Data-Driven Modeling and Model Predictive Control of Monoclonal Antibody Process

Data-Driven Modeling and Model Predictive Control of Monoclonal Antibody Process

by

Samardeepsingh Sarna, B.Tech.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfillment of the Requirements for
the Degree Master of Applied Science

Master of Applied Science (2022)	McMaster University
(Chemical Engineering)	Hamilton, Ontario, Canada

TITLE:	Data-Driven Modeling and Model Predictive Control of Monoclonal Antibody Process
AUTHOR:	Samardeepsingh Sarna, B.Tech.
	(McMaster University, Hamilton, ON)
SUPERVISOR:	Dr. Prashant Mhaskar
NUMBER OF PAGES:	xii, 79

LAY ABSTRACT

Biopharma processes involving live cells are utilised to produce several critical products such as monoclonal antibodies which are state-of-the-art cancer therapeutics. Improving productivity for these would require advanced process control methods which in turn would require good models. This thesis focuses on introducing a method to build control relevant data driven models by using input perturbation and data generation suitable for live cell systems. The data driven model is utilised in a model predictive control scheme which meets control objectives while respecting biological constraints with better performance than industrial standards. Further improvement is made by incorporating first principles knowledge into the data driven model through a novel implementation of constrained subspace identification. The approaches are showcased on an advanced simulator test bed created by Sartorius.

ABSTRACT

This thesis focuses on data-driven modeling and model predictive control for a monoclonal antibody process. The process uses live cell cultures such as Chinese Hamster Ovary (CHO) cells and thus needs special consideration. With the trend in the industry to move towards perfusion processes which are continuous allowing better productivity, advanced process control would play a vital role. Model Predictive Control (MPC) requires a suitable model to optimally control the process. First principles models for such live cell processes are complex and unsuitable for direct use in MPC. In this work, we focus on data-driven modeling given its suitability for use in control. The data-driven model is eventually also incorporated with some first principles knowledge for better performance and robustness.

Data-driven modelling requires some input perturbation and data generation methods such as Pseudo-Random Binary Sequence (PRBS) inputs. By themselves, these methods are unsuitable for live cells as they can shock the system. To account for this, a suitable, intensified design of experiments (DOE) approach is used to perturb the data frequently enough to build a model of reasonable accuracy without significantly impacting live cells. This method is general enough to be used to identify appropriate input perturbation and data generation for many bioprocesses, and for our particular process of study it identified an input perturbation frequency of once per three days. The data-driven model is used in a model predictive control (MPC) scheme which respects biological constraints and can meet desired objectives. Further improvement in model robustness are made possible by incorporating first principles knowledge into the data-driven model. The methods are demonstrated on an advanced simulator developed by Sartorius with significant improvement over current industry standard.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Dr. Prashant Mhaskar. Having had a great experience interning as a Mitacs Globalink scholar under his supervision in third year undergraduate summers, joining for masters at the same place was a given. He has been the absolute perfect supervisor one could ever hope for; helpful, knowledgeable, kind, friendly, supportive and a host of other positive adjectives would fit perfectly well and in addition he is a fantastic instructor, I had the pleasure of taking his course as well as TAing and loved both aspects. I had some difficult times and multiple unfortunate health issues during my masters and he was always there with his unwavering support and I could never thank him enough for that. He is also a great friend, has a great sense of humour and group events such as riddles with hidden puns and other social events were the highlight of my time as a graduate student. He is also exceptionally skilled at every sport I have played with him, be it physical sports or chess and it's always a pleasure to play a game with him even though its invariably a lost game for me. I would also like to thank Dr. Nikesh Patel, who has been a great friend and mentor and helped me become a better writer and researcher through his valuable insights. I want to thank Dr. Brandon Corbett for helping me really appreciate the field of big data through his course, which is one of the most exciting courses I ever took. In addition I would also thank him and Chris McCready for imparting biological knowledge and the opportunity to work on cutting edge biopharmaceutical space. I would like to thank my committee members, Dr. Christopher Swartz and Dr. Kamil Khan. I had the pleasure of taking Dr. Swartz's optimization course which was a great learning experience and his philosophical insights to not miss the forest for the trees. Although I did not have the chance to have professional interactions with Dr. Khan, every social interaction with him had been a great discussion, be it on anime or food. In addition, I would like to thank my great friends at McMaster for being absolute gems of people, making grad school a fun experience and bearing with my unfunny jokes. Finally, I would like

to thank my family, especially my mother, for all their support and values; I would not be here without them.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Advancements in technology, data availability and computational abilities have led to strong adaption of advanced process control and data analytics in traditional chemical engineering. Data driven modeling and machine learning methods have also seen a surge in popularity and relevance in the broader process industry. The adaptation in biopharmaceutical industry has however been much slower. The industry standard is very often batch recipe and operator led modes or Proportional-Integral (PI) control for process control applications. Further, biopharmaceutical processes have traditionally been batch or fed-batch. Recent advancements in capabilities and with rising demands of live cell therapeutics has generated strong interest in moving towards continuous or perfusion processing which would allow higher productivity. Continuous processing allows removal of toxic byproducts and waste allowing longer process runs and more time for cells to be in production phase and hence a potential for significantly more product over time and lower manufacturing footprint.

Advanced process control approaches such as Model Predictive Control (MPC) are well placed to meet multiple objectives for such a perfusion process such as maintaining Viable Cell Density (VCD) at a setpoint, maximising the volume specific concentration or total mass of the protein product and meeting various input constraints and suitable conditions in the bioreactor. Such control strategies are dependent on a good model for successful implementation. Models are broadly classified into first principles or physics based models (white box modeling) and data-driven models (black box modeling). A combination of features of both has also become an active area of research in recent times and such methods are known as hybrid models (grey box modeling). Biopharmaceutical processes involving live cells have a complex interplay of reactor conditions, inputs and outputs along with metabolites fed and consumed by cells, products and by-products produced and the impact of all factors of differ-

ent phases of growth, death and lysing of cells. As such, first principles modeling for these systems is especially convoluted and parameter estimation is challenging. More importantly, such complicated models would not be well suited for implementation in advanced model predictive control formulations. Data-driven modeling using methods such as neural networks need a lot of data to build a model with predictive capabilities and data is often limited due to high cost of running every single process run. This work focuses on data-driven modeling and model predictive control for biopharmaceutical processes with demonstration on an antibody process used by Sartorius. The method of choice is subspace identification for its strengths in modeling with reasonable accuracy even with limited data and the subspace state space model being relevant for control. Hybrid modeling is also a subject of this work with some first principles information incorporated in the data driven model for improved performance and robustness.

1.2 Outline of the thesis

The first manuscript of the thesis (Chapter 2) focuses on building a data-driven model for the perfusion bioreactor for monoclonal antibodies based on a proprietary advanced simulator by Sartorius which operates on a combination of true process conditions from their plant, first principles equations and proprietary latent variable modeling to replicate real-life bioreactor conditions with high fidelity. Given the significant costs as well as longer duration for perfusion process (over thirty days) to run every single run of the bioreactor process the simulator serves as the test bed to develop methods relevant for deployment to the actual plant. The key challenge for building the data-driven method is not just choosing a suitable method based on limited data but also the mode of data generation or input perturbation. Biopharmaceutical processes traditionally have very minimal changes in input condition during the process run to

avoid shocks to the system whereas traditional system identification would require very frequent perturbations such as Pseudo-Random Binary Sequence (PRBS) inputs and the manuscript is based on an intensified design of experiments approach to build a data driven model based on appropriate input perturbation frequency which would respect biological constraints as well as measurement frequency feasibility.

The second manuscript (Chapter 3) is based on model predictive control of the perfusion bioprocess based on the data-driven model developed in the first manuscript. The design of the model predictive controller is done to allow flexibility in meeting multiple objectives as required, including maximising the volume specific concentration of product, maximising the total mass of product produced and tracking the viable cell density. The inputs are constrained to respect biological requirements. The model predictive controller already showing improvement over the industry standard PI is further improved through incorporation of first principles knowledge in a hybrid or process aware modification of subspace identification which incorporates constraints on the sign of the gain matrix, allowing the causal input output relation for the biological process to be honored. The benefit of the approach towards robustness is also demonstrated by building the model with biological knowledge constraints but using input-output data from a different cell line and still be able to provide desired performance.

The fourth chapter discusses recent work done with summer students in the Mhaskar group which involves modern machine learning methods for improved process understanding. Bayesian inference using nested sampling for biological process parameters estimation and physics informed neural networks (PINNs) for hybrid modeling are the primary methods used. The last chapter summarizes the approaches, makes concluding remarks and discusses future work and ideas to explore.

Chapter 2

Determining Appropriate Input Excitation for Model Identification of a Continuous Bioprocess

This chapter introduces an approach to allow data driven modeling for live-cell continuous processes such as monoclonal antibody production. The method balances a high input perturbation frequency needed by traditional system identification and minimal perturbations to avoid shocking bioprocess. Based on the process dynamics and measurement feasibility, an intensified design of experiments approach is devised and subspace identification is used to identify the data-driven model.

This work was completed in collaboration with Sartorius who provided insights and discussions on the bioprocess as well as the advanced proprietary simulator. The manuscript has been submitted to Digital Chemical Engineering journal.

Sarna, S., Patel, N., Corbett, B., McCready, C., & Mhaskar, P. (2022).
Determining Appropriate Input Excitation for Model Identification of a Continuous Bioprocess, *Digital Chemical Engineering*, (Submitted)

2.1 Abstract

This manuscript addresses the problem of determining input excitation for data driven model identification appropriate for cell culture bioprocesses in general, and for an industrial bioreactor used for the production of monoclonal antibodies, in particular. The design space is set up to give us the operating parameters for the key objective of demonstrating the feasibility of using far more perturbations than typically done in bio process identification, although significantly less than other applications, to yield data rich enough for the purpose of data driven modeling (and subsequently, control). A proprietary mechanistic model developed by Sartorius for their Cellca cell line is first introduced to serve as a test bed, based on AMBR 250 ® (Sartorius registered trademark for integrated high throughput bioreactor systems). Subsequently, this test bed is used to address the question of determining the frequency of input perturbation sufficient to identify a data driven dynamic model. To this end, the test bed is used to generate data at various frequencies and a linear time invariant model identified. The predictive capability of the identified model is used to ascertain the frequency of changes in data generation such that the changes are acceptable from a biological standpoint, and yet generate sufficiently rich data. In particular, a frequency of perturbations at once every three days is found to balance these trade-offs for the monoclonal antibody process under consideration. The results from the manuscript are meaningful both from a specific results standpoint (as illustrated by subsequent adoption by Sartorius), but also by providing a mechanism to ascertain such information for other bioprocesses.

2.2 Introduction

Continuous bioprocessing is increasingly replacing more traditional batch and fed-batch modes of production in bio-pharmaceutical manufacturing due to several advantages such as significantly higher productivity (over 10 times), reduction in manufacturing footprint and cost and more opportunities for product quality and consistency control, inclusion of industry 4.0, digital twins and other advanced control strategies. [32] The productivity and economics of these continuous processes are affected by several factors including hydrodynamics and transport phenomena.[16] A majority of the literature on control and process systems engineering related work on bio-processes deals with batch or fed-batch operation. [22, 6, 25] This work deals with a perfusion process which differs from fed-batch and batch in that there is a continuous addition of fresh media and removal of spent media through bleed and harvest streams. The final product (volume specific monoclonal antibodies) is influenced by multiple factors like cell growth rate, feed rate and feed concentration and thus, these factors need to be accounted for in a model based control design.

The dependence of the process on the manipulated inputs and operating conditions is complex. Glucose being the key substrate is paramount, however, excess glucose is detrimental due to overflow metabolism, which is undesirable and leads to the production of excess lactate and ammonia. Other substrates such as glutamine play a similar role as glucose especially for promoting cell growth during periods of fast growth. [15] These metabolites alone are not the only factors affecting cell growth. Critical process parameters (CPPs) such as temperature and pH play a key role as well. [4, 31] [15] The production of a specific product such as a protein by these cells is heavily affected by the environment, such as the pH and glucose levels [3, 23, 25] and a diversity of variables affect the system dynamics with many of these having contradictory effects in different ranges.

One approach to quantify process understanding is to use first principles models. While a challenging problem, parameter estimation methods for first principles models do exists in literature [8, 26, 7, 21, 19, 1] which have been utilized to develop first principles models. [15] Most of the existing results have focused on batch or fed batch systems, and limited results exist for perfusion processes. Among other challenges, the availability of sufficiently rich data to estimate the parameters remains an issue with first principles modeling approaches.

An alternate to first principles models are data driven models, that are often well suited for ease of implementation.[6, 33] One suitable approach is subspace identification, which is a well established system identification method and has several advantages such as having only one decision variable (the order of the system) and its ability to handle large multi-input multi-output (MIMO) problems well. The model complexity of MIMO and the simpler single-input single-output (SISO) systems is similar when using subspace identification. This is in contrast to methods such as Auto-Regressive Moving Average with eXogenous inputs (ARMAX) models which have multiple 'tuning' parameters and hence in comparison to methods such as ARMAX, subspace identification is often easier to implement [11] .

Finally, hybrid semi-parametric models contain aspects of both models and utilize knowledge about the process and make use of the data available as well.[28] The results for hybrid models are encouraging and could outperform first principles and data driven (subspace) methods both individually. [12] Hybrid models have also been implemented in MPC based framework with good results [13] and would be considered in future work. Subspace identification [22] and other data driven and hybrid methods for bioprocesses [6, 33] have been presented in literature with promising results. However, the focus has generally been on modelling and prediction based on a given data set and not necessarily on determining the adequacy of input perturbations for model identification which is crucial for design of experiments to generate training

data for modeling of live-cell bioprocesses like the one under consideration.

From a system identification point of view, it is desirable to have frequent perturbations as system identification would need sufficiently excited, data-rich inputs [20], and is a regular practice in process industry. On the other hand, from an operational point of view for a bioprocess, minimal input perturbations are desired so as to not disturb the cell growth too much. Moreover, in implementing control or collecting data for model identification of bioreactors, one consideration is that the control action or input changes cannot be too drastic or extreme between sampling periods as the living cells are sensitive to minute changes.

In more recent results the notion of intensified design of experiments has been proposed in the context of biological processes, specifically for fed-batch processes.[29, 27, 2] The existing results are designed for fed-batch operation and as such not specifically suited for perfusion operations. More importantly the results in [29, 27, 2] focus on incorporating various consideration in the design, but do not specifically consider the suitability of the resultant data to yield an accurate control relevant model. Further, results on system identification on simulations of bioprocesses involving antibodies have mostly focused on single input-single output variables [9]. In summary, the question of determining an appropriate excitation level of data for bio processes remains unaddressed along with control relevant modeling for multiple input-multiple output variables.

Production of Monoclonal antibodies (mAbs) involving Chinese hamster ovary (CHO) cells has been an important part of the biotherapeutics industry in recent years. [17] Monoclonal antibodies are state of the art therapeutics which provide the most rapid route to clinical proof of concept for disease modulation target activation or inhibition. They are fast becoming popular with $75 Billion global sales revenue annually and over three times faster increase in sales growth than other recombinant therapeutics and biopharmaceutical products. [10] The current manuscript focuses on production

of monoclonal antibodies via a Sartorius Cellca cell line in a continuous bioreactor. For this process, a primary objective is to maximize its volume specific production by manipulating properties and variables in the bioreactor.

Motivated by above, this paper investigates the ability of data driven models to capture the process dynamics for control purposes, particularly focusing on comparing different input perturbation frequencies to enable identifying a sufficiently accurate model with fewest possible input changes. Thus the key contribution of this work is to illustrate a methodology to determine the tradeoff between input excitation that preserves biological stability and obtaining a sufficiently rich data set for model identification. Consistent with dynamics of the bio-process under consideration, the considered input change frequencies (over a nominal baseline) are three times per day, once per day, once per three days and once per five days. The trained models (with these different frequency of input changes during training) are then tested using validation data which has a singular temperature shift at day 10 of the process while keeping the other inputs constant which is an industrial practice adopted by Sartorius. The models are then compared for their predictive ability to determine the best frequency of input change. The rest of the paper is organized as following: Section 2.3 described the bioreactor process as well as the cell culture. The first principles model used as the test-bed is described briefly in section 2.3.2. Section 2.4 explains the reduced order modelling done via subspace identification and the application of the proposed method to the test bed. Concluding remarks are presented in Section 3.6.

2.3 Preliminaries

An overview of the bioreactor process is presented in this section followed by a description of the cell culture.

