Development of Soft Sensors for Monitoring of Chinese Hamster Ovary Cell Processes

by

Seyed Kaveh Ohadi

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

Waterloo, Ontario, Canada, 2014

AUTHOR'S DECLARATION

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

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

STATEMENT OF CONTRIBUTIONS

For Chapter 3 the author developed the data-driven soft sensors, performed the acquisition of the fluorescence data, and collaborated on the analytical measurements related to the culture variables. The cell cultivation work that was conducted in shake flasks was performed by Hengameh Aghamohseni.

For Chapter 5 the author formulated the glycosylation model, collaborated in the development of the nucleotide sugar model, coded the glycosylation and nucleotide sugar models in MATLAB, estimated the parameters of the nucleotide sugar and glycosylation models, connected three model-blocks and proposed an integration approach for accumulated glycosylation, and also tackled the issue of timely computation by trimming the network of reaction, enhancing the computer code, and adopting parallel calculations.

Abstract

The goal of this work was to develop monitoring techniques for use during the production of monoclonal antibodies (Mabs) in Chinese hamster ovary cell cultures. Such monitoring would enable real-time screening and control of key process variables both upstream and downstream so as to guarantee product quality and process consistency. The measurement techniques that are currently available are time and labor intensive and in some cases require frequent maintenance. Thus, they are not suitable for fast online monitoring of bioprocesses. Thus, with a goal of future real-time implementation, data-driven (empirical models) and model-driven (mechanistic models) soft sensors were developed.

The bioreactor is the key component of the upstream manufacturing phase. Continuous monitoring and control of this unit is critical in order to maximize production of the Mab with a desired quality (i.e. glycosylation pattern). Data-driven soft sensors were developed using intrinsic multi-wavelength fluorescence spectra of the culture broth in combination with partial least square regression (PLSR) for tracking viable cell, dead cell, recombinant protein, glucose, and ammonia concentrations. To better elucidate the relationship between the fluorescence spectra and process operating conditions, trajectories of fluorophore-peaks over the course of the culture were investigated and compared to changes in key process variables prior to model development. The proposed soft sensors were capable of predicting the aforementioned process variables with high accuracy.

To enhance the extrapolation accuracy of the data-driven soft sensor outside of the region of operating conditions used for model calibration and to better track the dynamics of the culture, an extended Kalman filter (EKF) was developed based on a combination of mechanistic and empirical models. To address the structural and parameter uncertainty of the models, non-stationary disturbances were introduced to the model through parameter adaptation. The resulting EKF-based soft sensor's predictions surpassed the accuracy of a standalone fluorescence based soft sensor and was capable of tracking process dynamics in between sampling instances with high precision.

N-linked glycosylation has a significant impact on the therapeutic properties of Mab and is an important quality attribute that is associated with the extracellular metabolic state of the culture. Based on the primary investigation it was revealed that the fluorescence spectroscopy is not capable of accurately tracking the glycosylation profile of the Mab. Thus, to enhance the controllability of the glycoprofile, a novel dynamic model was developed that relates the extracellular culture conditions to

the accumulated glycosylation pattern of Mab produced through the production of nucleotide sugars required for N-linked glycosylation in the Golgi apparatus. The model parameters were estimated using the experimental data. The resulting model was capable of accurately predicting the glycosylation extent in the form of a galactosylation index as well as individual glycan structures.

Another area of application of fluorescence was for monitoring protein aggregation. During downstream processing proteins are exposed to stress factors such as changes in temperature, pH, or shear stress that can increase the propensity of Mab to aggregate. Aggregation can trigger undesirable impacts including an increased immunogenicity response in the patient. Therefore, developing an *in situ* technique for fast quality and quantity control of protein aggregation is of great industrial interest. Fluorescence-based soft sensors, in conjugation with PLSR, were developed for quality control (product classification) and for quantitative monitoring (Mab monomer concentration) at different process conditions that typically occur in different stages of the purification process. To better elucidate the impact of stress factors on the degree of aggregation and identify the operating conditions for which the propensity to aggregate is minimal, a surface response model was fitted to the data prior to soft sensor development. The soft sensors were capable of accurately predicting monomer concentration of samples exposed to different levels of stress factors as well as for classifying the final product into different groupings according to their relative aggregation levels.

Acknowledgements

I would like to express my gratitude and appreciation to my supervisors: Drs. Hector Budman and Raymond Legge for their continual support and guidance. Both were tremendous mentors whose directions and encouragement during all these years allowed me to grow.

In addition, I would like to thank Drs. Thomas Duever, Ali Elkamel, William Taylor, and Carl Duchesne for agreeing to be part of my PhD defense.

I want to thank all of my friends, especially Navid, and colleagues, Hengameh, Ali, Jasdeep, Divya, and Vanessa at the Department of Chemical Engineering of University of Waterloo for all their help, support, and valuable hints.

In the end, I would like to extend my heartfelt appreciation to my lovely wife, Hadis, and my parents for their endurance, support, and encouragements especially during difficult times. Moreover, I like to appreciate the help and support of my in-laws during years of PhD studies.

Dedication

This dissertation is dedicated to my precious wife Hadis Zarrin whose unconditional love, encouragement, and support made it possible for me to finish this work. She has been the light and inspiration of my life during all difficulties.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Motivations

The biopharmaceutical industry has undergone substantial growth over the past few decades with recombinant proteins outpacing other biopharmaceutical products (Butler 2005; Butler and Meneses-Acosta 2012; Durocher and Butler 2009). Since the approval of insulin as the first recombinant protein, over 150 recombinant protein-based drugs have gained Food and Drug Administration's (FDA's) approval (Butler 2005; Butler and Meneses-Acosta 2012; Durocher and Butler 2009). Monoclonal antibodies (Mabs) constitute the dominant recombinant proteins in the biopharmaceutical market with their market value exceeding $100 billion in 2011 (Pais *et al.* 2014). A continuously growing market is forecasted for Mabs in the near future (Butler and Meneses-Acosta 2012; Pais *et al.* 2014). Among all the expression hosts for the commercial production of Mabs, Chinese hamster ovary (CHO) cells have been the preferable host. CHO cell lines are reported to grow very efficiently under single-cell suspension (Wurm 2004), are able to perform DNA transfection (Butler and Meneses-Acosta 2012; Wurm and Hacker 2011), can undergo human-like post-translational modification (Pais *et al.* 2014), and lack human viral-pathogens (Wurm and Hacker 2011). Glycosylation is the most important post-translational modification occurring in the Golgi apparatus inside the cell and it is believed to profoundly contribute to the Mab's therapeutic efficacy and its circulatory half-life (Hossler *et al.* 2007; Teixeira *et al.* 2009a).

To meet the growing demand for Mabs in the biopharmaceutical market, process analysis, control, and optimization are essential for scale-up and process development and to ensure product consistency. Process control requires real-time monitoring of process variables. Recent FDA (or other regulatory agencies) regulations have fostered biopharmaceutical manufacturers to adopt quality by design (QbD) and process analytical technology (PAT) guidelines to ensure product quality (Pais *et al.* 2014; Rathore *et al.* 2010; Read *et al.* 2010; Teixeira *et al.* 2009a). Time consuming, labor intensive, and expensive conventional measurement assays have hampered the ability to frequently monitor some key process variables thus motivating the adoption of alternative techniques such as soft sensors (Gernaey *et al.* 2012; Kadlec *et al.* 2009). Soft sensors can be classified into two classes (Kadlec *et al.* 2009): data-driven, which are based on empirical models, and model-driven, which are

1

based on mechanistic models. Electromagnetic spectroscopy based techniques have exhibited great potential for online monitoring of process variables due to their non-destructive, non-invasive, and informative properties (Lourenco *et al.* 2012; Teixeira *et al.* 2009a). Higher sensitivity and selectivity of fluorescence spectroscopy, as compared to vibrational techniques such as near infrared (NIR) and Raman spectroscopy, ranked this as a viable technique to be used for fast *in situ* monitoring of bioprocesses. The presence of intrinsically fluorescing compounds, which are sensitive to changes of their micro-environment, in cell culture's growth media as well as Mab's polypeptide chain reinforces the potential of this technique for real-time monitoring. The privilege of multi-wavelength fluorescence, over single wavelength, in encompassing various fluorophores is emphasized in the literature (Haack *et al.* 2004; Marose *et al.* 1998; Skibsted *et al.* 2001) and discussed in further detail in Chapter 3.

Due to the dependence of the fluoresence signal on the environment, it is possible to infer the dynamics of the cell culture process from the evolving spectra of the culture-broth (Li and Humphrey 1991; Ohadi *et al.* 2014a; Teixeira *et al.* 2011; Teixeira *et al.* 2009b). Chemometric methods are required to extract such correlations and generate empirical models (data-driven soft sensors) that relate multi-wavelength spectra to key process variables. Although this approach has been pursued for bacterial, yeast, and mammalian cell cultivation for upstream processes (Haack *et al.* 2004; Hagedorn *et al.* 2003; Jain *et al.* 2011; James *et al.* 2002; Li and Humphrey 1991; Teixeira *et al.* 2011; Teixeira *et al.* 2009b), a comprehensive soft sensor capable of predicting viable cell, dead cell, recombinant protein, and metabolites concentrations has not been developed. Thus, the development of such soft sensor(s) is one of the goals of the current work.

The data-driven soft sensors are developed using chemometric methods merely based on data and have no mechanistic insight into the process, and perform as a black-box. As a result they lack extrapolation accuracy when the process operating conditions deviate considerably from the conditions used for model calibration. On the other hand, properly calibrated mechanistic models, also referred as white-box models to contrast them with empirical or black-box models, can be utilized to forecast time-dependent changes of culture variables over a wide region of operating conditions. Metabolic flux analysis is a well-established approach used to develop dynamic models resembling the *in vivo* evolution of metabolic states (Zamorano *et al.* 2013). However, development of white-box models is taxing due to the complexities of mammalian cell metabolism (Kadlec and Gabrys 2009; Kadlec *et al.* 2009; Shioya *et al.* 1999) and the resulting large number of parameters

2

that need to be calibrated with data. Therefore, to suppress the individual drawbacks associated with white- and black-box models, a framework was pursued in this study that combines fluorescence based soft sensor with a dynamic metabolic model using a Kalman filter.

The quality of the antibodies produced in the cultivation process are highly dependent on the degree and extent of glycosylation (Del Val *et al.* 2011; Durocher and Butler 2009; Pais *et al.* 2014). Glycosylation can be defined as the sequential attachment of nucleotide sugars to the backbone of the polypeptide chain and it is quantifiable by the type and number of nucleotide sugars attached to the recombinant protein. N-linked glycosylation is the most conventional form of glycosylation in CHO cells (Del Val *et al.* 2011; Hossler *et al.* 2007), which are the focus of this study. It has been demonstrated in the literature that the glycosylation pattern can be affected by the *in vivo* metabolite availability (Del Val *et al.* 2011; Hossler *et al.* 2007; Liu *et al.* 2014). Accordingly, a model was developed to relate extracellular metabolite states to the glycosylation pattern of Mab produced through the synthesis of nucleotide sugars required for glycosylation.

The importance of real time monitoring is not limited to the upstream processes. In the downstream stages proteins are exposed to changes in pH, temperature, shear stress, and ionic strength making them prone to aggregation (Sadavarte and Ghosh 2014; Sahin *et al.* 2010). The propensity to aggregate is particularly high for large hydrophobic proteins such as Mabs. Loss of therapeutic efficiency and/or triggering of an immunogenicity response in the patient are listed as unfavorable impacts of recombinant protein unfolding and aggregation (Hawe *et al.* 2008; He *et al.* 2010; Mahler *et al.* 2009; Printz and Friess 2012). Consequently, it is crucial to develop techniques for fast online monitoring of protein aggregation under process conditions typically occurring during downstream processing. Most of the conventionally used techniques for measuring aggregation are time consuming and expensive assays; consequently they are unsuitable for real-time screening. Due to the presence of intrinsically fluorescent amino acids incorporated in the polypeptide chain of Mabs and the sensitivity of these amino acids to changes in their microenvironment, fluorescence spectroscopy can be implemented as a viable technique for *in situ* monitoring of protein aggregation (Abbas *et al.* 2013; Elshereef *et al.* 2006; Kumar *et al.* 2005). Fluorescence spectroscopy has been implemented to track α-lactoglobulin and β-lactoglobulin solubility under stressed conditions (Elshereef *et al.* 2006). Abbas *et al.* 2013 demonstrated the capability of second-derivative fluorescence to qualitatively track conformational changes under thermal, pH, and solvent stressed conditions for recombinant proteins. However, no soft sensor has been developed to classify and quantitatively assess Mab aggregation.

3

1.2 Research Objectives

The main objective of this research was to develop soft sensors for fast *in situ* monitoring of both the upstream and downstream monoclonal antibody production processes. Two different types of soft sensors were investigated. A data-driven soft sensor (black-box) was based on the multi-wavelength fluorescence spectra coupled with partial least squares regression (PLSR) and a model-driven soft sensors (white-box) that use mechanistic models that were developed based on the physical phenomena of the system under study. The following research approaches were pursued:

I. Fluorescence-based soft sensors were developed for fast *in situ* monitoring of viable cells dead cells, recombinant protein, glucose, and ammonia concentrations in batch CHO cell cultures.

II. An adaptive Kalman filter-based soft sensor that combines a fluorescence-based (data-driven) soft sensor and a dynamic metabolic model (model-driven soft sensor) was proposed to enhance the prediction accuracy and to better capture the dynamics of the process by providing estimates in between fluorescence samples.

III. A model was developed to relate the extracellular culture conditions to the glycosylation pattern of the Mab produced. This model is instrumental for identifying operating conditions that result in a particular glycan structure.

IV. Fluorescence-based soft sensors were developed for monitoring Mab monomer concentration and for classifying the level of Mab aggregation under different process conditions that typically occur during downstream processing.

1.3 Structure of Thesis

The thesis comprises 7 chapters, references, and appendices. Chapter 2 focuses on the background information of the topics covered in the thesis. Chapters 3 to 6 are presented in manuscript format each with an individual introduction and conclusion; thus, there might be some repetition between introduction of the thesis and introduction to the chapters. The purpose of Chapters 3 to 5 is to describe the soft sensors (data-driven and model driven) that were developed for real-time upstream quality and quantity monitoring. Chapter 6 presents a data-driven soft sensor for tracking Mab monomer levels and Mab aggregation under process conditions that typically occur in downstream processes. Finally, Chapter 7 presents the overall conclusions, recommendations and future research directions. Further details on the contents of the different chapters are as follows:

Chapter 3.   Tracking viable cell and Mab of CHO cell cultivation using intrinsic fluorescence spectroscopy has been previously studied in the literature (Teixeira *et al.* 2011; Teixeira *et al.* 2009b). However, no soft sensor has been developed to monitor dead cell and metabolites in addition to viable cell and recombinant protein concentration. In this chapter time profiles of different fluorophore peaks were studied to gain a better understanding of the correlations between the temporal evolution of the culture and changes in fluorescence spectra. Additionally, by implementing a supervised machine learning technique, i.e. PLSR, empirical models were developed to individually monitor viable cell, dead cell, recombinant protein, glucose, and ammonia concentrations over the course of the culture. This chapter has been published in *Biotechnology and Bioengineering Journal* (Ohadi *et al.* 2014a).

Chapter 4.   Although data-driven soft sensors, such as the one presented in Chapter 3, are fairly accurate within the region of operating conditions where data was collected for model calibration, their precision might diminish when predicting conditions that deviate substantially from this region. In this chapter limitations and advantages of data-driven and model driven soft sensors are compared. A data-driven soft sensor is developed to track viable cell, dead cell, recombinant protein, glucose, and ammonia concentrations using multi-wavelength fluorescence in combination with PLSR. Furthermore, a mechanistic model is formulated to dynamically track the metabolic state of the culture. Then, the mechanistic model is combined with the data-driven soft sensor using extended Kalman filter to enhance predictions at operating conditions that considerably deviate from the ones used for model calibration and to accurately reconstruct the dynamics of the estimates in between sampling instances. This chapter has been published in *Biotechnology and Bioengineering Journal (Ohadi et al. 2014b).*

Chapter 5.   One of the critical challenges that the biopharmaceutical industry is facing with is to ensure the products quality, i.e. therapeutic efficacy. Glycosylation as a post-translational modification plays a crucial role in the characteristics of the product and it is related to the extracellular state of the culture. In this chapter a novel model is developed consisting of three sub-models: metabolic model to capture the dynamics *in vivo*, production of nucleotide sugars in cytosol, and glycosylation inside the Golgi apparatus to relate extracellular state to cumulative glycoform profile. The second sub-model is presented as the link between extracellular and intracellular conditions. This work has been published

in IFAC-CAB (International Symposium on Computer Applications in Biotechnology) (Ohadi *et al.* 2013).

Chapter 6.   During purification proteins are constantly exposed to changes in operating conditions such as temperature, pH, shear stress, and ionic strength. Large individual or combined variation of these process conditions may result in disruption and consequently aggregation of the proteins. In this chapter experiments were performed by exposing samples of a Mab, at different concentrations, to changes in temperature and pH. These changes were intended to emulate corresponding typical changes in operating conditions occurring in downstream processing of Mab. Size exclusion chromatography was utilized to quantify protein aggregation. To better elucidate the impact of pH and temperature and to identify the operating regions with lower tendency of aggregation, a response surface model was calibrated and studied. Multi-wavelength fluorescence spectra of samples that experienced different levels of temperature and pH, combined with PLSR, were then utilized to develop an empirical model based soft sensor to predict monomer content of samples. Moreover, the treated samples were classified into four classes based on the ratio of their oligomer to monomer content. A soft sensor, for qualitative control, was formulated relating the fluorescence spectra to these classes. This chapter has been submitted to *Journal of Biotechnology*.

Chapter 2

Theoretical Background

2.1 Online Monitoring

A key challenge faced by the biopharmaceutical industry is to meet the ever-increasing demand of bio-drugs while maintaining product quality. Real-time analytical techniques are crucial for both the development of a bioprocess as well as for enhancing its controllability. Many methods used for bioprocess monitoring are time consuming, labor intensive, and very often require regular maintenance steps thus making them unsuitable for online use. Fast online monitoring techniques are playing an important role in increasing process efficacy, maximizing productivity, guaranteeing consistency, reproducibility and enabling automated control. Process analytical technology (PAT) is a systematic framework that encourages manufacturers to adopt tools and techniques for measuring quality and performance attributes so as to guarantee process reliability and consistency (Lourenco *et al.* 2012; Mandenius and Gustavsson 2014; Rathore *et al.* 2010; Teixeira *et al.* 2009a).

Monitoring techniques can be classified into three groups (Lourenco *et al.* 2012):

- Off-line: Measurements are performed at the laboratory following manual or automatic sampling. They are typically time consuming and cannot be used for real-time tracking of key variables.

- At-line: The measurements are done at the process, closer to the sampling location. Although these techniques can provide faster answers than the off-line measurements described above, they are not best suited for real-time monitoring and control due to the time delay between sampling and measurement and to the variability in duration of the manual measurement operation.

- Online: Online techniques can be classified into two subclasses: *in situ* and *ex situ*. The in situ configuration is a closed system, i.e. either no sample is removed from the vessel or it is returned to the vessel after the measurement. The sensor for this class can be invasive (in contact with the sample) or non-invasive. In contrast, the *ex situ* configuration is an open system where the sample is transferred to the measuring device and discarded afterwards. Both these techniques are useful for real-time monitoring and control.

7

The focus of this study is the development of non-invasive *in situ* method(s) for measuring key process variables of upstream and downstream process stages of Mab production during and following mammalian cell cultivation.

2.1.1 Soft Sensors

The term soft sensor (software sensor) refers to a combination of a hardware sensor that collects signal(s) (secondary measurements) from the process and a mathematical model (in a form of a computer model) that relates these signal(s) to a process variable(s) (primary variable(s)) of interest (Mandenius and Gustavsson 2014). From a wide perspective soft sensors are classified as data-driven and model-driven:

- Data-driven soft sensors are inferential empirical models that are developed only based on the data obtained from the process. Thus, they do not provide mechanistic insight of the process dynamics and perform as a black-box model (Kadlec *et al.* 2009). Since these soft sensors are established from the data of the process, they exhibit an accurate prediction within the calibration range (Kadlec *et al.* 2009). However, their prediction quality diminishes when operating conditions considerably diverge from the calibration range (Shioya *et al.* 1999). In order to obtain an accurate empirical model a large set of data, encompassing a wide range of operating conditions, need to be acquired for rigorous model calibration. Other key challenges in the development of the empirical models are measurement noise and missing data (Luttmann *et al.* 2012). The development of data-driven soft sensors is comparatively easier than model-driven soft sensors, by implementing machine learning techniques, thus became a popular approach in the context of real-time monitoring. It is worth noting that the development of data-driven soft sensors is mostly done on an ad hoc basis.

- Model-driven soft sensors are based on mechanistic models involving dynamic mass, components, and energy balances. Thus, these models are generally given in terms of differential or algebraic equations that capture both the steady state and dynamics of the physical or chemical phenomena of the process. Due to their ability to extrapolate behavior away from the region of calibration, these models can be utilized for process optimization and scheduling (Kadlec *et al.* 2009; Shioya *et al.* 1999). To contrast them with empirical or black-box models, mechanistic models used in model-driven soft sensors are also referred

to as white-box models. During the process development stage model-driven soft sensors often surpass the data-driven counter parts since a limited amount of data is available. However, development of white-box models is taxing due to the complexity of biopharmaceutical processes thus requiring expert knowledge of the system. In addition, the presence of a large number of model-parameters exacerbates the modeling challenges and increases the model prediction error through parameter uncertainty. Mechanistic biochemical models can be classified into structured or unstructured models. Since structured models are based on the actual physical/chemical phenomena of the system they are generally able to provide more accurate predictions over a wider range of operating conditions (Gernaey *et al.* 2012).

A soft sensor that is based on a hybrid model that combines a data-driven and a model-driven soft sensors, i.e. gray-box model, has been implemented in the current work to tackle individual drawbacks pertinent to white- and black-box models (Ohadi *et al.* 2014b).

2.1.2 Data-driven Soft Sensor

Several types of sensors have been investigated in the pharmaceutical industry for rapid process monitoring. Among these approaches, spectroscopic methods—such as UV-Vis absorbance, near infrared (NIR), mid infrared (MIR), Raman, dielectric, and fluorescence spectroscopy have gained popularity due to their noninvasive, nondestructive, and highly informative nature (Kara *et al.* 2011; Lourenco *et al.* 2012; Mandenius and Gustavsson 2014; Teixeira *et al.* 2009a). Additionally, their signal acquisition is considerably fast thus making them a natural choice of hardware for soft sensor development. Apart from these spectroscopic methods, nuclear magnetic resonance (NMR) has also been utilized for bioprocess monitoring (Kara *et al.* 2011). Changes in temperature, the presence of bubbles, background noise and variations in medium are some of the challenges that have to be overcome when implementing spectroscopic methods. To overcome these obstacles statistical methods (chemometric methods) and proper data preprocessing should be implemented (Lourenco *et al.* 2012). Chemometric methods are discussed in detail in subsequent sections.

These spectroscopic techniques are briefly discussed below:

- UV-Vis Absorbance: The samples are illuminated in a range of 200 to 780 nm (Lourenco *et al.* 2012; Teixeira *et al.* 2009a). In absorption spectroscopy the light passes through the sample and is attenuated based on a presence of a particular chemical in the sample. The

abundance of that chemical can be calculated using Beer's Law (Lourenco *et al.* 2012). UV-Vis spectroscopy has been reported to be capable of tracking conformational changes in proteins (Kumar *et al.* 2005); however, the presence of suspended particles in the sample can substantially deteriorate the quality of the signal. Moreover, this technique is not capable of providing structural information about the content of the sample under study (Kara *et al.* 2011). Although this technique has been extensively used because of its availability, its application for biopharmaceutical monitoring is limited.

- Infrared Spectroscopy: The NIR spectroscopy region ranges from 780 to 2500 nm and the MIR region ranges from 2500 to 40000 nm (Lourenco *et al.* 2012; Teixeira *et al.* 2009a). Upon absorption of the excitation light, molecules rotate or vibrate and subsequently transmit light that can be exploited to draw inferences about the molecular composition of the sample. Infrared spectra can be acquired using a probe without being in contact with the sample (noninvasive). A single probe can provide signals that can be used to predict various physical and chemical states of the process. Molecules that contain functional groups with a hydrogen bound including: C-H, N-H, O-H, and S-H are responsive to NIR exposure. NIR is by far the most extensively used technique for the development of soft sensors (Lourenco *et al.* 2012; Teixeira *et al.* 2009a). A major drawback of NIR is the low molar absorptivity of molecules and therefore, this method is not capable of detecting highly diluted components. Additionally, in aqueous solution which is the typical form that samples are collected in biopharmaceutical processes, the signal of water molecules highly interferes with other substances and diminishes the signal quality. Despite the better resolution of the MIR signal, its expensive and sensitive fiber probe made it less attractive for online monitoring (Lourenco *et al.* 2012; Mandenius and Gustavsson 2014). Infrared spectroscopy has been used for monitoring mammalian cell cultures (Harthun *et al.* 1998; Sellick *et al.* 2010) and fermentations (Nordon *et al.* 2008).

- Raman Spectroscopy: Raman is another form of vibrational spectroscopy involving a scattering light that provides information about the vibrational, rotational, and low frequency transitions of the molecule. A large portion of the scattered light (mostly laser light) is an incident light with the same specification as the excitation source (Rayleigh scattering). A small portion of it is scattered with a different wavelength from the excitation beam (Raman scattering). The Stokes-shift—i.e. the shift between the excitation light and the Raman scattering wavelength—can provide information about the molecules present in the sample

(Teixeira *et al.* 2009a). Raman has a lower propensity for absorption, in comparison with NIR and MIR, by polar substances (e.g. water). Raman spectroscopy can be implemented for different physical states (gas, liquid, and solid) of substances. The drawbacks of all vibrational spectroscopic techniques including Raman are weak signal and incapability of properly detecting dilute substances (Lourenco *et al.* 2012; Teixeira *et al.* 2009a). Use of high power laser source makes the instrument costly and in some cases this high-energy light can disrupt the sample (Lourenco *et al.* 2012). Furthermore, the presence of naturally fluorescing compounds in bioprocesses interferes with the Raman signal. Li *et al.* (2010) used Raman spectroscopy as a technique for evaluating the performance of culture media. Raman spectroscopy has also been implemented for monitoring of metabolites, viable cells, and total cell concentrations in mammalian cell culture (Abu-Absi *et al.* 2011).

- Dielectric Spectroscopy: dielectric spectroscopy is based on a radio frequency field that causes charge separation around the non-conductive cell membrane. Cells with undamaged membrane, i.e. viable cells, act as capacitors. Thus, this technique is capable of providing information about the size, abundance, and type of only viable cells (Teixeira *et al.* 2009a).

- Fluorescence Spectroscopy: Fluorescence spectroscopy is a highly sensitive and selective technique. Fluorescence spectroscopy can be intrinsic (through naturally fluorescing compounds) or extrinsic (through attaching a dye to the substance of interest). Due to the presence of intrinsically fluorescing substances in biological mixtures intrinsic fluorescence can be implemented non-invasively. An advantage of this technique over the vibration spectroscopies reviewed above is its capability of detecting trace substances (Lindemann *et al.* 1998; Ohadi *et al.* 2014a; Teixeira *et al.* 2009b). Details on fluorescence spectroscopy can be found in subsequent sections.

The purpose of this study was to develop soft sensors for monitoring the upstream and downstream processes of Chinese hamster ovary cell cultures producing Mab. To achieve this goal multi-wavelength fluorescence spectroscopy was exploited to develop data-driven soft sensors. Aside, model-driven soft sensors were also developed to either enhance the prediction the quality of data-driven soft sensors or be used standalone for quality control.

2.2 Protein Aggregation

Mabs, as any other protein-based product, are prone to aggregation which is considered one of the most significant causes of product instability and performance loss (Wang 2005). Protein aggregation

can lead to loss of activity of the product and/or adversely affect the patient by triggering an immune response (Hawe *et al.* 2008; He *et al.* 2010; Mahler *et al.* 2009). Protein aggregation can occur in various stages of the manufacturing process including upstream (fermentation), downstream (purification and formulation), and during storage and delivery. Aggregation can be induced by factors including, but not limited to, temperature changes/fluctuations, pH changes, shear stress, ionic strength changes, and exposure to light (Arosio *et al.* 2013; Mahler *et al.* 2009; Printz and Friess 2012; Wang 2005). Thus, it is crucial to establish methodologies for accurate real-time characterization and monitoring of recombinant protein aggregates. In the current study a fluorescence-based soft sensor was developed for quantitative and qualitative monitoring of the downstream processes (Chapter 6). Prior to analyzing protein aggregation and developing monitoring techniques it is crucial to understand the mechanisms by which the aggregates are formed, as well as the stress factors that induce such adverse phenomenon. This section briefly discusses the basics of protein aggregation and mechanisms, induction factors, and analytical measurement techniques.

2.2.1 Protein Aggregation Pathways

Protein aggregation can occur via a variety of pathways, forming either soluble or insoluble aggregates in the form of fibrillar or amorphous material, depending on the type of protein as well as environmental conditions (Mahler *et al.* 2009). Protein molecules can undergo physical aggregation by physical protein-protein interactions, or chemical aggregation involving the formation of covalent bonds. Chemically induced aggregation can occur via oxidation, which causes modification of the protein sequence that leads to aggregation, or via a Maillard reaction in which the amino acids of the protein react with reducing sugars, forming a dark-brown aggregate as the end product (Wang 2005). Aggregates may be classified according to different criteria: i-size (ranging from nm to μm in diameter)—smaller soluble aggregates or larger visible aggregates; ii- reversible versus irreversible; iii- non-covalent—bound by electrostatic forces and Van der Waals interaction, or covalent—formed by disulphide bond linkages or non-disulphide cross linking pathways and, iv- protein conformation—native structure, or mostly non-native structure (Mahler *et al.* 2009; Wang 2005). Aggregates may be soluble at first, and eventually become insoluble as they increase in size. The process of protein aggregation generally involves the unfolding of a native protein, causing a hydrophobic amino acid to become exposed. This exposed amino acid then forms new intermolecular bonds with other unfolded hydrophobic amino acids, resulting in the formation of protein aggregates (Elshereef *et al.* 2008; Wang 2005).

2.2.2 Aggregation Induction Factors

Protein aggregation can be induced by changes in a number of physiochemical factors, including temperature, pH, protein concentration, stresses/shaking, and freeze-thawing. The structure of a protein can also affect its aggregation, i.e. the more hydrophobic a protein is, the more it will aggregate (Arosio *et al.* 2013; Calamai *et al.* 2006). The presence of multiple regions and structural sequences of MAbs increase the propensity of these large proteins to form non-native aggregates under stressed conditions (Sahin *et al.* 2010). Sometimes two or more of these factors are simultaneously varied in order to study aggregation in greater depth (Elshereef *et al.* 2006; Elshereef *et al.* 2008).

The following is an explanation for the stress factors being studied in the current work:

- Temperature: Temperature has a profound effect on the formation of protein aggregates. Higher temperatures can contribute to unfolding of the secondary, tertiary and quaternary structure of polypeptide conformations, which can lead to the formation of non-native aggregates (Mahler *et al.* 2009; Printz and Friess 2012). Additionally it can trigger chemically induced aggregation (Mahler *et al.* 2009). Printz and Friess (2012) observed formation of oligomeric structure of IgG after being heated for 20 minutes at 60 °C and approximately 30% reduction in the monomer form. Hawe *et al.* (2008) concluded that higher temperatures can result in significantly large aggregates. Similar conclusions were drawn in Chapter 6. Figure 2-1 shows the chromatograms of heat stressed humanized IgG (0.5 mg/ml initial concentration) at 60 °C over a 60 minutes time span. The loss of monomeric structure and formation of trimmer/tetramer is evident temporally.
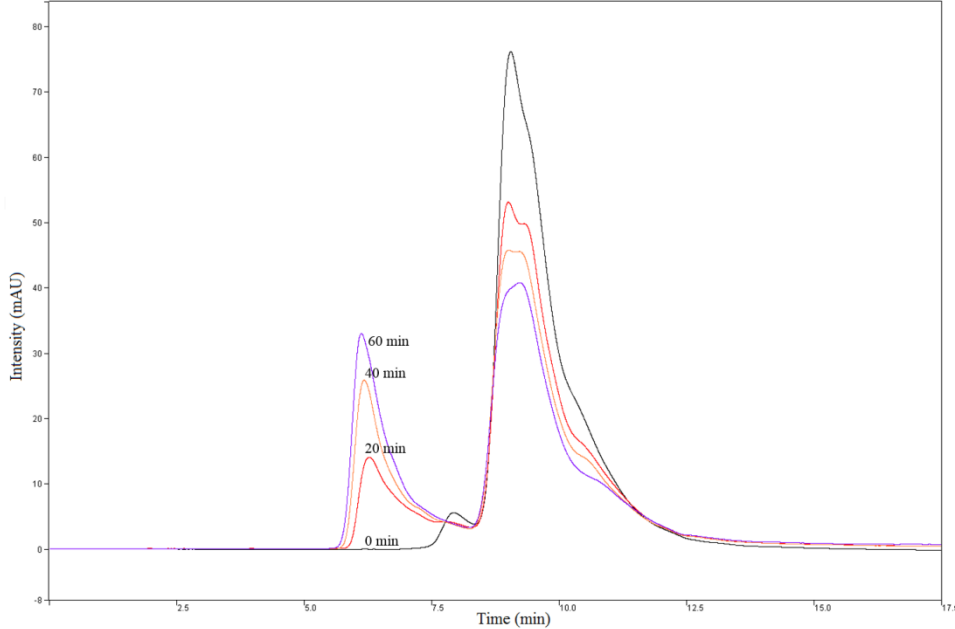
Figure 2-1. Chromatograms of heat stressed IgG (0.5 mg/ml initial concentration) at 60 °C over 60 minutes time span

- pH: Changes in pH can affect protein aggregation by changing the distribution of charges on the protein molecule (Wang 2005). The behavior of proteins under different pH conditions depends on a number of factors. Elshereef *et al.* (2008) found that the propensity of β-lactoglobulin to aggregate was higher near its isoelectric point. During the purification stages of recombinant proteins, such as affinity chromatography (Protein-A), low pH buffers are applied which result in the formation of large aggregates and a substantial decrease in the monomer concentration (Arosio *et al.* 2013; Hawe *et al.* 2008). Printz and Friess (2012) observed over 85% loss of the initial monomer IgG when subjected to a pH=1.1.

- Protein concentration: Protein aggregation is believed to increase proportionally to the protein concentration. Mahler *et al.* (2009) argued that high protein concentrations can increase aggregation through self-assembly and to also reduce the propensity for unfolding. Haw *et al.* (2008) observed that the effect of temperature at 80 °C on the formation of oligomeric structures was more pronounced at higher concentration (2 mg/ml) of IgG. However, at 75 °C the sample with lower concentration (0.2 mg/ml) experienced a more

significant increase in average particle size. Hence the effect of concentration is dependent on other factors.

2.2.3 Analytical Techniques for Measuring Protein Aggregation

Protein aggregates can range in size from nanometres in diameter, to being visible to the naked eye. In order to obtain further information, it is necessary to measure protein aggregation, either qualitatively or quantitatively. A number of methods have been used, including chromatography, spectroscopy and light scattering. An inconsistency might be observed in the measurements obtained via these techniques pertinent to the difference in the nature of the technique (Mahler *et al.* 2009). Mahler *et al.* (2009) emphasized the lack of suitable analytical techniques for measuring protein aggregation over the whole size range. A detailed discussion of all techniques for measuring aggregation can be found in the literature (Mahler *et al.* 2009). Appendix A summarises some of the conventional methods for detecting, analysing, and quantifying protein aggregation and highlights of the advantages, disadvantages, and size range for each technique.

2.2.3.1 Size exclusion chromatography

Size Exclusion Chromatography (SEC) is ranked as the most frequently utilized method for quantitative analysis as well as for particle size evaluation. SEC involves the use of a porous bed to fractionate molecules of different sizes which allows for both the determination of aggregates, as well as quantification of their size by implementing various detectors (Mahler *et al.* 2009; Printz and Friess 2012). This method is often used in conjunction with other techniques, such as HPLC (High Pressure Liquid Chromatography) that involves a column containing a porous separation medium. Molecules that are larger than the pore size do not penetrate the porous medium thus passing through the column and appearing as the first peak in the chromatogram (total exclusion). Molecules smaller than the pore size penetrate the porous medium in the column thus resulting in a longer transit time and eventually eluting after the larger particles (den Engelsman *et al.* 2011). Figure 2-2 illustrates the process of a sample moving through an SEC column. As seen in the elution curve, the higher molecular weight components elute first followed by the smaller molecular weight components. In this study a HP-SEC equipped with a UV absorbance detector was utilized for sample fractionation and estimation of monomer concentration of samples that had been subjected to different stressors (Chapter 6).
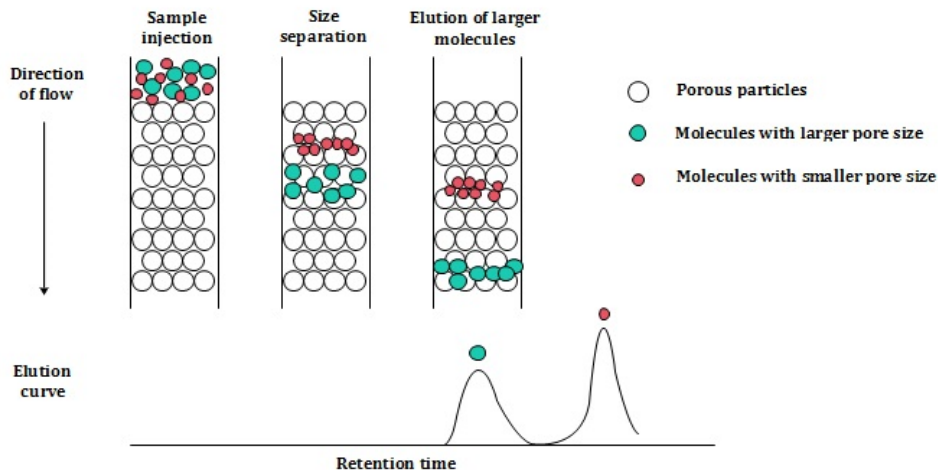
Figure 2-2. Schematic diagram illustrating the elution of molecules through a SEC column.

2.3 Fluorescence Spectroscopy

Since the early 1950s, fluorescence spectroscopy has become a popular analytical measurement tool in biochemistry, biophysics, and material science (Hof *et al.* 2005). Fluorescence techniques involve monitoring the emission of light from fluorophores resulting from the transfer of electrons from an excited energy level back to either one of the vibrational levels of the electronic ground state. This excitation is caused by the absorption of electromagnetic radiation (Herman 2003; Hof *et al.* 2005). The term photoluminescence refers to luminescence phenomena involving excitation by UV and/or visible light. Photoluminescence can be classified into two types: fluorescence and phosphorescence where the latter exhibits a longer excited lifetime than the former (Herman 2003). Not all molecules are capable of undergoing electronic transitions following absorption of light. Molecules that fluoresce following excitation by light are called *probes*, *fluorochromes*, or *dyes* (Herman 2003). A larger molecule with a probe bound to it, is known as a *fluorophore* and can be classified as *intrinsic* or *extrinsic*.

