Development and software implementation of modelling tools for rapid fermentation process development using a parallel mini-bioreactor system

Aarron Erbas

UCL

EngD Biochemical Engineering

Supervisors: Dr. Frank Baganz
Advisor: Dr. Tony Allman (Infors HT)

University College London
Department of Biochemical Engineering

Date of Submission: Dec 2017

Declaration

By submitting this coursework, I Aarron Erbas confirm that the work is my own and that any material derived or quoted from the published or unpublished work of other persons has been duly acknowledged. I confirm that I have read the UCL and Departmental guidance on plagiarism.

Acknowledgements

Abstract

In order to establish a generic framework for the rapid development and optimisation of scalable fermentation processes, a novel methodology for simplifying model building was explored. This approach integrates small-scale fermentations with model-based experimental design (DoE) and predictive control strategies. In this study, four 1.4 litre vessels were characterised for power input, volumetric oxygen transfer coefficient ($K_L a$) and mixing, to assess its potential for replicating cell culture rapidly. Engineering characterisation results showed excellent propeller operation over a range of 400-1200 rpm and up to the maximum motor output and under various air flow rates in fluid densities up to 4.21 Cp/mPa s (1.211 g/cm^3). Limits were reached using glycerol (99%) at fluid viscosities of 500Cp/mPa s (1.253g/cm^3) at 800 rpm and no air flow, hence experiencing the most resistance. This was the most taxing condition in terms of energy input into the system. Furthermore, we determined the efficient gas dispersion which is considered important for oxygen bubble dispersion in viscous fluids. The potential gas dispersion could be calculated as a function of both impeller speed, airflow rate, and the fluid viscosity. The calculations provided a working impeller speed of >263 rpm for >0.5 vvm air flow rate as preliminary parameters in our advanced modelling section. The key outcome of the $K_L a$ study was that the results showed suitable potential for mass transfer for high cell density fermentations, for each of the parallel stirred tank bioreactors. To assess the usability of the parallel bioreactors be used for bioprocess rapid development purposes *Escherichia coli* W3110 was characterised in the 1L WV vessels. So overall the experiments included testing the performance of the vessels engineering parameters and also the biological fermentations confirming that the system was suitable for parallel operation with high reproducibility.

For model building, especially suited for the 4-reactor set up the parallel bioreactors a fractional factorial design was used, in which models could be rapidly built and implemented for further research. The screening and model

optimisation helped to reduce the development time by using the parallel equipment. Batches of four reactors could be completed in parallel in which comparable experimental results were obtained rapidly for new fermentation models.

Optical density measurements provided a quick off-line analysis of the growth curve of microbial populations, as compared to cell plate counts or dry weights that require more time. For the model development and the establishment of our integrated software modelling tool, a modified logistic model was developed to predict microbial growth kinetics. First-order kinetic models, logistic, and Gompertz models were used and comparatively analysed to assess the model fit to test batch data. The logistic model was favourable for mapping and simulating the later phases of bacterial growth, while the well-established exponential growth model predicted the early lag phase in our stoichiometric growth simulation software tool better.

The initialisation of the previous fermentation model allowed us to build a statistical model, which was based on the engineering characteristics for optimisation of biomass. Therefore, batch nutrient supply with the aid of stoichiometric models could be tested and modelled. DoE model data was improved with metabolic flux analysis to develop an advanced feeding strategy by testing various metabolic pathways and the nutrients used in experimentation.

Bacterial growth predictions and media optimisation were tested for maximising microbial biomass yields. We then modelled the dissolved oxygen concentration and substrate utilisation. The techniques and principles of dynamic flux balance analysis, mechanistic modelling, and stoichiometric mass balancing were used. The aim was to create and validate our integrated software based on advanced modelling for the parallel bioreactor systems and tested through application for *E. coli* fermentations. Optimising microbial biomass was the main target in this project, with the data collected from

fermentation being the strongest comparator and validator. A new software for the integration of DoE and Dynamic flux balance analysis (DFBA) techniques with the intention of creating a working fermentation platform for the Multifors equipment via simulation and fermentation optimisation was the novel outcome of this research. The tool could provide functions for speeding up development time and control of parallel bioreactors.

Table of Contents

1 Introduction

1.1 Project Motivation

Miniature bioreactors have been receiving increasing attention because of their scale-up potential and volumetric advantages in experimentation and testing of suspended microbial cultures. Parallel miniature bioreactors have further advantages when applied to model-based experimental design (DoE) and rapid research completion. Several studies on oxygen transfer and power input have also been conducted for microbial fermentation processes (Gill et al., 2007; 2008; Betts et al., 2006). Previous research using miniature bioreactors focussed on the following key engineering parameters:

- Quantification and modelling of oxygen mass transfer rates (Doig et al., 2005)
- Impact of oxygen mass transfer rates on process performance (Doig et al., 2002; Elmahdi et al., 2003; Ferreira-Torres et al., 2005).
- Liquid phase mixing time (Micheletti et al., 2006).
- Energy dissipation rates (Micheletti et al., 2006).

The aim was to use the engineering principles established in previous research to identify the specific characteristics of a 4 x 1L parallel bioreactor system (Multifors). Thereafter, utilise the characteristics in two applications to explore methodologies that integrate small-scale microbial cell fermentations) with model-based experimental design and predictive control strategies. This approach will be used to optimise biomass production of *E.coli* grown in a defined medium.

The objectives of the project are outlined below:

1) Engineering characterisation of Multifors reactors
2) Rapid process optimisation using model-based DoE approaches
3) Integration of advanced modelling tools with Infors reactor control software (IRIS) and model development
4) Establishment of model-based control strategies and optimisation of fermentation protocols

Prior to this project, there was a need to restructure and develop a platform for bioprocess modelling. The ideal platform is one that can integrate the mathematical, computational, and biological fields. This cross-over is required to, for instance, integrate multidisciplinary bioinformatics resources with lab-scale miniature parallel bioreactors and control software. This project also needs to answer an important research question: "Is there any economic potential or impact?" Addressing this commercial impact is valid for engineering but not within the main scope of this research.

Lastly, the validation, critique, and review of the protocols for this project from an engineering perspective will improve the structure and strength of the research outcomes. Figure 1 below shows a diagrammatic overview of research needs as individual components:
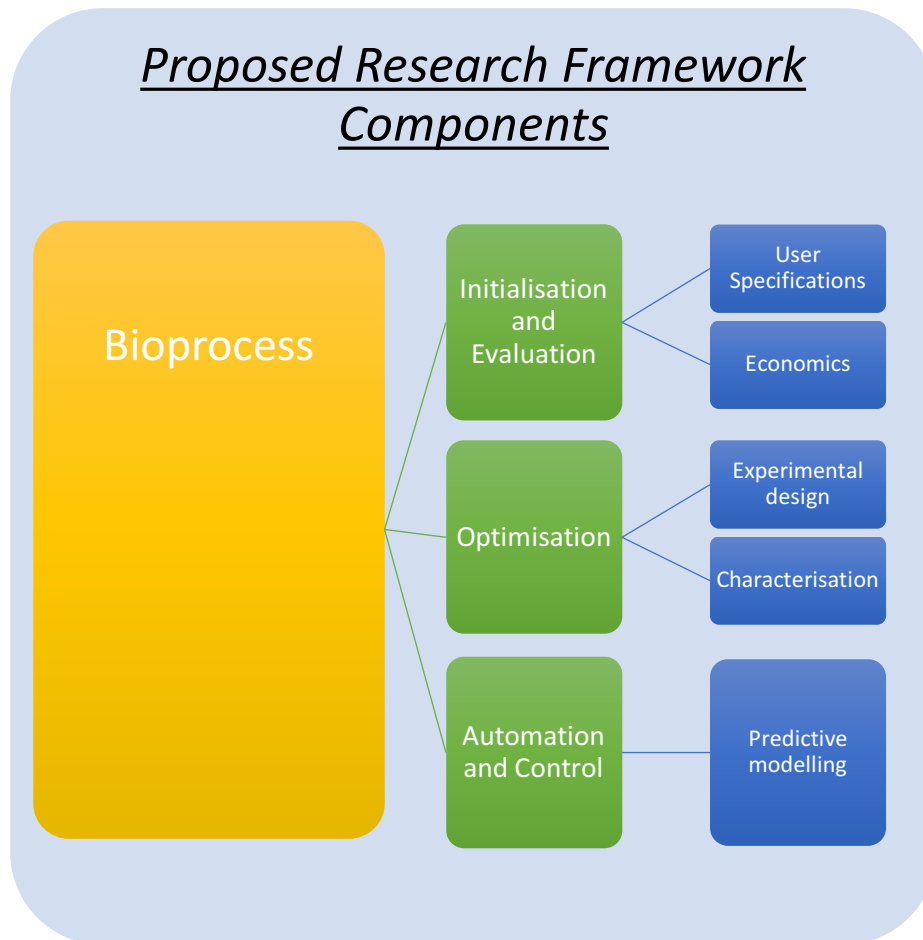
Figure 1 Proposed framework and constituent topics on which this research project is built upon

The diagram in Figure 1 indicates the framework for this research project, Evaluation of current software and hardware limitations, where applications can be optimised and lastly the current software and control can be enhanced. Research problems like those mentioned earlier improve bioreactor functionality, so Infors HT can target a wide audience. This diagram is also applicable to bioprocesses at various scales.

As there was a need to establish a robust framework for rapid fermentation process optimisation, this project used a small-scale parallel bioreactors

(Multifors) in combination with DoE and advanced predictive modelling approaches.

New process automations need to be validated and verified. Predictive extrapolation is a function of building models that mimics process data. Furthermore, this can also help control process parameters based on process trajectory and mathematical algorithms. However, each type of automation is unique to a process and can go in either of the two directions given below:

1) Towards a simplified or a generalised model-based design (black box design).

2) Towards the incorporation of different sets of information for complex modelling to gain a better insight into the process (as many examples of complex model designs exist).

Even though the ideas and field are well established, it is difficult to get useful information to construct online robust models for model predictive control. Therefore, bioprocess automation is arguably more challenging than for many other technical industries and it is difficult to carry out programmatically without accurate and validated information (Chang, Liu & Henson, 2016).

Implementing model predictive control is also a challenging factor for scale-up. Creating calibrated models for a range of bioreactor scales can lead to a cost benefit when predicting fermentations at scale, and if models are relevant and predictive, they may reduce the risk of product loss or batch failure with the use of better integrated controls.

This project attempts to overcome the difficulties faced with both modelling directions by looking at the implementation of generalised stoichiometric model and by advancing predictive techniques into a platform software. The project considers the engineering factors by defining appropriate engineering

bases for the rapid and accurate reproduction of cell growth and product kinetics.

Another goal was to elucidate the other problems faced at bench-top scale, namely high-density microbial growth, optimising product kinetics, and predicting the process trajectory with models created for defined microbial systems. For scale-up, closing the gap between miniature and larger scale is highly sought after, and this platform allows exploration by eventually implementing finite model predictive controls, process optimisation, and for future work scale translation.

1.1.1 Research aims and objectives

The initial project goals are to improve controls for suspended *E. coli* culturing and identify more effective performance solutions or *in silico* testing of metabolic models. There is a potential to work with engineered cells or various microorganisms in fermentations.
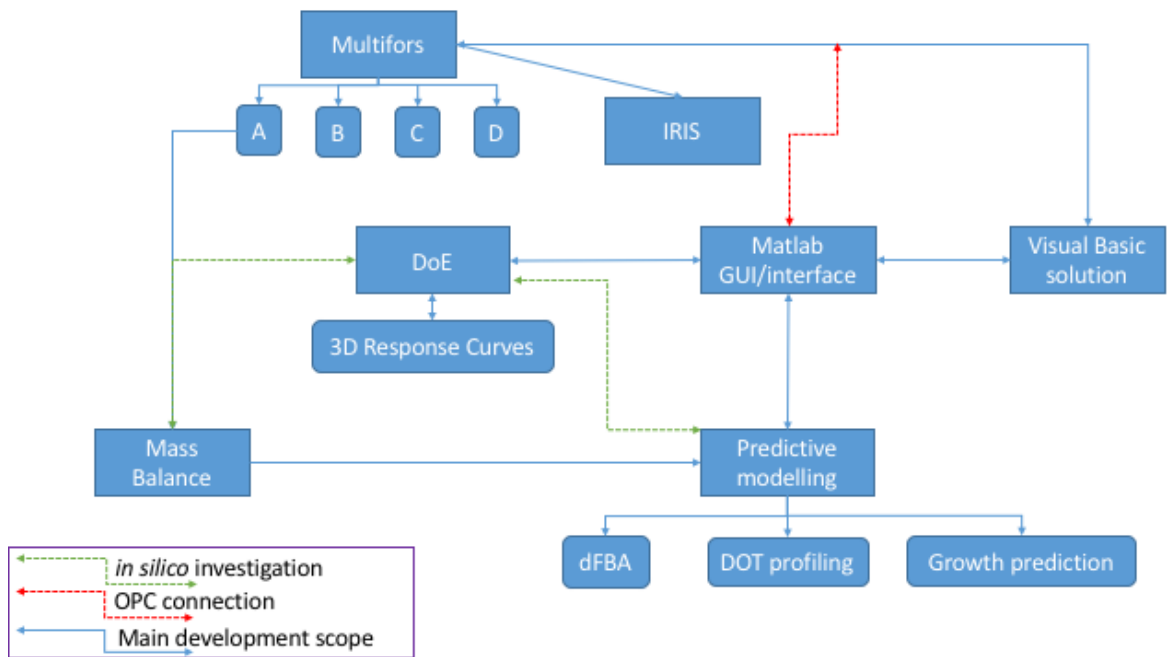
Figure 2 Overview of the project scope. This diagram follows the development of the integrated software and applications in this project (blue square boxes). The Multifors parallel bioreactor and control software IRIS was all that was established prior to this project.

This project will help to improve or help elucidate:

- Complex metabolic models
- Data visualisation
- Use of experimental design packages with reactor control software
- Using bench top reactors in parallel systems for exploring improvements in research capacity
- Reactor characterisation techniques to gain an advanced understanding of reactor performance
- Current operation of the Multifors reactors by automatically controlling parameters while aiming to improve interoperability between software programmes and reactor hardware with new interfaces, advanced bioprocess software, and integration
- Validation and data verification for software used with bioreactors and upstream applications

The objectives of the project are outlined below:

- Engineering characterisation of Multifors reactors
- Rapid process optimisation using DoE approaches
- Integration of advanced modelling tools with IRIS software and model development
- Establishment of model-based control strategies and optimisation of fermentation protocols
- Implement a Quality by Design approach with respect the software development and fermentation development as described in 6.1.1.

1.2 Thesis Structure

The thesis is structured as follows:

In chapter 1, the initial research problem is stated and the project motivation and relevance of the research questions covering academic, economic, and societal impacts are also established. In chapter 2, the materials and methods used for the entire project are described.

For the following results chapters, each results chapter (3-5) starts with the purpose and context within the scope of the project problem and/or objective being met. The principles, scientific theory, and other relevant information are included in each research section with further work being summarised

In chapter 3, the fundamentals of research experimentation are briefly reintroduced for engineering characterisation. These experiments were based on methods developed at UCL (e.g. Gill et al., 2008a). These established methods are then applied in practice to the Multifors parallel 1L stirred tank reactors. The Multifors reactor will be characterised in terms of power input, mixing times, and oxygen mass transfer and checked for consistency across

batch fermentations in the parallel bioreactor system. This is a crucial step as it tests modelling assumptions and establishes the parameters used in the subsequent sections. These are described in each section.

In chapter 4 these present our integrated software development and experimental testing, to build the initial fermentation model. It also includes a model-based DoE tool for the Multifors system and an appraisal for integrating statistical software packages for bioreactors.

In chapter 5, here the focus is on building upon the gained model knowledge and modelling methodologies for fermentation set up, operations, and outcomes in one integrated software tool. The protocols will be revisited and updated with the improvements from this model development.

Lastly, the research implementation and its impact are briefly investigated. Validation and commercialisation analysis of the new technology are also presented, covering industrial learning outcomes with future work suitable from this research.

2 Materials and Methods

2.1 Materials

All chemicals were obtained from Sigma Chemical Co, Ltd. (Dorset, UK) and was of analytical grade. Reverse osmosis (RO) water was used throughout.

2.2 Microbial Strain

Cell banks of *Escherichia coli* w3110, a laboratory strain of *E. coli* K-12, were prepared using two shake-flask cultures. Lysogeny broth (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl) was autoclaved at 121°C. Shake-flask cultures were incubated overnight at 37°C and 200 rpm on a shaking incubator (Manufacturer details). They were then aseptically transferred to sterile 1-mL Eppendorf tubes (0.5 mL culture aliquots with sterile 0.5 mL 50% glycerol solution) for freezing at -80°C.

2.3 Low-density Fermentation Medium

1 mL aliquots from the cell bank were thawed for seed culture/train in a shake flask culture (100 mL) for the final mini bioreactor 500-mL working volumes.

2.3.1 Media Composition and Preparation

Minimal media (M9) was used for low cell-density fermentations. For the 1L M9 media fermentation – Firstly 200ml of a sterilized M9 salt stock solution (consisting of reverse osmosis (RO) water, ammonium chloride (final concentration 1 g/L), disodium hydrogen phosphate (final concentration 6 g/L), potassium dihydrogen phosphate (final concentration 3 g/L), sodium chloride (final concentration 0.5 g/L) was added. Finally, 1% (v/v) 1M magnesium sulphate solution, and final 0.1% (v/v) 1M calcium chloride solution and more RO water to make the volume up to 1L. The pH was not adjusted for the seed train.

2.3.2 Fermentation set up and operation

Fermentation set up and operation was carried out according to the standard operating procedures provided in the appendix section 8.1. Rotation speed, air flow via feedback dissolved oxygen (DO) measurement using a polarographic electrode (manufacturer), pH using a pH probe (Details) and antifoam (PPG) were controlled by the Infors IRIS software, linked via a OPC connection to the Multifors touch screen control panel. The maximum speed that could be attained when working volume of media inside was 1200rpm. Where otherwise stated, pH was controlled to 7.0±0.05 by controlled addition of alkali and acid.

2.3.3 Optical density measurements and maximum specific growth rate

Optical density OD was measured off-line using a spectrophotometer (Thermo-Fisher scientific, USA) at 600 nm throughout the fermentation. Exponential regression was used in Microsoft Excel to determine the maximum specific growth rate from the OD data.

2.4 Power Measurement

Both the un-gassed and gassed torque/power measurements from the 1.4-L miniature bioreactor motor were carried out using a dynamometer located on the upper quartile section of the propeller shaft. Two Rushton turbine propellers were used in this experiment and the diameter of the glass vessel was d_t = 90 mm. The fill volume for each of the different tests was 1000 mL, carried out at 30°C. Glycerol was used to provide a range of viscosities in which motor torque was measured for each liquid medium.

To determine the power (W):

$$P(W) = Torque(Nm)* Speed(rpm)/9.5488$$

Equation 1

The gassed power was measured using the same liquid with air being sparged through the ring sparger.

2.5 K_La Measurements

The static dynamic gassing out method was utilised with an air and deionised water system (Stanbury & Whitaker, 1995). The probe response time was defined as the time taken to reach from 100% to 36.8% in a step-change from an oxygen-saturated environment to an unsaturated oxygen environment using nitrogen gas.

The DO data gathered from the IRIS software were outputted to Microsoft Excel and were corrected using the solver in Microsoft Excel by minimising the mean residuals between the square error of the raw data and corrected results (MSE), to produce k_La. The conditions tested for Van't Riet equation were for a range of aeration (vvm) and agitation rates (rpm) (Van't Riet, 1979; Stanbury and Whitaker, 1995):

$$k_L a = k.(Pg/V)^\alpha . (v_s)^\beta$$

Equation 2

2.6 Mixing Time

pH method:

The experiments were carried out at 20°C in a 1.4-L bioreactor with a 600 mL fill volume containing a Mettler Toledo pH probe for localised reading. The probe was submerged from the top plate. 95% homogeneity values were used for the calculation of the significant change in pH. The mixing time was

noted using a stop watch, from when the pH rose three levels from pH 4 to pH 7 using 1-M NaOH and to readjust the pH back down to pH 4 and below by adding 2-M phosphoric acid at a volume of 3 mL. The stirrer speeds for the local pH method were chosen to be 200 and 600 rpm and across aeration values of 0–1.5 vvm.

Decolourisation method:

Bromocresol purple pH indicator for global mixing method was added at 0.005 g per 600 mL fill volume in the reactor vessel. The measurements were again visually timed for a change in the colour. Bromocresol purple is purple at a range of pH 7 or above and it decolourises to yellow at ranges of 5.4 and below after addition of 1.5 mL of 3-M phosphoric acid. The pH was adjusted again to pH 7 with 1-M NaOH and the acid addition repeated (Bryant, 1977). The additions were made at the same height as the top plate to the working volume.

2.6.1 Logistic Model and Data Fitting

The mean squared error (MSE) (Equation 3) is calculated as the sum of the mean of the squared residuals between the experimental and the theoretical values. The smaller the MSE the closer the model fit to the data.

$$Mean\ squared\ error = (1/n) \times (sum\ of\ squared\ residuals)$$

Equation 3

2.7 DoE software and implementation

The DoE interface and final tool used for the creation of the experimental set up for parallel bioreactors was programmed in Matlab. Development in Matlab allowed implementation of the statistical toolboxes necessary for developing relevant fermentation models. The software components are diagrammatically shown in the Figure 2 project scope.

3 Engineering Characterisation of the Multifors

3.1 Summary

For this engineering characterisation section, the fundamentals used for research experimentation are described by the methods previously developed at UCL (e.g. Gill *et al.,* 2008a). The engineering characterisation methodologies were applied to the parallel 1-L Multifors stirred-tank reactors and the results and discussions focus on the specific characterisation of power input, mixing times, and oxygen mass transfer. The Multifors was checked for consistency across batch fermentations in the parallel bioreactor system. Growth conditions were based on pre-established protocols (see methods section 2).