2.3.1 Bioreactor Process Description

The bioprocess under consideration involves live cells in an enclosed enviornment meaning that the growth and death rates affect the environment in the reactor and consequently the titer (concentration in mg/L of final product). Thus, it is important to ensure that the environment is not disturbed too frequently and for the effect of the perturbations to be realized on the system before the next perturbation is implemented. In a perfusion process, especially, there are mixing and biological phenomenon that have to manifest before the effect of the process change is realized. This limits the frequency of perturbations that are meaningful in a given period of time to yield information rich data. Presenting a methodology to determine this frequency is the objective of the present work.

A simplified schematic of the bioreactor is presented in Figure 2.1. The recycle stream recycles live cells as a live cell retention filter does not allow live cells to leave through the harvest stream.

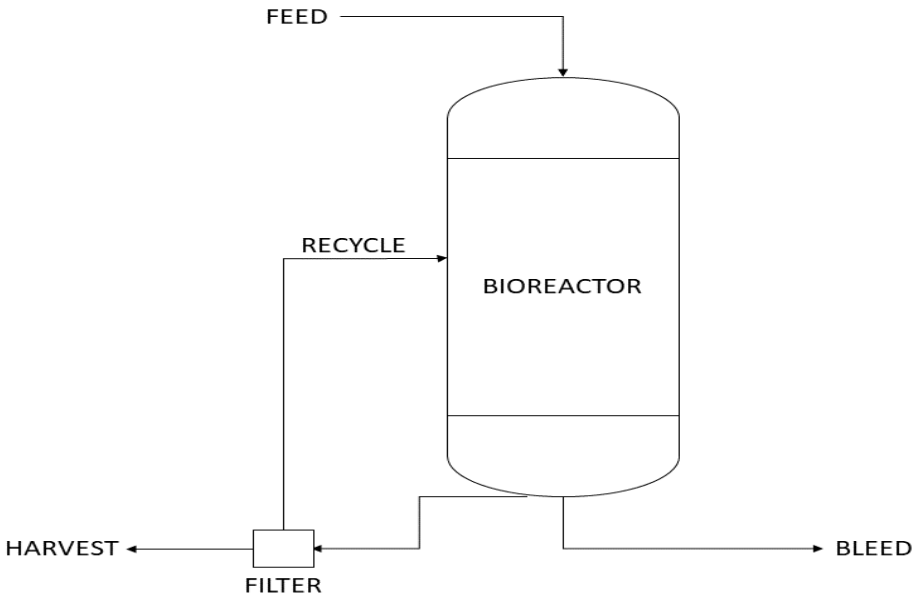


Figure 2.1: Simplified schematic of the bioreactor

An additional consideration is that the process runs in growth phase for 3 days followed by perfusion phase for 30 days hence it takes over a month of time (and significant costs) to generate data. Thus, the approach used in this work is illustrated utilizing a simulation test bed provided by Sartorius.

The nominal values for the temperature and pH are set at $36.8°C$ and 7.1 respectively. The working volume of the reactor is 0.2L and the feed rate is 0.3L/day or 1.5 volumes/day. The growth phase lasts for 3 days during which the system is operated as a fed batch process. This is followed by a perfusion phase for 30 days. For the purpose of subspace identification, the inputs are chosen as temperature (°C), pH, glucose feed concentration (g/L) and feed rate (L). The outputs are chosen as titer (mg/L), viability (%), glucose concentration (g/L) and viable cell density (VCD) (10^6 cells/mL). To reflect the current practice, the VCD is under PI control with a fixed setpoint of 50 $\times 10^6$ cells/mL. Further, the measurements for the metabolites and titer are only taken three times per day. Hence this is highest frequency of changes of inputs that is considered.

2.3.2 Cell Balance Model

The model can be broadly divided into three parts: 1) A first-principles cell balance model (growth/death), 2) A hybrid metabolic evolution model, and 3) A hybrid productivity (titer) model. Essentially the Sartorius test bed represents cells using a cell balance similar to mass balances from traditional chemical engineering. This balance tracks the population of cells as they move through three phases: live cells, dead cells, and lysed cells. Mathematically, the evolution of the live, dead, and lysed cells are tracked using ordinary differential equations as follows:

$$\frac{dx_v}{dt} = \left(u_{eff} - u_d - \frac{F_b}{V}\right) x_v \tag{2.1}$$

$$\frac{dx_d}{dt} = u_d x_v - \left(k_l + \frac{F_b}{V}\right) x_d \tag{2.2}$$

$$\frac{dx_l}{dt} = k_l x_d - \left(\frac{F_h + F_b}{V}\right) x_l \tag{2.3}$$

where x_v is the viable cell density (VCD), x_d is the dead cell density, and x_l is the lysed cell density (concentration of lysed cells). F_b is the bleed rate, F_h is the harvest rate, and V is the volume, u_{eff}, u_d, and k_l are the effective growth, effective death, and lysing rates respectively.

Live cells are formed at the effective growth rate. Live cells can either exit the reactor through the bleed stream or be transformed into dead cells. Note that this implies the existence of a perfect cell retention filter (i.e., no live cells exit the reactor through the harvest stream).

The growth and death rates are based on proprietary dynamics. The titer, which is a function of the final output concentration, evolves based on a hybrid dynamic model. While noting that the test bed developed by Sartorius has been validated against experimental data and captures the key complexities of the bio-process under consideration, it is important to recognize the results of this manuscript are neither dependent on the accuracy of the test bed, nor on the specific bio-process under consideration, and while the specific findings about the best excitation frequency are specific to the present bio-process, the approach is applicable to any other bio-process.

2.4 Determining Appropriate Input Excitation Frequency

This section describes the approach proposed in this manuscript to determine the appropriate frequency of input excitation for building a control-relevant reduced order model. The key idea is that candidate data sets are generated at different input frequencies (while keeping them within biologically acceptable ranges) and the appropriateness of a particular excitation frequency is determined via a metric that captures the prediction capability of the model identified using the candidate data set.

As described in sections 2.2 and 3.3.1, the cell growth and consequently the production of desired product/antibodies depends on the environment inside the reactor which is determined by the substrates and physical conditions. Since the eventual goal is to utilize the proposed approach for a model predictive control strategy, the variables which are relevant from a control perspective, i.e. variables which are reasonable to be considered manipulated inputs, were used as model inputs. Hence, pH, Temperature, Glucose Feed Concentration and Feed Rate were chosen as the inputs. The final product of interest is titer, hence it's presence as an output would be incontrovertible. Further, some of the key parameters which help get a sense of the cell growth inside the reactor are viability and viable cell density and hence those were also chosen as outputs. Finally, glucose is the key nutrient for cell growth and it is desirable to track the glucose levels inside the reactor hence it was also chosen as one of the outputs for the subspace model.

The inputs to the system are perturbed at four different frequencies to obtain four sets of dynamic input-output data and these are used to build dynamic linear time invariant models relevant for control applications using the technique of subspace identification described in the next section 2.4.1. The choice of frequencies to be

considered for input perturbation is based on the dynamics of the process under consideration. For the current system, laboratory measurements are available once every 8 hours or 3 times per day and that forms the upper bound of perturbation frequency considered. Though lower frequencies of perturbation would be preferable as long as a reasonably accurate model is identified, our upper bound would still be a candidate if the model identified is significantly better. The process runs for 30 days after the inoculation phase with a possible extension to 60 days in the future; hence an input perturbation of once per 5 days is considered as the lower bound of perturbation frequency under consideration to allow excitation of the system. Once per day and once per 3 days are considered in the intermediate frequency range between the above chosen higher and lower frequency bounds. These particular frequencies were based on the dynamics of the system under consideration but a similar method of determining possible perturbation frequencies and then building data driven model and validating its accuracy can be applied to other bioprocesses.

The accuracy of the models is compared using a root mean squared error metric to make a balanced choice between high accuracy and low perturbation frequency of input excitation and presented in section 2.4.2 for the test bed.

To generate the data, the input changes are made by random perturbations of a small magnitude around a chosen baseline profile and these are seeded, for replicability. The baseline profile is selected to ensure exploring the input space sufficiently. The perturbations themselves are done in a fashion similar to the pseudo random binary signal (PRBS) methods used for system identification and widely deployed for dynamic systems [14] with suitable modifications for a live cell bioprocess to avoid shocks to the system. These four different models or cases are summarised in the table 2.1 below:

Table 2.1: Summary of the four different data sets using four different frequencies of changes or perturbations made in inputs.

Case	*Description*
Case 0	Inputs are changed/perturbed three times per day
Case 1	Inputs are changed/perturbed once per day
Case 2	Inputs are changed/perturbed once per three days
Case 3	Inputs are changed/perturbed once per five days

One of the key contributions of the present work in the context of biological systems in general, and the Sartorius Bioreactor in particular, is to demonstrate that data sets generated with significant perturbations in the input profiles, and also where the perturbations are simultaneously made in all the inputs are feasible from both a data generation perspective, and from a biological perspective. This is significant, in that it reduces the number of independent experiments that need to be run, and can easily translate into savings of hundreds of thousands of dollars. The results of the present study have been implemented by Sartorius in their runs. In a significant departure from past strategies, experiments were carried out for input variations in the same frequency as suggested. In future work, data collected from the continuous bioreactor will be used to demonstrate the model technique as well as investigate the possibility of using hybrid models [12].

2.4.1 Subspace Model Identification

Methods such as neural networks, particularly recurrent neural networks (RNNs) are also established for use in model identification and control. Promising results were presented in literature such as [30]. RNNs also allow direct multiple step ahead predictions hence their use in MPC is attractive. However the performance of RNNs deteriorates with limited data which needs to be paid special attention to, especially with bioprocesses with limited data due to infrequent measurements of only upto three

times per day, and the exorbitant cost of running an experiment. Given the nature of such processes with limited availability and high costs of data, subspace identification is more well suited given its strength of building a model of reasonable accuracy from row space intersections.

For each data set, subspace identification is used to identify a Linear Time Invariant (LTI) model for the process [20] (adapted for batch processes in previous results[Corbett and Mhaskar]). The identification approach used in this paper identifies an LTI model as follows: Given N measurements (N is the length of the data) of the input $u[k] \in \mathbb{R}^m$ and the output $y[k] \in \mathbb{R}^l$ variables from each batch a model with order n can be identified using the following equations:

$$\begin{aligned} \hat{\mathbf{x}}[k+1] &= \mathbf{A}\hat{\mathbf{x}}[k] + \mathbf{B}\mathbf{u}[k], \\ \mathbf{y}[k] &= \mathbf{C}\hat{\mathbf{x}}[k] + \mathbf{D}\mathbf{u}[k], \end{aligned} \tag{2.4}$$

The objective is to identify the order n, which can be determined by cross validation, and the system matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{l\times n}$, $\mathbf{D} \in \mathbb{R}^{l\times m}$.

Identification of the system matrices is done in two stages, first stage involves identifying a state sequence and the second stage comprises of identifying the system matrices. Using subspace identification, the state sequence can be identified using methods such as SVD before knowing the A,B,C,D system matrices. The system matrices are later identified by least squares regression and any convergence or iterative algorithms are not needed.

Some algebraic manipulation on the state space equations leads to the following equation involving Hankel matrices.

$$\mathbf{Y}_p = \mathbf{\Gamma}_i \mathbf{X} + \mathbf{H}_t \mathbf{U}_p, \tag{2.5}$$

The block Hankel matrices are constructed for the input and output. The number of block rows (i) and columns (j) are chosen sufficiently large, typically i should be greater than or equal to n + 1 for the n identified later on and j >>max(mi,li).

The output and input block Hankel matrices are:

$$Y_p = \begin{bmatrix} y[k] & y[k+1] & \dots & y[k+j-1] \\ y[k+1] & y[k+2] & \dots & y[k+j] \\ y[k+2] & y[k+3] & \dots & y[k+j+1] \\ \vdots & & & \vdots \\ y[k+i-1] & y[k+i] & \dots & y[k+i+j-2] \end{bmatrix}$$

$$U_p = \begin{bmatrix} u[k] & u[k+1] & \dots & u[k+j-1] \\ u[k+1] & u[k+2] & \dots & u[k+j] \\ u[k+2] & u[k+3] & \dots & u[k+j+1] \\ \vdots & & & \vdots \\ u[k+i-1] & u[k+i] & \dots & u[k+i+j-2] \end{bmatrix}$$

Γ_i the extended observability matrix is:

$$\Gamma_i = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{i-1} \end{bmatrix}$$

H_t is a lower triangular block Topelitz matrix consisting of Markov parameters:

$$H_t = \begin{bmatrix} D & 0 & 0 & \dots & 0 \\ CB & D & 0 & \dots & 0 \\ CAB & CB & D & \dots & 0 \\ \vdots & & & & \vdots \\ CA^{i-2}B & CA^{i-3}B & & \dots & D \end{bmatrix}$$

Y_f and U_f are defined similar to Y_p and U_p. The state vector sequence can then be calculated from the intersection of the row spaces of two block Hankel matrices H1 and H2, constructed by concatenating the input-output vectors. Further details are present in [20]. Once the state vector sequence is calculated, the A,B,C,D matrices are determined by least squares regression as follows:

$$\begin{bmatrix} x[k+i+1] & x[k+i+2]\dots & x[k+i+j] \\ y[k+i] & y[k+i+1]\dots & y[k+i+j-1] \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x[k+i] & \dots & x[k+i+j-1] \\ u[k+i] & \dots & u[k+i+j-1] \end{bmatrix}$$

This approach identifies a linear state space model where the states are unmeasured but observable from measured outputs. For the present results, while testing, an initial state estimate was chosen based on the initial value for states observed while training. A Luenberger observer [18] was used starting at the beginning and run for either 30 time-steps or until the error (Euclidean norm) between the predicted output and actual/observed output was below a chosen threshold of 0.2, whichever was earlier.

The Luenberger observer takes the following form:

$$\hat{\mathbf{x}}[k+1] = \mathbf{A}\hat{\mathbf{x}}[k] + \mathbf{B}\mathbf{u}[k] + \mathbf{L}(\mathbf{y}[k] - \hat{\mathbf{y}}[k]) \tag{2.6}$$

where $\mathbf{L}$ is the observer gain and is chosen such that $(\mathbf{A} - \mathbf{LC})$ is stable. $\hat{\mathbf{y}}[k]$ is the predicted value given by state space equation $\mathbf{y}^{(b)}[k] = \mathbf{C}\hat{\mathbf{x}}[k] + \mathbf{Du}[k]$. After the observer has converged, the predictive ability and the potential use of the model in feedback control implementations such as Model Predictive Control (MPC) can be evaluated by determining the ability of the model to predict future values of the outputs based only on the current state estimate and future inputs.

2.4.2 Quantifying Model Performance for a Candidate Input Frequency

This section demonstrates the application of the method by presenting the modeling results for the different input frequencies. Recall that for each choice of input frequency, data was generated from the test bed, used to identify a subspace model and ultimately, the prediction capability of the subspace model was tested.

Note that the prediction error is compared after the first 30 time-steps because as described in subsection 2.4 the observer is run until a maximum of 30 time-steps for every case. Thus the error after 30 time-steps for all variables and all cases are being compared, resulting in a comparison over the same duration regardless of different cases possibly requiring the observer to run for a different number of steps.

The error metric of Root Mean Square Error (RMSE) was chosen as the basis for comparison. The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(x_i - \hat{x_i})^2} \tag{2.7}$$

where $\mathbf{x_i}$ is the ith observation and $\hat{\mathbf{x}}_\mathbf{i}$ is the corresponding prediction.

The discrete nature of the subspace identification approach was recognized by using the inputs and outputs at discrete times, at the same frequency at which the input

changes are made. Further, the model is specifically designed not for the growth phase, but for the perfusion phase, thus data is collected only after the perfusion mode starts (i.e. after three days).

The choice of the number of states is a hyperparameter and can be chosen in multiple ways, one of which is by cross-validation minimizing prediction error, which is the chosen method for this particular work. The number of states was chosen based on the best results for a particular case (input change frequency). The number of states for cases 0-3 were 6, 5, 4 and 7 respectively.