2.3.1 Basic Principles

The fluorescence process by which molecules are excited by light is depicted in Figure 2-3 and is known as a *Jablonski* diagram (Christensen *et al.* 2006; Hof *et al.* 2005). When at room temperature, molecules are at the lowest vibrational level of their ground electronic state (S_0). By absorbing an

external photon (i.e. light) with an energy of $E_{ex} = hv_{ex}$, where h is the Plank constant and v_{ex} is a frequency of excitation, fluorophores are excited from the electronic singlet state S_0 to either of the first excited state (S_1) or second excited state (S_2). This is followed by a rapid loss of excess energy to the lowest vibration level of the excited state, due to interactions with the molecular environment through vibrational relaxation, quenching and energy transfer. Subsequently, fluorophores will lose energy by emitting at a longer wavelength and consequently a lower energy ($E_{em} = hv_{em}$) photon until they reach the lowest vibrational level of their ground state (Christensen *et al.* 2006; Hof *et al.* 2005; Lakowicz 1999). These sequential events occur within nanoseconds after the sample is excited by light. In addition, a slower luminescence phenomena, phosphorescence, can also take place involving molecules that go through intermediate states but contrary to fluorescence it is temperature dependent (Christensen *et al.* 2006). The difference between the absorption and emission wavelength is referred to as Stokes shift (Christensen *et al.* 2006; Deshpande 2001).
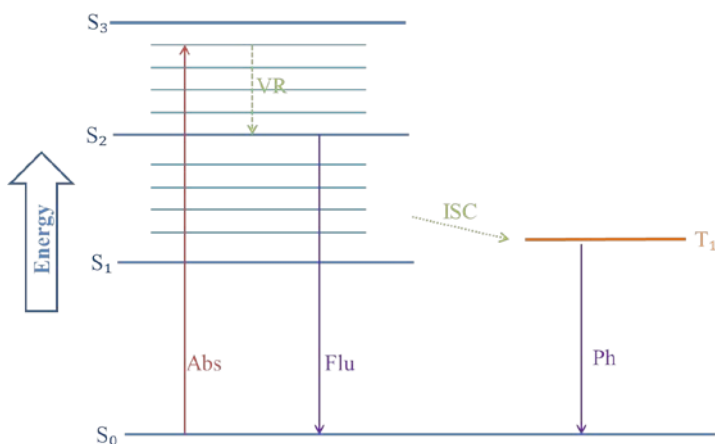
Figure 2-3. A simplified Jablonski diagram illustrating fluorescence and phosphorescence. In this figure, (Abs) absorption, (Flu) fluorescence, (Ph) phosphorescence, (ISC) intersystem crossing, and (VR) vibrational relaxation.

The fact that each electronic state has multiple vibrational sub-levels suggests that excitation can occur at various wavelengths associated with these vibrational transitions. Accordingly, the absorbed energy will result in a series of absorption bands. Also, since emission can be produced by electrons at several vibrational levels of the ground state, different emission wavelengths are expected. This phenomenon will result in a distinct excitation emission profile for a specific fluorophores (Christensen *et al.* 2006).

17

2.3.2 Extrinsic and Intrinsic Fluorophores

Fluorophores can be formally classified into *intrinsic* and *extrinsic* types. Extrinsic fluorophores are referred to as *dyes* or *fluorescent labels* that are attached to a particular substance and fluoresce with a specific spectrum. *Intrinsic* fluorophores are those that naturally exist in the system.

The medium, also referred to as the supernatant or broth, in mammalian cell cultures is a complex blend of nutrients required for cell growth. This blend includes intrinsic fluorophores, such as certain amino acids (tryptophan, tyrosine, phenylalanine), vitamins (pyridoxine and riboflavin), and co-factors (NAD(P)H and FAD). Thus, making intrinsic fluorescence spectroscopy a viable choice for *online* or *at-line* monitoring technique. The fluorescent behavior of these compounds is affected during the course of the culture through generation, consumption and/or incorporation inside the molecules. Each of these fluorophores exhibits a unique absorption and emission wavelength that make them distinguishable from each other. Table 2-1 summarizes the approximate region of the spectra corresponding to a particular fluorophore. It is important to notice that the quantum yield and Stokes-shift of these fluorophores can be affected drastically by their environment; thus, reinforcing the capability of fluorescence spectroscopy to serve as a tool to track the dynamics of the cell cultivation both qualitatively and quantitatively (Li and Humphrey 1991; Teixeira *et al.* 2009b). Due to peak overlaps, peak-shifts as well as effects imposed by solvent and other fluorophores multi-wavelength fluorescence spectroscopy is extensively favored over the use of single-wavelength data (Haack *et al.* 2004; Marose *et al.* 1998; Skibsted *et al.* 2001). Multi-wavelength fluorescence spectroscopy has been previously implemented for a variety of applications such as characterization of dissolved organic matter (Peiris *et al.* 2011a; Peiris *et al.* 2011b; Peiris *et al.* 2010), predicting the performance of growth media (Ryan *et al.* 2010), monitoring bacterial cultures (Hagedorn *et al.* 2003; Henneke *et al.* 2005; Jain *et al.* 2011; James *et al.* 2002; Skibsted *et al.* 2001), monitoring yeast cultures (Haack *et al.* 2004; Li and Humphrey 1991), and for monitoring of viable cell and dead cell in mammalian cell cultures (Teixeira *et al.* 2011; Teixeira *et al.* 2009b).

Table 2-1. Excitation-Emission wavelength for fluorophores with average stokes shift (Lindemann *et al.* 1998; Ryan *et al.* 2010; Teixeira *et al.* 2011)

Fluorophore	Excitation (nm)	Emission (nm)	Average Stokes Shift
Tyrosine	260~280 nm	300~310 nm	35 nm
Tryptophan	280~300 nm	350~360 nm	65 nm
Pyridoxine	320~330 nm	390~410 nm	75 nm
NAD(P)H	340~360 nm	440~460 nm	100 nm
FAD, Riboflavin	460~470 nm	500~520 nm	45 nm

In the context of protein aggregation, intrinsic fluorescence spectroscopy has been identified as a feasible technique for monitoring conformational changes and characterization of tertiary structure (Abbas *et al.* 2013; Kumar *et al.* 2005; Printz and Friess 2012). Three intrinsically fluorescing aromatic amino acids (phenylalanine, tyrosine, and tryptophan) exist in the protein chain that are sensitive to changes in their environment. Thus their fluorescence behavior can be exploited to draw inferences on the conformational changes as well as native and non-native characteristic of the protein (Eriksson L. 2006; Kumar *et al.* 2005; Printz and Friess 2012). The higher quantum yield and extinction coefficient of tryptophan makes it the fluorophore of choice to track protein aggregation (Abbas *et al. 2013;* Poole et al. 2012). The emission maxima of tryptophan shifts depending on the environment in terms of degree of exposure of the tryptophan residues to the solvent (Abbas *et al.* 2013; Poole *et al.* 2012). Second-derivative fluorescence has been utilized to investigate protein unfolding (Abbas *et al.* 2013; Kumar *et al.* 2005). Elshereef *et al.* (2008) developed a soft sensor based on multi-wavelength fluorescence to monitor α-lactoglobulin and β-lactoglobulin monomers' concentrations under stressed condition.

2.3.3 Excitation Emission Matrix

The 2D-fluorescence spectra can be organized into a matrix that is referred to as the Excitation-Emission Matrix (EEM). The fluorescence intensity values in the EEM are obtained by changing the excitation wavelength with specific increments and by measuring the corresponding emitted intensity over a desired range. Figure 2-4 is a visualization of a 3D-fluorescence spectrum of the growth media (Biogro) understudy in the form of a contour plot. A complete EEM can be divided into three major sections (Figure 2-4):

i. Fluorophores' Spectra: This is a region that includes intrinsic fluorophore peaks. Generally peaks in this region have an emission wavelength larger than their corresponding excitation wavelength ($\lambda_{em} > \lambda_{ex}$). These peaks are summarized in Table 2-1 and are indicated in Figure 2-4.

ii. Light Scattering Region: Not all radiation is absorbed by fluorophores. Instead, some might be reflected or scattered in different directions (Figure 2-4). This phenomenon can be due to solvent molecules or the presence of suspended colloidal particles and can have a considerable impact on the fluorescence EEM (Christensen *et al.* 2006; Lakowicz 1999). Scattering can happen in both clear and unclear (turbid) solutions. Rayleigh (first and second order) as well as Raman scatterings are two optical phenomena that occur in clear solution, while in opaque solutions Tyndall scatter, caused by large colloidal particles, can also be expected. Rayleigh scattering is due to excitation light scattering by air-cell interfaces, cell-wall solution interfaces and small particles in the sample (Christensen *et al.* 2006; Deshpande 2001). This type of scatter happens with $\lambda_{em} = \lambda_{ex}$ (first order) and $\lambda_{em} = 2\lambda_{ex}$ (second order) and therefore it normally does not interfere with the emission signal of the fluorophores (Christensen *et al.* 2006). Raman scattering is caused by the interaction of incident light with the solvent molecules, during which part of the incident energy can be converted to vibrational and rotational energy. Correspondingly, Raman scattering may be obtained at an emission wavelength with either lower energy or higher wavelength than the excitation one (Deshpande 2001). Raman scattering comes to the fore when dealing with low concentration samples while using high instrument sensitivity settings (Deshpande 2001). Some studies have made explicit use of different types of scattering to quantify protein aggregation (Elshereef *et al.* 2006; Elshereef *et al.* 2008).

iii. Non-informative Region: A triangular-shaped region on the top left-hand side of the Figure 2-4. This region exhibits emission wavelengths smaller than the excitation wavelengths ($\lambda_{em} < \lambda_{ex}$) which is physically impossible since it would imply that the emitted energy is larger than the excitation.
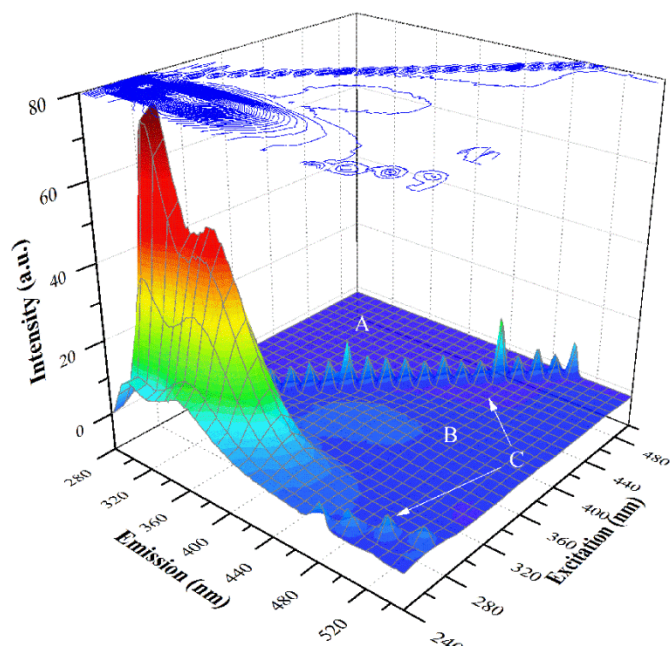
Figure 2-4. Multi-wavelength fluorescence spectra of the growth media in a form of a 3D plot: (A) Non-informative region, (B) Fluorophores' region, (C) Scattering region.

2.3.4 Optimization of Spectrofluorometer Settings

To guarantee the accuracy and reproducibility of the results the instrument settings of the spectrofluorometer should be optimized. This optimization was performed using Biogrow medium which has been specifically developed for use by members of MabNet. Three different spectrofluorometer settings including photomultiplier tube (PMT) voltage, scanning rate, and slit width (SW) were optimized to obtain consistent reproducible results. Additionally, the impact of dilution using different media to phosphate buffer saline (PBS) ratios was studied. Detail of this work can be found in section 3.2.3 and Appendix B.

2.4 Chemometric Methods

Although EEMs contain a large amount of useful information, they consist of a large number of variables (excitation-emission pairs) to sample ratio, highly correlated results, and an expected high sensitivity to noise. For such a complex dataset, univariate or bivariate methods will fail to extract sufficient information. Focusing on a particular peak (peak-picking) is not considered a viable choice because the position of the peaks can shift significantly as a result of changes in the environment of these fluorophores. Consequently, chemometric methods are crucial in the analysis of multivariate dataset and in the development of robust models. These methods can be used to relate changes in

21

multivariate measurements of a process to the actual conditions of a system by utilizing statistical approaches. Therefore, they have been widely used by researchers in bioprocess analysis (Rathore *et al.* 2011). Among these techniques data exploratory methods based on dimension reduction—e.g. principal component analysis, partial least square regression, and principal component regression—have gained more attention due to their simplicity as well as the interpretability of the results. In this study PLS is applied to relate the actual states of the system in upstream and downstream to the EEM obtained from fluorescence spectroscopy.

2.4.1 Principles of Multi-linear Regression

Generally, regression can be thought of as a relationship between independent parameters denoted as x_js and dependent ones (y_js). Linear regression is the simplest form of regression where the goal is to obtain a linear relationship between independent and dependent variables. Contrary to conventional wisdom, linear regression is not limited to only one dependent but can be extended for a system of p dependent variables as per Equation 2-1. Y and E are $n \times p$ matrices of dependent variables and residuals, respectively and B is a $m \times p$ matrix of coefficients which can be calculated from (Equation 2-2). Additional details on linear regression can be found in Montgomery (2009).

$$Y = XB + E$$

Equation 2-1

$$B = (X^T X)^{-1} X^T Y$$

Equation 2-2

In practice there are many cases where MLR as given by Equation 2-2 will not provide good results. Collinearities among independent variables (i.e. a highly correlated X matrix) combined with lower number of samples than variables and noisy measurements such as exhibited by 2-D fluorescence EEM, are a major cause for poor performance of MLR. Collinearity results in highly sensitive estimates of coefficients to noise (i.e. B`s will change significantly among models calibrated with different sets of data). Also, especially when using small data sets as compared to the number of parameters to be estimated, there is a possibility of over fitting. In that case the MLR based model may perfectly fit the calibration data set but it will fail to produce accurate predictions of data that were not used for model calibration (Geladi and Kowalski 1986; Wold *et al.* 1984).

2.4.2 Partial Least Squares

Partial Least Squares Regression (PLSR) was first developed by Herman Wold and it was initially used in the social sciences (Wold *et al.* 1984). In contrast with MLR, PLSR is of interest because of

its ability to handle collinear, noisy data and data sets with a very large number of variables. Moreover, PLSR can be used to tackle multiple-response problems, i.e. problems with many dependent variables, with a single regression model (Wold *et al.* 2001).

PLSR attempts to find new directions, i.e. Latent Variables (LV) in an independent input data space *X*, that both captures the variance in the X (independent variables) space and best describes the output (dependent variables) data Y. In other words, a multi-dimensional hyper-plane in X space is found such as the projection of independent variables on it (scores) correctly captures X, and at the same time, it can describe Y with a desirable degree of accuracy (Figure 2-5). This accuracy is achieved by maximizing the covariance between input and output while capturing the variance in the input. In PLSR the aim is to predict one set of data from another one and in that sense it is different from Partial Least Square Correlation (PLSC) which looks for similarities (correlations) between two matrices (Krishnan *et al.* 2011).
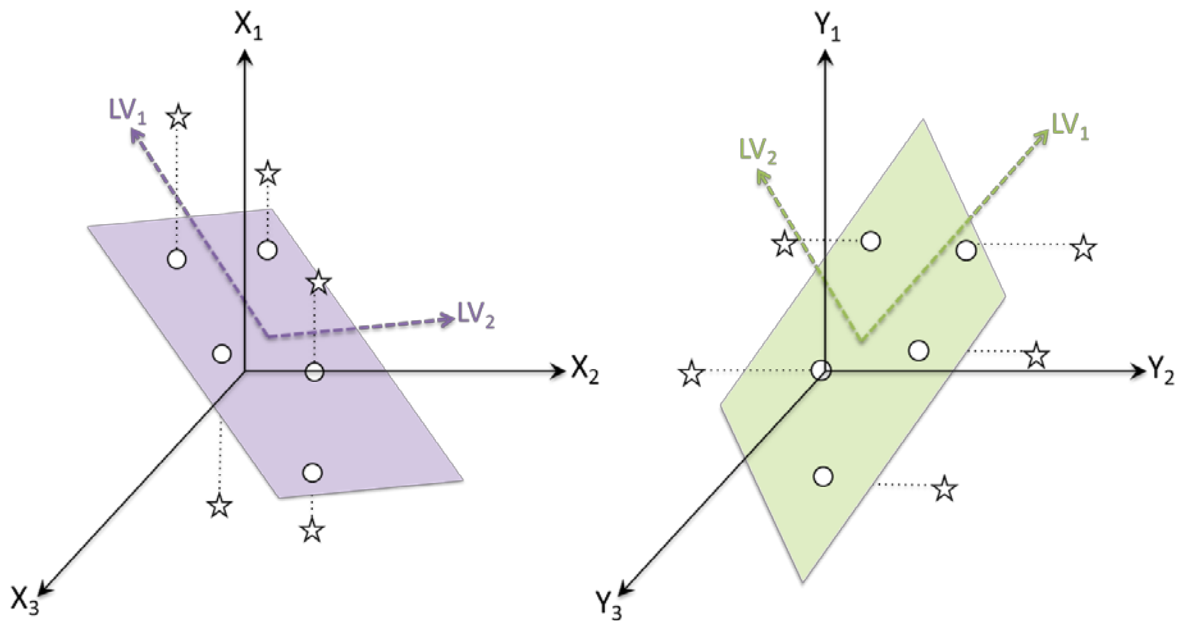
Figure 2-5. Schematic presentation of PLS derivation for an arbitrary 3-D X and Y space as well as hypothetical hyper-planes showing first and second LVs

Various algorithms have been investigated to calculate PLS regressions. Non-linear Iterative Partial Least Square (NIPALS) is the most commonly used algorithm for studying and understanding PLS method. NIPALS estimates scores, loadings and vector of weights iteratively. Detailed information

about the algorithm can be found in (Geladi and Kowalski 1986). The iterative nature of NIPALS results in numerous calculation steps and higher CPU time. Alternatively, SIMPLS is another method (Dejong 1993) that can be implemented. SIMPLS generates exactly the same results as NIPALS when only one response variable is considered and it produces slightly different results in the case of multivariate responses.

PLS consists of an outer relation for X and Y where these are given as follows:

$$X = TP^T + E$$

<div align="right">**Equation 2-3**</div>

$$Y = UQ^T + F$$

<div align="right">**Equation 2-4**</div>

X **is an** *n×m* **matrix,** *T* **is an** *n×l* **score matrix, and** *P* **is** *m×l* **loading matrix of the input data. Similarly,** *Y* **is an** *n×p*, *U* **is an** *n×l* **score matrix,** *and Q is a p × l* **loading matrix of the output data.** *l* **is number of latent variables. The Partial Least Square Regression equations are schematically illustrated in (**

Figure 2-6).

Finally it can be shown that the regression coefficient matrix *B* in Equation 2-1 can be calculated as Equation 2-5 (Wise 2006) and *T*, scores for independent matrix *X,* can be obtained as per Equation 2-6 (Wold *et al.* 2001).

$$B = W(P^T W)^{-1}(T^T T)^{-1} T^T Y$$

<div align="right">**Equation 2-5**</div>

$$T = X\, W(P^T W)^{-1}$$

<div align="right">**Equation 2-6**</div>

It can be shown that *W* contains the eigenvectors of $X^T Y Y^T X$ with descending order as per the corresponding eigenvalues. For example, Equation 2-7 shows the first eigenvector w_1 and its associated eigenvalue λ. It is also worth mentioning that the SIMPLS algorithm is based on this equation (Wold *et al.* 2001).

$$\lambda w_1 = X^T Y Y^T X w_1$$

<div align="right">**Equation 2-7**</div>
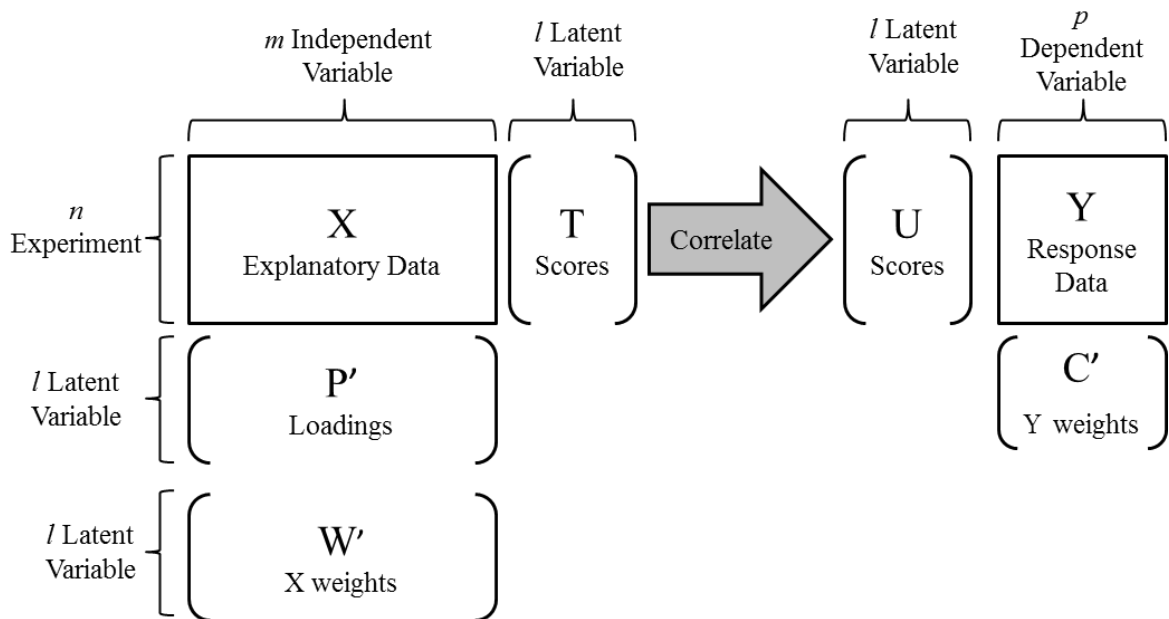
Figure 2-6. Use of PLS to correlate U (Y scores) to T (X scores) using MLR. Modified from (Wold *et al.* 2001).

In order to utilize the multi-wavelength fluorescence spectra as an input matrix (X), the EEM of each sample is organized in the form of a vector as shown in Figure 2-7. Subsequently, these vectors are appended together into a matrix, where the columns correspond to different emission and excitation wavelength combinations excitation-emission pairs) and rows correspond to different samples.
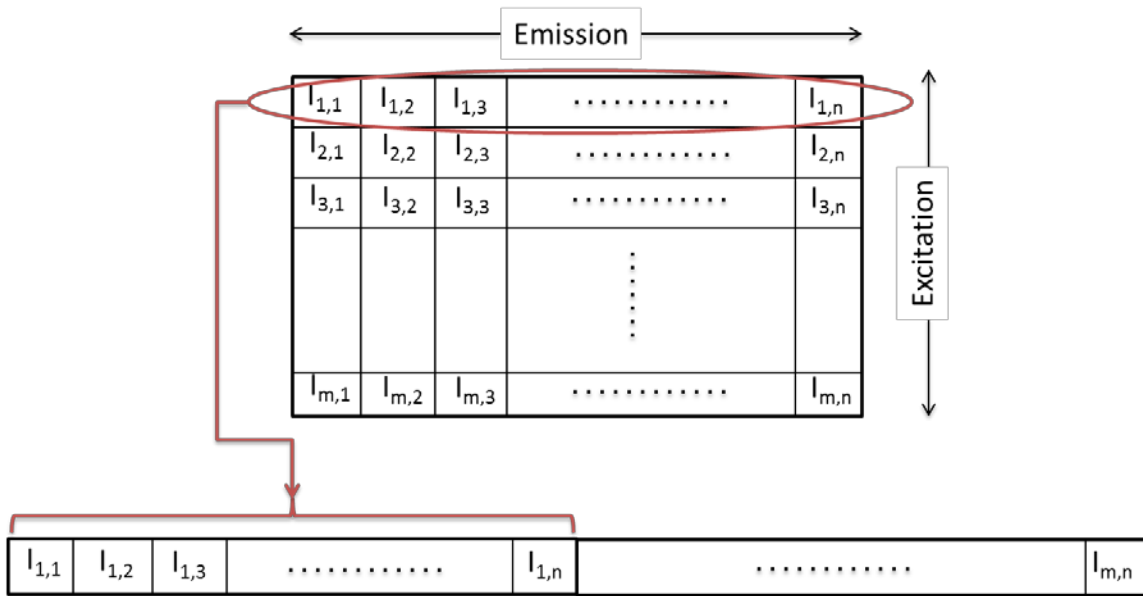
Figure 2-7. Schematic of organizing excitation emission matrix of a sample into a form of a row vector

2.4.2.1 Scaling

Pre-processing mathematical operations are often performed on the data in the input data set (X) and output data set (Y) matrices prior to projection to improve the accuracy of the regression. These pre-processing operations often help to enhance the extraction of relevant information or to improve the predictive capability of the resulting model on particular variables (Lourenco *et al.* 2012). Log and square transformations can be used to account for nonlinearity in the data. A derivative operation can be applied to suppress offset and a background slope. PLSR and methods based on projection in general are very sensitive to pre-scaling of the data. There is no requirement for scaling or even for applying the same scaling on X and Y. However, auto-scaling (centering and dividing each column by its variance) is generally applied when the relative influence of each variable is unknown *a priori*. In other words applying auto-scaling, assumes a similar impact for all variables involved in the regression model (Wold *et al.* 2001).

2.4.2.2 Number of latent variables

Although one might chose the number of LV`s as a function of the amount of variance that is explained by them, there is always a risk of over fitting. Cross Validation (CV) is considered a practical approach to quantitatively check the prediction quality of the model versus the number of

latent variables used and it has become widely accepted for calibrating PLSR and PCR models. CV procedures require that the data be divided into two sets, a calibration (Cal), i.e. training, set and a prediction (Pred), i.e. test, set. Partitioning of the data can be done in different ways as discussed by Shao, 1993. Using the calibration set, a regression model is developed and the Mean Squared Error (MSE) or Root Mean Square Error (RMSE) of calibration can be calculated. Then, the model is tested by using the prediction set, and the MSE-CV and RMSE-CV can be calculated, respectively to assess the prediction accuracy. These steps are done for a different number of LV`s and curves showing the RMSE-CV and RMSE-Cal vs. Number of LV`s can be used to choose an appropriate number of LV`s to capture a certain amount of variability. Alternatively, the model accuracy can be quantified by comparing $R_{CV}{}^2$ and $R_{Pred}{}^2$

2.5 Basics of Mab Glycosylation

Mabs are the fastest growing class of recombinant proteins in the biopharmaceutical industry (Butler 2005; Butler and Meneses-Acosta 2012) and are predominantly produced using mammalian cells as hosts. Antibodies, mostly immunoglobulins, are Y shaped glycoproteins typically composed of two heavy and two light polypeptide chains attached together with disulfide bounds, as depicted in Figure 2-8. The heavy chains and light chains are shown by purple and green colors respectively in Figure 2-8. The light chain and heavy chain of the antibody have two structural domains: i) constant domain and ii) variable domain, where the latter one, i.e. the antigen-binding domain, determines the affinity of a particular antibody towards an antigen. The top part of the Mab that includes the variable region of the heavy and light chain is known as the antigen-binding fragment (Fab) while the bottom domain, i.e. the two heavy chain terminals, is known as the crystallizable fragment (Fc). The Fab specifically binds to a certain antigen and the Fc domain either countervails the antigen or triggers a response in the host (Del Val *et al.* 2010).

Glycosylation is considered as a crucial post-translational modification of recombinant proteins, such as Mabs, and has been extensively specified as a key factor influencing the therapeutic properties of proteins (Del Val *et al.* 2010; Hossler *et al.* 2009; Stanley 2011). The extent and the structure of glycosylation, known as microheterogeneity, can profoundly affect the Mab's serum half-life and the response mechanisms evoked in the patient (Del Val *et al.* 2011; Hossler *et al.* 2007; Krambeck and Betenbaugh 2005; Umana and Bailey 1997). Hence, being able to predict the microheterogeneity in a timely manner is of interest for effective cell culture operation and control. Additionally, the extracellular metabolic state can also influence the glycan structure. As a result, a model relating the

extracellular culture conditions to the glycan structure can be implemented to calculate and implement operating conditions to achieve a desirable glycan structure.
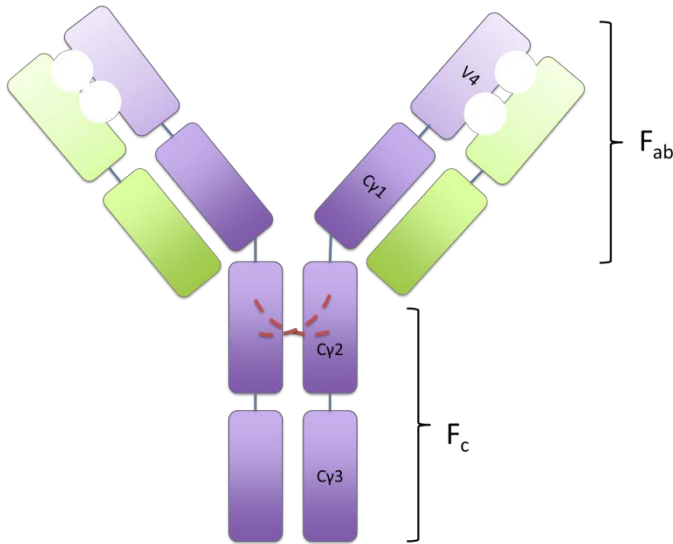
Figure 2-8. Schematic presentation of a monoclonal antibody (Mab).

2.5.1 Glycosylation Process

Most of the glycosylation process occurs within the intracellular Golgi apparatus. Based on their attachment sites, the glycans are characterized as O-linked (through serine or threonine link) and N-linked (through asparagine link) (Del Val *et al.* 2010). Several studies have emphasized the importance of N-linked glycans (Del Val *et al.* 2010; Hossler *et al.* 2009). O-linked glycosylation rarely occurs for antibodies (Hossler *et al.* 2009). The N-linked glycosylation can occur in both Fc and Fab regions, although the latter one reported to be unconventional (Del Val *et al.* 2010). The focus of the model presented in the current study is on the N-linked glycosylation occurring on the $C\gamma2$ domain of the heavy chain (as specified by red lines in Figure 2-8). The process of N-linked glycosylation initiates in the endoplasmic reticulum (ER) by attachment of an oligosaccharide (Glc3Man9GlcNAc2) to the nascent polypeptide chain (bottom left part of Figure 2-9). This is followed by proper folding of the protein and removal of three glucose residues and at least one mannose (top left part of Figure 2-9). Subsequently, the properly folded protein will be transferred to the Golgi apparatus. Prior to further expansion of glycan branches in the Golgi apparatus, more mannoses need to be removed. The glycosylation process occurring inside the Golgi can be simply defined as a sequential attachment/detachment of nucleotide sugars to the backbone of the Mab to

form a complex oligosaccharide. Examples of these attachments inside the Golgi are demonstrated in Figure 2-9 where different nucleotide sugars are sequentially connected to the oligosaccharide chain. The attachments/detachments of monosaccharides are triggered by a number of enzymes resulting in a large network of reactions and combination of oligosaccharide structures (Del Val *et al.* 2011; Hossler *et al.* 2009; Hossler *et al.* 2007). A software package referred to as GlycoVis (Hossler *et al.* 2006) can be utilized to visualize this network of glycoproteins and corresponding reactions.
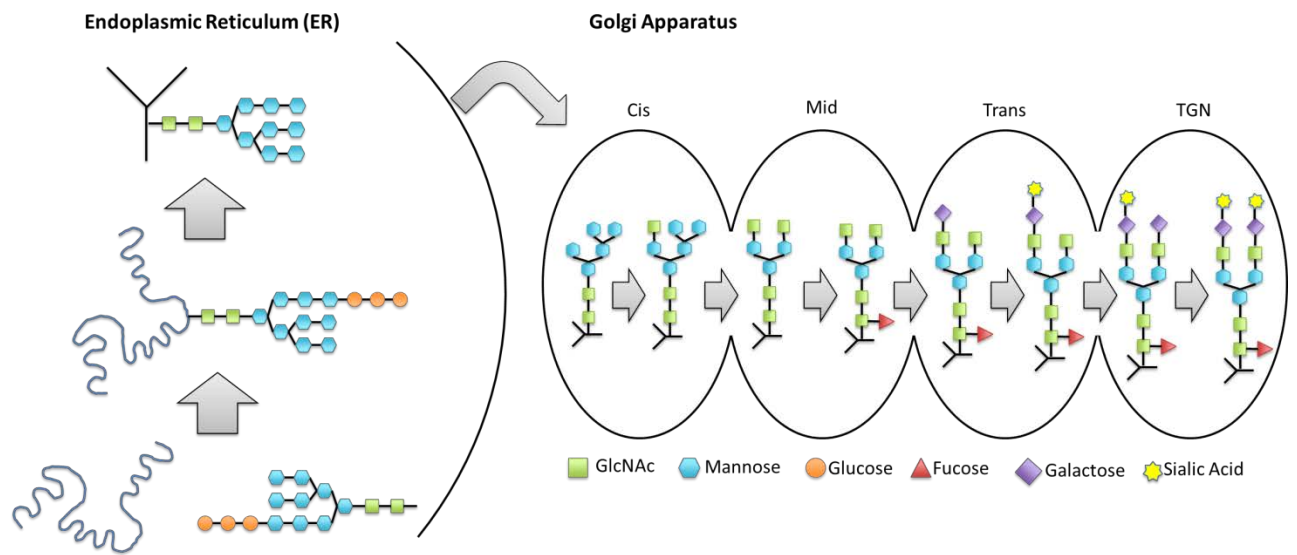
Figure 2-9. Schematic presentation of N-linked glycosylation initiating in the endoplasmic reticulum (left) and continuing in Golgi apparatus (right).

2.5.2 Mechanisms of transport in the Golgi Apparatus that were considered for model development

Structurally, the Golgi apparatus can be viewed as a set of biochemically heterogeneous membrane compartments, known as cisternae. These non-similar cisternae fall into four group: cis, medial, trans, and trans-Golgi network (TGN), in which a unique distribution of enzymes catalyze reactions in particular directions (Figure 2-9). Despite the debate regarding the nature of the Golgi apparatus, two main hypotheses, vesicular transport and Golgi maturation, have been used to formulate the glycosylation models for the Golgi apparatus (Del Val *et al.* 2011; Hossler *et al.* 2007).

The vesicular transport model assumes stationary compartments/cisternae with secretary cargo moving by vesicles, compartment by compartment. According to this assumption, the Golgi can be represented by a continuous stirred-tank reactors (CSTRs) in series where each reactor represents a cisterna containing a particular level of enzymes and the cargo (Mab) is transported across these reactors and is glycosylated according to its residence time within each reactor. In contrast, the Golgi maturation approach uses a constant secretary cargo assumption where the cargo is constant but each compartment or cisternae experiences maturation and transitions from early cisternae to a late cisternae status. Consequently, all Mabs will spend a similar time in the Golgi apparatus. In this case, plug flow reactors (PFRs) in series are used to represent the Golgi maturation process described above (Hossler *et al.* 2007). Although in earlier studies researchers favored the vesicular transport approach (Krambeck and Betenbaugh 2005; Umana and Bailey 1997), recent findings tend to support the Golgi maturation assumption (Del Val *et al.* 2011; Hossler *et al.* 2007). Thus, in the current work, the system is mathematically modeled using four plug flow reactors (PFRs) in series where each compartment represents one state of the Golgi during maturation.

2.5.3 Glycosylation Enzymes

N-glycosylation involves a series of successive reactions that are catalyzed by a small number of enzymes in the Golgi. Each enzyme is capable of catalyzing a number of reactions. Each glycan can be directed into different pathways by reacting with different enzymes. From a broad perspective these enzymes can be categorized into two classes: a) exoglycosidases that are enzymes acting on one substrate and detaching a mannose from the oligosaccharide chain and b) glycotransferases (GTs) that are enzymes acting on two substrates and attaching a particular monosaccharaide to the chain. GTs are reported in Stanley, (2013) to be over 250 for mammals, which can be classified in groups based on the type of monosaccharide they are acting on. Depending on the cell line some of these GTs are not expressed. Hence, the branches of the reaction network pertinent to these enzymes can be trimmed to reduce the complexity of the mathematical problem as explained later. Although some GTs can have more than one functional domain, a single type is assumed in this study. Due to the sequential nature of the glycosylation process, GTs are distributed along the Golgi (Hossler *et al.* 2007; Stanley 2011). Accordingly, the oligosaccharide structures can be profoundly affected by how the enzymes are localized. In the current study only reactions that occur in the Golgi apparatus are considered and reactions in the ER are assumed to proceed to completion.

Hossler *et al.* (2007) assumed 341 glycans and addressed this complex network of reactions with a *relationship-matrix* (Hossler *et al.* 2007). For a hypothetical reaction tree of five glycans the relationship-matrix is illustrated by Figure 2-10. Each non-zero element in the matrix appears as a possible reaction with its corresponding enzyme number.
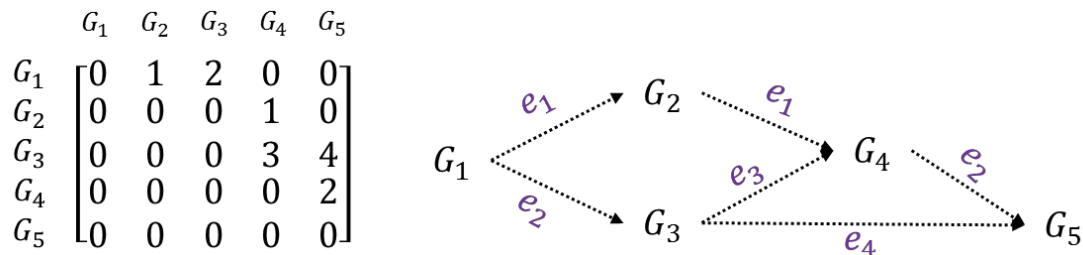
Figure 2-10. Relationship matrix (Hossler *et al.* 2006): a hypothetical example of 5 glycans and their corresponding relationship-matrix. G_is are representing glycans and $e_i s$ representing the type of enzyme.

Theoretical models for glycosylation have been proposed in the literature. However either no connection has been proposed in these models between the extracellular states and the glycosylation pattern (Del Val *et al.* 2011; Hossler *et al.* 2007) or they have not been properly verified by experimental data.

Chapter 3

Fluorescence-based Soft Sensor for *at situ* Monitoring of Chinese Hamster Ovary Cell Cultures[*]

Multi-wavelength fluorescence spectroscopy was investigated as a potential tool for use in monitoring key process variables that include: viable and dead cells, recombinant protein, glucose, and ammonia concentrations for Chinese hamster ovary (CHO) cells during cultivation. For the purpose of calibrating the fluorescence-based empirical model, cells were grown in batch mode with different initial glucose and glutamine concentrations. Spectrofluorometer settings were optimized to ensure reproducibility and accuracy of the acquired spectra. With the purpose of gaining qualitative insight into the evolution of the spectra, the trajectories of individual fluorophore peaks were studied during the cultivation process. Spectral changes related to biomass and secreted proteins were investigated by comparing the spectra at various stages during the downstream processing. A partial least square regression (PLSR) was used to formulate empirical models that related the input data set, i.e. the fluorescence excitation-emission matrix, to the actual state of the system including viable cell and dead cells and recombinant protein, glucose and ammonia concentrations. The models exhibited accurate prediction ability for the process variables of interest.

