong, D., Lam, C., Lee, Y., Yap, M., Pistikopoulos, E., and Mantalaris, A. Modeling amino acid metabolism in mammalian cells toward the development of a model library. pages 1261–1269, 2007. doi: 10.1021/bp070106z. URL http://dx.doi.org/10.1021/bp070106z.

Kosanovich, K., Dahl, K., and Piovoso, M. Improved process understanding using multiway principal component analysis. *Industrial & engineering chemistry research*, 35(1):138–146, 1996.

Kourti, T. and MacGregor, J. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, 28(1):3–21, 1995. ISSN 0169-7439. doi: 10.1016/0169-7439(95)80036-9.

Krambeck, F. J. and Betenbaugh, M. J. A mathematical model of n-linked glycosylation. *Biotechnology and Bioengineering*, 92(6):711–728, 2005. doi: 10.1002/bit.20645. URL http://dx.doi.org/10.1002/bit.20645.

Kramer, M., Thompson, M., and Bhagat, P. Embedding theoretical models in neural networks. In *American Control Conference, 1992*, pages 475–479. IEEE, 1992.

Krishnan, S., Verheij, E. E. R., Bas, R. C., Hendriks, M. W. B., Hankemeier, T., Thissen, U., and Coulier, L. Pre-processing liquid chromatography/high-resolution mass spectrometry data: extracting pure mass spectra by deconvolution from the invariance of isotopic distribution. *Rapid Commun. Mass Spectrom.*, 27(9):917–923, apr 2013a. doi: 10.1002/rcm.6517. URL http://dx.doi.org/10.1002/rcm.6517.

Krishnan, S., Verheij, E., Bas, R. C., Hendriks, M., Hankemeier, T., Thissen, U., and Coulier, L. Pre-processing liquid chromatography/high-resolution mass spectrometry data: extracting pure mass spectra by deconvolution from the invariance of isotopic distribution. *Rapid Communications in Mass Spectrometry*, 27(9):917–923, 2013b.

Kukharev, G. and Kuźmiński, A. An environment for recognition systems modeling. In *Enhanced Methods in Computer Security, Biometric and Artificial Intelligence Systems*, pages 157–164. Springer Science Business Media, 2005. doi: 10.1007/0-387-23484-5_15. URL http://dx.doi.org/10.1007/0-387-23484-5_15.

Kula, M. and Schütte, H. Purification of proteins and the disruption of microbial cells. 3:31–42, 1987. doi: 10.1002/btpr.5420030107.

Kunkel, J., Jan, D., Jamieson, J., and Butler, M. Dissolved oxygen concentration in serum-free continuous culture affects n-linked glycosylation of a monoclonal antibody. *Journal of Biotechnology*, 62(1): 55–71, 1998. ISSN 0168-1656. doi: 10.1016/S0168-1656(98)00044-3.

Labrou, Y., Peng, Y., Tolone, B., and Boughannam, A. A multi-agent system for enterprise integration. *Proceedings of the 3rd International Conference on the Practical Applications of Agents and Multi-Agent Systems (PAAM-98)*, 1998.

Lai, T., Yang, Y., and Ng, S. Advances in mammalian cell line development technologies for recombinant protein production. *Pharmaceuticals*, 6(5): 573–603, 2013. doi: 10.3390/ph6050579.

Laursen, K., Frederiksen, S. S., Leuenhagen, C., and Bro, R. Chemometric quality control of chromatographic purity. *Journal of Chromatography A*, 1217(42):6503–6510, oct 2010a. doi: 10.1016/j.chroma.2010.08.040. URL `http://dx.doi.org/10.1016/j.chroma.2010.08.040`.

Laursen, K., Frederiksen, S., Leuenhagen, C., and Bro, R. Chemometric quality control of chromatographic purity. *Journal of Chromatography A*, 1217(42):6503–6510, 2010b.

Laursen, K., Rasmussen, M. A., and Bro, R. Comprehensive control charting applied to chromatography. *Chemometrics and Intelligent Laboratory Systems*, 107(1):215–225, may 2011. doi: 10.1016/j.chemolab.2011.04.002. URL `http://dx.doi.org/10.1016/j.chemolab.2011.04.002`.

Laxalde, J., Ruckebusch, C., Devos, O., Caillol, N., Wahl, F., and Duponchel, L. Characterisation of heavy oils using near-infrared spectroscopy: Optimisation of pre-processing methods and variable selection. *Analytica Chimica Acta*, 705(1-2):227–234, oct 2011. doi: 10.1016/j.aca.2011.05.048. URL `http://dx.doi.org/10.1016/j.aca.2011.05.048`.

Le, H., Kabbur, S., Pollastrini, L., Sun, Z., Mills, K., Johnson, K., Karypis, G., and Hu, W. Multivariate analysis of cell culture bioprocess data—lactate consumption as process indicator. *Journal of biotechnology*, 162(2): 210–223, 2012.

Lécorché, P., Walrant, A., Burlina, F., Dutot, L., Sagan, S., Mallet, J., Desbat, B., Chassaing, G., Alves, I., and Lavielle, S. Cellular uptake and biophysical properties of galactose and/or tryptophan containing cell-penetrating peptides. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1818(3):448–457, 2012.

Lee, D. S., Lee, M., Woo, S., Kim, Y., and Park, J. Nonlinear dynamic partial least squares modeling of a full-scale biological wastewater treatment plant. *Process Biochemistry*, 41(9):2050–2057, 2006.

Lee, D., Jeon, C., Park, J., and Chang, K. Hybrid neural network modeling of a full-scale industrial wastewater treatment process. *Biotechnol. Bioeng.*, 78 (6):670–682, jun 2002. doi: 10.1002/bit.10247.abs. URL http://dx.doi.org/10.1002/bit.10247.abs.

Lee, K. and Gilmore, D. Statistical experimental design for bioprocess modeling and optimization analysis: Repeated-measures method for dynamic biotechnology process. *Applied Biochemistry and Biotechnology*, 135(2):101–116, 2006. doi: 10.1385/abab:135:2:101. URL http://dx.doi.org/10.1385/abab:135:2:101.

Lennox, B., Montague, G., Hiden, H., Kornfeld, G., and Goulding, P. Process monitoring of an industrial fed-batch fermentation. *Biotechnology and Bioengineering*, 74(2):125–135, 2001.

Levin, A. and Narendra, K. Identification of nonlinear dynamical systems using neural networks. In *Neural Systems for Control*, pages 129–160. Elsevier BV, 1997. doi: 10.1016/b978-012526430-3/50007-6. URL http://dx.doi.org/10.1016/b978-012526430-3/50007-6.

Li, B., Morris, J., and Martin, E. Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 64:70–89, 2002. URL http://www.sciencedirect.com/science/article/pii/S0169743902000515.

Li, B., Hu, Y., Liang, Y.-Z., Xie, P.-S., and Du, Y.-P. Quality evaluation of fingerprints of herbal medicine with chromatographic data. *Analytica Chimica Acta*, 514(1):69–77, jun 2004. doi: 10.1016/s0003-2670(04)00355-1. URL http://dx.doi.org/10.1016/s0003-2670(04)00355-1.

Li, B., Shanahan, M., Calvet, A., Leister, K. J., and Ryder, A. G. Comprehensive, quantitative bioprocess productivity monitoring using fluorescence EEM spectroscopy and chemometrics. *Analyst*, 139(7):1661, 2014. doi: 10.1039/c4an00007b. URL http://dx.doi.org/10.1039/c4an00007b.

Li, F., Vijayasankaran, N., Shen, A., Kiss, R., and Amanullah, A. Cell culture processes for monoclonal antibody production. *mAbs*, 2:466–477, 2010. doi: 10.4161/mabs.2.5.12720.

Li, J., Wong, C., Vijayasankaran, N., Hudson, T., and Amanullah, A. Feeding lactate for CHO cell culture processes: Impact on culture metabolism and performance. *Biotechnology and Bioengineering*, 109(5):1173–1186, 2012. doi: 10.1002/bit.24389. URL http://dx.doi.org/10.1002/bit.24389.

Li, Z., Gu, Y., and Gu, T. Mathematical modeling and scale-up of size-exclusion chromatography. *Biochemical engineering journal*, 2(2): 145–155, 1998.

Lienqueo, M., Shene, C., and Asenjo, J. Optimization of hydrophobic interaction chromatography using a mathematical model of elution curves of a protein mixture. *Journal of Molecular Recognition*, 22(2):110–120, 2009.

Lipman, N. and Jackson, L. Hollow fibre bioreactors: an alternative to murine ascites for small scale (¡ 1 gram) monoclonal antibody production. *In vivo and in vitro production of mAbs*, pages 571–576, 1998.

Listgarten, J. and Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular Cellular Proteomics*, pages 419–434, 2005. doi: 10.1074/mcp.R500005-MCP200. URL http://www.mcponline.org/content/4/4/419.short.

Liu, H., Ma, J., Winter, C., and Bayer, R. Recovery and purification process development for monoclonal antibody production. *mAbs*, 2:480–499, 2010. doi: 10.4161/mabs.2.5.12645. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2958570/.

Liu, Y., Lyon, B., Windham, W., Realini, C., Pringle, T., and Duckett, S. Prediction of color, texture, and sensory characteristics of beef steaks by visible and near infrared reflectance spectroscopy. a feasibility study. *Meat Science*, 65(3):1107–1115, nov 2003. doi: 10.1016/s0309-1740(02)00328-5. URL http://dx.doi.org/10.1016/s0309-1740(02)00328-5.

Loukas, Y. Artificial neural networks in liquid chromatography: efficient and improved quantitative structure–retention relationship models. *Journal of Chromatography A*, 904(2):119–129, 2000.

Low, D., O'Leary, R., and Pujar, N. Future of antibody purification. *Journal of Chromatography B*, 848:48–63, 2007. doi: 10.1016/j.jchromb.2006.10.033.

Lowe, C., Lowe, A., and Gupta, G. New developments in affinity chromatography with potential application in the production of biopharmaceuticals. *Journal of Biochemical and Biophysical Methods*, 49 (1-3):561–574, oct 2001. doi: 10.1016/s0165-022x(01)00220-2. URL http://dx.doi.org/10.1016/s0165-022x(01)00220-2.

Lu, N., Gao, F., and Wang, F. Sub-PCA modeling and on-line monitoring strategy for batch processes. *AIChE Journal*, 50(1):255–259, jan 2004. doi: 10.1002/aic.10024. URL http://dx.doi.org/10.1002/aic.10024.

Luna, A., da Silva, A., Pinho, J., Ferré, J., and Boqué, R. Rapid characterization of transgenic and non-transgenic soybean oils by chemometric methods using NIR spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 100:115–119, jan 2013. doi: 10.1016/j.saa.2012.02.085. URL http://dx.doi.org/10.1016/j.saa.2012.02.085.

Luna, A., da Silva, A., Pinho, J., Ferré, J., and Boqué, R. A novel approach to discriminate transgenic from non-transgenic soybean oil using FT-MIR and chemometrics. *Food Research International*, 67:206–211, jan 2015. doi: 10.1016/j.foodres.2014.11.011. URL http://dx.doi.org/10.1016/j.foodres.2014.11.011.

Luypaert, J., Heuerding, S., Heyden, Y. V., and Massart, D. The effect of preprocessing methods in reducing interfering variability from near-infrared measurements of creams. *Journal of Pharmaceutical and Biomedical Analysis*, 36(3):495–503, nov 2004. doi: 10.1016/j.jpba.2004.06.023. URL http://dx.doi.org/10.1016/j.jpba.2004.06.023.

Ma, C., Wang, H., Lu, X., Xu, G., and Liu, B. Metabolic fingerprinting investigation of artemisia annua l. in different stages of development by gas chromatography and gas chromatography–mass spectrometry. *Journal of Chromatography A*, 1186(1-2):412–419, apr 2008. doi: 10.1016/j.chroma.2007.09.023. URL http://dx.doi.org/10.1016/j.chroma.2007.09.023.

Madden, J. E., S., M. J., Dicinoski, G. W., Avdalovic, N., and Haddad, P. R. Simulation and optimization of retention in ion chromatography using virtual column 2 software. *Analytical Chemistry*, 74(23):6023–6030, dec 2002. doi: 10.1021/ac020280w. URL http://dx.doi.org/10.1021/ac020280w.

Maitra, S. and Yan, J. Principle component analysis and partial least squares: two dimension reduction techniques for regression, 2008.

Maleki, M., Mouazen, A., Ramon, H., and Baerdemaeker, J. D. Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. *Biosystems Engineering*, 96(3):427–433, mar 2007. doi: 10.1016/j.biosystemseng.2006.11.014. URL http://dx.doi.org/10.1016/j.biosystemseng.2006.11.014.

Mandenius, C. and Brundin, A. Bioprocess optimization using design-of-experiments methodology. *Biotechnol Progress*, 24(6): 1191–1203, nov 2008. doi: 10.1002/btpr.67. URL http://dx.doi.org/10.1002/btpr.67.

Martens, H. Reliable and relevant modelling of real world data: a personal account of the development of pls regression. *Chemometrics and intelligent laboratory systems*, 58(2):85–95, 2001.

Martinez, E. and Wilson, J. A hybrid neural network-first principles approach to batch unit optimisation. *Computers & Chemical Engineering*, 22: S893–S896, mar 1998. doi: 10.1016/s0098-1354(98)00174-4. URL http://dx.doi.org/10.1016/s0098-1354(98)00174-4.

MathWorks. trapz - trapezoidal numerical integration, July 2015. URL http://uk.mathworks.com/help/matlab/ref/trapz.html.

McCracken, N., Kowle, R., and Ouyang, A. Control of galactosylated glycoforms distribution in cell culture system. *Biotechnology Progress*, 30 (3):547–553, 2014. ISSN 1520-6033. doi: 10.1002/btpr.1906.

McGuffin, V. L. and Chen, S. Theoretical and experimental studies of the effect of pressure on solute retention in liquid chromatography. *Anal. Chem.*, 69(5):930–943, mar 1997. doi: 10.1021/ac960589d. URL http://dx.doi.org/10.1021/ac960589d.

Mehta, A., Tse, M., Fogle, J., Len, A., Shrestha, R., Fontes, N., Lebreton, B., Wolk, B., and Van Reis, R. Purifying therapeutic monoclonal antibodies. *Chemical Engineering Progress*, 104(5):S14, 2008.

Meireles, M., Lavoute, E., and Bacchin, P. Filtration of a bacterial fermentation broth: harvest conditions effects on cake hydraulic resistance. *Bioprocess and biosystems engineering*, 25(5):309–314, 2003.

Mercier, S., Diepenbroek, B., Wijffels, R., and Streefland, M. Multivariate pat solutions for biopharmaceutical cultivation: current progress and limitations. *Trends in biotechnology*, 32(6):329–336, 2014. ISSN 0167-7799. doi: 10.1016/j.tibtech.2014.03.008.

Mercier, S., Diepenbroek, B., Dalm, M., Wijffels, R., and Streefland, M. Multivariate data analysis as a PAT tool for early bioprocess development data. *Journal of biotechnology*, 167:262–270, 2013. doi: 10.1016/j.jbiotec.2013.07.006. URL `http://dx.doi.org/10.1016/j.jbiotec.2013.07.006`.

Meuwly, F., Weber, U., Ziegler, T., Gervais, A., Mastrangeli, R., Crisci, C., Rossi, M., Bernard, A., Von Stockar, U., and Kadouri, A. Conversion of a cho cell culture process from perfusion to fed-batch technology without altering product quality. *Journal of biotechnology*, 123(1):106–116, 2006.

Meuwly, F., Ruffieux, P., and Kadouri, A. Packed-bed bioreactors for mammalian cell culture: bioprocess and biomedical applications. *Biotechnology Advances*, 25:45–56, 2007.

Meyer, V. R. *Practical High-Performance Liquid Chromatography*. John Wiley & Sons, Ltd, apr 2010. doi: 10.1002/9780470688427. URL `http://dx.doi.org/10.1002/9780470688427`.

Milavec, P., Podgornik, A., Štravs, R., and Koloini, T. Effect of experimental error on the efficiency of different optimization methods for bioprocess media optimization. *Bioprocess and Biosystems Engineering*, 25(2):69–78, jun 2002. doi: 10.1007/s00449-002-0285-x. URL `http://dx.doi.org/10.1007/s00449-002-0285-x`.

Milo, R. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, oct 2002. doi: 10.1126/science.298.5594.824. URL `http://dx.doi.org/10.1126/science.298.5594.824`.

Minow, B., Rogge, P., and Thompson, K. Implementing a fully disposable mab manufacturing facility. *BioProcess International*, 10:48–57, 2012.

Moco, S., Forshed, J., De Vos, R., Bino, R., and Vervoort, J. Intra- and inter-metabolite correlation spectroscopy of tomato metabolomics data obtained by liquid chromatography-mass spectrometry and nuclear magnetic resonance. *Metabolomics*, 4(3):202–215, may 2008. doi: 10.1007/s11306-008-0112-8. URL `http://dx.doi.org/10.1007/s11306-008-0112-8`.

Mogk, G., Mrziglod, T., and Schuppert, A. Application of hybrid models in chemical industry. In *European Symposium on Computer Aided Process Engineering-12, 35th European Symposium of the Working Party on Computer Aided Process Engineering*, pages 931–936. Elsevier BV, 2002. doi: 10.1016/s1570-7946(02)80183-3. URL http://dx.doi.org/10.1016/s1570-7946(02)80183-3.

Monks, K., Molnár, I., Rieger, H.-J., Bogáti, B., and Szabó, E. Quality by design: Multidimensional exploration of the design space in high performance liquid chromatography method development for better robustness before validation. *Journal of Chromatography A*, 1232:218–230, apr 2012. doi: 10.1016/j.chroma.2011.12.041. URL http://dx.doi.org/10.1016/j.chroma.2011.12.041.

Monod, J. The growth of bacterial cultures. *Annual Review of Microbiology*, 3 (1):371–394, oct 1949. doi: 10.1146/annurev.mi.03.100149.002103. URL http://dx.doi.org/10.1146/annurev.mi.03.100149.002103.

Moreira, J., Feliciano, A., Santana, P., Cruz, P., Aunins, J., and Carrondo, M. Repeated-batch cultures of baby hamster kidney cell aggregates in stirred vessels. *Cytotechnology*, 15(13):337–349, 1994. ISSN 0920-9069. doi: 10.1007/BF00762409.

Morris, M., Saez-Rodriguez, J., Sorger, P., and Lauffenburger, D. Logic-based models for the analysis of cell signaling networks. *Biochemistry*, 49(15): 3216–3224, apr 2010. doi: 10.1021/bi902202q. URL http://dx.doi.org/10.1021/bi902202q.

Naciri, M., Kuystermans, D., and Al-Rubeai, M. Monitoring ph and dissolved oxygen in mammalian cell culture using optical sensors. *Cytotechnology*, 57 (3):245–250, 2008.

Naderi, S., Meshram, M., Wei, C., McConkey, B., Ingalls, B., Budman, H., and Scharer, J. Development of a mathematical model for evaluating the dynamics of normal and apoptotic chinese hamster ovary cells. *Biotechnology Progress*, pages 1197–1205, 2011. doi: 10.1002/btpr.647. URL http://dx.doi.org/10.1002/btpr.647.

Næs, T. and Mevik, B. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, 15(4):413–426, 2001.

Næs, T. Leverage and influence measures for principal component regression. *Chemometrics and Intelligent Laboratory Systems*, 5(2):155–168, 1989. ISSN 0169-7439.

Nelson, P., Taylor, P., and MacGregor, J. Missing data methods in pca and pls: Score calculations with incomplete observations. *Chemometrics and intelligent laboratory systems*, 35(1):45–65, 1996.

Neuman, S. Maximum likelihood bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment*, 17 (5):291–305, 2003.

Nfor, B., Verhaert, P., Van der Wielen, L., Hubbuch, J., and Ottens, M. Rational and systematic protein purification process development: the next generation. *Trends in Biotechnology*, 27(12):673–679, dec 2009. doi: 10.1016/j.tibtech.2009.09.002. URL http://dx.doi.org/10.1016/j.tibtech.2009.09.002.

Nielsen, N.-. V., Carstensen, J. M., and Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805(1-2):17–35, may 1998. doi: 10.1016/s0021-9673(98)00021-1. URL http://dx.doi.org/10.1016/s0021-9673(98)00021-1.

Nolan, R. and Lee, K. Dynamic model of CHO cell metabolism. *Metabolic engineering*, 13(1):108–124, 2011. doi: 10.1016/j.ymben.2010.09.003. URL http://dx.doi.org/10.1016/j.ymben.2010.09.003.

Nomikos, P. and MacGregor, J. Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems*, 30: 1361–1375, 1995a. ISSN 0169-7439. doi: 10.1016/0169-7439(95)00043-7.

Nomikos, P. and MacGregor, J. Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40:97–108, 1994. doi: 10.1002/aic.690400809. URL http://dx.doi.org/10.1002/aic.690400809.

Nomikos, P. and MacGregor, J. Multi-way partial least squares in monitoring batch processes. *Chemometrics and intelligent laboratory systems*, 30(1): 97–108, 1995b.

Northern Arizona Univeristy. Liquid chromatography. URL http://jan.ucc.nau.edu/~jkn/235A-Appendix.htm.

Nwana, H. Software agents: An overview. *The knowledge engineering review*, 11(03):205–244, 1996.

Nyberg, G., Balcarcel, R., Follstad, B., Stephanopoulos, G., and Wang, D. Metabolic effects on recombinant interferon-γ glycosylation in continuous culture of chinese hamster ovary cells. *Biotechnology and bioengineering*, 62(3):336–347, 1999.

Ödman, P., Johansen, C., Olsson, L., Gernaey, K., and Lantz, A. Sensor combination and chemometric variable selection for online monitoring of streptomyces coelicolor fed-batch cultivations. *Applied microbiology and biotechnology*, 86(6):1745–1759, 2010.

Oliveira, R. Combining first principles modelling and artificial neural networks: a general framework. *Computers & Chemical Engineering*, 28 (5):755–766, 2004.

Orellana, C., Shene, C., and Asenjo, J. Mathematical modeling of elution curves for a protein mixture in ion exchange chromatography applied to high protein concentration. *Biotechnology and bioengineering*, 104(3): 572–581, 2009.

Ozturk, S. and Hu, W.-S. *Cell culture technology for pharmaceutical and cell-based therapies*. Taylor and Francis, 2006.

Ozturk, S. and Palsson, B. Growth, metabolic, and antibody production kinetics of hybridoma cell culture: 1. analysis of data from controlled batch reactors. *Biotechnology Progress*, 7(6):471–480, 1991. doi: 10.1021/bp00012a001. URL http://dx.doi.org/10.1021/bp00012a001.

Ozturk, S., Riley, M., and Palsson, B. Effects of ammonia and lactate on hybridoma growth, metabolism, and antibody production. *Biotechnology and Bioengineering*, 39(4):418–431, 1992. doi: 10.1002/bit.260390408. URL http://dx.doi.org/10.1002/bit.260390408.

Pacis, E., Yu, M., Autsen, J., Bayer, R., and Li, F. Effects of cell culture conditions on antibody nlinked glycosylation—what affects high mannose 5 glycoform. *Biotechnology and Bioengineering*, 2011a. ISSN 1097-0290. doi: 10.1002/bit.23200.

Pacis, E., Yu, M., Autsen, J., Bayer, R., and Li, F. Effects of cell culture conditions on antibody n-linked glycosylation—what affects high mannose 5 glycoform. *Biotechnology and bioengineering*, 108(10):2348–2358, 2011b.

Pan, W., Xiao, G., and Huang, X. Using input dependent weights for model combination and model selection with multiple sources of data. *Statistica Sinica*, 16:523–540, 2006.

Piazza, G. and Garcia, R. Proteins and peptides as renewable flocculants. *Bioresource technology*, 101(15):5759–5766, 2010.

Potočnik, P. and Grabec, I. Empirical modeling of antibiotic fermentation process using neural networks and genetic algorithms. *Mathematics and Computers in Simulation*, 49(4-5):363–379, sep 1999. doi: 10.1016/s0378-4754(99)00045-2. URL http://dx.doi.org/10.1016/s0378-4754(99)00045-2.

Pravdova, V., Walczak, B., and Massart, D. A comparison of two algorithms for warping of analytical signals. *Analytica Chimica Acta*, 456(1):77–92, apr 2002. doi: 10.1016/s0003-2670(02)00008-9. URL http://dx.doi.org/10.1016/s0003-2670(02)00008-9.

Preedy, V. R., Srirajaskanthan, R., and Patel, V. B. *Handbook of Food Fortification and Health: From Concepts to Public Health Applications*, volume 1. Springer Science Business Media, july 2013.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 2007. ISBN 9780521880688. URL http://books.google.co.uk/books?id=1aAOdzK3FegC.

Psichogios, C. and Ungar, L. A hybrid neural network-first principles approach to process modelling. *AIChE Journal*, 38(10):1499–1511, oct 1992. doi: 10.1002/aic.690381003. URL http://dx.doi.org/10.1002/aic.690381003.

Qi, M. and Zhang, G. An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research*, 132(3):666–680, 2001.

Qin, S. and Dunia, R. Determining the number of principal components for best reconstruction. *Journal of Process Control*, 10(2):245–250, 2000.

Rajalahti, T. and Kvalheim, O. Multivariate data analysis in pharmaceutics: a tutorial review. *International journal of pharmaceutics*, 417:280–290, 2011a. ISSN 0378-5173. doi: 10.1016/j.ijpharm.2011.02.019.

Rajalahti, T. and Kvalheim, O. M. Multivariate data analysis in pharmaceutics: A tutorial review. *International Journal of Pharmaceutics*, 417(1-2): 280–290, sep 2011b. doi: 10.1016/j.ijpharm.2011.02.019. URL http://dx.doi.org/10.1016/j.ijpharm.2011.02.019.

Raju, T., Briggs, J., Chamow, S., Winkler, M., and Jones, A. Glycoengineering of therapeutic glycoproteins: in vitro galactosylation and sialylation of glycoproteins with terminal n-acetylglucosamine and galactose residues. *Biochemistry*, 40(30):8868–8876, 2001.

Ramadan, Z., Hopke, P., Johnson, M., and Scow, K. Application of pls and back-propagation neural networks for the estimation of soil properties. *Chemometrics and intelligent laboratory systems*, 75(1):23–30, 2005.

Rathore, A. Roadmap for implementation of quality by design (QbD) for biotechnology products. *Trends in Biotechnology*, 27(9):546–553, sep 2009. doi: 10.1016/j.tibtech.2009.06.006. URL http://dx.doi.org/10.1016/j.tibtech.2009.06.006.