The training input profile for the different inputs for case 2 (frequency of change once per three days) is shown in Figure 2.2. The inputs were changed by using random number deviations from a chosen baseline trajectory. The random numbers were seeded and the change in random numbers was made as per the mentioned frequency of three times per day to once per five days. As previously mentioned in sections 2.2 and 3.3.1, the environment in the reactor is key to cell growth and antibody production. Keeping in mind the presence of live cells, it is necessary to maintain an environment which is favorable to growth hence the temperature and pH are set close to the nominal values of 36.8 and 7.1 respectively. The temperature varies between 33 and 38.6 as values outside this range are more dangerous and can result in extinction events. The training trend itself is taken as a predetermined baseline trajectory and then perturbations are done around this baseline at a given frequency. Similarly, the pH varies between 6.9 and 7.3. The feed rate is conventionally set to 1.5 reactor volumes/day and it is varied from 1.35 to 1.65 reactor volumes/day. The glucose feed concentration was varied around the nominal value of 11 in the range of 9 to 13.

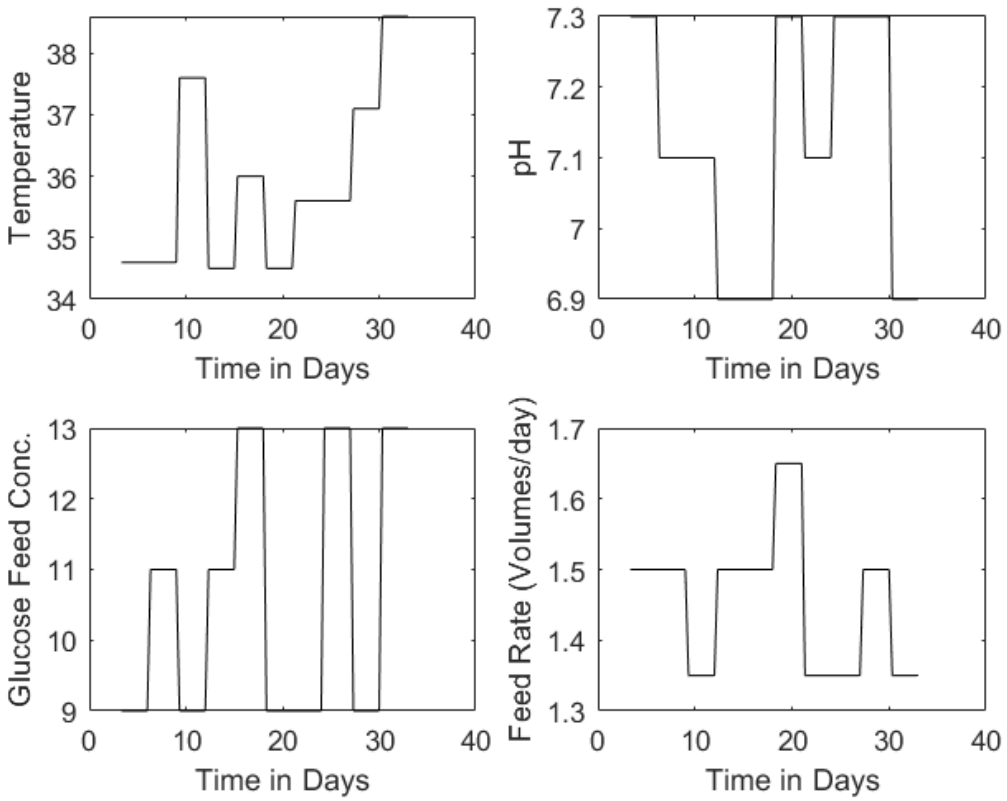


Figure 2.2: The input profiles for the four input variables for case 2 (frequency of change once per three days). Top left is temperature. Top right is pH. Bottom left is glucose feed concentration. Bottom right is feed rate.

The validation is done on a temperature shift profile consistent with Sartorius' industrial practices and is shown below in Figure 2.3. The pH was maintained at 7.1, feed rate at 1.5 volumes (300 ml) and glucose feed concentration at 13 g/L. The validation batch with a midway shift represents a batch intended to increase antibody production.

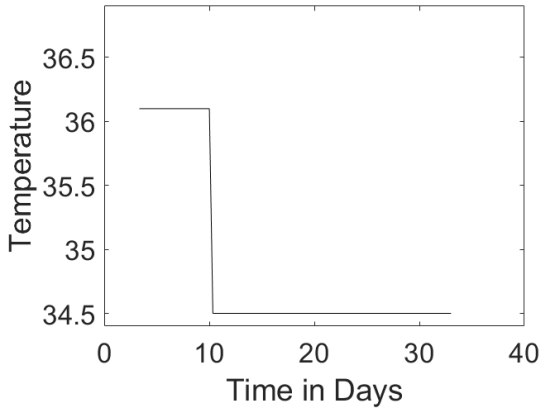


Figure 2.3: The validation input (Single Temperature Shift) profile, common for the four different trained models.

Glucose is closely related to the cell growth since it is a key nutrient, hence a dip in glucose also causes a dip in the cell viability as is evident from the plots presented. The viable cell density is directly proportional to the total viable cells present and the titer is being seen to be closely related to the viable cell density since the living cells produce titer. These trends are expected from the test bed and illustrate the validity of the underlying first principles equations. The subspace models identifying using the data for the various cases is able to capture this behavior as shown in Figure 2.5 and Figure 2.4. The results demonstrate the ability of a subspace model to accurately capture the bioprocess dynamics using only input output data.

The validation results are shown below in pre-processed scale in Figure 2.4. Note that the subspace model gives the output in pre-processed scale (necessary for model identification) which are then converted to the original scale in the end, the results for which are shown in Figure 2.5.

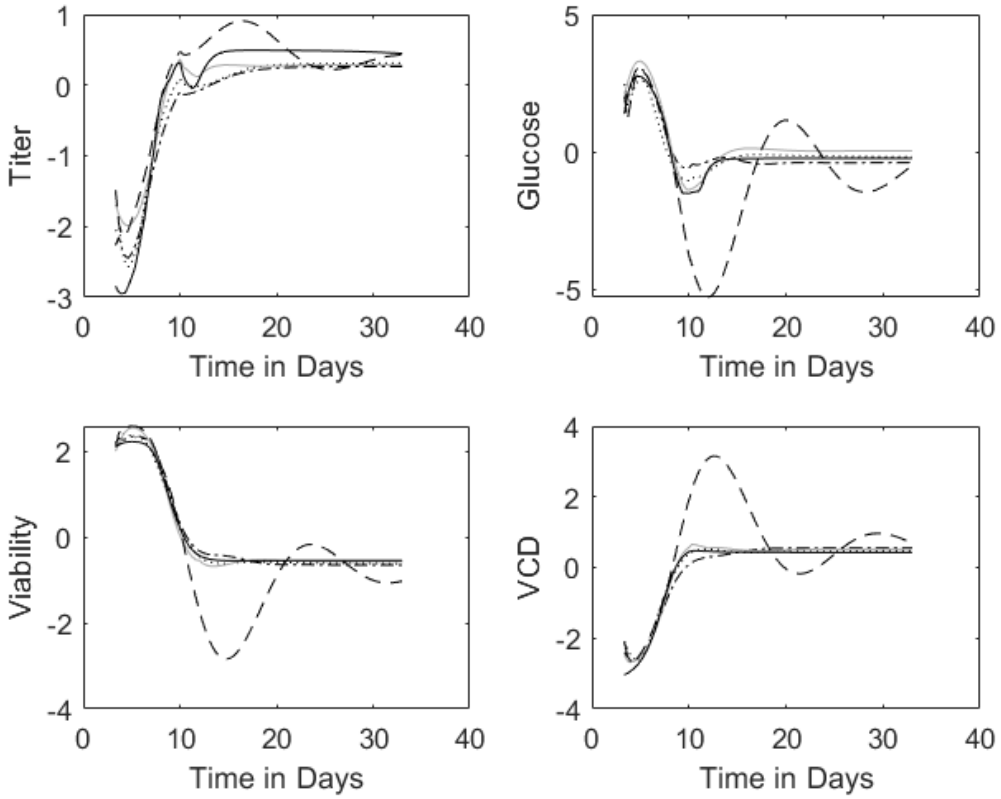


Figure 2.4: The output profiles for the four different trained models in pre-processed scale. Dotted line is prediction when training batch had input change three times per day. Dot-dash line had changes once per day. Solid grey line had changes once per three days. Dashed line had changes once per five days. Solid black lines are observation or actual data. Top left figure is for titer, top right is glucose, bottom left is viability and bottom right is viable cell density.

The table for prediction error in individual output variables for the four different cases is presented below:

Table 2.2: The prediction error between the subspace based model and the process for individual output variables, compared over models trained with four different input frequencies.

Model	*Titer Error*	*Glucose Error*	*Viability Error*	*VCD Error*
Case 0	0.2228	0.0885	0.0591	0.0305
Case 1	0.2598	0.1587	0.0950	0.1129
Case 2	0.1985	0.2876	0.0466	0.0823
Case 3	0.2447	1.4487	1.0705	0.9269

The overall prediction error (average over four different outputs) for the different models built on different training data (different frequencies of changes in input variables) is shown in table 2.3.

Table 2.3: The prediction error between the subspace based model and the process for the validation data, compared over models trained with four different input frequencies.

Model	*Validation Prediction Error (Average for four output variables)*
Case 0	0.1003
Case 1	0.1566
Case 2	0.1537
Case 3	0.9227

The results (validation plots) for all the output variables for all the four cases are shown below in original/actual scale in Figure 2.5

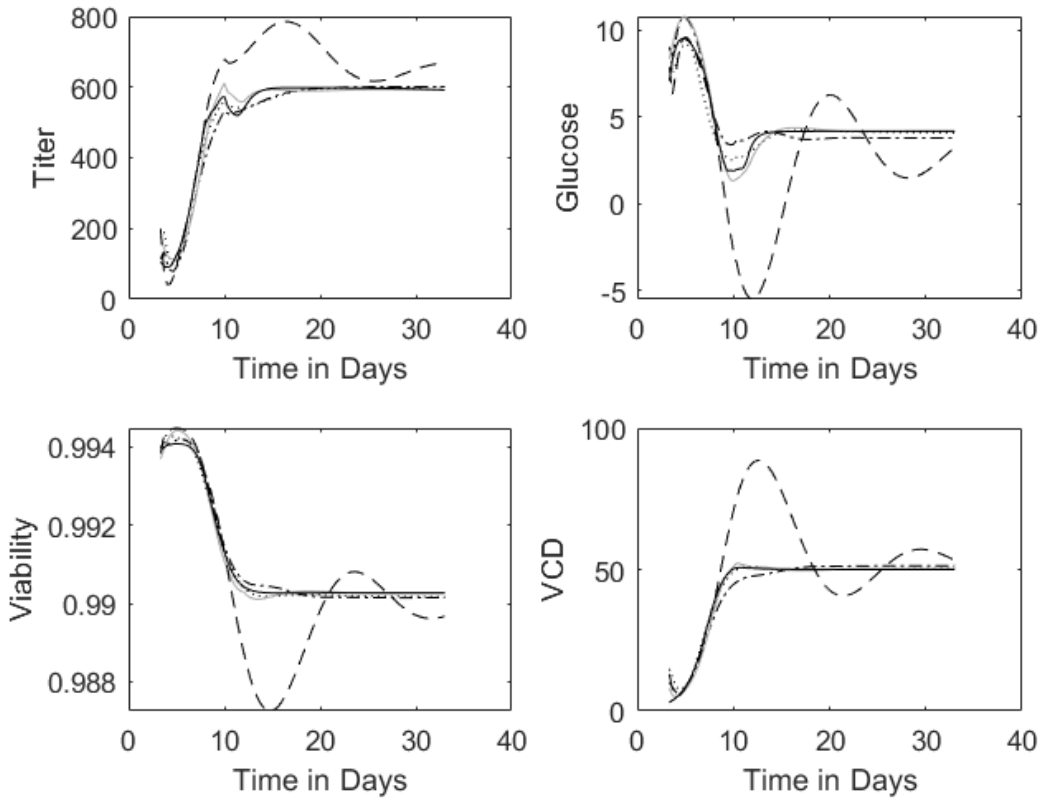


Figure 2.5: The output profiles for the four different trained models in actual scale. Dotted line is prediction when training batch had input change three times per day. Dot-dash line had changes once per day. Solid grey line had changes once per three days. Dashed line had changes once per five days. Solid black lines are observation or actual data. Top left figure is for titer, top right is glucose, bottom left is viability and bottom right is viable cell density.

As is evident from the prediction errors as well as can be visualized in the plots presented, the prediction accuracy for changing inputs once per three days (case 2) is the optimal solution. Having an input frequency lower than this, (case 3) had a significantly lower prediction accuracy. This is due to the effect of the input not being sufficiently rich as is required for subspace identification (in other words, the inputs did not end up exciting the low frequency dynamics). The cases with changes three times per day and once per day are not significantly better than once per three days, in fact the case for once per day is worse than once per three days except

for glucose in the pre-processed scale. This is due to the fact that the frequency of change for these cases is far too fast for this bio-process and since the dynamics of the bioreactor system are typically slower with several complex interactions, having the input changed too quickly results in the effect of the input on the output not being adequately expressed in the data. Also, based on the dynamics of cell growth and antibody production, from a biological perspective, it is preferred to have minimal perturbations. Keeping in mind the aforementioned reasons, case 2 is the optimal input frequency and is recommended to be applied online to the bioprocess.

A plot with the prediction for all four variables in case 2 (the final chosen frequency) with inputs changed once per three days is shown below in Figure 2.6

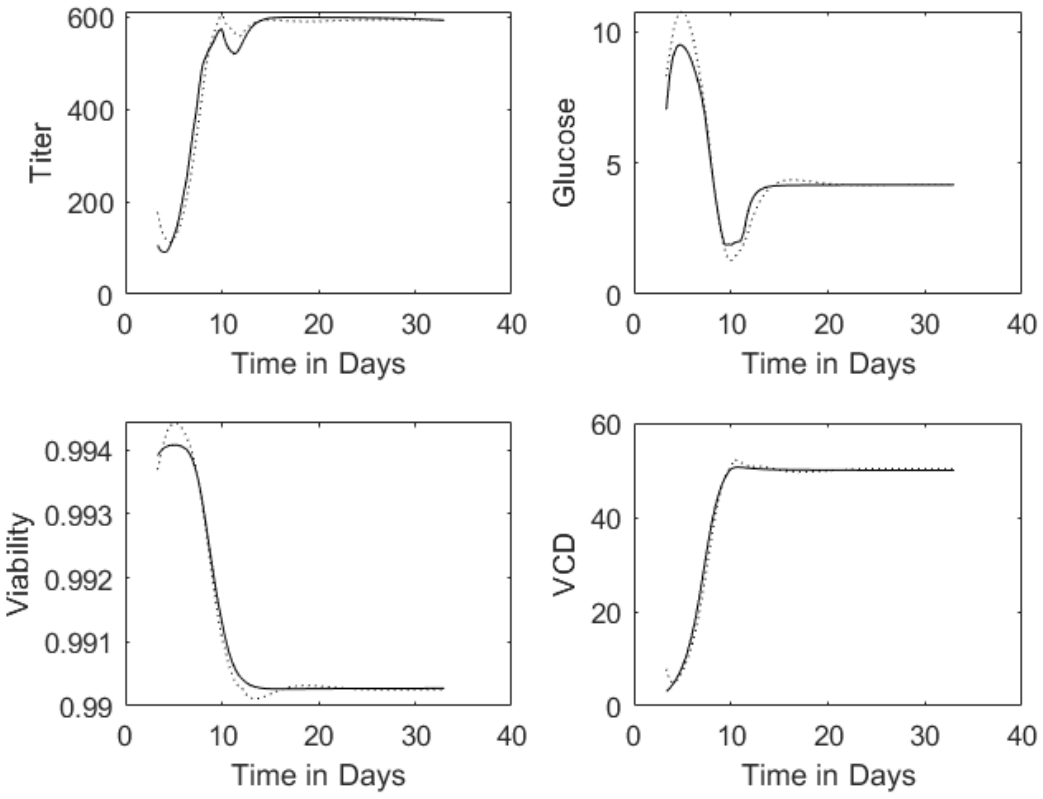


Figure 2.6: The prediction for all four outputs in the chosen case, i.e. case 2 (frequency of changes being made once per three days). The dotted line is prediction and the solid line is observation.

It is recognized that the first principles model of Sartorius is a test bed. Thus the key contribution of the work is not the specific frequency that was found to be the best for this particular case, but the approach of determining the best frequency that can be implemented with any other bioprocess easily and then used for building a control-relevant reduced order model from input-output data. Finally, note that the test-bed can be readily utilized as part of a hybrd model structure and embedded in an MPC implementation[13].