3.1 Introduction

Since the approval of recombinant insulin as the first recombinant drug product, the biopharmaceutical industry has experienced dramatic growth and advances in the production of recombinant proteins with humanized monoclonal antibodies (Mabs) surpassing the sales of other biologics (Butler 2005; Butler and Meneses-Acosta 2012). Chinese hamster ovary (CHO) cells are considered the dominant expression host for the commercial production of recombinant proteins due to their efficacious single-cell suspension growth (Wurm 2004), DNA transfection capabilities (Butler and Meneses-Acosta 2012; Wurm and Hacker 2011), inability for reproducing various human viral-pathogens (Wurm and Hacker 2011), and human-like glycosylation capacity. Rigorous process

---

[*] Ohadi, K., Aghamohseni, H., Legge, R. L. and Budman, H. M. (2014), Fluorescence-based soft sensor for at situ monitoring of Chinese hamster ovary cell cultures. Biotechnol. Bioeng., 111: 1577–1586.

32

analysis and optimization are required for scale-up to meet the five-fold proliferation of Mabs' demand that has occurred since 2000 (Butler 2005). Additionally, the essence of guaranteeing consistent production-quality reinforces the need for proper monitoring and control. The recently launched process analytical technology (PAT) by food and drug administration (FDA) has motivated manufacturers to utilize new *in situ* monitoring tools for real-time tracking of key process variables to ensure product quality (Rathore *et al.* 2010; Teixeira *et al.* 2009a). As a result of prolonged, laborious, and expensive measurement assays required for some of these key variables, manufacturers are urged to adopt alternative monitoring methodologies.

The nondestructive, noninvasive, and highly informative nature of electromagnetic spectroscopic techniques make these techniques attractive as potential tools for *in situ* monitoring of bioprocesses (Lourenco *et al.* 2012; Teixeira *et al.* 2009a). Among these techniques, multi-wavelength fluorescence spectroscopy is noteworthy due to its high sensitivity and selectivity, as compared to vibrational spectroscopic techniques, and that it encompasses a wide range of fluorophores that exist both in and outside of the cells (Lourenco *et al.* 2012; Marose *et al.* 1998). The adequacy of multi-wavelength fluorescence spectroscopy is supported because culture media and supernatant possess different intrinsic fluorophores, including amino acids (tyrosine and tryptophan), vitamins (pyridoxine and riboflavin), and cofactors (NAD(P)H and FAD) and that their fluorescent behavior changes through the cultivation process. These naturally fluorescent compounds are distinguishable by their excitation/emission wavelength maxima (Li and Humphrey 1991; Lindemann *et al.* 1998; Teixeira *et al.* 2009b). The quantum yield and Stokes shift of these compounds also exhibit a high dependency on their environment. Consequently, tracking culture-broth fluorescence over the course of the cell cultivation can provide qualitative and quantitative insight into the dynamics of the process (Li and Humphrey 1991; Teixeira *et al.* 2009b). Many studies emphasize the advantages of multi-wavelength spectrofluorometry over single-wavelength counterparts in handling peak overlaps, peak-maximum shifts, and signal quenching (Haack *et al.* 2004; Marose *et al.* 1998; Skibsted *et al.* 2001). Applications of multi-wavelength fluorescence spectroscopy for online monitoring of bacterial, (Hagedorn *et al.* 2003; Henneke *et al.* 2005; Jain *et al.* 2011; James *et al.* 2002; Skibsted *et al.* 2001), and yeast (Haack *et al.* 2004; Li and Humphrey 1991) cultivation have been well documented. Jain *et al.* (2011) monitored dry cell mass and non-fluorescing compounds (glucose concentration and carbon dioxide production rate) in recombinant bacterial cultures. Skibsted *et al.* (2001) argued that the ability to predict non-fluorescing compounds from the fluorescing ones is probably related to the stoichiometric relationship among them. Applications of multi-wavelength fluorometry for

mammalian cell cultivation have not been investigated as thoroughly. Rayan *et al.* (2010) used 2D-fluorescence fingerprints coupled with multivariate statistical techniques to predict mammalian growth-media performance for fed batch cultures. The capability of multi-wavelength fluorescence to track viable cell density and recombinant protein concentration, expressed in BHK21A, in batch and semi-batch modes has been described (Teixeira *et al.* 2009b). This same group also used a microwell plate reader for high-throughput prediction of biomass and secreted protein for a CHO culture (Teixeira *et al.* 2011).

Machine learning techniques have been utilized to generate empirical predictive-models (soft sensors) that relate the actual state of the system to 2D-fluorescence spectra obtained *in situ.* Mammalian cells can only incorporate, but not synthesize, vitamins and fluorescent amino acids that necessitate drawing an indirect inference about the key process variables from the fluorescence maps. Additionally, the large ratio between the number of fluorescence related variables, i.e. intensities measured at different combinations of excitation\emission wavelengths, to the number of samples, high collinearity among these intensities and the occurrence of measurement noise supports the use of chemometric methods for mining the fluorescence data (Rathore *et al.* 2010). Partial least square regression (PLSR) is the most frequently used supervised data mining technique capable of handling data sets that do not have a full rank (Wold *et al.* 1984).

The objective of this study was to develop an empirical predictive-model using multi-wavelength fluorescence spectroscopy coupled with PLSR as a tool for *at-line* monitoring of CHO cell cultivation in batch mode. Previous research documented the suitability of 2D-fluorometry for tracking viable cell and recombinant protein concentration trajectories (Teixeira *et al.* 2011; Teixeira *et al.* 2009b). The aim of the current study was to demonstrate the capability of multi-wavelength fluorescence spectroscopy for monitoring dead cell concentrations, a key nutrient (glucose) and inhibitor (ammonia) concentrations in addition to viable cell and Mab concentrations.

3.2 Materials and Methods

3.2.1 Cell Cultivation Conditions

A descendent of CHO-DXB11, specifically engineered by MabNet (National Science and Engineering Research Council Mab Network), was utilized as an expression host for producing EG2-hFc, a camelid-based Mab. The cell line is adapted to grow in a serum and protein free defined Biogro (provided by MabNet) media, supplemented with 0.9% HT (Invitrogen: 11067-030), under

suspension conditions with sub-culturing every 2 to 3 days with a seeding density of 0.25 million cells/ml to obtain seed cultures. Batch cultures, distinguished by their initial glucose and glutamine concentration (Table 3-1), were performed with an initial cell density of 0.2 million cells/ml in 500 ml polycarbonate shake-flasks (with 250 ml of working volume) and incubated at 37◦C with 5% CO_2 while agitated at 120 rpm. Viable and dead cell density (using hemocytometer and trypan-blue exclusion) and fluorescence maps were obtained at situ on a daily basis. To conduct off-line analytical measurements of recombinant protein and metabolite concentrations, additional samples were collected, centrifuged at 300×g for 10 min, and the supernatant stored at -20˚C.

Table 3-1. Batch cell cultivation notations and experimental design

Culture Notations	Media Type	Initial Concentration	
		Glucose (mM)	Glutamine (mM)
A	I	25	0
B	I	25	4
C	I	25	4
D	I	25	8
E	I	45	0
F	I	45	4
G	I	45	8
H	II	25	4
I	II	35	4
J	II	45	8

3.2.2 Analytical Methods

The extracellular recombinant protein concentration was measured using an Enzyme-Linked Immunosorbent Assay (ELISA) method developed by MabNet. 96-well plates were coated with an anti-human IgG1 (Fc specific) produced in goat (Sigma Aldrich, ON) and detected using IgG (Fc specific)-peroxidase antibody produced in goat (Sigma Aldrich, ON) after the addition of samples and standard. Addition of peroxidase's substrate, i.e. TMB, results in a colored product. The reaction was stopped using 2M sulfuric acid and the absorbance read in a Biotek Synergy 4 multi-plate reader at 450 nm. Glucose concentrations were determined utilizing a Megazyme assay kit (Cedarlane) with

the absorbance measured at 510 nm using a Biotek Synergy 4 multi-plate reader. An Ammonia Ion-Selective electrode connected to a pH/ISE meter model 710A of VWR was used to determine the ammonia concentration.

3.2.3 Fluorescence Map Acquisition

Fluorescence maps of the samples were acquired using a Cary Eclipse spectrofluorometer (Palo Alto, CA) equipped with a Peltier multi-cell holder using quartz cuvettes (Thermo Scientific, ON) at room temperature. The spectra were obtained over an excitation range from 240 to 500 nm at 10 nm increments and emission wavelengths of 280 to 600 nm with 2 nm increments. The resulting 2D-fluorescence maps were organized into a matrix referred to as excitation emission matrix (EEM). An initial investigation was conducted to obtain optimal settings of the instrument to guarantee accuracy and reproducibility of the signals using a three-factor Box-Behnken experimental design involving different levels of the photomultiplier tube (PMT) voltages, slit widths (SW), and scanning rates performed on Biogro media. The results of this optimization revealed that a higher excitation intensity results in a higher signal to noise ratio but the magnitudes of the peaks are limited by the detector's measuring range. The impact of PMT voltage and SW were found to be significant on the repeatability of the spectra in contrast with a small effect of the scanning rate. This corroborates the earlier findings observed for drinking water applications (Peiris *et al.* 2009). As a result, a PMT of 600V with a SW of 5 nm for both excitation and emission, and a scanning rate of 1200 nm/min were identified as optimum instrument settings for the media employed in this study. To determine the occurrence of signal quenching, a series of experiments were conducted with media at different levels of dilution using phosphate buffer saline (PBS). These experiments revealed the occurrence of signal quenching for peaks corresponding to tyrosine, tryptophan, and pyridoxine. A dilution corresponding to a sample to buffer ratio of 1:19 was found to be optimal to prevent the occurrence of signal quenching and was used for the remaining experiments.

3.2.4 Purification Effect

A set of experiments were conducted to identify whether the presence of cells and recombinant protein have a direct impact on the fluorescence map. This is relevant for understanding whether the predictions of cell and Mab concentrations to be presented later are based on a direct correlation or on an indirect inference from the supernatant composition. To investigate the contribution of the presence of the cells in the supernatant on the spectra, experiments were conducted by removing the

36

cells using centrifugation at 300×*g* for 10 min and collecting the spectra before (Stage I) and after centrifugation (Stage II). To assess correlations between the spectra and the recombinant protein concentration, culture samples obtained after centrifugation were eluted through a Protein A column (Sigma Aldrich, ON), which can specifically bind Mab, and the spectra before (Stage II) and after passage through the column (Stage III) were acquired.

3.2.5 Data Analysis

PLSR was exploited as the machine learning technique to generate predictive empirical-models that relate the input data set, i.e. fluorescence map, to the actual state of the system including viable and dead cells, glucose, ammonia and recombinant protein (Mab) concentrations. The principle behind PLSR is to find a compact set of new coordinates—for a highly redundant input data space X—onto which the data are projected, i.e. latent variables (LVs), that captures both the variance in X space and best describes the output data Y. In other words, a multi-dimensional hyper-plane, defined by the latent variables, in the X space is found such that the projection of independent variables on it (i.e. scores) correctly captures the input data X, and at the same time, accurately describes the output data Y (Geladi and Kowalski 1986; Wold *et al.* 1984; Wold *et al.* 2001). In the current work the input data X contained the fluorescence spectra at different emission excitation wavelengths and the output space Y is composed of the variables to be predicted: viable and dead cells, Mab, glucose and ammonia concentrations. For the purpose of training the PLSR models, the EEM elements of each sample were organized into a one row vector and then, the rows corresponding to all the samples were appended one above the other to form the matrix of input data (X). To eliminate background noise effects, the fluorescence map of the phosphate buffer saline (PBS) was obtained on a daily basis and subtracted from the samples' spectra. In order to prevent a dominant effect of fluorophores with high intensity peaks, as compared to parts of the spectra with lower intensity, data preprocessing was performed to improve the accuracy and to focus the predictive capability of the resulting model on particular variables. Similar weight was assigned to every variable corresponding to a particular excitation-emission pair using a column-wise auto-scaling (mean centering and dividing by variance) of input and output data matrices. An orthogonal projection to latent structures (O-PLS) was adopted to enhance the interpretability of the resulting predictive models through removing variation in input data that is orthogonal to the output data (Trygg and Wold 2002). In this study the SIMPLS algorithm (Dejong 1993) was pursued in preference to its computationally more expensive counterpart, NIPALS (Geladi and Kowalski 1986). To suppress model over-fitting, an optimal number of LVs was sought

37

by utilizing cross validation and minimizing the root mean square error of cross validation (RMSECV) as a measure of out of sample error and generalization capacity. Due to the time dependency, errors for samples in each batch are not utterly random and independent of each other but rather correlated; thus, an improper cross-validation (CV) procedure may result in a biased prediction of out of sample error and increase the sensitivity of the empirical model to noise. To address these issues the leave one batch out cross validation procedure (LOBO-CV) was utilized in this work (Hagedorn *et al.* 2003). Chemometric analyses were performed using the PLS-Toolbox 7.0.3 (Eigenvector Research Inc., Manson, WA) running in the MATLAB 8.0.0 (MathWorks, Natick, MA) environment.

3.3 Results and Discussion

Empirical predictive PLSR based models for viable and dead cells, recombinant protein, glucose, and ammonia concentrations were calibrated and tested using data collected from a set of batch cultures. Various initial glucose and glutamine concentrations (Table 3-1) were utilized with the objective of ensuring sufficient variability in culture conditions so as to produce robust models with predictive ability over a large range of operating conditions. Another factor that was evaluated was the effect of variations of the basal media on model prediction quality by growing cells in two different batches of media, to be referred as Media I and II (Table 3-1).

3.3.1 Analysis of Individual Fluorophore Peaks

According to the literature, different regions of the fluorescence map can be attributed to particular intrinsic fluorescent compounds (Lindemann *et al.* 1998; Ryan *et al.* 2010; Teixeira *et al.* 2009b). Although the merits of multi-wavelength, compared to single-wavelength spectrofluorometry, have been reported and were used for the current work (Haack *et al.* 2004; Marose *et al.* 1998; Skibsted *et al.* 2001), the trajectories of individual peaks (fluorophores) during the cultivation process was first investigated to gain a qualitative insight into the dynamics of the system. Accordingly, the temporal evolution of peak intensities were compared (Figure 3-1a-d) for two batches that were initiated with different glutamine concentrations (0 and 4 mM) and the same glucose concentration (25 mM). It is apparent from Figure 3-1a and b that the fluorophores exhibited similar trends for these two batches.
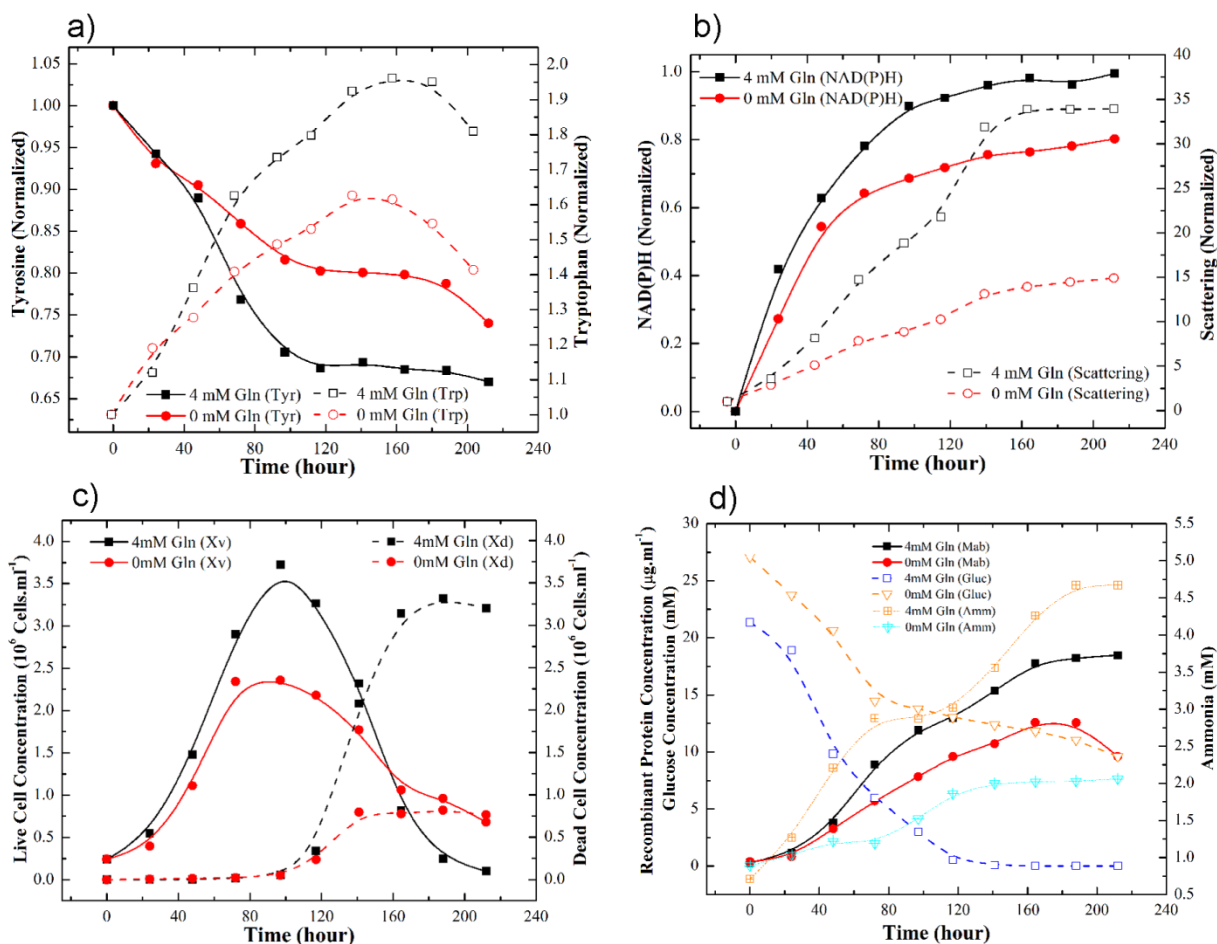
Figure 3-1. Comparison of trajectories of different peak intensities and attributes for cultures with the same initial glucose concentration (25 mM) and different glutamine concentrations (0 and 4 mM). a) Left axis: Trajectories of tyrosine peak ($\lambda_{ex} / \lambda_{em} = 270/305$ nm) intensity, Right axis: Trajectories of tryptophan peak ($\lambda_{ex} / \lambda_{em} = 300/350$ nm); b) Left axis: Trajectories of NAD(P)H peak ($\lambda_{ex} / \lambda_{em} = 360/450$ nm) intensity, Right axis: Trajectories of an scattering peak ($\lambda_{ex} = \lambda_{em} = 300$); c) Left axis: Trajectories of viable cell concentrations, Right axis: Trajectories of dead cell concentrations; d) Left axis: Trajectories of recombinant protein and glucose concentrations, Right axis: Trajectories of ammonia concentrations.

The tyrosine ($\lambda_{ex}/\lambda_{em} = 270/305$ nm) peak decreases (solid lines in Figure 3-1a) possibly due to an uptake of this amino acid by the cells for biomass synthesis and for protein production but then it plateaus when the maximum cell concentration was achieved (Figure 3-1c). This supports the notion that the uptake of tyrosine is growth associated and correlates with its higher consumption for the batch initiated with 4 mM glutamine which exhibits much greater biomass growth compared to the

39

culture initiated with no glutamine (Figure 3-1c). Mammalian cells are unable to synthesize tryptophan (Teixeira *et al.* 2009b) and the increase in the tryptophan peak could be explained by the incorporation of this amino acid into the protein structure. While tryptophan exposure to a hydrophobic environment results in increased emission-intensity, exposure to a hydrophilic environment or abundance of quenchers diminishes emission-intensity (Callis and Liu 2004; Teixeira *et al.* 2011). The tryptophan peak-intensity at $\lambda_{ex}/\lambda_{em} = 300/350$ nm (shown by dashed lines in Figure 3-1a) exhibited a more pronounced increase for the culture that commenced with the higher initial glutamine concentration in comparison to the culture with no glutamine other than residual glutamine from the inoculum. This conclusion further corroborates the contribution of tryptophan to biomass and recombinant protein concentrations, since higher cell growth and Mab productivity was observed for the batch initiated with 4 mM glutamine (Figure 3-1c-d). The slight decline in tryptophan intensities at the end of the culture might be explained by protein denaturation or a large amount of cell debris at the end of the cultivation process which can cause signal quenching. The peak related to NAD(P)H $\lambda_{ex}/\lambda_{em} = 360/450$ nm exhibits a temporal increase during the cultivation process until the culture reaches the stationary phase and remains approximately constant afterwards (Figure 3-1b). Production of NADH is mostly attributed to the glycolysis and citric acid cycles (Marose *et al.* 1998) and as such is proportional to the number of growing cells. Notably, the NAD(P)H peak`s slope declines after approximately 80 hours when glucose is consumed for the batch with 4 mM initial glutamine or glucose consumption is diminished for the 0 mM initial glutamine culture (Figure 3-1d). Accordingly, NAD(P)H might be used to track the process metabolic-state (Haack *et al.* 2004). Rayleigh scattering corresponds to $\lambda_{ex}/\lambda_{em}$ and is often reported to result from cell-solution interface and/or suspended particles (Christensen *et al.* 2006; Deshpande 2001), that might explain its increasing trajectory illustrated with dashed lines in Figure 3-1b corresponding to a scattering peak obtained at 300 nm over the course of the culture. Despite the reduction in the number of viable cells after the transition to the post-exponential phase, a surge was observed in the scattering peak intensity, possibly due to cell debris.

After obtaining the fluorescence spectra at different stages of purification, differences in spectra between stage I and II and between stages II and III were calculated (Figure 3-2a and b respectively; day 9 of culture B) and normalized by the maximum intensity to determine the relative changes before and after centrifugation and before and after purification through the Protein A column. From the difference of spectra for stage I and II, it was found that the biomass is significantly correlated with the scattering region intensities as well as to the tryptophan peak (Figure 3-2a-spectra of stage II

40

subtracted from I). By comparing the spectra of the sample before and after elution through the Protein A column it was observed that Mab only contributes slightly to the scattering and tryptophan intensities (Figure 3-2b). Similar dependencies were observed with different relative magnitudes for various days during cell culture (results not shown).
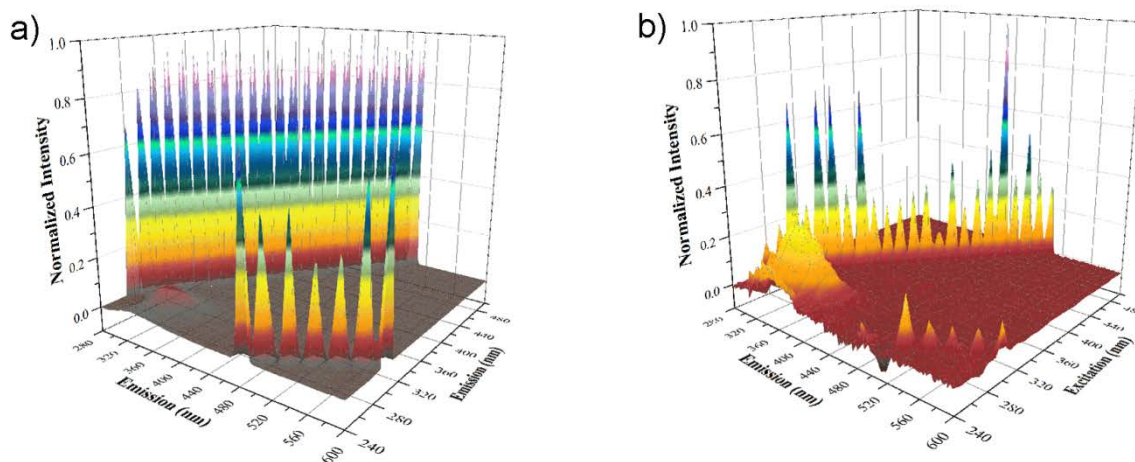
Figure 3-2. a) Surface plot of difference spectra for after centrifugation (Stage II) subtracted from before centrifugation (Stage I); b) Surface plot for difference spectra of after Protein A (Stage III) subtracted from after centrifugation (Stage II).

3.3.2 Predictive Models

The acquired spectra were utilized to formulate models for dynamic state estimation of various culture variables during the cultivation process using the PLSR method. As a direct correlation does not exist between the multi-wavelength fluorescence map and system-states, indirect inferences were drawn using PLSR. Individual models were developed for each attribute to facilitate interpretation
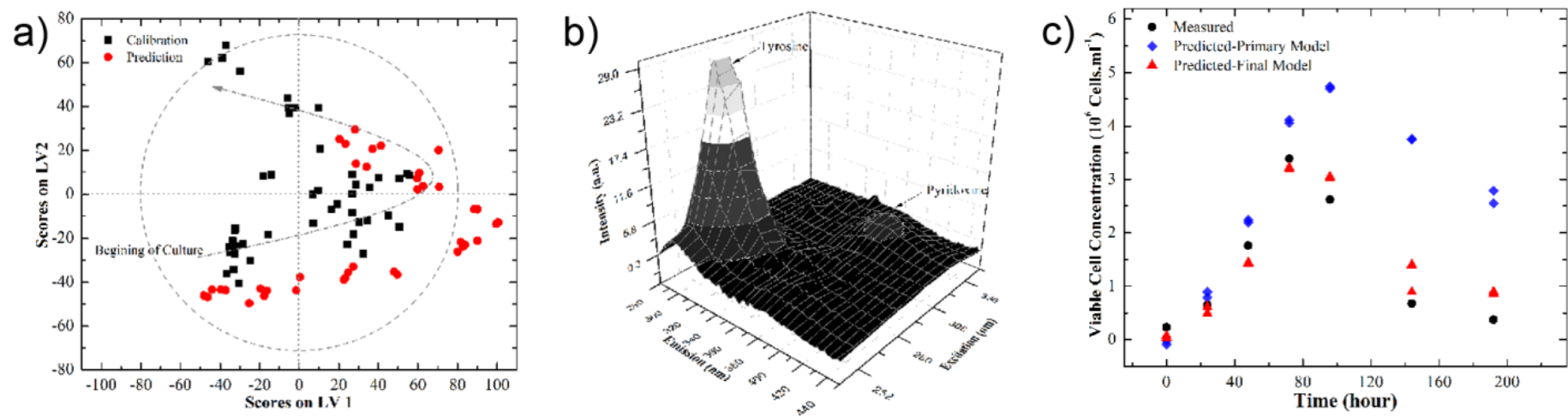
Figure 3-3. a) Scores of latent variable 2 vs. 1 for the viable cell concentration predictive-model (trained by cultures grown in Media I and tested with cultures grown in Media II); b) Surface plot for the fluorescence spectra of Media I subtracted from Media II; c) Comparison of viable cell density trajectories of culture I for two different model predictions and experimental results.

As a primary assessment, prediction accuracy of a model for viable cell concentrations, trained using data for batches grown in Media I, was tested with data for cultures grown in Media II (see Table 3-1 for notations). Using 5 LVs, selected based on LOBO-CV, the calibrated model is capable of capturing 92.9% of variation in output data with R^2_{CV}=81.5%; however, the model failed to accurately predict the viable cell concentrations (R^2_{CV}=49.5%) for cultures H, I, and J that were conducted with Media II. Figure 3-3a shows that the scores for cultures grown in Media II exhibit higher and lower values for LV1 and LV2 respectively than the corresponding scores obtained from the training set. This difference becomes more apparent at the end of the culture. The source of this deviation is evident from Figure 3-3b which shows the difference in spectra for pure Media I subtracted from pure Media II. A higher intensity peak is observed in the region corresponding to tyrosine and moderately higher in the region of pyridoxine for Media II. It was encouraging that despite this discrepancy between the pure media the model is still capable of properly predicting the trend in viable cell concentration (Figure 3-3c) implying that the fluorophores follow similar behavior in both media and reinforcing the capability of multi-wavelength fluorescence spectroscopy as a monitoring tool.

In order to formulate accurate models, the calibration of the models for each of the attributes was performed using data from cultures grown in both Media I and II. For a rigorous evaluation of prediction quality and out of sample error, data for one batch grown in Media I and one batch grown in Media II (culture E and I) were set aside for testing.

Table 3-2. PLSR results and goodness of fit for target process variables based on LOBO-CV strategy and prediction capability of model using cultures E and I.

Target Variable	Number of LVs	Calibration		Cross-Validation		Testing (Prediction)	
		$RMSE_{Cal}$	R^2_{Cal}	$RMSE_{CV}$	R^2_{CV}	$RMSE_P$	R^2_P
Viable Cells	7	0.213	96.3%	0.437	84.7%	0.383	89.2%
Dead Cells	5	0.227	94.5%	0.303	90.4%	0.174	90.4%
Mab (Eg2)	7	1.284	96.7%	2.157	87.2%	2.085	92.9%
Glucose	6	0.078	94.3%	0.120	86.7%	0.100	87.7%
Ammonia	4	0.262	93.6%	0.387	86.2%	0.540	93.5%

3.3.2.1 Viable cell concentration

Figure 3-4a shows the RMSE of calibration and cross validation for viable cell concentration versus the number of LVs. It clearly indicates that 7 LVs minimize the RMSE-CV to 0.44 million cell.ml^{-1} that corresponds to less than 25% of the average viable cell density. Considering a measurement error of approximately 15%, this result was deemed reasonable. The resulting model can capture 89.2% of the variability of the test set (Table 3-2). As illustrated in Figure 3-4b, the model accurately predicts both the calibration and test data sets with the residuals being normally distributed. Thus, by including cultures grown in Media II in the calibration set, the prediction capability of the model improved significantly (see Figure 3-3c for culture I). In Figure 3-4b most of the deviation of the predicted from the measured values is observed at the end of the culture, possibly due to cell debris that interferes with the fluorescence measurements and cell counting. The first two dimensions (LVs) accounts for over 60% of variance in the input data. LV1 scores increases to reach a maximum at a time pertinent to each culture's highest viable cell concentration, and then decreases until the end of the culture, i.e. LV1's trajectory resembles the temporal changes of viable cell concentration for each culture (Figure 3-4c). The contribution of the regions of the spectra for a specific LV is carried by its corresponding loading. Figure 3-4d is the contour plot of the first loading, obtained by performing the inverse of the reshape explained earlier. For the first loading, a strong correlation is observed with the region pertinent to tyrosine and tryptophan, while the excitation emission pairs corresponding to the cofactors are not as crucial, which corroborates a conclusion drawn by Marose *et al.* (1998) that viable cells are better correlated with the tyrosine and tryptophan regions of the spectra. Additionally, the contribution of the Rayleigh scattering should not be neglected, as it encompasses information regarding suspended particles. In contrast to the moderate impact of scattering on the first loading, the second loading is dominated by the region corresponding to the incident light (result not shown) which explains the persistent increase in the second LV's scores. A large contribution by the regions associated with tryptophan, NAD(P)H, and the flavines is observed in the regression coefficient for viable cell concentration predictive-model (Figure C- 1-Appendix C). These findings are in agreement with the earlier result regarding the direct contribution of cell mass to the spectra (Figure 3-2a).
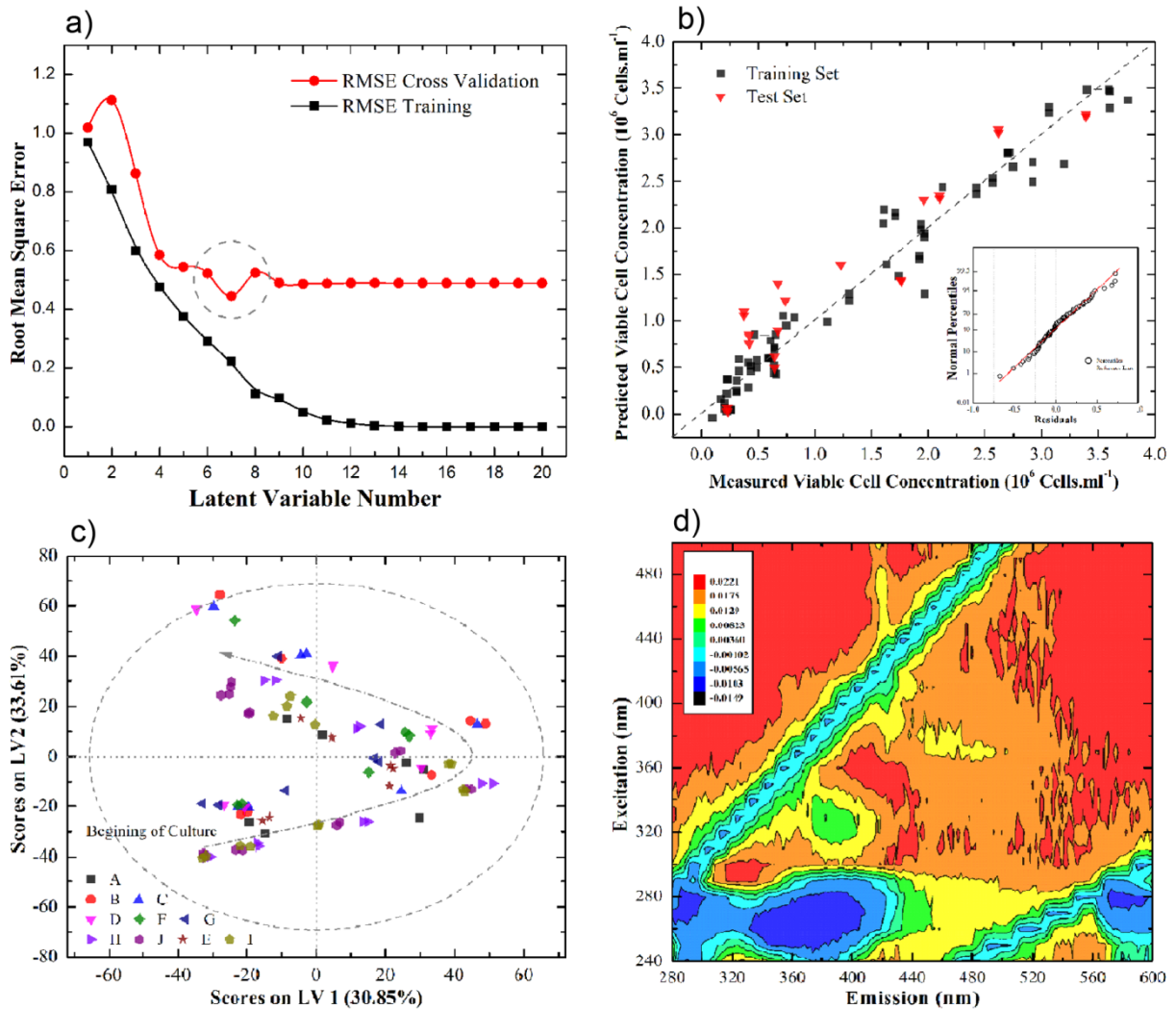
Figure 3-4. Results of predictive model for viable cell concentration: a) Evolution of RMSE for training and cross-validation set versus the number of latent variables; b) Model prediction vs. measured values for training and test set and normal percentiles of the residuals (inset); c) Latent variable (LV) 2 vs. 1; d) Contour plot of the first loading.

3.3.2.2 Dead cell concentration

Dead cell concentration is also a critical variable to be monitored as it provides insight into the growth stage of the culture and could be used for triggering feedback control corrections to mitigate death. A predictive model, established for dead cell concentration by using an approach similar to viable cell count, exhibited an adequate prediction quality for calibration, validation, and test sets, being summarized in Table 3-2 and Figure 3-5a. The increased prediction error for higher dead cell counts towards the end of the culture can possibly be attributed to cell lysis that results in an increase

in the number of suspended particles (Figure 3-5a) increasing both cell counting and signal acquisition noise. Figure 3-5b provides the regression coefficient of the PLSR model for dead cell concentration that is dominated by the region of the spectra pertinent to light scattering.
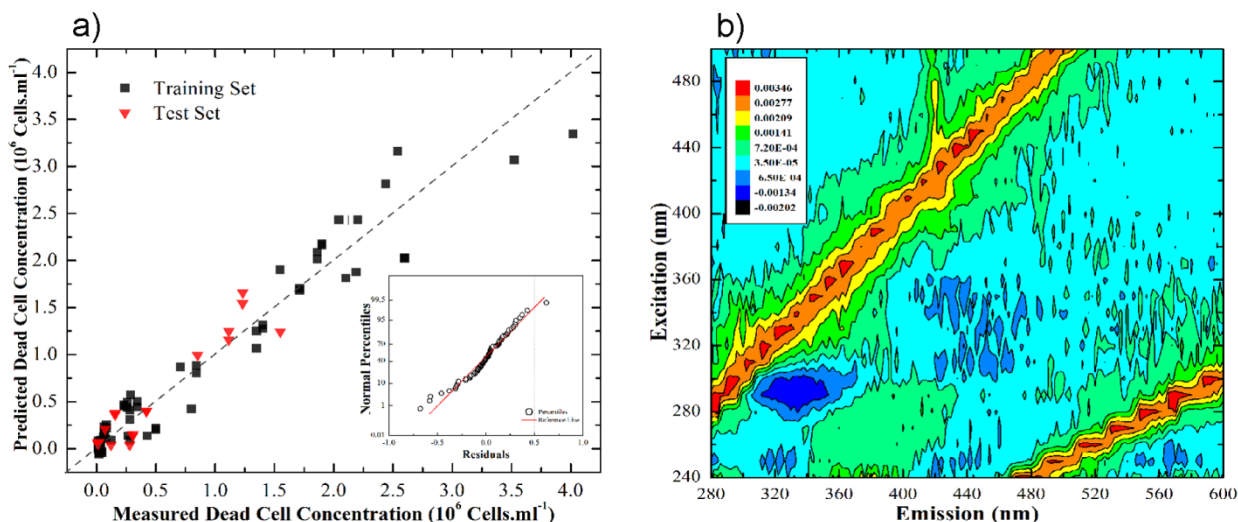
Figure 3-5. Results of predictive model for dead cell concentration: a) Model prediction vs. measured values for training and test set and normal percentiles of the residuals (inset); b) Contour plot of the regression coefficient.

3.3.2.3 Recombinant protein concentration

The PLS model—targeted at capturing the dynamics of the recombinant protein production—demonstrated an accurate calibration and cross-validation results with coefficient of determinations (R^2) equal to 96.7% and 87.2% respectively and predicted the test set with a precision of 92.9% (Table 3-2) with $RMSE_P$ equal to approximately 20% of the average Mab produced. It should be noted that measurement error of ELISA is approximately 15%. In Figure 3-6a the predicted vs. measured data fall very closely to the 45° line. The adequacy of this model is acceptable considering the ELISA's measurement error. LV1 was found to explain over 50% of variation in X and 94% of variation in Y with its corresponding loading, shown as a contour plot in Figure 3-6b, being dominated by a region associated with tyrosine, tryptophan and pyridoxine. A slight deviation from the normal distribution is observed in Figure 3-6a for the residuals, which might be due to a negligible nonlinear relation between Mab and the fluorescence spectra.
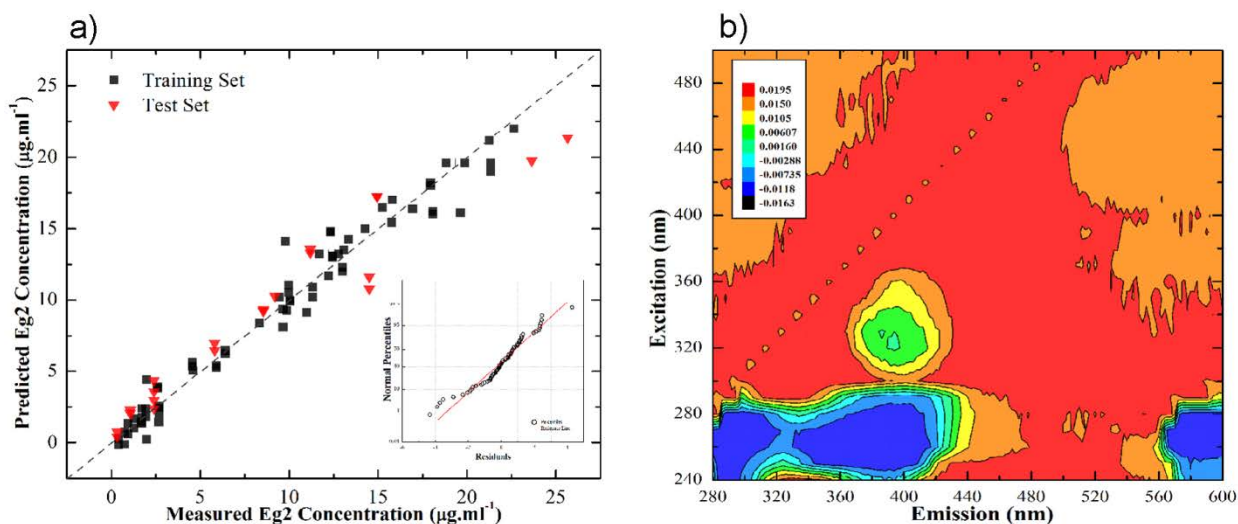
Figure 3-6. Results of predictive model for secreted recombinant protein concentration: a) Model prediction vs. measured values for training and test set and normal percentiles of the residuals (inset); b) Contour plot of the first loading.