3.2 Research Context

Many glass and stainless-steel bioreactors at a small scale (1L–7L) are available for the cultivation of microorganisms. The design and specifications of these reactors are varied so suitable characterisation is needed for good scale translation and comparisons between equipment. Some of the main differences occur with gassing, mixing strategies, and power input (Meusel *et al.*, 2016).

When comparing solutions for scale-up criteria, one of the leading causes of confusion seems to be the set engineering characterisation objectives, assumptions, and methods themselves. The problems that arise are common when engineering characterisation methods are conditioned to a given vessel(s) or when there are varied methods being employed between studies. As such, there should be a standardisation of the relevant parameters in the studies, validation, standard operating procedures (SOPs), models employed, and set criteria for scale translation. Guidelines could then be employed to

single-use systems as well as reusable glass and stainless-steel reactors (Meusel *et al.*, 2016). Picking the right operating conditions for a given application can be made easier through characterisation although potential applications may also be enhanced and optimised using characterisation and models in conjunction.

Biological data are not the only key information needed in each bioprocess as, for example, process-related information is also required. This is usually presented in the form of process parameters. Some are obtained through the initial selection, dimensioning, and design of the bioreactor; others (like the operation conditions) can be obtained through calculations (Meusel *et al.*, 2016).

The routine experiments that can be easily measured are summarised below:

Process parameters generated through operational conditions and medium properties:

- Flow regimes (Reynolds number, for given laminar and turbulent zones).
- Fluid velocity (e.g. tip speed).
- Superficial gas velocity (residence time of gas front and the distribution of gas phase).

Process parameters generated through empirical data:

- Power consumption
- Mixing time
- Particle (shear) stress
- Volumetric oxygen mass transfer coefficient ($K_L a$)

The measurements requiring specialised knowledge and expertise are listed below:

Most of the listed parameters above can be determined through computational fluid dynamics and experimentation including energy dissipation rate and residence time distribution.

Computational fluid dynamics is known for its detailed spatial and time-related dependencies that provide a lot of information about the process (Loffelholz *et al.*, 2011). With computational fluid dynamics and an engineering characterisation process, critical limitations (e.g. mixing and oxygen limitation) can be overcome quickly using verified and validated models albeit the limitations are worsened at a larger scale. More specialised knowledge is needed to carry out these experiments (Meusel *et al.*, 2016).

The dynamic methods used to measure the volumetric mass transfer coefficient K_La have become more widely accepted to determine K_La. One advantage is it allows measurements to be taken with the actual fermentation culture several times to address the efficiency across a number of time points (Doran, 1995). The mechanism involves adequate supply of oxygen to saturate the liquid broth. The supply is then shut off and the resulting drop of oxygen tension is measured. By using the polarographic oxygen probe in the broth, it is assumed K_La is inversely proportional to the specific oxygen uptake rate of the cells (assuming a closed system that is homogenous) (Garcia-Ochoa & Gomez, 2009; Gill et al., 2008b; Doran, 1995).

3.3 Fundamentals in Engineering Characterisation

3.3.1 Geometry and Design

One key function of a bioreactor is to control the biochemical environment for cells to survive and grow exponentially. This is achieved by facilitating mass transfer of nutrients, oxygen, and metabolic products to and from the cells (Chen and Hu, 2006). Difficulties in bioprocess operation may arise when there are poor design choices for a given process: mass transfer problems, incomplete knowledge of the system or incorrect implementation of control software (Salehi-Nik *et al.*, 2013). The mathematical modelling of these bioreactors can provide a greater understanding of how design can impact a given bioprocess. Experimental design methods for various bioreactor geometries and designs would normally flow into process optimisation.

The scale of the bioreactor is an important part of design and geometry because scale can be related to cost-determining factors for research and development and, ultimately, costs in overall production. The geometry of a bioreactor should be similar, for any additional scale-up/down equipment. Any necessary adjustments identified during the bioreactor characterisation and modelling for scale-up/down should be applied.
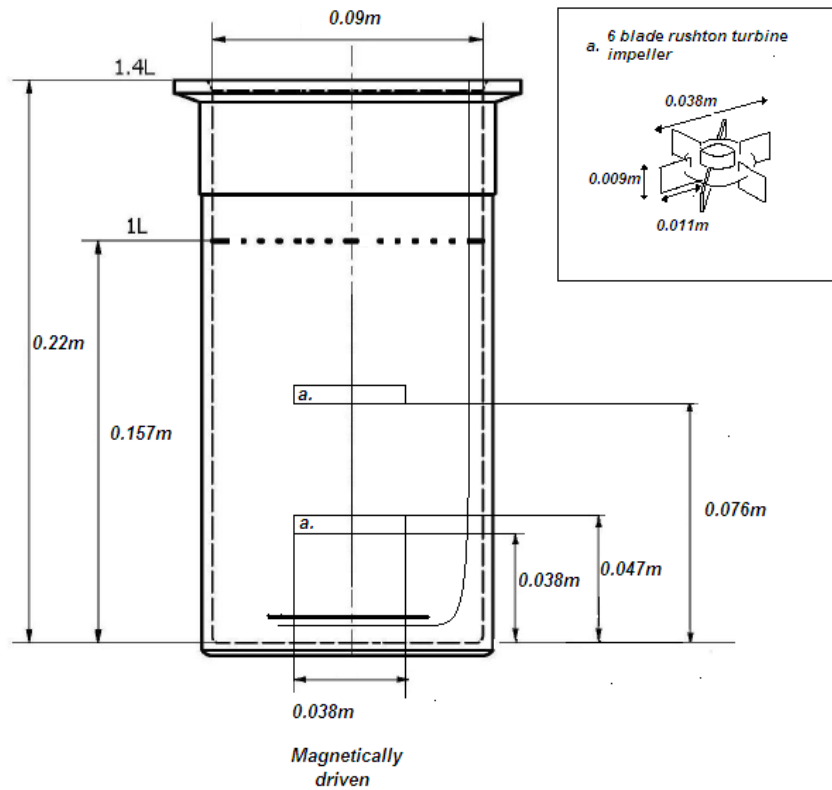
Figure 3 The STR vessel dimensions for the Multifors borosilicate glass vessels used for microbial cultures (scale 1L working volume) with permission for in-house purposes only (Adapted from Infors-HT, 2012a)

1: pO2 sensor
2: Air pipe/sparger
3: Exit gas cooler
4: Temperature sensor
5: Inoculation port
6: Spare port
7: Dip tube(sampling)
8: Antifoam sensor
9: pH sensor
10: Inlet needle for reagents
11: Ground for antifoam sensor

0.09m

0.012m port x5 (1,3,5,6,9)
0.010m port x4 (2,4,7,8)
0.007m port x4 (10)
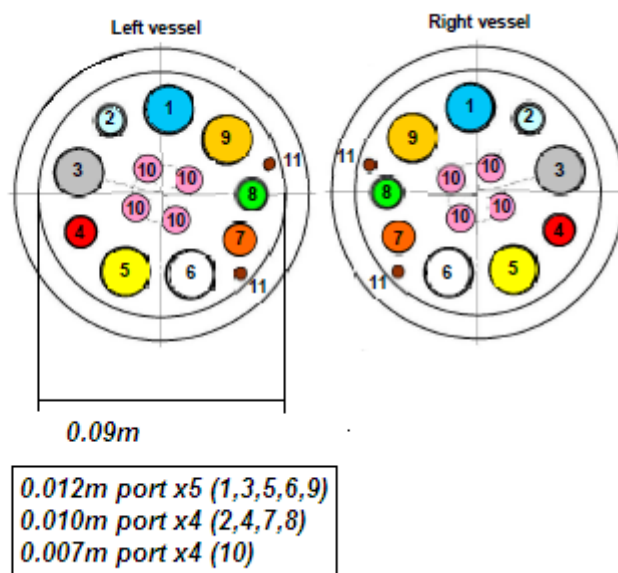
Figure 4 The top plate mechanical design for port space and location. Note the typical sensors/ports used for microbial culture listed. Used with permission and for in house purposes only (Adapted from Infors-HT, 2012a)

The baffled vessel used in this work was made from borosilicate glass, with a total volume of 1.4 L. The height to diameter ratio was 2.4:1 (Figure 3). Baffles were included as were two 6-bladed Rushton turbine impellers of diameter 0.038 m. Air was sparged through a bottom ring loop sparger located under the magnetic coupling attached to the bottom of the impeller shaft. The IRIS software, which was integrated into the control system and together with a flowmeter included with the Multifors™ (InforsHT, Switzerland), controlled airflow rate. A polarographic electrode was used for DOT measurements.

The reactor design shown in Figure 3 and Figure 4 allows for the accommodation of hardware sensors and probes on the top plate which is used for monitoring the culture during operation and for sampling. The top plate can be fastened onto the glass vessel via a clamp (not shown) and allows for the vessel to be sterilised in an autoclave. The exit gas port can be fixed to a mass spectrometer for gas analysis. The OPC client (networked controller) can then relay and receive information back from analysers to initiate automatic predefined controller inputs or cascade controls. Baffles were included in our experiments and these baffles fit into the glass vessel (not shown) (Infors HT, 2012a).

3.3.2 Power Consumption

Oxygen mass transfer coefficients (k_La) and power input per unit volume (P/V) are popular choices, among others, for scale-up (Gill *et al.*, 2008b). These criteria will be the basis on which a process framework for scale-up is developed. It is also assumed that the geometric height to diameter ratio is important for the empirical models to be valid. Other criteria that seem to be less important presently are maintaining mixing properties, shear stresses, impeller tip speed, gas velocity, and heat transfer capacity (Ju & Chase, 1992). Nonetheless, these are still useful parameters that are needed to keep

the microenvironment and physiological characteristics similar and make the cell physiology and CQA outputs repeatable. More complex models, where relevant, are needed to incorporate an increased number of scale-up criteria. This would entail the control software being updated and better applied. Predictive models could also be implemented.

Pg/V is one key parameter that can be used for process scale-up. The disadvantage of using this criterion for scale-up in process design is that the automatic controls usually cannot satisfy the high oxygen mass transfer demand from fast-growing organisms like *E. coli* (Gill *et al.*, 2008a), making it difficult to predefine the ranges for automatic control. Plain water and air systems are a useful way to test the maximum power input values for bioreactors. Conclusions can be based on the relationship between the empirical un-gassed to gassed power ratios (Gill *et al.*, 2008a). One disadvantage is that this power consumption may not necessarily reflect the power consumption with microbial cells at high cell densities although the same methods using torque measurements and gassing out techniques can be used to measure the un-gassed power input. Numerous operating conditions will have to be managed to better correlate engineering factors for high cell densities but this is still done through experimental means and is at best still an approximation (Gill et al., 2008b).

Power consumption and transfer of energy to kinetic motion and dissipation as heat to the liquid are key engineering parameters (Gill *et al.*, 2008b). Other parameters include the superficial gas velocity, bubble size, and viscosity needed for mass transfer when Pg/V is kept constant for scale-up (Gill *et al.*, 2008a; Miller, 1974).

The power number (P), also known as the Newton number, is a dimensionless number used to assess the performance of the stirrers used in a bioreactor:

$$N_p \frac{P}{\rho n^3 D^5}$$

Equation 4

Gill *et al.* (2008a) suggested that the bioreactor Pg/V performance was influenced by the overall number of impellers. In their investigation, the power number was multiplied by the number of stirrers in the calculation of the Pg/V to account for the new power efficiency input into liquid. There is a useful rule for a scale-up framework when using three impellers as outlined below as:

Power number of a single Rushton design impeller * the number of impellers used in the fermenter. N is power number of a single Rushton impeller. These values have been obtained before (Van't Riet and Smith, 1975). The factor 1.1 in Equation 5 is used for correct calculation of the power number for non-viscous liquids at high Reynolds numbers above >1 x 10^4 .

$$(n) * 3/1.1 = Np$$

Equation 5

The rule of thumb or correction was needed as previous studies looked at the power number of single Rushton impeller when a Di/Dt of 0.33 was used (Gill et al., 2008b). It was found that low agitation rates cause a drop in final biomass yields between different scales when the fermentation DOT ≤0 (Gill *et al.*, 2008b).

It was also found that their low energy input (657 W/m^3) was not optimal for achieving high oxygen transfer for efficient growth (DO dropped). This condition may be considered as being inappropriate for designs, especially for high cell densities in the Infors equipment. More specifically, effective oxygen mass transfer may not be achievable or reproducible and therefore not a good condition for scale-up. One course of action would be to compare the Infors-

HT data from experiments under high cell densities to those that are modelled by Hughmark (1980) or Cui et al. (1996), as cited in Gill *et al.* (2008b), using water/air systems to determine whether the assumptions are still valid.

One final area that can be characterised is the nature of gas dispersion during fermentation. It is challenging to accurately predict the hydrodynamic forces at various scales but this can be accomplished using computational fluid dynamics (CFD).

3.3.3 Fluid Flow and Mixing

Characterisation of the mixing efficiency is a useful step as it describes the physical process in which blending; dispersion; suspension; and transfer of mass, temperature, or immiscible liquids occurs (Doran, 1995). Power input into the system and fluid properties are key to understanding the mechanics behind fluid deformation and fluid flow. Fermentation liquid properties can affect the suitability and success of scale-up in the presence of cells, substrates, biological products, and air. Exponential growth and further survival in culture will occur unless there is very little access to substrates or huge concentrations of gradients are present. Equally so, survival is also affected if there is not a uniform suspension of biomass. Impeller mixing is the most typical method for carrying out this mixing operation.

3.3.3.1 Mixing Mechanisms

Maintaining and modelling turbulent flow patterns is important when large operating volumes are required and even more when the conditions inside a bioreactor are not at a steady state (Richardson & Peacock, 1994). Non-turbulent flow has been problematic within feedback control loops because mixing times and void times of solution additions can cause controller feedback issues, overall mixing inefficiencies, and further difficulties in modelling a system (Richardson & Peacock, 1994).

Turbulent flow regimes are currently used between laboratory, pilot scale and industrial scale to provide a basis for localised environments that are assumed to be well mixed and properly dispersed (Henzler *et al.*, 2000). This is especially important during batch and fed-batch fermentation to reduce the risk of any spatial gas-liquid or liquid-liquid mass transfer limitations. These mass transfers are crucial for microbial cell growth (Henzler *et al.*, 2000).
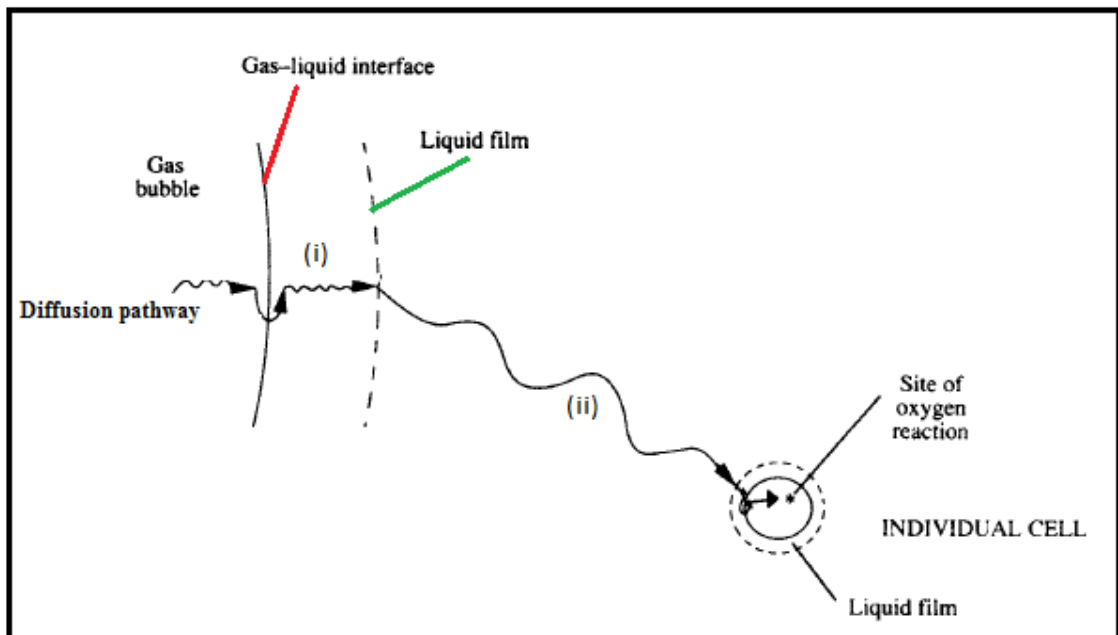
Figure 5 The oxygen-diffusion pathway theory from gas bubble through respective interfaces and films to reach bulk liquid and transportation to individual cells. Note: It is assumed the rate of oxygen mass transfer through the film layer around the individual cells is negligible. The (ii) pathway is assumed to be under well mixed conditions and, as such, the oxygen molecule diffusion is ultimately rate limited by the size of the liquid film layer surrounding the gas bubble (Adapted from Doran, 2012).

Diffusion-based micro-mixing is a deterministic model of a readily available supply of oxygen to the cell for aerobic cell metabolism, as indicated above. As such, physical stresses act upon particles in three dimensions owing to the space-dimensional relationship of particles – microbial cells in this case (Doran, 1995). This affects the cells' abilities to flocculate, adhere to surfaces, and rupture at cell scale.

Physical stresses also work within the broth as can be seen by the breaking up of gas bubbles to dissolve into liquid more easily. Gas dissolving is more efficient if the ratio of the surface area of the gas bubble is proportionally higher to the gas bubble volume. One way of providing this effect is through physical stresses via the impeller at high speeds (Doran, 1995).

Energy dissipation theory has been used to model the fluid dynamics expected during a unit operation (Lamping *et al.*, 2003). Energy dissipation characterisation is useful for building models to relate any correlation of power draw. Energy dissipation is specific to the type of liquid broth: whether it is aerated, its volume and its turbulent flow, the reactor geometry, and operational parameters that are to be controlled for scale up (Gill *et al.*, 2008a).

The next suitable step is to correlate models for high cell densities. Optimisation studies using computer software are becoming readily available and more popular for increasing throughput. More effective mathematical models are needed to increase the predictive simulations that can be used for advancing fermentation control at varying scales of operation.

3.3.3.2 Turbulence

The term "eddies" is used to describe kinetic energy dissipation from mechanical input via convective motion. This mechanical stirring and energy input creates the physical stresses to
the start flow and energy dispersion in liquid. Likewise, the particle stresses acting via turbulence can be modelled on a Kolmogorov scale as shown in Equation 6.

$$\lambda = (\frac{v^3}{\epsilon})^{1/4}$$

Equation 6

The smallest eddies provide the highest power dissipation over time and considering most microbial cells are likely to be smaller than these eddies, they will cause the highest particle stress. This is summarised in Figure 6 (Leipe, 1988, cited in Henzler *et al.*, 2000).

$$\tau_t^* = \frac{\tau_t}{\rho \sqrt{\nu \varepsilon}}$$

Dissipation range → ← Transition region → ← Inertial range

$$\tau_t^* = 1.9 \left(\frac{d_P}{\eta_L}\right)^{2/3}$$

$$\tau_t^* = 0.0676 \left(\frac{d_P}{\eta_L}\right)^2$$

$$\tau_t^* = 0.22 \left(\frac{d_P}{\eta_L}\right)^{4/3}$$

d_P/η_L
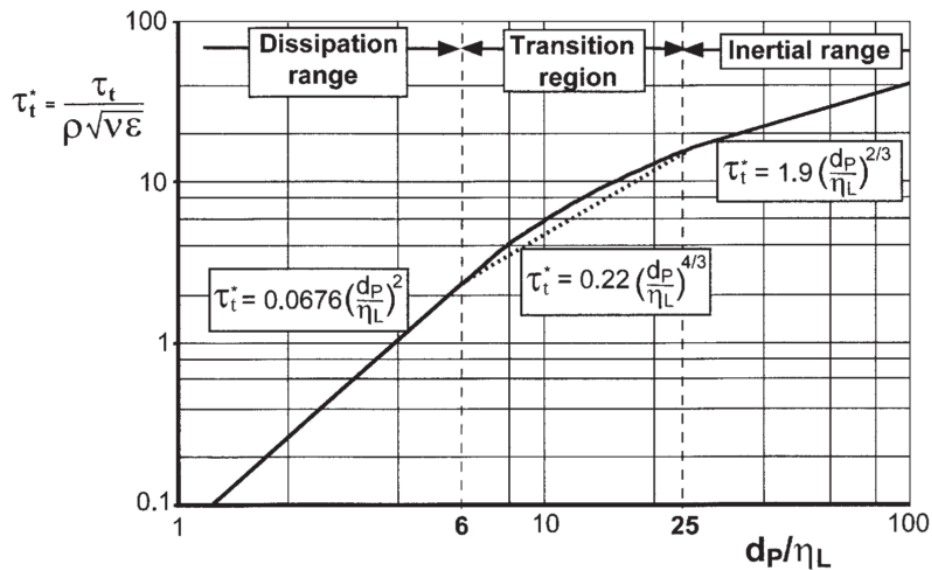
Figure 6 The simplified non-uniform distribution of energy associated with the smallest eddie formation at Kolmogorov micro scales. Note: The dissipation and inertial ranges can be modelled under linear (or close to linear) conditions (as it is based upon empirical evidence). The transitional period that has the non-linear region must be generalised (adding more inaccuracy) to characterise the isotropic turbulence (Referenced from Henzler *et al.,* 2000).