Rathore, A., Weiskopf, A., and Reason, A. Defining critical quality attributes for monoclonal antibody therapeutic products. *BioPharm International*, 27 (7):34–43, 2014a. URL http://www.biopharminternational.com/biopharm/Manufacturing/Defining-Critical-Quality-Attributes-for-mAb-Thera/ArticleStandard/Article/detail/848488?contextCategoryId=43048.

Rathore, A., Weiskopf, A., and Reason, A. Defining critical quality attributes for monoclonal antibody therapeutic products. *BioPharm International*, 27 (7):34–43, 2014b.

Reed, J., Devlin, D., Esteves, S., Dinsdale, R., and Guwy, A. Performance parameter prediction for sewage sludge digesters using reflectance FT-NIR spectroscopy. *Water Research*, 45(8):2463–2472, apr 2011. doi: 10.1016/j.watres.2011.01.027. URL http://dx.doi.org/10.1016/j.watres.2011.01.027.

Reichert, J. M. Marketed therapeutic antibodies compendium. *MAbs*, 4(3): 413–415, 2012. doi: 10.4161/mabs.19931. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3355480/.

Reichert, J. M. Therapeutic monoclonal antibodies approved or in review in the european union or united states, 2015. URL http://www.antibodysociety.org/news/approved_mabs.php.

Reid, R. Putting humpty dumpty together again. In *Evolutionary Theory*, pages 338–361. Springer Science Business Media, 1985. doi: 10.1007/978-1-4615-9787-2_18. URL http://dx.doi.org/10.1007/978-1-4615-9787-2_18.

Rhee, J. and Kang, T. On-line process monitoring and chemometric modeling with 2d fluorescence spectra obtained in recombinant e. coli fermentations. *Process Biochemistry*, 42(7):1124–1134, 2007.

Rinnan, A., Berg, F., and Engelsen, S. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, 28:1201–1222, 2009. doi: 10.1016/j.trac.2009.07.007.

Rito-Palomares, M. Bioseparation: The limiting step in bioprocess development. *J. Chem. Technol. Biotechnol.*, 83(2):115–116, 2008. doi: 10.1002/jctb.1886. URL http://dx.doi.org/10.1002/jctb.1886.

Rokach, L. and Maimon, O. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

Rubinson, K. A. and Rubinson, J. F. *Contemporary Instrumental Analysis*, volume 40. Prentice-Hall, may 2000.

Sadava, D., Orians, G., Heller, C., and Purves, W. *Life: The Science of Biology*. Macmillan Higher Education, 2007. ISBN 9781429208840. URL http://books.google.com.au/books?id=J-sTOAAACAAJ.

Sakoe, H. and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Sanderson, C., Barton, G., and Barford, J. Optimisation of animal cell culture media using dynamic simulation. *Computers & Chemical Engineering*, 19: 681–686, jun 1995. doi: 10.1016/0098-1354(95)87114-4. URL http://dx.doi.org/10.1016/0098-1354(95)87114-4.

Saraswat, M., Musante, L., Ravidá, A., Shortt, B., Byrne, B., and Holthofer, H. Preparative purification of recombinant proteins: Current status and future trends. *BioMed Research International*, 2013, 2013. ISSN 2314-6133. doi: 10.1155/2013/312709.

Sasorith, S. and Lefranc, M.-P. Glycosylation, 04 2004. URL http://www.imgt.org/IMGTeducation/IMGTlexique/G/Glycosylation.html.

Savitzky, A. and Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8): 1627–1639, 1964. doi: 10.1021/ac60214a047.

Schenker, B. and Agarwal, M. Online-optimized feed switching in semi-batch reactors using semi-empirical dynamic models. *Control Engineering Practice*, 8(12):1393–1403, dec 2000. doi: 10.1016/s0967-0661(00)00077-0. URL http://dx.doi.org/10.1016/s0967-0661(00)00077-0.

Schilling, C., Edwards, J., and Palsson, B. Toward metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnol. Prog.*, 15(3): 288–295, jun 1999. doi: 10.1021/bp9900357. URL http://dx.doi.org/10.1021/bp9900357.

Segovia-Juarez, J., Ganguli, S., and Kirschner, D. Identifying control mechanisms of granuloma formation during m. tuberculosis infection using an agent-based model. *Journal of Theoretical Biology*, 231(3):357–376, dec 2004. doi: 10.1016/j.jtbi.2004.06.031. URL http://dx.doi.org/10.1016/j.jtbi.2004.06.031.

Selvarasu, S., Kim, D., Karimi, I., and Lee, D. Combined data preprocessing and multivariate statistical analysis characterizes fed-batch culture of mouse hybridoma cells for rational medium design. *Journal of biotechnology*, 150 (1):94–100, 2010.

Selvarasu, S., Ho, Y., Chong, W., Wong, N., Yusufi, F., Lee, Y., Yap, M., and Lee, D. Combined in silico modeling and metabolomics analysis to characterize fed-batch cho cell culture. *Biotechnology and bioengineering*, 109(6):1415–1429, 2012.

SeparationsNow.com. A short primer on chemometrics for spectroscopists, 2014. URL http://www.separationsnow.com/details/education/sepspec10349education/A-Short-Primer-on-Chemometrics-for-Spectroscopists.html?&tzcheck=1.

Shaw, P., Buslig, B., and Moshonas, M. Classification of commercial orange juice types by pattern recognition involving volatile constituents quantified by gas chromatography. *J. Agric. Food Chem.*, 41(5):809–813, may 1993. doi: 10.1021/jf00029a025. URL http://dx.doi.org/10.1021/jf00029a025.

Shellie, R. A., Ng, B. K., Dicinoski, G. W., Poynter, S. D. H., O'Reilly, J. W., Pohl, C. A., and Haddad, P. R. Prediction of analyte retention for ion chromatography separations performed using elution profiles comprising multiple isocratic and gradient steps. *Analytical Chemistry*, 80(7): 2474–2482, apr 2008. doi: 10.1021/ac702275n. URL http://dx.doi.org/10.1021/ac702275n.

Shene, C., Lucero, A., Andrews, B., and Asenjo, J. Mathematical modeling of elution curves for a protein mixture in ion exchange chromatography and for the optimal selection of operational conditions. *Biotechnology and bioengineering*, 95(4):704–713, 2006.

Shoji, N., Matsui, A., Omote, M., Kuriyama, N., Blödorn, B., Kune, D., and White, C. A. Facing the challenges in bio-pharmaceutical production. Article, Nov 2009.

Shukla, A. and Thömmes, J. Recent advances in large-scale production of monoclonal antibodies and related proteins. *Trends in Biotechnology*, 28(5), 2010a. ISSN 0167-7799. doi: 10.1016/j.tibtech.2010.02.001.

Shukla, A. A. and Thömmes, J. Recent advances in large-scale production of monoclonal antibodies and related proteins. *Trends in Biotechnology*, 28(5): 253–261, may 2010b. doi: 10.1016/j.tibtech.2010.02.001. URL http://dx.doi.org/10.1016/j.tibtech.2010.02.001.

Shukla, A., Hubbard, B., Tressel, T., Guhan, S., and Low, D. Downstream processing of monoclonal antibodies—application of platform approaches. *Journal of Chromatography B*, 848(1):28–39, 2007.

Shuler, M. and Kargi, F. *Bioprocess Engineering: Basic Concepts*. Pearson new international edition. Pearson Education, Limited, 2010. ISBN 9781292025995. URL https://books.google.co.uk/books?id=aAMvnwEACAAJ.

Simutis, R., Oliveira, R., Manikowski, M., Feyo de Azevedo, S., and Hübbert, A. How to increase the performance of models for process optimization and control. *Journal of Biotechnology*, 59(1-2):73–89, dec 1997. doi: 10.1016/s0168-1656(97)00166-1. URL http://dx.doi.org/10.1016/s0168-1656(97)00166-1.

Singhania, R., Sramkoski, R., Jacobberger, J., Tyson, J., and Beard, D. A hybrid model of mammalian cell cycle regulation. *PLoS Comput Biol*, 7(2), 2011.

Skov, T. and Bro, R. Solving fundamental problems in chromatographic analysis. *Analytical and Bioanalytical Chemistry*, 390(1):281–285, 2008.

Skov, T., Van den Berg, F., Tomasi, G., and Bro, R. Automated alignment of chromatographic data. *Journal of Chemometrics*, 20(11-12):484–497, Nov 2006. doi: 10.1002/cem.1031. URL http://dx.doi.org/10.1002/cem.1031.

Skov, T. and Bro, R. Solving fundamental problems in chromatographic analysis. *Analytical and Bioanalytical Chemistry*, 390(1):281–285, oct 2007. doi: 10.1007/s00216-007-1618-z. URL http://dx.doi.org/10.1007/s00216-007-1618-z.

Sofer, G. and Hagel, L. *Handbook of Process Chromatography*. Academic Press, 1st edition, 1997. ISBN 0-12-654266-X.

Sohlberg, B. Hybrid grey box modelling of a pickling process. *Control Engineering Practice*, 13(9):1093–1102, sep 2005. doi: 10.1016/j.conengprac.2004.11.005. URL http://dx.doi.org/10.1016/j.conengprac.2004.11.005.

Soler, J., Julian, V., Rebollo, M., Carrascosa, C., and Botti, V. Towards a real-time mas architecture. *Proceedings of Challenges in Open Agent Systems, AAMAS*, 2, 2002.

Spearman, M., Rodrigues, J., Huzel, N., Sunley, K., and Butler, M. Effect of culture conditions on glycosylation of recombinant beta interferon in cho cells. *Cell technology for cell products*, 3:71–85, 2007.

Spik, G., Coddeville, B., Mazurier, J., Bourne, Y., Cambillaut, C., and Montreuil, J. *Primary and Three-Dimensional Structure of Lactotransferrin (Lactoferrin) Glycans*. Springer Science Business Media, 1994. doi: 10.1007/978-1-4615-2548-6_3. URL http://dx.doi.org/10.1007/978-1-4615-2548-6_3.

Sreedhara, A., Flengsrud, R., Prakash, V., Krowarsch, D., Langsrud, T., Kaul, P., Devold, T., and Vegarud, G. A comparison of effects of ph on the thermal stability and conformation of caprine and bovine lactoferrin. *International dairy journal*, 20(7):487–494, 2010.

Steijns, J. and Van Hooijdonk, A. Occurrence, structure, biochemical properties and technological characteristics of lactoferrin. *British Journal of Nutrition*, 84(S1):11–17, 2000.

Stein, A. and Kiesewetter, A. Cation exchange chromatography in antibody purification: pH screening for optimised binding and HCP removal. *Journal of Chromatography B*, 848(1):151–158, mar 2007. doi: 10.1016/j.jchromb.2006.10.010. URL http://dx.doi.org/10.1016/j.jchromb.2006.10.010.

Stordrange, L., Libnau, F. O., Malthe-Sørenssen, D., and Kvalheim, O. M. Feasibility study of NIR for surveillance of a pharmaceutical process, including a study of different preprocessing techniques. *Journal of Chemometrics*, 16(8-10):529–541, 2002. doi: 10.1002/cem.754. URL http://dx.doi.org/10.1002/cem.754.

Strandberg, L., Köhler, K., and Enfors, S. Large-scale fermentation and purification of a recombinant protein from escherichia coli. *Process Biochemistry*, 26(4):225–234, aug 1991. doi: 10.1016/0032-9592(91)85004-8. URL http://dx.doi.org/10.1016/0032-9592(91)85004-8.

Su, H. and McAvoy, T. Integration of multilayer perceptron networks and linear dynamic models: a hammerstein modeling approach. *Industrial & Engineering Chemistry Research*, 32(9):1927–1936, sep 1993. doi: 10.1021/ie00021a017. URL http://dx.doi.org/10.1021/ie00021a017.

Suhr, D. Principal component analysis vs. exploratory factor analysis. *SUGI 30 Proceedings*, pages 203–230, 2005.

Susanto, A., Knieps-Grünhagen, E., Von Lieres, E., and Hubbuch, J. High throughput screening for the design and optimization of chromatographic processes: assessment of model parameter determination from high throughput compatible data. *Chemical engineering & technology*, 31(12): 1846–1855, 2008.

Suzuki, E. and Ollis, D. Cell cycle model for antibody production kinetics. *Biotechnology and Bioengineering*, 34:1398–1402, 1989. ISSN 1097-0290. doi: 10.1002/bit.260341109.

Sycara, K. Multiagent systems. *AI magazine*, 19(2):79, 1998.

Tait, A., Aucamp, J., Bugeon, A., and Hoare, M. Ultra scale-down prediction using microwell technology of the industrial scale clarification characteristics by centrifugation of mammalian cell broths. *Biotechnology and Bioengineering*, 104(2):321–331, 2009. ISSN 1097-0290. doi: 10.1002/bit.22393. URL http://dx.doi.org/10.1002/bit.22393.

Tamhane, A. C. *Statistical Analysis of Designed Experiments*. Wiley-Blackwell, mar 2009. doi: 10.1002/9781118491621. URL http://dx.doi.org/10.1002/9781118491621.

Tang, Y., Ohashi, R., and Hamel, J. Perfusion culture of hybridoma cells for hyperproduction of igg2a monoclonal antibody in a wave bioreactorperfusion culture system. *Biotechnology Progress*, 2007. ISSN 1520-6033. doi: 10.1021/bp060299a.

Taschwer, M., Hackl, M., Hernández Bort, J., Leitner, C., Kumar, N., Puc, U., Grass, J., Papst, M., Kunert, R., Altmann, F., *et al.* Growth, productivity and protein glycosylation in a cho epofc producer cell line adapted to glutamine-free growth. *Journal of biotechnology*, 157(2):295–303, 2012.

Teixeira, A., Oliveira, R., Alves, P., and Carrondo, M. Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. *Biotechnology advances*, 27(6):726–732, 2009a. doi: 10.1016/j.biotechadv.2009.05.003. URL http://dx.doi.org/10.1016/j.biotechadv.2009.05.003.

Teixeira, A., Oliveira, R., Alves, P., and Carrondo, M. Advances in on-line monitoring and control of mammalian cell cultures: supporting the pat initiative. *Biotechnology advances*, 27(6):726–732, 2009b.

Thompson, M. and Kramer, M. Modeling chemical processes using prior knowledge and neural networks. *AIChE Journal*, 40(8):1328–1340, 1994.

Tomasi, G., Savorani, F., and Engelsen, S. B. icoshift: An effective tool for the alignment of chromatographic data. *Journal of Chromatography A*, 1218 (43):7832–7840, oct 2011. doi: 10.1016/j.chroma.2011.08.086. URL http://dx.doi.org/10.1016/j.chroma.2011.08.086.

Tomlin, C. and Axelrod, J. Biology by numbers: mathematical modelling in developmental biology. *Nat Rev Genet*, 8(5):331–340, may 2007. doi: 10.1038/nrg2098. URL http://dx.doi.org/10.1038/nrg2098.

Trummer, E., Fauland, K., Seidinger, S., Schriebl, K., Lattenmayer, C., Kunert, R., Vorauer, K., Weik, R., Borth, N., Katinger, H., and Muller, D. Process parameter shifting: Part i. effect of DOT, pH, and temperature on the performance of EpoFc expressing CHO cells cultivated in controlled batch bioreactors. *Biotechnology and Bioengineering*, pages 1033–1044, 2006. doi: 10.1002/bit.21013. URL http://dx.doi.org/10.1002/bit.21013.

Tsang, V., Wang, A., Yusuf-Makagiansar, H., and Ryll, T. Development of a scale down cell culture model using multivariate analysis as a qualification tool. *Biotechnology progress*, 30(1):152–160, 2014.

Tsen, A., Jang, S., Wong, D., and Joseph, B. Predictive control of quality in batch polymerization using hybrid ANN models. *AIChE Journal*, 42(2): 455–465, feb 1996. doi: 10.1002/aic.690420215. URL http://dx.doi.org/10.1002/aic.690420215.

Tucker, L. The extension of factor analysis to three-dimensional matrices. *Contributions to mathematical psychology*, pages 109–127, 1964.

Turing, A. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 237(641):37–72, 1952.

Ündey, C., Ertunç, S., Mistretta, T., and Looze, B. Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real time monitoring and control. *Journal of Process Control*, 20: 1009–1018, 2010. doi: 10.1016/j.jprocont.2010.05.008. URL http://www.ncbi.nlm.nih.gov/pubmed/24448474.

United States Food and Drug Administration (FDA). Guidance for industry: PAT — a framework for innovative pharmaceutical development, manufacturing, and quality assurance, 2004a.

United States Food and Drug Administration (FDA). Pharmaceutical cGMPs for the 21st century - a risk-based approach, 2004b.

Urmann, M., Graalfs, H., Joehnck, M., Jacob, L., and Frech, C. Cation-exchange chromatography of monoclonal antibodies: Characterization of a novel stationary phase designed for production-scale purification. *mAbs*, 2(4):395–404, 2010.

Van Berkel, P., Welling, M., Geerts, M., Van Veen, H., Ravensbergen, B., Salaheddine, M., Pauwels, E., Pieper, F., Nuijens, J., and Nibbering, P. Large scale production of recombinant human lactoferrin in the milk of transgenic cows. *Nat. Biotechnol.*, 20(5):484–487, may 2002. doi: 10.1038/nbt0502-484. URL http://dx.doi.org/10.1038/nbt0502-484.

Velayudhan, A. and Menon, M. Modeling of purification operations in biotechnology: Enabling process development, optimization, and scale-up. *Biotechnol. Prog.*, 23(1):68–73, feb 2007. doi: 10.1021/bp060378m. URL http://dx.doi.org/10.1021/bp060378m.

Vliegenthart, J. and Casset, F. Novel forms of protein glycosylation. *Current opinion in structural biology*, 8(5):565–571, 1998.

Von Lieres, E. and Andersson, J. A fast and accurate solver for the general rate model of column liquid chromatography. *Computers & chemical engineering*, 34(8):1180–1191, 2010.

von Stosch, M., Oliveira, R., Peres, J., and De Azevedo, S. A novel identification method for hybrid (n)PLS dynamical systems with application to bioprocesses. *Expert Systems with Applications*, 38(9):10862–10874, sep 2011. doi: 10.1016/j.eswa.2011.02.117. URL http://dx.doi.org/10.1016/j.eswa.2011.02.117.

von Stosch, M., Oliveira, R., Peres, J., and De Azevedo, S. A general hybrid semi-parametric process control framework. *Journal of Process Control*, 22 (7):1171–1181, 2012.

von Stosch, M., Davy, S., Francois, K., Galvanauskas, V., Hamelink, J., Luebbert, A., Mayer, M., Oliveira, R., O'Kennedy, R., Rice, P., and Glassey, J. Hybrid modeling for quality by design and PAT-benefits and challenges of applications in biopharmaceutical industry. *Biotechnology Journal*, 9(6): 719–726, 2014a. doi: 10.1002/biot.201300385. URL `http://dx.doi.org/10.1002/biot.201300385`.

von Stosch, M., Oliveira, R., Peres, J., and De Azevedo, S. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers & Chemical Engineering*, 60:86–101, 2014b.

Wahba, G. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 364–372, 1978.

Wahle, J. and Schreckenberg, M. A multi-agent system for on-line simulations based on real-world traffic data. In *Proceedings of the 34th Annual Hawaii International Conference on System Science (HICSS) (IEEE Computer Society 2001)*. IEEE, 2001.

Wall, M., Rechtsteiner, A., and Rocha, L. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

Wang, F. and Du, T. Using principal component analysis in process performance for multivariate data. *Omega*, 28(2):185–194, 2000.

Wang, G., Zhang, W., Jacklin, C., Freedman, D., Eppstein, L., and Kadouri, A. Modified celligen-packed bed bioreactors for hybridoma cell cultures. *Cytotechnology*, 9:41–49, 1992. ISSN 0920-9069. doi: 10.1007/BF02521730.

Wang, Y. and Kowalski, B. Calibration transfer and measurement stability of near-infrared spectrometers. *Applied Spectroscopy*, 46(5):764–771, may 1992. doi: 10.1366/0003702924124808. URL `http://dx.doi.org/10.1366/0003702924124808`.

Warnes, M., Glassey, J., Montague, G., and Kara, B. On data-based modelling techniques for fermentation processes. *Process Biochemistry*, 31(2): 147–155, 1996.

Weis, M., Avva, J., Jacobberger, J., and Sreenath, S. A data-driven, mathematical model of mammalian cell cycle regulation. *PLoS ONE*, 9(5): e97130, may 2014. doi: 10.1371/journal.pone.0097130. URL `http://dx.doi.org/10.1371/journal.pone.0097130`.

Westerhuis, J., Kourti, T., and MacGregor, J. Analysis of multiblock and hierarchical pca and pls models. *Journal of chemometrics*, 12(5):301–321, 1998.

Wheelwright, S. *Protein purification: design and scale up of downstream processing*. Hanser New York, 1991.

Wold, S. Spline functions in data analysis. *Technometrics*, 16:1–11, 1974.

Wold, S. Cross-validatory estimation of the number of components in factor and principal components model. *Technometrics*, 20(4):397–405, 1978a.

Wold, S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405, 1978b.

Wold, S. and Sjöström, M. Chemometrics, present and future success. *Chemometrics and Intelligent Laboratory Systems*, 44(1-2):3–14, dec 1998. doi: 10.1016/s0169-7439(98)00075-6. URL `http://dx.doi.org/10.1016/s0169-7439(98)00075-6`.

Wold, S., Geladi, P., Esbensen, K., and Öhman, J. Multi-way principal components and pls-analysis. *Journal of Chemometrics*, 1, 1987. doi: 10.1002/cem.1180010107. URL `http://dx.doi.org/10.1002/cem.1180010107`.

Wold, S., Antti, H., Lindgren, F., and Öhman, J. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 44(1-2):175–185, dec 1998. doi: 10.1016/s0169-7439(98)00109-9. URL `http://dx.doi.org/10.1016/s0169-7439(98)00109-9`.

Wold, S., Sjöström, M., and Eriksson, L. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2): 109–130, 2001a.

Wold, S., Trygg, J., Berglund, A., and Antti, H. Some recent developments in pls modeling. *Chemometrics and intelligent laboratory systems*, 58(2): 131–150, 2001b.

Wong, D., Wong, K., Goh, L., Heng, C., and Yap, M. Impact of dynamic online fed-batch strategies on metabolism, productivity and n-glycosylation quality in cho cell cultures. *Biotechnology and bioengineering*, 89(2): 164–177, 2005a.

Wong, J. W. H., Durante, C., and Cartwright, H. M. Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal. Chem.*, 77(17):5655–5661, sep 2005b. doi: 10.1021/ac050619p. URL `http://dx.doi.org/10.1021/ac050619p`.

Wooldridge, M. and Jennings, N. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(02):115–152, 1995.

Wu, W. and Manne, R. Fast regression methods in a lanczos (or pls-1) basis. theory and applications. *Chemometrics and intelligent laboratory systems*, 51(2):145–161, 2000.

Wurm, F. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nature Biotechnology*, 22(11):1393–8, 2004. ISSN 1087-0156. doi: 10.1038/nbt1026.

Xing, Z., Bishop, N., Leister, K., and Li, Z. J. Modeling kinetics of a large scale fed batch CHO cell culture by markov chain monte carlo method. pages 208–219, 2010.

Xu, Z., Li, J., and Zhou, J. Process development for robust removal of aggregates using cation exchange chromatography in monoclomal antibody purification with implementation of quality by design. *Preparative Biochemistry and Biotechnology*, 2012. ISSN 1082-6068. doi: 10.1080/10826068.2012.654572.

Yang, L., Harding, J. D., Ivanov, A. V., Ramasubramanyan, N., and Dong, D. D. Effect of cleaning agents and additives on protein a ligand degradation and chromatography performance. *Journal of Chromatography A*, 1385:63–68, mar 2015. doi: 10.1016/j.chroma.2015.01.068. URL `http://dx.doi.org/10.1016/j.chroma.2015.01.068`.

Yi, W., Clark, P., Mason, D., Keenan, M., Hill, C., Goddard, W., Peters, E., Driggers, E., and Hsieh-Wilson, L. Phosphofructokinase 1 glycosylation regulates cell growth and metabolism. *Science*, 337(6097):975–980, 2012.

Yoon, S., Choi, S., Song, J., and Lee, G. Effect of culture ph on erythropoietin production by chinese hamster ovary cells grown in suspension at 32.5 and 37.0 c. *Biotechnology and bioengineering*, 89(3):345–356, 2005.

Yu, L. Pharmaceutical quality by design: Product and process development, understanding, and control. *Pharm Res*, 25(4):781–791, jan 2008. doi: 10.1007/s11095-007-9511-1. URL `http://dx.doi.org/10.1007/s11095-007-9511-1`.

Zeng, A. Mathematical modeling and analysis of monoclonal antibody production by hybridoma cells. *Biotechnology and bioengineering*, 50(3): 238–247, 1996.

Zhang, X., Asara, J., Adamec, J., Ouzzani, M., and Elmagarmid, A. Data pre-processing in liquid chromatography–mass spectrometry-based proteomics. *Bioinformatics*, 21(21):4054–4059, 2005.

Zheng, Y.-B., Zhang, Z.-M., Liang, Y.-Z., Zhan, D.-J., Huang, J.-H., Yun, Y.-H., and Xie, H.-L. Application of fast fourier transform cross-correlation and mass spectrometry data for accurate alignment of chromatograms. *Journal of Chromatography A*, 1286:175–182, apr 2013. doi: 10.1016/j.chroma.2013.02.063. URL `http://dx.doi.org/10.1016/j.chroma.2013.02.063`.

Zhou, Y. and Titchener-Hooker, N. The application of a pareto optimisation method in the design of an integrated bioprocess. *Bioprocess and biosystems engineering*, 25(6):349–355, 2003.

Ziegelbauer, K. and Light, D. Monoclonal antibody therapeutics: Leading companies to maximise sales and market share. *Journal of Commercial Biotechnology*, 14(1), 2008. ISSN 1478-565X. URL `http://commercialbiotechnology.com/index.php/jcb/article/view/228`.