Remark 1. *While the common practice for biological systems is a traditional design of experiments without any perturbations, intensified design of experiments is a relatively new approach which aims to reduce experiments to get insightful data from the process by changing input multiple times [29, 2]. The proposed approach is significantly different from the existing technique in the context of bioreactors [29, 2] due to multiple reasons. First off, the notion of intensified design of experiments has focused on fed batch operation [27, 24] while the proposed approach addresses the problem for perfusion/continuous operation. More importantly the existing results focus on embedding nuanced constraints/requirements on the design, but do not consider the predictive capability of the resultant model. On the other hand, the proposed approach explicitly focuses on determining the input excitation that yields data that in turn gives the best predictive model while not perturbing the live cell system too much and is also applicable for multiple input-multiple output systems. Note that the present approach can use the notion of intensified design of experiments by making the characterization of the input changes more nuanced while still retaining the emphasis on the resultant model accuracy.*

2.5 Conclusions

In this work, a bioprocess under perfusion operation in the form of a high fidelity simulation testbed was utilized to present a method to determine appropriate input excitation. Specifically, the methodology addresses the problem of choosing an appropriate frequency of input changes based on the dynamics, sampling frequency and process run duration. The key objective for finding an appropriate frequency that maximizes the information content in the data to build a control relevant model while having minimal perturbations to the bioreactor. The optimal input frequency was determined to be once every three days after considering four different frequencies of input changes ranging from three times per day to once per five days. Utilising an appropriate frequency in conjunction with a data driven method suited for limited data provides a useful framework for control relevant modeling with a minimal number of experiments.

2.6 Acknowledgment

Financial support from Sartorius and the McMaster Advanced Control Consortium is gratefully acknowledged.

Bibliography

[1] Bernard, O., Mairet, F., and Chachuat, B. (2015). Modelling of microalgae culture systems with applications to control and optimization. In *Microalgae Biotechnology*, pages 59–87. Springer.

[2] Brendel, M. and Marquardt, W. (2008). Experimental design for the identification of hybrid reaction models from transient data. *Chemical Engineering Journal*, 141(1-3):264–277.

[3] Caramihai, M. and Severin, I. (2013). Bioprocess modeling and control. *Biomass Now: Sustainable Growth and Use*, page 147.

[4] Chusainow, J., Yang, Y. S., Yeo, J. H., Toh, P. C., Asvadi, P., Wong, N. S., and Yap, M. G. (2009). A study of monoclonal antibody-producing cho cell lines: What makes a stable high producer? *Biotechnology and bioengineering*, 102(4):1182–1196.

[Corbett and Mhaskar] Corbett, B. and Mhaskar, P. Subspace identification for data-driven modeling and quality control of batch processes. *AIChE Journal*, 62(5):1581–1601.

[6] Del Rio-Chanona, E. A., Cong, X., Bradford, E., Zhang, D., and Jing, K. (2019). Review of advanced physical and data-driven models for dynamic bioprocess simulation: Case study of algae–bacteria consortium wastewater treatment. *Biotechnology and bioengineering*, 116(2):342–353.

[7] Deschenes, J.-S., Desbiens, A., Perrier, M., and Kamen, A. (2006). Multivariable nonlinear control of biomass and metabolite concentrations in a high-cell-density perfusion bioreactor. *Industrial & engineering chemistry research*, 45(26):8985–8997.

[8] Dochain, D. and Perrier, M. (1997). Dynamical modelling, analysis, monitoring and control design for nonlinear bioprocesses. In *Biotreatment, Downstream Processing and Modelling*, pages 147–197. Springer.

[9] Downey, B., Schmitt, J., Beller, J., Russell, B., Quach, A., Hermann, E., Lyon, D., and Breit, J. (2017). A system identification approach for developing model predictive controllers of antibody quality attributes in cell culture processes. *Biotechnology progress*, 33(6):1647–1661.

[10] Ecker, D. M., Jones, S. D., and Levine, H. L. (2015). The therapeutic monoclonal antibody market. In *MAbs*, volume 7, pages 9–14. Taylor & Francis.

[11] Ferkl, L. and Širokỳ, J. (2010). Ceiling radiant cooling: Comparison of armax and subspace identification modelling methods. *Building and Environment*, 45(1):205–212.

[12] Ghosh, D., Hermonat, E., Mhaskar, P., Snowling, S., and Goel, R. (2019). Hybrid modeling approach integrating first-principles models with subspace identification. *Industrial & Engineering Chemistry Research*, 58(30):13533–13543.

[13] Ghosh, D., Moreira, J., and Mhaskar, P. (2021). Model predictive control embedding a parallel hybrid modeling strategy. *Industrial & Engineering Chemistry Research*.

[14] Godfrey, K. (1993). *Perturbation signals for system identification*. Prentice Hall International (UK) Ltd.

[15] Karra, S., Sager, B., and Karim, M. N. (2010). Multi-scale modeling of heterogeneities in mammalian cell culture processes. *Industrial & Engineering Chemistry Research*, 49(17):7990–8006.

[16] Leib, T. M., Pereira, C. J., and Villadsen, J. (2001). Bioreactors: a chemical engineering perspective. *Chemical engineering science*, 56(19):5485–5497.

[17] Li, F., Vijayasankaran, N., Shen, A., Kiss, R., and Amanullah, A. (2010). Cell culture processes for monoclonal antibody production. In *MAbs*, volume 2, pages 466–479. Taylor & Francis.

[18] Luenberger, D. (1971). An introduction to observers. *IEEE Transactions on automatic control*, 16(6):596–602.

[19] Mairet, F., Bernard, O., Cameron, E., Ras, M., Lardon, L., Steyer, J.-P., and Chachuat, B. (2012). Three-reaction model for the anaerobic digestion of microalgae. *Biotechnology and Bioengineering*, 109(2):415–425.

[20] Moonen, M., De Moor, B., Vandenberghe, L., and Vandewalle, J. (1989). On-and off-line identification of linear state-space models. *International Journal of Control*, 49(1):219–232.

[21] Morel, E., Tartakovsky, B., Guiot, S., and Perrier, M. (2006). Design of a multimodel observer-based estimator for anaerobic reactor monitoring. *Computers & chemical engineering*, 31(2):78–85.

[22] Patel, N., Corbett, B., Trygg, J., McCready, C., and Mhaskar, P. (2020). Subspace based model identification for an industrial bioreactor: Handling infrequent sampling using missing data algorithms. *Processes*, 8(12):1686.

[23] Pörtner, R., Barradas, O. P., Frahm, B., and Hass, V. C. (2017). Advanced process and control strategies for bioreactors. In *Current Developments in Biotechnology and Bioengineering*, pages 463–493. Elsevier.

[24] Schaepe, S., Kuprijanov, A., Simutis, R., and Lübbert, A. (2014). Avoiding overfeeding in high cell density fed-batch cultures of e. coli during the production of heterologous proteins. *Journal of biotechnology*, 192:146–153.

[25] Simutis, R. and Lübbert, A. (2015). Bioreactor control improves bioprocess performance. *Biotechnology journal*, 10(8):1115–1130.

[26] Sirois, J., Perrier, M., and Archambault, J. (2000). Development of a two-step segregated model for the optimization of plant cell growth. *Control Engineering Practice*, 8(7):813–820.

[27] von Stosch, M., Hamelink, J.-M., and Oliveira, R. (2016). Hybrid modeling as a qbd/pat tool in process development: an industrial e. coli case study. *Bioprocess and biosystems engineering*, 39(5):773–784.

[28] Von Stosch, M., Oliveira, R., Peres, J., and de Azevedo, S. F. (2014). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers & Chemical Engineering*, 60:86–101.

[29] von Stosch, M. and Willis, M. J. (2017). Intensified design of experiments for upstream bioreactors. *Engineering in Life Sciences*, 17(11):1173–1184.

[30] Wong, W. C., Chee, E., Li, J., and Wang, X. (2018). Recurrent neural network-based model predictive control for continuous pharmaceutical manufacturing. *Mathematics*, 6(11):242.

[31] Xie, L. and Wang, D. I. (1996). High cell density and high monoclonal antibody production through medium design and rational control in a bioreactor. *Biotechnology and bioengineering*, 51(6):725–729.

[32] Yang, O., Prabhu, S., and Ierapetritou, M. (2019). Comparison between batch and continuous monoclonal antibody production and economic analysis. *Industrial & Engineering Chemistry Research*, 58(15):5851–5863.

[33] Zhang, D., Del Rio-Chanona, E. A., Petsagkourakis, P., and Wagner, J. (2019). Hybrid physics-based and data-driven modeling for bioprocess online simulation and optimization. *Biotechnology and bioengineering*, 116(11):2919–2930.

Chapter 3

Process-Aware Data Driven Modeling and Model Predictive Control of Bioreactor for Production of Monoclonal Antibodies

The previous chapter presented a novel design of experiments approach combined with subspace identification to identify appropriate input frequency and build data driven model for monoclonal antibody perfusion process. This chapter builds upon that to build a model predictive control scheme capable of meeting biological requirements and input constraints. Further improvement to the modeling and control solution is made by a novel constrained subspace method incorporating first principles knowledge through constraints on signs on gain matrix. Robustness of the method is also demonstrated by good performance on a constrained model despite being built using data from a different cell line.

This work was completed in collaboration with Sartorius Inc who provided insights into the bioprocess and provided a high fidelity advanced simulator as a testbed.

The work has been accepted for publication in the Canadian Journal of Chemical Engineering.

Sarna, S., Patel, N., Corbett, B., McCready, C., & Mhaskar, P. (2022).
Process-Aware Data Driven Modeling and Model Predictive Control of Bioreactor for Production of Monoclonal Antibodies. *Canadian Journal of Chemical Engineering*, (Accepted)

3.1 Abstract

This manuscript addresses the problem of controlling a bio-reactor to maximize the production of a desired product while respecting the constraints imposed by the nature of the bio-process. The approach is demonstrated by first building a data driven model and then formulating a model predictive controller (MPC) with the results illustrated by implementing on a detailed monoclonal antibody production model (the test bed) created by Sartorius Inc. In particular, a recently developed data driven modeling approach using an adaptation of subspace identification techniques is utilized that enables incorporation of known physical relationships in the data driven model development (constrained subspace model identification) making the data-driven model process aware. The resultant controller implementation demonstrates significant improvement in product compared to the existing PI controller strategy used in the monoclonal antibody production. Simulation results also demonstrate the superiority of the process aware or constrained subspace model predictive controller compared to traditional subspace model predictive controller. Finally, the robustness of the controller design is illustrated via implementation of a model developed using data from a test bed with a different set of parameters, thus showing the ability of the controller design to maintain good performance in the event of changes such as a different cell line or feed characteristics.

3.2 Introduction

The need for bio-based and pharmaceutical products is on the rise with advancements in healthcare and the demand of an ever-increasing global population. Bioreactors form an important part of this industry by allowing for the mass production of these bio-pharmaceutical products. One such product is a monoclonal antibody which is produced by Sartorius and is used to demonstrate the model based controller design approach. The Sartorius Bioreactor is designed to produce this protein in a perfusion processing setup.

There exist several challenges associated with control of bio reactors in general, and the monoclonal antibody production in consideration, in particular (herewith referred to as the Sartorius Bioreactor). First off, in contrast to batch or fed-batch processing [24, 6, 28] the Sartorius Bioreactor is operated in perfusion mode (thus is a continuous removal of bleed and harvest streams). In addition to the perfusion mode of operation, there are several other factors, such as hydrodynamics and transport phenomena [16], that affect the volumetric production of the monoclonal antibodies. Factors like cell growth rate, feed rate and feed concentration are all key variables in bioreactor operation and thus, these factors need to be accounted for in order to maximize the final product. The final product is a combination of the volumetric flow rate (referred to as harvest rate) and a high volume specific concentration of the antibody (referred to as titer). In order to maximize the final product the individual interactions between different inputs, outputs and other parameters must first be examined. The first and most important parameter to consider is glucose concentration, as glucose is the key energy source, however, excess glucose is also detrimental due to lactate production which increases cell death. Similarly, Glutamine plays an important role especially for promoting cell growth during periods of fast growth. Lactate and more so, ammonia, are inimical to cell growth[14]. These metabolites are not the only factors affecting

cell growth. Both temperature and pH play a key role. Increasing temperature has been shown to increase cell growth rate. However, high temperatures can also lead to cell death. To handle this issue, increased antibody production is achieved with a midway temperature shift [4, 31]. A more complex variable is pH since pH levels also affect ammonia and lactate levels. Often, a shift in pH in later stages is necessary [14]. Further, due to operational considerations, it is preferable to decrease the pH rather than increase it since it can be decreased by sparging CO_2 but increasing pH would require the addition of a base that could potentially disturb the cell environment negatively. In essence, since the production of a specific product such as a protein by these cells is heavily affected by the environment in the reactor, such as the pH and glucose levels [2, 26, 28] and with such a diversity of variables affecting the system dynamics with many of these having contradictory effects in different ranges, the modelling and control problem is a challenging one.

With the increasing recognition of the flexibility provided by process control in process operation, process control is being adopted within the bio-processing industry [27]. One popular and successful control strategy that has been used in large scale production is model predictive control (MPC). MPC relies on a process model to calculate the optimal input trajectory to meet desired objectives while respecting constraints or bounds. MPC has been implemented in chemical industries and the energy sector with favorable results. In recent years it has also been implemented for biochemical and fermentation processes [21, 15, 3]. However, MPC of bioreactors is not common in industry due to the sensitive nature of the cells and the set batch recipes available. Instead proportional integral (PI) control is used to follow a batch trajectory. The use of PI control however, potentially limits the productivity of the process (as illustrated by the results in this manuscript) motivating the need to explore the implementation of MPC.

In an MPC implementation, the process model forms the heart of the entire strat-

egy therefore identifying a good model is critical to improved control. When modelling a system, first principles models are valuable since they provide a direct insight into the process. Although parameter estimation for first principles models is challenging, parameter estimation methods for first principles models exists in literature [8, 29, 7, 22, 19, 1], and this has been applied to bioreactors [14]. More recently, Sartorius Inc. has developed a high fidelity simulator for the monoclonal antibody process, and is used in the present manuscript to illustrate the control approach. The detailed simulator, while being a good representation of the bioreactor, is not very suitable for direct incorporation in an MPC formulation due to model complexity. More importantly, it is of much more benefit to the practitioner to demonstrate the implementation of a control approach that can readily utilize process data directly for model development and control implementation.

Data driven and black box models are one choice for ease of implementation[6, 32]. Reduced order models can also achieve high performance control if it is possible to capture basic and fundamental dynamical features of the system. The performance of the controller is often the main objective for model building in these instances and thus such kind of models are valuable [12]. Within data-driven methods, there are several different approaches; however, not all such approaches are suitable for the Sartorius bioreactor problem. One particular concern is that the complex metabolite interactions require specific gains that must be adhered to in the data driven model. To that end any modeling approach must be capable of incorporating these constraints with minimal complexity. Techniques such as Partial Least Squares (PLS) do not explicitly differentiate between inputs and outputs or handle multiple batches without additional complexity [11, 10]. To that end, an approach involving Linear Time Invariant (LTI) models would be better suited to handle this problem. One such approach is subspace identification, which is a well established system identification method and has several advantages such as having only one decision variable (the order of the system) and its ability to handle large multi-input multi-output (MIMO) prob-

lems well. The model complexity of MIMO and the simpler single-input single-output (SISO) systems is similar when using subspace identification. This is in contrast to methods such as Auto-Regressive Moving Average with eXogenous inputs (ARMAX) models which have multiple 'tuning' parameters. In comparison to methods such as ARMAX, subspace identification is often easier to implement, faster and more accurate, including cases with white noise. [9] Additionally, recent results have allowed imposing constraints in subspace identification at the modeling stage with minimal additional computational complexity [24], to enable the model to be more 'aware' of the process.

Motivated by the above considerations, the present work addresses the problem of maximizing the production in a Sartorius bioreactor using MPC with a process aware or constrained subspace model. Specifically, a process aware subspace MPC is implemented on the simulation test-bed and compared against existing PI control. Next the need to implement process aware MPC is demonstrated by comparing against a traditional subspace model based MPC. Finally, the robustness of the MPC approach is tested by comparing the MPC against a new process with different system dynamics. The rest of the paper is arranged as following: Section 3.3 described the bioreactor process, reviews subspace identification and constrained subspace identification. The model predictive control scheme which is developed and used is presented in section 3.4. Section 3.5 presents the application of the proposed method to the Sartorius Bioreactor. Concluding remarks are presented in Section 3.6.