The small micro-scale eddies provide the least amount of distribution, known as macromixing, so depending on the type of impeller, turbulent zones occur mostly around and behind the impeller blades and have the highest particle stresses locally. Owing to this localisation effect, it can be considerably difficult to characterise the exact maximum energy dispersion (Henzler *et al.*, 2000). Additionally, the further the cells are from the blade the lower the impact of particle collisions and stress on the cells (Henzler *et al.*, 2000).

Macro- and micro-scale turbulence should be treated as independent forces. This ensures an insignificant diffusion rate limit whilst warranting sufficient bulk mixing at high Reynolds numbers. This can validate the models described in Figure 6 (Henzler *et al.*, 2000; Miller, 1974). Therefore, scale translation provides macro-mixing times that are relatively similar. This relationship for power input and mixing can be used for scaling-up purposes when calculating the maximum energy dissipation that a particular bioreactor can theoretically achieve (Henzler *et al.*, 2000; Doran, 2012).

The maximum energy dissipation is then proportional to the mass mean power input in (Equation 7), as cited by Henzler et al. (2000).

$$\overline{\in} = \frac{P}{\rho V}$$

The power number must be determined experimentally and correlated into Equation 8 (Garcia-Ochoa & Gomez, 2009).

$$P = \frac{Newton\ number(\ \text{Ne or } P_{ug})}{\rho * n^3 * d_i^5}$$

To use Equation 8, the Reynolds number must be under turbulent regime within a non-aerated system. In aerobic conditions, when modelling for the aerated parameter, dimensionless equations can be used to relate particle stress in reactors at various scales with the Froude number (Fr) and gas throughput number (Q) (Garcia-Ochoa & Gomez, 2009):

$$Fr = stirrer\ speed(\frac{1}{s})^2 * \frac{diameter\ of\ impeller\ (m)}{gravity\ (\frac{m}{s^2})}$$

$$Q = \frac{volumetric\ gas\ flow\ rate\ \left(\frac{m^3}{s}\right)}{stirrer\ speed\ \left(\frac{1}{s}\right) * impeller\ diameter\ (m)^3}$$

$$\frac{Pg}{Pug} = \frac{1-\left(\frac{Pmin}{Pug}\right)}{1+c*Fr^a*Q^b} = \frac{Pmin}{Pug}$$

where P_{ug} is f(x)=Froude number

$$\frac{P_g}{P_0} = 0.10\left(\frac{F_g}{N_iV}\right)^{-0.25}\left(\frac{N_i^2D_i^4}{gW_iV^{2/3}}\right)^{-0.20}$$

To relate the gassed-to-ungassed power input, the ratio model could be used (Equation 11) (Miller, 1974; Henzler *et al.*, 2000; Garcia-Ochoa & Gomez, 2009) although it should be noted that P_{min} is the lowest possible Newton number achieved by allowing the gas throughput number to be significantly high. In this way, P_{min} becomes relatively constant (and reaches its asymptote) with any further increases to Q.

Another correlation previously described by Hughmark (1980) included the additional impeller width (W) parameter as seen in Equation 12. At this current technological point, a model like this seems reasonable to use to correlate gassed to un-gassed power requirements (Henzler *et al.*, 2000: Hughmark, 1980; Doran, 2012). This equation was also used by Gill *et al.* (2008) to correlate their miniature scale-down bioreactor power requirements.

Once the power of the aerated vessel is determined, then the power input per unit volume can be simply calculated to be used for scale-up. One additional point is that "mixing times will increase while keeping Pg/V constant" and can be accounted for during controller feedback (Doran, 2012).

Mixing time as a scale-up criterion is often not used for scaling up owing to the amount of extra power needed to drive larger impellers at larger scales, making some microbial bioprocesses uneconomical (Miller, 1974; Garcia-Ochoa & Gomez, 2009).

Finally, dissipation modelling is dependent on the particles' ability to deal with those physical stresses exerted in the turbulent region.

Characterisations are needed to determine good mixing for specific types of microorganisms at various scales. This could highlight the dependence that shear stress has on cell viability when fermentation liquid is turbulent and high cell densities are required. This also relies on empirical means to correlate and create relationships but gives some degree of information to use for simulation studies (Henzler *et al.*, 2000). Gassed power input per unit volume is also a volumetric empirical correlation which has been reported to be easily used for rapid scaling owing to the simplification of the environment characteristics into scalable terms (Gill *et al.*, 2008b).

3.3.3.3 Mixing Time Assessment

Mixing time, i.e. the rate of bulk flow mixing, is the parameter used to compare the efficiency of macro mixing at various scales (Doran, 1995). Homogenous mixtures require a certain time (t) to reach the desired non-gradient state. Liquid tracers and relevant detectors are used to measure the mixing time. These can be in the form of addition of acids, bases, or concentrated salt solutions (Doran, 1995). The difficulty lies in deciding the

mixing time required for a given degree of homogeneity and uniformity. Usually, this is when the tracer concentration is less than 10% from when it initially started (C_{final}–C_{start}) (Doran, 1995). The factors influencing the mixing time of large-scale equipment is based on the reactor design. This links the need to characterise the reactor and the mixing time as scale-up parameters. Estimated predictions can be made by understanding the relationship between the size of the tank and the impeller, rotational speed, and fluid properties in Equation 13 (Doran, 1995). This correlation was chosen because high Reynolds numbers can be achieved at the scale of the parallel bioreactors used.

$$N_i t_m = \frac{1.54\, V}{D_i^3} \; at\; high\; Re_i$$

Equation 13

3.3.3.4 Oxygen Mass Transfer

Cell growth, cell maintenance, and metabolite production in aerobic bioprocesses require oxygen as a substrate for some specific metabolic reactions. It is important to note that reaction rates could either be rate limiting to the metabolic reaction kinetics or to the transport of substrate to the site of reaction (Garcia-Ochoa & Gomez, 2009).

Keeping the substrate (oxygen) concentration above the minimum rate for determining mass transfer threshold can be challenging at a large scale. As such, process performance can be affected by a poorly aerated liquid growth medium (Garcia-Ochoa & Gomez, 2009). Maximising overall process performance has a large relevance in the industrial bioprocesses.

Unsurprisingly, the aeration potential and metabolic pathways can be impacted by the scale and vessel geometry (ratio of height:diameter ratio of the vessel) in the very first phase of vessel design. The sensitivity of the

reactor design should be considered early on. Therefore, it is important that the oxygen transfer rate be correctly estimated for each vessel used in each bioprocess (Garcia-Ochoa & Gomez, 2009). Measurement of the mass transfer rate can also be challenging owing to many phenomena occurring at the same time during fermentation.

Aeration potential and oxygen uptake of cells can be limited by the fact that oxygen is a notably poorly dissolvable gas both in water and liquid medium containing salts and organic substances. Other factors like the fluid properties also affect the dissolvability of gases (Garcia-Ochoa & Gomez, 2009).

3.3.3.5 $K_L a$ Determination

The maximum mass transfer rate of oxygen transfer (OTR) from gas bubble to bulk liquid can be calculated by the product of the volumetric mass transfer coefficient ($K_L a$) and the saturation concentration of oxygen in liquid medium (C*) (Garcia-Ochoa & Gomez, 2009).

$$OTR = K_L a.C^*$$

Equation 14
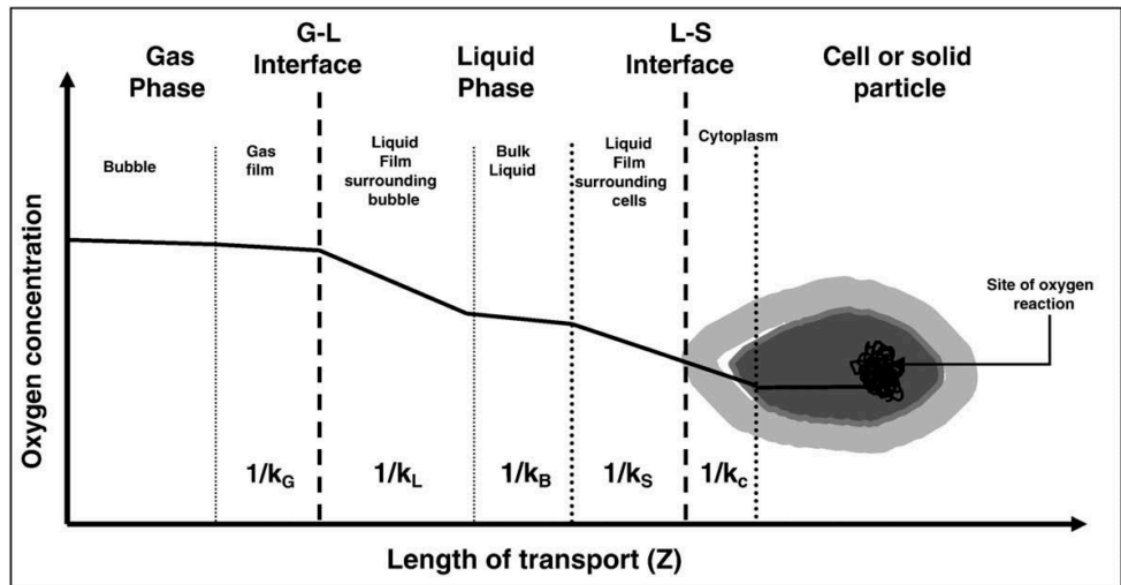
Figure 7 Garcia-Ochoa & Gomez's (2009) schematic view of the rates of transportation influencing oxygen transfer rate in a bioprocess

In Figure 7, the oxygen diffusion pathway schematic shows the sum of the rates at which oxygen is diffused. Oxygen diffusion through phase boundaries, layers, and films into the bulk liquid and to the cell is defined as the oxygen transfer rate (Equation 14) (Doran, 1995). This rate is a function of the concentration known as the driving force between the global saturated concentration of oxygen and the actual mean concentration of oxygen in the liquid. Considering that during fermentation, oxygen is constantly being consumed, coupled with the poor solubility of oxygen, the driving force remains rather small (Garcia-Ochoa & Gomez, 2009). The overall oxygen mass transfer coefficient is then a good criterion to match for scale-up as $K_L a$ is influenced by both engineering and biological factors. The factors include gas interfacial area, mass transfer coefficients, concentration gradients, bubble hold up, bubble diameter as well as operational conditions (Garcia-Ochoa & Gomez, 2009)

3.3.3.6 Antifoam

It has been noted that antifoam, such as polypropylene glycol (PPG), is more likely to reduce the capacity for oxygen transfer than simple electrolyte solutions (Stanbury, 1995). This is due to the antifoam increasing the air bubbles' surface tension, thereby making it harder for small bubbles to be created. Oxygen mass transfer is then influenced by resistance of the size of the gas-liquid film layer which slows down oxygen dissolution, as shown in Figure 7 (Garcia Ochoa & Gomez, 2009).

3.3.3.7 Gas Blending

Gas blending with a higher content of oxygen is usually needed for high cell density culturing. This method allows the supply of enough oxygen at later stages of fed-batch or continuous fermentations to maintain cell viability and metabolism by providing a DOT that does not fall below a critical level. Blending on a cascade control was recommended as a suitable strategy in this study as it is easier to operate automatically instead of judging when to manually start this type of control.

3.3.3.8 Functions of K_La

A well-established empirical correlation has been used to relate the K_La with the superficial gas velocity and mean power input over time (see Equation 15). The constant "c" can be found from previous literature studies, e.g. Gill *et al.* (2008) or Van't Riet (1979) or data fitting. The exponents α and β can be determined experimentally.

$$K_L a = c * \left(\frac{P_g}{V}\right)^{\alpha} * (v_s)^{\beta}$$

Equation 15

It has been suggested that K_La is a function of Pg/V, superficial gas velocity, Froude number (Equation 9), gas throughput number (Equation 10), kinematic viscosity, oxygen solubility/diffusivity, broth density, and the hydrodynamic pressure inside the vessel (Gill *et al.*, 2008b). These functions have led previous studies to include the gas throughput number instead of the superficial gas velocity (Moresi & Patete, 1988). It has been found to provide better dependence for active cells in larger volume vessels, which has been corroborated in recent studies (Cants *et al.*, 2005).

3.3.3.9 Functions for Larger Reactors

K_La-determining functions for larger vessels can be used with equations based on averages across the fermenter, mainly at the inlet and outlet as represented below:

$$K_L a = \frac{OTR}{\dfrac{(\sigma^* in - C_L) - (C^* out - C_L)}{ln \dfrac{(C^* in - C_L)}{(C^* out - C_L)}}}$$

Equation 16

C^* represents the total oxygen concentration saturation and C_L is the local concentration after converting local oxygen partial pressures to dissolved oxygen concentrations (Doran, 1995).

$$K_L a = \frac{ln \dfrac{C^* - C_1}{C^* - C_2}}{t_2 - t_1}$$

Equation 17

(C* is oxygen saturation conc. in liquid. "Cin" is equal to the local electrode reading of DO at gas inlet and "Cout" is the local electrode reading of DO at gas outlet. C_1 is DO at time point t_1 and C_2 is DO reading at time point $(t)_2$).

It will be beneficial if K_La can be determined via the computer/controller during fermentation with the option to graphically show the values. This would be ideal in conjunction with dynamic gassing out or dynamic pressure measurement techniques (Doran, 1995).

3.3.3.10   Probe Response Characterisation

Methods used to determine the K_La include the static gassing out technique based on oxygen tension from 100% saturation to 66%. This is done to assess the mass transfer capability, assuming the DOT probe response time is negligible. As oxygen diffuses locally close to its membrane and crosses the polarographic layer to be detected, errors can be introduced from the probe response time. This must be accounted for when measuring the partial pressure of oxygen in the reactor (Garcia-Ochoa & Gomez, 2009).

Other methods that are often overlooked for bioreactor K_La characterisation are chemical methods using sodium sulphate, primarily because of the hydrodynamic changes and the need to control the concentration of ions in the liquid. These changes usually occur at higher cell densities and tend to overestimate the K_La as compared to other techniques. It has not been commonly adopted as a framework for scale-up (Doran, 1995). This is the case with the CO_2 absorption-determination method (Garcia-Ochoa & Gomez, 2009).

As long as the mass transfer of $1/K_La$ is at least a magnitude higher than the probe response time, no further corrections incorporating the lag into the formula are required. However, when dealing with fast-growing organisms, there has been the need to correct for probe response times, necessitating

the need to get accurate values of the mass transfer coefficient (Garcia-Ochoa & Gomez, 2009). Gill *et al.* (2008b) utilised this dynamic gassing out method to determine the mass transfer coefficient. Further studies will be needed to characterise the K_La for higher cell density cultures with this dynamic gassing out method.

3.4 Results and Discussion

3.4.1 Power Input and Energy Dissipation

Understanding the gassed to un-gassed power ratio is important to find correlations specific to the Multifors system. We need to understand how much energy is being dissipated into the mixed system. This experimentation allows for an insight into the maximum resistant forces that the reactor can operate at to create suitable correlations, which can be used to evaluate how best to control a batch operation at this scale or for scale-up/down. Based on previous research, the gassed to un-gassed power ratio can vary between 30% and 100% (Gill *et al.*, 2008). The aim of the experiment is to define the operating parameters for *E. coli* batch fermentation and provide input parameters for the simulation tool such as the power input for advanced modelling. The experiment was set up by making fresh glycerol solutions (1L) w/w for each density, bringing solution to temperature, setting agitation and aeration conditions, then take motor current reading after 10 minutes using a calibrated ampmeter and have a total of 36 experiments for the reactor Two standard Rushton Turbines in fully equipped vessel was used and a standard bacterial ring sparger.

The results in Figure 8, Figure 9, and Figure 10 show the measured power requirement of the gassed and un-gassed liquids for each of the various viscosities: 0.8, 4.21, and 500 Cp/mPa·s. For all aeration rates tested, the relationship of the power requirements was positive and almost exponential,

there was a difference in the correlations between the highest viscosities and the low viscosities, which indicates that viscosity has impact on the aeration in the vessel. This means that an increase in both rpm and viscosity led to increased power requirements. The maximum power consumption established was 10.68 W for 1 vvm at 1200 rpm for 500 Cp/mPa·s.

At two points, too much fluid resistance was experienced from the most concentrated glycerol experiments (see Figure 10). After operating conditions were set and increased to the higher rotational speeds they gradually led to the failed agitation measurements as the shaft stopped rotating. Moreover, this was seen visually when too much stress was applied to the magnetic coupling on the base of the shaft. This occurred under the conditions at 500 Cp/mPa·s, glycerol density of 1.253 g/cm^3, and 1200 rpm.
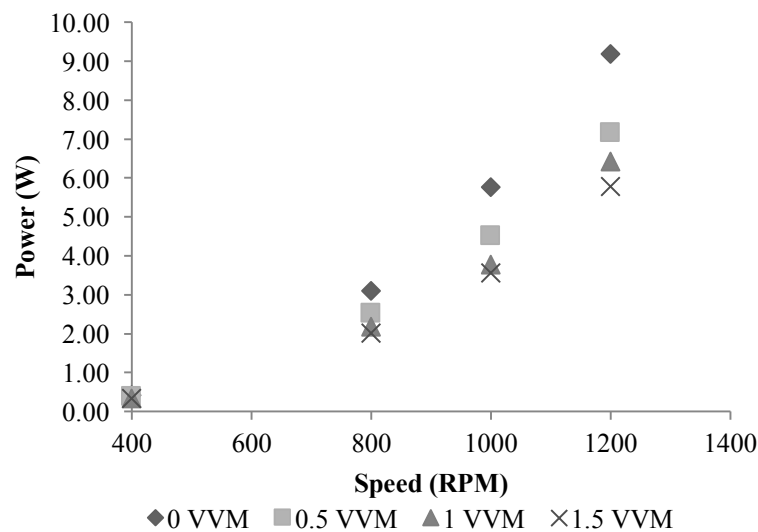
Figure 8 Un-gassed and gassed power requirements for the two impellers setup with a 1-L fill volume for water in the Multifors™ vessel. The water viscosity measured was 0.8007 Cp/mPa s. The water density was 0.996 g/cm^3. The experiment was carried out at 30°C.
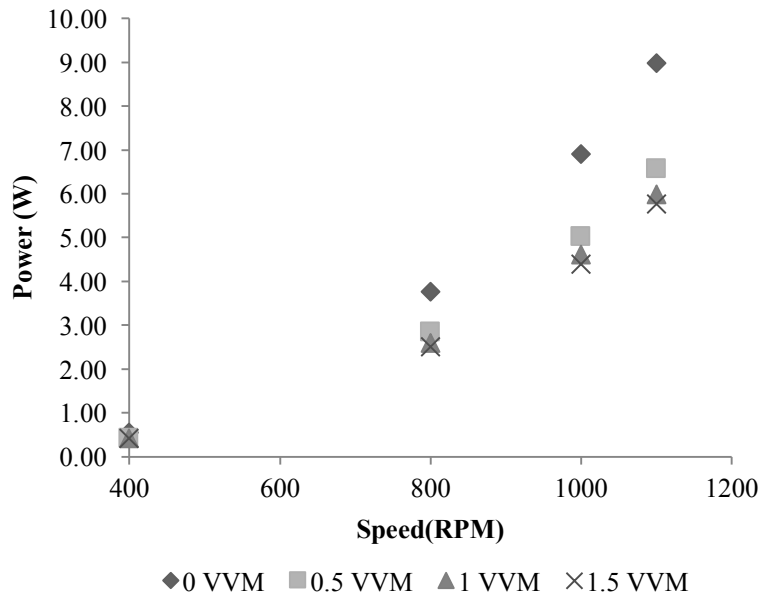
Figure 9 Glycerol (50%); ungassed and gassed power requirements for the two impellers setup with a 1-L fill volume for glycerol solution in the Multifors™ vessel. The viscosity measured was 4.21Cp/mPa·s. The density was 1.211 g/cm³. The experiment was carried out at 30°C.
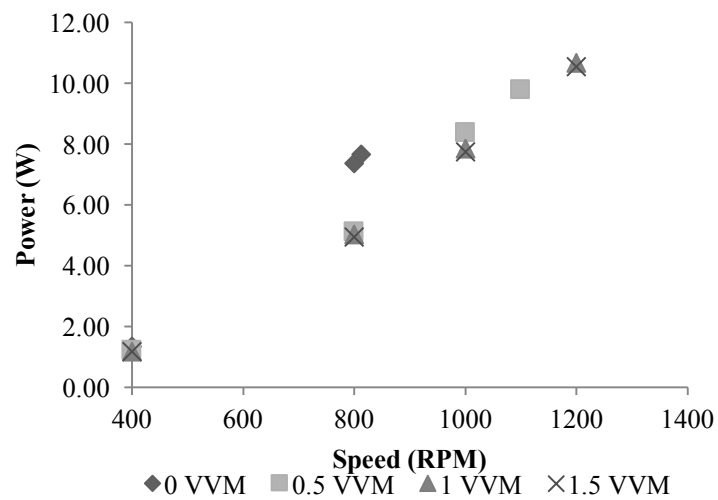
Figure 10 Glycerol (99.5%); Ungassed and Gassed power requirements for the two impellers setup with a 1-L fill volume for water in the Multifors™ vessel. The water viscosity measured was 500 Cp/mPa s. The water density was 1.253 g/cm³. The experiment was carried out at 30°C.