Appendix A

Part 1

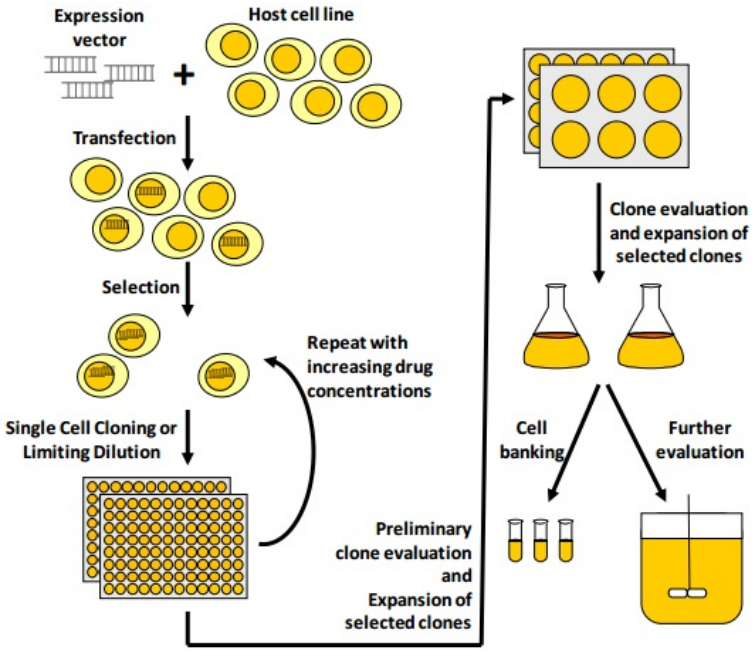


Figure A1: Development of a mammalian cell line for recombinant protein manufacture. A vector containing the gene of interest (GOI) and a selection marker is transfected into the host cell line. The transfected cells undergo selection and cloning to ensure cells are derived that contain the GOI. This amplification process can be repeated with increasing amounts of selection drug to derive cell clones which are more productive. Cell clones with high product titre are expanded before cell banking and further evaluation. This further evaluation can include tests which look at cell stability and quality or protein produced (Lai *et al.*, 2013).

Part 2

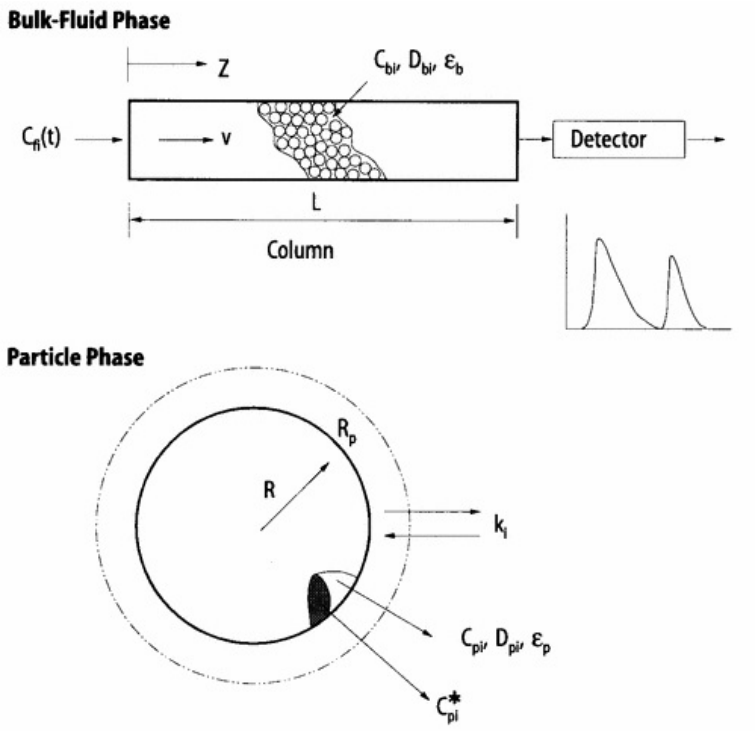


Figure A2: Modelling of fixed bed axial flow chromatography (Gu, 1995)

Mass balance for mobile phase (Gu, 1995)

$$-D_{bi}\frac{\partial^2 C_{bi}}{\partial Z^2}+v\frac{\partial C_{bi}}{\partial Z}+\frac{\partial C_{bi}}{\partial t}+\frac{3k_i(1-\varepsilon_b)}{\varepsilon_b R_p}(C_{bi}-C_{pi,R=R_p})=0 \qquad \text{(A1)}$$

$$-D_{bi}\frac{\partial^2 C_{bi}}{\partial Z^2} = \text{Transport by axial dispersion in the mobile phase}$$

$$v\frac{\partial C_{bi}}{\partial Z} = \text{Convective transport in the mobile phase}$$

$$\frac{\partial C_{bi}}{\partial t} = \text{Accumulation in the mpbile phase}$$

$$\frac{3k_i(1-\varepsilon_b)}{\varepsilon_b R_p}(C_{bi}-C_{pi,R=R_p}) = \text{Accumulation in the stationary phase}$$

Mass balance for solid phase (Gu, 1995)

$$(1-\varepsilon_p)\frac{\partial C^*_{pi}}{\partial t}+\varepsilon_p\frac{\partial C_{pi}}{\partial t}+\varepsilon_p D_{pi}\left[\frac{1}{R^2}\frac{\partial}{\partial R}\left(R^2\frac{\partial C_{pi}}{\partial R}\right)\right]=0 \qquad \text{(A2)}$$

$$(1-\varepsilon_p)\frac{\partial C_{pi}^*}{\partial t} = \text{Accumulation in the micropores}$$

$$\varepsilon_p \frac{\partial C_{pi}}{\partial t} = \text{Accumulation in the macropores}$$

$$\varepsilon_p D_{pi}\left[\frac{1}{R^2}\frac{\partial}{\partial R}\left(R^2\frac{\partial C_{pi}}{\partial R}\right)\right] = \text{Radial diffusion in particle}$$

Appendix B

Part 1

Table B1: A minimum resolution IV design for five factors (Load concentration, flow rate, load pH, elution pH, and gradient), showing the experimental conditions for the 15 experimental runs used to generated the lactoferrin data set.

Run identifier	Load concentration (mg/ml)	Flow rate (ml/min)	Load pH	Elution pH	Gradient (Column volumes)
Run 1	20	1.0	7	7	12
Run 2	10	1.5	8	8	8
Run 3	30	1.5	8	8	8
Run 4	30	0.5	8	8	16
Run 5	10	0.5	8	8	16
Run 6	20	1.0	7	7	12
Run 7	30	1.5	6	6	16
Run 8	10	1.5	6	6	8
Run 9	30	0.5	6	6	16
Run 10	10	0.5	6	6	16
Run 11	10	1.5	6	8	16
Run 12	30	0.5	6	8	8
Run 13	30	0.5	8	6	8
Run 14	10	1.5	8	6	16
Run 15	20	1.0	7	7	12

Part 2

$$\frac{d(X_v)}{dt} = \mu f_{gr}(X_v - X_{ap}) - k_D X_{ap} \frac{1}{1 + k_2 e^{-P\alpha N}} \tag{B1a}$$

$$\frac{d(X_{ap})}{dt} = k_{ap}(X_v - X_{ap})^2 - k_D X_{ap} \frac{1}{1 + k_2 e^{-P\alpha N}} \tag{B1b}$$

$$\frac{df_{gr}}{dt} = -\mu f_{gr}(1 - f_{gr}) \tag{B1c}$$

$$\frac{d(mAb)}{dt} = \alpha_g X_v \tag{B1d}$$

$$\frac{d[GLC]}{dt} = -a_{11} X_g - \left(\frac{a_1[GLC]}{k_{glc} + [GLC]} + 0.5 \frac{a_7[GLC] \cdot [GLN]}{(k_{glc}[GLC]) \cdot (k_{gln} + [GLN])} \right) X_{ng} \tag{B1e}$$

$$\frac{d[LAC]}{dt} = a_{22} X_g - \left(2 \cdot \frac{a_2[GLC]}{k_{glc} + [GLC]} + \frac{a_3[GLU]}{k_{glu} + [GLU]} \right) X_{ng} \tag{B1f}$$

$$\frac{d[GLN]}{dt} = -a_{33} X_g - \left(\frac{a_5[GLN]}{k_{gln5} + [GLN]} + \frac{a_7[GLC] \cdot [GLN]}{(k_{glc} + [GLC]) \cdot (k_{gln} + [GLN])} \right) X_{ng} \tag{B1g}$$

$$\frac{d[GLU]}{dt} = a_{44} X_g + \left(-\frac{a_3[GLU]}{k_{glu} + [GLU]} + \frac{a_7[GLC] \cdot [GLN]}{(k_{glc} + [GLC]) \cdot (k_{gln} + [GLN])} \right) x_{ng} \tag{B1h}$$

$$\frac{d[ASN]}{dt} = -a_{55} X_g - \frac{a_4[ASN]}{k_{asn} + [ASN]} X_{ng} \tag{B1i}$$

$$\frac{d[ASP]}{dt} = a_{66} X_g + \left(\frac{a_4[ASN]}{k_{asn} + [ASN]} - \frac{a_8[ASP]}{k_{asp} + [ASP]} \right) X_{ng} \tag{B1j}$$

$$\frac{d[ALA]}{dt} = a_{77} X_g + \left(-\frac{a_6[ALA]}{k_{ala} + [ALA]} + \frac{a_7[GLC] \cdot [GLN]}{(k_{glc} + [GLC]) \cdot (k_{gln} + [GLN])} \right) X_{ng} \tag{B1k}$$

$$\frac{d[AMM]}{dt} = a_{88} X_g + \left(\frac{a_3 + [GLU]}{k_{glu} + [GLU]} + \frac{a_4 + [ASN]}{k_{asn} + [ASN]} + 2 \frac{a_5 + [GLN]}{k_{gln5} + [GLN]} + \frac{a_6 + [ALA]}{k_{ala} + [ALA]} + \frac{a_8 + [ASP]}{k_{asp} + [ASP]} \right) X_{ng} \tag{B1l}$$

Part 3

$$\frac{d(V)}{dt} = F_{in} + f_{glc} - F_{out} \tag{B2a}$$

$$\frac{d(VX_v)}{dt} = \mu VX_v - \mu_d VX_v - F_{out}X_v \tag{B2b}$$

$$\frac{d(VX_d)}{dt} = \mu_d VX_v - K_{lysis}VX_d - Fd - FoutX_d \tag{B2c}$$

$$X_t = X_v + X_d \tag{B2d}$$

$$\mu = \mu_{min} + (\mu_{max} - \mu_{min})(\frac{[GLC][GLN]}{(k_{glc} + [GLC])(k_{gln} + [GLN])} + \frac{[ARG][VAL][LYS][THR]}{(k_{arg} + [ARG])(k_{val} + [VAL])(k_{lys} + [LYS])(k_{thr} + [THR])} \cdot \frac{[HIS][SER][ILE][PHE][LEU]}{(k_{his} + [HIS])(k_{ser} + [SER])(k_{ile} + [ILE])(k_{phe} + [PHE])(k_{leu} + [LEU])}) \tag{B2e}$$

$$\mu_d = \mu_{d,max}\frac{k_{d,amm} + [AMM] - [AMM]_{cr}}{k_{d,amm}} \cdot \frac{(k_{d,lac} + [LAC] - [LAC]_{cr})}{k_{d,lac}} \tag{B2f}$$

$$\frac{d(V[GLC])}{dt} = Q_{glc}VX_v + F_{in}[GLC]_{in} - F_{out}[GLC] \tag{B3a}$$
$$\frac{d(V[LAC])}{dt} = Q_{lac}VX_v - F_{out}[LAC] \tag{B3b}$$
$$\frac{d(V[AMM])}{dt} = Q_{amm}VX_v + k_{d,gln}V[GLN] - F_{out}[AMM] \tag{B3c}$$
$$\frac{d(V[ALA])}{dt} = Q_{ala}VX_v + F_{in}[ALA]_{in} - F_{out}[ALA] \tag{B3d}$$
$$\frac{d(V[ARG])}{dt} = Q_{arg}VX_vF_{in}[ARG]_{in} - F_{out}[ARG] \tag{B3e}$$
$$\frac{d(V[ASN])}{dt} = Q_{asn}VX_vF_{in}[ASN]_{in} - F_{out}[ASN] \tag{B3f}$$
$$\frac{d(V[ASP])}{dt} = Q_{asp}VX_vF_{in}[ASP]_{in} - F_{out}[ASP] \tag{B3g}$$
$$\frac{d(V[CYS])}{dt} = Q_{cys}VX_vF_{in}[CYS]_{in} - F_{out}[CYS] \tag{B3h}$$
$$\frac{d(V[GLU])}{dt} = Q_{glu}VX_vF_{in}[GLU]_{in} - F_{out}[GLU] \tag{B3i}$$
$$\frac{d(V[GLN])}{dt} = Q_{gln}VX_v - k_{d,gln}V[GLN] + F_{in}[GLN]_{in} - F_{out}[GLN] \tag{B3j}$$
$$\frac{d(V[GLY])}{dt} = Q_{gly}VX_vF_{in}[GLY]_{in} - F_{out}[GLY] \tag{B3k}$$
$$\frac{d(V[HIS])}{dt} = Q_{his}VX_vF_{in}[HIS]_{in} - F_{out}[HIS] \tag{B3l}$$
$$\frac{d(V[ILE])}{dt} = Q_{ile}VX_vF_{in}[ILE]_{in} - F_{out}[ILE] \tag{B3m}$$
$$\frac{d(V[LEU])}{dt} = Q_{leu}VX_vF_{in}[LEU]_{in} - F_{out}[LEU] \tag{B3n}$$
$$\frac{d(V[LYS])}{dt} = Q_{lys}VX_vF_{in}[LYS]_{in} - F_{out}[LYS] \tag{B3o}$$
$$\frac{d(V[MET])}{dt} = Q_{met}VX_vF_{in}[MET]_{in} - F_{out}[MET] \tag{B3p}$$
$$\frac{d(V[PHE])}{dt} = Q_{phe}VX_vF_{in}[PHE]_{in} - F_{out}[PHE] \tag{B3q}$$
$$\frac{d(V[PRO])}{dt} = Q_{pro}VX_vF_{in}[PRO]_{in} - F_{out}[PRO] \tag{B3r}$$
$$\frac{d(V[SER])}{dt} = Q_{ser}VX_vF_{in}[SER]_{in} - F_{out}[SER] \tag{B3s}$$
$$\frac{d(V[THR])}{dt} = Q_{thr}VX_vF_{in}[THR]_{in} - F_{out}[THR] \tag{B3t}$$
$$\frac{d(V[TYR])}{dt} = Q_{tyr}VX_vF_{in}[TYR]_{in} - F_{out}[TYR] \tag{B3u}$$
$$\frac{d(V[VAL])}{dt} = Q_{val}VX_vF_{in}[VAL]_{in} - F_{out}[VAL] \tag{B3v}$$

$$Q_{ala} = -\frac{mu}{Y_{x,ala}} + Y_{ala,x} \tag{B4a}$$

$$Q_{arg} = -\frac{mu}{Y_{x,arg}} + Y_{arg,glu}Q_{glu} + Y_{arg,pro}Q_{pro} - Y_{arg,asp}Q_{asp} \tag{B4b}$$

$$Q_{asn} = -\frac{mu}{Y_{x,asn}} + Y_{asn,asp}Q_{asp} \tag{B4c}$$

$$Q_{asp} = -\frac{mu}{Y_{x,asp}} + Y_{asp,arg}Q_{arg} + Y_{asp,x} \tag{B4d}$$

$$Q_{cys} = -\frac{mu}{Y_{x,cys}} + Y_{cys,ser}Q_{ser} \tag{B4e}$$

$$Q_{glu} = -\frac{mu}{Y_{x,glu}} + Y_{glu,pro}Q_{pro} - Y_{gly,his}Q_{his} - Y_{glu,gln}Q_{gln} - Y_{glu,arg}Q_{arg} + Y_{glu,x} \tag{B4f}$$

$$Q_{gln} = -\frac{\mu}{Y_{x,gln}} - M_{gln} + Y_{gln,glu}Q_{glu} \tag{B4g}$$

$$M_{gln} = \frac{\alpha_1[GLN]}{\alpha_2 + [GLN]} \tag{B4h}$$

$$Q_{gly} = -\frac{mu}{Y_{x,gly}} + Y_{gly,ser}Q_{ser} \tag{B4i}$$

$$Q_{his} = -\frac{mu}{Y_{x,his}} \tag{B4j}$$

$$Q_{ile} = -\frac{mu}{Y_{x,ile}} \tag{B4k}$$

$$Q_{leu} = -\frac{mu}{Y_{x,leu}} \tag{B4l}$$

$$Q_{lys} = -\frac{mu}{Y_{x,lys}} + Y_{lys,x} \tag{B4m}$$

$$Q_{met} = -\frac{mu}{Y_{x,met}} \tag{B4n}$$

$$Q_{phe} = -\frac{mu}{Y_{x,phe}} \tag{B4o}$$

$$Q_{pro} = -\frac{mu}{Y_{x,pro}} + Y_{pro,glu}Q_{glu} + Y_{pro,arg}Q_{glu} - Y_{pro,arg}Q_{arg} \tag{B4p}$$

$$Q_{ser} = -\frac{mu}{Y_{x,ser}} + Y_{ser,gly}Q_{gly} \tag{B4q}$$

$$Q_{thr} = -\frac{mu}{Y_{x,thr}} \tag{B4r}$$

$$Q_{tyr} = -\frac{mu}{Y_{x,tyr}} + Y_{tyr,phe}Q_{phe} \tag{B4s}$$

$$Q_{val} = -\frac{mu}{Y_{x,val}} \tag{B4t}$$

$$Q_{glc} = -\frac{mu}{Y_{x,glc}} - M_{glc} \tag{B4u}$$

$$Q_{lac} = -Y_{lac,glc}Q_{glc} \tag{B4v}$$

$$Q_{amm} = -Y_{amm,gln}Q_{gln} \tag{B4w}$$

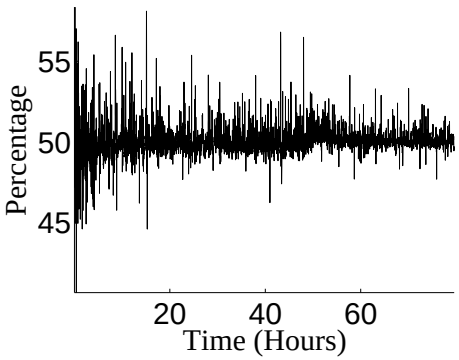
Table B2: Summary of the constants used in the Naderi *et al.* (2011) first principles model, values were obtained from literature.

Parameter	Value	Parameter	Value
a_{11}	0.1317	K_{GLC}	2.4100
a_{22}	0.2451	K_{GLN}	0.5090
a_{33}	0.0416	K_{GLN5}	4.9680
a_{44}	0.0148	K_{GLU}	0.2230
a_{55}	0.0023	K_{ASN}	0.0960
a_{66}	0.0099	K_{ASP}	0.0605
a_{77}	0.0235	K_{ALA}	0.1090
a_{88}	0.0416	M	0.0420
a_1	0.0486	M_D	0.0090
a_2	0.0115	K_2	5.4200
a_3	0.0001	A	1.3970
a_4	0.0011	K_{D2}	0.1500
a_5	0.0012	K_3	0.0500
a_6	0.0001	K_4	0.0170
a_7	0.0012	K_{ap}	0.0093
a_8	0.0010	a_g	0.0250

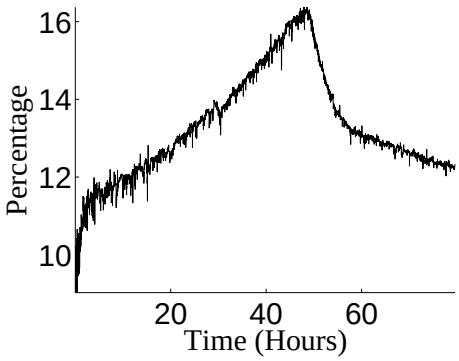
Table B3: Summary of the constants used in the Kontoravdi *et al.* (2007) first principles model, values were obtained from literature.

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
$[AMM]_{cr}$	5	K_{VAL}	0.015	$Y_{GLU,HIS}$	0.1	$Y_{X,GLC}$	$7.7x10^{7}$	α_1	$2x10^{-14}$
K_{ARG}	0.06	$[LAC]_{cr}$	20	$Y_{GLU,PRO}$	0.01	$Y_{X,GLN}$	$8x10^{8}$	α_2	2
$K_{d,AMM}$	0.05	M_{GLC}	$1x10^{-14}$	$Y_{GLY,SER}$	0.065	$Y_{X,GLU}$	$9.6x10^{8}$	$\mu_{d,min}$	$5x10^{-4}$
$K_{d,GLN}$	0.009	$Y_{ALA,X}$	$5.5x10^{-12}$	$Y_{GLU,X}$	$4.2x10^{-13}$	$Y_{X,GLY}$	$1.6x10^{9}$	μ_{min}	$2x10^{-3}$
$K_{d,LAC}$	4.5	$Y_{AMM,GLN}$	1.2	$Y_{LAC,GLC}$	1.45	$Y_{X,HIS}$	$4.6x10^{9}$	μ_{max}	$5.2x10^{9}$
K_{GLC}	0.15	$Y_{ARG,ASP}$	0.001	$Y_{LYS,X}$	$1x10^{-13}$	$Y_{X,ILE}$	$2x10^{9}$		
K_{GLN}	0.22	$Y_{ARG,GLU}$	0.01	$Y_{PRO,ARG}$	0.6	$Y_{X,LEU}$	$1.5x10^{9}$		
K_{HIS}	0.005	$Y_{ARG,PRO}$	0.001	$Y_{PRO,GLU}$	0.5	$Y_{X.LYS}$	$1.3x10^{9}$		
K_{ILE}	0.025	$Y_{ASN,ASP}$	0.01	$Y_{SER,GLY}$	$1x10^{-13}$	$Y_{X,MET}$	$5.2x10^{9}$		
K_{LEU}	0.02	$Y_{ASP,ARG}$	0.0001	$Y_{TYR,PHE}$	0.5	$Y_{X,PHE}$	$4.1x10^{9}$		
K_{LYS}	0.013	$Y_{ASP,X}$	$2x10^{-16}$	$Y_{X,ALA}$	$1x10^{9}$	$Y_{X,PRO}$	$2.1x10^{9}$		
K_{LYSIS}	0.0001	$Y_{CYS,SER}$	0.1	$Y_{X,ARG}$	$2x10^{9}$	$Y_{X,SER}$	$2.5x10^{9}$		
K_{PHE}	0.04	$Y_{GLN,GLU}$	0.1	$Y_{X,ASN}$	$1.5x10^{9}$	$Y_{X,THR}$	$1.7x10^{9}$		
K_{SER}	0.03	$Y_{GLU,ARG}$	0.001	$Y_{X,ASP}$	$1.1x10^{9}$	$Y_{X,TYR}$	$2.6x10^{9}$		
K_{THR}	0.05	$Y_{GLU,GLN}$	0.7	$Y_{X,CYS}$	$6x10^{8}$	$Y_{X,VAL}$	$3x10^{9}$		

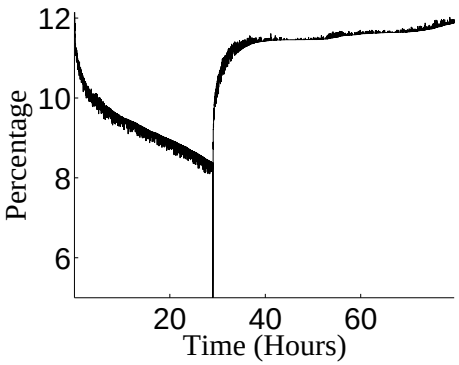
Appendix C



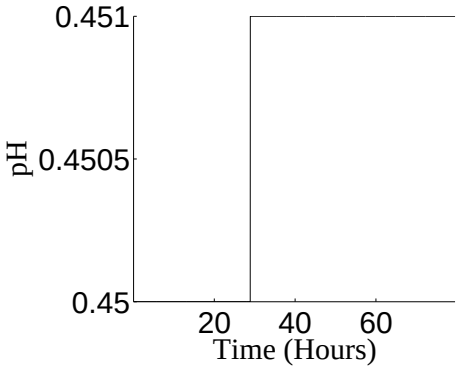
(a) Dissolved oxygen



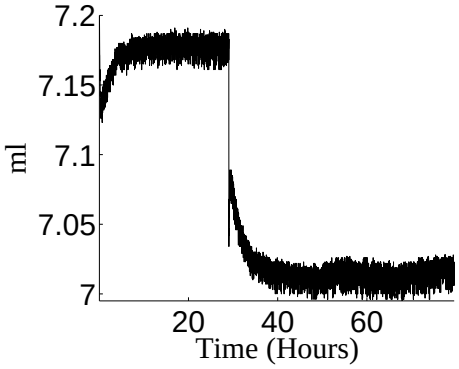
(b) Oxygen



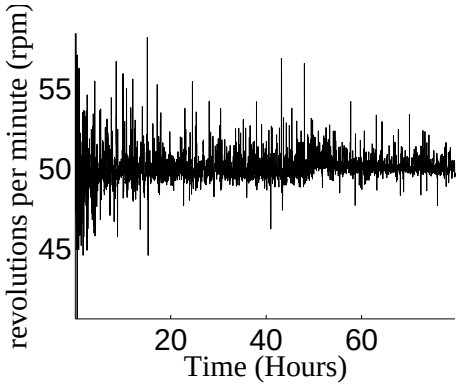
(c) Carbon dioxide



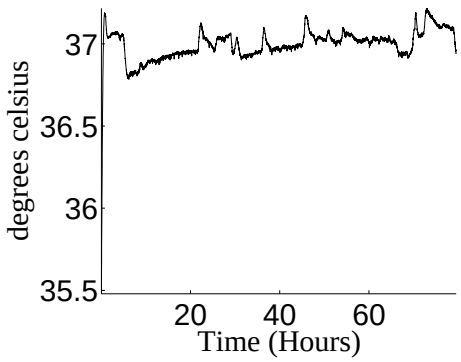
(d) pH



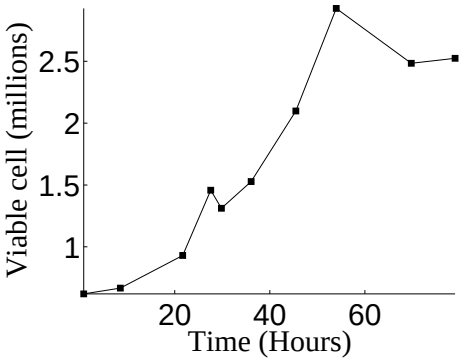
(e) Base



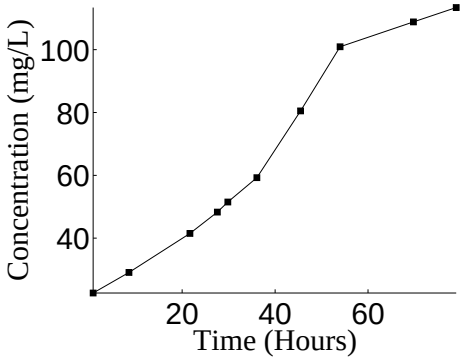
(f) Stirrer speed



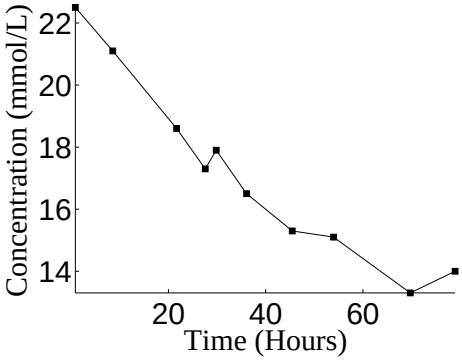
(g) Temperature



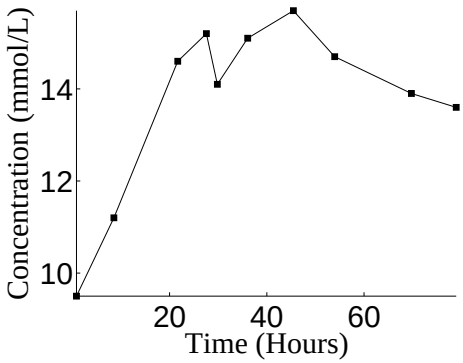
(h) Viable cell count



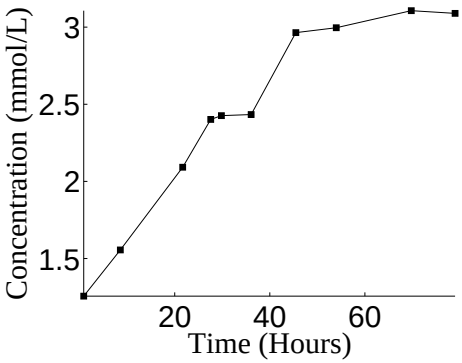
(i) Product titre



(j) Glucose concentration

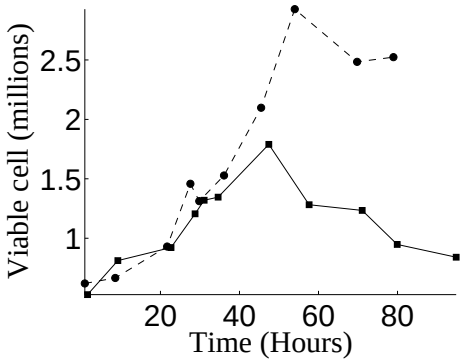


(k) Lactate concentration

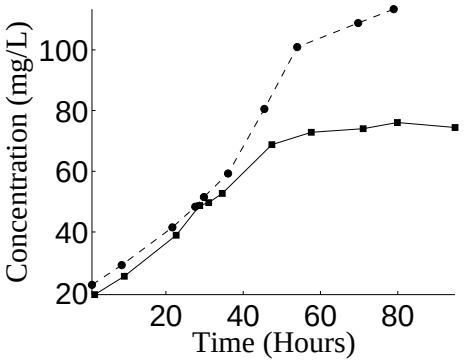


(l) Ammonia concentration

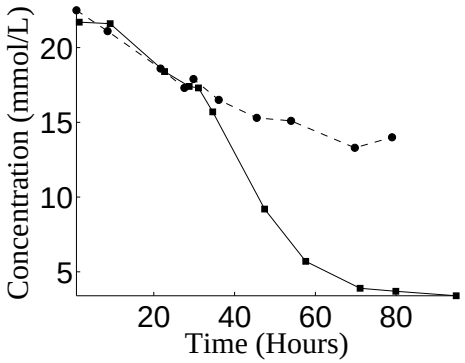
Figure C1: Example of the raw data collected for Run 9 (table 5.1 on page 93) the murine Hybridoma cell cultivation. Sub-figures (a-g) are contained within matrix (a) and are on line measurements; Sub-figures (h-l) are contained within matrices (b and c) and are off line measurements