3.3.2.4 Glucose and ammonia concentrations

In cell culture operations it is highly desirable to monitor key nutrients and inhibitors such as glucose and ammonia, as they hold a vital role in culture sustainability. Since these metabolites are not intrinsically fluorescent it is not possible to directly track their concentrations in culture broth using multi-wavelength fluorescence spectroscopy. However, since these metabolites are correlated through stoichiometric relations with fluorescent compounds, it is possible to infer their consumption or production rate (i.e. their normalized evolvement) from changes in the spectra (Skibsted *et al.* 2001). With this logic, the fluorescent spectra should be correlated to changes in glucose and ammonia with respect to their known initial concentrations. For glucose, the change with respect to initial concentration is normalized for model training by subtracting the glucose concentration from the initial batch concentration and dividing by the initial concentration. For ammonia, a non-normalized change with respect to the measured initial concentration (approximately equal to zero) is used for model training. The normalization was used for each batch prior to auto-scaling of the data.

Table 3-2 summarizes the prediction models obtained for glucose and ammonia concentrations and Figure 3-7a-b provides the model prediction versus measured values. Probability plots showing the normalized percentiles as a function of the residuals are within Figure 3-7a-b. Based on the results, fluorescence spectra exhibited a potential to accurately fit the calibration data set and predict the test

set with the residuals following a normal trend for both ammonia and normalized glucose concentrations (Figure 3-7a-b).
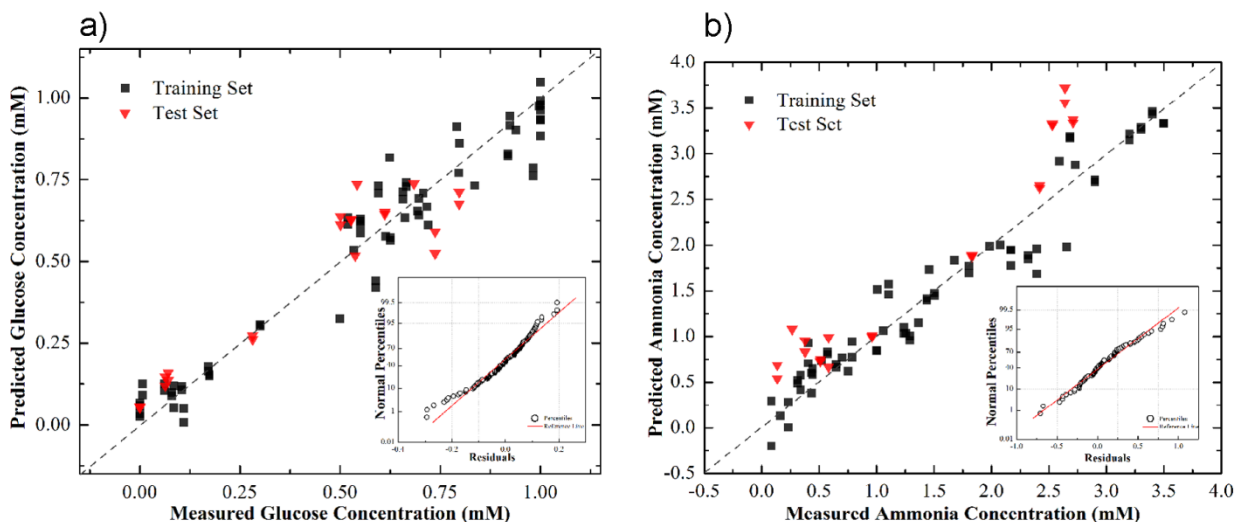
Figure 3-7. Results of predictive model for glucose and ammonia concentrations: a) Normalized glucose concentration model prediction vs. measured values for training and test set and normal percentiles of the residuals (inset); b) Ammonia concentration model prediction vs. measured values for training and test set and normal percentiles of the residuals (inset).

3.4 Conclusion

Evidence is presented which demonstrates the potential for multivariate PLSR based models to monitor different mammalian culture variables from 2-D fluorescence spectra. As a first step, to guarantee reproducibility and accuracy of the acquired spectra, the spectrofluorometer settings were optimized by performing a three level Box-Behnken design. For the purpose of better understanding the relationship between dynamics of mammalian cell culture and changes in the spectra over time, time profiles of specific fluorophore peaks were studied. The impact of biomass and secreted protein on the multi-wavelength spectra was investigated by comparing the spectra acquired during downstream processing that included centrifugation and Protein A purification. Empirical predictive PLSR based models were developed that provided accurate estimates of viable and dead cell concentrations, recombinant protein, glucose, and ammonia concentrations. This combined approach is an attractive candidate for *in situ* monitoring of mammalian cell culturing.

3.5 Acknowledgements

Chapter 4

Development of a Soft-Sensor Based on Multi-Wavelength Fluorescence Spectroscopy and a Dynamic Metabolic Model for Monitoring Mammalian Cell Cultures[*]

A soft-sensor based on an Extended Kalman Filter (EKF) that combines data obtained using a fluorescence-based soft-sensor with a dynamic mechanistic model, was investigated as a tool for continuous monitoring of a Chinese hamster ovary (CHO) cell cultivation process. A standalone fluorescence based soft-sensor, which uses a combination of an empirical multivariate statistical model and measured spectra, was designed for predicting key culture variables including viable and dead cells, recombinant protein, glucose, and ammonia concentrations. The standalone fluorescence sensor was then combined with a dynamic mechanistic model within an EKF framework, for improving the prediction accuracy and generating predictions in-between sampling instances. The dynamic model used for the EKF framework was based on a structured metabolic flux analysis and mass balances. In order to calibrate the fluorescence-based empirical model and the dynamic mechanistic model, cells were grown in batch mode with different initial glucose and glutamine concentrations. To mitigate the uncertainty associated with the model structure and parameters, non-stationary disturbances were accounted for in the EKF by parameter-adaptation. It was demonstrated that the implementation of the EKF along with the dynamic model could improve the accuracy of the fluorescence-based predictions at the sampling instances. Additionally, it was shown that the major advantage of the EKF-based soft-sensor, compared to the standalone fluorescence-based counterpart, was its capability to track the temporal evolution of key process variables between measurement instances obtained by the fluorescence-based soft-sensor. This is crucial for designing control strategies of CHO cell cultures with the aim of guaranteeing product quality.

4.1 Introduction

The past few decades have witnessed significant growth and development in biopharmaceutical industries that produce recombinant proteins via mammalian cell cultivation. Among these protein

---

products, monoclonal antibodies (Mabs) constitute the fastest growing segment using Chinese hamster ovary (CHO) cells as the dominant expression host (Butler 2005; Butler and Meneses-Acosta 2012). The importance of real-time monitoring of cell cultivation processes to guarantee productivity and consistent production-quality has been repeatedly emphasized (Pohlscheidt *et al.* 2009; Teixeira *et al.* 2009a). Due to time-intensive, labor-intensive, and costly nature of the assays available for some of the crucial cell culture variables, soft-sensors have been often proposed and adopted as an alternative approach to conventional measurement analytical techniques (Gernaey *et al.* 2012; Kadlec *et al.* 2009). Soft-sensors can be classified into two classes: model-driven based on mechanistic models and data-driven based on empirical predictive models (Kadlec and Gabrys 2009).

Multi-wavelength fluorescence spectroscopy has shown promise as a potential non-destructive and non-invasive tool for online or offline monitoring of bacterial (Hagedorn *et al.* 2003; Skibsted *et al.* 2001), yeast (Haack *et al.* 2004; Li and Humphrey 1991), and mammalian (Ohadi *et al.* 2014a; Ryan *et al.* 2010; Teixeira *et al.* 2009b) cell cultures. Multivariate statistical techniques have been utilized to draw an indirect inference about the key process variables from 2D-fluorescence spectra (Lourenco *et al.* 2012) and formulate an empirical predictive model (data-driven soft-sensor). Thus, this soft-sensor is strictly developed based on data and it does not use mechanistic knowledge of the process; thus, performing as a black-box. Despite the precision of such data-driven soft-sensors to predict the process attributes close to the calibration conditions of the model, their accuracy diminishes for process conditions that substantially deviate from the region used for model calibration. In other words, data-driven soft-sensors often lack proper extrapolation capabilities (Shioya *et al.* 1999). Accordingly, a large data set is required for model calibration in order to obtain a robust model over a large range of operating conditions. Measurement noise in fluorescence data and missing data are additional issues that reduce the reliability of the data-driven model predictions (Luttmann *et al.* 2012).

Lack of mechanistic insight into the process requires utilization of model-driven soft-sensors (white-box model), specifically at the process development stage when scarce process data are available (Pohlscheidt *et al.* 2009), whereas a dynamic mechanistic model can be implemented for forecasting the temporal evolution of key culture variables. Structured models are generally preferred, over their unstructured counterparts, since they properly account for the actual metabolic reactions occurring in the process thus often resulting in more accurate models (Gernaey *et al.* 2012). Metabolic flux analysis (MFA) is a well-established technique in the biopharmaceutical industries

51

used in variety of applications such as simulation, optimization, and control (Zamorano *et al.* 2013) to elucidate *in vivo* metabolic states of the cell cultivation as well as cell growth dynamics. In spite of the advantages of white-box models in capturing the dynamics of the process, it is challenging to formulate and calibrate such models due to complexities of mammalian cell metabolism (Kadlec *et al.* 2009; Shioya *et al.* 1999). Additional challenges arise from the significant uncertainty associated with the large number of parameters and model-structure, which results in susceptibility of the model predictions to high levels of noise (Wang *et al.* 2010).

To address the respective limitations of black- and white-box models, recursive state observers can be implemented to combine a dynamic metabolic model with a data-driven soft-sensor (Dochain 2003). The Kalman filter (KF) and its nonlinear version, the extended Kalman filter (EKF), are recognized as the most extensively used state estimator implemented in process industries (Dewasme *et al.* 2013; Dochain 2003). The accuracy of a KF-based soft-sensor is significantly impacted by the accuracy of the mechanistic model (de Assis and Maciel 2000); thus, reinforcing the superiority of structured metabolic models over their unstructured counterparts.

A data-driven soft-sensor, based on multi-wavelength fluorescence spectroscopy, was proposed previously by the authors (Ohadi *et al.* 2014a) which exhibited a potential for *at situ* monitoring of CHO cell cultures. However, this soft-sensor as a black-box model, is inherently limited to predicting key process variables in discrete space and hence cannot accommodate the dynamic evolution of the variables in between measurement instances. The objective of this study was to formulate a KF-based soft-sensor, which combines a mechanistic metabolic model and a multi-wavelength fluorescence-based soft-sensor, in order to: (i) filter noise in the fluorescence data by using the mechanistic model and (ii) produce estimates of culture variables in-between fluorescence data samples that are acquired at discrete time intervals. For this purpose, a simplified structured metabolic-model was initially developed to capture the dynamics of the extracellular metabolite concentrations. The structure of this model was identified based on a preliminary metabolic flux model to establish the most important reactions relating the main nutrients to by-products. Since the previously proposed fluorescence based soft-sensor was not amenable for the formulation of the KF-based soft-sensor, a new model was developed for the current study. For the data-driven soft-sensors, partial least square regression (PLSR) was used to draw an indirect inference about the actual state of the CHO cell cultivation process from the 2D-fluorescence spectra and consequently generate an empirical predictive-model for *at situ* measurement of viable and dead cells, recombinant protein, glucose, and ammonia

concentrations. Then, the fluorescence based sensor was integrated together with the mechanistic model using an EKF with adaptation of the model parameters.

4.2 Materials and Methods

4.2.1 Cell Culture

A CHO cell line provided by MabNet (National Science and Engineering Research Council Mab Network) was utilized for expressing EG2-hFc Mab. To obtain the seed culture, cells were sub-cultured every 2 to 3 days, with a seeding density of 0.25 million cells/ml, and then grown in a serum and protein-free defined media (Biogro) supplemented with 0.9% HT (Invitrogen: 11067-030). To increase the richness of the data for model calibration purposes, batch cultures were performed at different initial glucose and glutamine concentrations (Table 4-1) in 500 ml polycarbonate shake-flasks (250 ml of working volume at 37 $^\circ$C with 5% CO_2 and 120 rpm agitation rate) with an initial cell density of 0.2 million cells/ml. Two different media formulations to be referred as Media I and II were used as shown in Table 4-1 for the corresponding cultures. Samples were withdrawn for analytical measurements and fluorescence spectra acquisition.

Table 4-1. Experiment Design (Actual concentrations might slightly deviate).

Culture Notations	Media Type	Initial Concentration	
		Glucose (mM)	Glutamine (mM)
A	I	25	0
B	I	25	4
C	I	25	4
D	I	25	8
E	I	45	0
F	I	45	4
G	I	45	8
H	II	25	4
I	II	35	4
J	II	45	8

4.2.2 Analytical Measurements

Cells were counted under suspension using a hemocytometer and trypan-blue exclusion *at situ*. An Enzyme-Linked Immunosorbent Assay (ELISA) method, developed by MabNet (NSERC Monoclonal Antibody Network), was implemented to determine recombinant protein concentration. Briefly, 96-well plates were coated with an anti-human IgG1 (Fc specific) produced in goat (Sigma Aldrich, ON) followed by the addition of samples and standard. Samples and standard were detected using IgG (Fc specific)-peroxidase antibody produced in goat (Sigma Aldrich, ON) and subsequently the addition of peroxidase's substrate (TMB) resulted in a colored product. Two M sulfuric acid was added to stop the reaction and the absorbance read using a Biotek Synergy 4 multi-plate reader at 450 nm. Glucose concentrations were measured using a Megazyme assay kit (Cedarlane, ON) protocol and the absorbance measured at 510 nm using a Biotek Synergy 4 multi-plate reader. Ammonia concentrations were determined using an ammonia ion-selective electrode—connected to a VWR pH/ISE meter (model 710A).

4.2.3 Fluorescence Signal Acquisition

2D-fluorescence spectra for samples were obtained *at situ* by employing a Cary Eclipse spectrofluorometer (Palo Alto, CA) equipped with a Peltier multi-cell holder using quartz cuvettes

(Thermo Scientific, ON) over an excitation and emission range from 240 to 500nm at 10 nm increments and 280 to 600 nm with 2 nm increments, respectively. Based on a preliminary investigation (Ohadi *et al.* 2014a), a photomultiplier tube (PMT) voltage of 600 V, slit width (SW) of 5 nm for both excitation and emission, scanning rate of 1200 nm/min and a dilution proportional to a sample to buffer ratio of 1:19 were identified to be optimal for signal acquisition. The measured 2D-fluorescence spectra was organized for analysis in the form of a matrix being referred hereafter to as an excitation emission matrix (EEM). The growth media was a complex mixture of components including amino acids, vitamins, and co-factors, some of which intrinsically fluoresce when being excited by light. Table 4-2 summarizes the approximate excitation/emission range of known fluorophores in the 2D-spectra.

Table 4-2. Summary of approximate excitation/emission range of various components of media in the 2D-spectra.

Fluorophore	Excitation (nm)	Emission (nm)
Tyrosine	260~280	300~310
Tryptophan	280~300	350~360
Pyridoxine	320~330	390~410
NAD(P)H	340~360	440~460
FAD, Riboflavin	460~470	500~520

4.3 Model Development

4.3.1 Data-driven Predictive Model

Due to the high dimensionality of the 2D-fluorescence spectra, required in order to encompass regions pertinent to informative fluorophore peaks, compared to the limited number of samples available, high collinearity among the spectral variables and fluorescence-associated noise, multivariate statistical methods are required to formulate a reliable empirical model. For this purpose the partial least square regression (PLSR) method (Geladi and Kowalski 1986; Wold *et al.* 1984) with SIMPLS algorithm (Dejong 1993) was exploited to develop predictive empirical-models for *at-line* monitoring of viable and dead cells, glucose, ammonia and recombinant protein (Mab) concentrations. A comprehensive review of PLSR was conducted by Wold *et al.* (1984). For formulation of the PLSR model, the samples EEMs were individually re-arranged into a row vector and appended one above the other to form the input data matrix (X). The fluorescence spectra of the

phosphate buffer saline (PBS) was acquired parallel to that of the samples and subtracted from them so as to filter background noise. For simplicity, data for viable and dead cells, glucose, ammonia and Mab concentrations were combined into a single matrix to be referred hereafter as the output data set (*Y*). Both the input and output data sets were auto-scaled (mean centered and divided by variance) column-wise to avoid dominance by columns involving variables of relatively large magnitudes. In order to obtain an optimal number of latent variables (LVs) and avert model over-fitting, the leave one batch out cross validation (LOBO-CV) approach was applied based on the minimization of the root mean square error of cross validation (RMSECV). This cross validation procedure is expected to mitigate the problem of biased of out of sample error prediction and the sensitivity of the data-driven model to noise in the spectra (Hagedorn *et al.* 2003; Ohadi *et al.* 2014a). Chemometric analyses were performed using the PLS-Toolbox 7.0.3 (Eigenvector Research Inc., Manson, WA) running in the MATLAB 8.0.0 platform (MathWorks, Natick, MA).

Although metabolites such as glucose and ammonia are not intrinsically fluorescent, their consumption or production rates are expected to be associated with media-fluorophores through stoichiometric correlations (Ohadi *et al.* 2014a; Skibsted *et al.* 2001). Previous work by the authors (Ohadi *et al.* 2014a) have confirmed that the evolution of the fluorescence-spectra over time can be utilized to track normalized—with respect to initial concentration—glucose and ammonia trajectories. For glucose the normalization was performed by subtracting the current concentration from the initial culture concentration and dividing by the initial concentration (Ohadi *et al.* 2014a). On the other hand, since batch cultures were initiated with a negligible amount of ammonia, this metabolite was not normalized with respect to an initial value. The normalizations were executed prior to auto-scaling. Replicated spectra were included in the model regression step to improve the accuracy of the predictions. To ensure richness of data for calibration and to produce a robust model, the soft-sensor was calibrated using cultures initiated with various glucose and glutamine concentrations specified in Table 4-1. For a rigorous evaluation of prediction quality for interpolation and extrapolation, data from cultures A, C, G, and I (Table 4-1) were excluded from the calibration set and used to estimate the prediction precision.

4.3.2 Dynamic Metabolic Model

A structured model based on the metabolic network was used to describe the evolution of extracellular metabolites over time. However, the network of possible reactions for mammalian cells is very large and modeling the entire cell metabolism network is challenging, requiring a large

amount of data for parameter estimation and most probably resulting in an inaccurate model due to noise and scarcity of data. Hence, the network was trimmed based on metabolic flux analysis (MFA) to embrace merely metabolites and fluxes that mostly explain the distribution of carbon and nitrogen in the process (Naderi *et al.* 2011; Nolan and Lee 2011). The MFA analysis is based on the pseudo-steady states mass balances, i.e. the internal metabolite accumulation is assumed to be negligible during the culture process (Nolan and Lee 2011). The metabolic network consist of pathways that are associated with the consumption of carbon including glycolysis, the TCA cycle, energy-related amino acid metabolism, and production of biomass and MAb (Naderi *et al.* 2011). A systematic approach was implemented to eliminate insignificant fluxes and reduce the network (Naderi *et al.* 2011). By assuming that there is no accumulation of intermediate compounds it is possible to eliminate intermediate reactions thus generating a set of macro-reactions that directly relate the main nutrients to main by-products (Table 4-3). It is noteworthy that a single metabolic network was used throughout the whole course of the culture encompassing both the exponential and post-exponential phases of growth. Based on the identified macro-reactions, a set of dynamic mass balances was formulated for each one of the metabolites involved in these reactions as reactants or products. Monod-kinetic models were assumed for the majority of the reactions involved in the dynamic balances, unless stated otherwise.

Table 4-3. Set of macroscopic-reactions relating extracellular metabolite concentrations.

Reaction No.	Macro-reaction
R1	$Glc \rightarrow 2\ Lac$
R2	$Glc \rightarrow 6\ CO_2$
R3	$Asn \rightarrow Asp + NH_3$
R4	$Ala \rightarrow NH_3 + 3\ CO_2$
R5	$Gln + 0.5\ Glc \rightarrow Asp + Ala + CO_2$
R6	$Asp \rightarrow NH_3 + 4\ CO_2$
R7	$Glc + 2\ NH_3 \rightarrow 2\ CO_2 + Gln$
R8	$Gln + Asp \rightarrow Asn + Ala + 2\ CO_2$

Viable cell temporal trajectory is a crucial factor profoundly affecting the rate of metabolite consumption and production. Based on the observation in a previous study by the authors (Ohadi *et al.* 2014a) that some of the viable cells are not proliferating during different culture stages, the viable cell population was divided into growing and non-growing fractions. Then, it was hypothesized that the viable cell population is dependent on glucose availability as shown in Equation 4-1. It has been

substantiated from the experimental data and corroborated by the literature (Ahn and Antoniewicz 2011; Jang and Barford 2000) that CHO cell growth and metabolism are generally characterized by high uptake rates of glucose and glutamine, growth is suppressed by lactate and ammonia while ammonia accumulation and glucose starvation stimulate cell death rate (Zustiak *et al.* 2008). It was also assumed that dead cells lyse along the culture time (Jang and Barford 2000). Following these observations, growth and death rates were mathematically described as (Equation 4-2 and Equation 4-3).

$$\frac{dfgr}{dt} = -K_{11}\frac{fgr}{1+[Glc]/K_{12}}$$

Equation 4-1

$$\frac{dX_v}{dt} = \mu_{\max}.fgr.X_v\left(\frac{[Glc]}{K_{21}+[Glc]}\times\frac{[Gln]}{K_{22}+[Gln]}\times\frac{1}{1+[Amm]/K_{23}}\times\frac{1}{1+[Lac]/K_{24}}\right)$$

Equation 4-2

$$-k_d.(1-fgr)X_v\left(\frac{1}{1+\left([Amm]/K_{25}\right)^n}+\frac{K_{26}}{[Glc]}\right)$$

$$\frac{dX_d}{dt} = k_d.(1-fgr).X_v\left(\frac{1}{1+([Amm]/K_{25})^n}+\frac{K_{26}}{[Glc]}\right)-k_{lys}X_d$$

Equation 4-3

The lactate production rate is assumed to be proportional to the rate of glucose consumption. Following the MFA analysis ammonia is assumed to be partially produced from asparagine, aspartate, and alanine as well as glutamine consumption. Based on the lumped macro-reactions obtained from the preliminary metabolic flux analysis, the dynamic balances for the main extracellular metabolites are given by equations (4-4, 5, 6, 7, 8, 9, and 10).

$$\frac{d[Glc]}{dt} = -X_v\left(\frac{K_{31}[Glc][Gln]}{(K_{32}+[Glc])(K_{33}+[Gln])}+\frac{K_{34}[Glc]}{(K_{35}+[Glc])}\right)-K_{36}X_v$$

Equation 4-4

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

Equation 4-5

$$\frac{d[Lac]}{dt} = -K_{51}X_v\frac{d[Glc]}{dt}$$

<div align="right">**Equation 4-6**</div>

$$\frac{d[Asn]}{dt} = -K_{61}X_v\frac{[Asn]}{K_{62}+[Asn]}$$

<div align="right">**Equation 4-7**</div>

$$\frac{d[Asp]}{dt} = X_v\left(\frac{K_{61}[Asn]}{K_{62}+[Asn]} - \frac{K_{63}[Asp]}{K_{64}+[Asp]} + \frac{K_{65}[Glc][Gln]}{(K_{66}+[Glc])(K_{67}+[Gln])}\right)$$

<div align="right">**Equation 4-8**</div>

$$\frac{d[Ala]}{dt} = X_v\left(\frac{K_{65}[Glc][Gln]}{(K_{66}+[Glc])(K_{67}+[Gln])} - \frac{K_{68}[Ala]}{K_{69}+[Ala]}\right)$$

<div align="right">**Equation 4-9**</div>

$$\frac{d[Amm]}{dt} = -K_{71}\frac{d[Gln]}{dt} + K_{72}X_v\left(\frac{K_{61}[Asn]}{K_{62}+[Asn]} + \frac{K_{63}[Asp]}{K_{64}+[Asp]} + \frac{K_{68}[Ala]}{K_{69}+[Ala]}\right)$$

<div align="right">**Equation 4-10**</div>

Based on the experimental results, it was found that cultures with lower glutamine concentration consumed glucose at a slower rate. Consequently, viability was sustained for a longer duration resulting in higher recombinant protein productivity. Also, cultures initiated with higher amount of glucose accumulated more Mab in comparison to their counterparts at equal initial glutamine concentrations. Accordingly, the dynamic balance for the recombinant protein as the main product was proposed as Equation 4-11.

$$\frac{dMab}{dt} = X_v(K_{81} + K_{82}[Glc])$$

<div align="right">**Equation 4-11**</div>

4.3.2.1 Parameter estimation

The parameters of the model were determined by minimizing the sum of squared error (SSE) between model predictions and experimental results (analytically measured as explained in *4.2.2 Analytical Measurements* section) of culture B, with recursive cross validation with data from culture D (Table 4-1), conducted by a constrained optimization algorithm (*fmincon*) in the MATLAB environment. To avoid convergence to a local minimum a global search method (*GlobalSearch* function in MATLAB) was implemented to systematically generate different initial guesses for the *fmincon* algorithm. The

GlobalSearch function is also capable of performing parallel computation to reduce the overall run time of the optimization search. Table 4-4 summarizes the estimated-parameters.

Table 4-4. Parameter values of Equations 4-1 to 11.

Parameter	Value	Parameter	Value
K_{11}	0.25	K_{41}	4.56
K_{12}	9.99	K_{42}	10.18
μ_{max}	1.38	K_{43}	4.96
k_d	1.34	K_{51}	0.53
k_{lys}	0.2	K_{61}	1.91
K_{21}	1	K_{62}	27.03
K_{22}	0.037	K_{63}	3.03
K_{23}	6.24	K_{64}	7.9
K_{24}	110	K_{65}	0.39
K_{25}	2.37	K_{66}	1
n	2.85	K_{67}	0.031
K_{26}	0.006	K_{68}	5.5
K_{31}	20.9	K_{69}	34.75
K_{32}	46	K_{71}	0.45
K_{33}	0.044	K_{72}	0.25
K_{34}	0.92	K_{81}	1.1
K_{35}	47.51	K_{82}	0.062
K_{36}	0.16		

4.3.3 Kalman Filter Model Development

The conventional Kalman observer (Kalman 1960) recursively updates the states' estimates when using noisy measurements and it gradually minimizes the estimation-error covariance on the assumptions of a linear process and/or measurement models and Gaussian distribution for all error terms. Instead, the EKF is a variant of the KF observer, suitable for nonlinear systems, that is based on the piecewise linearization of the process around the time trajectories of the variables. In the context of this work, the process dynamics is formulated in a continuous time domain while in contrast, measurements are obtained in a discrete fashion with low sampling frequencies. To tackle this, a hybrid (continuous-discrete) EKF formulation is utilized.

Following the large number of assumptions and simplifications that were done to formulate the dynamic model, kinetic parameters and the model structure of the metabolic model have a significant level of uncertainty that necessitates parameter adaptation so as to suppress state-estimates bias

(Dochain 2003; Kozub and Macgregor 1992). In this study non-stationary metabolic-model parameters ($\mu_{max}, k_d, and\ K_{31}$) were allowed to change with time by considering them as additional states. These specific parameters were selected for adaptation based on the sensitivity of the results to these parameters. Accordingly, the vector of states to be used by the EKF is composed of the states of the dynamic model (observed and non-observed) and the 3 adapting parameters as:

$$\boldsymbol{\xi} = [\,X_v, X_d, Glc, Amm, Mab, fgr, Gln,\ Lac, Asn, Asp, Ala, , \mu_{max}, k_d, K_{31}]^T.$$

Consequently, the process and measurement models can be represented as Equation4-12 and Equation4-13, where the term $\varphi(\boldsymbol{\xi}(t), t)$ is zero for the adapting parameter, i.e. for the last three elements of the vector of states.

$$\frac{d\hat{\boldsymbol{\xi}}}{dt} = \varphi(\boldsymbol{\xi}(t), t) + \boldsymbol{\omega}(t) \qquad\qquad \textbf{Equation4-12}$$

$$\mathbf{Z}_k = \mathbf{H}\boldsymbol{\xi}_k + \mathbf{v} \qquad\qquad \textbf{Equation4-13}$$

The terms $\boldsymbol{\omega}$ and $\mathbf{v}$ are independent random Gaussian process noise and measurement noise with zero mean covariance matrices $\mathbf{Q}$ and $\mathbf{R}$, respectively. Using the linear terms of a Taylor expansion the covariance of the states can be formulated in the continuous time domain by Equation4-14, where $\mathbf{J}$ is the Jacobian matrix.

$$\frac{d\hat{\mathbf{P}}}{dt} = \mathbf{J}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{J}(t)^T + \mathbf{Q}(t) \qquad\qquad \textbf{Equation4-14}$$

$$\mathbf{J} = \left.\frac{\partial\varphi(\boldsymbol{\xi}(t), t)}{\partial\xi_i}\right|_{\xi=\hat{\xi}(t)} \qquad\qquad \textbf{Equation4-15}$$

The *a priori* estimate of states ($\boldsymbol{\xi}_{k|k-1}$) and covariance of states ($\mathbf{P}_{k|k-1}$) at time k, when a new measurement is acquired, are estimated by solving Equation4-12 and Equation4-14 from t_{k-1} to t_k. The posterior estimates are attained by exploiting the measurements obtained at t_k and the Kalman gain $\mathbf{K}$ in Equation4-16. The expressions for the time-varying Kalman gain ($\mathbf{K}$) as well as the recursive equations for updating states ($\boldsymbol{\xi}_{k|k}$) and state covariance matrix ($\mathbf{P}_{k|k}$) are as follows:

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T (\mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T + \mathbf{R})^{-1} \qquad\qquad \textbf{Equation4-16}$$

$$\hat{\boldsymbol{\xi}}_{k|k} = \hat{\boldsymbol{\xi}}_{k|k-1} + \mathbf{K}_k (\mathbf{Z}_k - \mathbf{H}\hat{\boldsymbol{\xi}}_{k|k-1}) \qquad\qquad \textbf{Equation4-17}$$

61

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)\mathbf{P}_{k|k-1}$$ **Equation4-18**

It is noteworthy that when using the multi-wavelength fluorescence data as the only available measurements, the inference of observed states are indirect; thus, $\mathbf{Z}$ is assumed to be equal to a function of the entire fluorescence spectra using the regression model explained in the *4.3.1 Data-driven Predictive Model* section. Accordingly, $H = \begin{bmatrix} \mathbf{I}_{5\times5} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ and $\mathbf{v}$ is the model noise associated to the PLSR regression model.

4.3.3.1 Estimation of covariances to be used in the EKF

As a result of the complexity of the mammalian cell cultivation process and the uncertainty associated with model structure and parameters, the mechanistic model is susceptible to noise (Wang *et al.* 2010). The data-driven soft-sensor also suffers from an inaccuracy being exacerbated while predicting values far from the calibration conditions (Shioya *et al.* 1999). Proper evaluation of the process model noise, $\boldsymbol{\omega}$ in Equation4-12, as well as the fluorescence-based soft-sensor noise, $\mathbf{v}$ in Equation4-13, are required to guarantee the performance of the EKF and the precision of the posterior estimates. However, it is a challenging task to properly estimate the out-of-sample error, i.e. the error of the prediction set, for both fluorescence and mechanistic models due to the limited amount of data. Therefore, simplification assumptions were done to estimate these errors and their corresponding covariance. The process and measurement model-noise were assumed constant and uncorrelated, i.e. the noise-covariance matrices are constant and diagonal, with the diagonal elements being the noise variances.

Variance of residuals of the data-driven soft-sensor ($\mathbf{R}$) for the calibration set was assumed to be equal to the *in sample* measurement noise. The diagonal elements of the matrix ($\mathbf{R}$) for the prediction set was set equal to the variance of the residuals of the LOBO-CV procedure that was used as an approximation of the *out of sample* error. To obtain the covariance of process noise ($\mathbf{Q}$), curves were fitted to data that were collected by analytical methods other than fluorescence as explained in section 4.2.2 for the cultures used for calibration of the fluorescence based soft-sensor (Table 4-1). The deviations of the slopes of these fitted curves from the derivatives of the dynamic metabolic model gradients were utilized to construct the process-noise covariance matrix ($\mathbf{Q}$) for metabolic states. This calculation was performed for each of the cultures. The process noise variances for the three non-stationary model parameters were optimized to minimize the SSE of the posterior estimates.

The initial covariance of states (**P**) is assumed diagonal with zero variances for all metabolic states since actual data were used for the initial point. For the non-stationary parameters, the corresponding elements of **P** were acquired through minimization of the SSE of the posterior estimates.

4.4 Results and Discussions

PLSR was exploited to draw indirect inference between the multi-wavelength fluorescence map and key system-states that included viable and dead cells, recombinant protein, glucose, and ammonia concentrations. For simplicity a single model was calibrated for all these key attributes.

Based on the RMSECV, 6 LVs were selected capable of capturing 96.35% of variation in X matrix and explaining 87.51% of variation in Y. The loadings were visually analyzed to interpret the contribution of different regions of the fluorescence spectra on the LVs. The loading of the first LV, that captures 69.35% of the input variability, was found to be dominated by the region of spectra attributed to amino acids (tryptophan and tyrosine) and, with less pronounced magnitude, pyridoxine (Figure 4-1a) that corroborates previous findings (Marose *et al.* 1998). While these same regions also contributed to the second LV that explains 12.1% of input variance, this second LV is significantly impacted by the scattering region (Figure C- 2-Appendix C). To better understand the correlation of the spectra with the viable cell density, Figure 4-1b provides the contour plot of the regression coefficient corresponding to the viable cell concentration as a function of the emission and excitation wavelengths. From this figure, a strong inverse correlation can be observed with the region associated with tryptophan and NAD(P)H, which coincides with previous observations (Marose *et al.* 1998). Although not shown for brevity, other correlations between the regression coefficients of the other inferred variables with the intensities of the fluorescence spectra were also identified. For example, the Rayleigh scattering dominated the regression coefficient of dead cell concentration, vitamins contributed to the regression coefficient of glucose, amino acids and co-factors contributed to the Mab's regression coefficient, and NAD(P)H region played a critical role in the ammonia's regression coefficient.

63
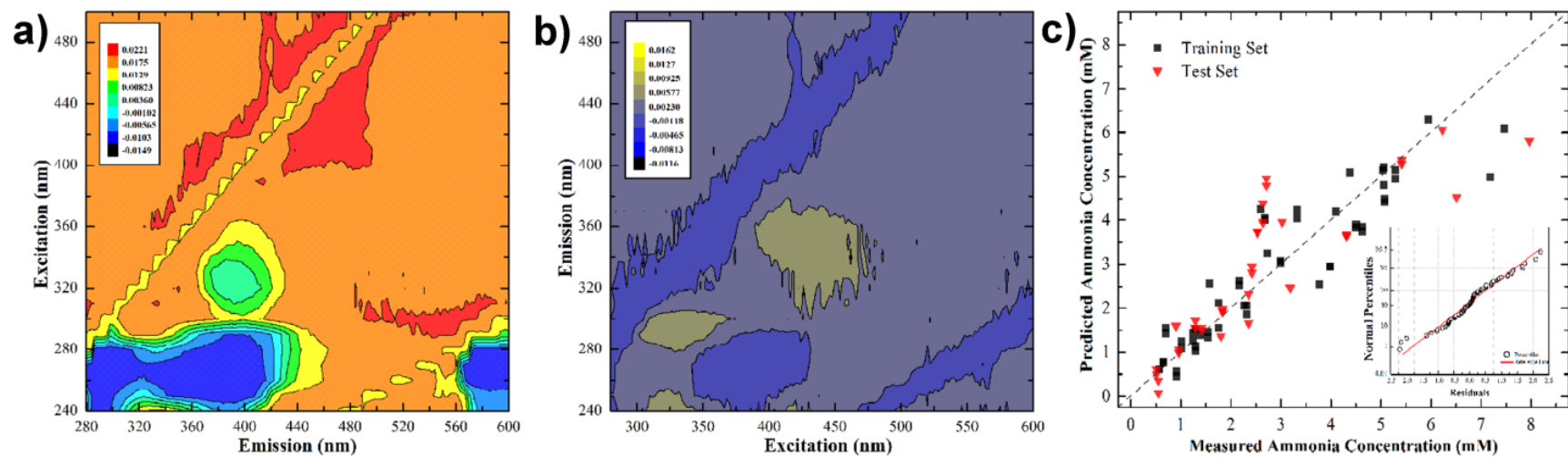
Figure 4-1. Results of fluorescence based predictive model: a) Contour plot of the first loading; b) Contour plot of regression coefficient for viable cell concentration; c) Model prediction vs. measured values for training and test set and normal percentiles of the residuals (inset) for viable cell concentration.

Table 4-5 exhibits the PLSR model results and goodness of fit for the key process variables. Figure 4-1c presents the data-driven soft-sensor prediction vs. measured (as explained in *4.2.2 Analytical Measurements* section) values for viable cell concentration for the calibration and test set with normal percentiles of residuals. From the figure it can be seen that the points fall into a region close to 45° line and residuals are normally distributed. A similar conclusion was drawn for other key process variables of interest (for brevity graphs are not provided). The PLSR model predicts the calibration and test data sets fairly well with the residuals being approximately normally distributed. However, it is clear that the PLSR model can only infer the value of the variables at the instances of sampling but it cannot predict the behavior in between these instances.

Table 4-5. Data-driven soft-sensor results and goodness of fit for key process variables based on LOBO-CV strategy and prediction precision of model using cultures A, C, G, and I.

Target Variable	Number of LVs	Calibration R^2_{Cal}	Cross-Validation R^2_{CV}	Testing/Prediction R^2_P
Viable Cells	6	84.01%	64.8%	79.31%
Dead Cells	6	94.23%	90.31%	92.49%
Glucose	6	90.2%	60.08%	89.04%
Ammonia	6	83.42%	42.21%	79.62%
Mab (Eg2)	6	85.26%	27.56%	85.9%

The EKF, which is based on the combination of the dynamic mechanistic model with the fluorescence-based soft-sensor, is implemented to improve the prediction precision at the point of the data–driven based measurements, enhance the extrapolation (out of calibration region) capability, and better predict the dynamic profiles between the sampling instances until the next measurement becomes available.