It was found that the un-gassed power requirements were, on average, 32% higher than the maximal gassed power input at 1.5 vvm. Under highly viscous conditions, gassed power requirements were 24% less at 500 Cp/mPa·s than

un-gassed ones. When comparing the results to other research, there was a small difference between established observations of 0.3–0.4 of the P_g:P_{ug} ratio (Gill *et al.,* 2008). This is further explained in the discussion.
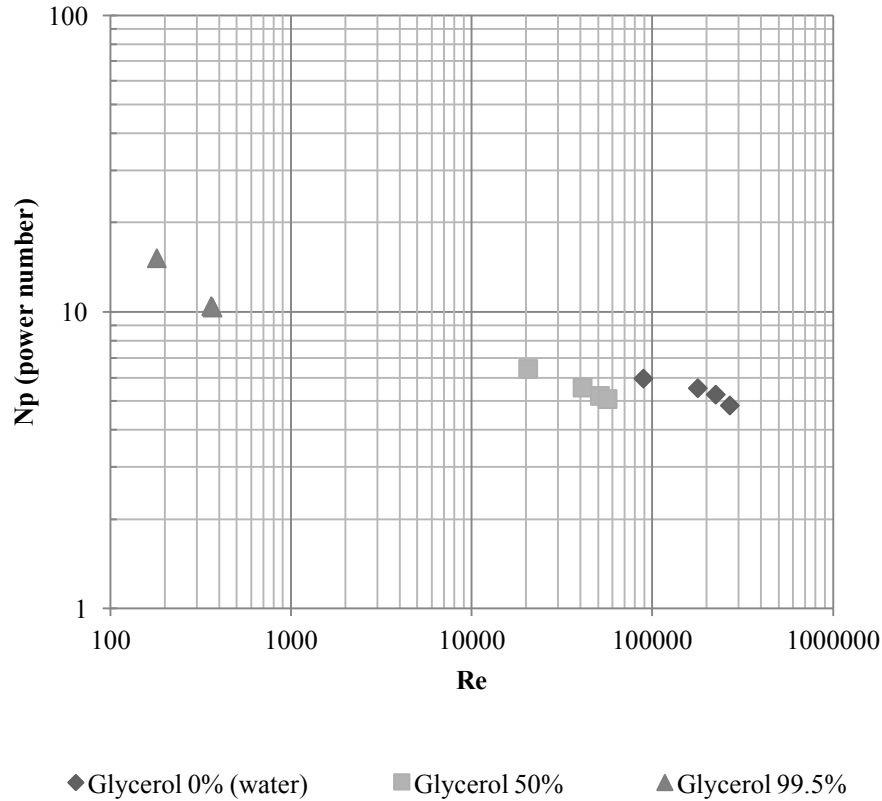
Figure 11 Power number and Reynolds dimensionless number correlation between different glycerol concentrations

For the un-gassed power requirements, the power number relationship in Figure 11 was correlated to the dimensionless Reynolds number. The power number (Np) was calculated using Equation 18.

$$Pug = Np.\rho.N^3.di^5$$

Equation 18

Here, Pug (W) is the measured un-gassed power input, ρ (kg/m^3) is the density of the liquid, N (m/s) is the agitation rate, and di (m) is the impeller diameter (Gill et al., 2008a).

3.4.2 Gas Dispersion

The goal of this part of the study was to understand the gas dispersion for the parallel bioreactors under different fluid viscosities. This is a follow-on study from the power input characterisation. The gassed to un-gassed ratio can give insight on the gas dispersion in the types of media used to characterize the power input. The types of impeller and aeration rate are factors affecting the power decrease for gassed operation (Aiba *et al.*, 1973). The gas flow number can be used to describe the nature of the gas dispersion within the liquid (see Equation 19)

$$Fl=Q/Nd_i^3$$

Where the Q is the volumetric gas flow rate, N is the agitation, d_i is the impeller diameter (Gill *et al*, 2008a). The following figures show the gassed to ungassed ratio with the Fl (flow number). The experimental values determined were all between 0.5-1. Gassed to un-gassed ratios have been previously described within a range of 0.3-1 (Gill *et al*, 2008a).

Table 1 Calculated minimum agitation rates and flow numbers for complete gas dispersion using Equation 19.

	Air flow rate (vvm)		
	0.5	**1**	**1.5**
Calculated minimum agitation rate for successful gas dispersion (rpm) (see Equation 20).	263	372	456
Minimum flow number that should be attained for complete gas dispersion, using only respective air flow rate and **minimum** agitation speed in Equation 19.	0.035	0.049	0.060
At **higher** agitation speeds (1000 rpm), the flow number for the experimental data for gas dispersion achieved for water (0% glycerol).	0.008	0.018	0.027
At **higher** agitation speeds (1000 rpm), the flow number for the experimental data for gas dispersion achieved for 50% glycerol.	0.009	0.018	0.027

For Equation 20, a correlation for the minimum agitation rate for which complete gas dispersion is achieved. This is valid for vessel dimensions up to 1.8m in diameter (Nienow *et al,* 1977).

$$N_c = 4*(Q_g^{0.5} d_T^{0.25})/d_i^2$$

Equation 20

Where N_c is the rotational speed of the minimum agitation rate for complete gas dispersion (rps). Q_g the volumetric gas flow rate, d_T is vessel diameter, and d_i is the diameter of the impeller.

This correlation was chosen because of the importance of the vessel dimensions of the parallel bioreactors.
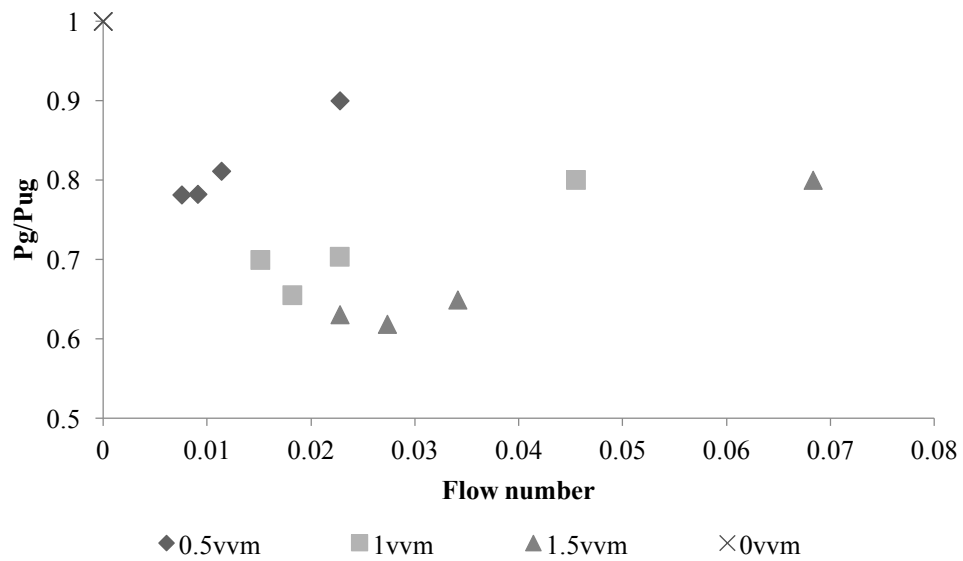
Figure 12 Flow number against gassed/ungassed ratio for water
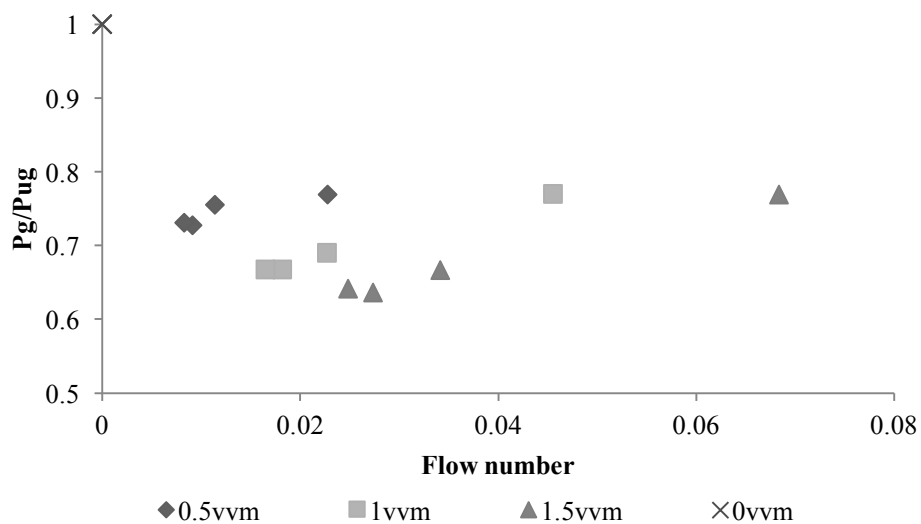
Figure 13 Flow number against gassed/ungassed ratio for 50% glycerol concentration

Flow numbers were calculated using minimum set points of the air flow rate and the agitation rate (see Table 1 and Figure 12, Figure 13). Gas dispersion was correlated to be fully dispersible and hence the most effective for mass transfer at the point where Pg/Pug was the lowest (Gill *et al.*, 2008). Gill *et al.* (2008) suggest that a difference in the flow number needed for gas dispersion is likely due to the impeller blade thickness.

Reynolds numbers below 10,000 would not have desirable gas dispersion at those speeds. For the pure glycerol (99.5%) concentration, the Pg/Pug showed the lowest Reynolds numbers; at an agitation rate of 800 rpm, laminar flow was observed (see Figure 11). Therefore, concluding that complete gas dispersion for 99.5% glycerol concentration was not achieved.

Small differences in the minimal flow number were found shown in Figure 12 and Figure 13and table 1 for each of the experiments. These small differences are accounted for by the correlation chosen, such that, increased agitation speeds N in equation 19 affects the flow number. This suggests that the agitation rate and vessel dimensions have the most influence on gas dispersion.

Measured Pg/V for water at 1.5 vvm

Measured Pg/V for Glycerol (50% v/v) at 1.5 vvm

Measured Pg/V for Glycerol (99.5% v/v) at 1.5 vvm
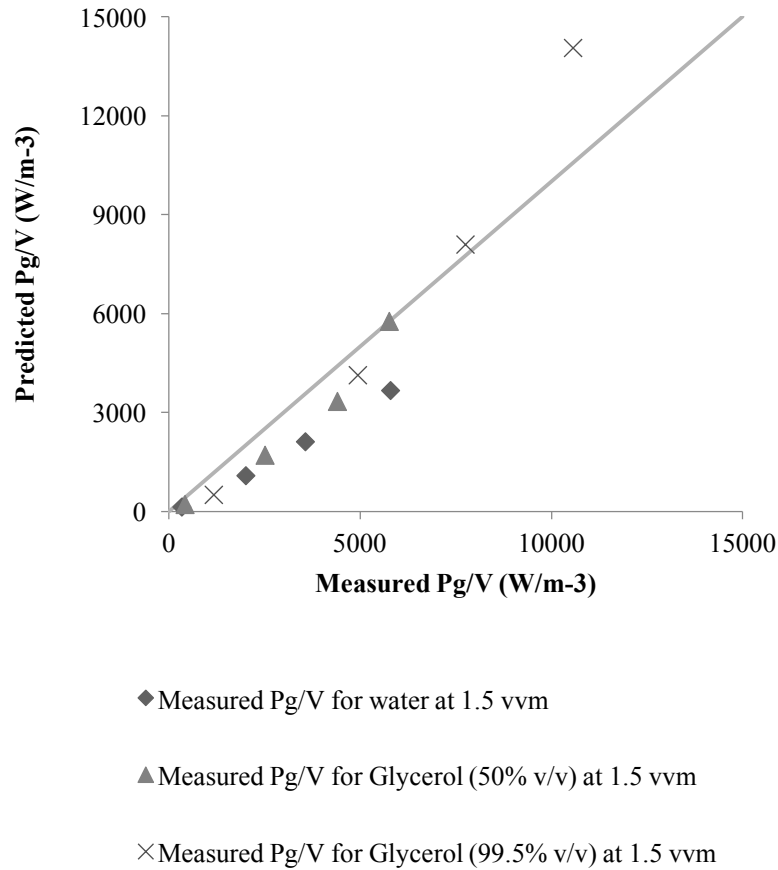
Figure 14 Parity plot of measured Pg/V against the predicted Pg/V using Equation 21

The next part of the characterisation was to describe the measured power input values versus the calculated values for different liquid viscosities. (see Figure 14). Luong and Volesky (1979) described a correlation (see Equation 21) that was used to estimate Pg/Pug with a viscosity component. When substituting the correlation Pug = $(N_p . \rho . N^3 d_i^5)$ into it and with it a power number of 5.8 (originally determined by Van't Riet, 1979) for 3 impellers. This third impeller is an important choice in our parity plot in Figure 14. It was added because in our experimental set up we had one magnetic stirrer bar attached inside the vessel and concluded that this had impact on the estimation of the power input, since viscous liquids were used the correlations show that there was impact on the power input. Two Rushton impeller turbines were included in the reactor set up as normal. This new estimated Pg/V using Equation 14 was calculated and compared against the measured

Pg/V in a parity plot. Figure 14 indicates how well the correlation predicted the empirical data except for the highest power input and high viscosity.

$$P_g = 0.497 \left(\frac{Q}{Nd_i^3}\right)^{-0.38} \left(\frac{N^2 d_i^3 \rho}{\sigma}\right)^{-0.18} (N_p . \rho . N^3 d_i^5)$$

Furthermore, Figure 14 highlights the differences in the rheological properties of the measured gassed power input to those of the correlation for Pg/V. It does consider rheological factors in the correlation and makes it easier to relate the correlation for gassed power input to viscous liquids. This correlation was a better fit to more viscous liquids of our measured data typical of non-coalescing systems that are similar to media broths of high viscosity (Luong & Volesky, 1979).

3.4.3 Power Input discussion

The gassed-to-un-gassed ratio gave insight into the gas dispersion for the types of media used to characterise the power input. Impeller type and aeration rate are factors in the power decrease for gassed operation (Aiba *et al.*, 1973, cited in Gill *et al.*, 2008a). The gas flow number (Fl) can be used to describe the nature of the gas dispersion within the liquid as shown in the equation below:

$$Flow\ number = \frac{Q}{Nd_i^3}$$

Q is the volumetric gas flow rate, N is the agitation, and d_i is the impeller diameter (*Gill et al.*, 2008). The following figures show the gassed-to-un-gassed ratio with the Fl. All of them are between 0.5 and 1, but the gassed-to-un-gassed ratio has previously been found within the range of 0.3–1 (Gill *et*

al., 2008). Equation 23 shows the correlation for the minimum agitation rate for which complete gas dispersion is achieved. This is for vessels up to 1.8 m in diameter (Nienow *et al.*, 1977).

$$N_c = 4*(Q_g^{0.5} d_T^{0.25})/d_i^2$$

Equation 23

N_c is the rotational speed of the minimum agitation rate for complete gas dispersion (rps). Q_g denotes the volumetric gas flow rate, d_T, the vessel diameter, and d_i, the diameter of the impeller.

The Multifors system (InforsHT, Switzerland) has an upper recommended rotational limit of 1200 rpm. The system has this limitation owing to the power rating of the motor, and the control system has been programmed to not exceed this limit. It is beneficial to limit the rotational speed so as not to cause overflow of the contents inside the reactor during operation, mainly during aeration.

The gassed power measurements were lower than the corresponding values under un-gassed conditions. When the process parameters were being generated, the parameters were dependent on changes in fluid density occurring between conditions. There could have been a risk of gas bubbles not being adequately dispersed in highly aerated systems (Van't Riet & Smith, 1973). Hence, three repeats were conducted to mitigate the errors being introduced into the calculations. Another effect of having an aerated system is less resistance at the point of the impeller as the impeller hits the gas for easy rotation (less power to rotate at a maintained speed) (Van't Riet & Smith, 1973).

After the experiments, the calculations for the power number with respect to the impeller type were approximated. This is a good way to verify that the power number is in line with other literature. Careful consideration was given to the influence of the impeller bar on the mixing. There was a concern that the magnetic bar, located at the bottom of the shaft, influenced the overall power distribution. This bar was likely to contribute to the distribution of power input into the liquid.

After the careful consideration of the influence of the magnetic stirrer bar a rule of thumb or correction was needed for the power input studies. Previous studies have looked at the power input using a single Rushton impeller (Gill *et al.*, 2008b). It was found that a low energy input (657 W/m^3) was used which was not optimal for achieving high oxygen transfer for efficient growth (DO dropped). and the low agitation rates cause a drop in final biomass yields between different scales when the fermentation DOT ≤0 (Gill *et al.*, 2008b).

Their operating conditions would be considered as being inappropriate for certain processes, especially for high cell density cultures in the Infors equipment. More specifically, effective oxygen mass transfer may not be achievable or reproducible and therefore not a good condition for scale-up. One course of action would be to compare the Infors-HT data from experiments under high cell densities to those that are modelled by Hughmark (1980) or Cui *et al.* (1996), as cited in Gill *et al.* (2008b), using water/air systems to determine whether the assumptions are still valid.

One final area that was characterised was the nature of gas dispersion during operation. It is challenging to accurately predict the hydrodynamic forces at various scales without using computational fluid dynamics (CFD). The correlation of P_g/P_{ug} against the gas flow number was conclusive by indicating the gas dispersion in the experiments because of the power input varies more so under aerobic conditions; it decreases from 1 to about 0.3. Furthermore,

this can be applied to rapid bioreactor design for prediction of aerobic culture conditions (Gill *et al.*, 2008b).

The un-gassed power requirements were finally divided by three (because of the two Rushton impellers and the magnetic stirrer bar) to give a better approximation of the power number for each impeller.

The magnetic bar/impeller would still naturally exhibit resistance and create agitation because it is in contact with the liquid; it did not matter whether it was water or a viscous solution. The fact that the stirrer bar is a uniquely cut and coated magnet with a plated disk and with opposite facing 45° circular wings on top, it was likely designed to help in the distribution of gas. The gas sparger was located below the magnetic impeller. All of the additional objects in the bioreactor are likely to cause concerns related to the distribution of gas, power, and overall homogeneity.

During the experiment, the magnetic bar was going to increase torque in the shaft, giving higher torque values than would be expected if the impellers were top driven and not magnetically coupled to the motor at the bottom of the vessel. Further investigation would be required in that case.

This experiment included an estimation of the force being exerted on the shaft during fermentation by using viscous liquids, as the experiment required a torque meter to measure the resistance. The vessel was set up by removing the top plate and fitting an all-in-one shaft and torque meter into the vessel.

The gassed-to-un-gassed ratio is dependent on the impeller type and gas flow rate. It has been observed that blade thickness, depth, and location also affect the gassed-to-un-gassed ratio (Gill *et al.*, 2008a; Betts *et al.*, 2006).

The estimated power number for the water environment in Figure 11 in turbulent flow was between 4.8 and 5.9, with an average of 5.39 for a single

impeller. However, this value was slightly lower than that found in the literature of 5.7 (Nienow *et al.*, 1994). It was reported that the geometric parameters of the propellers could be characterised with respect to their effect on the power number (Bujalski *et al.,* 1987, as cited in Gill *et al.*, 2008; Rutherford *et al.*, 1996). A correlation between the blade thickness and the diameter impeller was described by Rutherford *et al.* (1996), with a new power number in the following form:

$$Np = 6.57 - 54.771 \ (B_t/D_i)$$

This calculation using a blade thickness of 1.6 mm and a 38-mm impeller diameter gave the power number 4.26 which is much lower than the value shown in Figure 11. The increase in the additional power requirements is likely due to the thickness of the magnetic coupling exerting a higher degree of resistance. This highlights the difference and the need for further (individual) characterisation of the power input considering the number and type of the agitators, and magnetic coupling. It will be necessary to characterise the power requirements in terms of the location of these impellers on the shaft (Caberet *et al.*, 2008).

3.4.4 Volumetric Mass Transfer Coefficient Determination

The static gassing out method used relies on the use of polarographic electrode/oxygen sensors. They accurately measure the partial pressure of the dissolved oxygen once conditioned and calibrated. However, there is a limitation where sudden changes to the partial pressure will experience some degree of delay in measurement, which is referred to as the probe response time. For these sets of experiments, it was assumed that the mixture was set under well mixed conditions. The volumetric oxygen mass transfer coefficient, $k_L a$, was determined from the measured dissolved oxygen–time profiles accounting for probe response time according to (Gill *et al,* 2008) in equation 22;

$$C_p = \frac{1}{t_m - \tau_p} \left[t_m \exp\left(\frac{-t}{t_m}\right) - \tau_m \exp\left(\frac{-t}{\tau_p}\right) \right]$$

Equation 25

C_p is the normalised dissolved oxygen concentration measured by the probe at time t, t_m equals $1/k_L a$ and τ_p is the probe response time (calculated from experimental findings see Figure 15, Figure 16, Figure 17 at 37 °C) (Gill *et al.*, 2008). This Equation 22 was used as the first order solution and was necessary to obtain the correct mass transfer coefficient. Other assumptions made here were that the position of the probe in the reactor was suitable and any other hydrodynamic conditions were transient during the measurements.