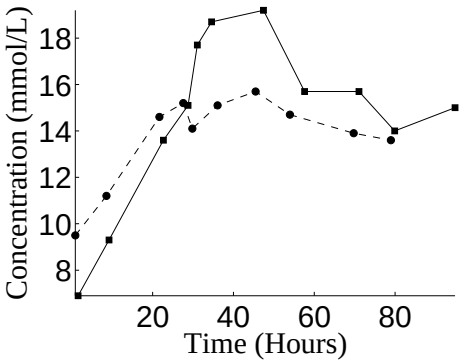
(a) Viable cell count



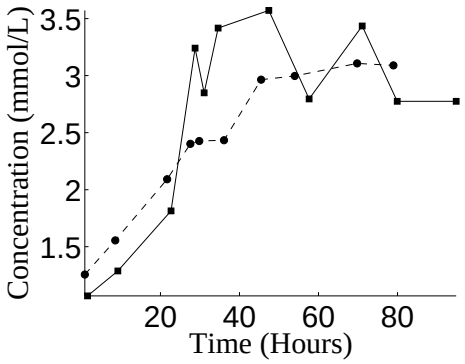
(b) Product titre



(c) Glucose concentration

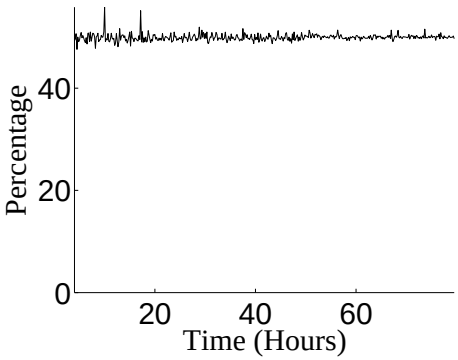


(d) Lactate concentration

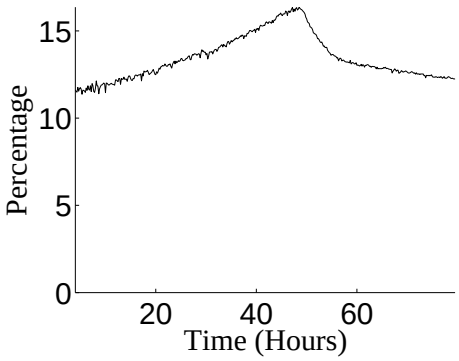


(e) Ammonia concentration

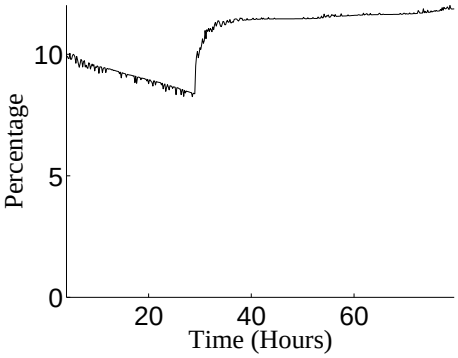
Figure C2: Comparison of raw off line measurements for runs 9 (circular marker with dashed line) and 11 (square marker with solid line) (table 5.1 on page 93) post cubic spline and data cut.



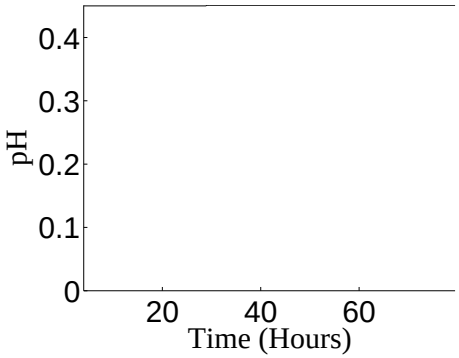
(a) Dissolved oxygen



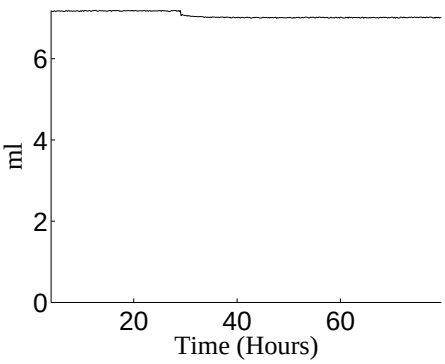
(b) Oxygen



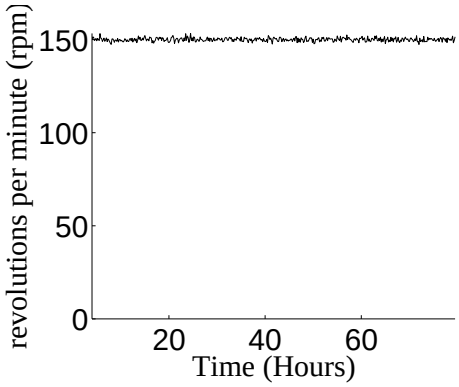
(c) Carbon dioxide



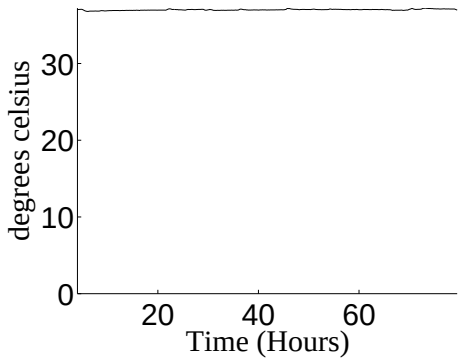
(d) pH



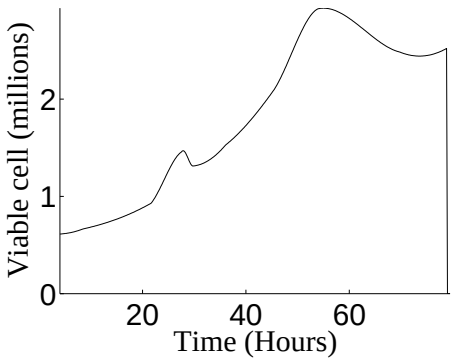
(e) Base



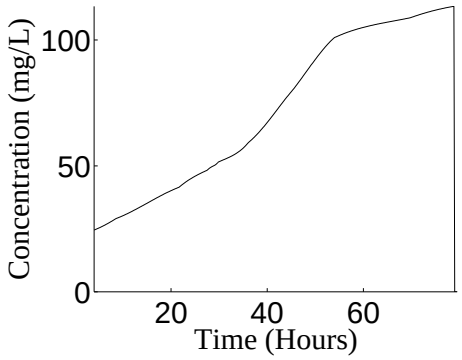
(f) Stirrer speed



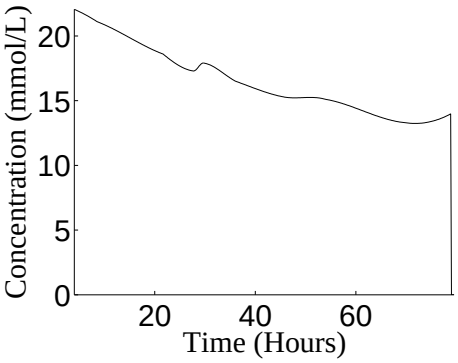
(g) Temperature



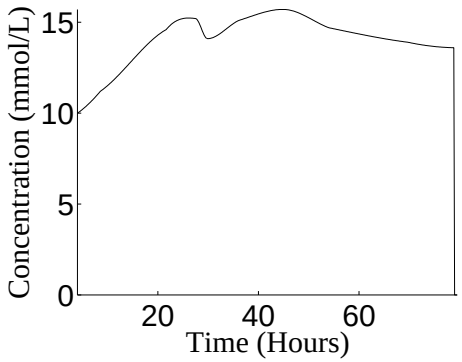
(h) Viable cell count



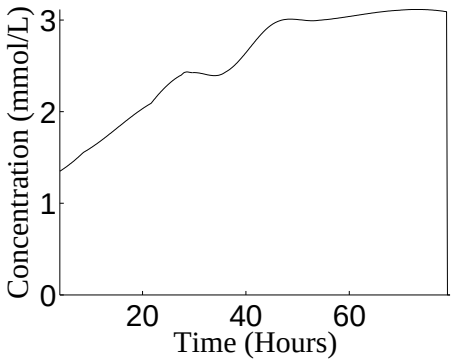
(i) Product titre



(j) Glucose concentration

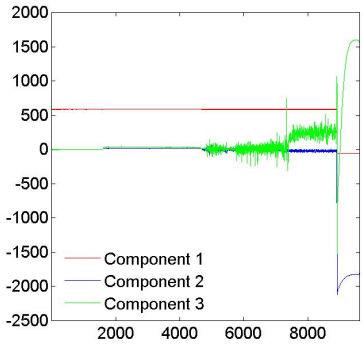


(k) Lactate concentration

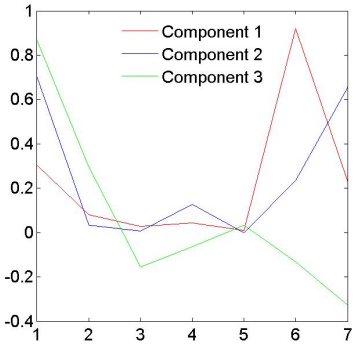


(l) Ammonia concentration

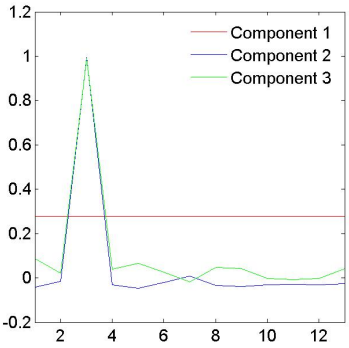
Figure C3: Data for Run 9 (table 5.1 on 93) post cubic spline and data cut.



(a) Mode one

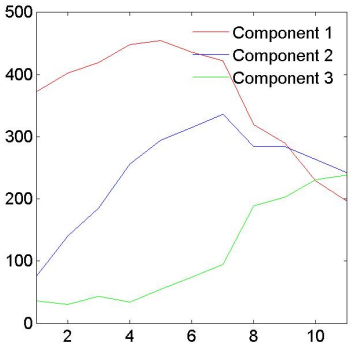


(b) Mode two

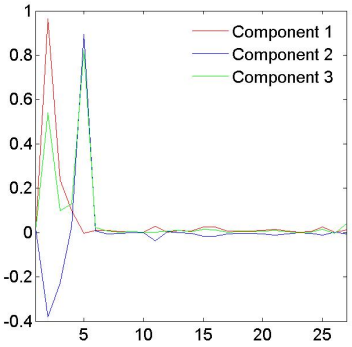


(c) Mode three

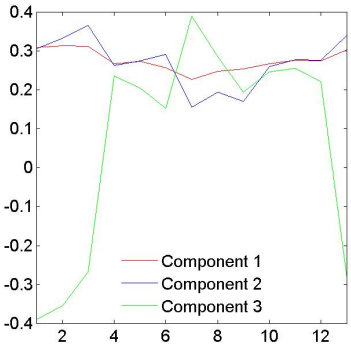
Figure C4: Loadings vectors for a three component model constructed for the on-line data measurements. Model 1 (Table ?? on page ??) showing (a) Mode one (over time) (b) Mode two (variables) (c) Mode 3 (Batches)



(a) Mode one



(b) Mode two



(c) Mode three

Figure C5: Loadings vectors for a three component model constructed for the on-line data measurements. Model 2 (Table ?? on page ??) showing (a) Mode one (over time) (b) Mode two (variables) (c) Mode 3 (Batches)

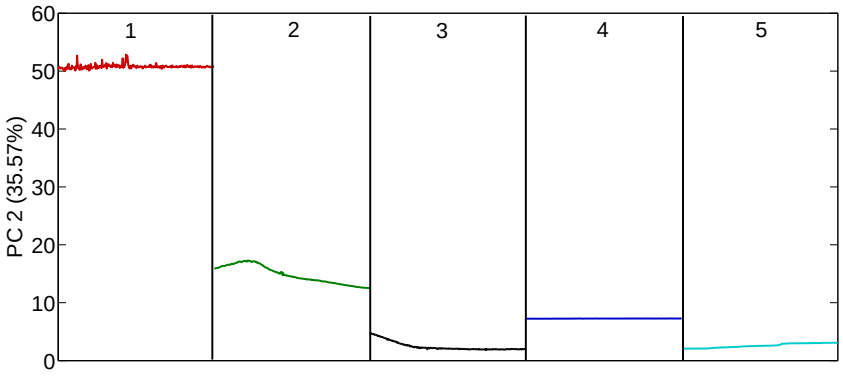


Figure C6: Loadings plot for PC2 showing 5 on-line variables; (1) DO, (2) O_2, (3) CO_2, (4) pH, and (5) base.

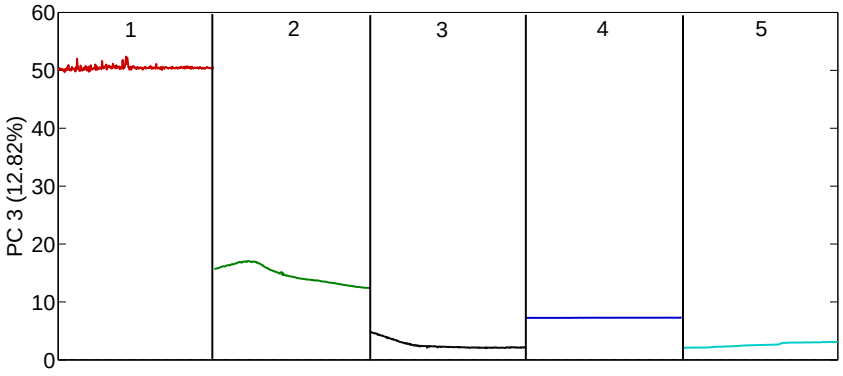


Figure C7: Loadings plot for PC3 showing 5 on-line variables; (1) DO, (2) O_2, (3) CO_2, (4) pH, and (5) base.

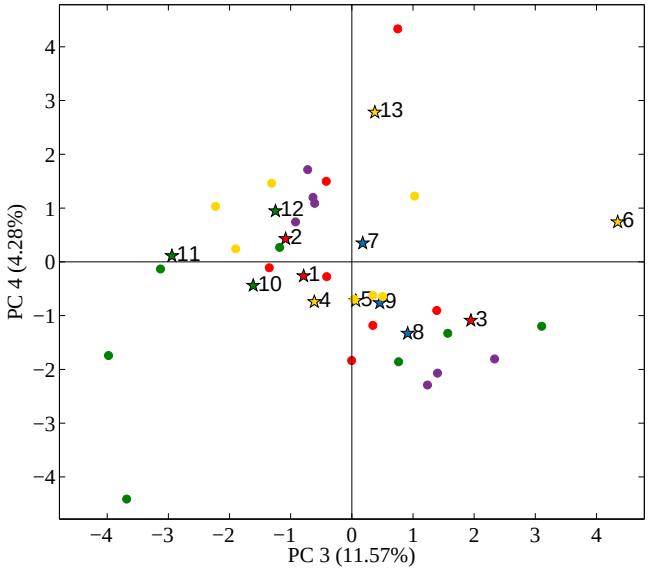


Figure C8: Bi-plot showing PC3 and PC4 for the PCA analysis containing the off-line data for glucose (yellow circles), lactate (red circles), titre (purple circles), and viable cell (green circles). The scores are grouped to show dissolved oxygen (red stars), osmolality (yellow stars), pH (blue stars), and sparger (green stars).

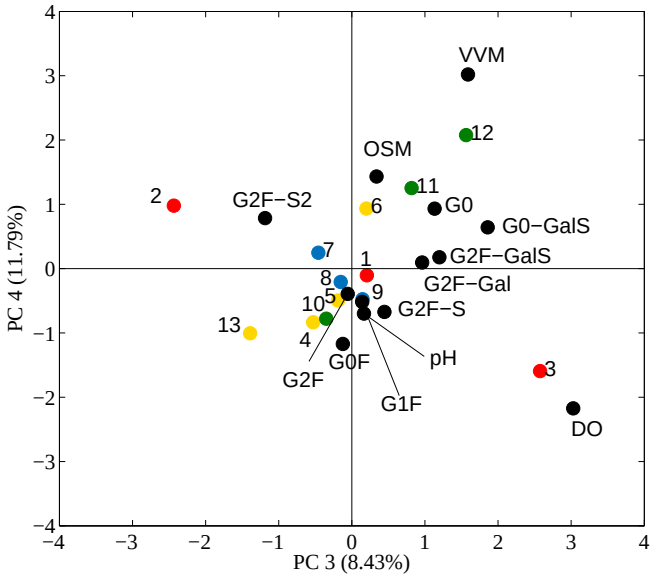


Figure C9: Bi-plot showing PC3 and PC4 for the PCA analysis containing the off-line data for the operating parameter set points and final product glycosylation profile. The scores are coloured to show dissolved oxygen (red), osmolality (yellow), pH (blue), and sparger (green). The glycans and the off-line operating parameters (black) are labelled to identify which item they are.

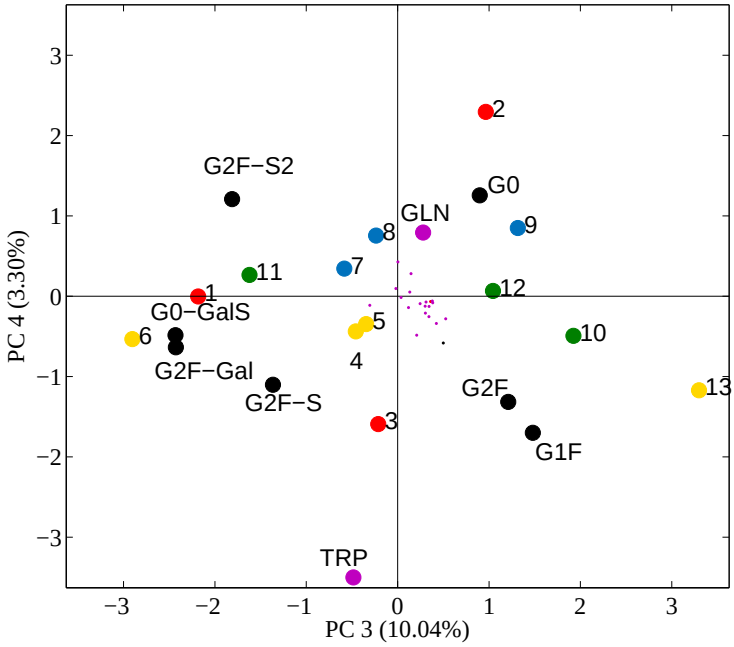


Figure C10: Bi-plot showing PC1 and PC2 for the PCA analysis containing the off-line data for the amino acid concentrations and final product glycosylation profile. The scores are coloured to show dissolved oxygen (red), osmolality (yellow), pH (blue), and sparger (green). The loadings are colour coded to show glycans (black) and amino acids (pink).

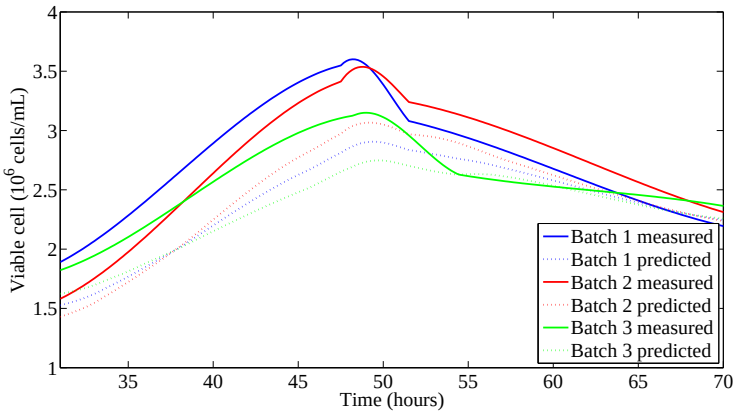


Figure C11: Measured and predicted viable cell count for dissolved oxygen experiments (batches 1-3) for model A.

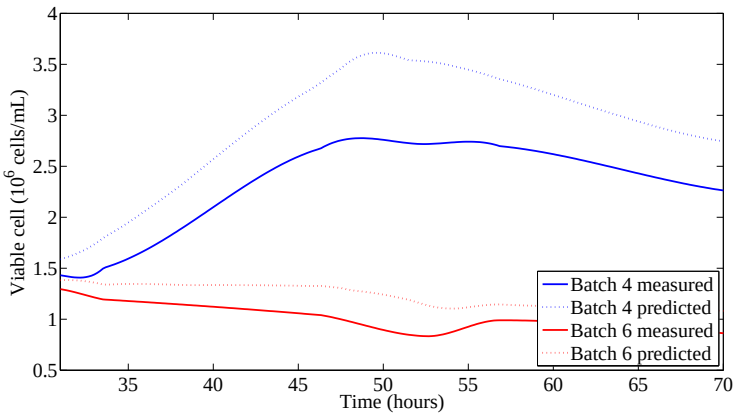


Figure C12: Measured and predicted viable cell count for osmolality experiments (batches 4 and 6) for model A.

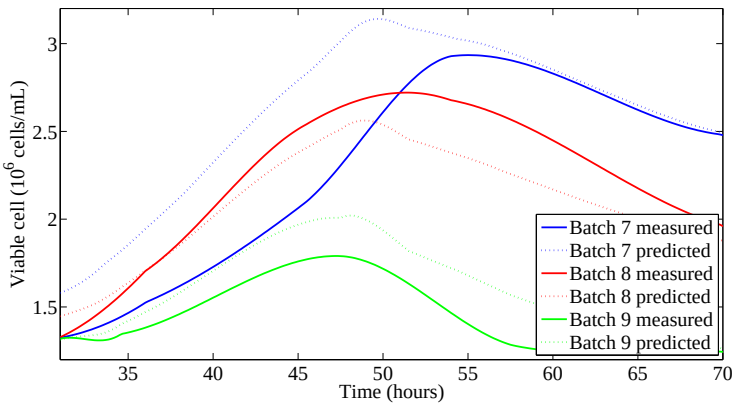


Figure C13: Measured and predicted viable cell count for pH experiments (batches 7-9) for model A.

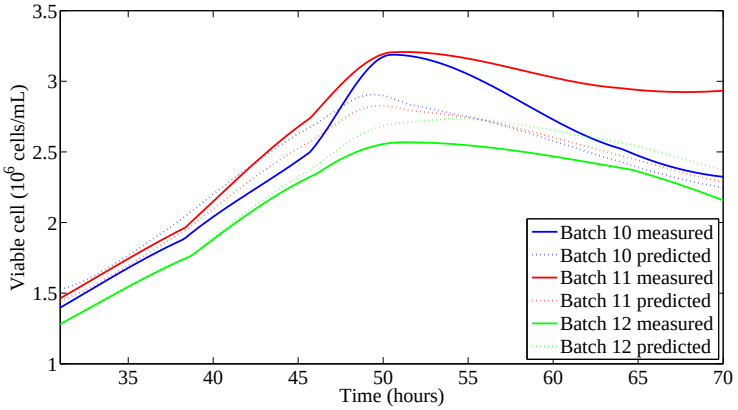


Figure C14: Measured and predicted viable cell count for sparger (vvm) experiments (batches 10-12) for model A.