Table 4-6. Comparison of the key process variables SSEs of the Fluorescence-based (Flu) with that of the EKF based soft-sensors for the performed batch cultures.

	Sum of Squared Error (SSE)							
	Viable Cell Concentration		Glucose Concentration		Ammonia Concentration		Mab Concentration	
Culture Notations	Flu	EKF	Flu	EKF	Flu	EKF	Flu	EKF
A	2.07	0.68	116.04	112.97	0.68	0.36	46.94	16.95
B	0.76	0.56	45.56	22.05	2.65	1.42	18.35	18.71
C	3.14	2.20	48.79	9.34	3.25	1.59	59.45	60.31
D	1.93	1.94	71.80	34.36	8.91	7.96	10.31	9.80
E	0.33	0.18	67.91	69.32	0.23	0.34	77.81	58.39
F	2.15	1.75	232.32	277.71	0.57	0.132	8.9	2.54
G	0.90	0.31	237.23	204.86	9.71	6.59	69.73	28.01
H	0.32	0.13	48.82	28.89	5.07	5.41	68.15	66.93
I	3.29	3.12	26.18	30.17	8.66	8.28	27.52	26.56
J	1.56	1.44	167.36	197.06	2.33	2.33	39.38	40.01

After tuning the unknown variances, as explained in the *4.3.3.1 Estimation of covariances to be used in the EKF* section, the EKF was implemented to obtain estimates for the prediction set. The performance of the EKF was investigated by calculating the SSE of each state at sampling points, assuming correct analytical measurements, and comparing them with the SSE of these attributes obtained using the data-driven soft-sensor. Table 4-6 compares the SSE over the entire course of a culture of the data-driven soft-sensor with that of the EKF-based sensor. For the calibration set (cultures B, D, E, F, H, and J in Table 4-6) only a slight improvement is observed, while for the prediction set (A, C, G, and I in Table 4-6) a considerable reduction is observed in the SSE of each variable when the EKF model is implemented. For instance, the average reduction in the SSE for the cultures used for validation is 51% in favor of the EKF-based approach.
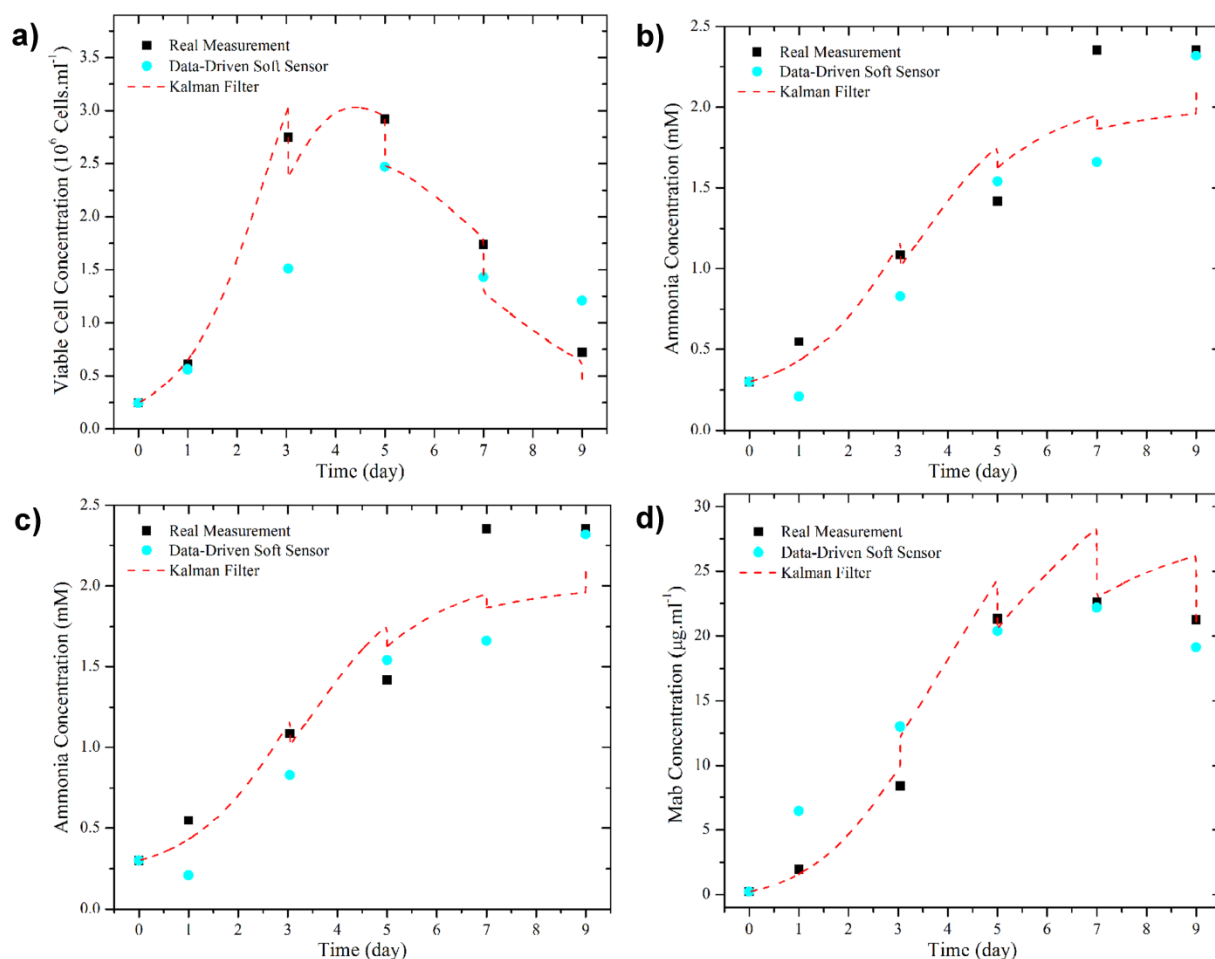
Figure 4-2. Comparison between data-driven soft-sensor, Kalman Filter based soft-sensor, and real measurement results of culture A for: a) Viable cell concentration; b) Glucose concentration; c) Ammonia concentration; d) Recombinant protein concentration.

As described in the materials and methods section, two different growth media formulations were used in the experiments. These two media resulted in very significant differences of growth pattern and evolution of metabolites. It was observed that the EKF provides only a slight improvement in the SSE of the cultures grown in Media II (cultures H, I, and J in Table 4-6) due to the considerable differences between the two media which is not well captured by the mechanistic model. For the dead cell concentration the data-driven soft-sensor was already accurate thus the EKF based sensor did not noticeably improve the dead cell concentration prediction accuracy. Figure 4-2 and Figure 4-3 compare the mechanistic model, the data-driven soft-sensor, the EKF soft-sensor, and analytical measurement (as explained in *4.2.2 Analytical Measurements* section) results for cultures A and G, respectively, for viable cell (a), glucose (b), ammonia (c), and Mab (d) concentrations. As illustrated

in Figure 4-2a and Figure 4-3a for A and G respectively, the EKF generally improved the estimates at the sampling instances. The EKF based soft-sensor was capable of reducing the overall SSE of viable cell concentration, over the time course of the culture, for approximately 67% for culture A and G and for approximately 30% for culture C (Table 4-6). However, the true advantage of using the EKF-based sensor is for predicting the evolution of the variables in between the sampling instances. For the glucose, the temporal evolution is well captured for culture A (Figure 4-2b), although the EKF-based soft-sensor did not improve significantly the precision at measurement points, as compared to the stand-alone fluorescence-based soft-sensor. For culture G initiated with a higher glucose concentration, the superior performance of the EKF soft-sensor is evident at the beginning of the culture where the glucose concentration decreases rapidly (Figure 4-3). The EKF reduced the overall SSE of glucose for roughly 14% for culture G (Table 4-6). For both cultures A and G the EKF soft-sensor accurately approximated the trajectories of ammonia and Mab and considerably improved their estimation at the measurement instances. For culture A, the EKF reduced the overall SSE at measurement instances for about 47% and 64% for ammonia and Mab concentrations, respectively (Table 4-6). Similarly, the implementation of the EKF enhanced the prediction accuracy for ammonia by 32% and Mab by 60% at measurement instances for culture G. The large jumps, e.g. observed in Figure 4-2, between the prior to the posterior estimates of the EKF is due to the accumulation of state covariance ($\mathbf{P}$) along a large time span caused by the relatively low sampling frequency of the fluorescence measurements. It can be concluded (Figure 4-2 & Figure 4-3; Table 4-6), that the EKF soft-sensor can enhance the accuracy of the predictions outside of the calibration range.
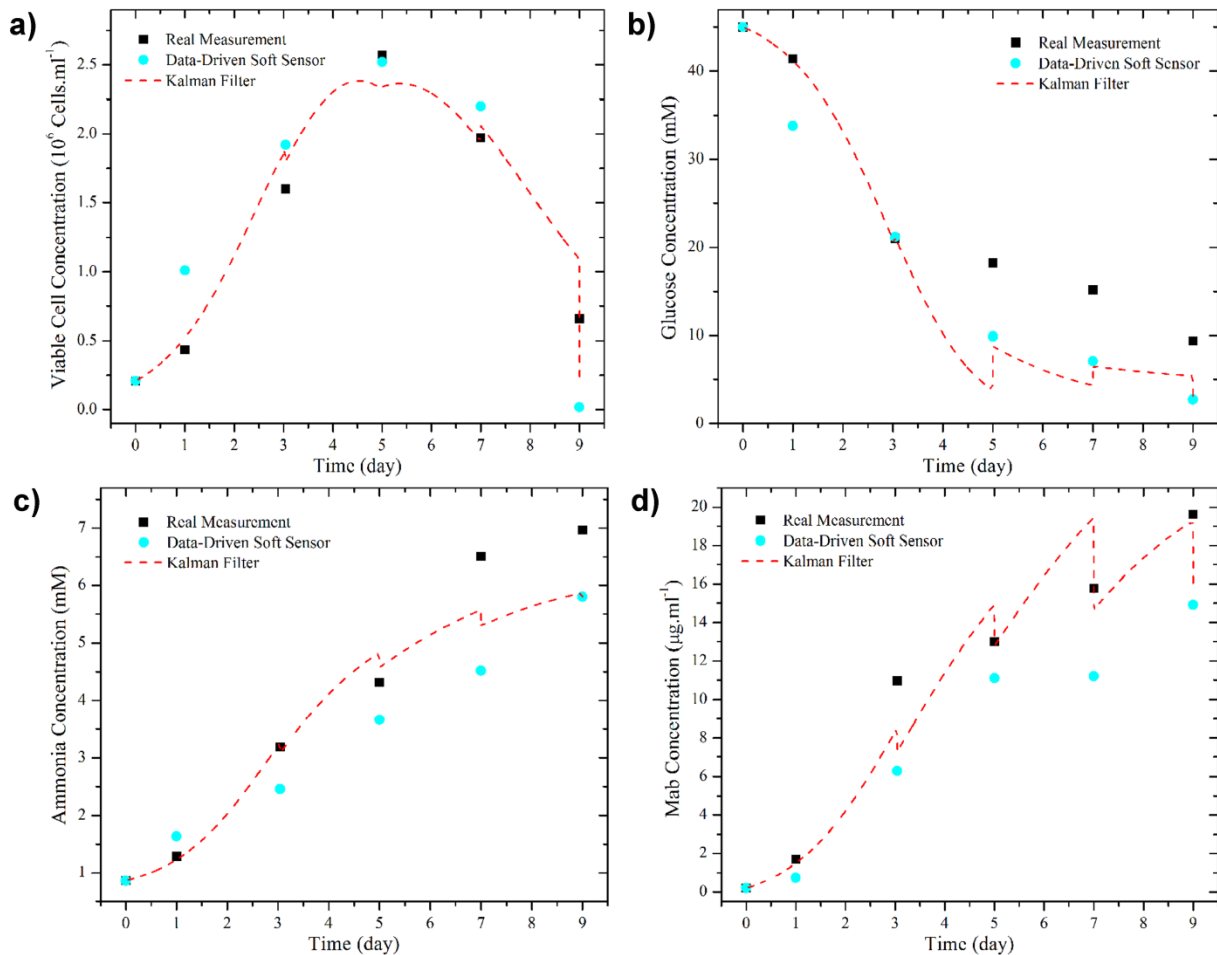
Figure 4-3. Comparison between data-driven soft-sensor, Kalman Filter based soft-sensor, and real measurement results of culture G for: a) Viable cell concentration; b) Glucose concentration; c) Ammonia concentration; d) Recombinant protein concentration.

Low sampling frequency of measurements, missing data, and/or outliers associated with analytical techniques, including fluorescence spectra acquisition, diminish the controllability of the bioprocess (de Assis and Maciel 2000) thus affecting process consistency. In view of this, the key advantage of the EKF-based sensor, apart from enhancing the precision at measurement instances, is its ability to predict between samples, which is particularly important for monitoring and automatic control.

To better elucidate this advantage a simplistic first order hold extrapolation approach was implemented based on the fluorescence based soft-sensor measurements to predict the dynamic behavior between measuring instances and these predictions are then compared to the ones obtained with the EKF based sensor. A first order hold extrapolation is based on a simple linear extrapolation

using a present and past measurement to infer future behavior until the next measurement becomes available. This operation is illustrated in Figure 4-4 for the temporal evolution of the viable cells and the recombinant protein concentrations of cultures A, C, and G. It is evident that the data-driven soft-sensor has an inferior temporal extrapolation capability and cannot be relied on for automated control, while in contrast the EKF-based soft-sensor better tracks the actual dynamics of the system. A similar conclusion can be drawn for the temporal evolutions of the glucose and ammonia concentrations (results not shown for brevity). For quantitative evaluation of the improvement achieved by the EKF-based soft-sensor compared to the first order hold extrapolation approach, the integrals of absolute error (IAE) between a two-point linear extrapolation and real dynamics of the culture (obtained by fitting a curve to process data measured by conventional analytical techniques) was compared to the IAE between EKF-based soft-sensor and the actual dynamics of the culture (Figure 4-4). For the viable cell concentration of culture A (Figure 4-4a) the total sum of the IAE between the predicted behavior by the first order hold and real dynamics of the culture is 5.76, while in contrast, 1.94 was obtained between EKF-based soft-sensor and real dynamics of the culture. For the recombinant protein concentration of culture A (Figure 4-4d), the IAE between the predictions with the first order hold and the real dynamic was 40.7 whereas for the EKF approach was 17.03. Similarly for culture G, the EKF based soft-sensor, compared to the first order hold approximation, reduces the IAE for approximately 61% for both viable cells (from 4.72 to 1.83) and Mab (from 30.56 to 11.98) concentrations.
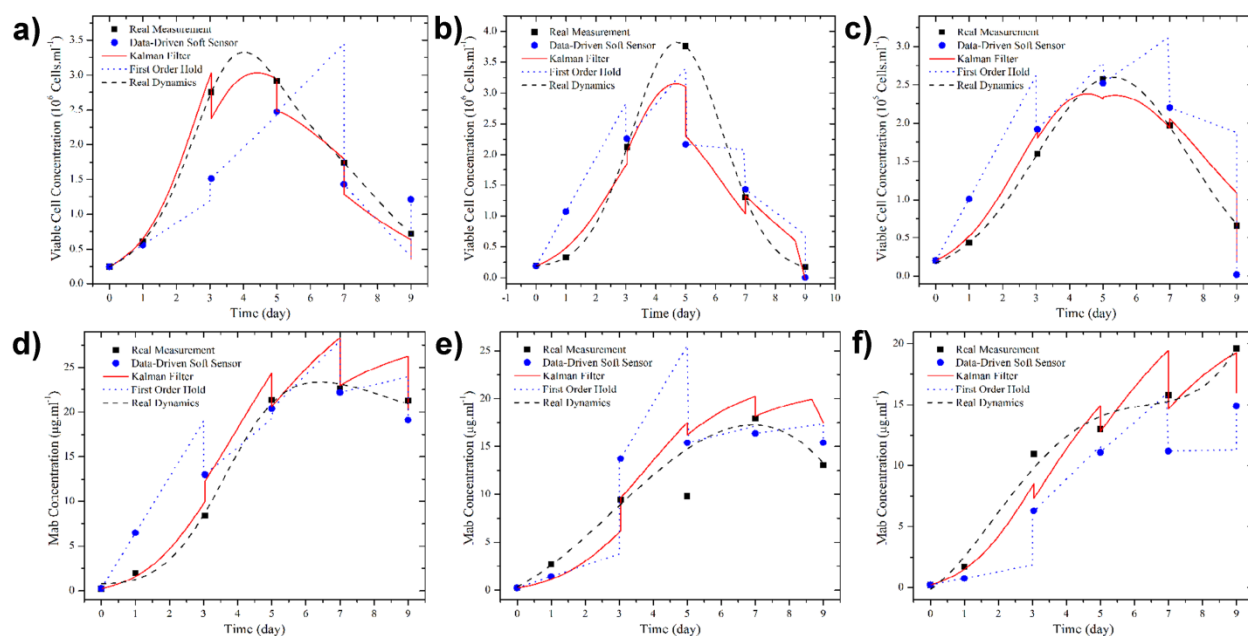
Figure 4-4. Comparison between (temporal) extrapolation performance of Kalman Filter based soft-sensor, first order hold and real dynamics for viable cell concentration: a) Culture A; b) Culture C; c) Culture G, and recombinant protein concentration for: d) Culture A; e) Culture C; f) Culture G.

4.5 Conclusion

An extended Kalman filter-based soft-sensor that combines fluorescent measurements and a dynamic mechanistic model is presented. Components of the fluorescence spectra were related to key process variables using a multivariate statistical model. The output from this regression model was combined together with the mechanistic model using an EKF observer framework. It was shown that the dynamic model can generally improve the accuracy of the fluorescence-based predictions at the instances of sampling. Beyond these improvements, the true advantage of the EKF model is its ability to generate accurate predictions of the temporal evolution of the culture variables in between sampling instances. Thus, the proposed approach is valuable for process monitoring and implementation of automated strategies of the bioprocess.

4.6 Acknowledgement

4.7 Nomenclature

μ_{max}	Specific growth rate
$\boldsymbol{\nu}$	Error matrix of measurement-model noise
$\boldsymbol{\xi}$	Vector of states
$\hat{\boldsymbol{\xi}}$	Vector of estimated states
$\boldsymbol{\omega}$	Error matrix of process-model noise
fgr	Fraction of growing cells
k_d	Death rate constant
k_{lys}	Constant of cell lysis
T	Time
Ala	Alanine
Amm	Ammonia
Asn	Asparagine
Asp	Aspartate
Glc	Glucose
Gln	Glutamine
H	Matrix relating states to the measured variables
J	Jacobian matrix of the function φ
$K_{i,j}$	Reaction constants
$\mathbf{K}_k$	Matrix of Kalman gain at time step k
Lac	Lactate
Mab	Monoclonal antibody
P	Covariance matrix of states
$\hat{\mathbf{P}}$	Estimate of covariance matrix of states
Q	Error covariance matrix of process-model noise
R	Error covariance matrix of measurement-model noise
X_d	Dead cell
X_v	Viable cell

Z	Vector of measured variables
Subscript k	Time step k
Superscript T	Transpose of matrix
Superscript -1	Inverse of matrix

Chapter 5

Novel Dynamic Model to Predict the Glycosylation Pattern of Monoclonal Antibodies from Extracellular Cell Culture Conditions[*]

Glycosylation is a critical protein post-translational modification with a profound impact on the therapeutic properties of Mab and research indicates that it depends on extracellular culture conditions. A novel dynamic model was developed to relate extracellular metabolites' concentrations to a cumulative glycoprofile. The model has three components: dynamic evolution of extracellular metabolites, production of nucleotide sugars in the cytosol, and glycosylation inside the Golgi apparatus. Following comparisons with experimental data obtained from batch CHO cell cultures, the model was found capable of predicting the glycoform profile of Mab temporally, as well as the extent of galactosylation given in the form of galactosylation index. The model has the potential for use in controlling the glycoform profile by manipulating culture conditions.

5.1 Introduction

Monoclonal antibodies (Mabs) comprise the dominant products in the fastest growing segment of the biopharmaceutical market. Chinese hamster ovary (CHO) cells are widely used as hosts for Mab production. Several studies have reported the profound impact of glycosylation, as post-translational modification, on Mabs' therapeutic properties. The degree and extent of glycosylation can be quantified in terms of the type and number of nucleotide sugars attached to the Mab (Durocher and Butler 2009).

N-linked glycosylation, the most common form of glycosylation in mammalian cells, is initiated in the endoplasmic reticulum (ER) by a covalent attachment of a sugar to the polypeptide chain that is followed by proper folding of the protein and removal of three glucose molecules and at least one mannose. The resulting glycoprotein is then transferred into the Golgi apparatus where the final glycosylation process takes place (Del Val *et al.* 2011; Hossler *et al.* 2007; Stanley 2011).

It has been reported (Del Val *et al.* 2011; Hossler *et al.* 2009) and experimentally observed by the authors that culture conditions can be manipulated to obtain a specific oligosaccharide (OS) structure attached to the Mab. Accordingly, a mathematical model that could describe the effect of culture conditions on glycosylation may be instrumental for manufacturing a Mab with a desired glycoprofile. Such a model should be able to relate the extracellular culture conditions to intracellular glycosylation mechanisms at the cell level. To the knowledge of the authors, no such model has yet been developed.

Quality by Design (QbD) is a recent FDA initiative that promotes a systematic approach to drug development, emphasizing that product quality considerations should be introduced at the design stage of the manufacturing process. Predictive mathematical models, such as the one proposed in this study, have been identified as key for designing novel manufacturing operations within the QbD framework (Hossler *et al.* 2009).

In the current work a specifically engineered CHO cell line, provided by MabNet (National Science and Engineering Research Council of Canada Mab Network), has been used to produce glycosylated Mab. A comprehensive model is developed for this cell line that is composed of three main parts as shown in Figure 5-1. The first part consisting of a metabolic flux analysis (MFA) model, developed to describe changes in extracellular metabolites concentrations. The second part bridges the gap between metabolite uptake and glycosylation in the Golgi apparatus. This part, shown by the middle block in Figure 5-1, involves modelling the synthesis of the essential nucleotide sugars from glucose and glutamine in the cytosol (or the nucleus for CMP-Neu5Ac). In the third part, to the right of Figure 5-1, the glycosylation inside the Golgi apparatus is mathematically modelled based on the Golgi maturation assumption (Del Val *et al.* 2011; Hossler *et al.* 2007). Since many of the parameters of three parts of the model have not been explicitly documented in the literature or they have not been reported for the specific cell line under consideration, a comprehensive sensitivity analysis with respect to parameters was conducted based on experimental results to identify which parameters are more critical for model calibration. Preliminary model calibration was done with data for extracellular metabolites' concentrations and glycosylation pattern at different days of the cell cultures. It should, be emphasized that while the first block in Figure 5-1 is used to describe average extracellular metabolites' concentration in the culture media, the third block of the model describes glycosylation at the intracellular level. Since samples were collected at different days of culture, the glycoprofiles reported in this work reflect a cumulative time average of producing cells undergoing glycosylation

under varying culture conditions. Consequently, as further explained in the following sections, for comparing model predictions with measurements, it was necessary to translate the intracellular OS concentration values, calculated by the third component shown in Figure 5-1, to a cumulative average value over culture time based on the instantaneous Mab productivity along the time course of the culture. The subsequent sections describe the experimental techniques followed by theoretical details about the mathematical model and comparisons of model predictions with experimental data.

Nucleotide Sugar
Model

Glycosylation Model

MFA Model
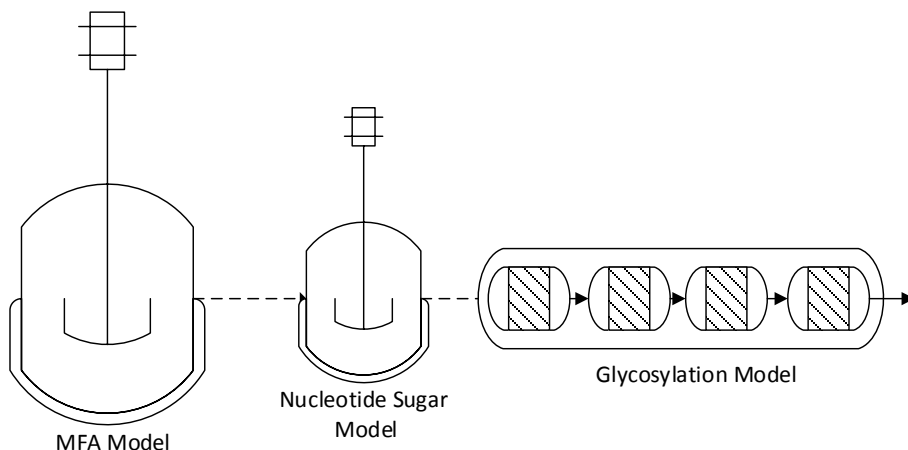
Figure 5-1. Schematic representation of the comprehensive-model: Metabolic Flux Analysis (left), Nucleotide Sugar (centre), Glycosylation based on Golgi Maturation (right).

5.2 Materials and Methods

5.2.1 Culture condition

The cell line used in the present study is a CHO cell line derived from the parental CHO-DXB11 cell line, specifically engineered and provided by MabNet expressing EG2-hFc, a camelid Mab. The cells were maintained in regular serum-free Biogro medium (Provided by MabNet) supplemented with 0.9 % HT (Invitrogen:11067-030). Seed cultures were produced by sub-culturing the cells every 2 to 3 days with a seeding density of 0.25 million cells/ml. Cultures were grown in 500 ml polycarbonate shake-flasks with 250 ml working volume at initial cell density of 0.2 million cell/ml. Flasks were agitated at 120 rpm, and incubated at 37°C with 5% CO_2. Batch experiments were performed at initial glucose concentration of 25 mM at different levels of initial glutamine concentration (0 and 4 mM). A glucose concentration of 25 mM was found to be optimal for growth. Samples were taken on daily basis for *at situ* analysis of viable and dead cell concentration using trypan-blue exclusion method. In

addition, samples were collected, centrifuged at 300×g for 10 minutes, and their supernatant were stored at -20˚C to perform off-line measurements.

5.2.2 Analytical methods

An Enzyme-Linked Immunosorbent Assay (ELISA) method, developed by MabNet, was applied to obtain recombinant protein concentration. The Mab concentration was calculated conventionally by comparing the optical density of each sample with the standard curve.

Glucose, glutamine, lactate, and ammonia concentration of culture broth were measured using a multi-parameter bioanalytical system, Bioprofile 400 (Nova Biomedical, Waltham, MA), off-line.

In order to obtain the composition of the glycans/OSs attached to Mab, a hydrophilic interaction chromatography, following by an exoglycosidase enzyme array digestion method was implemented (Brockhausen 2006). This analysis was performed on samples obtained every other day, starting from the third day of the culture.

5.3 Model Development

The development of the dynamic metabolic flux equations commenced with a detailed steady state flux analysis for this CHO line using the procedure described previously (Naderi *et al.* 2011). An important conclusion of the flux balance analysis, not shown here for brevity, was that all non-essential amino acid synthesis could be described in terms of co-metabolism of glucose $[glc]$ and glutamine $[gln]$. Obviously, this co-metabolism also affects biomass (X_v) and ultimately Mab synthesis as well.

The viable biomass population is differentiated as growing (fgr) and a non-growing ($1 - fgr$) fraction. The non-growing fraction includes both resting cells arrested in the G phase as well as apoptotic cells. In our previous study (Naderi *et al.*, 2011) these two sub-populations were modelled separately. However, in the present case no significant differences in either metabolite uptake or Mab productivity were observed to necessitate separation. Consequently the dynamic model for cell growth is as Equation 5-1.

$$\frac{dX_v}{dt} = \mu \, fgr \, X_v - k_d (1 - fgr) X_v \frac{1}{1 + K_{11} e^{-(\frac{[N]}{K_{12}})}}$$

Equation 5-1

Where µ is specific growth rate, k_d is specific death rate and K_{ij}s are model parameters. As in case of our previous models (Naderi *et al.* 2011) the rate of death is enhanced by the ammonia concentration, [N]. The total ammonia concentration, in turn, was calculated from the metabolic flux balance equations (not shown here).

The estimation of the temporal dependence of the growing cell fraction is a key aspect of the model. A number of plausible model structures were considered and the following choice was made by invoking Akaike's information criterion for model selection (Burnham and Anderson 2002) using a comparison of predicted versus simulated total viable biomass concentration, being formulated as Equation 5-2.

$$\frac{dfgr}{dt} = -K_{21}\frac{fgr}{1+[glc][gln]/K_{22}}$$

Equation 5-2

Equation 5-2 simulates the dependence of the observed growth rate on the product of the glutamine and glucose concentration. Also, as observed experimentally, the growth did not stop immediately if either glucose or glutamine were exhausted from the medium.

The dynamic metabolic flux models for glucose and glutamine metabolism are shown in Equation 5-3 and Equation 5-4 respectively.

$$\frac{d[glc]}{dt} = -K_{31}\,fgr\,X_v\frac{[glc][gln]}{K_{32}+[glc][gln]} - K_{33}X_v$$

Equation 5-3

$$\frac{d[gln]}{dt} = -K_{41}\,fgr\,X_v\frac{[glc][gln]}{K_{42}+[glc][gln]} - K_{43}X_v$$

Equation 5-4

As shown above, both glucose and glutamine metabolism consists of growth associated and a non-growth associated terms. Glucose and glutamine appear to be co-metabolized. It is noteworthy that an identical half-saturation constant for the co-metabolism (K_{32} and K_{42}) satisfied both equations (15.77 mM^2).

The selection of a robust model that adequately describes the system requires careful examination of the parameter values. The parameter values (not shown for brevity) and their distributions in this study were generated by the Gibbs sampler version of the Metropolis-Hastings algorithm (Jitjareonchai *et al.* 2006). Markov Chain Monte Carlo (MCMC) methods such as the Metropolis-

Hastings algorithm used here have been proven to be powerful tools, especially when some prior knowledge about the parameters is available from the literature.

The middle block in Figure 5-1 consists of a simplified model to simulate the production of nucleotide sugars from glucose and glutamine in the cytosol (or nucleus for CMP-Neu5Ac). This model is formulated to connect between the extracellular metabolites' concentration and the glycoform profile of produced MAb. Figure 5-2 exhibits the simplified reaction network connecting the extracellular glucose and glutamine concentrations to nucleotide sugars based on the database Kyoto Encyclopaedia of Genes and Genomes (KEGG) where the intermediate reactions are lumped together for simplicity (Kanehisa and Goto 2000; Kanehisa *et al.* 2012). The amount of glucose and glutamine being consumed towards the production of nucleotide sugar is negligible compared to the consumption of these nutrients towards cell growth and maintenance. Thus this consumption (Figure 5-2) is not included in the MFA model. The reaction rates are assumed to be irreversible following a one or two-substrate Michaelis-Menten kinetics, as shown in Equation 5-6. Additionally, a first order dissociation kinetic rate is utilized to depict the dissociation of nucleotide sugars following their incorporation into biomass production. The mass balance for each of the nucleotide sugars is given in general form by Equation 5-5, assuming that the cytosol (or nucleus) acts as an ideal batch reactor. The model uses the extracellular glucose and glutamine concentration and accounts for the mass transfer across the cell membranes through the calibrated parameters.

$$\frac{d[Sug]_i^{Cyt}}{dt} = \sum v_i r_i^{Cyt} - K_{[Sug]}^{disociation}[Sug]_i^{Cyt} \qquad \textbf{Equation 5-5}$$

$$r_i^{Cyt} = \frac{[Sug]_p^{Cyt}[Sug]_q^{Cyt}}{(K_p + [Sug]_p^{Cyt})(K_q + [Sug]_q^{Cyt})} \qquad \textbf{Equation 5-6}$$

Although the glycosylation starts in the ER, in the current study it is assumed that the proteins are properly folded in ER and only contain nine-mannose attached to their constant heavy chain. Thus the glycosylation model, represented by the rightmost block in Figure 5-1, simulates the transfer of nucleotide sugars inside the Golgi only. In the Golgi, a small number of enzymes will catalyse a large number of sequential reactions that shift the high mannose structure of the saccharide exiting the ER towards complex and hybrid conformations; hence, each enzyme can trigger various reactions (Del Val *et al.* 2011; Hossler *et al.* 2007; Stanley 2011). This large network of reactions has been mathematically described in a form of *relationship-matrix* by Hossler, et al, 2007. Previous studies

have reported that the Golgi apparatus, consisting of four different cisternae, can be approximated as if each compartment converts to a successive one through a maturation procedure (cisternal maturation). Consequently, and following (Del Val *et al.* 2011; Hossler *et al.* 2007) in this study the Golgi was represented by four ideal (no radial and axial dispersion) plug flow reactors (PFR) in series, each representing one cisterna, at steady state conditions. Different overall retention time in the Golgi has been assumed in literature. In this work the overall residence time of Golgi (all four compartments) is set to be forty minutes as postulated in Hossler *et al.* 2007.
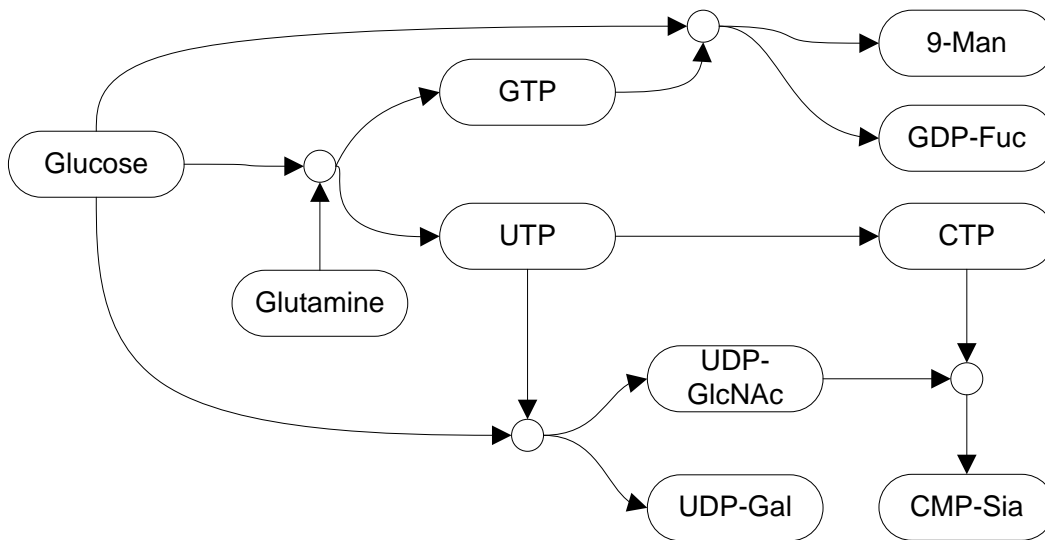
Figure 5-2. Simplified reaction-tree of nucleotide sugars inside the cytosol (or nucleus). Man: mannose, Gal: galactose, GlcNAc: N-acetylglucosamine, Fuc: Fucose, Sia: sialic acid.

The glycosylation reactions occurring in the Golgi are catalyzed by two main groups of enzymes: exoglycosidases and glycosyl-transferases (GTs) (Del Val *et al.* 2011; Hossler *et al.* 2007; Stanley 2011). For simplicity, exoglycosidases were assumed to follow irreversible one-substrate Michaelis-Menten kinetics shown in Equation 5-7 while the GTs follow irreversible two-substrate kinetic, as shown in Equation 5-8 (Hossler *et al.* 2007). It is worth mentioning that GTs were assumed to have only one functional group and each can only transfer an specific nucleotide sugar (Stanley 2011).

The superscript n specifies the compartment/cistern number, k determines the corresponding enzyme and its nucleotide sugar (for GTs). The $\alpha-$s in Equation 5-7 and Equation 5-8 are coefficients that define the spatial distribution of enzymes along the Golgi compartments (Hossler *et al.* 2007; Stanley 2011). Thus, they can determine the direction along the network of reactions.

Kinetic parameters presented in Equation 5-7 and Equation 5-8 strongly depend on the cell line, thus different values are reported in literature. In this work K_k^G, K_k^S, and r_k^{max} values were selected as per the values documented by Hossler, et al, 2007.

$$r_k^n = \frac{\alpha_k^n \, r_k^{\max} [Glyc_i]^n}{K_k^G (1 + \sum_{z=1}^{NC} \frac{[Glyc_z]^n}{K_k^G})}$$

Equation 5-7

$$r_k^n = \frac{\alpha_k^n \, r_k^{\max} [Glyc_i]^n [Sug_k]^n}{K_k^G (1 + \sum_{z=1}^{NC} \frac{[Glyc_z]^n}{K_k^G} + \frac{[Sug_k]^n}{K_k^S} + \sum_{z=1}^{NC} \frac{[Glyc_z]^n [Sug_k]^n}{K_k^G K_k^S})}$$

Equation 5-8

Concentration of 9-Mannose, UDP-Gal, UDP-GlcNAc, GDP-Fuc, and CMP-Sia (Neu5Ac), depend on the extra cellular availability of glucose and glutamine. After being produced in the cytosol (or the nucleus), these sugars are fed into the Golgi apparatus.

The glycosylation model, as presented in previous studies (Del Val *et al.* 2011; Hossler *et al.* 2007), depicts the instantaneous OS pattern of the Golgi in one individual cell. However, the experimental results provide a cumulative average of the glycosilated Mab accumulated in the supernatant over the time of the culture up to the time of the measurement. To account for this effect, Equation 5-9 was utilized to convert the instantaneous glycan composition calculated as per Golgi maturation model to the accumulated one measured experimentally.

$$[Glyc_i]^{Acc} = \frac{\int_0^t \dot{Mab} [Glyc_i]^{Ins} \, dt}{\int_0^t \dot{Mab} \, dt}$$

Equation 5-9

For the purpose of comparisons and for simplicity it was hypothesized that the experimental glycan composition at day three is identical to the instantaneous OS composition at the beginning of the culture. This claim can be supported by the fact that a negligible amount of glycoprotein was measured during the first three days of the culture. For the purpose of model calibration, the α- s were obtained by minimizing the sum of square error (SSE) of OSs using genetic algorithm (GA), while the nucleotide sugars were assumed to be in excess (at the beginning of the culture). The nucleotide-sugar model's parameters were calibrated by minimizing the SSE between comprehensive-model predictions and measured glycoprofiles along the culture time, using data from the batch with no glutamine added to it initially.

81

In order to simplify the quantification of glycosylation extent, the relative abundance area of predominant glycan structures can be lumped together in a form of glycosylation indices: galactosylation index (GI), sialiation index (SI), and fucosylation index (FI). However, the focus of current work is on GI since this cell line is highly galactosylated. The agalctosylated (G_i^0), monogalctosylated (G_i^1), and digalctosylated (G_i^2) OSs were lumped within the GI as shown by Equation 5-10. This grouping is further motivated by the fact that the therapeutic properties of Mabs have been often correlated to these indices (Majid *et al.* 2007).

$$GI = \frac{\sum_i G_i^2 + 0.5 \times \sum_i G_i^1}{\sum_i G_i^2 + \sum_i G_i^1 + \sum_i G_i^0}$$

Equation 5-10

Due to the large number of OSs proposed in Hosslet *et al.* 2007 (341 OSs), the computation is very costly. To reduce computations, the large network of reactions was trimmed based on the experimental data, to encompass only the plausible Mab's OSs for this particular cell line. At current stage the model solves the mass balance equations for 100 OSs and four nucleotide sugars inside the Golg apparatus. This has significantly reduced the run time as compared to the original model. To further decrease the CPU time, conditional statements were avoided. By perusing the matrix calculation and the use of binary matrices, instead of conditional statements, the CPU time was drastically reduced. Additionally, the program was modified to take advantage of parallel computation toolbox in MATLAB that reduced the run time by 1/10. On a hex-core Intel CPU computer equipped with 16 gigabyte of RAM, it takes approximately 30 seconds to run the program for simulating 9 days of culture.