After a DOT response curve was obtained, a chosen section of the curve was used to calculate the $K_L a$ based on empirical data for a given set of conditions. Firstly, the first-order correction was applied after which $K_L a$ could be calculated. It is important to note that no more than 30% of the data points

were excluded from the lower end of the response curve as recommended in the literature (Garcia-Ochoa *et al.,* 2009).
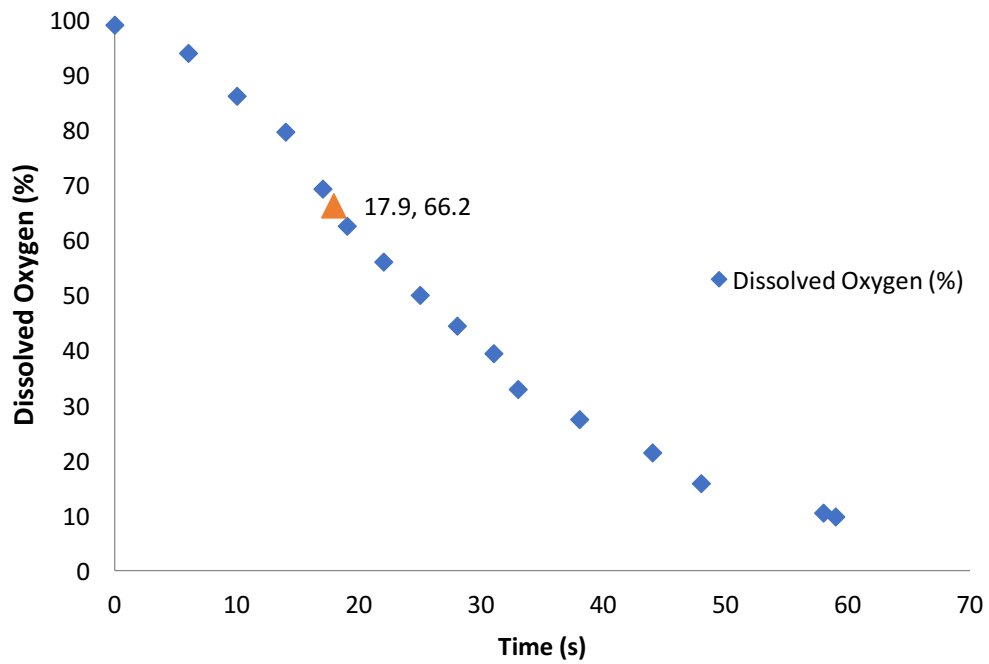
Figure 15 Probe response time for parallel bioreactors were calculated to determine whether a correction would need to be applied. The response curve can be determined via experimentation by a step change from an aerated environment to an immediate non-aerated environment and measuring the Dissolved oxygen over time. The orange triangle is the calculated probe response time to be used in the correction equation 22.
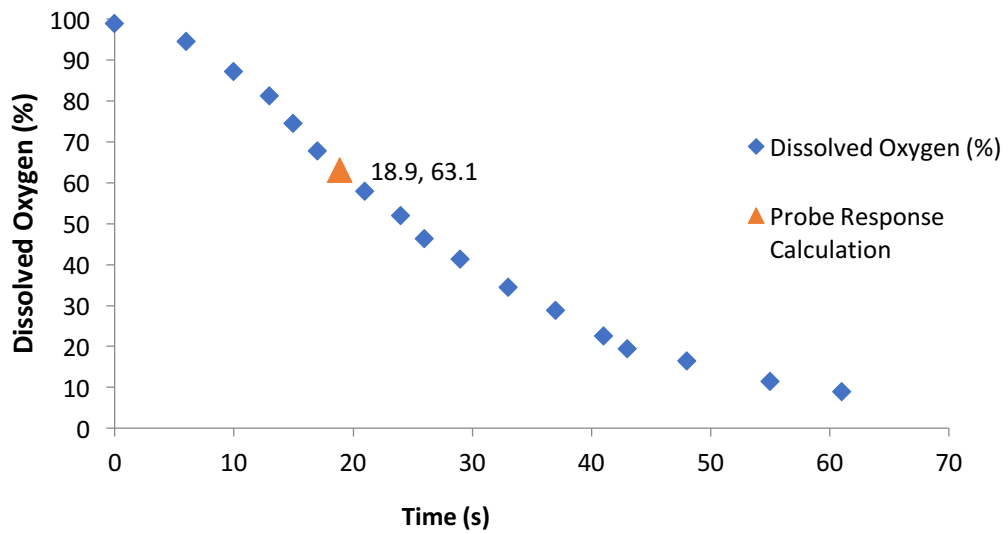
Figure 16 Probe response time for parallel bioreactors were calculated to determine whether a correction would need to be applied. The response curve can be determined via experimentation by a step change from a aerated environment to an immediate non aerated environment and measuring the Dissolved oxygen over time. The orange triangle is the calculated probe response time to be used in the correction equation 22.
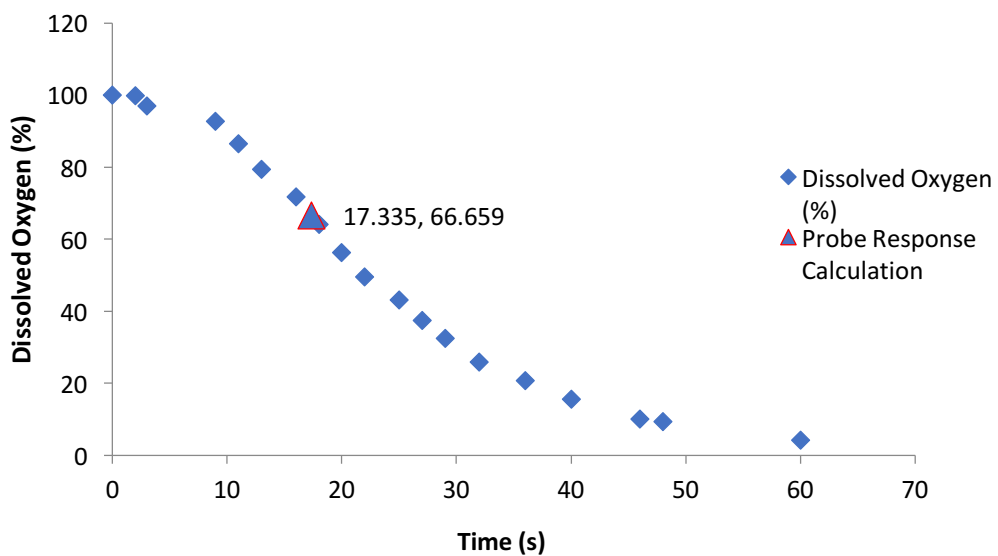
Figure 17 Probe response time for parallel bioreactors were calculated to determine whether a correction would need to be applied. The response curve can be determined via experimentation by a step change from a aerated environment to an immediate non aerated environment and measuring the Dissolved oxygen over time.. The orange triangle is the calculated probe response time to be used in the correction equation 22. The difference of these three figures are each reactors probe response time.

Probe response times were calculated above in Figure 15, Figure 16, and Figure 17 with the highest calculated value of 18.9 s at the highest 1200 rpm agitation rate. This agitation was supplied by the magnetically coupled bottom-driven stirrer bar. The highest stirrer speed was used owing to the known effect that agitation has on probe response times (PRT) where higher rotational speeds have been previously shown to decrease the PRT. This effect is likely due to the gas-liquid layer boundary getting thinner as agitation increases (Betts *et al.*, 2006).
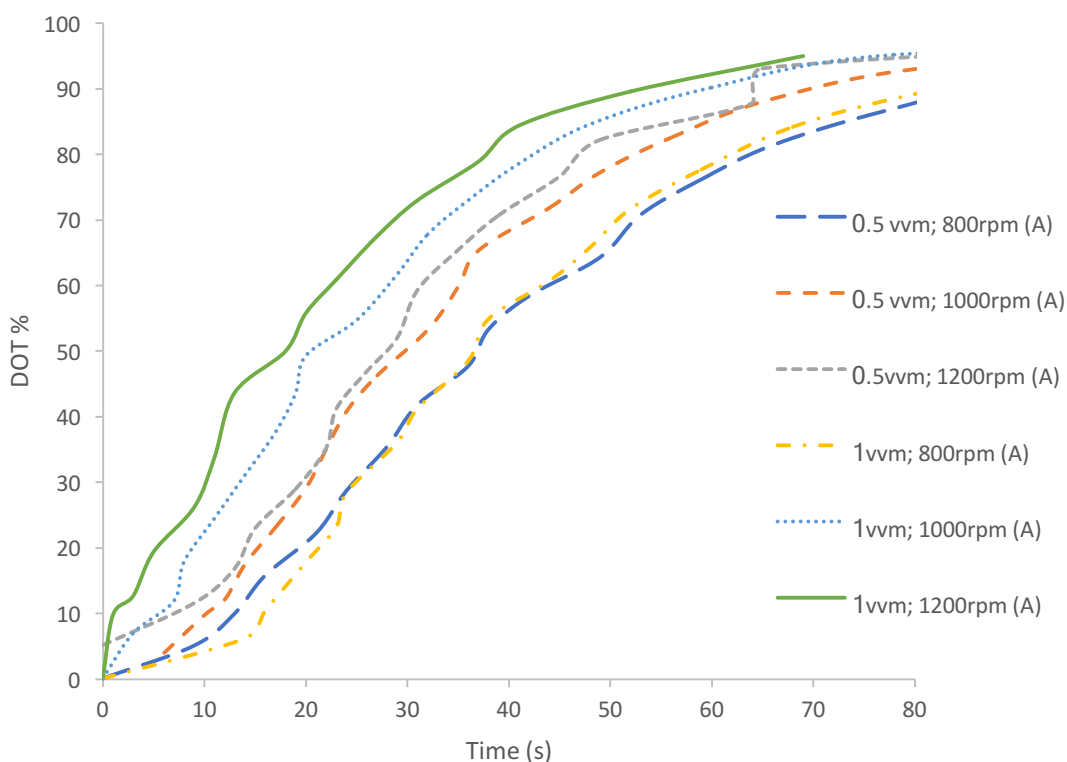
Figure 18 Measurement of the DOT% with respect to the effect of changing VVM and RPM on the oxygenation of the Multifors vessel. Experiments were performed at 37 degrees Celsius in water.

Measurement of the DOT with respect to the effect of changing aeration and agitation rates on the oxygenation of the Multifors vessel was measured next (Figure 18 and Figure 19). Figure 18 and Figure 19 illustrate the two different types of systems: using either water or defined M9 salts medium that exhibit significantly different lengths of time (averaging M9 results to be >1.5 longer

than the water experiments) to reach the DOT plateau. When the probe is precisely calibrated, the liquid is almost completely saturated with oxygen.

For the water system, the profiles and the curvature of the response curves are characteristically like those in other studies, whereby the changes in agitation have the largest impact on the DOT. This means the DOT curves of function of agitation rate become steeper at a faster rate than the DOT profiles as function of aeration rate. Again, this is probably due to the thinning of the liquid boundary (Betts *et al.*, 2006).
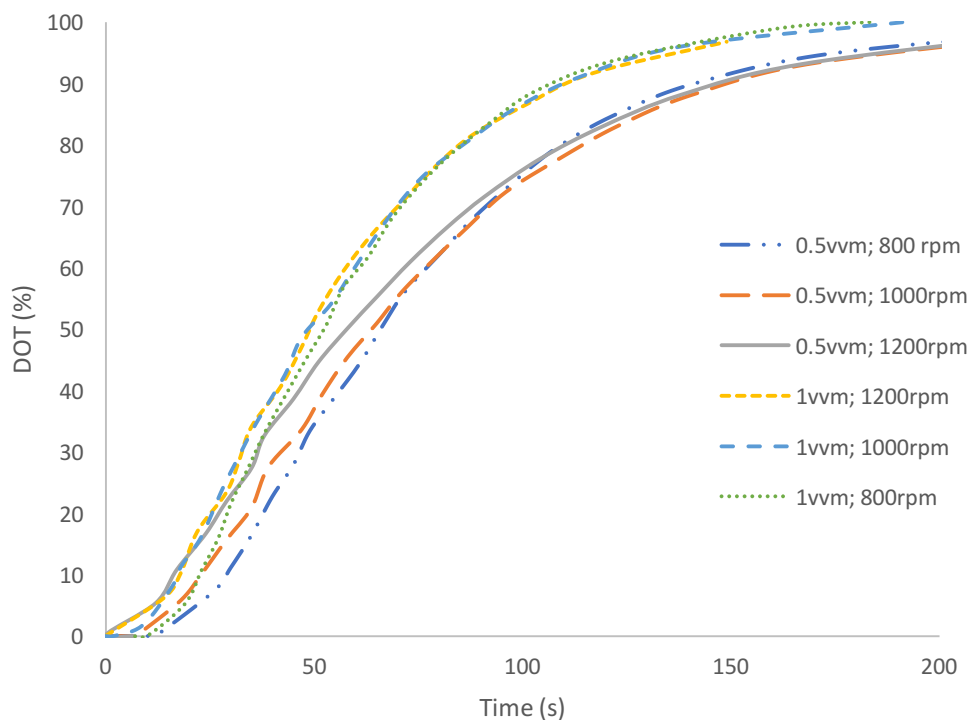
Figure 19 Measurement of the DOT% with respect to the effect of changing VVM and RPM on the oxygenation of the Multifors vessel. Experiments were performed at 37° Celsius in defined M9 media.

The media system in this case exhibited a larger impact when changing the aeration rate while keeping the agitation rate constant. This was likely since the non-aqueous surfactant layer (which would be typically used in a fermentation to prevent foaming) would create resistance to the transfer of

oxygen from gas bubble to liquid. Since a higher agitation rate creates a larger surface area for the surfactant to impact mass transfer, it is likely to impede the mass transfer. Therefore, more gas bubbles were present at the higher aeration rates, resulting in better mass transfer under these conditions.
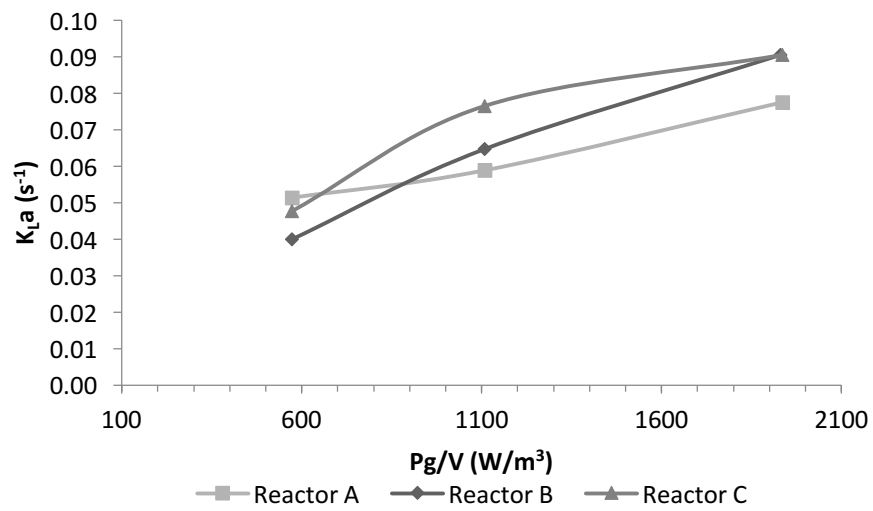
Figure 20 K$_L$a determination for 1 L at 1 vvm. Calculated mass transfer coefficients against power input for the in-parallel bioreactors using water as the fluid in the experimental set-up.
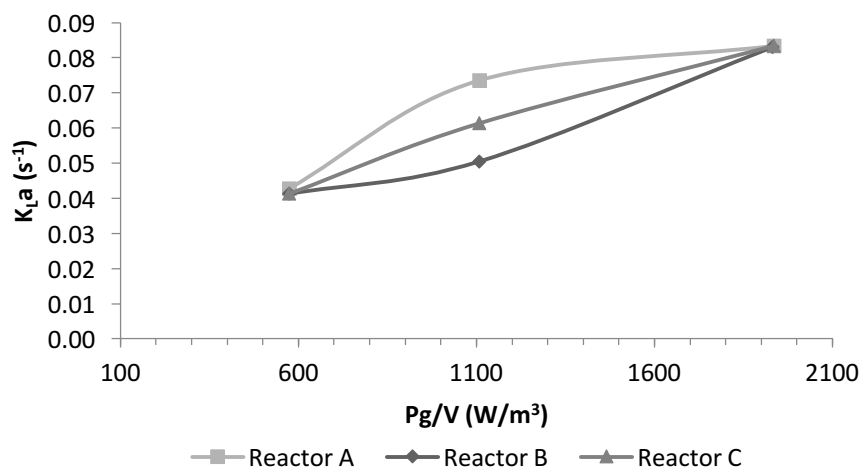
Figure 21 K$_L$a determination for 1L at 0.75 vvm. Calculated mass transfer coefficients against power input for the in-parallel bioreactors using water as the fluid in the experimental set-up..
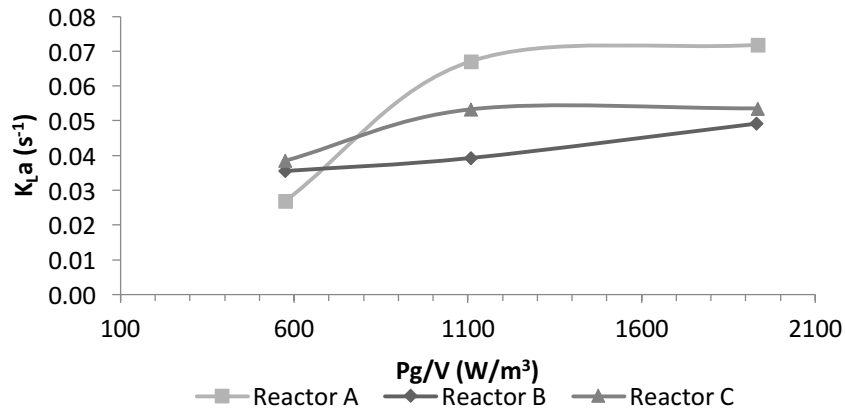
Figure 22 K_La determination for 1 L at 0.5 vvm. Calculated mass transfer coefficients against power input for the in-parallel bioreactors using water as the fluid in the experimental set-up..

Experimentally derived K_La values, over the range of conditions used, resulted in a close linear trend. Figure 20 and Figure 21 show an approximate linear trend for all three reactors. In Figure 20 there were high aeration rates providing gas bubbles and in general increasing the potential available oxygen to the system ready to be dissolved, however the agitation has had a larger effect upon the K_La. Indicating agitation having a greater impact on the oxygen mass transfer. This we measured an increase in the DOT% over time in Figure 18, when agitation rates were kept the same and aeration rates were increased. The DOT% over time effect with the most impact was at the higher (1000, 1200 rpm) agitation speeds.

Table 2 Corrected volumetric mass transfer coefficient averages, standard deviation, and percentage error for 1 vvm and ~1200 rpm conditions in the water system at 37°C

	~1200 rpm	1 L/min	Alpha	Beta			
	Pg/V	Vs			K	$K_La=K((Pg/V)^\alpha) \times (Vs^\beta)$	
							K_La (per hour)
Reactor A	1935	0.0026	0.564	0.379	0.013	0.104	373
Reactor B	1930	0.0026	0.539	0.479	0.028	0.095	343
Reactor C	1954	0.0026	0.460	0.566	0.083	0.094	338
Reactor D	1935	0.0026	0.434	0.519	0.082	0.099	357
						Average	353
						Standard deviation	13.48
						Percentage deviation	3.82

Further experiments were conducted using the four bioreactors in parallel. The objective was to show the reproducibility with statistical relevance. Table 2 shows an average K_La of 353/h with the highest value being 373/h. Furthermore, overall differences between the reactors were experimentally determined to be at 3.82% deviation. Deviation was calculated by subtracting the experimental K_La value from the mean and then dividing the result by the experimental value and multiply by 100. Since the deviation was less than 5% it was a suitable threshold of significance, the differences between the K_La values were insignificant for the four bioreactors. These values show good reproducibility across the reactors.

The volumetric oxygen mass transfer coefficient (K_La) was described in the fundamentals section as the reciprocal measure of time for the transfer of oxygen from the gaseous to the liquid phase (Meusel *et al.*, 2016). The investigation looked at the stirrer speed and airflow rate on K_La. It is known

that these parameters affect the $K_L a$ through changes to gas bubble size, number of gas bubbles, and gas distribution. Since many fermentation processes require oxygen as a substrate for respiration, this study used aerobic microbial fermentation, as it was important to determine the efficiency in the Multifors bioreactors. This was done to enable comparison of the criteria for scale-up.

It is important to note that no more than 30% of the data points were excluded from the lower end of the response curves, as recommended in the literature (Garcia-Ochoa & Gomez, 2009). Calculating the $K_L a$ from the response curve should be carried out at the steepest parts of each response curve. The beginning of the curve is flatter making it not entirely useful in the calculation of the $K_L a$ profile. This leads to some concerns about the accuracy of the $K_L a$ estimation, especially where aqueous and broth solutions are not properly represented in the $K_L a$ estimation.

During fermentation, liquid broth would contain salts and since it is a non-coalescing liquid, it is likely to experience a decrease in oxygen solubility and have a detrimental effect on the overall mass transfer coefficient (Garcia-Ochoa & Gomez, 2009). It is best to characterise the system using both water and media to better represent the likely conditions of fermentation.

3.4.5 Mixing Characterisation

The aim of the mixing experiment was to characterise the time taken for fluid distribution and whether this could be used as scale up criterion and determine the impact at this scale using agitation speeds set below and above the optimal operating conditions determined from section 3.4.2 and the power input studies. The experiment was set up for mixing using either concentrated acid, base, or colour dye.