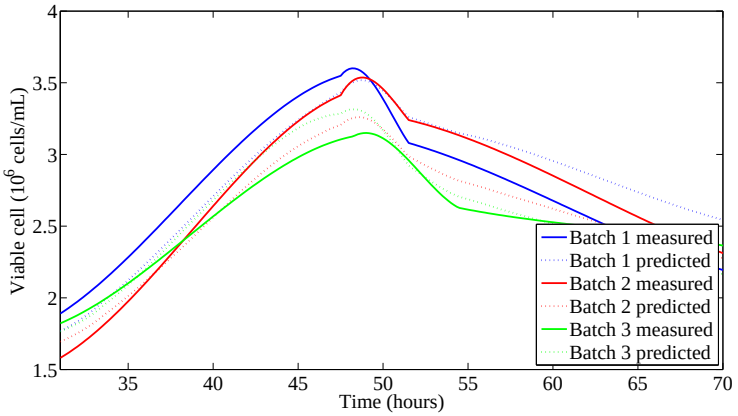


Figure C15: Measured and predicted viable cell count for dissolved oxygen experiments (batches 1-3) for model B.

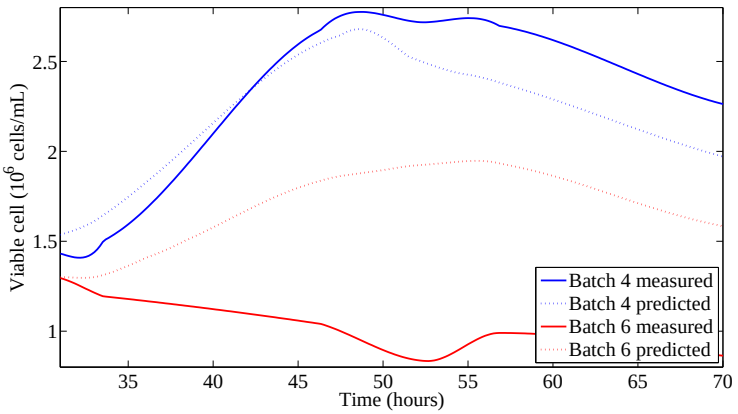


Figure C16: Measured and predicted viable cell count for osmolality experiments (batches 4 and 6) for model B.

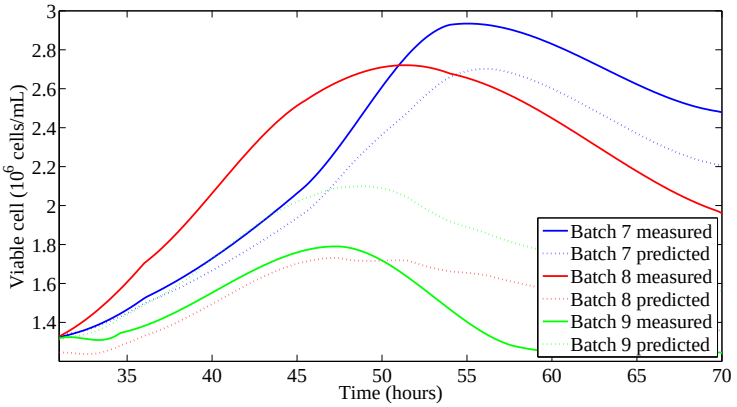


Figure C17: Measured and predicted viable cell count for pH experiments (batches 7-9) for model B.

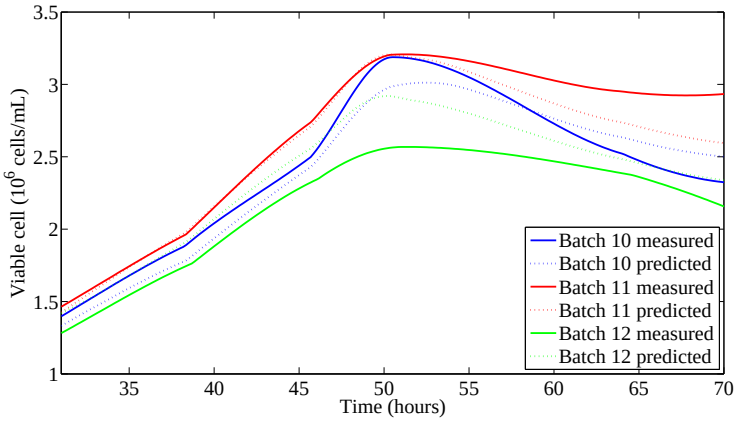


Figure C18: Measured and predicted viable cell count for sparger (vvm) experiments (batches 10-12) for model B.

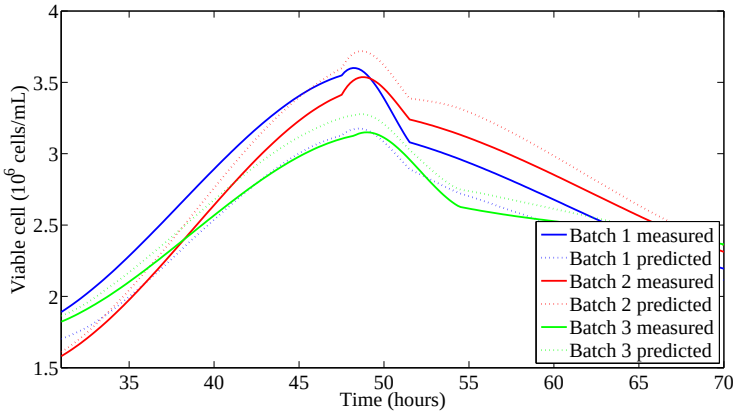


Figure C19: Measured and predicted viable cell count for dissolved oxygen experiments (batches 1-3) for model C.

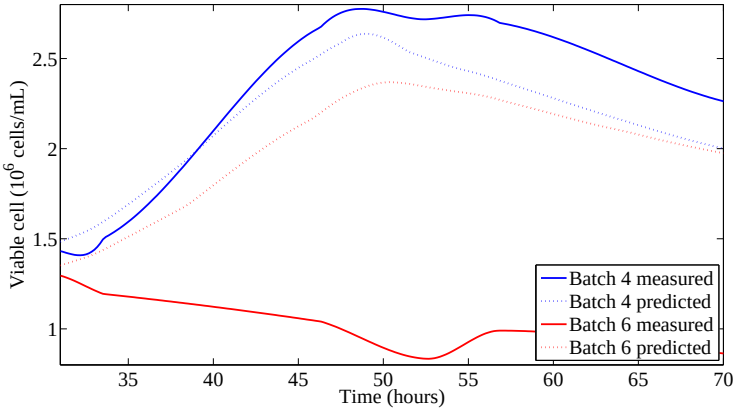


Figure C20: Measured and predicted viable cell count for osmolality experiments (batches 4 and 6) for model C.

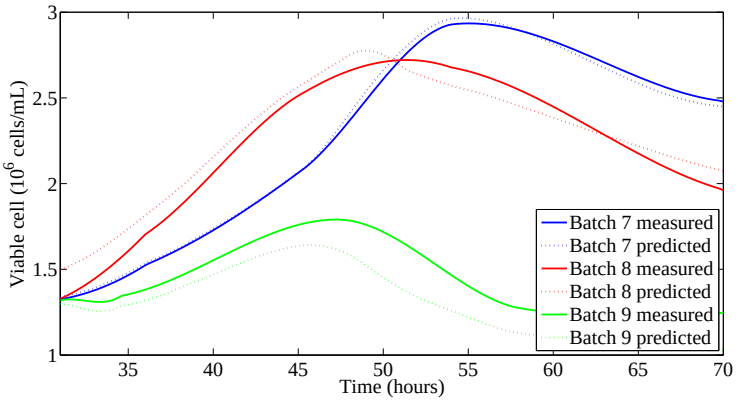


Figure C21: Measured and predicted viable cell count for pH experiments (batches 7-9) for model C.

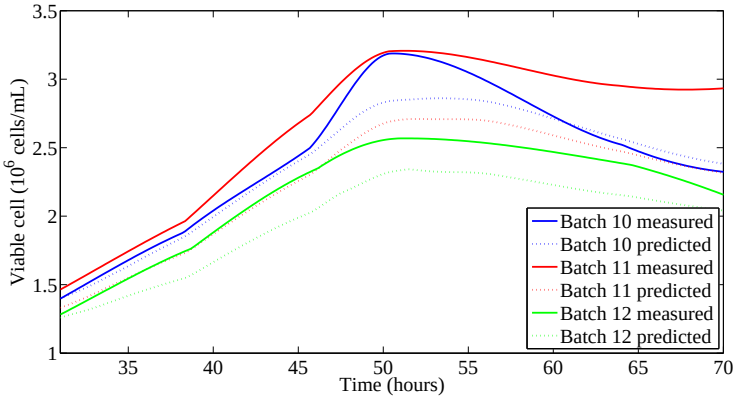


Figure C22: Measured and predicted viable cell count for sparger (vvm) experiments (batches 10-12) for model C.

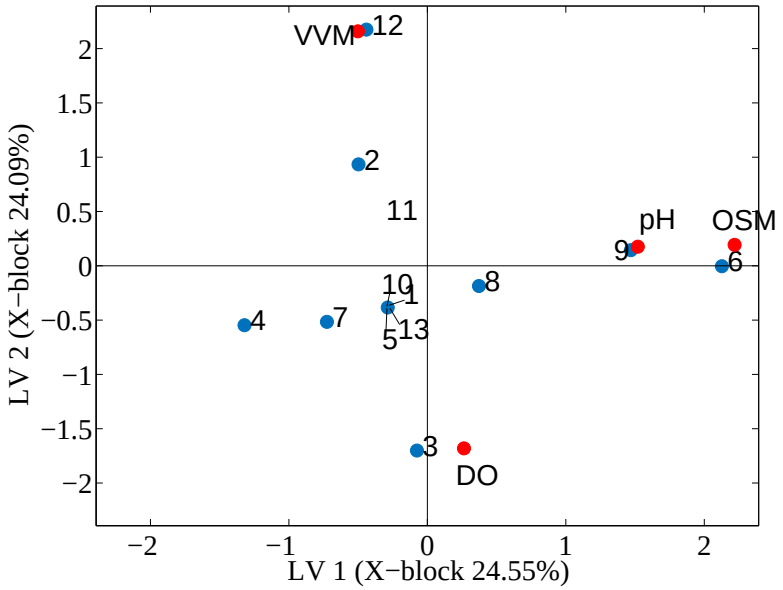


Figure C23: Bi-plot for LV1 and LV2 for X-block data for model A to predict viable cell count.

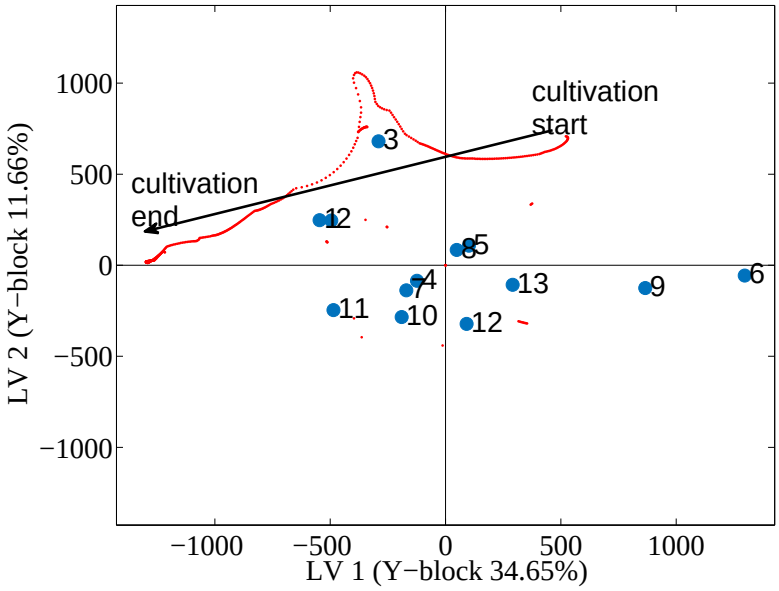


Figure C24: Bi-plot for LV1 and LV2 for Y-block data for model A to predict viable cell count.

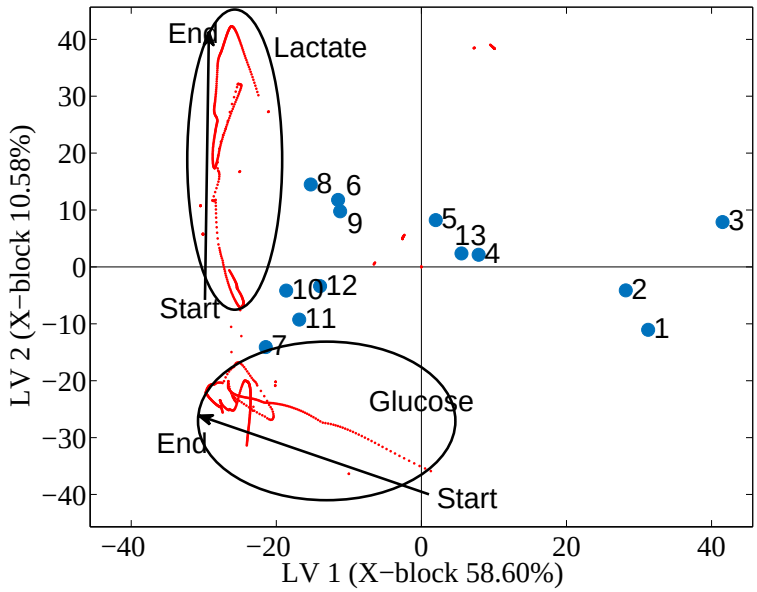


Figure C25: Bi-plot for LV1 and LV2 for X-block data for model B to predict viable cell count.

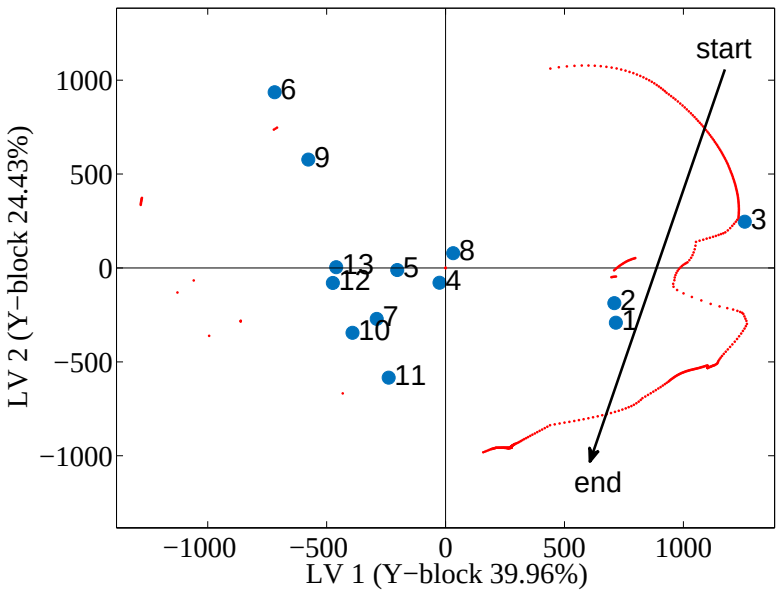


Figure C26: Bi-plot for LV1 and LV2 for Y-block data for model B to predict viable cell count.

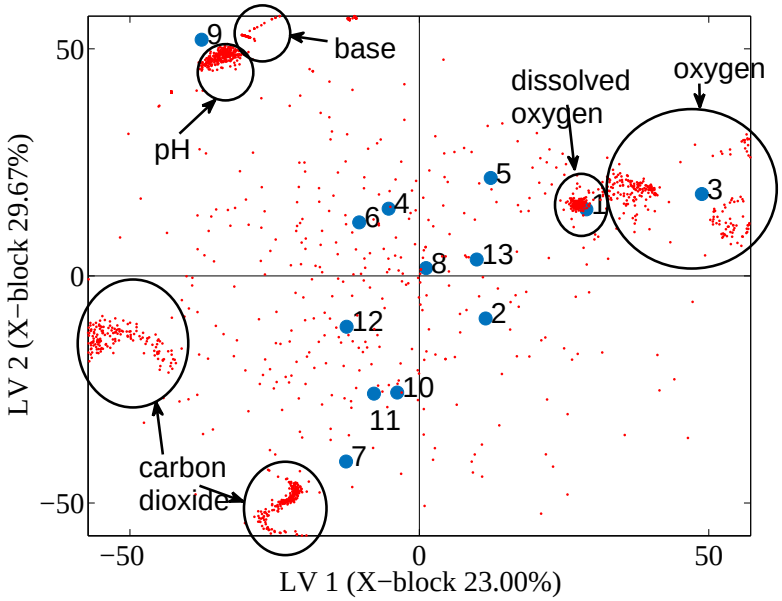


Figure C27: Bi-plot for LV1 and LV2 for X-block data for model C to predict viable cell count.

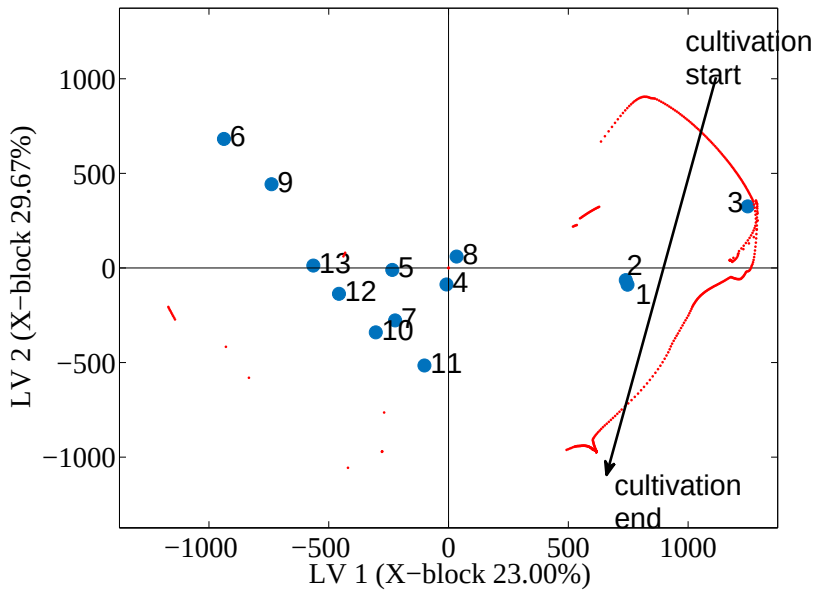


Figure C28: Bi-plot for LV1 and LV2 for Y-block data for model C to predict viable cell count.

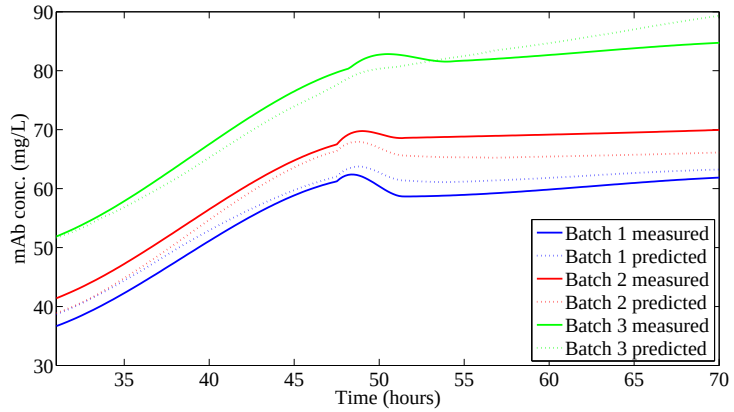


Figure C29: Measured and predicted product titre for dissolved oxygen experiments (batches 1-3) for model A.

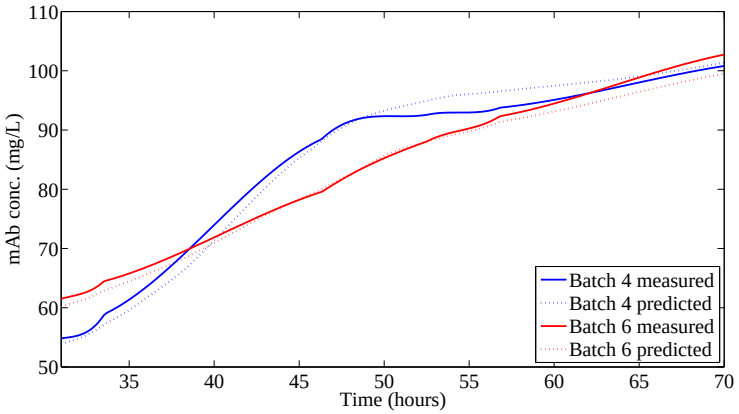


Figure C30: Measured and predicted product titre for osmolality experiments (batches 4 and 6) for model A.

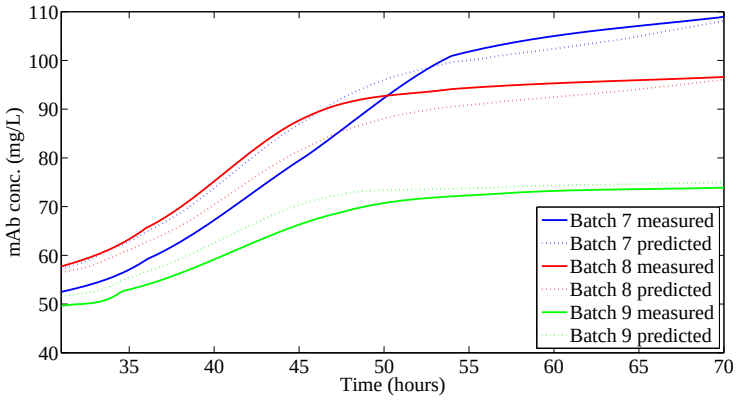


Figure C31: Measured and predicted product titre for pH experiments (batches 7-9) for model A.

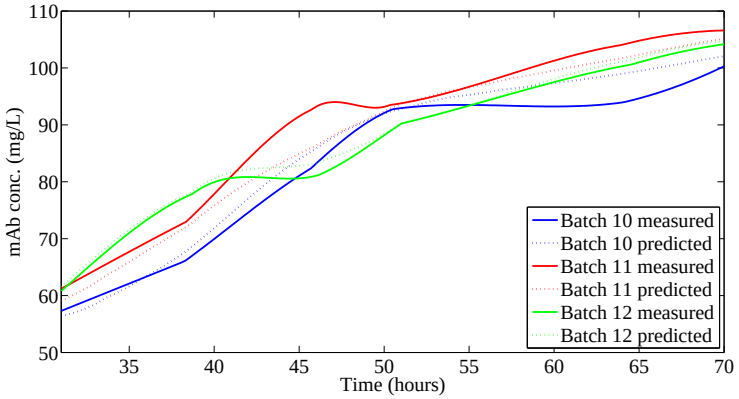


Figure C32: Measured and predicted product titre for sparger (vvm) experiments (batches 10-12) for model A.

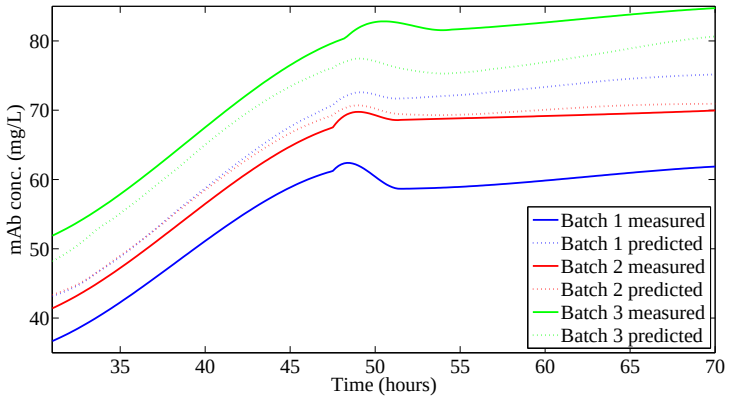


Figure C33: Measured and predicted product titre for dissolved oxygen experiments (batches 1-3) for model B.

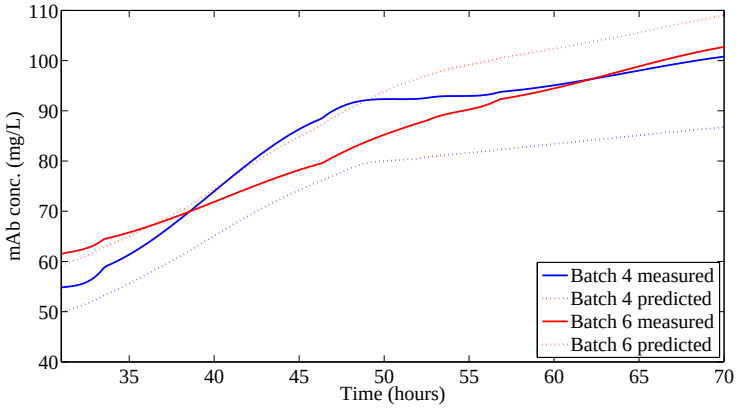


Figure C34: Measured and predicted product titre for osmolality experiments (batches 4 and 6) for model B.

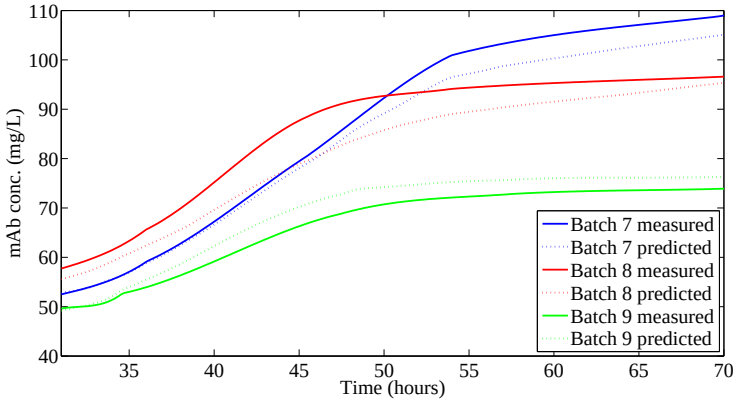


Figure C35: Measured and predicted product titre for pH experiments (batches 7-9) for model B.