5.4 Results and Discussion

The simulations of the viable cell concentration with 4 and 0 mM glutamine added are shown in Figure 5-3. This particular CHO cell line is relatively fast growing, having a minimum generation time of approximately 16.5 hours. With 4 mM glutamine initially, a maximal viable cell density of over 3.7 million cells/ml is reached after 5 days of incubation. In contrast, the maximum cell density when no glutamine is added to the medium is about one-half of that obtained with the 4 mM.
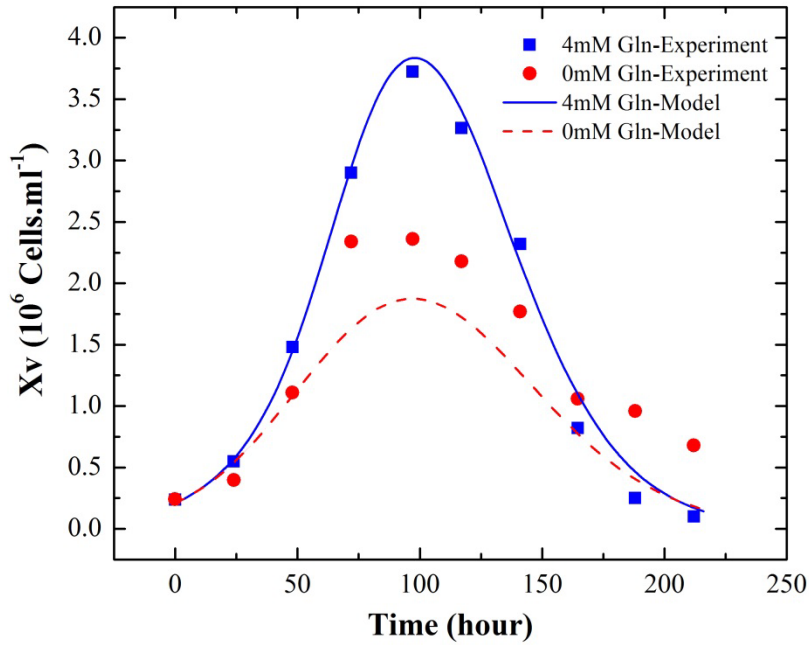
Figure 5-3. Viable cell concentration time profile of two different batch cultures; simulated vs. experimental.

The model slightly underestimates the viable cell density concentration for the culture when no glutamine is added. It should be emphasized that the 0 mM glutamine experiments have not been used for model calibration so as to test the predictive capabilities of the model. Although the culture commencing with 4mM glutamine reaches a higher viable cell concentration, the viability drops sharply after day five. This is probably due to the glucose concentration being depleted more quickly, compared to the culture with no glutamine added at the beginning (Figure 5-4).

The time profiles of the extracellular glucose and glutamine concentrations are shown in Figure 5-4 and Figure 5-5, respectively. With 4 mM glutamine present initially, glutamine exhaustion coincided with the maximum cell concentration. Glucose was exhausted somewhat later, after 6 days of culture.
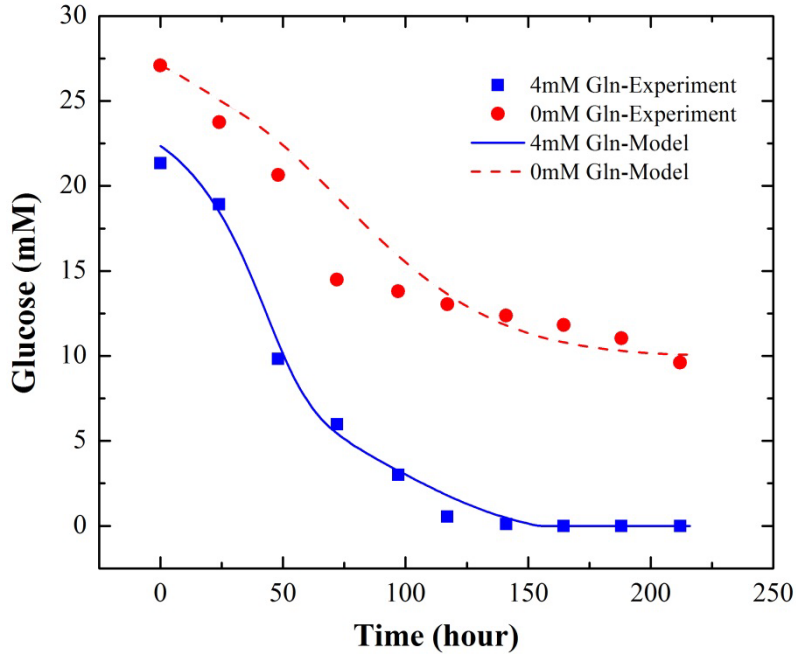
Figure 5-4. Extracellular glucose concentration time profile of two different batch cultures; simulated vs. experimental.

When no glutamine is added, the initial glutamine concentration reflects the residual concentration introduced with the inoculum. In this case, glutamine is rapidly consumed and after 2 days it is below detection limits. It is noteworthy that the minimum measurement threshold of the bioprofile for glutamine is $0.2\ mM$. Due to the observed co-metabolism of glutamine and glucose, glycosylation was expected to stop after glutamine depletion. However, in reality glycosylation continued after the point that glutamine fell below a non-detectable amount. Accordingly, it was necessary to assume a nonzero intracellular glutamine concentration beyond this point, possibly indicating that glutamine is consumed at very slow rates at low concentrations. On the other hand, beyond that point, the glucose concentration was consumed at a much slower rate and significant concentrations were observed and predicted even after nine days of culture. This can be attributed to the co-metabolism with glutamine.

Based on the experimental results (not shown), it can be concluded that this cell line produces a highly galacosylated Mab. The concentration of nucleotide sugars was found to be strongly correlated with glucose and glutamine concentrations. Following the parameter estimation exercise, explained in the previous section, the model was found to be less sensitive to the glutamine compared to the glucose concentration.
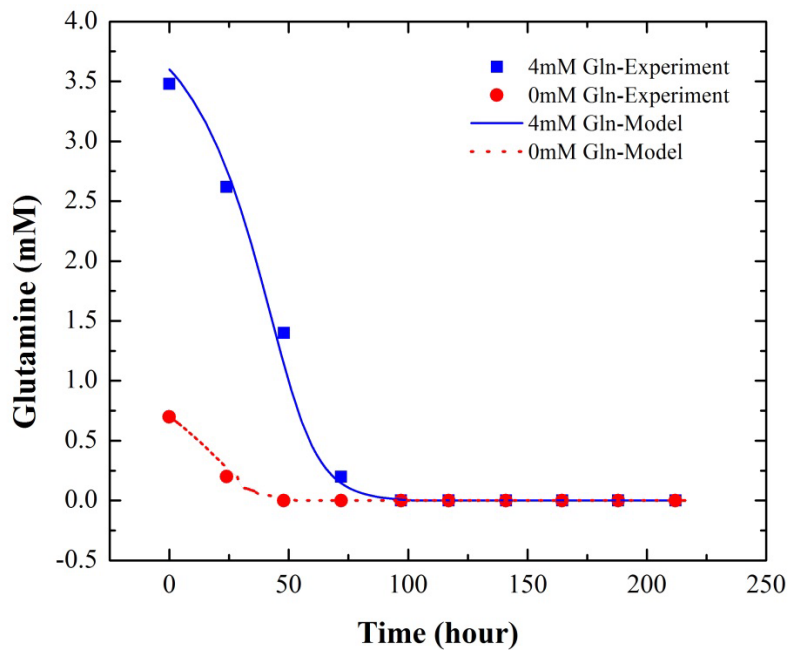
Figure 5-5. Extracellular glutamine concentration time profile of two different batch cultures; simulated vs. experimental.

The evolution of individual glycans strongly depends on the nucleotide sugar levels. For example, Figure 5-6 compares the accumulated composition of F(6)A2G2, a dominant OS structure with two galactose, and F(6)A2G0, with no galactose, along the culture time. The F(6)A2G2 availability diminishes along the culture time. The steepest decline can be seen from approximately the 3rd to the 5th day of the culture, which might be attributed to the glucose reduction exhibited in Figure 5-4. On the contrary, an increase is observed in F(6)A2G0 abundance, supporting the dependency of OS structure on the extracellular metabolites. The model correctly predicted these trends.
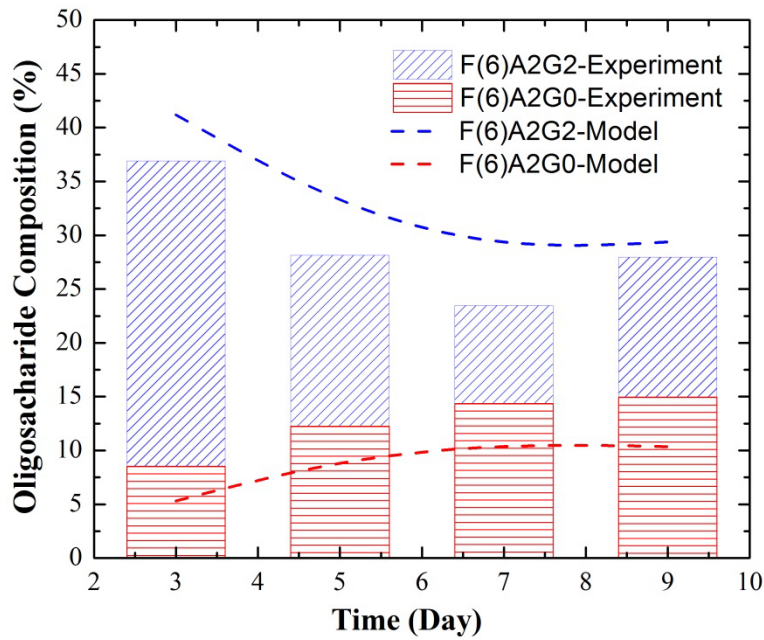
Figure 5-6. Time profile of simulated and measured cumulative oligosaccharides (F(6)A2G0 and F(6)A2G0) concentrations for a batch culture with no additional glutamine at the beginning.

Figure 5-7 provides the time profile of accumulated GI in the culture, illustrating the ability of the model to predict the trends in the index. As explained earlier, due to the consumption of glucose and glutamine, the nucleotide sugar concentration decline with culture time. This leads to an increase in the abundance of agalctosylated OSs and a relative decrease in digalactosylated OSs and hence, a decrease in GI.
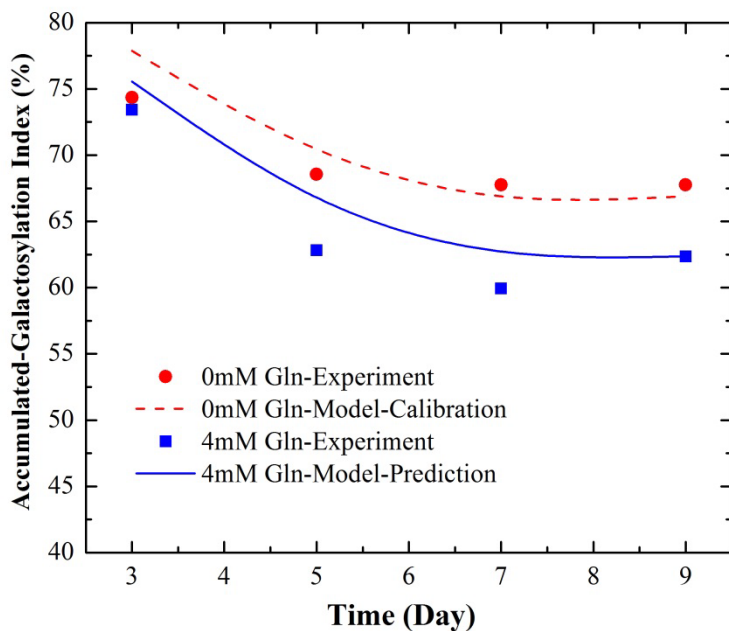
Figure 5-7. Accumulated galactosylation index (GI) time profile of two distinguishable batch cultures. Red: calibration vs. experiment, Blue: prediction vs. experiment.

   As glucose depletes faster in the presence of glutamine (squares in Figure 5-7), the accumulated GI for the culture commenced with initial 4mM glutamine lies beneath the one with no glutamine. The dashed line shows the calibration quality of the model for the GI of culture with no initial glutamine, while the solid line depicts the prediction capability of the model for forecasting the GI for the culture with initial glutamine concentration of 4mM. It is worth mentioning that the measurements of OSs and recombinant protein concentrations suffer from considerable experimental error, due to the large number of experimental steps required for analysis.

5.5 Conclusion

The dynamic metabolic flux model provided an adequate approximation of the extracellular metabolite concentrations, assuming glucose and glutamine co-metabolism. The parameters of this model were calibrated using experimental results for the culture with 4 mM glutamine presented initially and validated by comparing the simulation results with experimental results of the culture with an initial glutamine concentration of zero. The production of nucleotide sugars in the cytosol (or nucleus) as a function of extracellular glucose and glutamine concentrations were formulated. This model plays a crucial role in linking the glycosylation pattern of Mab to the corresponding

extracellular culture conditions. Ultimately, the glycosylation of Mab inside the Golgi apparatus has also been formulated based on the Golgi maturation hypothesis and successfully linked to the above mentioned models. Additionally, for comparing the model predictions to experiments, an averaging calculation was implemented to convert instantaneous glycoprofiles within the cell to accumulated ones. The model correctly predicted dynamic trends in glycosylation indices as well as specific glycan species.

Chapter 6

Intrinsic Fluorescence-based *at Situ* Soft Sensor for Monitoring Monoclonal Antibody Aggregation[†]

Intrinsic fluorescence spectroscopy, in conjunction with partial least squares regression (PLSR), was investigated as a potential technique for online quality control and quantitative monitoring of Immunoglobulin G (IgG) aggregation that occurs following exposure to conditions that emulate those that can occur during protein downstream processing. Initially, the impact of three stress factors (temperature, pH, and protein concentration) on the degree of aggregation determined using size exclusion chromatography (SEC) data, was investigated by performing a central composite design (CCD) experiment and fitting a response surface model (RSM). This investigation identified the influence of the factors as well as the operating regions with minimum propensity to induce protein aggregation. Spectral changes pertinent to the stressed samples were also investigated and found to corroborate the high sensitivity of the intrinsic fluorescence to conformational changes of the proteins under study. Partial least squares regression (PLSR) was implemented to formulate fluorescence-based soft sensors for quality control—product classification—and quantitative monitoring—concentration of monomer. The resulting regression models exhibited accurate prediction ability and good potential for *in situ* monitoring of monoclonal antibody downstream purification processes.

6.1 Introduction

Monoclonal antibodies (MAbs) are the dominant products in the biologics market (Butler and Meneses-Acosta 2012) with the immunoglobulin Gs (IgGs) outpacing other bioproducts (Sahin *et al.* 2010). Hydrophobic recombinant proteins (e.g. IgGs) are inherently susceptible to aggregation (Sadavarte and Ghosh 2014; Sahin *et al.* 2010) during downstream processing following exposure to stress factors such as pH, temperature, and physical shear (Mahler *et al.* 2009; Printz and Friess 2012). Protein aggregation is considered as the most common hindrance for process development (Wang 2005) with such adverse effects as loss of efficacy and/or provoking an immunogenicity

---

[†] Ohadi, K., Legge, R. L. and Budman, H. M. (2014), Intrinsic Fluorescence-based at Situ Soft Sensor for Monitoring Monoclonal Antibody Aggregation. Submitted to *Journal of Biotechnology*-manuscript number: JBIOTEC-D-14-01402

response (Hawe *et al.* 2008; He *et al.* 2010; Mahler *et al.* 2009; Printz and Friess 2012). These issues are fueling the development of *in situ* techniques that enable real-time accurate quality and quantity control of protein aggregation.

Conventional techniques for monitoring protein aggregation, such as size exclusion chromatography, are comparatively time consuming, while intrinsic fluorescence spectroscopy is a fast, non-invasive, and non-destructive technique with high sensitivity and a signal to noise ratio which is amenable to online monitoring (Abbas *et al.* 2013; Elshereef *et al.* 2006; Hawe *et al.* 2008; Ohadi *et al.* 2014a). Fluorescence spectroscopy has been identified as a plausible technique for monitoring conformational changes and characterization of protein tertiary structure (Abbas *et al.* 2013; Kumar *et al.* 2005; Printz and Friess 2012).

Three intrinsically fluorescent aromatic amino acids (phenylalanine, tyrosine, and tryptophan) have been found to be sensitive to the micro-environment they are exposed to in the protein. During protein disruption and aggregate formation, the surrounding environment of the fluorophores change. Accordingly, changes in their fluorescence behavior can be exploited to draw inferences on conformational changes as well as native and non-native characteristics of the protein (Elshereef *et al.* 2006; Kumar *et al.* 2005; Printz and Friess 2012). The higher quantum yield and extinction coefficient of tryptophan makes it a desirable probe to track protein aggregation (Abbas *et al.* 2013; Poole *et al.* 2012). The emission maxima and intensity of tryptophan shift depending on its environment indicating whether it is exposed to the surrounding solvent or buried within the protein (Abbas *et al.* 2013; Poole *et al.* 2012). Multi-wavelength intrinsic fluorescence has previously been investigated for use as a soft sensor for monitoring α-lactoglobulin and β-lactoglobulin solubility under stressed conditions (Elshereef *et al.* 2006; Elshereef *et al.* 2008) and for discriminating between different types of cheeses (Herbert *et al.* 2000). Kumar *et al.* (2005) demonstrated the suitability of second-derivative fluorescence spectra of tryptophan to identify subtle structural changes in β-lactoglobulin and interferon alpha-2a upon exposure to various solvent conditions. Second-derivative fluorescence has also been implemented to qualitatively assess MAb conformational changes under thermal, pH, and solvent stressed conditions (Abbas *et al.* 2013). However, the utilization of intrinsic multi-wavelength fluorescence spectroscopy in the development of a soft sensor for both qualitative and quantitative monitoring of MAb aggregation has not been explored.

The implementation of chemometric methods has been explored in the context of fluorescence-based soft sensor development (Ohadi *et al.* 2014a; Rathore *et al.* 2010; Teixeira *et al.* 2009a). Partial

least square regression (PLSR), one of the most commonly used data exploratory techniques (Wold *et al.* 1984), was utilized to draw indirect inferences from the fluorescence spectra to generate predictive models for monitoring aggregation.

The objective of this study was to develop soft sensors based on *intrinsic* fluorescence, coupled with PLSR, for *at-line* quality control—classification of product—and also for quantitative monitoring—prediction of monomer concentration—at different stages of a downstream purification process. For simplicity and to have better control of experimental conditions, the measurements were collected by exposing the samples to operating conditions analogous to those that occur during protein purification. Given the diverse aggregation patterns induced by various stress factors imposed along downstream processes, IgG samples of different concentrations were subjected to different temperature and pH levels. A response surface method (RSM) was formulated to describe the aggregation patterns induced by different stress factors. RSM was performed to better comprehend the effect and interaction of the stress factors as well as to determine the regions with minimum propensity to cause aggregation. Additional experiments were also conducted to produce a diverse data set for soft sensor development. High-pressure size exclusion chromatography (HP–SEC) was utilized for IgG-sample fractionation and independent quantification. Finally, PLSR-based models were formulated between HP-SEC measurements and fluorescence spectra collected for the corresponding samples. These models can be used as a soft sensor for predicting monomer concentration and inferring conformational changes from measured fluorescence spectra.

6.2 Materials and Methods

6.2.1 Sample Preparation

ChromPure Human IgG (Cedarlane, ON) with a concentration of approximately 11.3 mg/ml in 0.01 M phosphate buffer saline (PBS) was utilized for the experiments. To study the impact of temperature, pH, and protein concentration on the aggregation, an asymmetric central composite design (CCD) with 4 center points was implemented (Table 6-1). Due to the sample limitations and a need for a certain combination of conditions for soft sensor development purposes, it was not possible to perform symmetric CCD. Sample dilution was performed using HyClone 0.01 M PBS (Fisher Scientific, ON) at pH 7. For the experiments performed at different pH levels, the pH of the buffer was adjusted using HCl (1 M) and NaOH (1 M) prior to sample dilution. Samples were thermally stressed at different temperatures in a water bath for 20, 40, and 60 min then cooled for 20 min at

room temperature in a water bath before being centrifuged at 400×g for 5 min. The resulting supernatant was decanted and used for the fluorescence and HP-SEC measurements. For the purpose of developing a soft sensor for *in situ* monitoring of protein aggregation, additional experiments were carried out (Table 6-1).

Table 6-1 Summary of experimental design. Experiment 1-18: Asymmetric central composite design (CCD) with 4 center points and experiment 19-29: Complimentary set of experiments.

Experiment No.	Temperature (°C)	pH	Concentration (mg/ml)
1	70	5.1	0.2
2	70	8.1	0.2
3	70	8.1	1
4	70	5.1	1
5	50	5.1	0.2
6	50	8.1	0.2
7	50	8.1	1
8	50	5.1	1
9	80	7	0.5
10	40	7	0.5
11	60	9.1	0.5
12	60	3.1	0.5
13	60	7	1.5
14	60	7	0.1
15-18	60	7	0.5
19	60	7	1
20	60	7	0.2
21	60	8.1	0.5
22	60	5.1	0.5
23	60	5.1	0.2
24	60	5.1	1
25	60	3.1	0.2
26	60	3.1	1
27	60	9.1	0.2
28	70	7	0.5
29	70	7	0.2

6.2.2 High-Pressure Size Exclusion Chromatography

The resulting supernatant was analyzed by HP-SEC on an Agilent 1200 chromatography system (Palo Alto, CA) equipped with a UV detector. 200 μl of sample were injected into a Speax Zenix-C SEC-300 (Sepax Technologies, Newark, DE) column with a flow rate of 1 ml/min for 20 min with a mobile phase of 1.5 M PBS at pH 7.4 and detected at 280 nm. Measurements were carried out in duplicate. A calibration curve was prepared with various concentrations of pure IgG_1 ranging from 1.5 to 0.01 mg/ml. Protein concentration calculations were based on the area under the curve (AUC) pertinent to the peak of the monomeric form of the IgG_1. To avoid errors related to elution caused by column saturation or attachment of highly hydrophobic molecules, a new calibration curve was generated after column regeneration on a regular basis. The aggregation percentage was defined as the difference between the concentration of the protein in monomeric form before and after being subjected to the imposed stresses. This difference was then normalized with respect to the initial monomer concentration of unstressed sample. As such, the calculated aggregation percentage is equivalent to a measure of product loss.

6.2.3 Fluorescence Spectroscopy

Multi-wavelength fluorescence spectra of the supernatant were acquired at room temperature utilizing a Cary Eclipse spectrofluorometer (Palo Alto, CA) equipped with a Peltier multi-cell holder in 700 µl far UV quartz cells with path length of 10 mm (Mandel Scientific, ON). The measurements were collected over an excitation range from 260 to 350 nm at 5 nm increments and emission range from 280 to 450 nm with 1 nm increments. The photomultiplier tube (PMT) voltage of 600 V, slit width (SW) for excitation and emission of 5 nm, and scanning rate of 600 nm/min were set for signal acquisition. The measured fluorescence spectra were arranged in a form of a matrix (19 excitation×121 emission) known in the literature as an excitation emission matrix (EEM).

6.2.4 Chemometric Analysis

With the goal of developing empirical models (soft sensors) based on multi-wavelength fluorescence spectra for quality control and quantitative monitoring, the PLSR (Geladi and Kowalski 1986; Wold *et al.* 1984) technique was applied. The input data matrix (X) is formed by individual re-arrangement of EEMs of samples into row vectors followed by their row-wise attachment. Sample duplicates were included in the X matrix separately. The formation of the response matrices (Ys) for each soft sensor is explained separately in the corresponding sections. The optimal number of latent variables (LVs)

can be obtained by minimizing the root mean square error of cross validation (RMSE-CV) to avoid model over-fitting. In the current study, a random subset routine with 20 iterations was carried for cross validation (CV). Chemometric analyses were performed utilizing the PLS-Toolbox 7.0.3 (Eigenvector Research Inc., Manson, WA) running in the MATLAB 8.0.0 (Mathworks, Natick, MA) platform.

6.3 Results and Discussion

6.3.1 Impact of Stress Factors

To better understand the impact of stress factors and investigate diverse aggregation patterns, an asymmetric central composite design with a complementary set of experiments (to encompass a wide range of stress factors) was conducted. Additionally, the relevance of the tryptophan fluorescence-spectra of the samples for tracking Mab conformational changes was thoroughly studied. It is worth noting that the samples that were not subjected to stress factors contained less than 5% dimers which are neglected in the current study.

Temperature plays a crucial role to evoke aggregation. Higher temperatures contribute to the protein conformational changes, exposure of hydrophobic regions of the polypeptide chain to the external medium, and consequently to the formation of non-native aggregates (Mahler *et al.* 2009; Printz and Friess 2012).
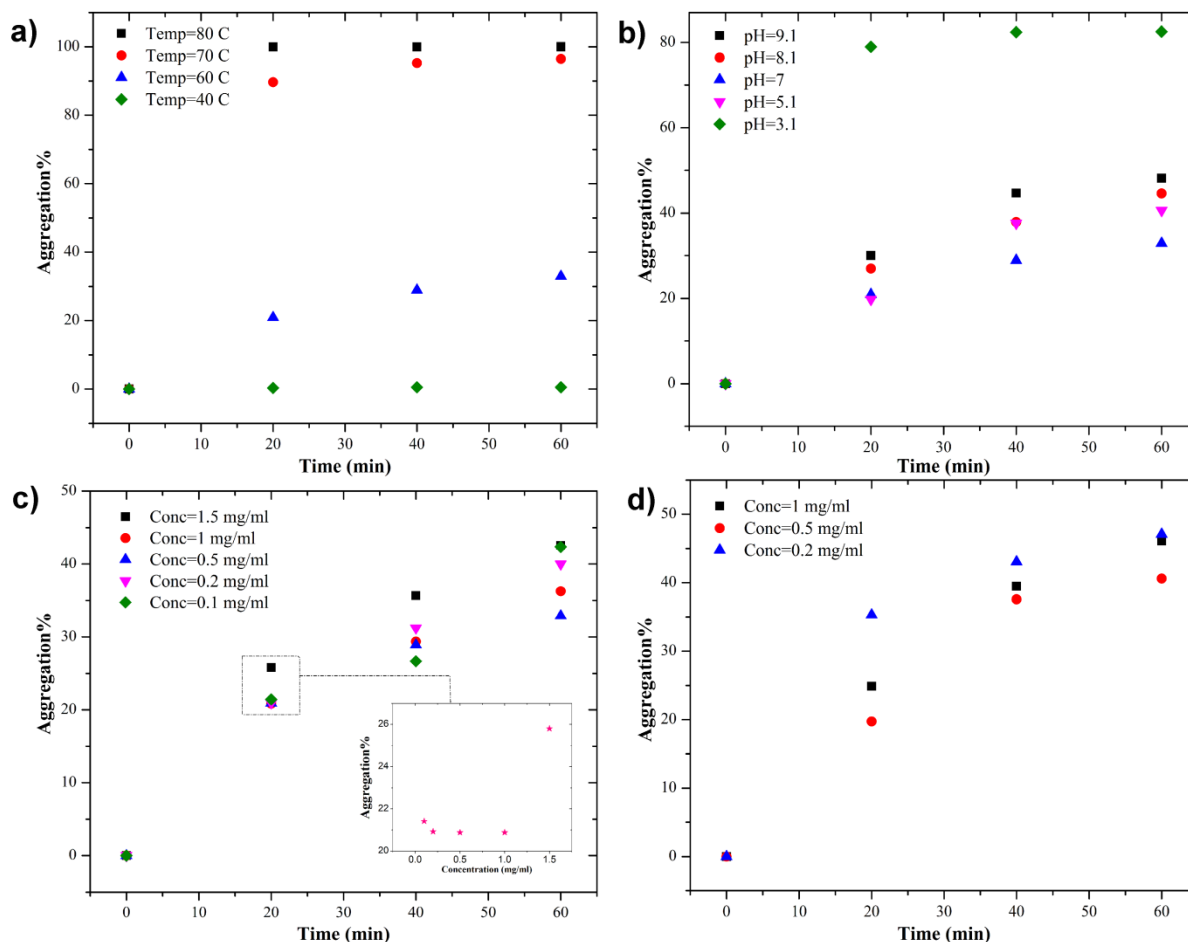
Figure 6-1 Temporal evolvement of the degree of aggregation under stressed conditions for: (a) Samples with 0.5 mg/ml initial IgG at pH=7 under different temperatures; (b) Samples with 0.5 mg/ml initial IgG stressed at 60 °C under different pHs; (c) Samples with different initial IgG concentrations stressed at 60 °C and pH=7; (d) Samples with different initial IgG concentrations stressed at 60 °C and pH=5.1.

Figure 6-1a exhibits the changes in aggregation percentage of IgG samples (pH=7 and concentration=0.5 mg/ml) with time at different temperatures. From the results it can observed that the propensity for aggregation is proportional to the temperature and occurs when samples are heated above a minimum temperature. For instance, over 20% of the IgG monomer is aggregated after 20 min of heat treatment at 60 °C, while at 40 °C the level of aggregation is negligible. Higher temperatures (70 and 80 °C) resulted in a significant loss of the monomer structure as well as induced formation of large insoluble oligomers that precipitate as supported by the chromatograms in Figure

95

6-2a for samples subjected to these higher temperatures (pH=7 and concentration=0.5 mg/ml) for 20 min. This phenomenon is attributed to the formation of greater levels of unfolded IgG at higher temperatures (Hawe *et al.* 2008). Regardless of the initial IgG concentration and pH, similar trends were observed for temperature-induced aggregation (Figure C- 3-Appendix C) which is in accordance with the literature (Hawe *et al.* 2008; Printz and Friess 2012).
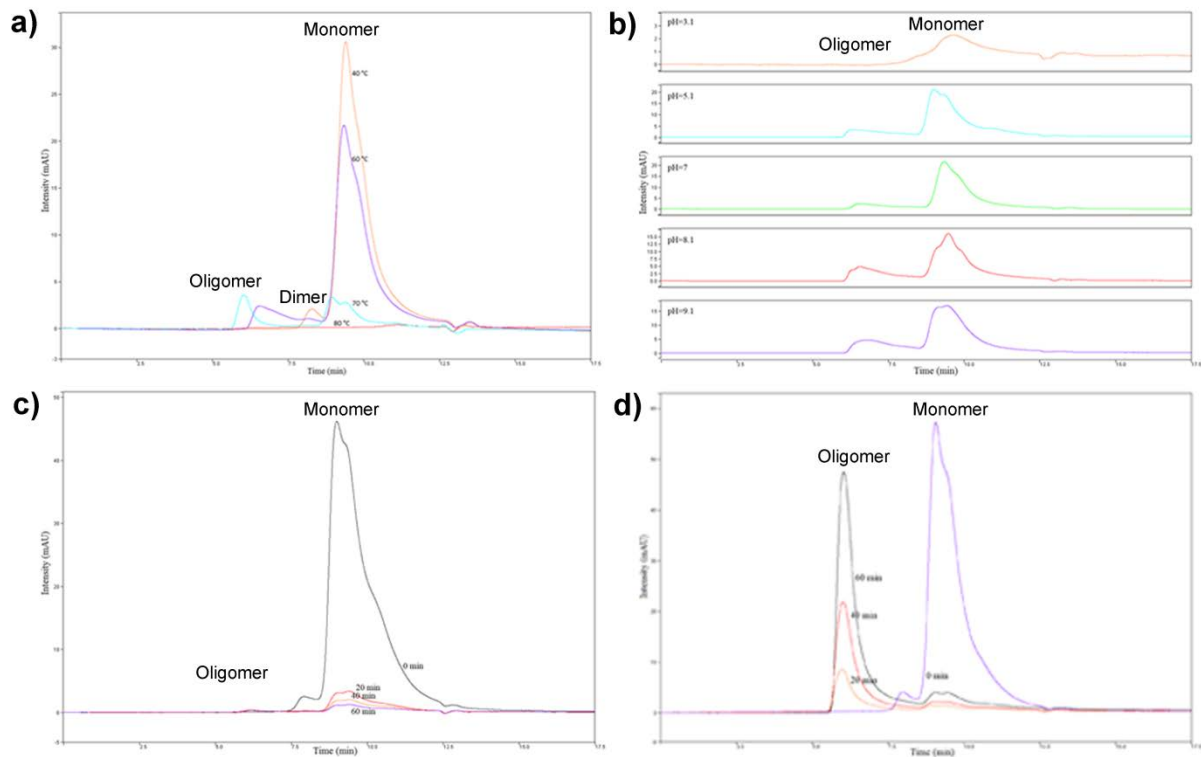
Figure 6-2 Chromatograms of samples obtained from SEC for: (a) Samples with 0.5 mg/ml initial IgG at pH=7 after 20 min of treatment under different temperatures; (b) Samples with 0.5 mg/ml initial IgG stressed at 60 °C for 20 min under different pHs; (c) Samples with 1 mg/ml initial IgG and pH=5.1 stressed at 70 °C at different sampling time; (d) Samples with 1 mg/ml initial IgG and pH=8.1 stressed at 70 °C at different sampling time.

Changes in pH can trigger aggregation by altering the polypeptide charge-distribution (Wang 2005) and the aggregation response to pH depends on the type of the protein. Based on the experimental results obtained at different pHs, it was concluded that samples with pH close to neutral levels had a

very small tendency to aggregate while deviations from pH 7 resulted in an increase in the rate of aggregation (Figure 6-1b). Figure 6-2b is a comparison between the heat-treated samples, at 60 °C monomer concentration of 0.5 mg/ml after 20 min, at different pHs. It can be seen that under acidic conditions (pH=3.1) the rate of loss of monomer is significantly higher. The substantial increase in the intensity of tryptophan fluorescence is in agreement with this conclusion. Additionally, it was observed that at 70 °C and an initial IgG concentration of 1 mg/ml, the sample with pH 5.1 experienced more pronounced loss of monomer and formation of insoluble oligomeric structures over time, in contrast with samples subjected to pH 8.1 that exhibited a higher tendency of forming trimer and tetramers (Figure 6-2c and d).

Although protein aggregation is generally found to increase with protein concentration (Mahler *et al.* 2009; Wang 2005), the pattern of protein aggregation is found to be inconsistent (for the concentration range of this study) under different pH conditions. For instance, at pH 7 a sample with concentration of 1.5 mg/ml experienced the greatest loss of monomers and at the same time formed more trimers/tetramers after 20 and 40 min exposure to 60 °C (Figure 6-1c), while at pH 5.1 the aggregation percentage of the sample with initial concentration of 0.2 mg/ml outpaced that of 0.5 and 1 mg/ml samples (Figure 6-1d). In spite of the high level of aggregation observed under acidic conditions (pH=3.1), the sample with an initial concentration of 1 mg/ml (at 60 °C after 20 min) aggregated 10% and 7% more than samples with initial IgG of 0.5 and 0.2 mg/ml respectively (results not shown).
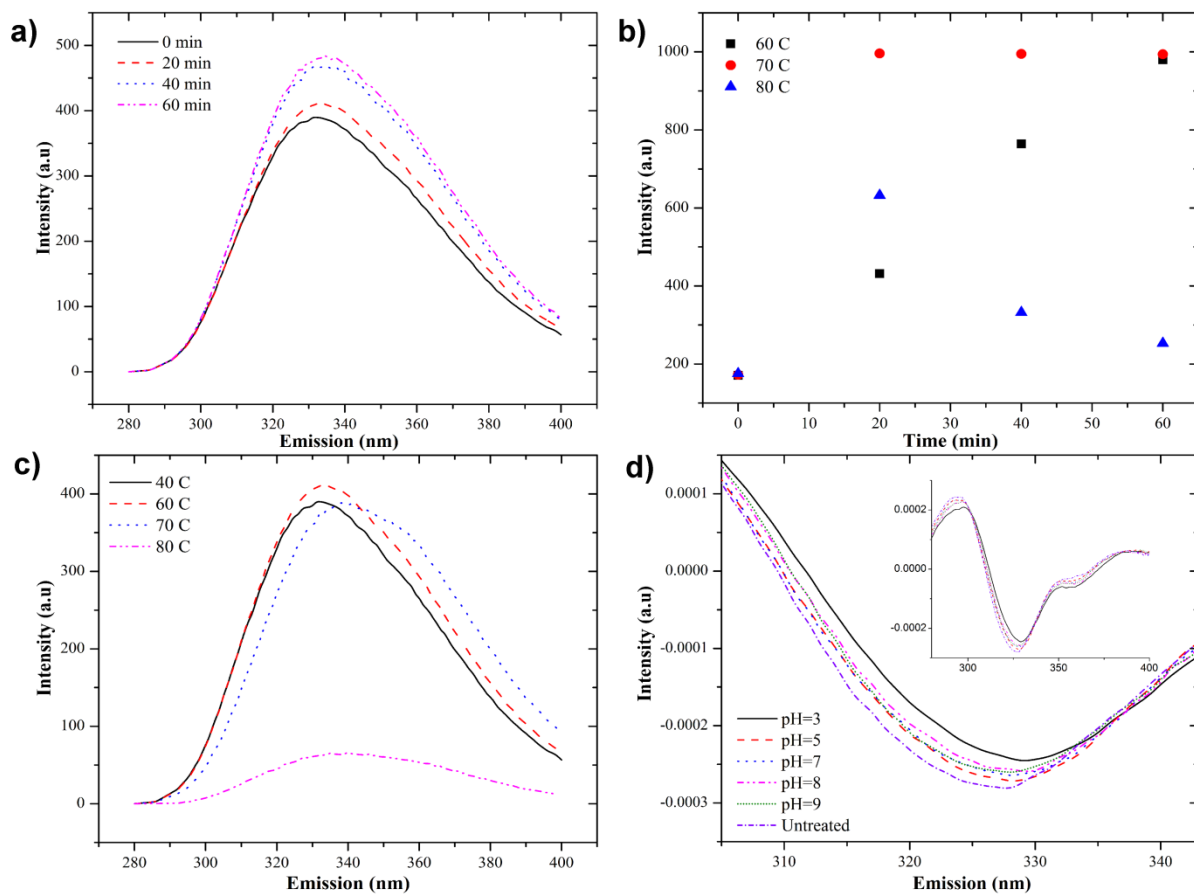
Figure 6-3 Fluorescence signal of stressed samples: (a) Emission spectra (at excitation 280 nm) for samples with 0.5 mg/ml initial IgG and pH=7 stressed at 60 °C at different sampling time; (b) Intensity of scattering-maxima (at excitation 305 nm) for samples with 0.5 mg/ml initial IgG and pH=7 stressed at different temperatures over time; (c) Emission spectra (at excitation 280 nm) for samples with 0.5 mg/ml initial IgG and pH=7 stressed after 20 min at different temperatures; (d) Second derivative of emission spectra (at excitation 280 nm) for samples with 0.5 mg/ml initial IgG after 20 min treatment at 60 °C at different pHs.