The determination of the mixing time was the focus of this study. The mixing time is related to the geometry, impeller features, size, and liquid properties of a bioreactor (Meusel *et al.*, 2016). Inhomogeneous addition of solution to a 95% homogeneity level was sufficient for determining the mixing time.
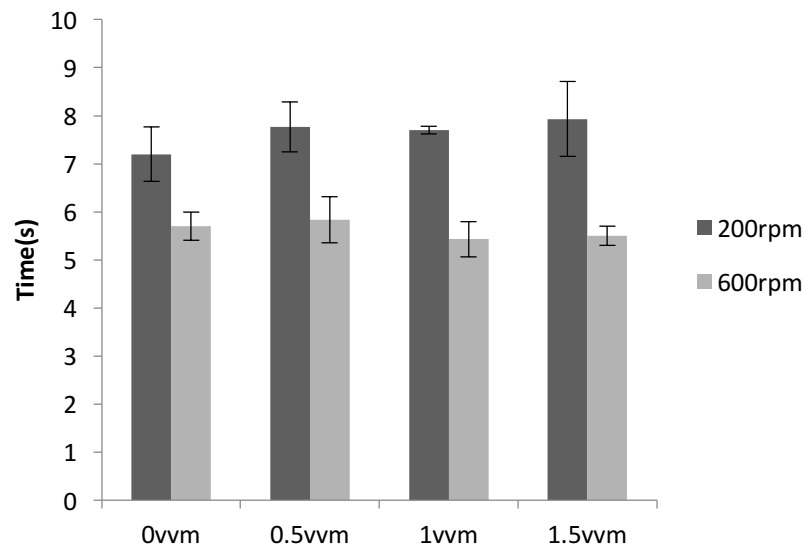
Figure 23 Mixing time using the pH global method in 600 mL fill volume (with 200 rpm, maximum Pg/V's were estimated at 57 W/m[3] with 0 vvm and 600 rpm, Pg/V estimated at 1531 W/m[3] again with 0 vvm. Error bars represent the standard deviation about the mean (n =3).
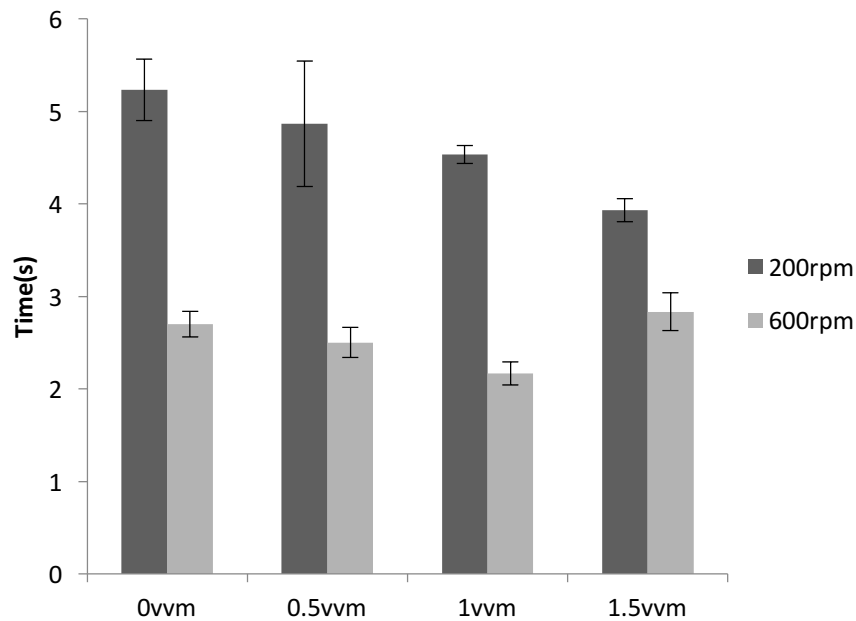
Figure 24 Mixing time using global colour method in 600 mL fill volume (with 200 rpm Pg/V estimated at 57 W/m³ and 600 rpm Pg/V estimated at 1531 W/m³). Error bars represent the standard deviation about the mean (n=3).

It is necessary to check whether the mixing time is sufficient by being well mixed and completely homogenous. This is also the assumption used in the K_La determination using the dynamic gassing out method. Figure 23 and Figure 24 demonstrate that with an increase in the P_g/V, the mixing time was significantly reduced. The effect of the aeration rate on mixing time was minimal at higher speeds, in correlation to the gas flow characteristics being above the minimum speed for complete gas dispersion. At lower speeds, the increase in aeration rate allowed for a slight reduction in the overall mixing time, as shown in Figure 24.

In this research section, various methodologies were used to characterise the vessel. The aim was to determine what the bioreactors can achieve in terms of mixing and mass transfer under expected operation conditions and the ideal criteria for scale-up. So, firstly established correlations that improve the understanding of the equipment with power input, volumetric mass transfer,

and mixing and how each element affects the performance of the bioreactor, i.e. power needed to maintain rotational speed under stress and the break-up of coalescing bubbles used for mass transfer. The power number was calculated specifically with the three impellers, i.e. two Rushton-type impellers and one non-standard impeller, because of the inclusion of the magnetic stirrer bar acting as an impeller when being driven by the motor, as it is thought to be contributing to the power input into the liquids involved, especially at higher viscosities.

Gas flow was examined where the minimum speeds for sufficient gas dispersion were calculated from the flow number. A parity plot was used to show the comparison of an empirical gassed power correlation (Hughmark, 1980) to the experimental values measured. There was a fairly good fit to the experimental data, but it better described liquids with viscosities higher than just typically water, which is relevant in the culturing of cells. Again, the Pg/V was recommended as the criterion for scale-up in future fermentations.

A limited set of conditions were tested which would likely be used for mixing and controls for cell culture tests designed for water, glycerol, and defined media. We included the media environment to make the tests relevant in this section which involves the use of microbial cell culturing. This section concludes with the understanding of significant consistency between the multiple bioreactors ready for cell culturing in parallel operation.

4 Model-based Experimental Design Screening Experiment

4.1 Introduction

4.1.1 Addressing the Research Problem

The next question or research area addressed is "What can we do next to take advantage of the parallel bioreactor system? What can we improve upon?" A model-based design specifically geared towards taking advantage of the Multifors capacity will be established in this chapter.
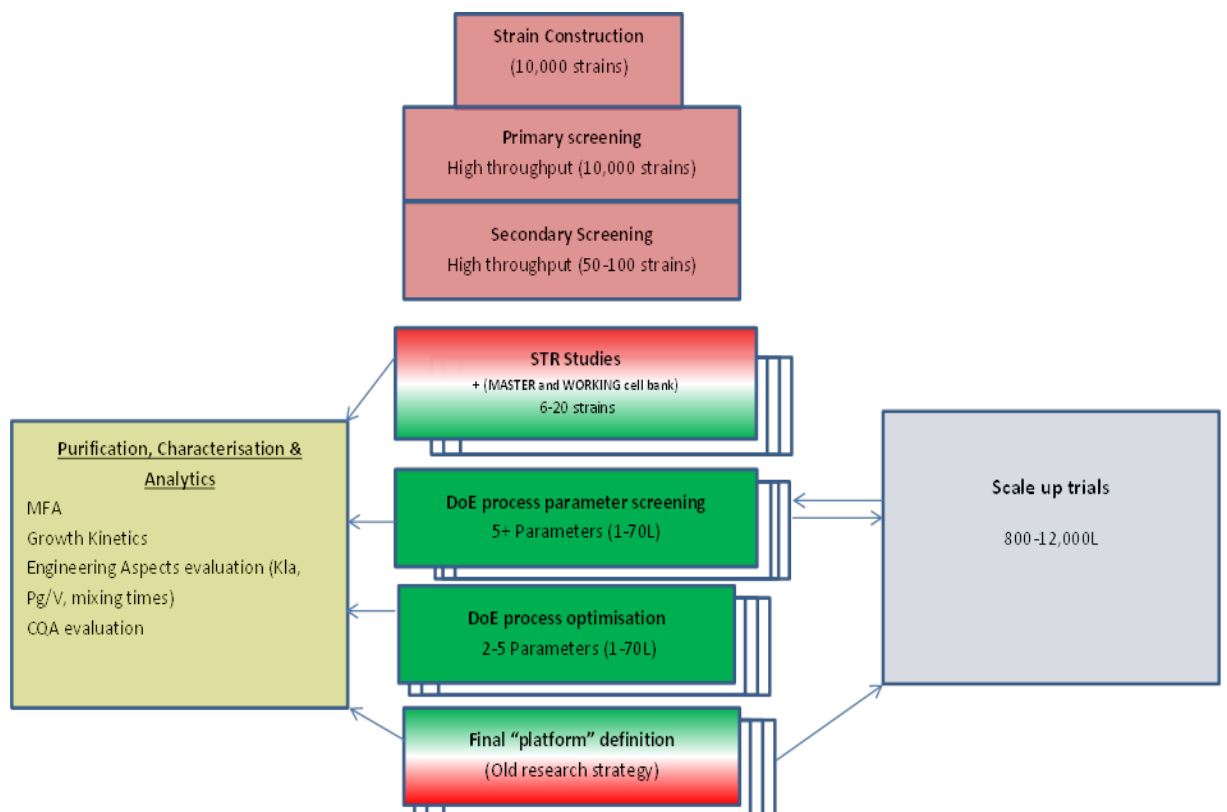
Figure 25 A review of advanced small-scale parallel bioreactor technology for accelerated process development: current state and future need (Adapted from Bareither, R. & Pollard, D., 2011).

The Multifors is perfectly suited to "close the gap" between high throughput technologies and pilot plant operations outlined in Figure 25. Small-scale parallel bioreactor technology can help accelerate process development. Relevant insights are needed to make good assumptions and research conclusions so that significant parameters that affect culturing/operation can be identified at this stage. We can establish a working model and then seek to improve upon the knowledge gained from this section.

In this research section, parallel bioreactors for the next platform hardware for rapid experimentation and in turn lead to rapid bioprocess development were used. Here it was proposed to integrate DoE software tool based in Matlab for specific application for this parallel bioreactor format.

4.2 Fundamentals of Experimental Design

DoE is a statistical tool used in the planning of an experiment. Input process variables are typically called factors. DoE has been successfully used to reduce large number of variables to meaningful, biologically relevant information (Wechselberger *et al.*, 2012). This refinement of critical variables allows for better model fitting and mechanistic characterisation. Parallel systems like the Infors HT Multifors speed up the overall time frame for DoE experiments when the design is implemented effectively. This technology is somewhat comparable to that of parallel micro-well technologies described by Lye *et al.* (2003). There is an opportunity to integrate DoE into parallel systems.

4.2.1 DoE: Factors, Responses, and Design space

DoE is also useful when the user plans an experiment and the operating variables may be controlled, but the effect on the output response is unknown. DoE aims to elucidate the unknown effects in a type of process

71

model called "black box" model. In some cases, this could be carried out in fewer fermentation processes than first calculated. Experimental models link input variables and output responses and move from conceptualisation to using real data. Responses can be numerous. Typical modelling-type responses encompass

- growth kinetics,
- mass transfer, or
- product formation.

It is necessary to quantify and analyse samples and data effectively. The effectiveness of the tool is only limited to the input of data and statistical relevance in the first place. It is important to assess this in the planning and analysis stages. Furthermore, effective DoE may occur when a user intends to change or control a small number of process factors. Some examples have been identified before by Islam *et al.* (2007) for biologically relevant factors (i.e. media composition). Engineering-based factors (i.e. stirrer speed, geometry, or temperature) are also relevant. Wechselberger *et al.* (2012) also focused on a DoE design to establish a regime for the feeding of *E. coli* expressing recombinant proteins.

A DoE tool provides the exact set-up and plans for exploring a design space. This is a window of operation which shows the dimensional factor interactions. Although it may be preferential to use a small set of factors, DoE can be used to handle a larger design space and many factor interactions (one that has many variables and/or levels). This reduction of factors is commonly known as the screening design.

For example, a 3D space would make use of 3 factors and 3-"level" design in the experiment (e.g. one main and two other interacting variables). Each level represents a calculated upper, lower, and middle operable parameter. These should be realistic and relevant.

A scaling factor and the "real" conditions are converted to create these new statistical levels for the chosen design space (the black dots in Figure 26). These conditions are normalised operating parameters. Screening designs can provide solutions for identifying significant factors from a list of many potential ones. If the purpose is to identify significant main effects, rather than interaction effects then there are several design options available. Fractional factorial designs are one example of this.

Next optimisation designs are the next logical step in identifying interactions between the important factors. The basic design options for optimisation are summarised below:

- Box-Behnken designs are good for reducing the number of runs and predicting a response curve at the centre of the design space and not at the corners.

- Central composite designs are advantageous for better modelling at the corners of the design space.
- One factor at a time (not a DoE design) is a classical design that does not explore the full range and number of resources in a short time (Figure 26).

Factors are firstly chosen by the operator. The aim of the experiment would be to find the main and interaction effects, for example temperature, media composition or inoculation volume. The advantage in all the methods is the inclusion of a statistical model using analysis of variance (ANOVA). The statistics provide a way of basing predictions and behaviour of significance and provides a mathematical model in the end. The limitations for modelling lay upon a black box model. Ultimately the design does not detail the mechanistic behaviours or chemical pathways of a complex biological system, for example biomass formation, protein production, of protein glycosylation patterns well enough. Overall the DoE methodology relies on feeding the mathematical model inputs and experimental outputs data. Therefore, it is advisable that this black box modelling should be used as an initial study for rudimentary relationships.

The "black box" models may contain several continuous or discrete factors that should be identified and used as initial conditions for experimenting, measurable outputs are collected as data. Data collected from the experiments known as 'outputs' are uploaded to the software, transformed, polynomial models are assessed and qualified using a statistical significance test and then resultant data and graphs are interpreted by the user. Most experimental models contain terms of the first and second order in their formulas leading, respectively, to creating a new linear or quadratic empirical model for a given biological system.

A user could draw conclusions and identify any of the main effects, interactions, or quadratic effects from a statistical model. It is important to

remember that experimental error is likely to be part of a model, so randomisation of the order in which experiments are run can help mitigate this inherent effect. To help with reducing the error introduced into the design randomised distribution of operating parameters and inclusion of a number of centre points can be used. There is a need for repeat centre points, specifically where the effects caused by uncontrollable process variables are mitigated as much as possible. Overall, if the model is insufficient, then it would signal the need to further characterise the process, requiring more money and resources (Konstantinidis *et al.*, 2012).

4.3 Set-up for Screening Experiment and Creation of the Theoretical Model

The objective of this screening experiment is to build a statistically relevant mathematical model that can help identify the main operating variables effecting biomass production. The resultant model would ideally provide the basis for bioreactor operation for later simulation and rapid process development.

The choice of DoE design was the fractional factorial design (FFD) at a resolution IV design. This level of DoE design was chosen because it can estimate up to four main effects. It is advantageous as none of the individual main effects are aliased with two-factor interactions. However, other two-factor interactions are aliased with each other. Therefore, the initial model will have five terms: the intercept term and with four main effects up for investigation.

Below are the experimental set up for the chosen fractional factorial design, including the output responses and input variable factors. Calculation dictates for this chosen 2 level- FFD IV design to have a total of 8 observations. These

observations are calculated by two (the number of set levels to be investigated) to the exponent power of whatever resolution of the experimental design chosen. In this case a resolution of IV equals three. Two to the power three equals a total of eight observations in this design choice. The summary is listed below in table 3, describing the set up for this experiment.

Methods and materials are briefly described here for use in *E. coli* fermentations which were carried out according to the standard operating procedures provided in the appendix section 8.1. Rotation speed, air flow via feedback dissolved oxygen (DO) measurement using a polarographic electrode (manufacturer), pH using a pH probe and no antifoam (PPG) was controlled by the Infors IRIS software, linked via a OPC connection to the Multifors touch screen control panel. The standard operating procedure and the batch fermentation parameters and conditions used for the DoE, was initialised by the software tool in Table 4.

Table 3: Description of experimental design choices

	Values	Description
Suitable Fractional factorial design	$2^{(4-1)}$	No main effects are confounded with any 2-factor interactions; main effects are confounded with 3-factor interactions.
Number of levels for each factor	2	
Number of factors	4	
Number of observations	8	
Resolution	4	

Response Variables:

1) Response Variable Y = Optical Density
2) Response Variable Y2 = Packed Cell Volume

Experimental input factors:

1) Factor 1 (X1) = Glycerol Concentration (final concentration of carbon source used in the rapid growth experiment, levels were 60 and 100 g/L, centre point level of 80).

2) Factor 2 (X2) = Temperature (a controlled heating system built into the Multifors provides heating to the glass vessel; the levels were 30 and 37°C, with the centre level of 33.5°C).

3) Factor 3 (X3) = Inoculum cell concentration (the seed train concentration being grown to a certain OD before inoculating the fermentation vessel; levels were 10% and 20%.

4) Factor 4 (X4) = Trace element inclusion (this was a categorical include or exclude trace element, levels were 0 and 1).

The Design Matrix is shown below in (randomised) run order

Table 4 Design matrix for batch fermentation of *E.coli* using different glycerol conc., temp., inoculum conc. and addition of trace elements.

Run	"Glycerol Concentration"	"Temperature"	"Cell Concentration"	"Trace Elements"	"Original Run Order"
1	100	30	0.2	1	6
2	60	30	0.1	0	1
3	100	30	0.1	1	5
4	60	37	0.1	0	3
5	100	37	0.1	0	7
6	60	30	0.2	1	2
7	100	37	0.2	1	8
8	60	37	0.2	0	4

4.4 Results of Screening Experiment

We plotted the raw data in several ways to check if any trends or anomalies appear that would not be accounted for by the models.

- A quantile-quantile (Q-Q) plot, plots the sample quantiles of OD versus theoretical quantiles from a normal distribution. If the distribution of the input samples is normal, the plot will be close to linear. Also, the solid

line in the Q-Q plot joins the first and third quartiles. A dashed line extrapolates the solid line.

- A normal probability plot graphically assesses whether the data in the input sample comes from normal distribution. If the data are normal, the plot will be linear. Other distribution types introduce curvature in the plot. This normal plot uses midpoint probability plotting positions.

Three of the factors were continuous and one, "TraceElements", was discrete (i.e. categorical using 0 and 1 notation). There are no any centre points for this type of two-level design. Fractional factorial designs for two-level experiments typically have the desirable properties of being both balanced and orthogonal (Islam *et al.,* 2007).

Table 5 Model refinement, raw data in left most column, data refinement, and data transformation in the right column Each colour represents each analysis which was carried out on the resultant data, known as each 'Try'. Each colour is then processed as a gradient on how high a value it is. This makes it easier to see outliers in our data.

Response Try 1 (Optical Density)	Response Try 2 After analysis and removal of outliers	Log (Response) Try 3 after analysis and removal of outliers
3.38	3.37	1.21
3.76	3.76	1.32
3.54	3.54	1.26
16.25	1.33	0.28
62.76	Not applicable (data removed for analysis)	Not applicable (data removed for analysis)
4.11	4.11	1.41
5.16	5.16	1.64
0.17	0.17	-1.77

4.4.1 Quantile and Probability plots
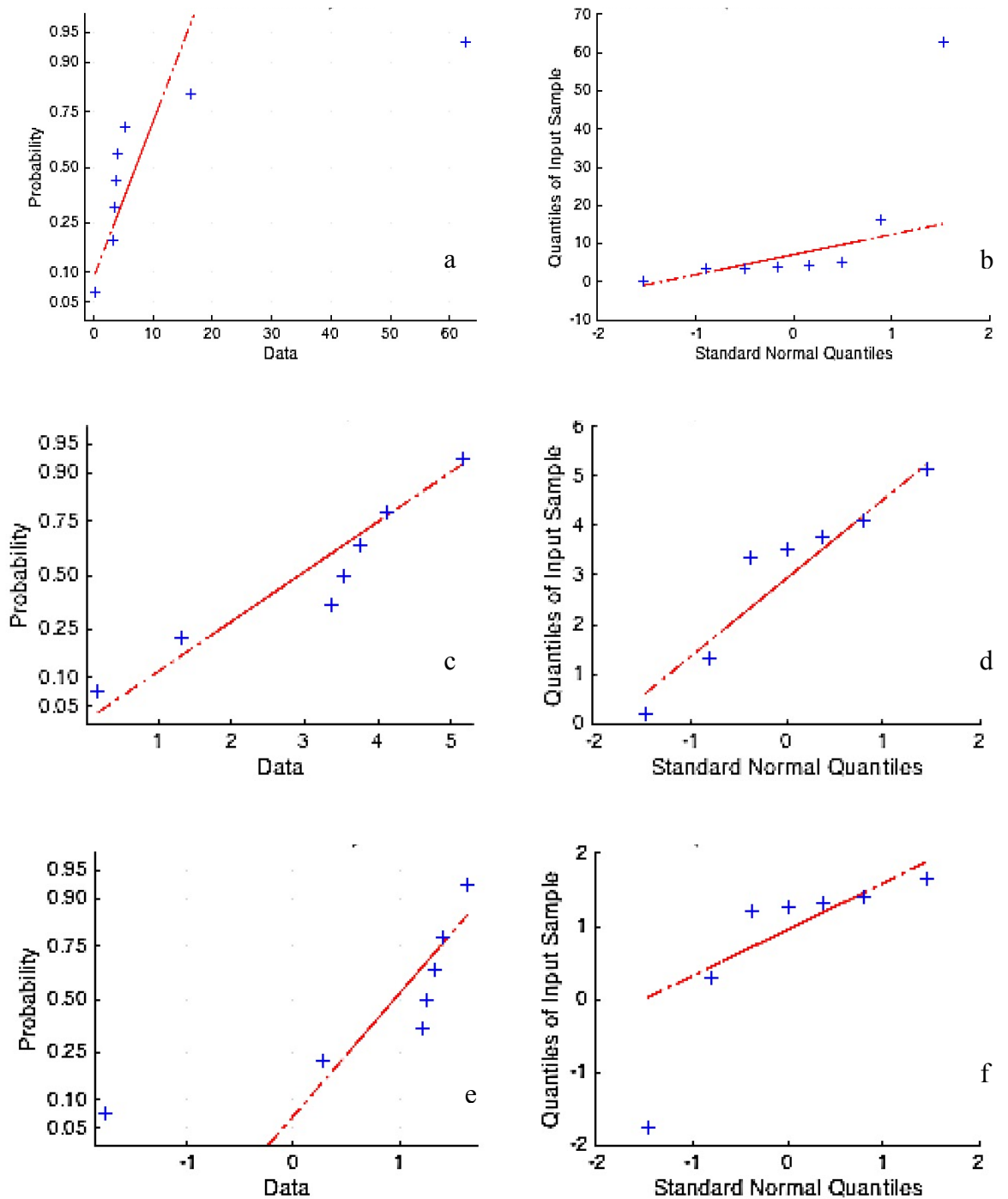
Figure 27 Normal probability plots graphically assess whether the data in the input sample could come from a normal distribution left (a,c,e). Quantile-Quantile plots show sample quantiles of optical density (input samples) right (b,d,f)

The plots in Figure 28 were divided into three models, the first two being the unrefined model (a and b), the second being the model with applying data refinement (c and d), and the third having the data transformed using the (log y) transformation (e and f).