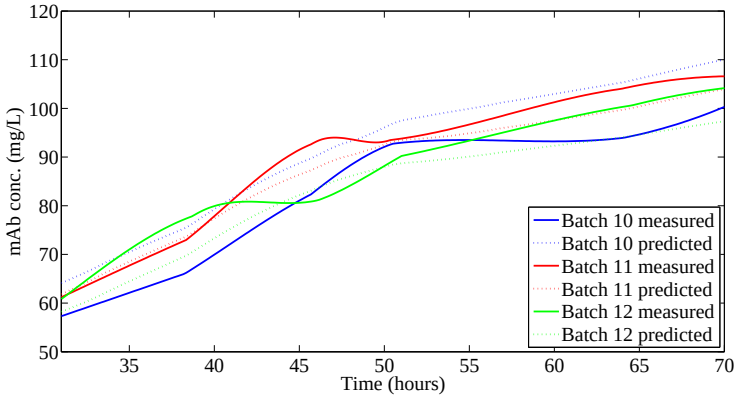


Figure C36: Measured and predicted product titre for sparger (vvm) experiments (batches 10-12) for model B.

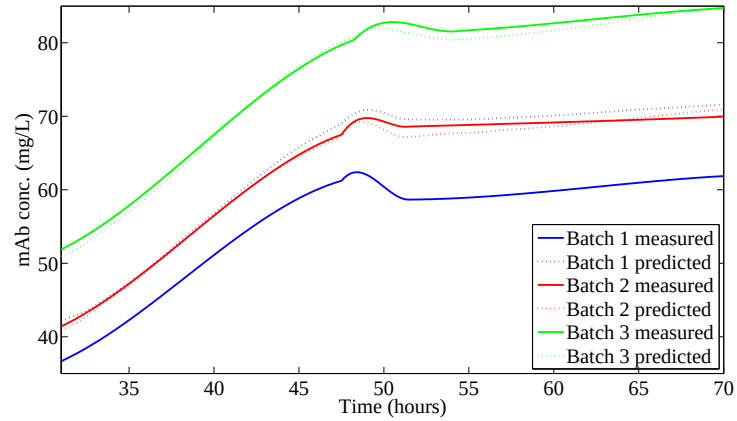


Figure C37: Measured and predicted product titre for dissolved oxygen experiments (batches 1-3) for model C.

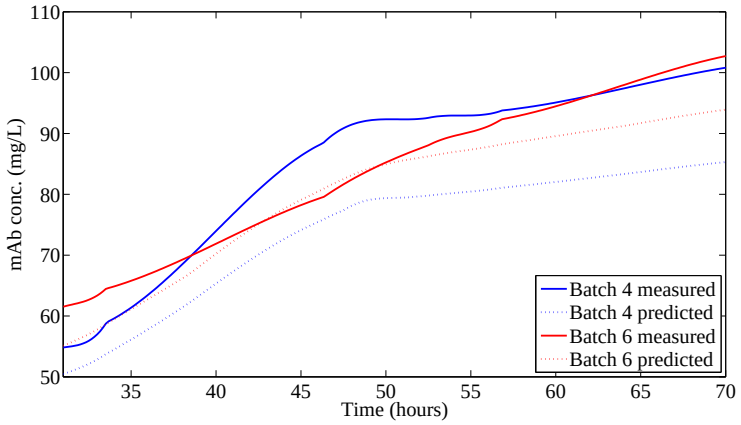


Figure C38: Measured and predicted product titre for osmolality experiments (batches 4 and 6) for model C.

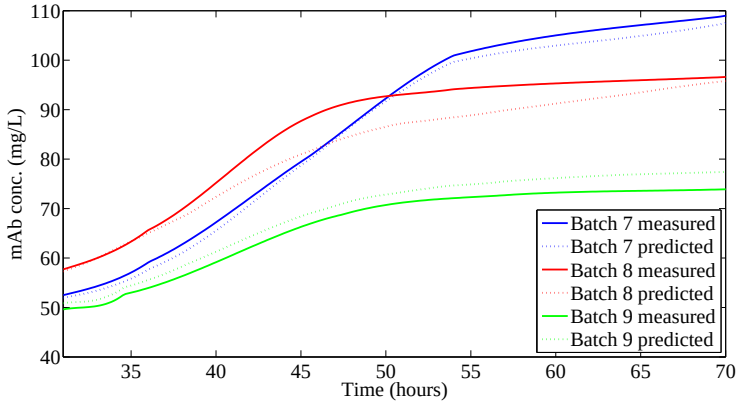


Figure C39: Measured and predicted product titre for pH experiments (batches 7-9) for model C.

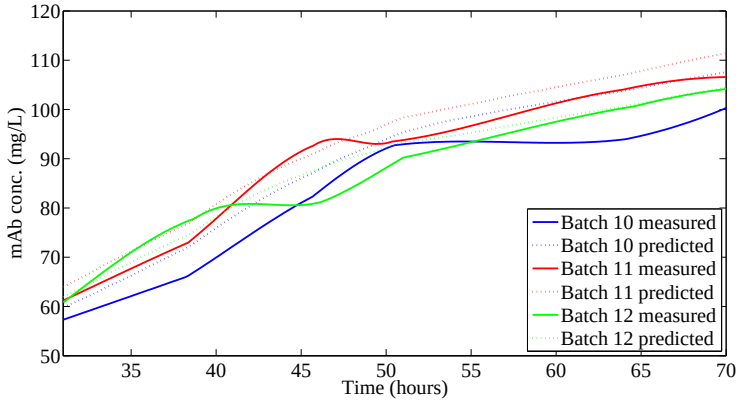


Figure C40: Measured and predicted product titre for sparger (vvm) experiments (batches 10-12) for model C.

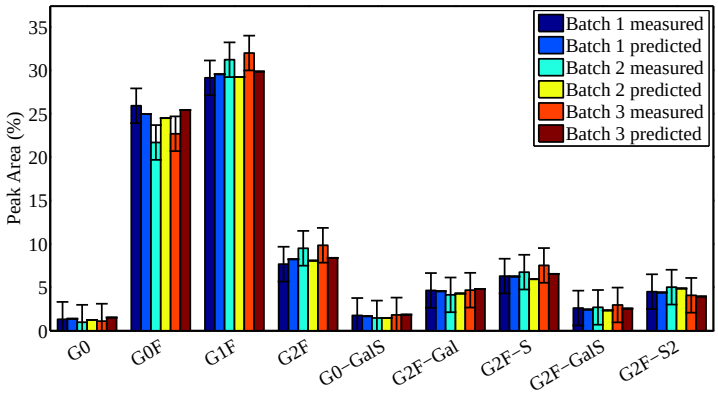


Figure C41: Measured and predicted product glycosylation profile for dissolved oxygen experiments (batches 1-3) for model A.

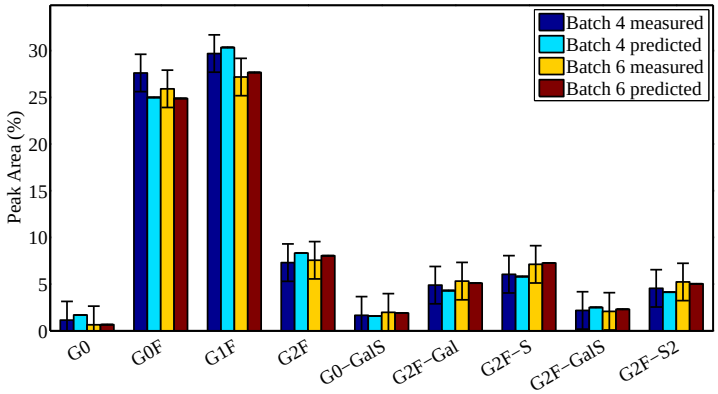


Figure C42: Measured and predicted product glycosylation profile for osmolality experiments (batches 4 and 6) for model A.

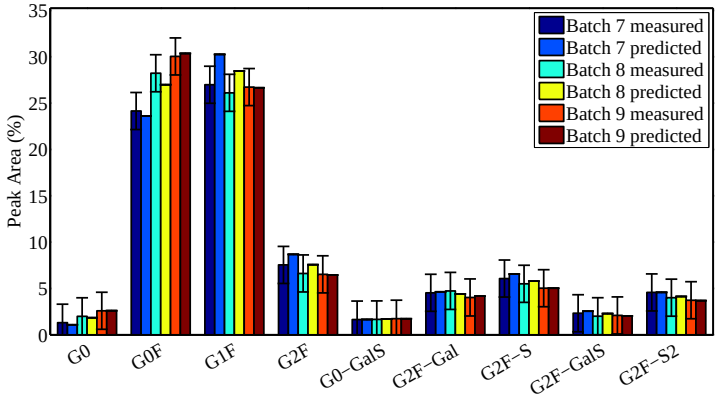


Figure C43: Measured and predicted product glycosylation profile for pH experiments (batches 7-9) for model A.

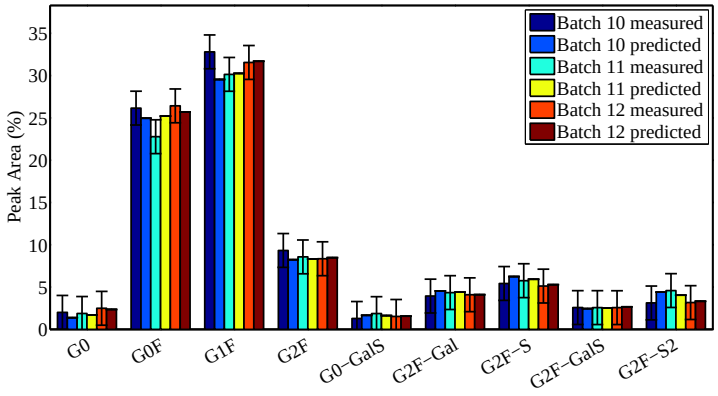


Figure C44: Measured and predicted product glycosylation profile for sparger (vvm) experiments (batches 10-12) for model A.

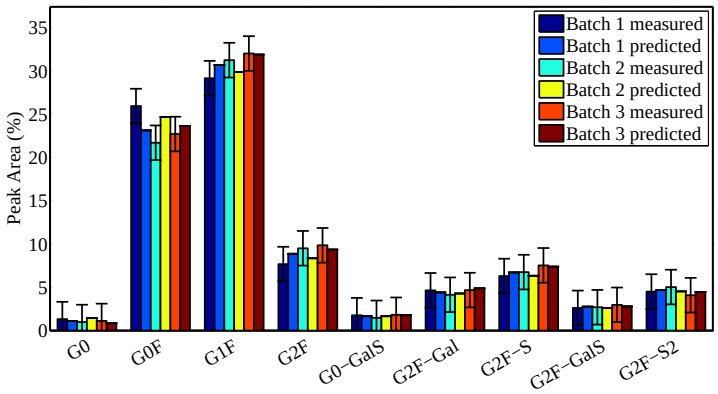


Figure C45: Measured and predicted product glycosylation profile for dissolved oxygen experiments (batches 1-3) for model B.

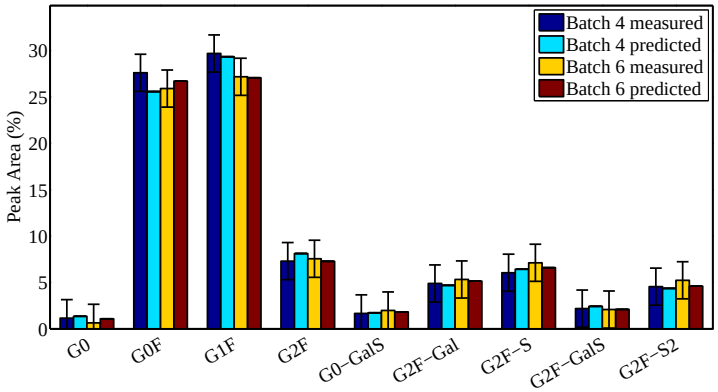


Figure C46: Measured and predicted product glycosylation profile for osmolality experiments (batches 4 and 6) for model B.

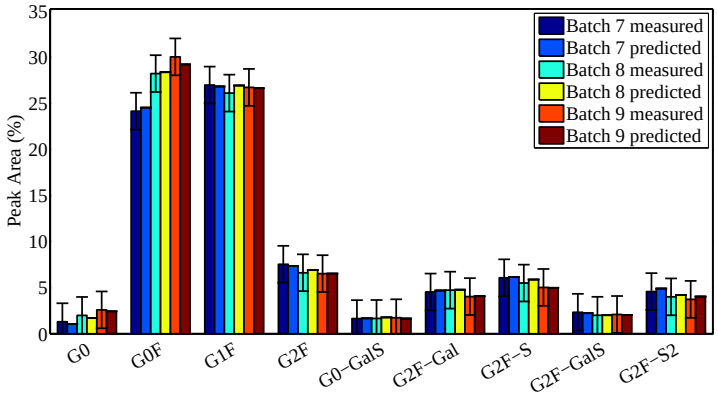


Figure C47: Measured and predicted product glycosylation profile for pH experiments (batches 7-9) for model B.

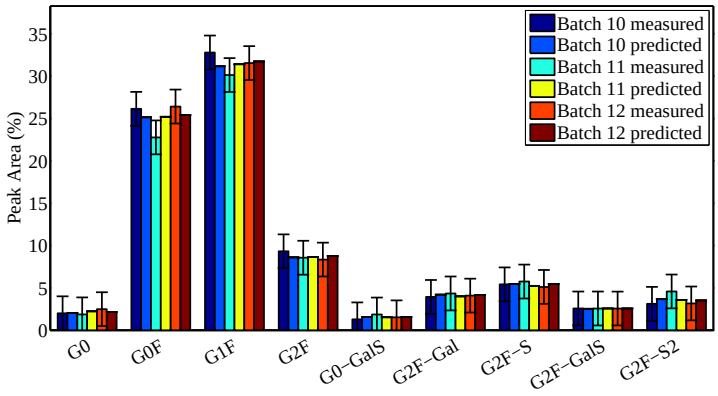


Figure C48: Measured and predicted product glycosylation profile for sparger (vvm) experiments (batches 10-12) for model B.

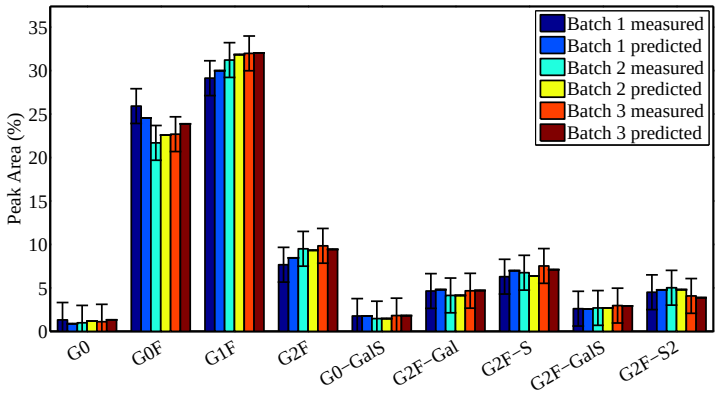


Figure C49: Measured and predicted product glycosylation profile for dissolved oxygen experiments (batches 1-3) for model C.

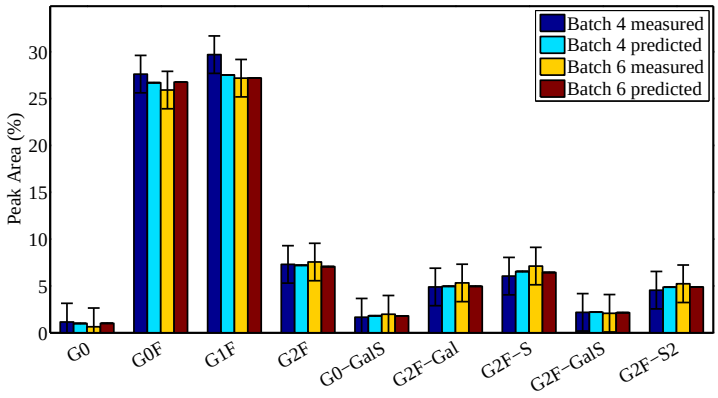


Figure C50: Measured and predicted product glycosylation profile for osmolality experiments (batches 4 and 6) for model C.

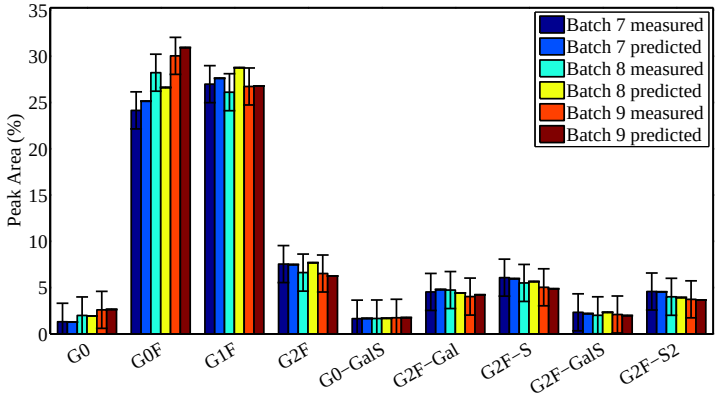


Figure C51: Measured and predicted product glycosylation profile for pH experiments (batches 7-9) for model C.

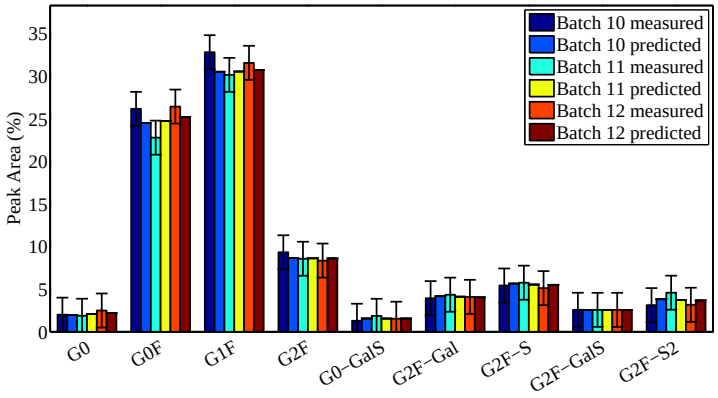


Figure C52: Measured and predicted product glycosylation profile for sparger (vvm) experiments (batches 10-12) for model C.

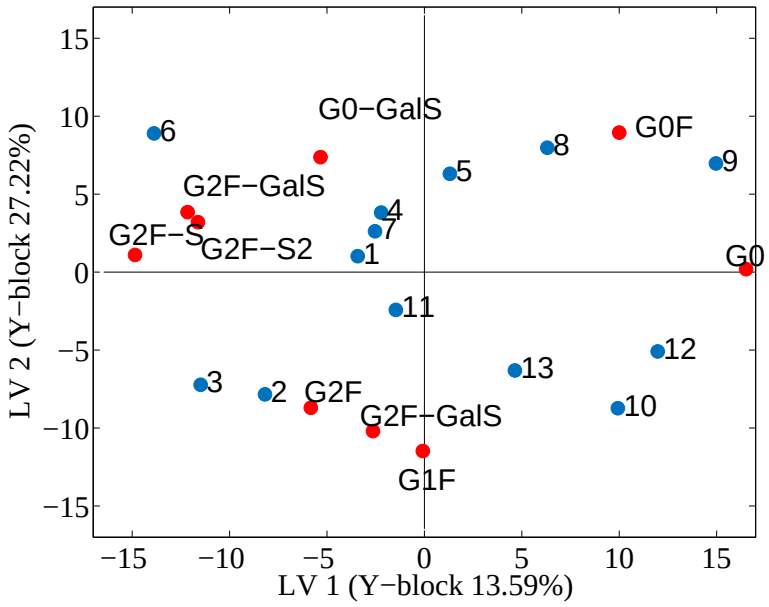


Figure C53: Bi-plot for LV1 and LV2 for Y-block data for model A to product glycosylation profile.

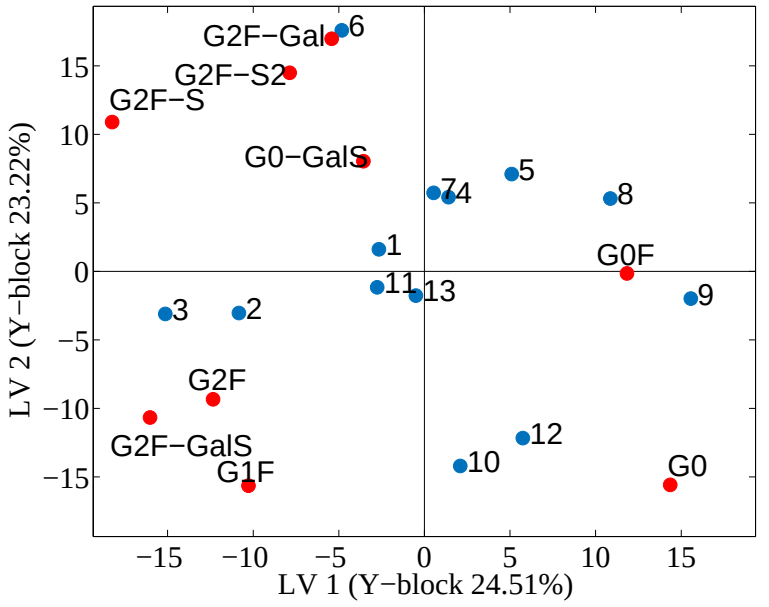


Figure C54: Bi-plot for LV1 and LV2 for Y-block data for model B to product glycosylation profile.

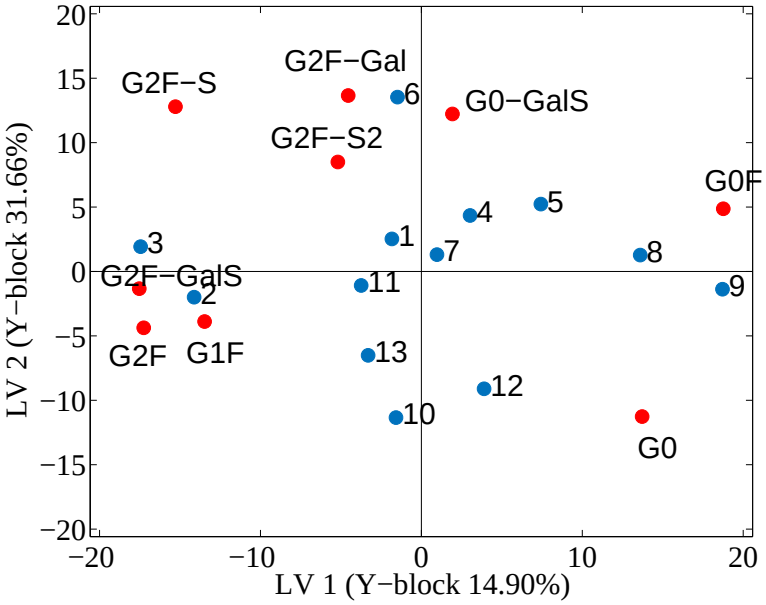


Figure C55: Bi-plot for LV1 and LV2 for Y-block data for model C to product glycosylation profile.

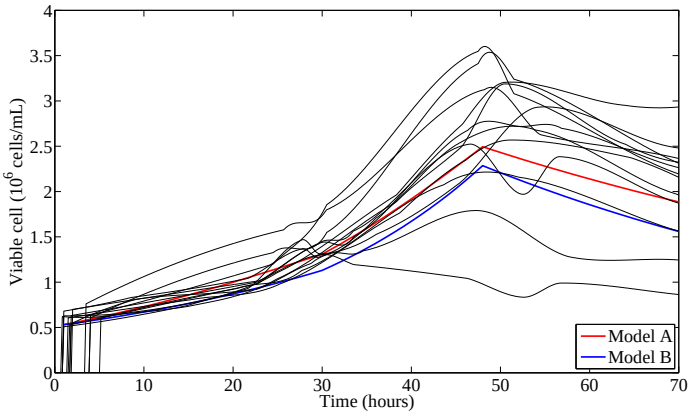


Figure C56: Measured and prediction data for viable cell count for all batches. The predictions shown are the first principles model derived from the Naderi equations (FP model 1) and the first principles model derived from the Kontoravdi equations (FP model 2).

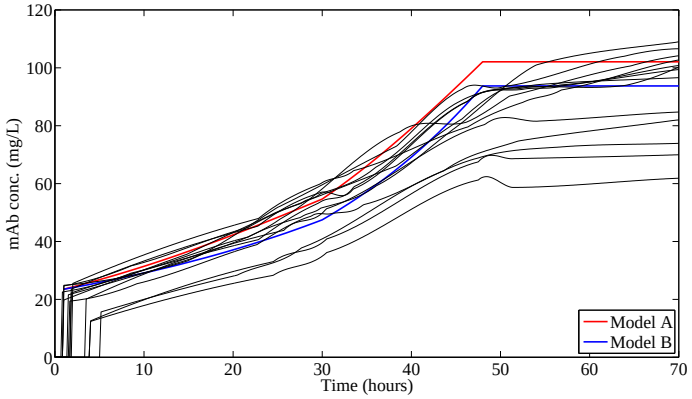
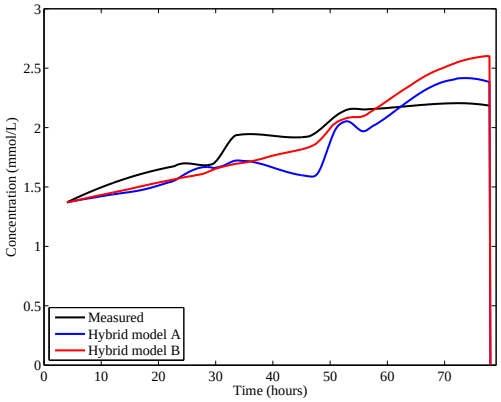
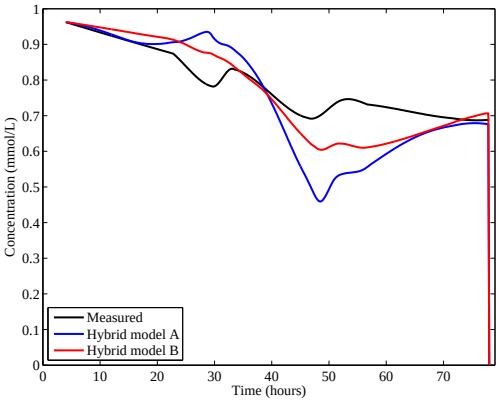


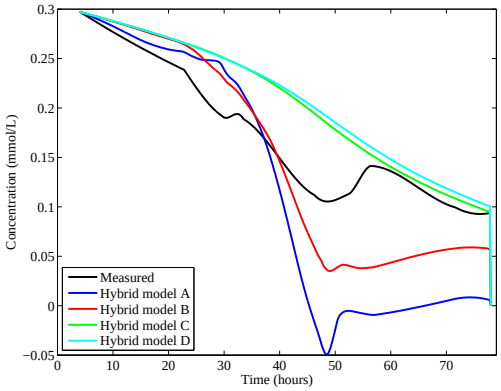
Figure C57: Measured and prediction data for product titre for validation batches. The predictions shown are the first principles model derived from the Naderi equations (FP model 1) and the first principles model derived from the Kontoravdi equations (FP model 2).