Figure 6-3a provides the tryptophan fluorescence (Excitation=280 nm) of IgG samples (pH=7 concentration=0.5 mg/ml) stressed at 60 °C over a 60 min time span. An increase in the tryptophan intensity over time and a red-shift of peak maxima are indicative of protein unfolding and the formation of oligomeric structures that are observed in the HP-SEC results for the same samples during the course of heat induced aggregation. A pronounced change in the tryptophan fluorescence between 20 to 40 min and a subtle change occurring between 40 to 60 min are in agreement with the

98

results in Figure 6-1a. Rayleigh light scattering is an incident light, occurring at equal excitation and emission wavelengths and it is generally attributed to the presence of suspended particles and their size (Ohadi *et al.* 2014a). The scattering intensity (Excitation=Emission=305 nm) for the aforementioned samples continually increases (Figure 6-3b) over time and is attributed to the formation of larger oligomers in the samples in agreement with the protein fractionation results (results not shown). At higher temperatures (80 ℃) the overall protein concentration decreases due to the precipitation of large oligomers. This is evident in the sudden reduction in the tryptophan fluorescence intensity and a higher red-shift in the peak maxima of samples (pH=7 concentration=0.5 mg/ml) after 20 min of heat treatment (Figure 6-3c) as well as the scattering levels (Figure 6-3b) which agrees with the HP-SEC results in Figure 6-2a. Abbas et al. (2013) argued that the second derivative of the normalized tryptophan signal can be implemented to track induced conformational changes. Figure 6-3d provides the second derivative (calculated using the Savitzky-Golay algorithm) of the normalized emission signal obtained at an excitation of 280 nm for a sample with 0.5 mg/ml initial IgG after 20 min of heat treatment at 60 °C under various pHs. A reduction in the peak value at approximately 327 nm is an indicator of the loss of the native structure of IgG with larger disruption occurring for the samples exposed to pH=3. From Figure 6-3d a similarity can be found between the aggregation pattern of samples at pH=8 and 9 and pH=5 and 7, which is in agreement with previous observations inferred from Figure 6-1b and 2b and agrees with the findings of Abbas et al. (2013). The subtle changes in the intrinsic fluorescence-spectra of stressed samples reinforce the capability of fluorescence spectroscopy to serve as a tool for quantitative and qualitative monitoring of aggregation.
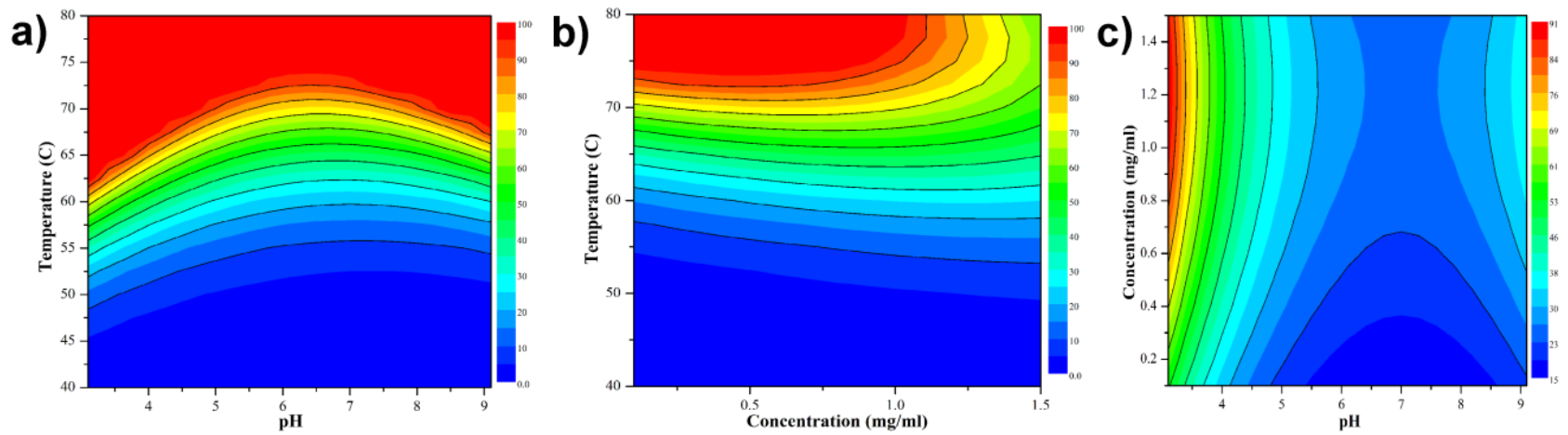
Figure 6-4 Contour plots of fitted surface model for Agg% after 20 min: (a) Temperature vs. pH at initial IgG of 0.5 mg/ml; (b) Temperature vs. concentration at pH=7; (c) Concentration vs. pH at 60 °C

6.3.2 Response Surface Model

An RSM was developed to describe the impact of stress factors on the percentage of aggregation. The experimental error was calculated based on a replicated center point and is estimated to be approximately 12%. The response variable was defined as the percentage of aggregation after 20 min. For better accuracy, a logarithmic transformation of the response variable was used for model calibration. Since the aggregation was bounded to 0% and 100%, the model output was forced to be equal to 0 for prediction values below 0 and 100 for prediction values above 100. Analysis of variance (ANOVA) along with an F-test was performed for various model structures to evaluate the lack of fit for each proposed model. Consequently, a function containing quadratic, interactions, and linear terms was found to be the best fit to the experiments. ANOVA with t-test was then performed to determine significant effects and fine tune the final model. Table 6-2 summarizes the ANOVA of final RSM including only the significant effects. Temperature was found to have the most profound impact on aggregation. It was also noted that the impact of pH and concentration are slightly dependent on the temperature, while no considerable interaction was observed between pH and protein concentration. Additionally, a curvature effect was noted for all three effects which were evident from the conclusion drawn previously. Figure 6-4 provides the contour plots of aggregation percentage after 20 min of treatment, obtained from the surface model, with respect to two factors while the third is kept constant. These contour plots can be employed to evaluate the operating regions where the propensity of aggregation formation is lower. For example, it was found that the tendency of aggregation at pH values ranging from 6 to 7 is comparatively lower than pH levels above or below this range at any temperature (Figure 6-4a) and monomer concentration (Figure 6-4c) investigated. At a protein concentration of 0.5 mg/ml, the impact of basic pH levels on aggregation is more substantial at higher temperatures, compared to lower ones (Figure 6-4a). A similar behavior was observed for other concentrations. Regardless of the sample pH, the degree of aggregation at different concentrations was found to be dependent on temperature (Figure 6-4b for pH=7). To improve the extrapolation accuracy of the surface model at high temperatures and monomer concentrations, more data were acquired for further calibration.

Table 6-2 Analysis of Variance (ANOVA) of the final response surface model (RSM) after excluding the insignificant factors.

	Degree of Freedom (df)	Sum of Square (SS)	Mean Square (MS)	$F_{Observed}$	F_{Actual}
SSE	10	394.72	39.47		
SSLF	7	373.65	53.38	7.596	8.89
SSPE	3	21.08	7.02		

6.3.3 Soft Sensor for Quantitative Monitoring

An empirical model (soft sensor) was developed based on multi-wavelength fluorescence spectra, by utilizing the PLSR method, in order to predict the concentration of the monomeric IgG of pure (unstressed) and stressed samples. This soft sensor can be used *at situ* to track desirable product concentrations as well as monitor loss of monomers. Since the aim of this soft sensor is to estimate the monomeric content, only the part of spectra pertinent to the tryptophan peak were implemented as input to the model. Thus, the scattering region, which is believed to provide information about the suspended particles and is probably more relevant to oligomeric structures, was eliminated from the spectra using a MATLAB program. The concentrations of the IgG in monomeric structure obtained via protein fractionation (as explained in section 6.2.2) were appended together to form the output matrix (Y). To ensure the model-prediction accuracy, experiments 2, 13, 17, 20, 23, and 25 (Table 6-1) were randomly selected and eliminated from the calibration set. These experiments were then used to test the prediction accuracy of the resulting model. Five LVs were found to minimize the RMSE-CV resulting in a minimal error of 0.053 mg/ml of monomer concentration which is equivalent to approximately 16% of the average monomer-IgG concentration; in accordance with 13% SEC measurement error. The resulting model was capable of capturing roughly 100% of variation in the input (X) and simultaneously explains 97.1% of the output (Y) for the calibration set which is deemed highly reasonable considering the R^2_{CV} equal to 96.9%. The resulting model predicted the test set with a precision of R^2_P=97.6% that verifies the accuracy of the cross-validation procedure necessary to avoid model over-fitting. The model precision is depicted in Figure 6-5a where the calibration and validation sets are aligned closely to a 45° line with the residuals being

approximately normally distributed (Figure 6-5a inset). The first latent variable of the PLSR model accounts for 98% and 57% of variance of the input and output data sets, respectively. Although the second LV only contributes to 1% of variation in the X matrix, it can describe roughly 34% of the response matrix Y. Figure 6-5b shows the scores of LV2 versus LV1 with the 95% confidence region. As shown in Figure 6-5b, the scores on the LV1 are discriminated into four regions that correspond to a different range of monomer concentrations. This correlation becomes more apparent in Figure 6-5c that shows the monomer concentration versus LV1 scores. In spite of the group of outliers with extremely low LV2-scores (Figure 6-5b and c), a correlation is observed between monomer concentration and the scores on the LV1. The outliers observed in Figure 6-5b have very low leverage; hence, impose a minimal effect on the model accuracy. Figure 6-5d shows the trajectories of LV2-scores for three different experiments with an initial IgG concentration of 0.5 mg/ml. For experiments performed at 60 °C at pH 7 and 9.1, the LV2-scores are constantly decreasing which is in agreement with the trend of formation of oligomers noted earlier in section 6.3.1. On the other hand, for the experiment performed at 70 °C and pH 7, an increase in LV2-scores is observed after reaching a minimum that can be explained by the formation of precipitates that were eliminated by the centrifugation step performed before the measurement.
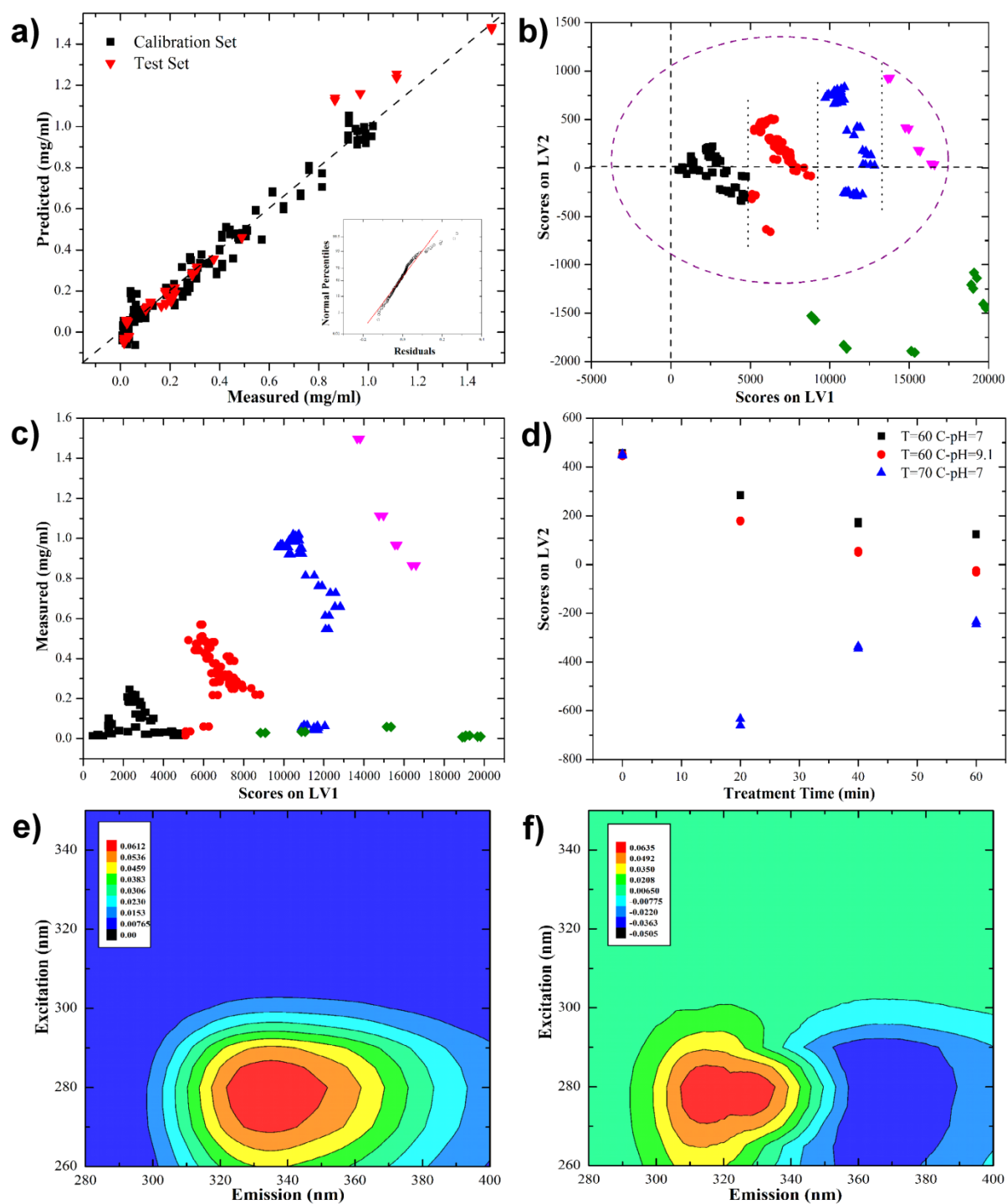
Figure 6-5 Results of the soft sensor for quantitative analysis: (a) Model predictions vs. measured values for calibration and test set and normal percentiles of residuals (inset); (b) Scores of LV2 vs LV1; (c) Measured values vs. scores on LV1; (d) Scores on LV2 for three

different experiments over time; (e) Contour plot of the first loading; (f) Contour plot of second loading.

The physical relevance of the latent variables can be explained by investigating the matrix of loadings. By performing an inverse of the matrix transformation, explained in section 6.2.4, the loadings pertinent to each latent variable were converted to the form of an EEM. Figure 6-5e and f are the contour plots of the first and second LVs. The first loading (Figure 6-5e) is related to the region corresponding to tryptophan with an emission peak at approximately 330 nm, which resembles the tryptophan buried in the hydrophobic core in the native protein and in agreement with previous observation by Elshereef at al. (2006). This conclusion reinforces the stronger relevance of LV1 to the monomeric-IgG concentration. The LV2 encompasses regions of spectra attributed to tryptophan in the native and in the non-native protein with positive and negative correlations, respectively (Figure 6-5f).

Table 6-3 Confusion table for calibration and test set for the soft sensor for quality control.

	Calibration Set				Test Set			
	Class 0	Class 1	Class 2	Class 3	Class 0	Class 1	Class 2	Class 3
Predicted as Class 0	74	8	0	0	13	2	0	0
Predicted as Class 1	0	43	4	0	3	18	3	0
Predicted as Class 2	0	3	26	2	0	0	3	0
Predicted as Class 3	0	0	0	32	0	0	0	6

6.3.4 Soft Sensor for Quality Control

The presence of oligomeric structures of Mab substantially diminishes the quality of the final product and can have profound negative impacts in the patient (Hawe *et al.* 2008; He *et al.* 2010; Mahler *et al.* 2009; Printz and Friess 2012). It imposes a need for a non-invasive approach for rapid quality control of final product. Fluorescence spectra, in conjunction with PLSR, were implemented to develop an empirical model capable of *at situ* classification of the final product into classes with different degrees of aggregation. In order to calibrate the model, samples were grouped into 4 classes that were defined based on the ratio of the percentage of the area under the SEC chromatograms' peak pertinent to tri/tetramers and that of the monomer IgG. For the purpose of model calibration, classes were

assigned corresponding integer values: 0, 1, 2, and 3. A possible description of the classes corresponding to each one of these integer values could be as follow: 0) no to very small aggregates-acceptable sample; 1) mostly monomer with few aggregates-acceptable sample with caution; 2) equal amount of aggregates and monomers-require further evaluation; and 3) over half of the sample are aggregates-rejected sample. Clearly, the description of the classes is subjective and it could be modified if additional information such as the therapeutic efficacy of the Mab were available from other sources. The integer values describing the classes were used to form the response matrix Y that was regressed with respect to the input matrix containing the measured fluorescence spectra of the samples. As mentioned before, the scattering region of the spectra is attributed to larger particles and can provide information about the presence of oligomeric structures. Thus, for the formulation of this soft sensor the complete fluorescence spectra, encompassing both the scattering regions and region attributed to tryptophan, were implemented as the input matrix (X). To validate the model precision, randomly selected experiments were excluded from the calibration step. The cross validation procedure revealed that 6 LVs minimizes the RMSE-CV and generates a model capable of capturing 99.99% of input data set and explains 92.12% of output data with R_{CV}^2 being equal to 90.1%. The subsequent model is capable of predicting the test set with the R_P^2 of approximately 79%.

Figure 6-6a illustrates the model predictions for the calibration set and validation set distinguished based on their specified class. Since the predictions of the model are continuous variables, i.e. continuous values between 0 to 3, thresholds were selected as boundaries of classes in order to assign to each predicted value one of the integer values between 0 to 3. It can be concluded from the Figure 6-6a that the model properly discriminates between classes with few samples being misclassified. Table 6-3 summarizes the confusion table for the calibration and validation sets where this table indicated the percentages of correct and incorrect detections obtained for each data set under study. From this table, the maximum probability for a true positive prediction (sensitivity) is observed for classes 0 and 3 for both calibration and validation sets, which supports the fact that the definite acceptance or rejection of the final product can be done with a high degree of certainty. The confusion table also indicates that the probability of true negative (selectivity) of different classes for calibration and validation sets are over 90%, except for class 1 for the validation set which is roughly 80%. Despite the presence of some misclassified samples, the model demonstrates high selectivity and sensitivity which strengthens the potential of this model for quality control. The first LV plays the most crucial role accounting for approximately 84% of variation in X and predicting 90% of the response Y. From Figure 6-6b, the scores of LV2 vs. LV1, it is observed that there is a strong

106

correlation between the LV1 scores and classes. The physical relevance of LV1 is demonstrated by the contour plot of its loading, Figure 6-6c, which is dominated by the regions of spectra pertinent to scattering and, to a lesser degree, tryptophan-related emission.
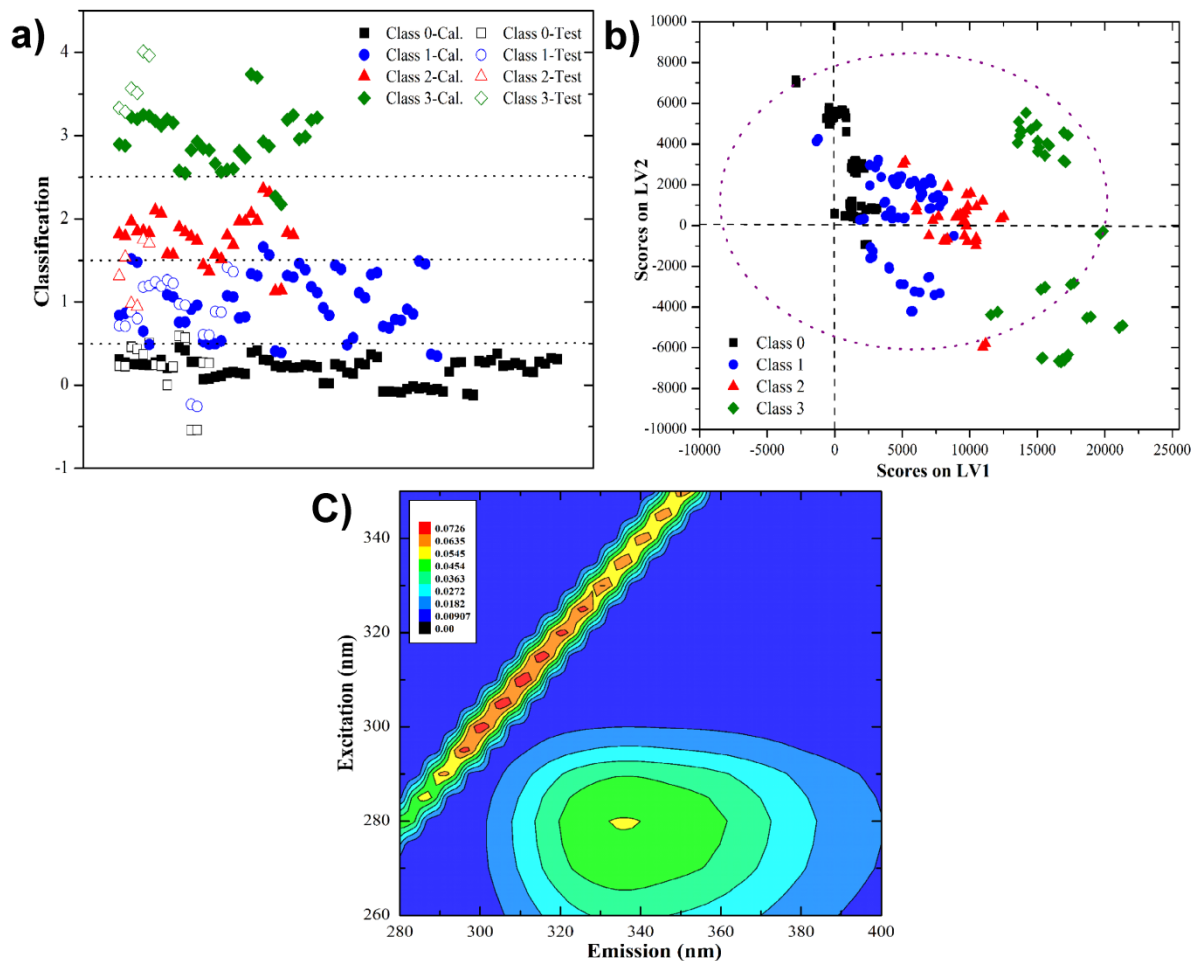
Figure 6-6 Results of the soft sensor for qualitative analysis: (a) Model classification of different samples (initial classes specified by symbols and colors); (b) Scores of LV2 vs LV1; (c) Contour plot of first loading.

6.4 Conclusion

Based on a comparison of HP-SEC and multi-wavelength fluorescence data it is demonstrated that intrinsic multi-wavelength fluorescence spectroscopy can be utilized for at situ quality control and quantitative monitoring of protein aggregation for different process conditions that typically occur during the downstream purification of monoclonal antibodies. To simulate the diverse processing

conditions during downstream processing, IgG samples were stressed under various conditions that included concentration, pH, and temperature that resulted in different aggregation patterns. The impact of these factors on the fluorescence signal was investigated and compared with the results obtained from HP-SEC-based sizing. To systematically comprehend the influence of these operating factors individually or in conjugation with each other, on the degree of aggregation, a non-symmetrical CCD was performed. A surface response model was fitted to the data which helped to identify the operating regions where the propensity to aggregation was lower. An empirical predictive model was then established providing fast and accurate predictions of the monomer-IgG concentration in the sample. Finally, a model was developed using PLSR that successfully discriminated the samples into different classes corresponding to different degrees of aggregation. In view of the adverse impact of aggregation on therapeutic efficacy of antibodies, such a PLSR based model could be extremely beneficial for future online quality control. The feasibility of a fluorescence-based soft sensor for predicting both monomer concentration and aggregation patterns for a diverse set of operating conditions was demonstrated. The proposed approach offers potential for at situ monitoring of downstream processes for monoclonal antibody production.

Chapter 7

Conclusions and Recommendations

The goal of this research was to develop soft sensors for real-time monitoring of monoclonal antibodies production by Chinese hamster ovary (CHO) cells both upstream (bioreactor) and downstream (purification). Two types of soft sensors were developed: data-driven (empirical model or black box) and model-driven (mechanistic model or white-box). Multi-wavelength spectra acquired by spectrofluorometry were used for the development of the data-driven soft sensors. The following discussion summarizes the challenges and the approaches adopted to overcome them.

7.1 Upstream Monitoring

In Chapter 3 evidence was presented that shows the ability of multi-wavelength fluorescence spectroscopy to serve as an analytical tool for *in situ* monitoring of CHO cell cultures. The importance of the spectrofluorometer settings on signal quality and experimental error were investigated by performing a Box-Behnken design. The optimum settings that lead to reproducible and accurate spectra were obtained through a systematic experimental design exercise. With the purpose of model calibration batch cell cultures were performed in shake-flasks with various initial glucose and glutamine concentrations in two different growth media (Media I & II) that resulted in significantly different growth patterns. The temporal evolution of the fluorophore peaks intensities were compared to the dynamics profile of culture variables in order to gain a better understanding about the correlations between them. The conclusion from this step was that the fluorescence spectra can be used for tracking cell culture variables and complemented a previous investigation (Teixeira *et al.* , 2009b). Moreover, the impact of biomass and recombinant protein on the spectra was studied by comparing spectra for culture broth before and after centrifugation as well as following protein A purification. These results were in agreement with the conclusions drawn from comparing the time evolution of fluorophore peaks and culture variables that indicated strong correlations between these variables. To develop soft sensors capable of predicting the actual state of the culture, PLSR was implemented to draw inferences about process variables from the fluorescence spectra. Preliminary modeling revealed that soft sensors developed using data from cultures grown in Media I resulted in less accurate predictions of culture variables for cells grown in Media II since the variation in growth Media had not been incorporated into the models. However, the models did track the trend in

109

variables properly thus reinforcing the notion that there are strong correlations between the dynamic evolution of fluorophores and other culture variables. To capture more accurately the dynamic behavior of the culture variables, the data set was randomly sliced into calibration and prediction sets (each containing cultures grown in both Media 1 & 2). Individual soft sensors were calibrated using PLSR to track viable cell, dead cell, recombinant protein, glucose, and ammonia concentrations. The physical relevance of the spectra used to predict each attribute was investigated in depth by studying the loading plots of PLSR models that corroborated the prior conclusions regarding the existence of clear correlations between fluorescence spectra and other culture variables. The soft sensors demonstrated accurate prediction quality after being tested using the prediction set. No previous study has investigated the impact of diverse growth media on model development and soft sensors capable of tracking dead cell density and metabolite concentrations for mammalian cell cultures as developed in this work.

Chapter 4 developed the hypothesis that data-driven soft sensors, such as the fluorescence-based soft sensors presented in Chapter 3, are not accurate for extrapolation since they lack mechanistic insight into the process dynamics and are merely empirical. Thus, their prediction quality might diminish while predicting conditions that considerably diverge from the ones used for model calibration. Additionally, since fluorescence measurements are acquired at discrete sampling intervals the resulting soft sensors based on these infrequent measurements are incapable of capturing dynamics of the process in between measurement instances. A framework was proposed in Chapter 4 based on an extended Kalman filter (EKF) that combines the results obtained from a fluorescence-based soft sensor with a mechanistic model with the goals of: (a) filtering the noise from the estimates of the fluorescence-based soft sensor and (b) improving the accuracy of predictions of the culture variables in between sampling instances. Although the Kalman filter is a well-established technique, the proposed approach involving the combination of a fluorescence-based soft sensor and a metabolic flux based model has not been previously reported for mammalian cell cultures. Accordingly, a hybrid (continuous-discrete) EKF was implemented in this study to combine model-driven and data-driven soft sensors. To account for structural and parameter uncertainty a non-stationary error was assumed for three model parameters. A key challenge in formulating the EKF is the calculation of the covariance matrices of measurement and process noise. Approaches were presented in this chapter to systematically identify these values from experimental data. The EKF-based soft sensor significantly enhanced accuracy of predictions at measurement instances. A critical advantage of EKF-based soft sensor is to track process variables in between sampling instances. This advantage is beneficial to

address the loss of accuracy associated with less frequent sampling, missing data, and outliers. To better comprehend this advantage, predictions of the dynamic behavior in between sampling points of the EKF-based soft sensor were compared with a simplistic first order hold extrapolation approach obtained from the fluorescence-based soft sensor's measurements. The prediction accuracy of the EKF-based soft sensor substantially surpassed the first order hold extrapolation approach.

The importance of the N-linked glycosylation on the therapeutic efficacy of Mab was discussed in Chapter 2. Preliminary investigations conducted revealed that the multi-wavelength fluorescence spectra cannot directly predict the glycosylation profile of the recombinant protein produced. However, as reported in the literature and experimentally observed here, the instantaneous glycosylation profile is correlated with the extracellular metabolic state of the culture. Thus, in Chapter 5 a mechanistic model was proposed relating the extracellular glucose and glutamine concentrations to the extent of the glycosylation. The model consisted of three blocks: (i) a dynamic metabolic model based on metabolic flux analysis (MFA) that predicts the extracellular state of the culture, (ii) a nucleotide sugars' model that predicts the production of the nucleotide sugars, required for the glycosylation, in cytosol (or nucleus), and (iii) a dynamic model that simulates the glycosylation reactions occurring in the Golgi apparatus. The dynamic metabolic model and the glycosylation model (blocks i and iii) have been individually investigated in the literature. However, connections have not been made between the extracellular metabolites' models to the dynamic glycosylation processes. Accordingly, a key novelty of this work was the development of the nucleotide sugar model that links the extracellular condition to the glycosylation inside the Golgi. This model was developed based on the reaction network of nucleotide sugars—relating the extra cellular glucose and glutamine to nucleotide sugars—obtained from the database Kyoto Encyclopaedia of Genes and Genomes (KEGG), where the intermediate reactions are lumped together for simplicity. Due to the unavailability of experimental data of nucleotide sugars the parameter estimation was a challenging task. To be able to compare the prediction results to the experimental data an averaging step was added to the model that converts the simultaneous glycoform profile to the accumulated form. In Chapter 5 assumptions and approaches were presented to overcome the parameter estimation related challenges for the proposed model. The computation time was another complexity in the parameter estimation problem that resulted from the large number of glycosylation reactions, unknown parameters, and number of iterations to predict over 9 days of culture. Systematic trimming of the network of reactions, utilizing matrix calculation, favoring binary matrices instead of conditional statements, and adopting parallel calculation by modifying the program were approaches

taken to drastically reduce the CPU time. The model performance was tested with a set of experimental data that was not used for model calibration. The resulting model was found capable of tracking the glycosylation profile of the Mab over the course of culture and it will be instrumental for identifying operating conditions that lead to a desirable glycan profile. It is also worth mentioning that in the context of process monitoring the first block of this model could be substituted with the EKF-based soft sensor, presented in Chapter 4, for at-line or online estimation of the extracellular state of the culture. Then, the extracellular state could be used to infer the corresponding glycosylation pattern by using the model proposed in Chapter 5.

7.2 Downstream Monitoring

In a standard manufacturing process the Mabs produced need to be purified using various downstream processes, which expose the product to various stress factors such as pH, temperature, ionic strength, and shear stress. In Chapter 6 the capability of multi-wavelength fluorescence spectroscopy for qualitative and quantitative monitoring of Mab aggregation was demonstrated. The use of intrinsic fluorescence spectroscopy for the development of predictive models for Mab aggregation has not been previously documented in the literature. In order to emulate the process conditions occurring during downstream processing. For the sake of simplicity, samples were exposed to changes similar to the conditions that occur during purification rather than conducting an actual purification process. An asymmetric central composite design (CCD), consisting of three factors: temperature, pH, and initial recombinant protein concentration, was used to obtain diverse aggregation patterns. Mab samples were exposed to these factors over a 60 min time period (with 20 min sampling instances). The stressed samples were then analyzed using multi-wavelength fluorescence spectroscopy and high-pressure size exclusion chromatography (HP-SEC). The qualitative impact of stress factors on the formation of oligomers (or loss of monomers) and conformational changes were studied by comparing the chromatograms obtained from the HP-SEC and tryptophan spectra obtained with the spectrofluorometry. The changes in fluorescence spectra resulting from conformational changes in recombinant protein structure of samples exposed to pH and temperature changes were in agreement with the corresponding changes observed in the chromatograms obtained from HP-SEC thus reinforcing the hypothesis that fluorescence spectra can be used to track protein aggregation. To better understand the impact of these parameters on the degree of aggregation (calculated based on HP-SEC results) and to identify a set of conditions with less propensity to cause aggregation, a response surface model was formulated and studied. With the

112

purpose of developing soft sensors for predicting aggregation the experimental data were randomly divided into calibration and test sets. Using the fluorescence spectra in combination with a PLSR model, a soft sensor was developed capable of predicting the monomeric content of samples (exposed/not exposed to changes in temperature or pH). This empirical based model predicted the test set with high precision. Finally, a soft sensor was formulated using fluorescence spectra in combination with PLSR regression model for classifying the final product into classes with different degrees of aggregation. To calibrate this model, samples were classified into 4 classes pertinent to their level of tri/tetramer and monomer Mab with integer values assigned into each class. The model was found to accurately predict the different levels of aggregation by using the test set data. In summary, the study performed in Chapter 6 demonstrated that fluorescence spectroscopy is a potential tool for *at situ* monitoring of downstream processes of recombinant protein production.

From the work presented in this thesis it can be concluded that intrinsic fluorescence spectroscopy is a viable non-invasive and non-destructive technique for real time monitoring of mammalian cell cultivation at different stages of the process. The inherent limitations of this technique can be addressed by introducing process knowledge through mechanistic models in combination with empirical ones.

7.3 Future Work Suggestions

7.3.1 Use of a Fiber-Optic Probe for Online Fluorescence Acquisition

The fluorescence-based soft sensor developed in this research (Chapter 3 and 6) relies on cuvette-based signal acquisition and therefore only suitable for *at situ* monitoring. The spectrofluorometer can be equipped to acquire fluorescence signals using a fiber optic-based probe. This may allow for extending the capabilities of the proposed soft sensors to online monitoring. For bioreactor applications where sterility is critical, the fluorescence signals could be acquired through a quartz window. The implementation of such an online measurement would require the mitigation of perturbations arising from the presence of air bubbles and motion due to stirring.

7.3.2 Use of Sensor-Fusion to Improve Predictability of the Soft Sensor

Vibration techniques such as near infrared (NIR) spectroscopy are capable of detecting materials such as metabolites, e.g. glutamine (Teixeira *et al.* 2009a). NIR has been formerly investigated as a potential tool for monitoring mammalian cell cultivation (Harthun *et al.* 1998; Sellick *et al.* 2010).

The NIR data for the culture broth could be combined with the fluorescence data to form the input matrix (X) to be used for soft sensor calibration. Such a combination of different types of measurements, referred to in the literature as sensor fusion, could lead to improved prediction accuracy of the soft sensor developed in Chapter 3. Additionally, with this approach, the calibrated soft sensor could be used to predict the concentrations of different metabolites that could not be properly inferred from fluorescence signals alone.

Such a combination of measurements could be also implemented with an extended Kalman filter (EKF) framework as presented in Chapter 4. Due to the larger number of observed variables, the consequent EKF-based soft sensor prediction accuracy, at and in between sampling instances, would be expected to improve.

As demonstrated in this thesis, fluorescence spectroscopy is capable of providing information about the conformational changes of the monoclonal antibody, i.e. protein tertiary structure. In contrast, NIR spectroscopy would provide information on the secondary structure of the protein (van de Weert and Jørgensen 2012). The applicability and prediction accuracy of the soft sensor developed in Chapter 6 could be enhanced by incorporating the NIR signal along with the fluorescence spectra in the input data set (X) for model calibration using partial least square regression (PLSR). In order to tackle the possible nonlinear correlation between the input data set (X), that would consist of the combination of fluorescence and NIR data, and the output data set (Y) involving protein aggregation data as measured by SEC, support vector machine (SVM) could be utilized in combination with partial least square (PLS) or principal component analysis (PCA). SVM is a supervised learning technique capable of capturing nonlinear relationships between the input and output without increasing the dimensionality of the model thus avoiding over-fitting.

7.3.3 Improvement of the Glycosylation Model

The culture broth pH and cultivation temperature have been reported to have a significant impact on the growth pattern of the cells as well as the glycosylation pattern of the Mabs (Aghamohseni *et al.* 2014). The effects of pH and temperature could be integrated into the models presented in Chapter 5. The metabolic model could be modified to account for the diverse growth pattern due to changes in culture pH and temperature. Additionally, the glycosylation processes occurring in the Golgi apparatus could be affected through changes in the enzymatic activity.

7.3.4 Inference of the Mab Glycosylation Pattern

The presence of glycans on the backbone of the Mab has been reported to have a significant impact on their aggregation pattern. This phenomenon could be used as an indirect approach for inferring the glycan structure of the Mab. For instance, Mabs with different glycosylation levels could be subjected to temperature and pH changes such as to cause protein aggregation. The resulting aggregation levels could then be correlated to the degree of glycosylation. This could be a more practical and more inexpensive approach for monitoring glycosylation compared to the time consuming methods currently available, e.g. hydrophilic interaction liquid chromatography (HILIC) followed by exoglycosidase enzyme array digestion. Also, the empirical model presented in Chapter 5 could be used in combination with the aggregation results to enhance the prediction of the glycosylation pattern.

Appendix A

Protein Aggregation: Analytical Techniques

Method	Size range	Application	Advantages	Disadvantages	References
Visual inspection	50 µm - mm	Determination of the absence or presence insoluble aggregates	Easy to perform Provides information on size/ shape of particle and clarity of solution	Does not give quantitative results Cannot be used with soluble aggregates	(den Engelsman *et al.*, 2011)
SDS-PAGE	kDa - MDa	Separation in a gel, by an electric field Size (molecular weight) estimation	Differentiates between covalent and non-covalent aggregates	Separated protein bands must be stained in order to be visible, making quantification of results difficult	(den Engelsman *et al.*, 2011) (Wang, 2005) (Mahler *et al.*, 2009)
AF4	1 nm - 50 µm	Size estimation and quantification of soluble aggregates	Does not need a stationary phase Can be combined with other techniques (UV, MALLS) Has a large detection range	Difficult to validate Dependent on size of aggregate Less rigorous than chromatography	(Mahler *et al.*, 2009) (den Engelsman *et al.*, 2011)
MALLS	kDa - MDa	Detect different aggregate species & determine their molecular mass	No calibration required Independent of order of elution Can be combined with SEC to give useful results	Highly sensitive to contamination of sample Concentration of elution fraction, and differential refractive index increment must be known	(den Engelsman *et al.*, 2011) (Mahler *et al.*, 2009)
DLS	1nm - 5µm	Determination of size distribution of soluble aggregates	Non-destructive Non-invasive Allow samples to be reused Very small volume of	Does not yield quantitative results Highly sensitive to contamination (dust, bubbles) &	(Mahler *et al.*, 2009) (den Engelsman *et al.*, 2011)

				sample required for analysis Rigorous sample preparation not needed	large aggregates	
CD	N/A	Analysis and identification of secondary, tertiary or quaternary structures of proteins	Non-destructive Small sample volume required Samples can be dissolved in aqueous buffers without need for co-solvents	Data obtained is difficult to interpret, must be deconvoluted before structures can be quantitatively assessed Light scattering particles interfere with results	(den Engelsman *et al.* 2011; Li *et al.* 2006)
FT-IR	N/A	Identification and characterization of secondary structures	Non-destructive Convenient Absorption spectra can be obtained for proteins in a range of environments Small sample volume required Does not require calibration	Low sensitivity Aqueous solvent may interfere with absorption High (>10mg/ml) concentration of protein & small path length needed	(Kong and Yu 2007; Mahler *et al.* 2009)
NMR	N/A	Identification and analysis of secondary protein structures	Non-invasive Used instead of x-ray techniques – eliminate exposure to radiation	NMR spectra of large proteins difficult to interpret Concentration dependent Insensitive Only works if aggregate soluble	(Li *et al.* 2006; Surewicz *et al.* 1993)
Fluorescence spectroscopy	N/A	Highly sensitive method for protein quantification and characterization	Rapid Non-invasive Non-destructive Can be combined with microscopic tools Possible online detection	Low resolution Can be affected by polarity of solvent Presence of dissolve oxygen can reduce emission due to oxidation of the protein Use of dyes may affect protein conformation Impurities may affect fluorescence	(Abbas *et al.* 2013; Elshereef *et al.* 2006)

Appendix B

Optimization of Spectrofluorometer Settings

B. 1 Fluorescence Measurements

A Varian Carry Eclipse Fluorescence spectrofluorometer equipped with a Peltier multi-cell holder, was used to acquire spectra from samples of Bio-Grow media. To monitor the evolution of various amino acids, vitamins, and cofactors, that may fluoresce in a wide range of excitation emission the instrument settings have to be carefully selected to generate statistically accurate results. In this study, fluorescence spectrometry was performed at room temperature (~25˚C) within an excitation and emission range of wavelengths of 240-500 nm (with increments of 10 nm) and 280-550 nm respectively to encompass the emission-excitation wavelengths' combinations corresponding to the main fluorophores (Lindemann *et al.* 1998; Peiris *et al.* 2009; Ryan *et al.* 2010; Teixeira *et al.* 2011). Three different parameters of the spectrofluorometer, including photomultiplier tube (PMT) voltage, scanning rate, and slit width (SW), were manipulated to obtain consistent results with low noise to signal ratio as well as a high sensitivity spectrum. MATLAB 7.11.0 software (Mathwork Inc., Natick, Massachusetts, U.S.A.) was utilized for data processing. In some cases to increase the accuracy with respect to the fluorophores' peaks, scattering related signals were eliminated using MATLAB based software (Zepp *et al.* 2004).