3.3 Preliminaries

3.3.1 Bioreactor Process Description

The Sartorius Bioreactor grows live cells in an enclosed environment meaning that the growth and death rates affect the environment in the reactor and consequently the titer (final product). A simplified schematic of the bioreactor is shown in figure 3.1. The recycle stream shown in the figure recycles live cells as a cell retention filter does not allow live cells to leave in the harvest stream.

A detailed first principles model developed by Sartorius is used as a test bed in the present manuscript. The Sartorius simulator comprises a system of 10 ordinary differential equations to describe the time evolution of variables including the cells and metabolites (characterized by viable cell density (VCD), dead cell density, lysed cell density, biomaterial, titer, glucose, glutamine, lactate, ammonia and glutamate). The parameters, and various function describing growth rates etc are determined by fitting the model to experimental data from twelve AMBR 250 ® (Sartorius registered trademark for integrated high throughput bioreactor systems) fed batch runs to yield a biologically meaningful and fairly accurate description of the bioreactor. Transferrability of this model structure from fed-batch to perfusion operation has been established by Sartorius researchers, and as such, the present model is being utilized to demonstrate the data driven modeling and control approach.

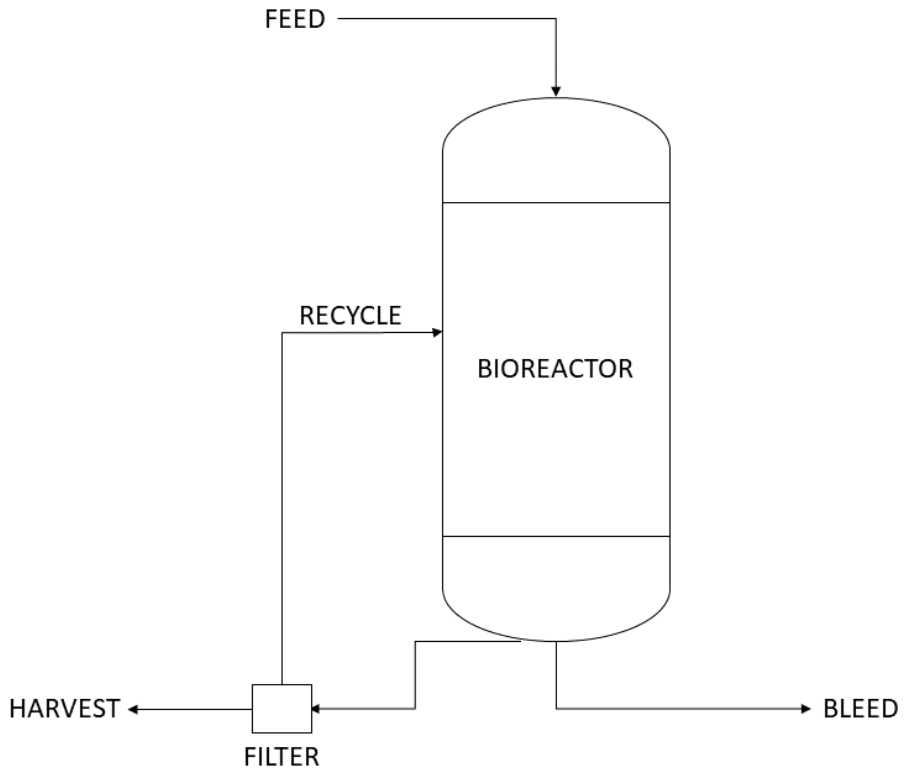


Figure 3.1: Schematic of Sartorius Bioreactor

The process initiates in growth phase for 3 days during which the system is operated in a fed batch fashion. This is followed by a perfusion phase for 30 days. In this work, based on the specific process used by Sartorius, the nominal values for the temperature and pH are set at $36.1°C$ and 7.1 respectively. The reactor temperature (°C), pH, glucose feed concentration (g/L), feed rate (vols/day) and bleed rate(L/day) are available as potential inputs. The measured outputs are viability (%), viable cell density (VCD) (10^5 cells/mL)], titer (mg/L) and glucose concentration (g/L). By design the bioreactor has control over and the ability to manipulate all of these inputs and hence these are chosen as manipulated variables in the control scheme. Note that plant model mismatch, the natural 'drifting' of parameters over time act as disturbances and in the present manuscript, robustness to such disturbances is demonstrated via a successful control of a 'new cell line' using an MPC designed

using data from a previous operation. The inputs and outputs are organized in the following vectors:

$$u = \begin{bmatrix} Reactor\ Temp \\ pH \\ Feed\ Conc \\ Feed\ Rate \\ Bleed\ Rate \end{bmatrix}$$

$$y = \begin{bmatrix} Viability \\ VCD \\ Titer \\ Glucose\ Conc \end{bmatrix}$$

The process objective is to maximize bioreactor production over the course of the perfusion phase which is currently done by putting VCD under PI control where with a fixed setpoint of 50. The PI controller that is currently employed adjusts the the bleed rate in order to control the VCD. With the feed rate kept constant at 0.25L/day or 1.25 volumes/day and under constant volume operation, the harvest rate can be computed as:

$$Harvest\ Rate = Feed\ Rate - Bleed\ Rate \tag{3.1}$$

As the bleed rate is the most significant contributor to cell growth, it is utilized as the control variable with additional shifts in temperature or pH being applied by the operators manually. The objective of the present work is to demonstrate the possibility of using a data driven MPC to control and improve the bioprocess operation.

The current industry state of the art practice is using proportional-integral (PI) controller. For discrete time implementations, Sartorius has tuned a velocity-form PI

controller described below:

$$\begin{aligned} e[k] &= y[k] - y_{sp}[k] \\ \Delta u &= (e[k] - e[k-1]) \times Kc + \frac{Kc}{\tau_I \times e} \\ u[k] &= max(0, u[k-1] + \Delta u) \end{aligned} \tag{3.2}$$

Where y_{sp} is the set-point of the controlled output VCD, u corresponds to the manipulated input bleed rate and the tuned values of Kc and τ_I are 0.003 and 10.

3.3.2 Subspace Identification Description

Subspace identification is one model identification technique that is used to identify a Linear Time Invariant (LTI) model [17] of the form:

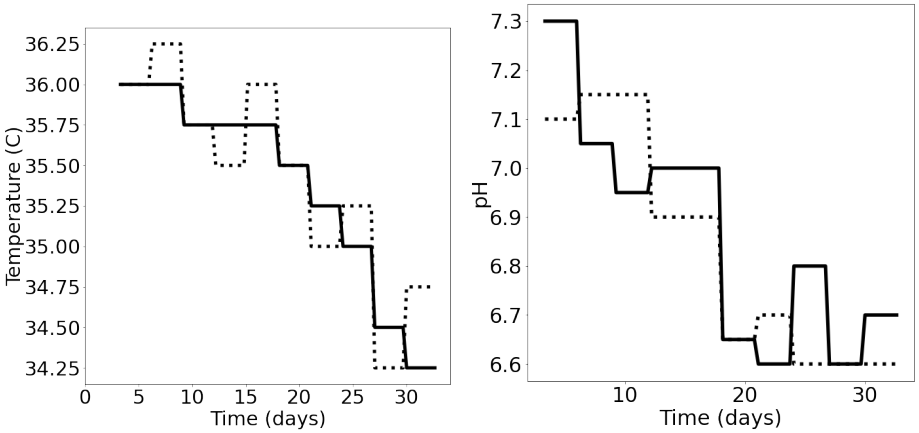
$$\begin{aligned} \hat{\mathbf{x}}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k], \\ \mathbf{y}[k] &= \mathbf{C}\hat{\mathbf{x}}[k] + \mathbf{D}\mathbf{u}[k], \end{aligned} \tag{3.3}$$

where the objective is to identify the order n, and the system matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$, $\mathbf{C} \in \mathbb{R}^{l \times n}$, $\mathbf{D} \in \mathbb{R}^{l \times m}$. The particular subspace adaptation utilized in the present work is originally based on [20], which was later adapted for batch processes[Corbett and Mhaskar].

An important consideration for the process under consideration is to ensure that the bioreactor is not disturbed too frequently otherwise the cell balance will be disrupted leading to inefficient protein production and additional costs. Thus an appropriate frequency of input changes is utilized in collecting data such that it has the fewest number of perturbations while meeting reasonable prediction accuracy. Based on a preliminary analysis, perturbation of inputs three times per day is utilized in the

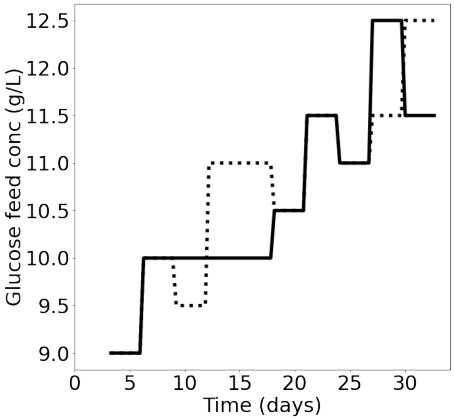
present work. An additional consideration is that the process runs in growth phase for 3 days followed by perfusion phase for 30 days thus it takes over a month of time (and significant costs) to generate data. To be able to demonstrate the approach, data is generated from a detailed simulation test bed provided by Sartorius. The data was generated by gradual shifts in the inputs over their appropriate constrained ranges along with small perturbations. Data was obtained from a single batch run over thirty days with measurements available thrice a day (for a total of 90 measurements). Note that the ability to use this relatively modest dataset is extremely important for the process under consideration where each run is prohibitively expensive.

The Sartorius simulator is used to generate input-output trajectory for one run, and this data is assumed to be available for building the data driven MPC, and shown in figures 3.2 and 3.3.

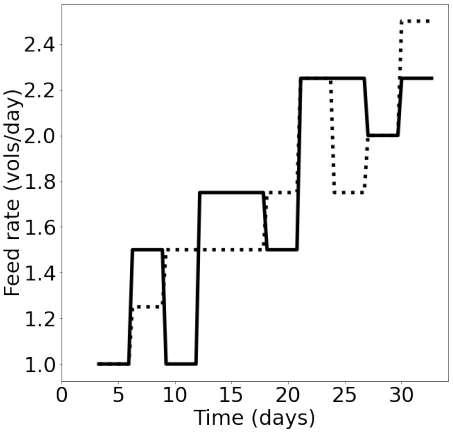


(a) Temperature

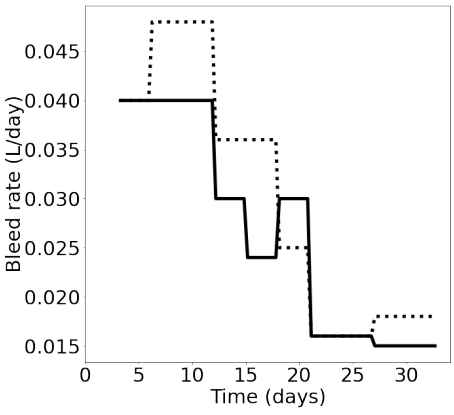
(b) pH



(c) Glucose feed concentration

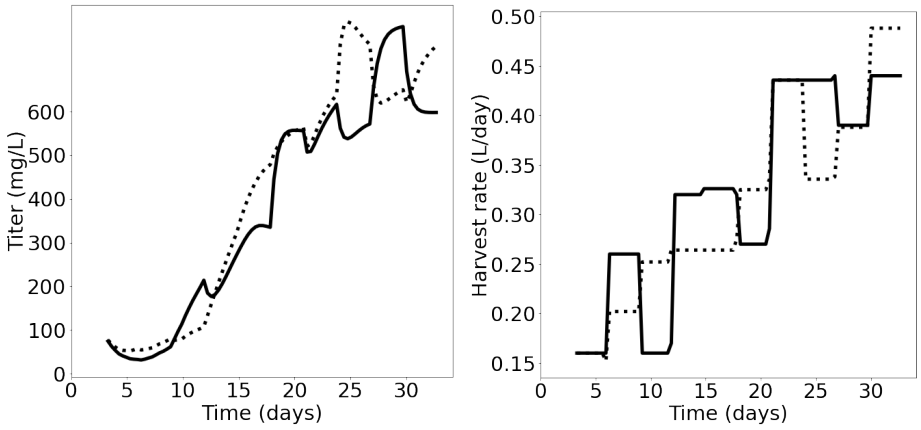


(d) Feed rate



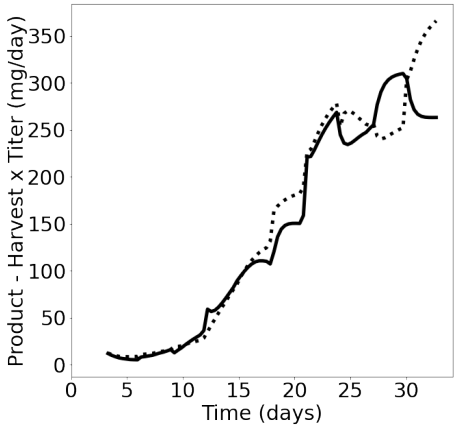
(e) Bleed rate

Figure 3.2: Input data for building model with training (dotted) and validation data (solid).

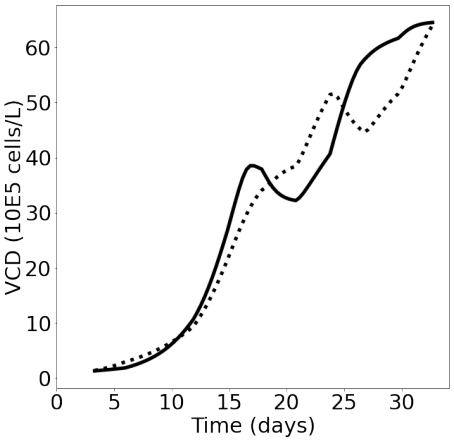


(a) Titer

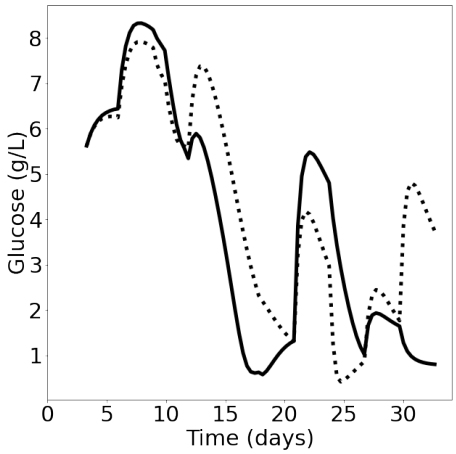
(b) Harvest



(c) Product (Titer × Harvest)



(d) Viable Cell Density (VCD)



(e) Glucose Concentration in Reactor

Figure 3.3: Output data for building model with training (dotted) and validation data (solid).

Identification of the system matrices is done in two stages, first stage involves identifying a state sequence and the second stage comprises of identifying the system matrices [20]. Using subspace identification, the state sequence can be identified using methods such as SVD before knowing the A,B,C,D system matrices. The system matrices are later identified by least squares regression.

In solving for the state sequence, block Hankel matrices are constructed for the inputs and outputs. The number of block rows (i) and columns (j) are chosen sufficiently large, typically i should be greater than or equal to $n+1$ and $j >> max(mi, li)$.

The output and input block Hankel matrices are:

$$Y_p = \begin{bmatrix} y[k] & y[k+1] & \dots & y[k+j-1] \\ y[k+1] & y[k+2] & \dots & y[k+j] \\ y[k+2] & y[k+3] & \dots & y[k+j+1] \\ \vdots & & & \vdots \\ y[k+i-1] & y[k+i] & \dots & y[k+i+j-2] \end{bmatrix}$$

$$U_p = \begin{bmatrix} u[k] & u[k+1] & \dots & u[k+j-1] \\ u[k+1] & u[k+2] & \dots & u[k+j] \\ u[k+2] & u[k+3] & \dots & u[k+j+1] \\ \vdots & & & \vdots \\ u[k+i-1] & u[k+i] & \dots & u[k+i+j-2] \end{bmatrix}$$

Y_f and U_f are defined similar to Y_p and U_p except the values are offset by i. These matrices are used to identify the state vector which can be organized in a Hankel matrix allowing the A,B,C,D matrices to be solved by least squares regression.

$$\begin{bmatrix} x[k+i+1]\dots & x[k+i+j] \\ y[k+i]\dots & y[k+i+j-1] \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x[k+i]\dots & x[k+i+j-1] \\ u[k+i]\dots & u[k+i+j-1] \end{bmatrix} \tag{3.4}$$

Note that the approach identifies a linear state space model where the states are unmeasured but intrinsically observable from measured outputs. As subspace states are not measured, it is necessary to estimate the states before using the subspace model for prediction/validation. To this end, during model validation an initial state estimate is chosen (can be based on a state estimate from identification batch or a random initialization) and a state observer is utilised. The state observer is run until the error (Euclidean norm) between the predicted output and actual/observed output is below a chosen tolerance, from which point on the model can be utilized for prediction purposes. This same approach is utilized when the model is used for feeback control. Note that Sartorius Bioreactor operation has the unique advantage of an initial 3 day growth phase (without any feedback control) that can be used to converge the states allowing the controller to be used online immediately after the growth phase ends.