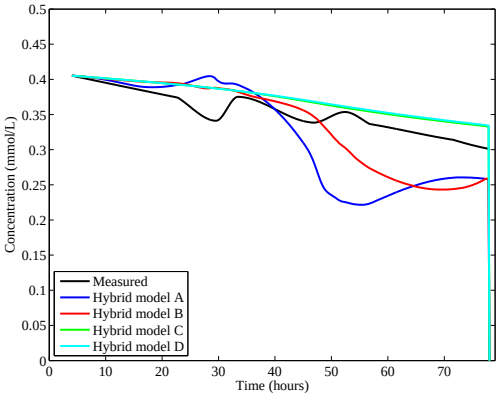
(a) Alanine



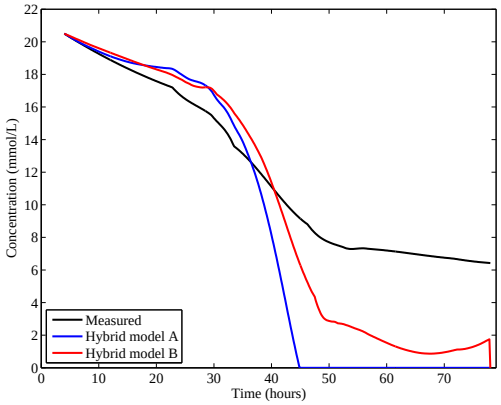
(b) Arginine



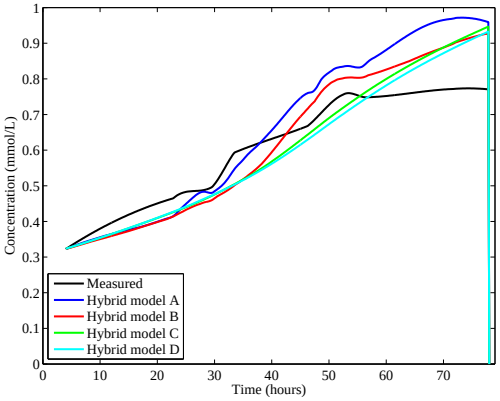
(c) Asparagine



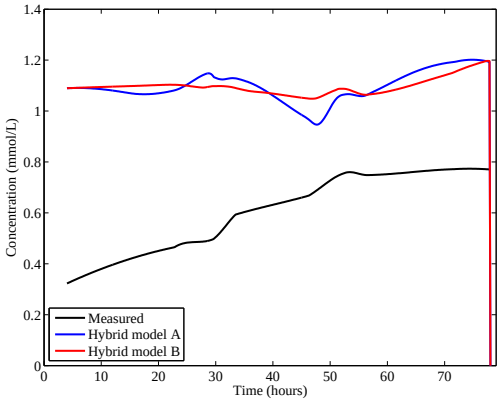
(d) Aspartic acid



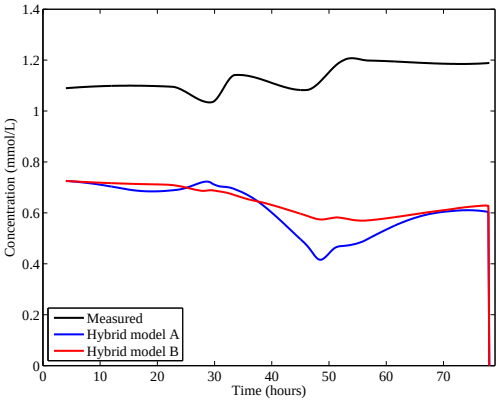
(e) Cysteine



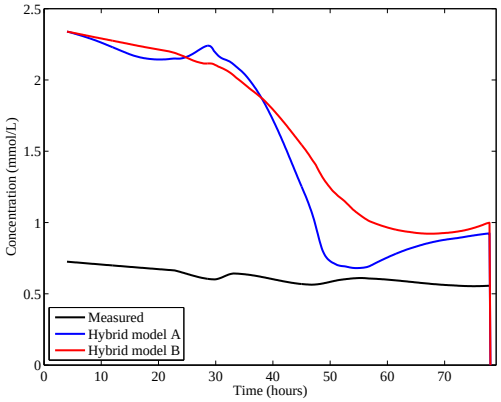
(f) Glutaminc acid



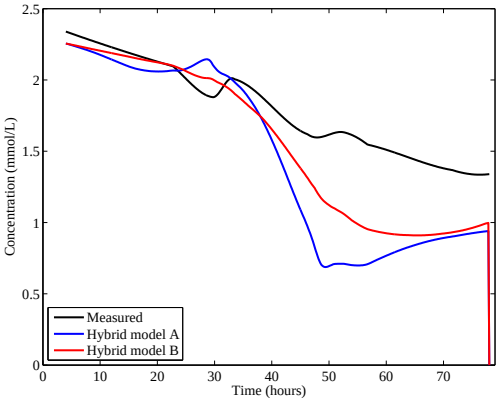
(g) Glycine



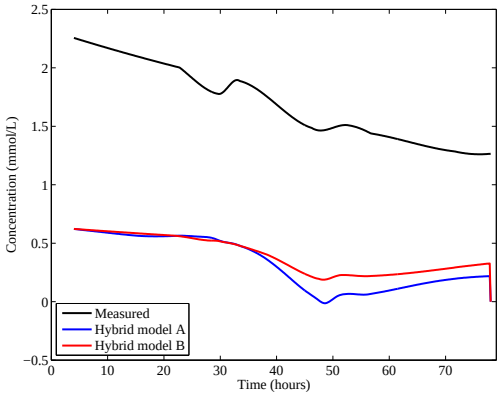
(h) Histidine



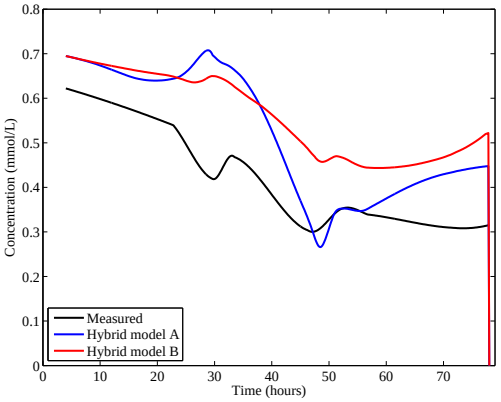
(i) Isoleucine



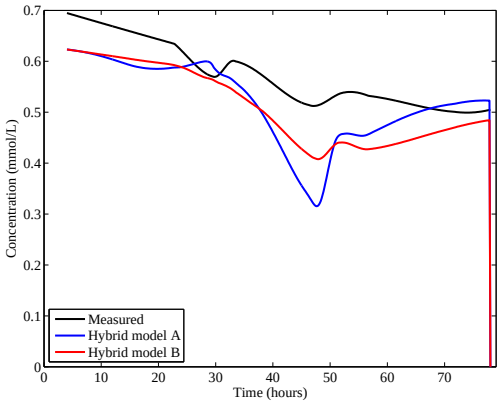
(j) Leucine



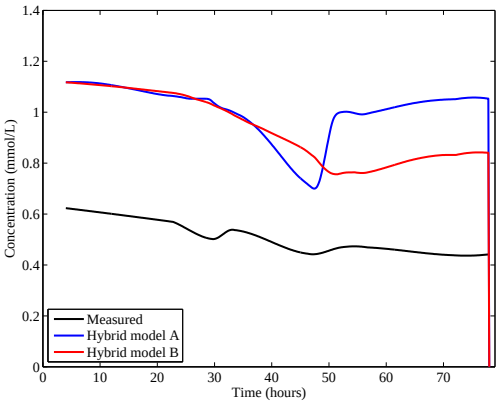
(k) Lysine



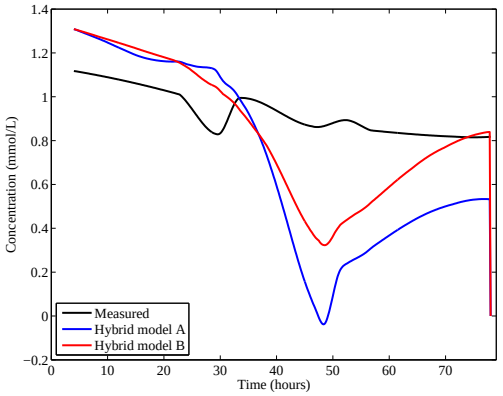
(l) Methoionine



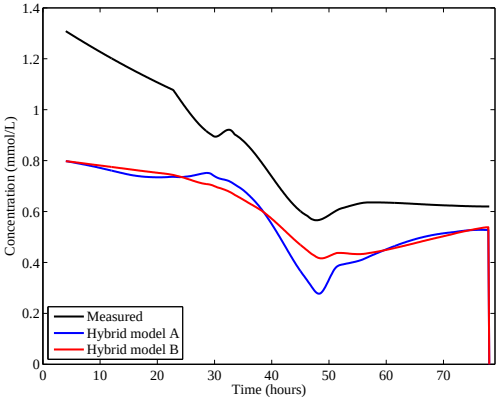
(m) Phenylalanine



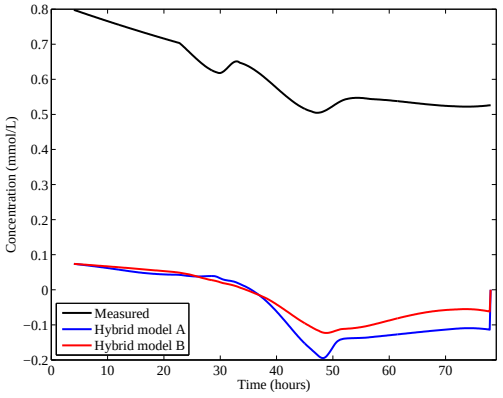
(n) Proline



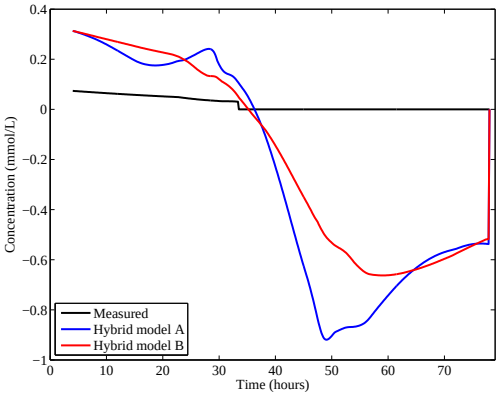
(o) Serine



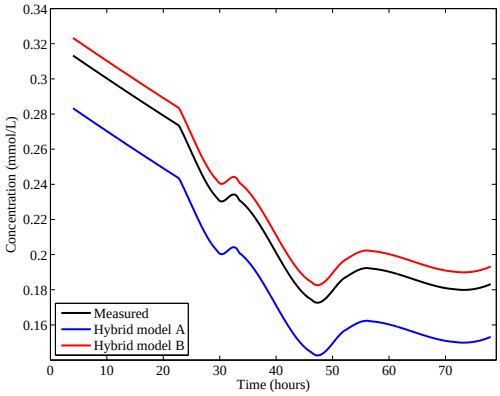
(p) Threonine



(q) Tryptophan

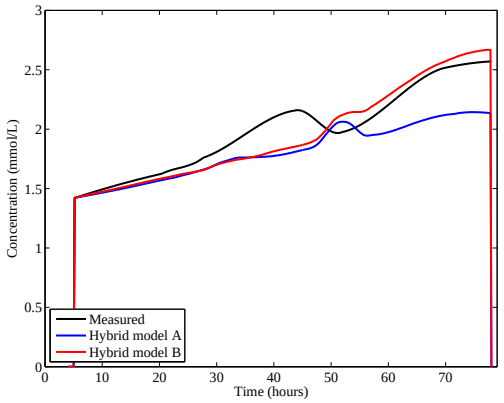


(r) Tryosine

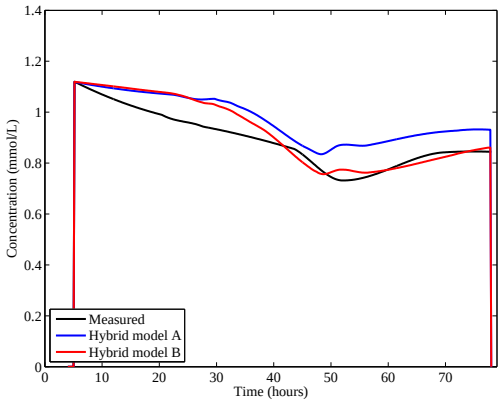


(s) Valine

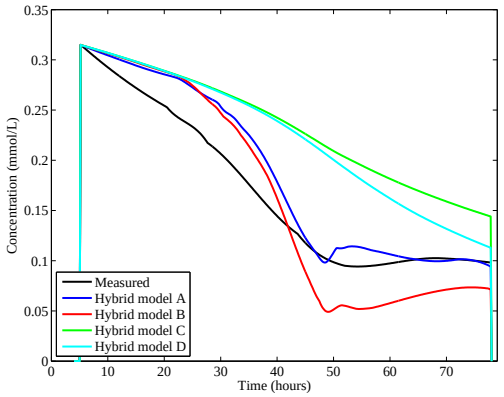
Figure C56: Predictions for metabolites from four hybrid models for batch 5. Hybrid model one uses on-line X-block data and Kontoravdi *et al.* (2007) ODEs, hybrid model two uses operational parameter X-block data and Kontoravdi *et al.* (2007) ODEs, hybrid model three uses on-line X-block data and Naderi *et al.* (2011) ODEs, and hybrid model four uses operational parameter X-block data and Naderi *et al.* (2011) ODEs. The corresponding model assessment values are reported in Table 5.5.



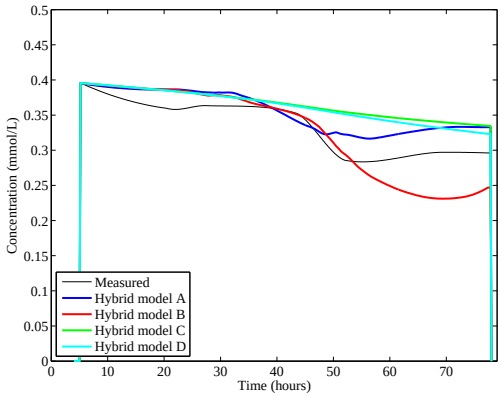
(a) Alanine



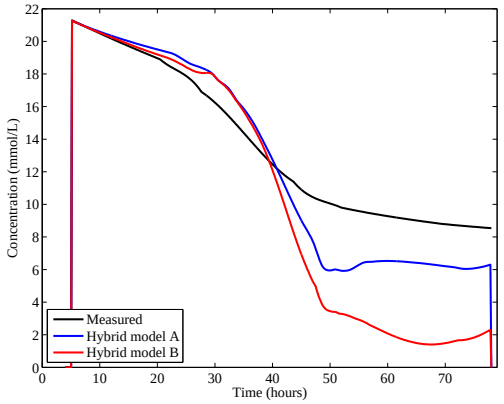
(b) Arginine



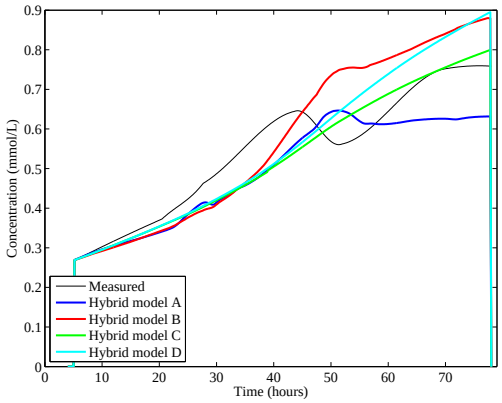
(c) Asparagine



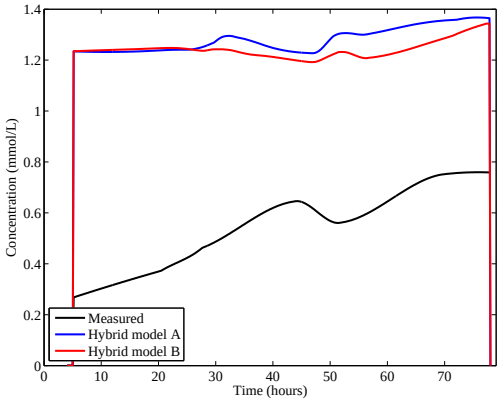
(d) Aspartic acid



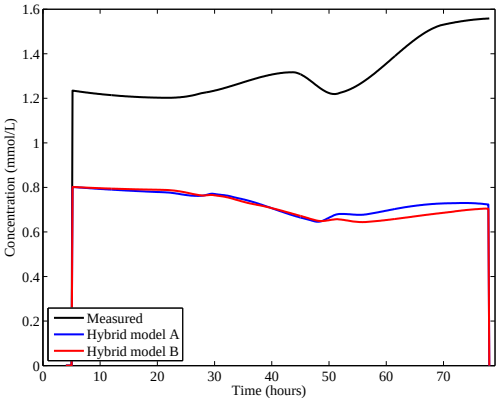
(e) Cysteine



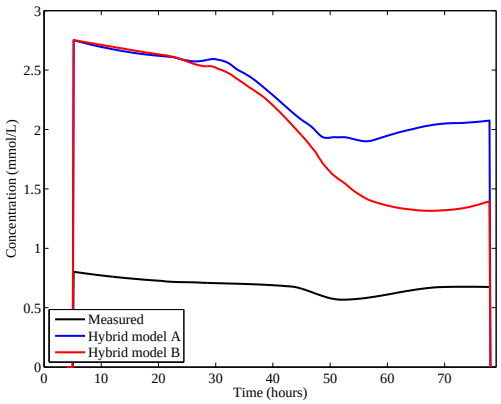
(f) Glutaminc acid



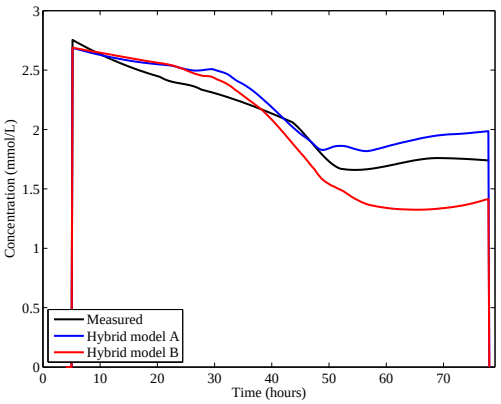
(g) Glycine



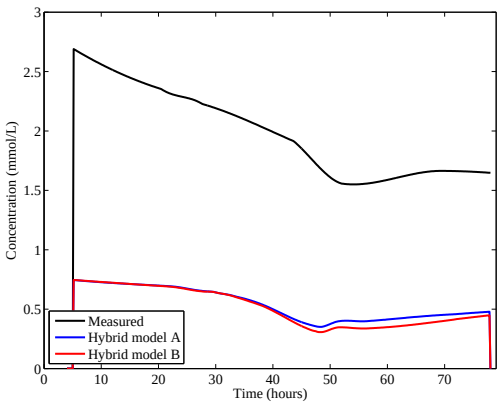
(h) Histidine



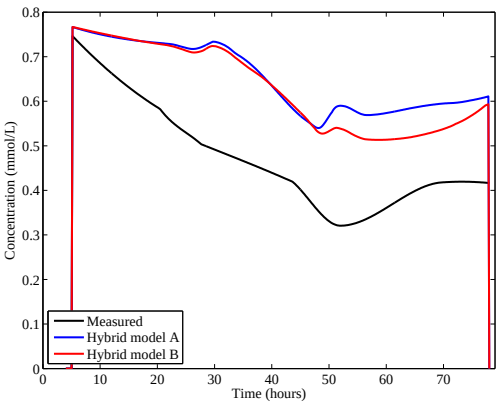
(i) Isoleucine



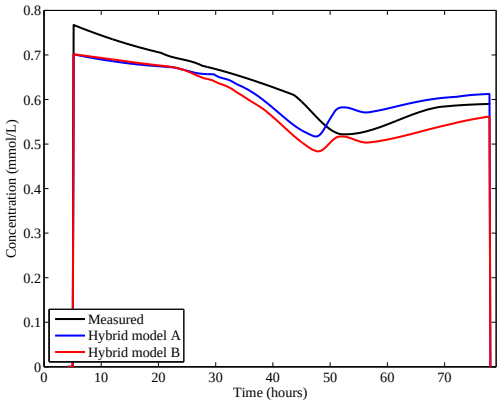
(j) Leucine



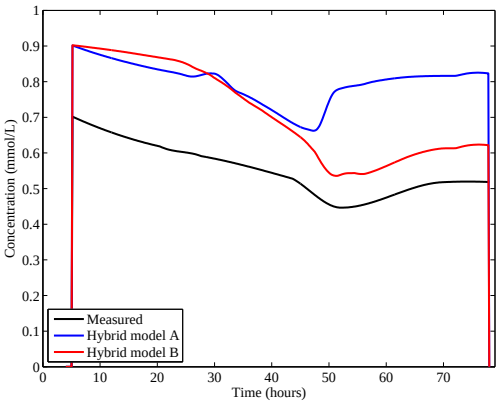
(k) Lysine



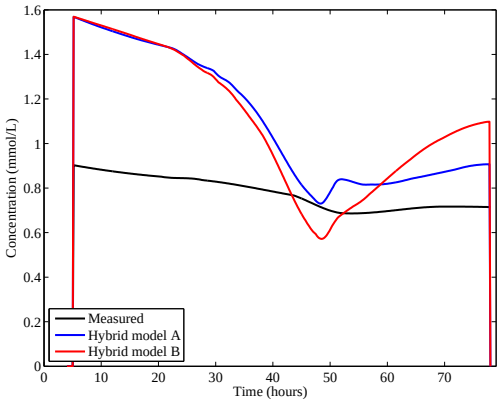
(l) Methoionine



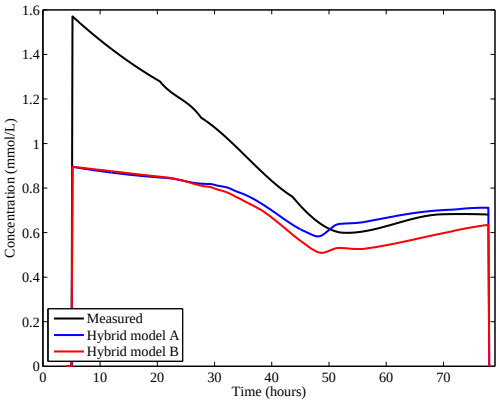
(m) Phenylalanine



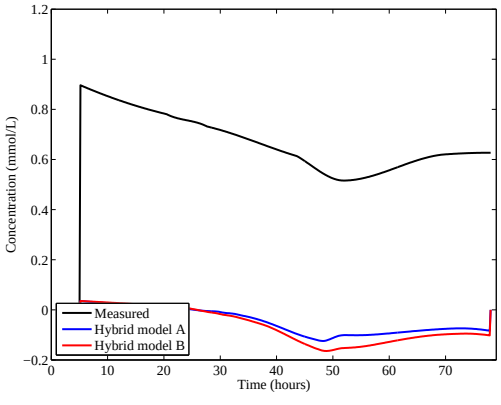
(n) Proline



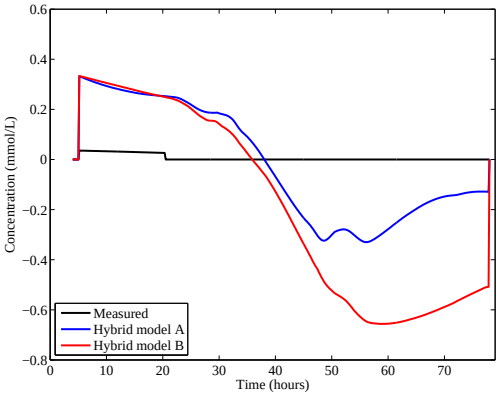
(o) Serine



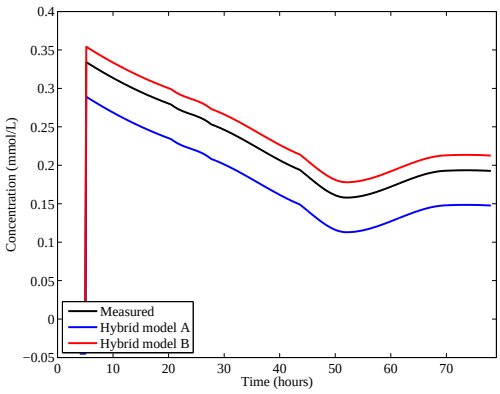
(p) Threonine



(q) Tryptophan



(r) Tryosine



(s) Valine

Figure C55: Predictions for metabolites from four hybrid models for batch 13. Hybrid model one uses on-line X-block data and Kontoravdi *et al.* (2007) ODEs, hybrid model two uses operational parameter X-block data and Kontoravdi *et al.* (2007) ODEs, hybrid model three uses on-line X-block data and Naderi *et al.* (2011) ODEs, and hybrid model four uses operational parameter X-block data and Naderi *et al.* (2011) ODEs. The corresponding model assessment values are reported in Table 5.5.

Appendix D

Table B1: Run information for 15 experiments comprising the DoE, included are the process parameter set points and the off-line recorded peak data.

Run identifier	Flow rate	Load pH	Elution CV	Load concentration	Elution pH	Retention time	Peak area	Peak height	Peak width
1	1	7	12	20	7	23.35	3883.66	1828.41	7.35
2	1.5	8	8	10	8	14.09	469.07	421.97	3.47
3	1.5	8	8	30	8	15.16	4107.62	2541.96	4.60
4	0.5	8	16	30	8	50.15	9174.28	2441.52	9.93
5	0.5	8	8	10	8	41.74	1587.07	574.25	8.49
6	1	7	12	20	7	23.54	2152.11	1136.84	6.64
7	1.5	6	16	30	6	17.34	3641.47	1697.60	7.39
8	1.5	6	8	10	6	14.41	123.93	118.18	2.91
9	0.5	6	16	30	6	51.95	10220.59	2208.52	12.03
10	0.5	6	16	10	6	49.51	402.87	120.86	9.30
11	1.5	6	16	10	8	16.43	127.71	89.28	3.72
12	0.5	6	8	30	8	46.11	5326.82	2923.32	5.95
13	0.5	8	8	30	6	44.99	7754.81	2913.67	7.05
14	1.5	8	16	10	6	16.28	40.80	28.25	3.35
15	1	7	12	20	7	23.48	2722.93	1386.37	7.09

Table B2: Off-line yield data calculated for the 15 experiments in the DoE, yields were determined by recording the UV at each stage.