Starting from the unrefined model (figure 28, graphs a and b):

We can see a potential non-normality within both our Q-Q plot and the normal probability plot as they both potentially display non-normal distribution. The next best step in the model creation was to remove the outliers in the data input. The advantage in excluding the outlier data is to give a much more refined mathematical model to work with. The disadvantage was the small risk of reducing the overall resolution of the model and introducing error in the interpretation of the results. Careful consideration was taken into account for this model refinement.

Using coded variables for our models

A = Glycerol Concentration

B = Temperature

C = Initial Cell concentration

D = Trace Elements

General linear regression method results (model 1)

Optical Density ~ -45.113 + 0.572*A + 1.02*B – 81.25*C - 20.5*D

~51% degree of confidence

Equation 26

81

Stepwise Linear Regression results (model 1)

Optical Density ~ -28.2183 + 0.699*A - 30.6767*D

~88% degree of confidence

With the outlier data removed from the initial model a second round of modelling started (Figure 28, graphs c and d):

General linear regression method results (model 2)

Optical Density ~ 4.527 - 0.005*A– 0.046*B – 6.65*C + 2.58*D

~32% degree of confidence

Stepwise Linear Regression results (model 2)

Optical Density ~ 1.7533 + 2.2917*D

~92% degree of confidence

With the outlier data removed from the initial model and log transformation of the data (Figure 28, graphs e and f) yielded:

On the next approach, it was thought that data transformation would help resolve this negative OD feedback issue. There were two advantages that using the natural logarithm of OD as the response might lead to a better model. A linear model fit to ln(Y) will always predict a positive distance when converted back to the original scale for any possible combination of X factor values. This would solve the 'models negative OD value' issue. Physical considerations suggest that a realistic model for OD might require quadratic terms since mixing and settling and gravity play a key role when taking offline samples, as taking logarithms often reduces the impact of non-linear terms.

So again, with same run order, row number 5 and 8 (the outliers) omitted the raw data would be transformed using natural log functions and fitting by using the R^2 value (see appendix section 8.2.2). In summary, the resultant graphs still show good data spread (see Figure 28 graph f), with data being positive when transformed back to the original data, a criterion we wanted to fulfil.

Table 6 Generalised linear regression model and formula

```
Generalized Linear regression model:
    OpticalDensity ~ [Linear formula with 5 terms in 4 predictors]
    Distribution = Normal

Estimated Coefficients:
                            Estimate        SE        tStat       pValue

                            _____     _____    _____    _____

    (Intercept)                3.302      5.2823       0.6251     0.59572
    GlycerolConcentration  −0.0072238    0.039793     −0.18154    0.87268
    Temperature             −0.043801     0.17613     −0.24868    0.82681
    CellConcentration         −10.532      12.329      −0.8542     0.48298
    TraceElements_1            1.9653      2.0134       0.97609     0.43196


7 observations, 2 error degrees of freedom
Estimated Dispersion: 1.52
F-statistic vs. constant model: 0.917, p−value = 0.581
```

83

```
                        SumSq      DF      MeanSq        F         pValue

                        _____      __      _____      _____      _____

        Total           8.6186      6       1.4364
        Model           7.9877      3       2.6626      12.661      0.032875
        . Linear        5.3462      2       2.6731      12.711      0.034293
        . Nonlinear     2.6415      1       2.6415      12.561      0.038252
        Residual        0.6309      3       0.2103
```

Table 7 Final Linear regression model and formula

```
Linear regression model:
    OpticalDensity ~ 1 + CellConcentration*TraceElements

Estimated Coefficients:
                                        Estimate      SE        tStat      pValue

                                        _____    _____   _____   _____

    (Intercept)                          3.3816     0.79429     4.2573    0.023762
    CellConcentration                   -25.768      5.6165    -4.5878    0.019451
    TraceElements_1                      -2.2764     1.2419     -1.8331    0.16417
    CellConcentration:TraceElements_1    27.357      7.7191     3.5441    0.038252


Number of observations: 7, Error degrees of freedom: 3
Root Mean Squared Error: 0.459
R-squared: 0.927,   Adjusted R-Squared 0.854
F-statistic vs. constant model: 12.7, p-value = 0.0329
```

Using coded variables for our models

A = Glycerol Concentration

B = Temperature

C = Initial Cell concentration

D = Trace Elements

General linear regression method results (model 3)

Ln(Optical Density) ~ -3.302 -0.007*A -0.043*B − 10.532*C + 1.965*D

~42% degree of confidence

Equation 30

Stepwise Linear Regression results (model 3)

Ln(Optical Density) ~ 3.382 + 27.357*(A *D)

~96% degree of confidence

Equation 31

4.5 Discussion

Simply put the model that gave the factor interactions that had the highest significance, based from our experimental results was model 3. Model three experimental data were pre-processed for outliers in the data and logarithmically transformed. The methods for all of our models were calculated using either the general linear regression method (GLRM) or stepwise linear methods. The general linear regression method calculates a generalisation of multiple linear regression models. The 'stepwise' method starts with a constant, and adds or subtracts terms one at a time, choosing an optimal term each time, until it cannot improve further. Stepwise fitting was used to find a good model that had only relevant terms., since this method gives a higher degree of confidence.

The original criterion for the model was that it should be logical i.e. that the estimate OD output response values from the model would result in positive OD values. To determine which method could provide more significance and meet the original criterion R^2 values were used. As with model calculation using experimental results and this 'black box' approach, there was always a risk that some comparisons may not have had favourable degrees of significance. It is upon the user to interpret the quality of the model. Further analysis of models 1 and 2 are provided in the appendix.

Recommendations at this stage were to check how significant the lack of fit was to the experimental data. We can then say with high degree of

confidence how inadequate a model being fitted to the experimental data really is in statistical terms.

After pre-processing the data for outliers (Model 2), we used the stepwise general linear modelling method to create the model from the actual data; in this instance, the non-significant terms were removed and allowed to refit the model. It was found that the R^2 and R^2 adjusted values were still not ideal. The ANOVA table showed that the model was not a significant fit, i.e. the lack-of-fit test further showed this as a significant calculation (<0.05). In ideal cases, we would be looking for no lack of fit in the model. What we would be able to do is rerun the model using repetitions and carry out some key calculations, such as average, standard deviation, and signal-to-noise ratio.

The quality of the model and results of a good model fit depend on the starting model. Starting with more terms can lead to a more complex model but one that has a lower mean squared error. The Matlab DoE toolbox is limited in that robust options (which is method used to find the optimal model) cannot be used alongside stepwise fitting. Robust fitting saves one the trouble of manually discarding outliers, so after a stepwise fit, it is necessary to examine the model for outliers

For the final model to be used (Model 3), it was possible to introduce the stepwise regression method with the transformed data. Again, this allowed the fit of the model with a much better R^2 value fit a good indicator of a model predicting real life data. There was no significant lack of fit of our model, which, as discussed, was ideal.

4.6 Rapid Microbial Growth Response Surface Model Experiment

The goal of this experiment was to optimise the fermentation process based upon the screening results in section 4.4. The plan of this experiment was to use response surface modelling to see if we could determine the optimal conditions for optical density (OD) and packed cell volume (PCV). Packed cell volume was used as a way to rapidly take samples which is inexpensive and time efficient method of determining increase to packed cell volumes during experimentation.

The response variables would be expressed as a function of variables, glycerol concentration (between 60 and 100g/L) and ratio of trace elements (0, 0.5, 1). This was obtained from the results from initial screening modelling. These mentioned factors are our controllable variables for rapid optimisation of *E. coli* fermentations, in an approx. 24-hour period of operation. Temperature was kept at 37°c and inoculum volume kept as 0.2 (v/v).
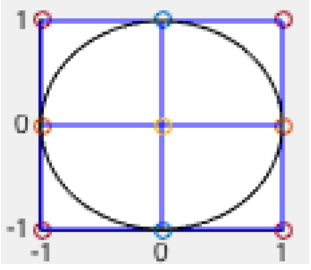
	Value	Notes
Suitable Design	Response Surface model (Faced)	
Number of levels for each factor	3	
Number of factors	2	
Number of observations	11	
Number of design points	(8)	
Number of centre points	(3)	

Table 8 Description of experimental design choices

Response Variables:

1) Response Variable Y = Optical Density
2) Response Variable Y2 = Packed Cell Volume

Experimental Variables:

1) Factor 1 (X1) = Glycerol concentration (final concentration of carbon source used in the rapid growth experiment; levels were 60 and 100 g/L. Note a centre point level of 80).

2) Factor 2 (X2) = Trace elements included ratio levels between 0 and 1. Note a centre point level of 0.5 which is set as a volume ratio of trace elements stock solution to be added. For example, the centre point 0.5 would include only half the volume of trace elements compared to a 1.0 level in the design.

4.6.1.1 The Design Matrix

The chosen design matrix from the properties given in Table 9 is outlined below. These runs appear in randomised run order. Included are the first response(s) and log(response) Y = optical density values.

Table 9 Design matrix for recipes. NaN = data excluded from the analysis.

Run	"Glycerol Concentration"	"Trace Elements"	Block Number	"Original Run Order"	Response Y = (Optical Density)	Log(Y) (Optical Density)
1	60	0	1	1	41.7	1.62
2	100	0	1	3	29.267	1.466
3	60	1	1	2	23.7	NaN
4	100	1	1	4	NaN	NaN
5	80	0.5	1	9	25.6	1.408
6	80	0.5	2	11	24.33	1.386
7	80	0.5	2	10	11.467	1.059
8	100	0.5	2	6	24.867	1.396
9	80	1	2	8	15.267	1.184
10	80	0	2	7	15.033	1.177
11	60	0.5	2	5	39.833	1.6

Since response curvature, especially for both responses, was a distinct possibility for this experiment, an experimental surface design that allowed estimating a second order (quadratic) model was needed. A central composite-faced (CCF) design was chosen for this purpose. For two factors, this design is typically recommended to have 8 runs with 5 centre point runs. However, to conserve upon a limited amount of time and resources, the design choice was chosen to include only 3 centre point runs. The design is still rotatable, but the uniform precision property has been sacrificed.

Recommended steps for fitting a response surface:

1) Fit the full model to the first response, using stepwise regression to identify important variables. When selecting variables for inclusion in the model, follow the hierarchy principle and keep the main effects that

are part of significant higher-order terms or interactions, even if the main effect p-value is larger than desirable.

2) Generate diagnostic residual plots (histograms, box plots, normal plots, etc.) for the model selected.

3) Examine the fitted model plot, interaction plots, and ANOVA statistics (R^2, adjusted R^2, lack-of-fit test, etc.). Use all these plots and statistics to determine whether the model fit is satisfactory.

4) Use contour plots of the response surface to explore the effect of changing factor levels on the response.

5) Repeat all the above steps for the second response variable.

6) After satisfactory models have been fitted to both responses, one can overlay the surface contours for both responses.

7) Find optimal factor settings.

4.6.2 Analysis of Experiment

At the start, the data were plotted in several ways to determine whether any trends or anomalies appear that would not be accounted for by the models.

In Figure 28 both the Q-Q plot and normal probability plots for the experimental runs are shown alongside the run order (which was randomised) and the box plot of the response values from the experiment. These graphs display a non-normal distribution. The run order plot clearly confirms that the CCF design was randomised without obvious problems (i.e. there is no discernible trend) except for a failed batch in Run 4. There was a wide range of data which could likely indicate and further argue that the data may have some distribution problems or lack of fit issues. In the appendix section 8.2.4 to 8.2.6.2 we investigated with further analysis data quality or suitability for our CCF model, it was found that the data could be fitted well without transforming or pre-processing the data.
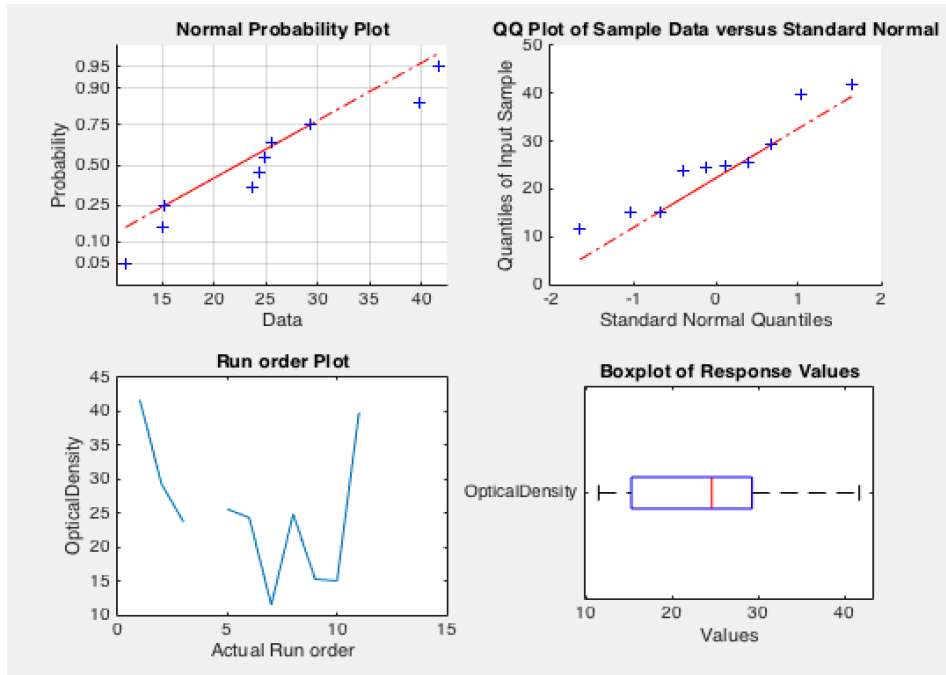
Figure 28 Data plot diagnostics for response surface design experiment. Run order plot is included so as to show that our run order was randomised so analysis of the data can be carried out.
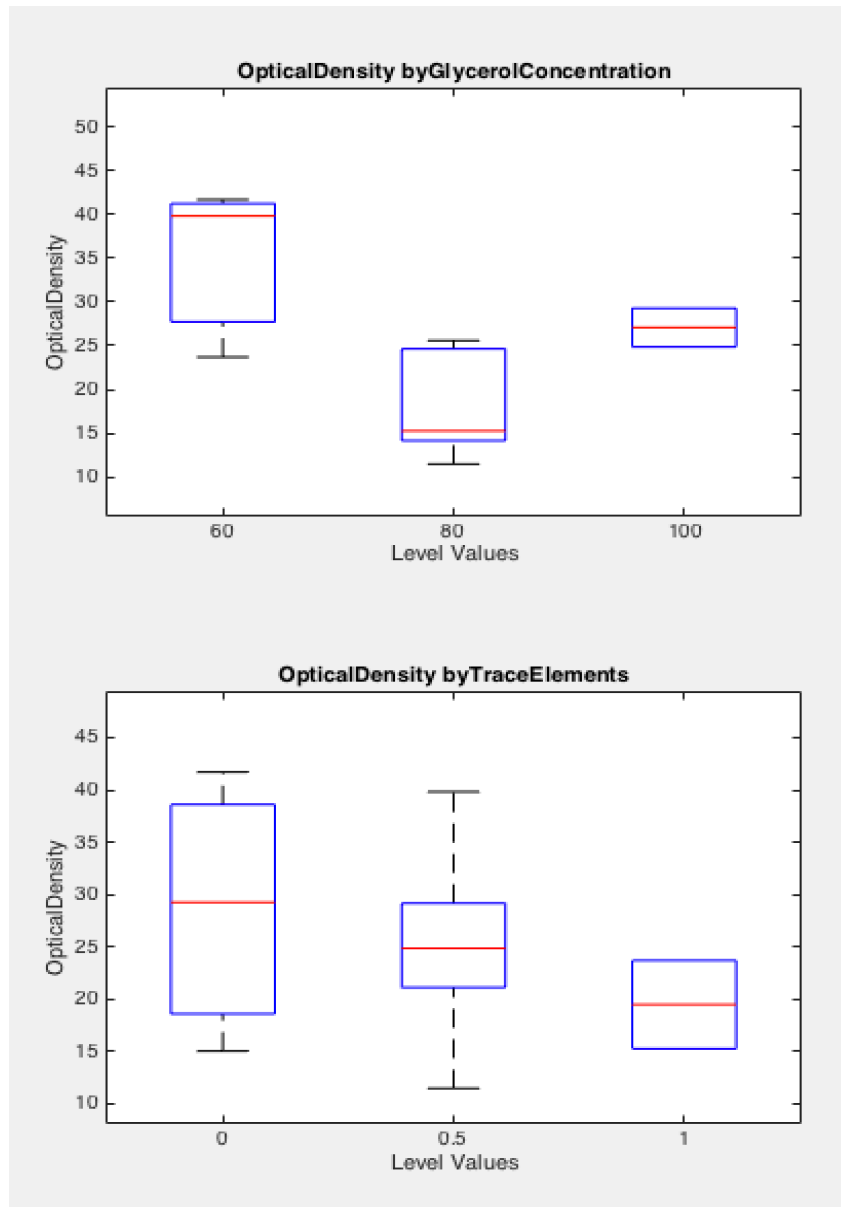
Figure 29 Output response by factor plots

Several factors appear to change the average response level, glycerol concentration showing the biggest change by far. With all response observations in Figure 29 having a large spread at each of the levels, similar trend may be found later by the analysis of residuals in appendix section 8.2.6.2.

4.6.2.1 Fitting to a Full Model

Here, the full model to the first response was initially fitted (the experimental data using robust fit and regression modelling techniques). Here, several unsatisfactory estimated coefficients and score values (the p values) occured. This gives us confidence that some of our factors have significance in this quadratic model. This depends on whether or not the p values are below 0.1 (90% confidence of significance) or 0.05 (95% confidence of significance).

Linear regression model using robust fit method

```
Linear regression model (robust fit):
    OpticalDensity ~ [Linear formula with 6 terms in 2 predictors]

Estimated Coefficients:
                                       Estimate       SE       tStat      pValue
                                       _____    _____   _____   _____

    (Intercept)                          263.5      87.085     3.0258    0.038942
    GlycerolConcentration              -5.7044      2.1869    -2.6084    0.059517
    TraceElements                       -18.124     52.667   -0.34412     0.74808
    GlycerolConcentration:TraceElements  0.33031    0.57562    0.57384     0.5968
    GlycerolConcentration^2             0.033226   0.013387      2.482    0.068063
    TraceElements^2                     -15.835     21.419   -0.73928     0.50077


Number of observations: 10, Error degrees of freedom: 4
Root Mean Squared Error: 7.74
R-squared: 0.744,  Adjusted R-Squared 0.424
F-statistic vs. constant model: 2.32, p-value = 0.217
```

In the tabulated analysis above, a relatively weak R^2 value and the R^2 adjusted value was significantly lower than the R^2 value. This indicated that the model had a weak fit with all factors and interaction terms being included in this quadratic model. Therefore, it was needed to simplify the model by using techniques known as stepwise regression methods.

Using coded variables for our models

A = Glycerol Concentration
D = Trace Elements

93

The initial model was

$$\text{Optical density} \sim 263.5 - 5.704*A - 18.124*D + 0.33*(A*D) + 0.033*A^2 - 15.835*D^2$$

~78% degree of confidence

Equation 32

4.6.2.2 Fitting the Model to Our Data using Stepwise Regression

After the initial fit, the function examines a set of available terms and adds the best one to the model. This was based on an F-test by adding the "term" which has a p-value 0.05 or less. If no terms can be added, it examines the terms currently in the model and removes the worst one, if, an F-test for removing it has a p-value of 0.10 or greater. The method was repeated until no terms can be added or removed. The function never removes the constant term. It may add terms defined by linear, quadratic, or two-way interactions among the design factors see initial model in Equation 32.

```
1. Removing GlycerolConcentration:TraceElements, FStat = 0.38881, pValue = 0.56672
2. Removing TraceElements^2, FStat = 1.2265, pValue = 0.3185
3. Removing TraceElements, FStat = 1.9454, pValue = 0.21253
```

```
Linear regression model:
    OpticalDensity ~ 1 + GlycerolConcentration + GlycerolConcentration^2

Estimated Coefficients:
                              Estimate      SE       tStat      pValue
                             _____   _____   _____   _____

    (Intercept)               238.09     70.563     3.3741    0.011853
    GlycerolConcentration     -5.2934    1.8214     -2.9062   0.022784
    GlycerolConcentration^2   0.031832   0.011485   2.7716    0.02763


Number of observations: 10, Error degrees of freedom: 7
Root Mean Squared Error: 7.19
R-squared: 0.597,  Adjusted R-Squared 0.481
F-statistic vs. constant model: 5.17, p-value = 0.0417
```

Here again, a relatively weak R^2 value (0.597) and the R^2 adjusted (0.481) was calculated. The difference this time was that the value gap was now closer, giving an indication that it was a little more reliable fit than its previous initial quadratic fitting. Of note, there is now a quadratic term in the model in Equation 33

The 'stepwise' model was

$$\text{Optical density} \sim 238.09 - 5.293*A - 0.033*A^2$$

~96% degree of confidence

Special circumstances can be explained when stepwise regression method selects a model containing the interaction term and only **one** main effect. It is widely accepted by statisticians to go back in the same stepwise fashion until all main effects are included. Since "Glycerol Concentration * Trace Elements" interaction term was removed from the regression model fitting, this was sensible for our modelling purposes.

Testing the significance of this model's fit and how confident we are against these models' lack of fit is next. Testing the model assumptions using standard residual graphs were included in the appendix 8.2.6.2.