A 3^3 Box-Behnken (Montgomery 2009) experimental design was implemented to get a preliminary assessment of the impact of the instrument settings as well as finding the feasible range of values for these settings. The experimental design is summarized in Figure B-1 and was performed for samples of three different ratios of media to buffer (1:0, 1:3, 1:19).
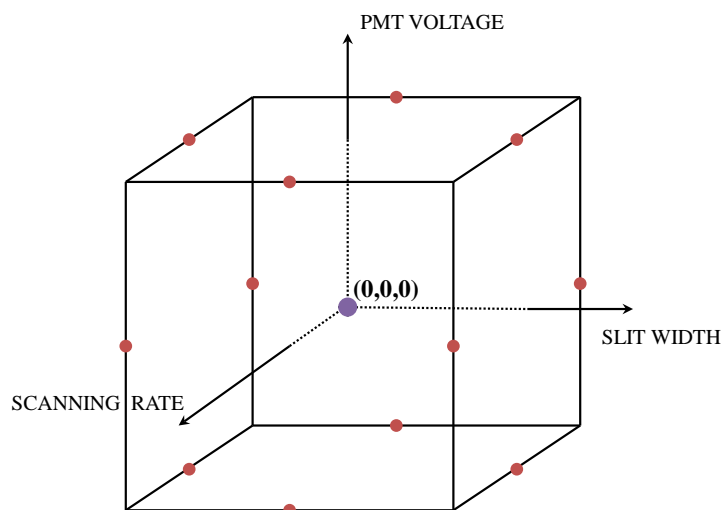
Figure B-1. Box Behnken design for 3 instrument parameters (A) PMT voltage: Low=400V, Medium =600V, High =800V; (B) SW: 2.5 nm, 5 nm, 10 nm for both excitation and emission; (C) Scanning rate: Medium (600 nm/min, Em interval of 1 nm), Fast (1200 nm/min, Em interval of 2 nm), Fastest (9600 nm/min, Em interval of 2 nm)

Based on the preliminary measurements, it was found that low intensity signals are associated with higher noise to signal ratio, while signal intensities above a particular strength exceed the detector's measuring range. Signal resolution is another factor that is affected by instrument settings. It was also observed as shown in Figure B- 2 that following dilution peak intensities related to some fluorophores increase until a certain dilution and then decrease with further dilution. It is believed that this initial increase of peak intensity followed by a decrease may be due to quenching of the fluorescence signal and dilution of quenching agents.
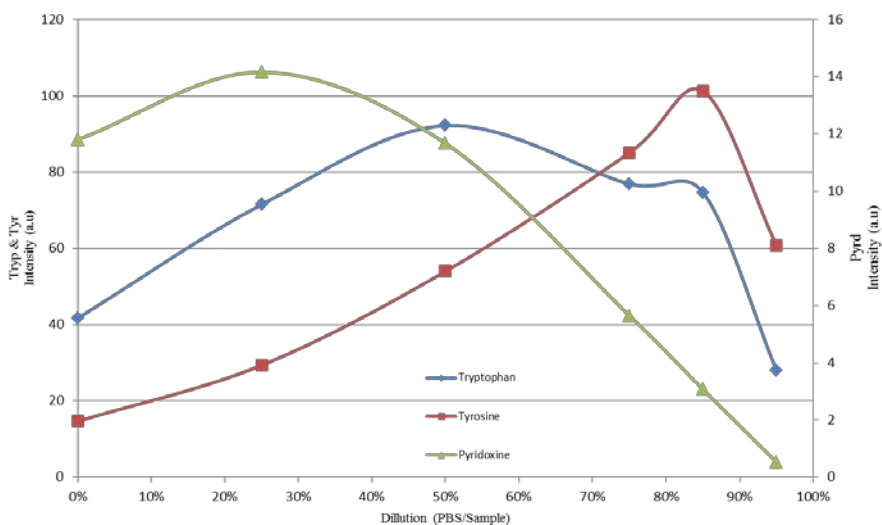
Figure B- 2. Tyrosine, tryptophan, and pyridoxine peak intensity (a.u) vs. media to PBS ratio.

B. 2 Results and Discussions

To further assess the effect of the device settings, samples were collected in triplicates and 95% confidence intervals error as well as signal to noise ratio have been calculated and compared with different instrument settings. Impacts of instrument settings are as follow:

B. 2.1 PMT Voltage

At constant slit width of 5 nm and medium scanning rate an increase in PMT voltage from 400 V to 600 V, significantly improved signal`s noise but a subtle improvement was observed with increase in PMT voltage from 600 V to 800 V. Additionally, the increase in PMT voltage from 400 to 600V decreased error dramatically but further increase of PMT voltage to 800V caused the error to rise significantly. These observations are consistent with previous studies (Casado-Terrones *et al.* 2007; Peiris *et al.* 2009).

B. 2.2 Slit Width

Possibly the most influential instrument setting is the emission and excitation slit-width (SW) as it significantly affects the trade-off between noise to signal ratio and signal resolution. It is important to notice that SW cannot be increased above a certain point, constrained by Equation B-1 (Peiris *et al.* 2009). Respectively, tyrosine with a Stokes-shift of approximately 40 nm, suggesting that both emission and excitation SW should not exceed 10 nm.

$$2 \times (SW_{ex} + SW_{em}) \leq Stokes.Shift$$

Equation B- 1

With constant PMT of 600 V and Fast scanning rate measurement it was revealed tha, error tends to diminish when shifting to higher SW, although the improvement from SW equal to 5 nm to 10 nm is negligible.

Regarding the signal resolution, it was concluded that lower SW contributes to better resolution but higher noise, which corroborates with previous findings (Peiris *et al.* 2009). These results also suggest that SW above 10 nm is not satisfactory from the *resolution* point of view.

B. 2.3 Scanning Rate

At constant PMT of 600 V and SW of 5 nm, no significant difference was observed in the results acquired in different scanning rates. Thus, since fluorescence spectroscopy is aimed to be used as an *in situ* monitoring technique; higher scanning rates are more favorable.
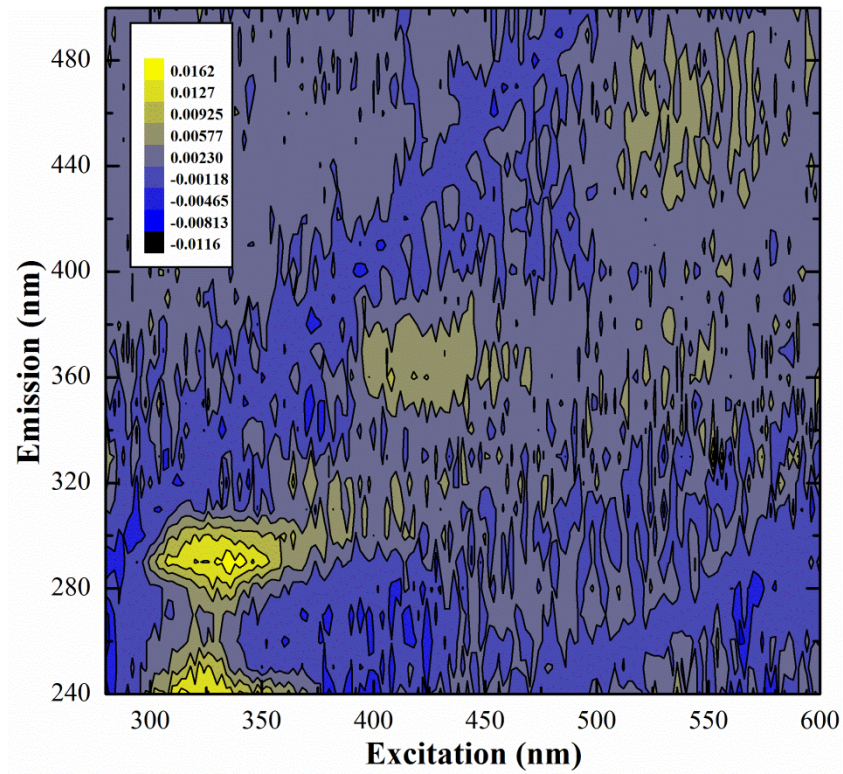
Appendix C

Additional Graphs

Figure C- 1. Contour plot of regression coefficient for viable cell density model presented in Chapter 3
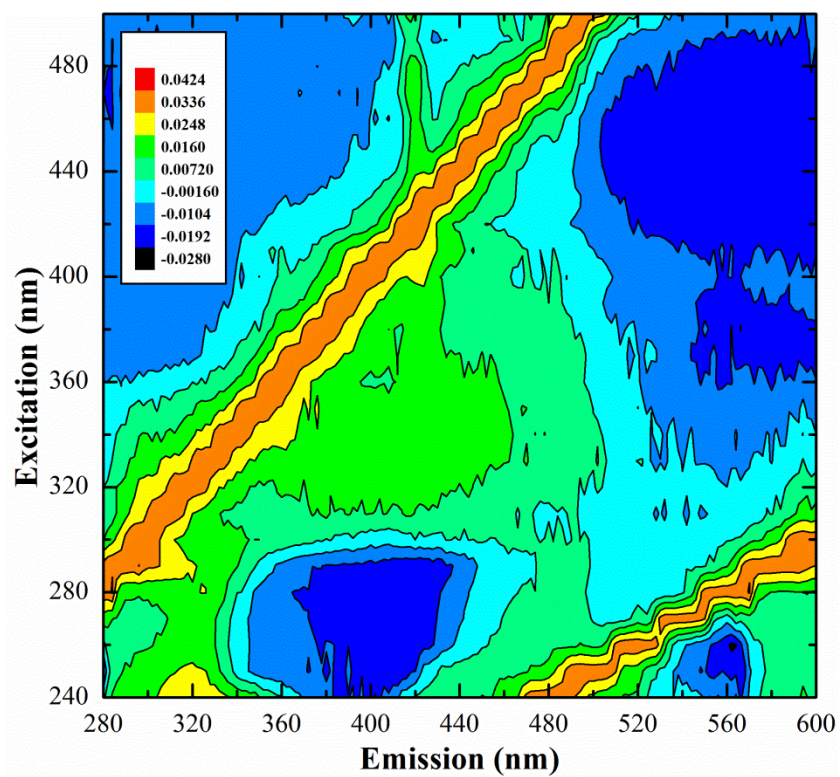
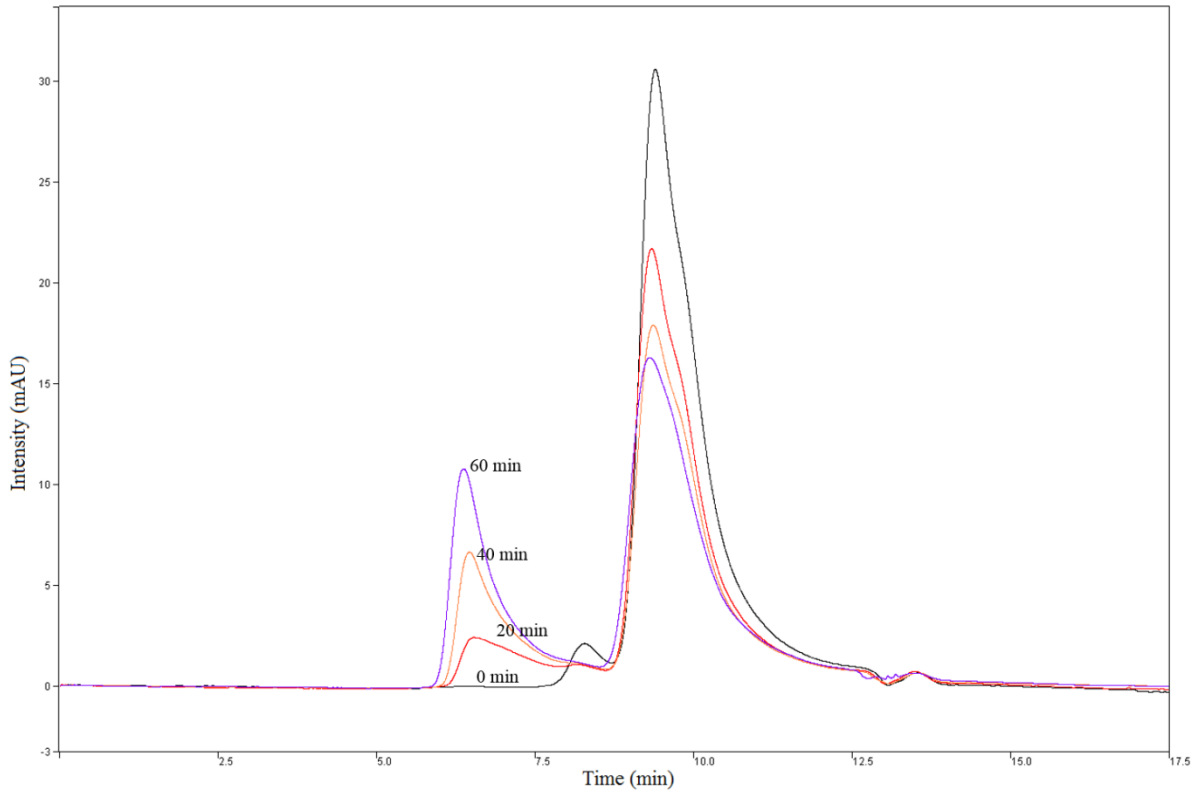**Figure C- 2. Contour plot of second latent variable of model presented in Chapter 4**

Figure C- 3. Het treated samples with 0.5 mg/ml initial IgG at 60 °C and pH 7 for 60 min

Appendix D

Supplementary Information for Chapter 5

Below are the set of equations related to the production of nucleotide sugars:

$$\frac{d[UDP-Gal]}{dt} = \frac{r_1^{max} \times [Glc] \times [UTP]}{(K_1^{Glc} + [Glc]) \times (K_1^{UTP} + [UTP])} - K_1^{Diss}[UDP-Gal]$$

$$\frac{d[UDP-GlcNAc]}{dt}$$
$$= \frac{r_2^{max} \times [Glc] \times [UTP]}{(K_2^{Glc} + [Glc]) \times (K_2^{UTP} + [UTP])}$$
$$- \frac{r_3^{max} \times [UDP-GlcNAc] \times [CTP]}{(K_3^{UDP-GlcNAc} + [UDP-GlcNAc]) \times (K_3^{CTP} + [CTP])} - K_2^{Diss}[UDP-GlcNAc]$$

$$\frac{d[CMP-Sial]}{dt} = \frac{r_3^{max} \times [UDP-GlcNAc] \times [CTP]}{(K_3^{UDP-GlcNAc} + [UDP-GlcNAc]) \times (K_3^{CTP} + [CTP])} - K_3^{Diss}[CMP-Sial]$$

$$\frac{d[UTP]}{dt} = \frac{r_4^{max} \times [Glc] \times [Gln]}{(K_4^{Glc} + [Glc]) \times (K_4^{Gln} + [Gln])} - \frac{r_1^{max} \times [Glc] \times [UTP]}{(K_1^{Glc} + [Glc]) \times (K_1^{UTP} + [UTP])}$$
$$- \frac{r_2^{max} \times [Glc] \times [UTP]}{(K_2^{Glc} + [Glc]) \times (K_2^{UTP} + [UTP])} - \frac{r_5^{max} \times [UTP]}{(K_5^{UTP} + [UTP])} - K_4^{Diss}[UTP]$$

$$\frac{d[CTP]}{dt} = \frac{r_5^{max} \times [UTP]}{(K_5^{UTP} + [UTP])} - \frac{r_3^{max} \times [UDP-GlcNAc] \times [CTP]}{(K_3^{UDP-GlcNAc} + [UDP-GlcNAc]) \times (K_3^{CTP} + [CTP])} - K_5^{Diss}[CTP]$$

$$\frac{d[GTP]}{dt} = \frac{r_6^{max} \times [Glc] \times [Gln]}{(K_6^{Glc} + [Glc]) \times (K_6^{Gln} + [Gln])} - \frac{r_7^{max} \times [Glc] \times [GTP]}{(K_7^{Glc} + [Glc]) \times (K_7^{GTP} + [GTP])}$$
$$+ \frac{r_8^{max} \times [Glc] \times [GTP]}{(K_8^{Glc} + [Glc]) \times (K_8^{GTP} + [GTP])} - K_6^{Diss}[GTP]$$

$$\frac{d[GDP-Fuc]}{dt} = \frac{r_7^{max} \times [Glc] \times [GTP]}{(K_7^{Glc} + [Glc]) \times (K_7^{GTP} + [GTP])} - K_7^{Diss}[GDP-Fuc]$$

$$\frac{d[Man]}{dt} = \frac{r_8^{max} \times [Glc] \times [GTP]}{(K_8^{Glc} + [Glc]) \times (K_8^{GTP} + [GTP])} - K_8^{Diss}[Man]$$

Bibliography

Abbas SA, Gaspar G, Sharma VK, Patapoff TW, Kalonia DS. 2013. Application of second-derivative fluorescence spectroscopy to monitor subtle changes in a monoclonal antibody structure. Journal of Pharmaceutical Sciences 102(1):52-61.

Abu-Absi NR, Kenty BM, Cuellar ME, Borys MC, Sakhamuri S, Strachan DJ, Hausladen MC, Li ZJ. 2011. Real Time Monitoring of Multiple Parameters in Mammalian Cell Culture Bioreactors Using an In-Line Raman Spectroscopy Probe. Biotechnology and Bioengineering 108(5):1215-1221.

Aghamohseni H, Ohadi K, Spearman M, Krahn N, Moo-Young M, Scharer JM, Butler M, Budman HM. 2014. Effects of nutrient levels and average culture pH on the glycosylation pattern of camelid-humanized monoclonal antibody. Journal of Biotechnology 186:98-109.

Ahn WS, Antoniewicz MR. 2011. Metabolic flux analysis of CHO cells at growth and non-growth phases using isotopic tracers and mass spectrometry. Metabolic Engineering 13(5):598-609.

Arosio P, Rima S, Morbidelli M. 2013. Aggregation Mechanism of an IgG2 and two IgG1 Monoclonal Antibodies at low pH: From Oligomers to Larger Aggregates. Pharmaceutical Research 30(3):641-654.

Brockhausen I. 2006. Glycobiology protocols. Totowa, N.J.: Humana ; Arkholme : Quantum [distributor].

Burnham KP, Anderson DR. 2002. Model selection and multi-model inference : a practical information-theoretic approach. New York ; London: Springer.

Butler M. 2005. Animal cell cultures: recent achievements and perspectives in the production of biopharmaceuticals. Applied Microbiology and Biotechnology 68(3):283-291.

Butler M, Meneses-Acosta A. 2012. Recent advances in technology supporting biopharmaceutical production from mammalian cells. Applied Microbiology and Biotechnology 96(4):885-894.

Calamai M, Kumita JR, Mifsud J, Parrini C, Ramazzotti M, Ramponi G, Taddei N, Chiti F, Dobson CM. 2006. Nature and significance of the interactions between amyloid fibrils and biological polyelectrolytes. Biochemistry 45(42):12806-12815.

Callis PR, Liu TQ. 2004. Quantitative prediction of fluorescence quantum yields for tryptophan in proteins. Journal of Physical Chemistry B 108(14):4248-4259.

Casado-Terrones S, Fernández-Sánchez JF, Segura-Carretero A, Fernández-Gutiérrez A. 2007. Simple luminescence detectors using a light-emitting diode or a Xe lamp, optical fiber and charge-coupled device, or photomultiplier for determining proteins in capillary electrophoresis: A critical comparison. Analytical Biochemistry 365(1):82-90.

Christensen J, Norgaard L, Bro R, Engelsen SB. 2006. Multivariate autofluorescence of intact food systems. Chemical Reviews 106(6):1979-1994.

de Assis AJ, Maciel R. 2000. Soft sensors development for on-line bioreactor state estimation. Computers & Chemical Engineering 24(2-7):1099-1103.

Dejong S. 1993. Simpls - an Alternative Approach to Partial Least-Squares Regression. Chemometrics and Intelligent Laboratory Systems 18(3):251-263.

Del Val IJ, Kontoravdi C, Nagy JM. 2010. Towards the Implementation of Quality by Design to the Production of Therapeutic Monoclonal Antibodies with Desired Glycosylation Patterns. Biotechnology Progress 26(6):1505-1527.

Del Val IJ, Nagy JM, Kontoravdi C. 2011. A dynamic mathematical model for monoclonal antibody N-linked glycosylation and nucleotide sugar donor transport within a maturing Golgi apparatus. Biotechnology Progress 27(6):1730-1743.

den Engelsman J, Garidel P, Smulders R, Koll H, Smith B, Bassarab S, Seidl A, Hainzl O, Jiskoot W. 2011. Strategies for the Assessment of Protein Aggregates in Pharmaceutical Biotech Product Development. Pharmaceutical Research 28(4):920-933.

Deshpande SS. 2001. Principles and applications of luminescence spectroscopy. Critical Reviews in Food Science and Nutrition 41(3):155-224.

Dewasme L, Goffaux G, Hantson AL, Vande Wouwer A. 2013. Experimental validation of an Extended Kalman Filter estimating acetate concentration in E. coli cultures. Journal of Process Control 23(2):148-157.

Dochain D. 2003. State and parameter estimation in chemical and biochemical processes: a tutorial. Journal of Process Control 13(8):801-818.

Durocher Y, Butler M. 2009. Expression systems for therapeutic glycoprotein production. Current Opinion in Biotechnology 20(6):700-707.

Elshereef R, Budman H, Moresoli C, Legge RL. 2006. Fluorescence spectroscopy as a tool for monitoring solubility and aggregation behavior of beta-lactoglobulin after heat treatment. Biotechnology and Bioengineering 95(5):863-874.

Elshereef R, Budman H, Moresoli C, Legge RL. 2008. Fluorescence-based soft-sensor for monitoring beta-lactoglobulin and alpha-lactalbumin solubility during thermal aggregation. Biotechnology and Bioengineering 99(3):567-577.

Eriksson L. JE, Kettaneh-Wold N., Wold S. 2006. Multi- and Megavariate Data Analysis. Umea, Sweden: Umetrics Academy.

Geladi P, Kowalski BR. 1986. Partial Least-Squares Regression - a Tutorial. Analytica Chimica Acta 185:1-17.

Gernaey KV, Cervera-Padrell AE, Woodley JM. 2012. A perspective on PSE in pharmaceutical process development and innovation. Computers & Chemical Engineering 42:15-29.

Haack MB, Eliasson A, Olsson L. 2004. On-line cell mass monitoring of Saccharomyces cerevisiae cultivations by multi-wavelength fluorescence. Journal of Biotechnology 114(1-2):199-208.

Hagedorn A, Legge RL, Budman H. 2003. Evaluation of spectrofluorometry as a tool for estimation in fed-batch fermentations. Biotechnology and Bioengineering 83(1):104-111.

Harthun S, Matischak K, Friedl P. 1998. Simultaneous prediction of human antithrombin III and main metabolites in animal cell culture processes by near-infrared spectroscopy. Biotechnology Techniques 12(5):393-397.

Hawe A, Friess W, Sutter M, Jiskoot W. 2008. Online fluorescent dye detection method for the characterization of immunoglobulin G aggregation by size exclusion chromatography and asymmetrical flow field flow fractionation. Analytical Biochemistry 378(2):115-122.

He F, Phan DH, Hogan S, Bailey R, Becker GW, Narhi LO, Razinkov VI. 2010. Detection of IgG Aggregation by a High Throughput Method Based on Extrinsic Fluorescence. Journal of Pharmaceutical Sciences 99(6):2598-2608.

Henneke D, Hagedorn A, Budman HM, Legge RL. 2005. Application of spectrofluorometry to the prediction of PHB concentrations in a fed-batch process. Bioprocess and Biosystems Engineering 27(6):359-364.

Herbert S, Riou NM, Devaux MF, Riaublanc A, Bouchet B, Gallant DJ, Dufour E. 2000. Monitoring the identity and the structure of soft cheeses by fluorescence spectroscopy. Lait 80(6):621-634.

Herman B, Lakowicz, J.R., Murphy, D.B., Spring, K.R., Davidson, M.W. 2003. Fluorescence Microscopy - Basic Concepts in Fluorescence. cited on April 4, 2008 from http://micro.magnet.fsu.edu/primer/techniques/fluorescence/fluorescenceintro.html.

Hof M, Hutterer R, Fidler V. 2005. Fluorescence spectroscopy in biology : advanced methods and their applications to membranes, proteins, DNA, and cells. Berlin: Springer.

Hossler P, Goh LT, Lee MM, Hu WS. 2006. GlycoVis: Visualizing glycan distribution in the protein N-glycosylation pathway in mammalian cells. Biotechnology and Bioengineering 95(5):946-960.

Hossler P, Khattak SF, Li ZJ. 2009. Optimal and consistent protein glycosylation in mammalian cell culture. Glycobiology 19(9):936-949.

Hossler P, Mulukutla BC, Hu WS. 2007. Systems Analysis of N-Glycan Processing in Mammalian Cells. Plos One 2(8).

Jain G, Jayaraman G, Kokpinar O, Rinas U, Hitzmann B. 2011. On-line monitoring of recombinant bacterial cultures using multi-wavelength fluorescence spectroscopy. Biochemical Engineering Journal 58-59:133-139.

James S, Legge R, Budman H. 2002. Comparative study of black-box and hybrid estimation methods in fed-batch fermentation. Journal of Process Control 12(1):113-121.

Jang DJ, Barford JP. 2000. An unstructured kinetic model of macromolecular metabolism in batch and fed-batch cultures of hybridoma cells producing monoclonal antibody. Biochemical Engineering Journal 4(2):153-168.

Jitjareonchai JJ, Reilly PM, Duever TA, Chambers DB. 2006. Parameter estimation in the Error-in-Variables models using the Gibbs Sampler. Canadian Journal of Chemical Engineering 84(1):125-138.

Kadlec P, Gabrys B. 2009. Soft Sensor Based on Adaptive Local Learning. Advances in Neuro-Information Processing, Pt I 5506:1172-1179.

Kadlec P, Gabrys B, Strandt S. 2009. Data-driven Soft Sensors in the process industry. Computers & Chemical Engineering 33(4):795-814.

Kalman RE. 1960. A New Approach to Linear Filtering and Prediction Problems. Journal of Fluids Engineering 82(1):35-45.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28(1):27-30.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research 40(D1):D109-D114.

Kara S, Mueller JJ, Liese A. 2011. Online analysis methods for monitoring of bioprocesses. Chimica Oggi-Chemistry Today 29(2):38-41.

Kong J, Yu S. 2007. Fourier transform infrared spectroscopic analysis of protein secondary structures. Acta Biochimica Et Biophysica Sinica 39(8):549-559.

Kozub DJ, Macgregor JF. 1992. State Estimation for Semibatch Polymerization Reactors. Chemical Engineering Science 47(5):1047-1062.

Krambeck FJ, Betenbaugh MJ. 2005. A mathematical model of N-linked glycosylation. Biotechnology and Bioengineering 92(6):711-728.

Krishnan A, Williams LJ, McIntosh AR, Abdi H. 2011. Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. Neuroimage 56(2):455-475.

Kumar V, Sharma VK, Kalonia DS. 2005. Second derivative tryptophan fluorescence spectroscopy as a tool to characterize partially unfolded intermediates of proteins. International Journal of Pharmaceutics 294(1-2):193-199.

Lakowicz JR. 1999. Principles of fluorescence spectroscopy. New York ; London: Kluwer Academic/Plenum.

Li H, Rahimi F, Sinha S, Maiti P, Bitan G, Murakami K. 2006. Amyloids and Protein Aggregation—Analytical Methods. Encyclopedia of Analytical Chemistry: John Wiley & Sons, Ltd.

Li JK, Humphrey AE. 1991. Use of Fluorometry for Monitoring and Control of a Bioreactor. Biotechnology and Bioengineering 37(11):1043-1049.

Lindemann C, Marose S, Nielsen HO, Scheper T. 1998. 2-dimensional fluorescence spectroscopy for on-line bioprocess monitoring. Sensors and Actuators B-Chemical 51(1-3):273-277.

Liu B, Spearman M, Doering J, Lattova E, Perreault H, Butler M. 2014. The availability of glucose to CHO cells affects the intracellular lipid-linked oligosaccharide distribution, site occupancy and the N-glycosylation profile of a monoclonal antibody. Journal of Biotechnology 170:17-27.

Lourenco ND, Lopes JA, Almeida CF, Sarraguca MC, Pinheiro HM. 2012. Bioreactor monitoring with spectroscopy and chemometrics: a review. Analytical and Bioanalytical Chemistry 404(4):1211-1237.

Luttmann R, Bracewell DG, Cornelissen G, Gernaey KV, Glassey J, Hass VC, Kaiser C, Preusse C, Striedner G, Mandenius C-F. 2012. Soft sensors in bioprocessing: A status report and recommendations. Biotechnology Journal 7(8):1040-1048.

Mahler HC, Friess W, Grauschopf U, Kiese S. 2009. Protein Aggregation: Pathways, Induction Factors and Analysis. Journal of Pharmaceutical Sciences 98(9):2909-2934.

Majid FAA, Butler M, Al-Rubeai M. 2007. Glycosylation of an immunoglobulin produced from a murine hybridoma cell line: The effect of culture mode and the anti-apoptotic gene, bcl-2. Biotechnology and Bioengineering 97(1):156-169.

Mandenius C-F, Gustavsson R. 2014. Mini-review: soft sensors as means for PAT in the manufacture of bio-therapeutics. Journal of Chemical Technology & Biotechnology:n/a-n/a.

Marose S, Lindemann C, Scheper T. 1998. Two-dimensional fluorescence spectroscopy: A new tool for on-line bioprocess monitoring. Biotechnology Progress 14(1):63-74.

Montgomery DC. 2009. Design and analysis of experiments. Hoboken: John Wiley & Sons.

Naderi S, Meshram M, Wei C, McConkey B, Ingalls B, Budman H, Scharer J. 2011. Development of a Mathematical Model for Evaluating the Dynamics of Normal and Apoptotic Chinese Hamster Ovary Cells. Biotechnology Progress 27(5):1197-1205.

Nolan RP, Lee K. 2011. Dynamic model of CHO cell metabolism. Metabolic Engineering 13(1):108-124.

Nordon A, Littlejohn D, Dann AS, Jeffkins PA, Richardson MD, Stimpson SL. 2008. In situ monitoring of the seed stage of a fermentation process using non-invasive NIR spectrometry. Analyst 133(5):660-666.

Ohadi K, Aghamohseni H, Legge RL, Budman HM. 2014a. Fluorescence-Based Soft Sensor for at Situ Monitoring of Chinese Hamster Ovary Cell Cultures. Biotechnology and Bioengineering 111(8):1577-1586.

Ohadi K, Legge RL, Budman HM. 2014b. Development of a soft-sensor based on multi-wavelength fluorescence spectroscopy and a dynamic metabolic model for monitoring mammalian cell cultures. Biotechnology and Bioengineering:n/a-n/a.

Ohadi SK, Aghamohseni H, Gädke J, Moo-Young M, Legge RL, Scharer J, Budman HM. 2013. Novel Dynamic Model to Predict the Glycosylation Pattern of Monoclonal Antibodies from Extracellular Cell Culture Conditions. 12th IFAC Symposium on Computer Applications in Biotechnology. IIT Bombay, Mumbai, India: ELSEVIER.

Pais DAM, Carrondo MJT, Alves PM, Teixeira AP. 2014. Towards real-time monitoring of therapeutic protein quality in mammalian cell processes. Current Opinion in Biotechnology 30(0):161-167.

Peiris BRH, Budman H, Moresoli C, Legge RL. 2009. Acquiring reproducible fluorescence spectra of dissolved organic matter at very low concentrations. Water Science and Technology 60(6):1385-1392.

Peiris RH, Budman H, Moresoli C, Legge RL. 2011a. Development of a species specific fouling index using principal component analysis of fluorescence excitation-emission matrices for the

ultrafiltration of natural water and drinking water production. Journal of Membrane Science 378(1-2):257-264.

Peiris RH, Budman H, Moresoli C, Legge RL. 2011b. Identification of humic acid-like and fulvic acid-like natural organic matter in river water using fluorescence spectroscopy. Water Science and Technology 63(10):2427-2433.

Peiris RH, Halle C, Budman H, Moresoli C, Peldszus S, Huck PM, Legge RL. 2010. Idendifying fouling events in a membrane-based drinking water treatment process using principal component analysis of fluorescence excitation-emission matrices. Water Research 44(1):185-194.

Pohlscheidt M, Charaniya S, Bork C, Jenzsch M, Noetzel TL, Luebbert A, Flickinger MC. 2009. Bioprocess and Fermentation Monitoring. Encyclopedia of Industrial Biotechnology: John Wiley & Sons, Inc.

Poole RA, Hawe A, Jiskoot W, Braeckmans K. 2012. Fluorescence Spectroscopy to Characterize Protein Aggregates and Particles. Analysis of Aggregates and Particles in Protein Pharmaceuticals: John Wiley & Sons, Inc. p 201-226.

Printz M, Friess W. 2012. Simultaneous detection and analysis of protein aggregation and protein unfolding by size exclusion chromatography with post column addition of the fluorescent dye BisANS. Journal of Pharmaceutical Sciences 101(2):826-837.

Rathore AS, Bhambure R, Ghare V. 2010. Process analytical technology (PAT) for biopharmaceutical products. Analytical and Bioanalytical Chemistry 398(1):137-154.

Rathore AS, Bhushan N, Hadpe S. 2011. Chemometrics Applications in Biotech Processes: A Review. Biotechnology Progress 27(2):307-315.

Read EK, Park JT, Shah RB, Riley BS, Brorson KA, Rathore AS. 2010. Process Analytical Technology (PAT) for Biopharmaceutical Products: Part I. Concepts and Applications. Biotechnology and Bioengineering 105(2):276-284.

Ryan PW, Li BY, Shanahan M, Leister KJ, Ryder AG. 2010. Prediction of Cell Culture Media Performance Using Fluorescence Spectroscopy. Analytical Chemistry 82(4):1311-1317.

Sadavarte RH, Ghosh R. 2014. A Thermal-Cycling Method for Disaggregating Monoclonal Antibody Oligomers. Journal of Pharmaceutical Sciences 103(3):870-878.

Sahin E, Grillo AO, Perkins MD, Roberts CJ. 2010. Comparative Effects of pH and Ionic Strength on Protein-Protein Interactions, Unfolding, and Aggregation for IgG1 Antibodies. Journal of Pharmaceutical Sciences 99(12):4830-4848.

Sellick CA, Hansen R, Jarvis RM, Maqsood AR, Stephens GM, Dickson AJ, Goodacre R. 2010. Rapid Monitoring of Recombinant Antibody Production by Mammalian Cell Cultures Using Fourier Transform Infrared Spectroscopy and Chemometrics. Biotechnology and Bioengineering 106(3):432-442.

Shioya S, Shimizu K, Yoshida T. 1999. Knowledge-based design and operation of bioprocess systems. Journal of Bioscience and Bioengineering 87(3):261-266.

Skibsted E, Lindemann C, Roca C, Olsson L. 2001. On-line bioprocess monitoring with a multi-wavelength fluorescence sensor using multivariate calibration. Journal of Biotechnology 88(1):47-57.

Stanley P. 2011. Golgi Glycosylation. Cold Spring Harbor Perspectives in Biology 3(4).

Surewicz WK, Mantsch HH, Chapman D. 1993. Determination of Protein Secondary Structure by Fourier-Transform Infrared-Spectroscopy - a Critical-Assessment. Biochemistry 32(2):389-394.

Teixeira AP, Duarte TM, Oliveira R, Carrondo MJT, Alves PM. 2011. High-throughput analysis of animal cell cultures using two-dimensional fluorometry. Journal of Biotechnology 151(3):255-260.

Teixeira AP, Oliveira R, Alves PM, Carrondo MJT. 2009a. Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. Biotechnology Advances 27(6):726-732.

Teixeira AP, Portugal CAM, Carinhas N, Dias JML, Crespo JP, Alves PM, Carrondo MJT, Oliveira R. 2009b. In Situ 2D Fluorometry and Chemometric Monitoring of Mammalian Cell Cultures. Biotechnology and Bioengineering 102(4):1098-1106.

Trygg J, Wold S. 2002. Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics 16(3):119-128.

Umana P, Bailey JE. 1997. A mathematical model of N-linked glycoform biosynthesis. Biotechnology and Bioengineering 55(6):890-908.

van de Weert M, Jørgensen L. 2012. Infrared Spectroscopy to Characterize Protein Aggregates. Analysis of Aggregates and Particles in Protein Pharmaceuticals: John Wiley & Sons, Inc. p 227-248.

Wang JL, Zhao LQ, Yu T. 2010. On-line Estimation in Fed-batch Fermentation Process Using State Space Model and Unscented Kalman Filter. Chinese Journal of Chemical Engineering 18(2):258-264.

Wang W. 2005. Protein aggregation and its inhibition in biopharmaceutics. International Journal of Pharmaceutics 289(1-2):1-30.

Wise BM, Gallagher, N. B., & Windig, W. . 2006. Chemometrics Tutorial for PLS _ Toolbox and Solo. Eigenvector Research. Eigenvector Research Corporation.

Wold S, Ruhe A, Wold H, Dunn WJ. 1984. The Collinearity Problem in Linear-Regression - the Partial Least-Squares (Pls) Approach to Generalized Inverses. Siam Journal on Scientific and Statistical Computing 5(3):735-743.

Wold S, Sjostrom M, Eriksson L. 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58(2):109-130.

Wurm FM. 2004. Production of recombinant protein therapeutics in cultivated mammalian cells. Nature Biotechnology 22(11):1393-1398.

Wurm FM, Hacker D. 2011. First CHO genome. Nature Biotechnology 29(8):718-720.

Zamorano F, Wouwer AV, Jungers RM, Bastin G. 2013. Dynamic metabolic models of CHO cell cultures through minimal sets of elementary flux modes. Journal of Biotechnology 164(3):409-422.

Zepp RG, Sheldon WM, Moran MA. 2004. Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating Rayleigh and Raman scattering peaks in excitation-emission matrices. Marine Chemistry 89(1-4):15-36.

Zustiak MP, Pollack JK, Marten MR, Betenbaugh MJ. 2008. Feast or famine: autophagy control and engineering in eukaryotic cell culture. Current Opinion in Biotechnology 19(5):518-526.