In this work a Luenberger observer [18] was used which takes the following form:

$$\hat{\mathbf{x}}[k+1] = \mathbf{A}\hat{\mathbf{x}}[k] + \mathbf{B}\mathbf{u}[k] + \mathbf{L}(\mathbf{y}[k] - \hat{\mathbf{y}}[k]) \tag{3.5}$$

where $\mathbf{L}$ is the observer gain and is chosen such that $(\mathbf{A} - \mathbf{LC})$ is stable. $\hat{\mathbf{y}}[k]$ is the predicted value given by the state space equation $\mathbf{y}[k] = \mathbf{C}\hat{\mathbf{x}}[k] + \mathbf{D}\mathbf{u}[k]$.

The identified system matrices by regular subspace identification are:

$$Y_p = \begin{bmatrix} y[k] & y[k+1] & \dots & y[k+j-1] \\ y[k+1] & y[k+2] & \dots & y[k+j] \\ y[k+2] & y[k+3] & \dots & y[k+j+1] \\ \vdots & & & \vdots \\ y[k+i-1] & y[k+i] & \dots & y[k+i+j-2] \end{bmatrix}$$

$$U_p = \begin{bmatrix} u[k] & u[k+1] & \dots & u[k+j-1] \\ u[k+1] & u[k+2] & \dots & u[k+j] \\ u[k+2] & u[k+3] & \dots & u[k+j+1] \\ \vdots & & & \vdots \\ u[k+i-1] & u[k+i] & \dots & u[k+i+j-2] \end{bmatrix}$$

3.3.3 Constrained Subspace Identification

In this subsection the approach used to impose the physical constraints on the subspace model is described [25]. The key idea is to include the first-principles knowledge of the system at the model identification stage through the use of constraints to make the data-driven model process aware. In this approach, instead of using regression to determine the model parameters (3.4) an optimization problem with the first principles based constraints is posed and solved. Thus, while the initial state trajectory may have been determined without considering the physical constraints, the resultant matrices do account for the presence of constraints. To minimize the discord between the state trajectory and the constraints, the state trajectory is re-estimated using the newly computed system matrices, and this iterative process terminated when a pre-decided tolerance is achieved (see [25] for further details).

For the Bioreactor, it is understood that the steady state gain between the bleed rate and titer (product) should be negative. Similarly a positive relation holds for temperature, glucose feed concentration and feed rate and is incorporated into the constrained subspace model through constraints. The constraints are therefore mathematically formulated as:

$$\begin{aligned} dcgain(3,5) + 0.01 <= 0 \\ -dcgain(3,1) + 0.01 <= 0 \\ -dcgain(3,3) + 0.01 <= 0 \\ -dcgain(3,4) + 0.01 <= 0 \\ norm(eig(A)_j) - 0.99 <= 0, j = 1, ...N \end{aligned} \tag{3.6}$$

where $dcgain(i,j)$ refers to the $(i,j)^{th}$ index of the steady state gain matrix, which for a discrete linear time invariant (LTI) system is:

$$dcgain = D + C(I - A)^{-1}B \tag{3.7}$$

Thus the first constraint specifies $dcgain(3,5)$ to be negative. That is, the gain between the third output (titer) and the fifth input (bleed rate) should be negative. Similarly, the other gain constraints enforce the positive steady state gain relationship between the titer and the temperature, feed concentration and feed rate, respectively. The final constraint is for the eigenvalues of the identified A matrix to lie within the unit circle.

The identified matrices for the constrained case are given below:

$$A_c = \begin{bmatrix} 0.9623 & -0.0439 & -0.0433 & -0.1074 & 0.0584 & 0.1023 \\ 0.0567 & 0.9112 & -0.0932 & -0.0227 & -0.0259 & 0.1783 \\ -0.0420 & 0.0921 & 0.9340 & -0.0567 & -0.1187 & -0.1246 \\ 0.1216 & 0.0526 & 0.0753 & 0.8233 & -0.0959 & 0.2526 \\ -0.0385 & 0.0314 & -0.0295 & 0.1431 & 0.8519 & 0.0116 \\ -0.0089 & 0.0058 & 0.019 & -0.0134 & -0.0060 & 0.8722 \end{bmatrix}$$

$$B_c = \begin{bmatrix} 0.0092 & -0.0242 & 0.0295 & -0.2165 & -1.3273 \\ -0.0018 & -0.0828 & 0.006 & 0.339 & -2.1627 \\ 0.0083 & 0.0808 & -0.031 & -0.3866 & 0.7003 \\ -0.011 & -0.0455 & -0.0276 & -0.601 & 0.8038 \\ -0.0071 & 0.1129 & -0.0219 & 0.2169 & 0.4335 \\ 5.0176e-05 & 0.0124 & -0.005 & 0.1087 & 0.5853 \end{bmatrix}$$

$$C_c = \begin{bmatrix} -0.1644 & -1.0734 & -1.6297 & -0.6979 & 0.3026 & 0.7797 \\ -0.2343 & 0.9999 & -0.2667 & -0.4763 & 0.0552 & 0.208 \\ 0.2704 & 1.1325 & -0.0565 & 0.1059 & 0.4200 & 0.6313 \\ -0.2525 & -1.597 & 0.2053 & -0.366 & -1.8811 & -0.7681 \end{bmatrix}$$

$$D_c = \begin{bmatrix} -0.0051 & 0.3442 & -0.085 & 1.1634 & -3.356 \\ 0.0395 & 0.1102 & 0.006 & 0.4413 & -2.5792 \\ 0.0724 & -0.695 & 0.0234 & -2.2726 & 5.1213 \\ -0.1012 & 0.3344 & 0.268 & 3.0322 & -8.577 \end{bmatrix}$$

The identified matrices for the regular case are given below:

$$A = \begin{bmatrix} 0.8829 & -0.0224 & 0.1269 & -0.0128 \\ -0.0362 & 0.946 & -0.1341 & -0.0008 \\ -0.0828 & 0.0476 & 0.9136 & 0.1849 \\ 0.1446 & 0.0316 & -0.0397 & 0.9412 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.0073 & 0.0448 & 0.0131 & -0.1117 & -1.726 \\ -0.0031 & 0.0424 & -0.0054 & -0.0052 & 1.405 \\ -0.0058 & 0.024 & 0.027 & 0.5971 & -0.7986 \\ -0.0185 & -0.0008 & -0.0082 & -0.136 & 1.5365 \end{bmatrix}$$

$$C = \begin{bmatrix} -0.452 & 0.9336 & 1.5757 & 0.1143 \\ -0.3125 & -1.1584 & 0.8583 & -0.0731 \\ 0.5158 & -1.009 & 0.5781 & 0.9361 \\ -0.323 & 1.541 & -0.915 & -2.4586 \end{bmatrix}$$

$$D = \begin{bmatrix} -0.0175 & 0.3966 & -0.0082 & 0.8305 & -1.219 \\ 0.0081 & 0.2437 & -0.0258 & 0.5727 & -5.6478 \\ 0.0437 & -0.534 & 0.0142 & -2.2901 & 5.762 \\ -0.015 & -0.3498 & 0.3135 & 3.0965 & 0.59434 \end{bmatrix}$$

The number of subspace states were chosen as 6 and 4 respectively for constrained and regular (unconstrained) based on best model accuracy for their respective cases.

Remark 2. *Subspace identification was chosen as the model identification approach due to the following reasons: 1) The method results in a linear time invariant model which in turn makes the resultant control problem easy to solve and implement, 2) Compared to other approaches such as Projection to*

Latent Spaces (PLS), the method explicitly accounts for the presence of input and output variables, consistent with a control implementation and 3) Even though the model parameters are eventually determined using a regression, the state trajectory first invokes the key property of a state- that future outputs should be able to be completely determined using the current states and the future inputs, thus avoiding potential overfitting issues. The implementation does not require a first principles model- in the present manuscript, the first principles model is simply used as a test bed. Finally, the method can be readily adapted to incorporate first principles information either explicitly, as is done in the present manuscript, or through a hybrid model [13].

Remark 3. *Yes another possibility for model identification would be to use the resurgent techniques of artificial neural networks. For developing a good neural network model, large amounts of data is needed. While it is possible in principle with the test bed, it would not be very feasible when this technique is implemented in practice on the Sartorius Bioreactor. Note that the data set size utilized for training in the present work was chosen while being cognizant of the cost and effort needed to generate data from bioreactor.*

3.4 Model Predictive Controller Formulation

In this section, the state space MPC formulation adapted to use the feedthrough matrix [23] utilizing the subspace model of Eqn 3.4 is described. A Python script utilising scipy.optimize [30] is used to solve the optimization problem. At the l^{th} time step, with the observer determining $\hat{x}[l]$, the following optimization problem is solved to compute the control action:

$$
\begin{aligned}
\min_{\bar{u}} \sum_{i=1}^{P} (y_3[i])^T Q(u_4[i] - u_5[i]) \\
+(\Delta u^T) R(\Delta u) + S\Delta u
\end{aligned}
$$

s.t.

$$
\begin{aligned}
u_{min} \leq \bar{u}[i] \leq u_{max} \ , \ i = 1, \ldots, P \\
\Delta u = \bar{u}[i] - \bar{u}[i-1] \ , \ i = 2, \ldots, P \\
\Delta u[1] = \bar{u}[1] - u[l-1] \\
x[1] = \hat{x}[l] \\
x[i+1] = Ax[i] + B\bar{u}[i] \ , \ i = 1, \ldots, P \\
y[i] = Cx[i] + D\bar{u}[i] \ , \ i = 1, \ldots, P
\end{aligned}
\tag{3.8}
$$

where y denotes the predicted output obtained from Eqn 3.4 and as described in section 3.3.1 y_3 corresponds to the titer, u is the input vector and $\bar{u}$ is the optimisation variable i.e. the inputs MPC computes, and sets $u[l] = \bar{u}[1]$, and u_{min} and u_{max} are the vectors corresponding to the lower and upper bounds respectively for the inputs (see Table 3.1). The bounds are kept commensurate to the usual practice in industry and hence feed rate, albeit important and strongly related to maximising product, has been given a smaller upper bound. An effort has been made to impute any increase in product to the other strongly related variables such as glucose by tuning the weight appropriately. P denotes the prediction horizon, Q is a negative value picked to appropriately weigh the product maximization in the objective function. R is a diagonal matrix with appropriate penalties for input change. S is a scalar weight to additionally penalise a positive change in pH. This term has been chosen to not be a quadratic term specifically to penalize only positive changes. Note that the implementation of a positive pH change is done via using a buffer, which 'shocks' the cells and is preferably avoided.

Table 3.1: Input Constraints

u	units	u_{min}	u_{max}	u_{nom}
Temp	degC	35	36.8	36.1
pH		6.95	7.15	7.1
Glucose Feed Conc.	mg/L	6	12	9
Feed Rate	vols/day	1	1.6	1.25
Bleed Rate	L/day	0.01	0.05	0.025

Remark 4. *Note that in this formulation the predicted outputs are determined using a state space model with a feedthrough term. When identifying a data driven model it has been shown that retaining the feedthrough term provides more accurate control comparing to dropping it, motivating the use of the recent MPC formulation in the present work [23].*

The elements of the matrix R in the term $\Delta u^T R \Delta u$ are taken as the inverse of the nominal value of that input variable to compensate for the different scales of the manipulated inputs. The value S is utilized to specifically penalize positive changes in the pH. To accomplish this we adjust the elements in R and S such that the increase to the objective due to pH change from term R is higher than the decrease due to term S when the pH decreases. In case of a candidate positive pH change, since R and S have positive weights, the pH terms add up from $(\Delta u^T)R(\Delta u)$ which is always positive due to being quadratic and $S\Delta u$ which is positive since Δu is positive. In the case of a negative pH change, the positive contribution from $(\Delta u^T)R(\Delta u)$ would outweigh the negative contribution from $S\Delta u$ for any reasonable change in pH when s is small. The value of s is taken as 0.05. The value of the R term corresponding to pH is $\frac{10}{7.1}$. For a Δu of -0.05, the $(\Delta u^T)R(\Delta u)$ term is +0.0035 and $S\Delta u$ term is -0.0025 resulting in a net +0.001 penalty. Thus, a change to the pH would be only made if it results in a net benefit to the product quality. On the other hand, for a candidate increase in pH, for a Δu of +0.05, the $(\Delta u^T)R(\Delta u)$ term is +0.0035 and

$S\Delta u$ term is +0.0025 resulting in a net +0.006 penalty, a six times higher penalty than a corresponding decrease.

The net affect of such a choice of the tuning parameters and the formulation is that any significant pH changes are penalized but increases in pH are penalised more than decreases in pH.

The MPC is initialized when the error between predicted outputs and observed outputs (using the Luenberger observer) becomes smaller than a user specified tolerance. The tolerance is chosen such that there is minimal plant-model mismatch but also enough time left to implement an MPC strategy. Before the state observer converges, a constant nominal input is applied to the process.

$$Q = q_1 \tag{3.9}$$

$$R = \begin{bmatrix} \frac{r}{u_{nom,1}} & 0 & 0 & 0 & 0 \\ 0 & \frac{r}{u_{nom,2}} & 0 & 0 & 0 \\ 0 & 0 & \frac{r}{10 \times u_{nom,3}} & 0 & 0 \\ 0 & 0 & 0 & \frac{r}{u_{nom,4}} & 0 \\ 0 & 0 & 0 & 0 & \frac{r}{u_{nom,5}} \end{bmatrix} \tag{3.10}$$

$$S = s \tag{3.11}$$

Table 3.2 reports q_1, r and s while u_{nom} values are reported in table 3.1. Since the inputs vary in orders of magnitude, the weights on input change penalty are also adjusted as such with their respective nominal values. The values of weights are chosen by the user based on requirements or objectives to be met and can be changed as needed. The specific values were weighted keeping in mind the presented objectives.