Run identifier	Wash (yield %)	Elution (yield %)	Strip (yield %)	Caustic (yield %)	Total (yield %)
1	0.18	67.64	3.99	0.17	*71.98*
2	0.30	6.60	18.77	2.04	*27.70*
3	1.00	83.65	6.24	1.13	*92.03*
4	0.38	81.07	0.10	0.66	*82.21*
5	0.15	6.37	20.61	1.28	*28.41*
6	0.17	35.41	2.46	0.27	*38.31*
7	-0.15	64.58	0.14	-0.11	*64.46*
8	-1.15	1.12	6.89	1.08	*7.94*
9	-0.29	63.16	-0.04	0.14	*62.97*
10	0.12	7.85	0.06	0.23	*8.28*
11	-0.13	7.19	0.55	0.17	*7.79*
12	0.10	95.29	2.32	1.21	*98.92*
13	0.39	94.06	2.14	1.04	*97.64*
14	0.34	1.85	-0.10	-0.12	*1.96*
15	0.11	44.18	2.96	0.32	*47.56*

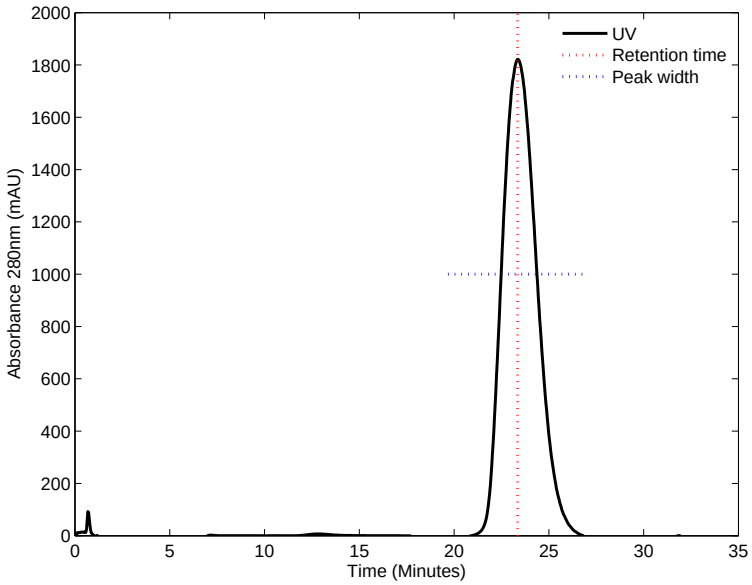


Figure D1: UV absorbance (280nm) for run one (Table B1) showing the elution peak to be isolated along with the retention time and peak width.

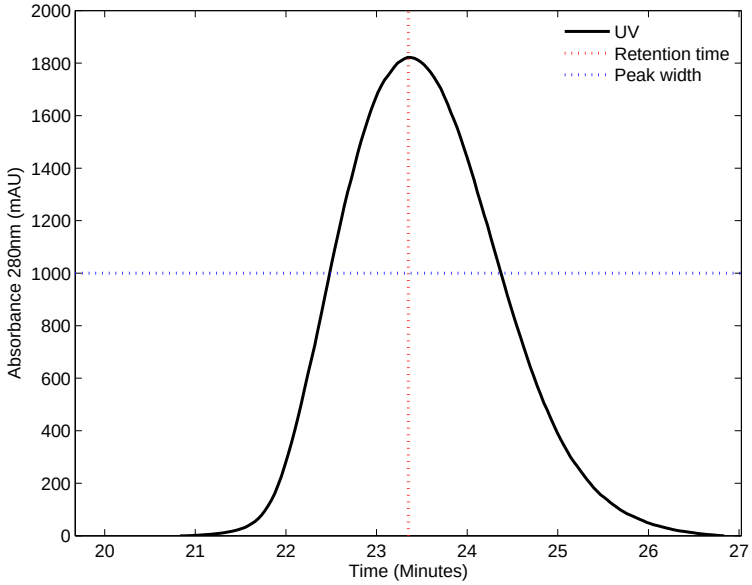


Figure D2: UV absorbance (280nm) for run one (Table B1) showing the isolated elution peak, peak isolation was determined based upon the shown retention time and peak width.

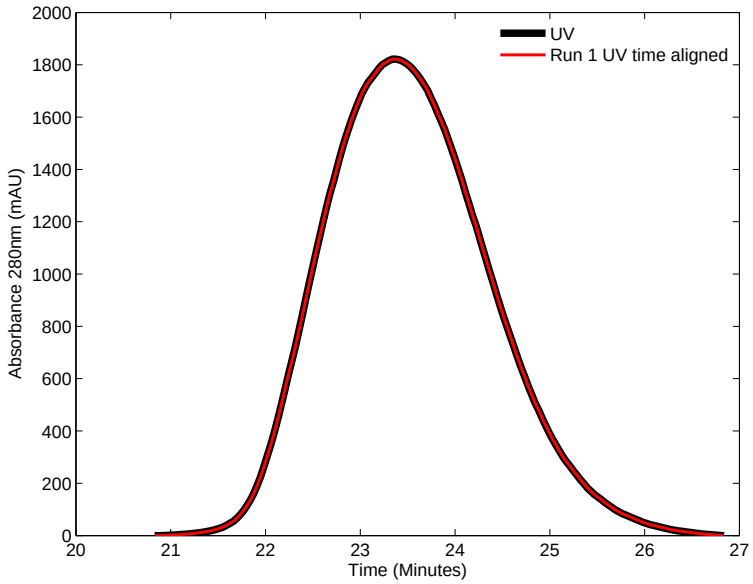


Figure D3: UV absorbance (280nm) for run one (Table B1) showing the isolated elution peak. The black line shows the original data points, the red line shows the interpolated data points. As can be seen the interpolation does not effect the UV trace.

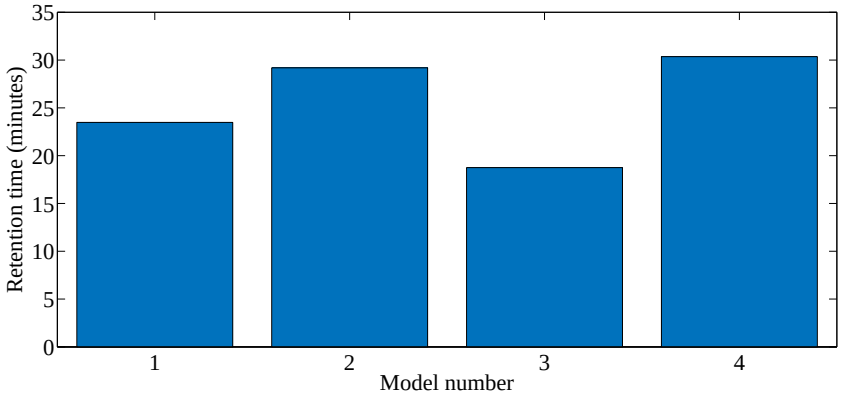


Figure D4: Predictions for retention time of batch 15. Column one is the original measured data. Column 2 is a PLS model (2 LVs) constructed using the operating parameters form the DoE as the X-block data. Column three is a PLS model (5 LVs) constructed using the conductivity, concentration, and pH as the X-block data. Column four is a PLS model (4LVs) constructed using the absorbance as the X-block data.

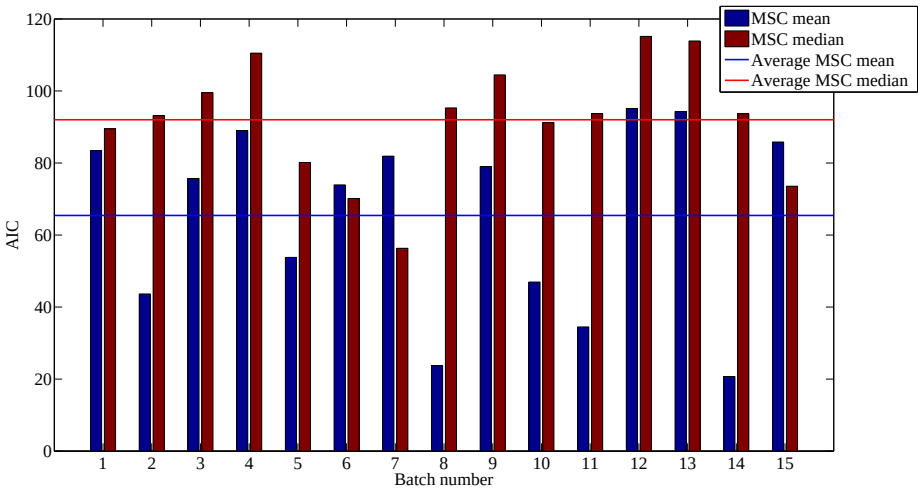


Figure D5: AIC values for the training batches in models 2 and 3 where MSC mean and MSC median techniques are applied respectively. The figure shows the variation between batches and how the mean is used as an overall measure of model fit.

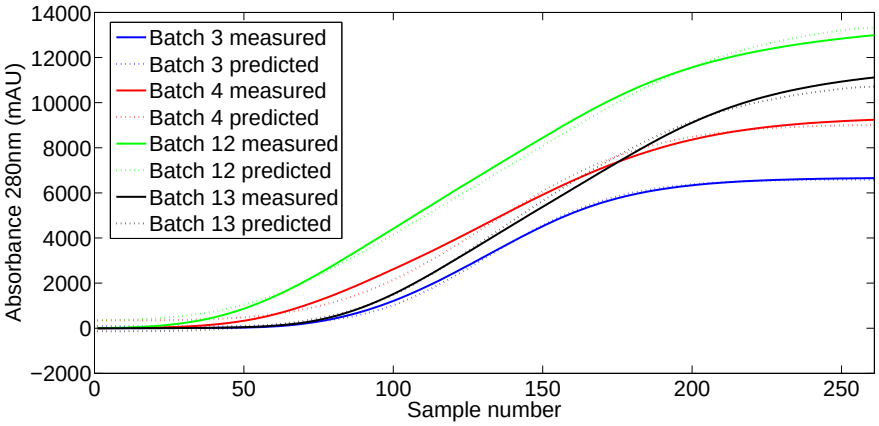


Figure D6: Measured and predict values for high yield batches for peak area PLS model containing 14 training batches and 1 validation batch.

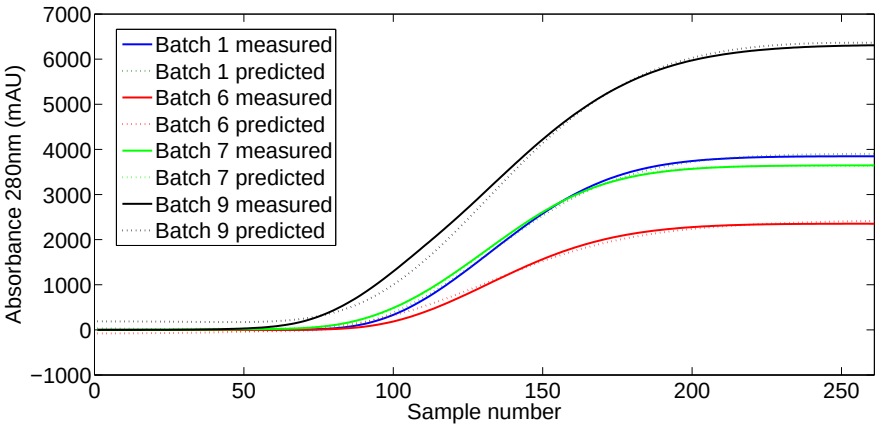


Figure D7: Measured and predict values for medium yield batches for peak area PLS model containing 14 training batches and 1 validation batch.

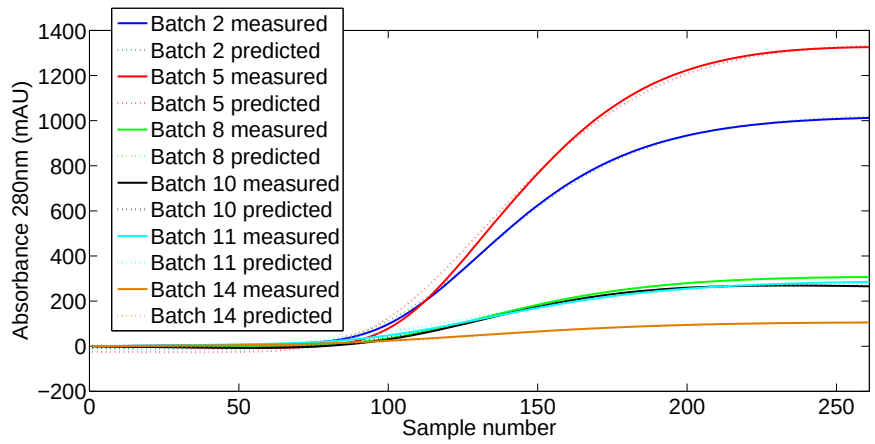


Figure D8: Measured and predict values for low yield batches for peak area PLS model containing 14 training batches and 1 validation batch.

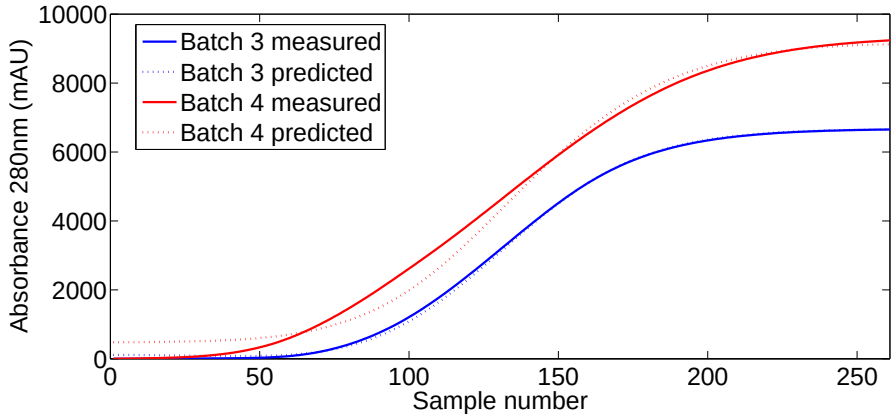


Figure D9: Measured and predict values for high yield batches for peak area PLS model containing 12 training batches and 1 validation batch.

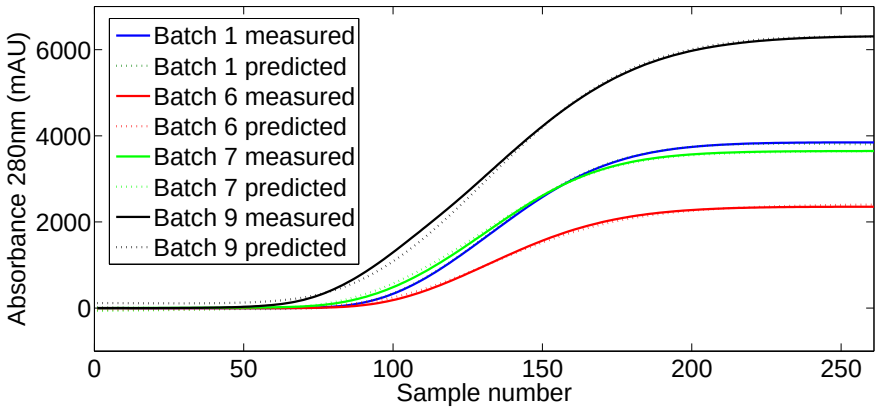


Figure D10: Measured and predict values for medium yield batches for peak area PLS model containing 12 training batches and 1 validation batch.

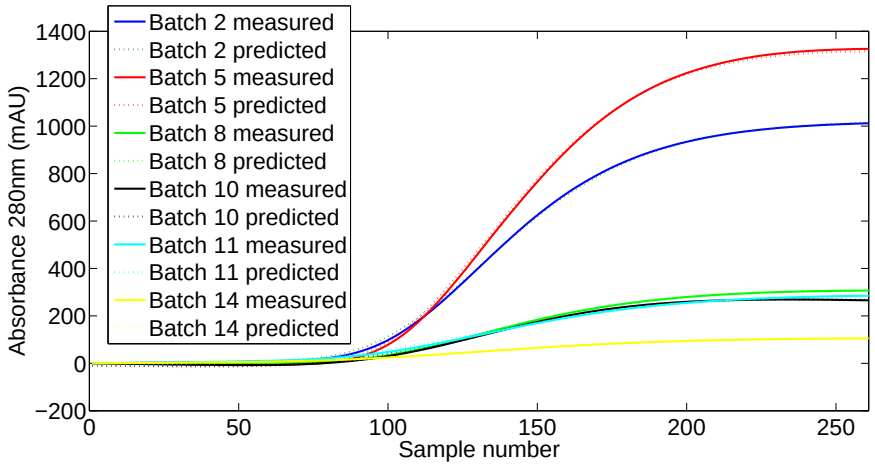


Figure D11: Measured and predict values for medium yield batches for peak area PLS model containing 12 training batches and 1 validation batch.

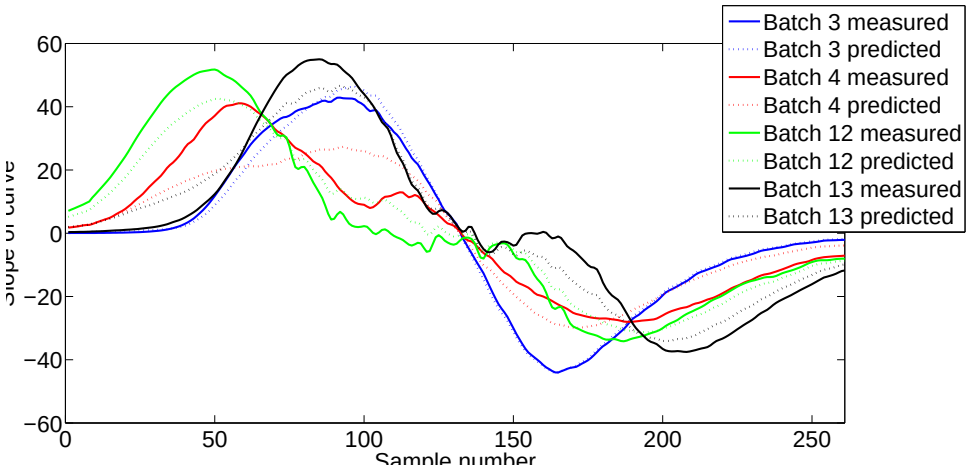


Figure D12: Measured and predict values for high yield batches for slope gradient PLS model containing 14 training batches and 1 validation batch.

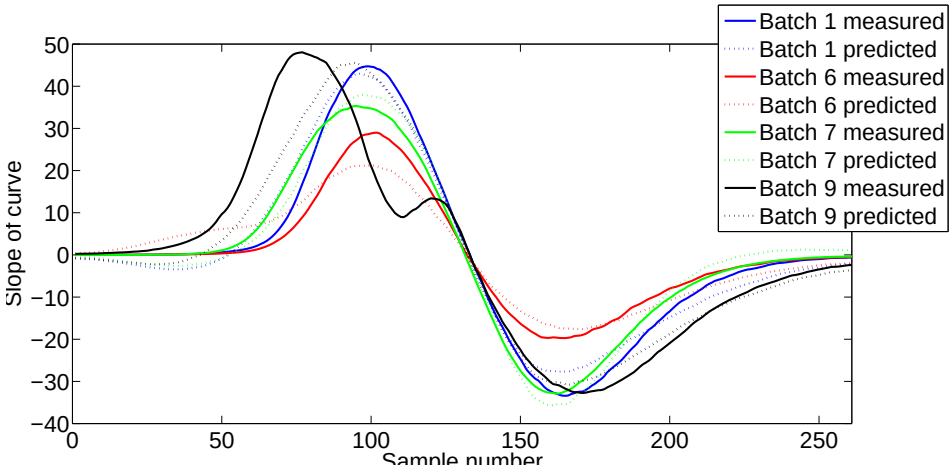


Figure D13: Measured and predict values for medium yield batches for slope gradient PLS model containing 14 training batches and 1 validation batch.

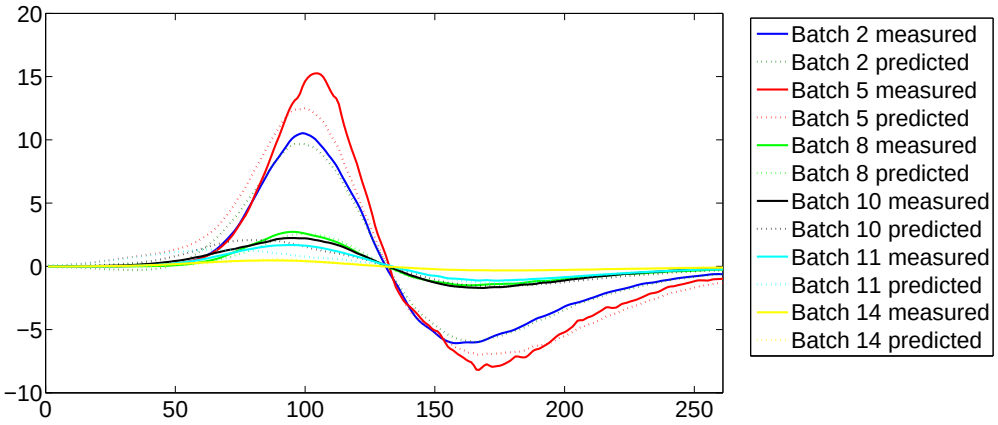


Figure D14: Measured and predict values for low yield batches for slope gradient PLS model containing 14 training batches and 1 validation batch.

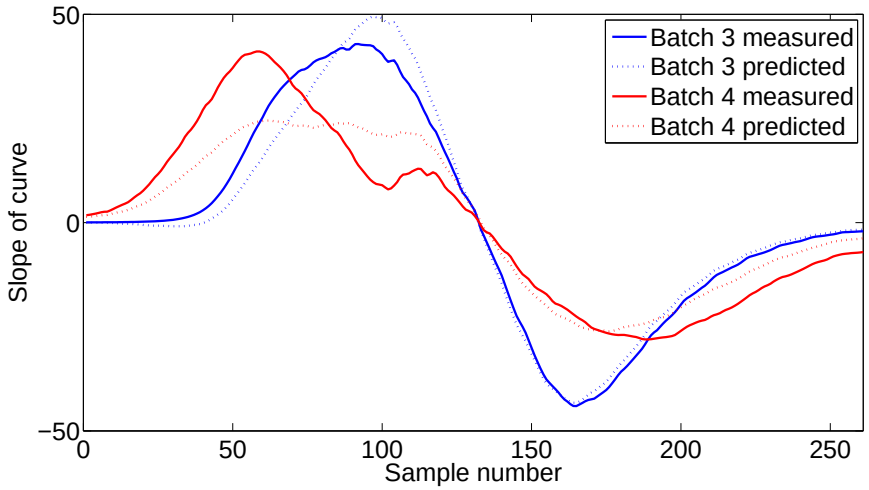


Figure D15: Measured and predict values for high yield batches for slope gradient PLS model containing 14 training batches and 1 validation batch.

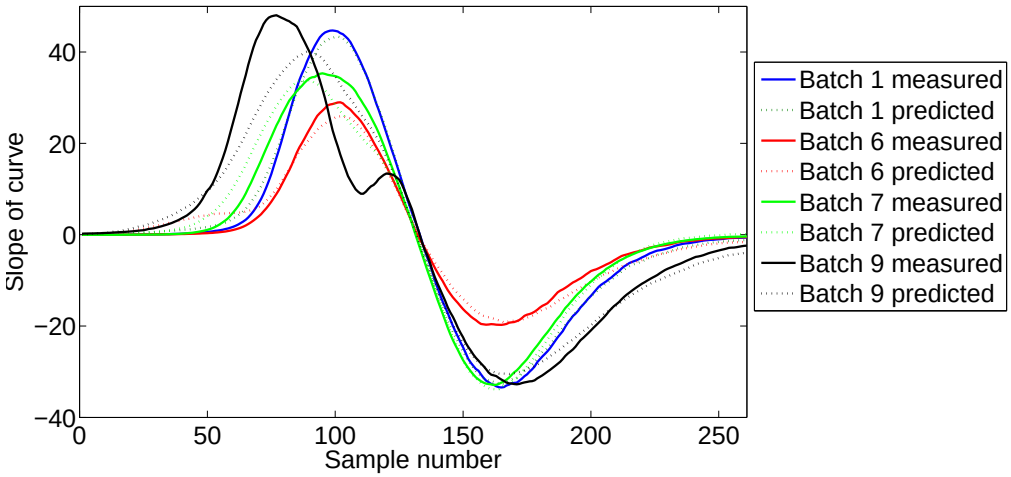


Figure D16: Measured and predict values for medium yield batches for slope gradient PLS model containing 14 training batches and 1 validation batch.

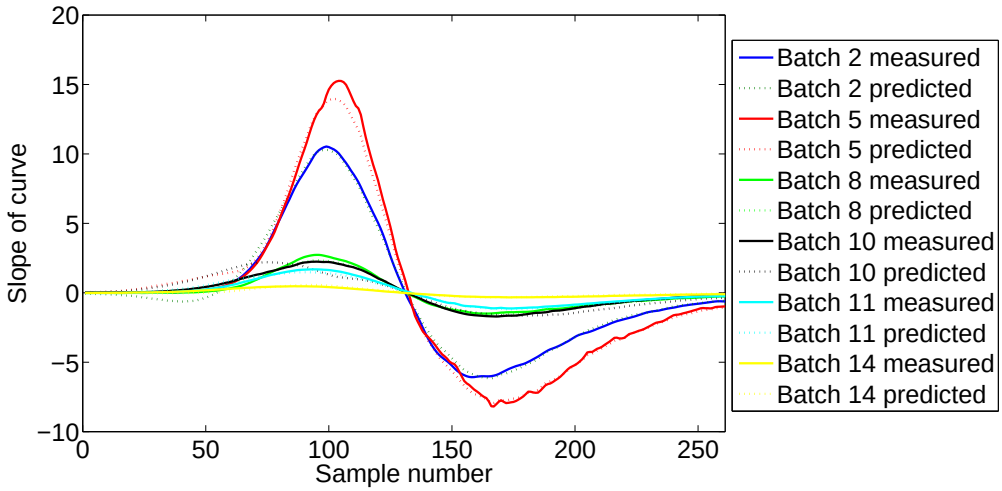


Figure D17: Measured and predict values for low yield batches for slope gradient PLS model containing 14 training batches and 1 validation batch.