Table 3.2: Tuning parameters

q_1	r	s
-1	10	0.05

Remark 5. *The objective function in the MPC formulation focuses on maximising the final product which depends on both the titer (similar to output concentration) and the harvest rate (similar to output flow rate). The specific objective function can readily be altered. The key contribution of the present manuscript are not the input profiles that the controller implements, but to demonstrate that a data driven model based MPC, with a meaningful objective function, can be implemented on the system (the test bed in this case) and a biologically acceptable control action and system behavior achieved. This objective function can be further fine tuned or changed based on the specific needs of the process operation.*

Remark 6. *The measurements from biological data available from the lab are available once every 8 hours or 3 times per day. This has been dealt with in the present work by using state observers to predict the next state as well as the sampling time of 8 hours per day being sufficient due to the slow dynamics of the biological system which works well with changing the inputs less frequently. Explicit handling of the time delay between the drawing of the sample and the measurement being available would be considered in future work.*

3.5 Results and Discussion

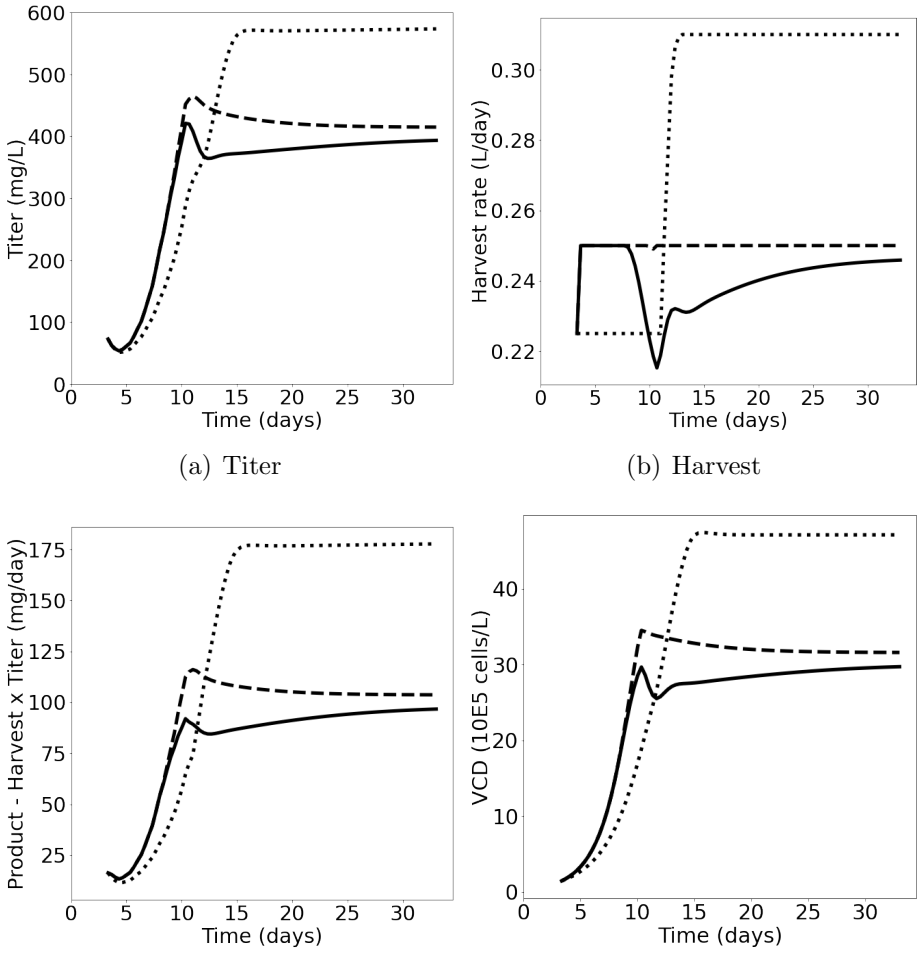
The first contribution of the work is to demonstrate the improved performance achievable using a data driven MPC implementation over the current industrial practice of

PI control. In current practice, a PI control is used with a fixed VCD setpoint of 30 which it is able to achieve at the twenty five day mark as shown in Figure 3.4. Under the PI control, the bleed rate is initially kept low, and as the VCD starts to peak, the bleed rate is increased in order to hold a VCD setpoint of 30 as seen in Figure 3.5. As expected the other variables are held at their nominal values since only one PI controller is used which is linked to the bleed rate. This controller reaches a final product of 97 mg/day. In contrast, the implementation of the MPC results in a VCD over 47 but more importantly, results in the production of 178 mg/day of product as mentioned in Table 3.3. This is due to the controller's ability to shift all of the input variables while utilizing an appropriately identified process aware model. Figure 3.5 shows that increasing temperature, decreasing pH, increasing glucose feed concentration and decreasing bleed rate leads to optimal bioreactor operation and a clearly superior control strategy. Yet another benefit of the MPC implementation is the ability to simply ask it to maximize the product (through the objective function) instead of specifying a set-point.

In contrast, if the PI implementation was used to arbitrarily increase the PI setpoint to 60 (in an effort to achieve comparable product) the set-point is not met (see Figure 3.5). This of course is due to the limited control action available to the PI (the bleed rate), which it does push to zero. While the resulting final product is slightly higher than the original PI implementation at 104mg/day, the increase is marginal. Of course, in practice, the bleed rate would never be set to zero because without removing any waste from the bioreactor the build up would lead to increased cell death (thus under MPC implementation, the bleed rate is not allowed to go below 0.01 L/day as reported mentioned in Table 3.1).

Table 3.3: Constrained Subspace MPC vs PI control

Case)	Improvement (%)
Current PI	0 (current)
Higher VCD setpoint PI	7.2
Constrained subspace MPC	83.5



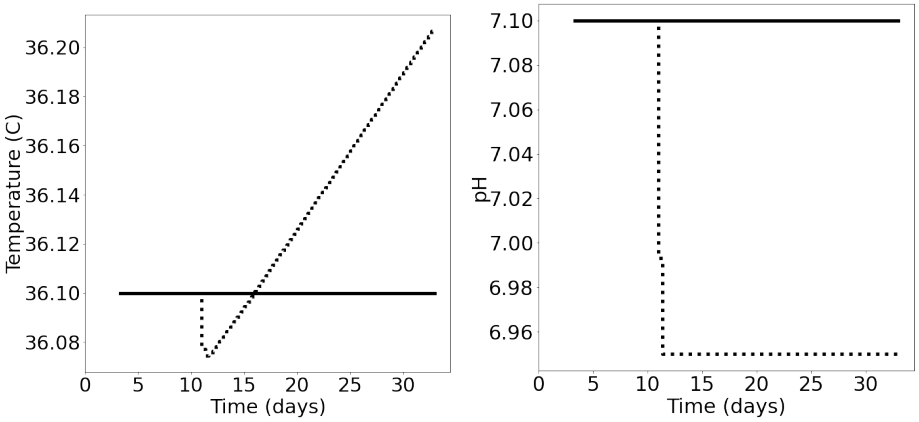
(a) Titer

(b) Harvest

(c) Product (Titer × Harvest)

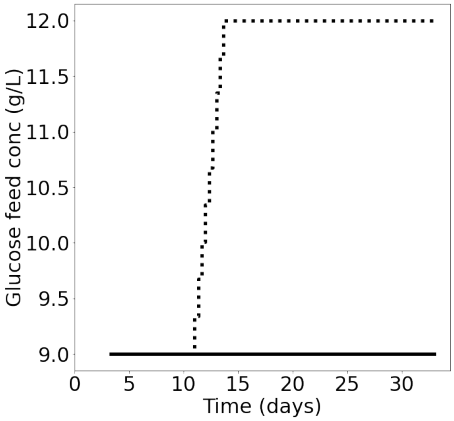
(d) Viable Cell Density (VCD)

Figure 3.4: Comparison of performance of the best MPC (dotted lines) with existing PI (solid lines) as well as PI with higher VCD setpoint (dashed lines).

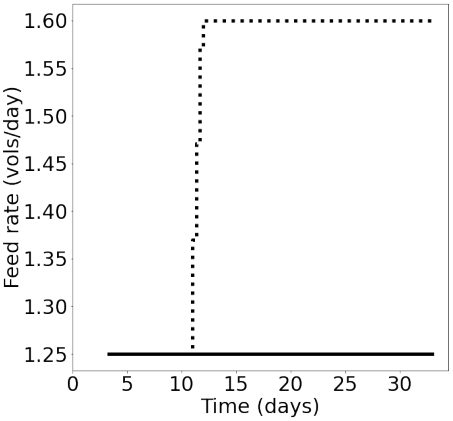


(a) Temperature

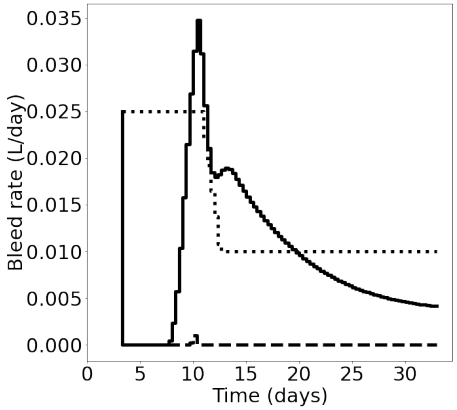
(b) pH



(c) Glucose feed concentration



(d) Feed rate



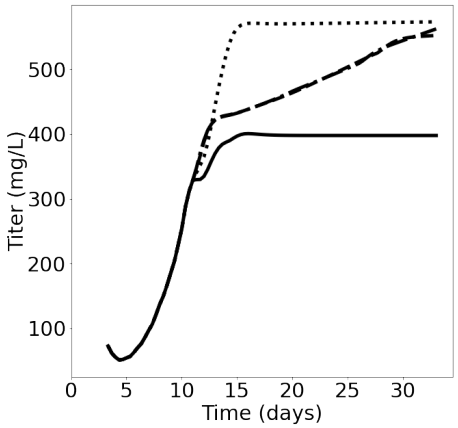
(e) Bleed rate

Figure 3.5: Comparison of inputs of the best MPC (dotted lines) with existing PI (solid lines) as well as PI with higher VCD setpoint (dashed lines).

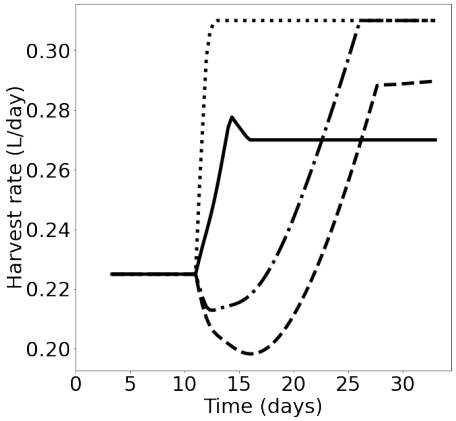
Remark 7. *The performance of PI shows higher outputs over the MPC in the first 10 days due to the PI being implemented from the beginning of the run whereas the MPC is only implemented after the convergence of the observer. This demonstrates the ability of the MPC to have a superior overall performance despite being utilized later in the run.*

The second objective of this work is to demonstrate the necessity of a process aware constrained subspace identification technique when identifying the plant model. The key advantage in utilizing the constrained model is that a traditional unconstrained model may have incorrect steady state process gains. The process awareness comes from constrained subspace identification approach applied biologically relevant knowledge in the model identification stage by having the correct and relevant signs in the steady state gain between inputs and outputs as constraints during identification of the system matrices. Figure 3.6 clearly shows the advantage of utilising the constrained subspace method compared to regular subspace identification (see Table 3.4). The MPC utilising the **u**nconstrained subspace model with **s**hort **h**orizon is referred to as USH in the table for brevity. Similarly, ULH, CSH and CLH represent unconstrained long horizon, constrained short horizon and constrained long horizon respectively. Note that, in short horizon control these differences aren't as noticeable especially in the titer concentration as evident in Figure 3.6. However, when longer control horizons are utilized, the unconstrained model MPC performance deteriorates due to the effect of wrong gain signs identified, causing it to move inputs in a wrong direction. This difference is highlighted in Figure 3.7 where the unconstrained MPC fails to decrease the bleed rate as it has the wrong sign in the process gain. The longer horizon unconstrained subspace MPC thus performs very slightly better than existing PI control while the constrained model far outperforms both. Longer control horizons lead to improved performance with the process aware (constrained subspace) MPC as the controller is able to optimize the input trajectory over a longer period. Not only does the long horizon constrained subspace based MPC get the highest final

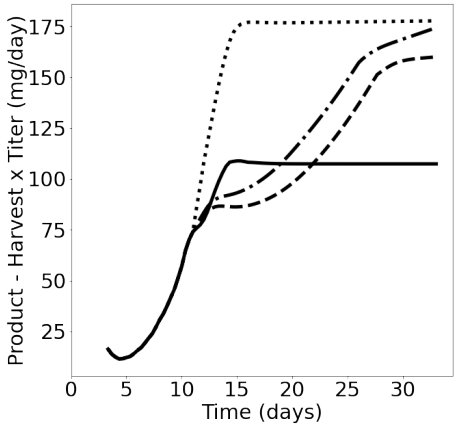
product, it achieves the higher product and higher concentration both much earlier whereas the shorter horizon approaches only are able to reach values towards the end, leading to a significantly lower cumulative product compared to the longer horizon constrained MPC as shown in Table 3.5.



(a) Titer



(b) Harvest



(c) Product (Titer × Harvest)

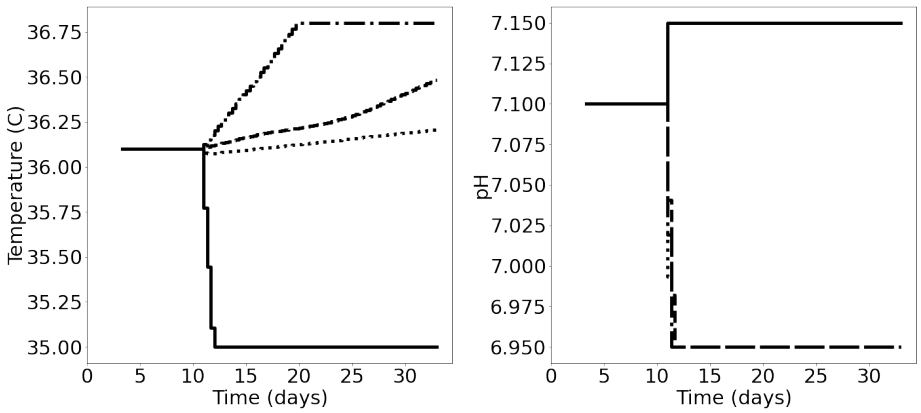
Figure 3.6: Comparison of performance of the best case i.e.longer horizon constrained subspace MPC (dotted) with MPCs based on shorter horizon constrained subspace (dash-dotted), longer horizon unconstrained i.e. regular subspace (solid) and shorter horizon unconstrained i.e. regular subspace (dashed).

Table 3.4: Unconstrained Subspace MPC vs Constrained Subspace MPC - Final Product

Case	Final Product (mg/day)	Improvement over current PI (%)	Improvement over USH MPC (%)
USH	160	65	0
ULH	107.4	10.7	-33
CSH	174	79.4	9
CLH	178	83.5	11.3

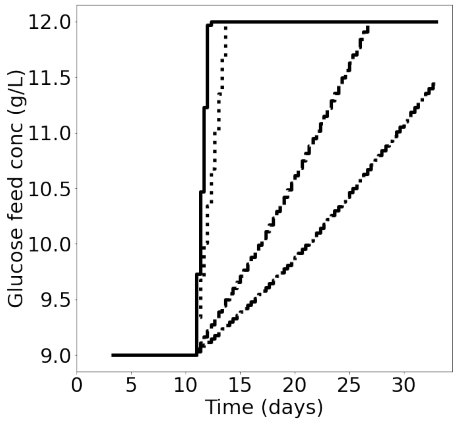
Table 3.5: Unconstrained Subspace MPC vs Constrained Subspace MPC - Average Product

Case	Average Product (mg/day)	Improvement over current PI (%)	Improvement over USH MPC (%)
USH	94	18.2	0
ULH	85.3	7.3	-9.3
CSH	103.2	29.8	9.8
CLH	132.8	67	41.3

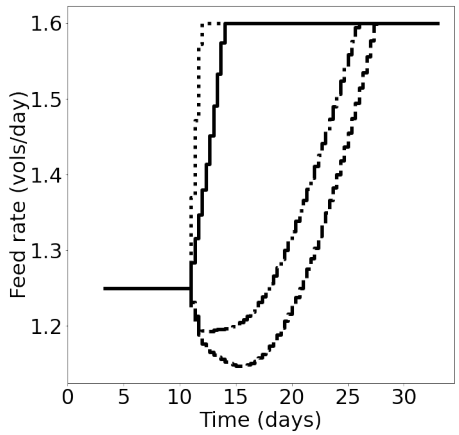


(a) Temperature

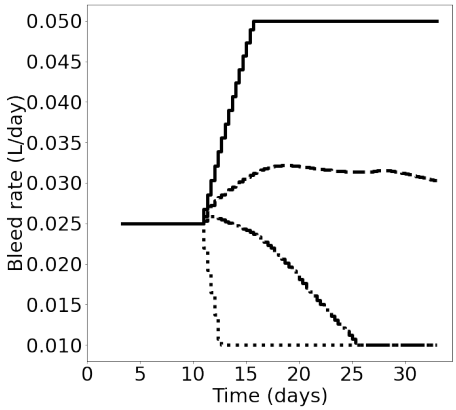
(b) pH



(c) Glucose feed concentration



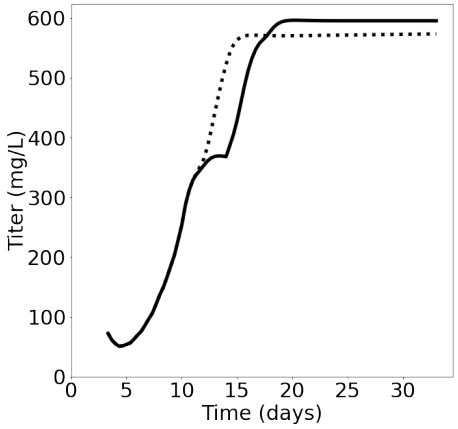
(d) Feed rate



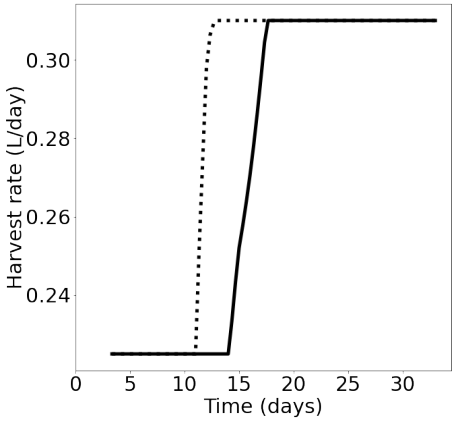
(e) Bleed rate

Figure 3.7: Comparison of inputs of the best case i.e.longer horizon constrained subspace MPC (dotted) with MPCs based on shorter horizon constrained subspace (dash-dotted), longer horizon unconstrained i.e. regular subspace (solid) and shorter horizon unconstrained i.e. regular subspace (dashed).

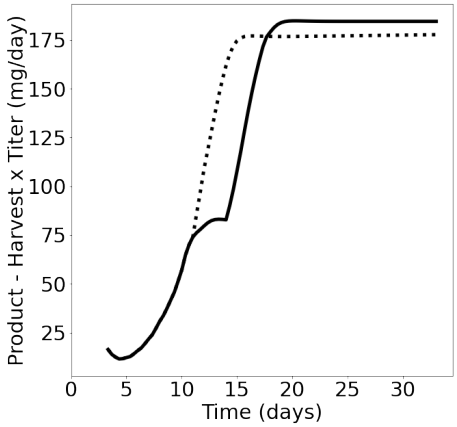
The final contribution of this work is to show the robustness of model predictive control to maximize the final product. In order to test the robustness of the controller the MPC using the constrained subspace model is compared against a new bioreactor process. In the new bioreactor the parameters such as growth rates and death rates are now different than the training data used to identify the constrained subspace model. This creates additional plant model mismatch and represents scenarios where the reactor may be processing new batches of cells. Figure 3.8 shows how the constrained MPC is able achieve a similar final product. When comparing the input changes made by the MPC, Figure 3.9 shows that the constrained MPC makes similar input moves in both the constrained model built on new or current data as well as a constrained model which was built on data from an older system. The MPC utilising the constrained subspace model built on old plant data is also able to achieve a high final product though at a cost of higher temperatures which is not very desirable. But overall, the control performance remains acceptable when using the constrained subspace MPC on a different system demonstrating robustness



(a) Titer

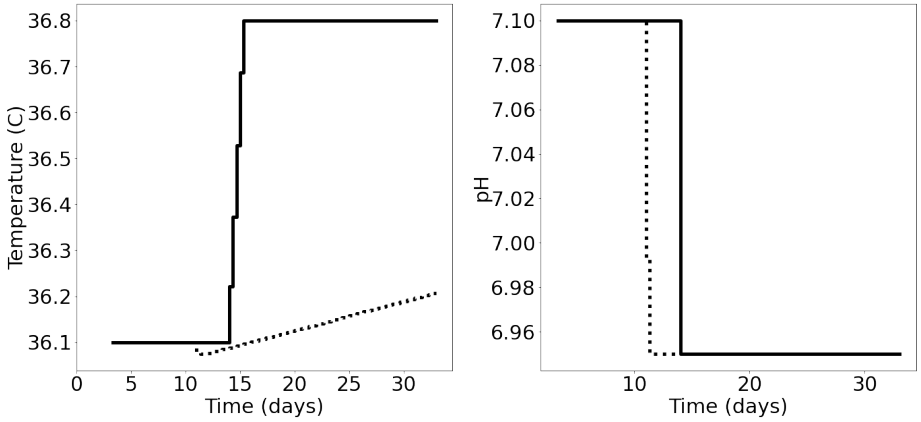


(b) Harvest



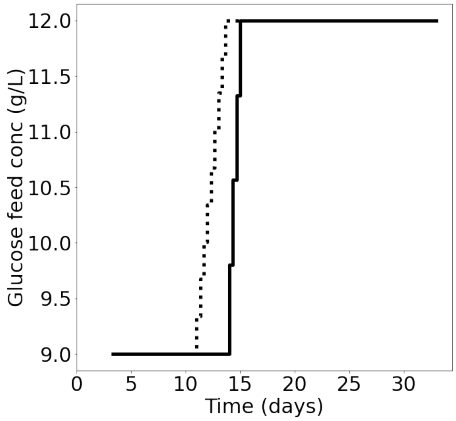
(c) Product (Titer × Harvest)

Figure 3.8: Comparison of performance of the constrained subspace MPC trained on old model plant system (solid) with performance of constrained subspace MPC trained on current model plant system (dotted) to demonstrate robustness.

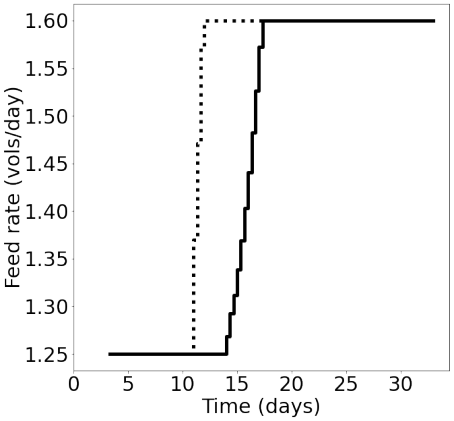


(a) Temperature

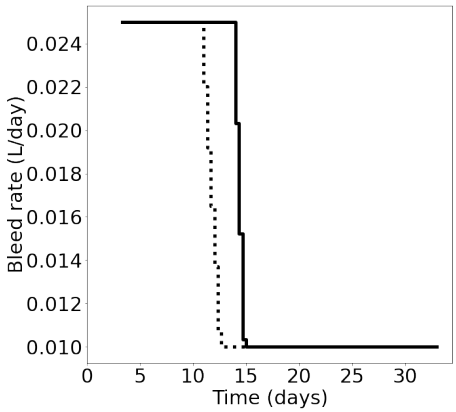
(b) pH



(c) Glucose feed concentration



(d) Feed rate



(e) Bleed rate

Figure 3.9: Comparison of inputs of the constrained subspace MPC trained on old model plant system (solid) with performance of constrained subspace MPC trained on current model plant system (dotted).

3.6 Conclusions

The present manuscript demonstrated the possibility of using a process aware data driven model predictive control scheme for bioreactors to enable performance improvement compared to industry standard of proportional-integral controller schemes. The importance of using a process aware model within model predictive control schemes was illustrated by comparing subspace model with process knowledge based constraints to standard subspace model based MPC implementation. Finally, the ability of the MPC to handle process changes was illustrated, with the MPC performance continuing to be acceptable under process changes.

3.7 Acknowledgment

Financial support from Sartorius and the McMaster Advanced Control Consortium is gratefully acknowledged.

Bibliography

[1] Bernard, O., Mairet, F., and Chachuat, B. (2015). Modelling of microalgae culture systems with applications to control and optimization. In *Microalgae Biotechnology*, pages 59–87. Springer.

[2] Caramihai, M. and Severin, I. (2013). Bioprocess modeling and control. *Biomass Now: Sustainable Growth and Use*, page 147.

[3] Chang, L., Liu, X., and Henson, M. A. (2016). Nonlinear model predictive control of fed-batch fermentations using dynamic flux balance models. *Journal of Process Control*, 42:137–149.

[4] Chusainow, J., Yang, Y. S., Yeo, J. H., Toh, P. C., Asvadi, P., Wong, N. S., and Yap, M. G. (2009). A study of monoclonal antibody-producing cho cell lines: What makes a stable high producer? *Biotechnology and bioengineering*, 102(4):1182–1196.

[Corbett and Mhaskar] Corbett, B. and Mhaskar, P. Subspace identification for data-driven modeling and quality control of batch processes. *AIChE Journal*, 62(5):1581–1601.

[6] Del Rio-Chanona, E. A., Cong, X., Bradford, E., Zhang, D., and Jing, K. (2019). Review of advanced physical and data-driven models for dynamic bioprocess simulation: Case study of algae–bacteria consortium wastewater treatment. *Biotechnology and bioengineering*, 116(2):342–353.

[7] Deschenes, J.-S., Desbiens, A., Perrier, M., and Kamen, A. (2006). Multivariable nonlinear control of biomass and metabolite concentrations in a high-cell-density perfusion bioreactor. *Industrial & engineering chemistry research*, 45(26):8985–8997.

[8] Dochain, D. and Perrier, M. (1997). Dynamical modelling, analysis, monitoring and control design for nonlinear bioprocesses. In *Biotreatment, Downstream Processing and Modelling*, pages 147–197. Springer.

[9] Ferkl, L. and Širokỳ, J. (2010). Ceiling radiant cooling: Comparison of armax and subspace identification modelling methods. *Building and Environment*, 45(1):205–212.

[10] Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89(425):122–127.

[11] Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17.

[12] Gevers, M. (2005). Identification for control: From the early achievements to the revival of experiment design. *European journal of control*, 11(4-5):335–352.

[13] Ghosh, D., Hermonat, E., Mhaskar, P., Snowling, S., and Goel, R. (2019). Hybrid modeling approach integrating first-principles models with subspace identification. *Industrial & Engineering Chemistry Research*, 58(30):13533–13543.

[14] Karra, S., Sager, B., and Karim, M. N. (2010). Multi-scale modeling of heterogeneities in mammalian cell culture processes. *Industrial & Engineering Chemistry Research*, 49(17):7990–8006.

[15] Lee, J. H. (2011). Model predictive control: Review of the three decades of development. *International Journal of Control, Automation and Systems*, 9(3):415–424.

[16] Leib, T. M., Pereira, C. J., and Villadsen, J. (2001). Bioreactors: a chemical engineering perspective. *Chemical engineering science*, 56(19):5485–5497.

[17] Ljung, L. (1998). *System Identification: Theory for the User*. Pearson Education.

[18] Luenberger, D. (1971). An introduction to observers. *IEEE Transactions on automatic control*, 16(6):596–602.

[19] Mairet, F., Bernard, O., Cameron, E., Ras, M., Lardon, L., Steyer, J.-P., and Chachuat, B. (2012). Three-reaction model for the anaerobic digestion of microalgae. *Biotechnology and Bioengineering*, 109(2):415–425.

[20] Moonen, M., De Moor, B., Vandenberghe, L., and Vandewalle, J. (1989). On-and off-line identification of linear state-space models. *International Journal of Control*, 49(1):219–232.

[21] Morari, M. and Lee, J. H. (1999). Model predictive control: past, present and future. *Computers & Chemical Engineering*, 23(4-5):667–682.

[22] Morel, E., Tartakovsky, B., Guiot, S., and Perrier, M. (2006). Design of a multimodel observer-based estimator for anaerobic reactor monitoring. *Computers & chemical engineering*, 31(2):78–85.

[23] Patel, N., Corbett, B., and Mhaskar, P. (2021). Model predictive control using subspace model identification. *Computers & Chemical Engineering*, 149:107276.

[24] Patel, N., Corbett, B., Trygg, J., McCready, C., and Mhaskar, P. (2020a). Subspace based model identification for an industrial bioreactor: Handling infrequent sampling using missing data algorithms. *Processes*, 8(12):1686.

[25] Patel, N., Nease, J., Aumi, S., Ewaschuk, C., Luo, J., and Mhaskar, P. (2020b). Integrating data-driven modeling with first-principles knowledge. *Industrial & Engineering Chemistry Research*, 59(11):5103–5113.

[26] Pörtner, R., Barradas, O. P., Frahm, B., and Hass, V. C. (2017). Advanced process and control strategies for bioreactors. In *Current Developments in Biotechnology and Bioengineering*, pages 463–493. Elsevier.

[27] Seborg, D. E., Mellichamp, D. A., Edgar, T. F., and Doyle III, F. J. (2010). *Process dynamics and control*. John Wiley & Sons.

[28] Simutis, R. and Lübbert, A. (2015). Bioreactor control improves bioprocess performance. *Biotechnology journal*, 10(8):1115–1130.

[29] Sirois, J., Perrier, M., and Archambault, J. (2000). Development of a two-step segregated model for the optimization of plant cell growth. *Control Engineering Practice*, 8(7):813–820.

[30] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

[31] Xie, L. and Wang, D. I. (1996). High cell density and high monoclonal antibody production through medium design and rational control in a bioreactor. *Biotechnology and bioengineering*, 51(6):725–729.

[32] Zhang, D., Del Rio-Chanona, E. A., Petsagkourakis, P., and Wagner, J. (2019). Hybrid physics-based and data-driven modeling for bioprocess online simulation and optimization. *Biotechnology and bioengineering*, 116(11):2919–2930.

Chapter 4

Conclusions and Recommendations

Biopharmaceutical processes especially monoclonal antibodies (mAbs) have been an active area of interest in recent decades and have also been the subject of Nobel Prizes in Physiology/Medicine in 1984 and 2018. Industry trends are moving towards perfusion or continuous processes owing to the significantly increased productivity and reduced manufacturing footprint for the same amount of product. With it's increased process run length, maintaining desired input and output conditions would require advanced control strategies which are not as well adapted in the biopharma industry as traditional process industry due to a combination of complicated processes affecting first principles modeling and limited data and limitations on input perturbations for identification. Chapter 2 focused on an intensified DOE approach to identify an appropriate input perturbation mechanism that respected biological requirements and feasibility depending on process run duration and dynamics. The presented result of once per three days was chosen for this process, but the primary contribution of the manuscript is the methodology and approach used to choose an appropriate data generation method that can be implemented in combination with a data-driven method such as SSID to build a control-relevant model that can be extended to any similar bioprocess. The next stage of work was to develop a model predictive controller (MPC) to meet various objectives and respect biological input constraints. This was covered in chapter 3. The chapter also covered an enhancement to the subspace-based modelling by incorporating first principles knowledge through the signs of the long-term gain matrix that is added as a constraint in the identification step, allowing the causal effect of a specific input to a specific output to be respected. This approach also allowed robustness; for example, including the knowledge of the process gains allowed reasonable performance even when the training data was based on a different cell line with different growth and death rates.

4.1 Recent and Upcoming Work

This section presents work done recently in collaboration with summer students Nigel Mathias and Sarah Rasmussen, also from the Mhaskar Group, under my supervision.

The first area is using Bayesian Inference through methods such as nested sampling to allow estimation of biological parameters such as growth and death rates directly from input-output data. A parameterised neural network is built and trained offline and then used in a nested sampling approach to sample different values of the parameters and iteratively improve on estimation of the critical parameters purely from data, giving insights into the cell culture, saving time and money on intensive laboratory experiments. This will be expanded to involve parallel estimation of multiple parameters.

The second area is using Physics Informed Neural Networks (PINNs) to provide a robust solution including both data and first principles knowledge. This also is trained offline and the first principles knowledge is incorporated through the loss function which now penalises not just the deviation of prediction from process data but also the gradients from the gradients in the first principles ODEs. This approach would be further augmented by introducing mismatch in the parameters in the ODE known to the neural network and actual system and the combination of loss terms from data and first principles would potentially still be able to predict with reasonable accuracy since without model plant mismatch the physics based loss would help guide the solution to a better accuracy and with model plant mismatch the input-output data based loss would be able to able to bring the solution to the correct values.

The control solution for the bioreactor did not use neural networks initially as we wanted to use a method being capable of accurate modeling and advanced process control from limited data such as from actual plants. The neural network here is built

offline from the simulator and the purpose of Bayesian inference and physics informed neural networks is to provide a high fidelity alternative to the advanced simulator which can be run in under a second to give an entire output trajectory unlike the simulator which takes a few minutes. A large number of runs of this fast model can then be run for parameter estimation of nonlinear models like biological parameters such as maximum growth rate. Further, physics informed neural networks also help prevent problems with overfitting by virtue of having knowledge of the first principles equations.

4.2 Future Work

This section presents recommendations for further development and exploration of process improvement for such processes, focusing on machine learning. This section discusses recommendations for future areas. Section 4.1 covered present and upcoming near future work in detail and this section briefly expands upon the same with other future possibilities. The machine learning methods have the potential to be expanded to include parallel parameter estimation for multiple parameters and to include model plant mismatch in building the physics informed model. The bioreactor used in the work had the initial three days in inoculation mode where no control action was made. Future work would involve including the three day inoculation into the state estimation and potentially control which would especially be useful for early detection of batch quality problems.

The core of the thesis was to utilize subspace based methods to develop process models and use them in control strategies. These can be expanded into further data driven and hybrid models like incorporating them with PLS and PCA for monitoring and potentially using a linear simplified model for the data loss in the physics informed neural network to allow building models with limited data which traditional neural

networks used in data loss struggle with.