4.6.3 Test results of our model confidence (model 1)

Table 11 Testing the confidence levels using statistical ANOVA techniques

	SumSq	DF	MeanSq	F	pValue
Total	896.73	9	99.637		
Model	534.93	2	267.46	5.1747	0.04172
. Linear	137.88	1	137.88	2.6676	0.14643
. Nonlinear	397.05	1	397.05	7.6818	0.02763
Residual	361.81	7	51.686		

From the output table 22, if a linear model were chosen to help describe the process control, then it would have had a confidence level of significance just above 85%, meaning up to an additional 7.5% (half of the remaining significance approx.) of our estimated output responses from this linear model could be known to be erroneous and could have had an impact on the process controls chosen for our final optimisation steps. The non-linear, quadratic model was a much more likely fit to our experimental data and with a higher confidence, i.e. above 95%.

4.6.4 Refining the Model (model 2)

4.6.4.1 Design Matrix and Response Values for First Response Variable

Purpose

To improve the fitting of the model by using the R-squared value and then refit the model to the experimental data after data transformation (log).

Model Refining

In some cases, it may be beneficial to transform the data. A simple transformation of the response variable (Y = "Optical Density") was tried to see if the model could be improved. The steps for model fitting were repeated for the transformed data using a log to the base 10 transformation.

A linear model fit to $\log_{10}$ (Y) would always predict a positive distance when converted back to the original scale for any possible combination of X factor values. Once again, physical considerations suggest that a realistic model for OD might require quadratic terms since mixing and settling and gravity play a key role when taking offline samples – taking logarithms often reduces the impact of non-linear terms.

We fit the initial full model using all types of interactions and main effects from the transformed data. We obtained p values that lay just below 90% confidence of significance. There are terms that are not significant in this model, so we repeated the steps involved with the stepwise regression.

```
Linear regression model (robust fit):
    Log_1_0OpticalDensity ~ [Linear formula with 6 terms in 2 predictors]

Estimated Coefficients:
                                           Estimate          SE        tStat      pValue
                                           _____      _____     _____    _____

    (Intercept)                             5.5308        1.9159       2.8867     0.04471
    GlycerolConcentration                  -0.10191      0.048114     -2.1182     0.10155
    TraceElements                          -0.096374      1.1587     -0.083172    0.93771
    GlycerolConcentration:TraceElements     0.0035316    0.012664      0.27887    0.79417
    GlycerolConcentration^2                 0.00060651   0.00029452     2.0593     0.10854
    TraceElements^2                        -0.30062       0.47124     -0.63795     0.5582


Number of observations: 10, Error degrees of freedom: 4
Root Mean Squared Error: 0.17
R-squared: 0.63,  Adjusted R-Squared 0.167
F-statistic vs. constant model: 1.36, p-value = 0.394
```

Using the same coded variables, the first round of model refinement using robust fit methods for the first response variable (model 2):

$$\text{Log}_{10} \text{ (Optical density)} \sim 5.530 - 0.101*A - 0.096*D + 0.0035*(A *D) + 0.0006*A - 0.301*D^2$$

~61% degree of confidence
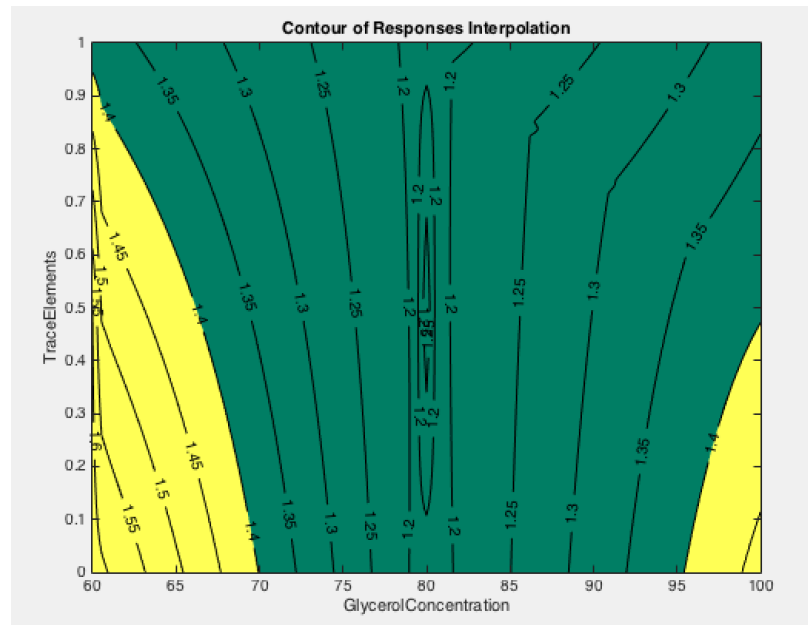
Figure 30 Contour response plot derived from model 2 in eq. 33

The contour plot in Figure 30 shows two highlighted areas. The yellow area is the operating space recommended to give above mean average response output, in this case, above the mean average output for Log (optical density). The green is the opposite: below mean average output. We can adjust this threshold later in the optimisation area and visualise the effect it has on the ideal operating area and the protocols for bioprocess control optimisation. We can use the regions highlighted to set new statistically relevant control set points on the fermentation control device for further experiments.

4.6.4.2 *Design Matrix and Response Values for Second Response Variable packed cell volume*

Below, is a summary table of the factor set points. The runs appear in randomised order. Included are the second response(s) Y2 = PCV values.

99

Table 13 Chosen design matrix summary for experiment with packed cell volume as our response variable Y. NaN are no data points for that run.

Run	"Glycerol Concentration"	"Trace Elements"	Block Number	"Original Run Order"	Y = (PCV)
1	60	0	1	1	0.025
2	100	0	1	3	0.03
3	60	1	1	2	0.05
4	100	1	1	4	NaN
5	80	0.5	1	9	NaN
6	80	0.5	2	11	0.04
7	80	0.5	2	10	0.03
8	100	0.5	2	6	0.045
9	80	1	2	8	NaN
10	80	0	2	7	NaN
11	60	0.5	2	5	0.05

We start by plotting the data several ways to check if any trends or anomalies appear that would not be accounted for by the models.

4.6.4.3 Analysis of Experiment

In the initial analysis (shown in Figure 31), it had suitable normality within both our Q-Q plot and the normal probability plot as they both display an almost linear fitting. The run order plot clearly confirms very weak data quality as 4 of the data observations were failures i.e. no data points (for run 4, 5, 9, and 10); also, with data that was included in the model, there were no obvious trends. There was a wide range of data which could likely indicate that the data has some distribution problems or lack of fit issues.
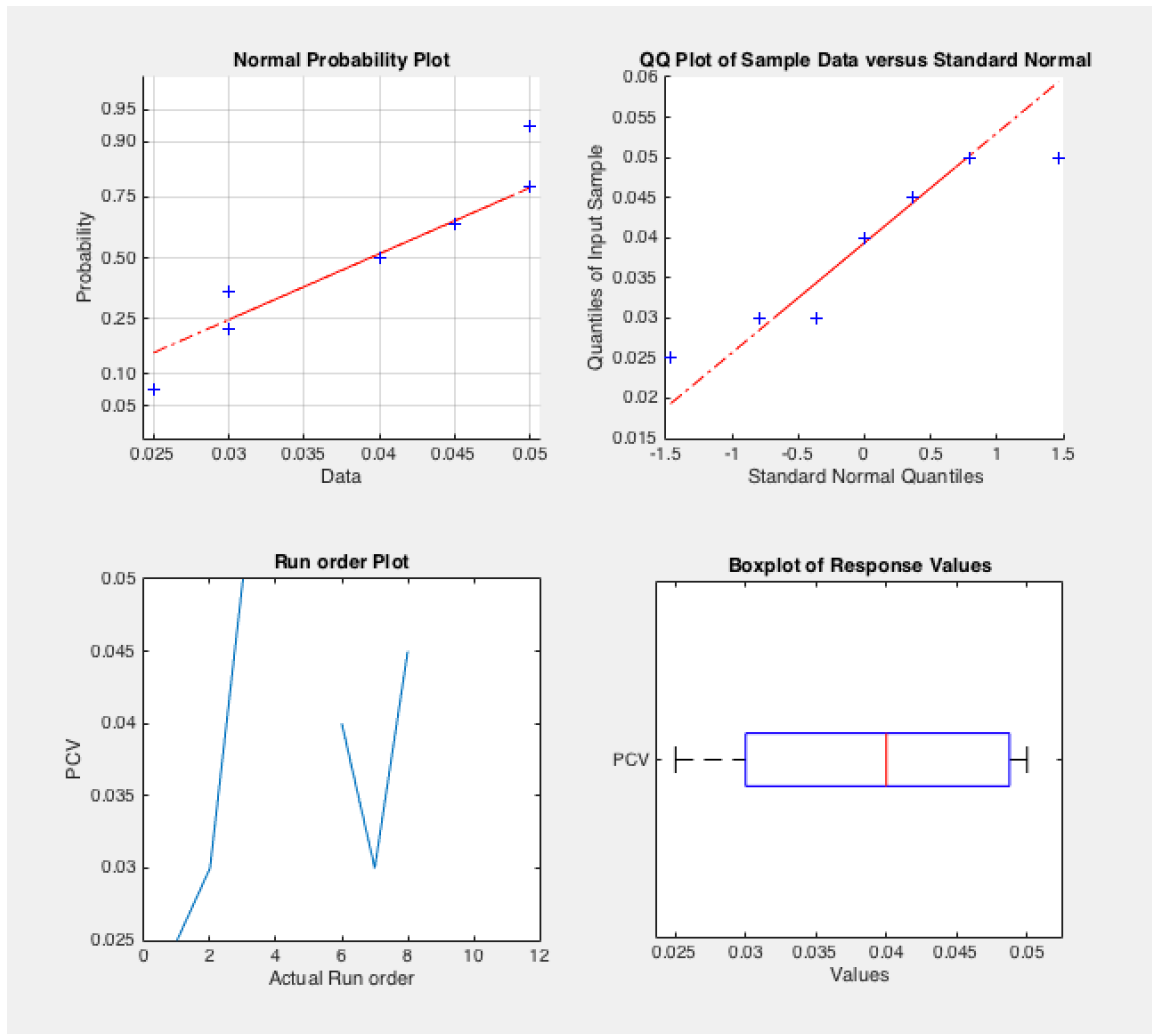
Figure 31 Data plot diagnostics for second response experimental values.

First fit the initial full model using all types of interactions and main effects from the second response data. All the presented terms and interactions presented were not significant. This could be evidence for a poor model fitting, i.e. it will be difficult to know the response output to the actual controllable factors. Still proceeding with modelling, repeating the steps involved with the stepwise regression (following up below).

Using coded variables for our models

A = Glycerol Concentration
D = Trace Elements

101

The initial model 3 build from table 14 was

$$\text{Packed cell volume} \sim 0.081 - 0.001*A + 0.071*D - 0.0005*(A*D) + 0.000008*A^2 - 0.027*D^2$$

~4.2% degree of confidence

Table 14 Linear regression model output using robust fit methods

```
Linear regression model (robust fit):
    PCV ~ [Linear formula with 6 terms in 2 predictors]

Estimated Coefficients:
                                      Estimate        SE        tStat      pValue
                                     _____   _____   _____   _____

    (Intercept)                          0.08087      0.1803    0.44853   0.68419
    GlycerolConcentration             -0.0012589   0.0045375   -0.27744   0.79946
    TraceElements                       0.071807     0.10616    0.67638   0.54728
    GlycerolConcentration:TraceElements -0.0005673  0.0011422   -0.49666   0.65354
    GlycerolConcentration^2           8.4646e-06   2.7794e-05    0.30454   0.78062
    TraceElements^2                    -0.027166    0.044471   -0.61088   0.58445


Number of observations: 9, Error degrees of freedom: 3
Root Mean Squared Error: 0.0153
R-squared: 0.218,  Adjusted R-Squared -1.08
F-statistic vs. constant model: 0.168, p-value = 0.958
```

Again, the stepwise regression method was used to model our experimental data. We attempted to eliminate the non-significant factor interactions.

```
Linear regression model:
    PCV ~ 1 + TraceElements

Estimated Coefficients:
                    Estimate        SE        tStat       pValue
                   _____   _____   _____   _____

    (Intercept)       0.0285     0.0043932    6.4873     0.0012981
    TraceElements     0.0235     0.0082189    2.8593      0.035437


Number of observations: 7, Error degrees of freedom: 5
Root Mean Squared Error: 0.00695
R-squared: 0.621,  Adjusted R-Squared 0.545
F-statistic vs. constant model: 8.18, p-value = 0.0354
```

Mathematical model 4 derived from the stepwise linear regression method:

Packed cell volume ~ 0.028 —+ 0.0235*D

~96% degree of confidence

Equation 36

Table 15 ANOVA tabulated results for second variable response and model 4

```
                     SumSq        DF      MeanSq         F         pValue
                  _____     __     _____    _____    _____

    Total         0.00063571      6     0.00010595
    Model         0.00039446      1     0.00039446     8.1754     0.035437
    Residual      0.00024125      5      4.825e-05
    . Lack of fit       1e-05      1          1e-05     0.17297     0.69883
    . Pure error  0.00023125      4     5.7813e-05
```

The stepwise regression modelling method was used to create our model from the actual experimental data see Equation 35. In this instance, the non-significant terms were removed and allowed to refit our model. It was found

that the R^2 and R^2 adjusted values were not ideal. However, there was an improvement in the difference between the R^2 and R^2 adjusted when fitted to the full model from this experimental data. The ANOVA in Table 15 shows that the model has been fitted as best as possible but was still very weak; the lack-of-fit test further shows this as significant with a confidence level of just above 30%. Ideally, one would be looking for no lack of fit in a said model.

The regions highlighted could be used to set new control set points on the fermentation controller for further experiments. However, these may not be statistically reliable, as described by our results; there had been a noticeable lack of fit with this model to the experimental response data. The fact that the stepwise procedure selected a model for "Trace Elements" containing a term that was not significant in the full model indicates that all output generated by statistical software should be carefully examined and we did this via the diagnostic plots and by testing our assumptions. In this case, the stepwise procedure identified the model with the lowest p value, but did not consider contributions by individual terms.

4.6.5 Response Surface Contours for Both Responses

Preceeding to overling the contour plots, the two responses lead to visually comparing the surface responses over the region of interest.
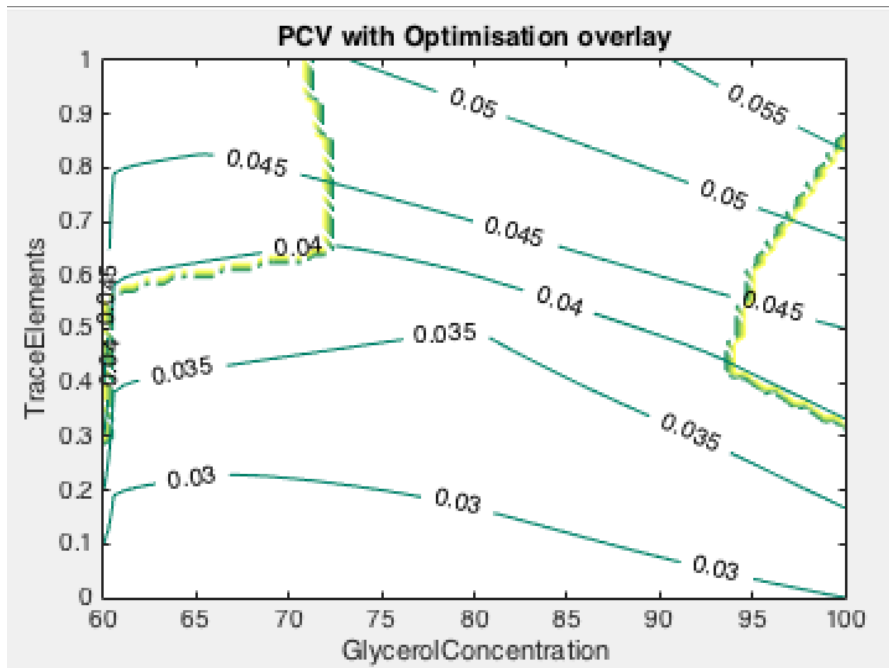

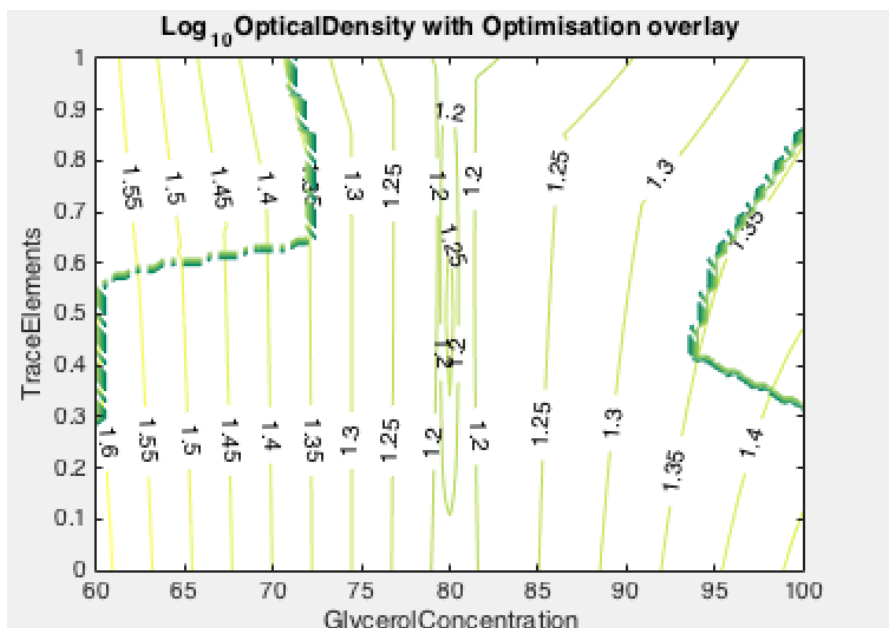**Figure 32 Final optimisation parameter chart output from DoE experiment (PCV)**

Figure 33 Optimisation overlay plots
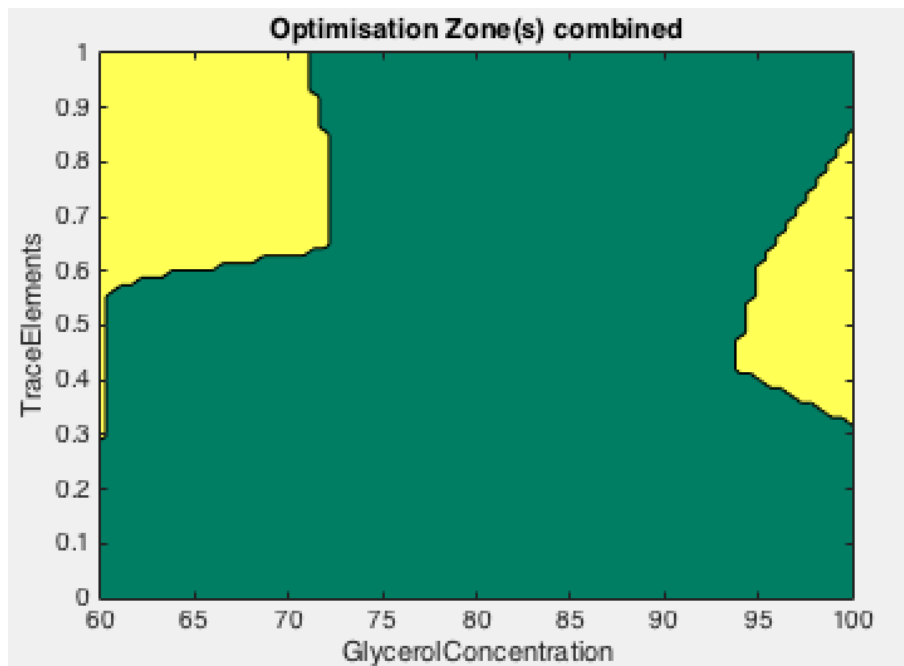
Figure 34 Optimisation combined plot and recommendations. Yellow regions show the conditions in which both output from OD and PCV is optimised. Green areas are the areas where optimisation is not achieved based on the response output of our experimental data.

The models and corresponding contour plots show that a little trade-off can be made when trying to achieve above mean average output response for both OD and packed cell volumes. The areas coloured in yellow show the conditions of operation put in place to achieve the desired optimised responses.

4.6.6 Closing statement for this research section

Specifically evaluating the model, these models only show a black box type of modelling and does not describe the interactions mechanistically. It would be advisable to seek more detail in what the cell was doing within these optimised regions and at the condition boundaries. If stoichiometric modelling of a population of cells could be achieved then the motivation behind further experimentation and advancing the modelling in chapter 5 becomes important.

One example question is what nutrients are bottlenecking the overall growth and optimisation in these regions? Can it be better described? This question arises because of two regions of response optimisation in the Figure 34 and at the boundaries between the yellow and green regions. A higher level of glycerol concentration is acceptable for cell growth, but it is expected to perform better than average OD and packed cell volume responses. This is not surprising, as we expect that OD and PCV to increase with adding more carbon source for growth under suitable growth conditions.

However, the output response space for lower carbon glycerol concentration had a larger output space. Some may consider this to be a more robust operating range, as lower concentrations of carbon source with higher ratios of trace elements over a short period of fermentation are likely to have the same optimised effect. This was not as expected and eludes to other metabolic interactions that need describing. Here metabolic flux balance analysis would be ideal as we can predict the flow of nutrients and possibly answer these questions.

Verification runs at the chosen conditions should be carried out to confirm the output goals of the experiment, as there was some concern over the quality of the model for the packed cell volume response. More verification runs will confirm the robustness of the proposed process controls, optimisation, and model fits. However, these factors have room for improvement in terms of elucidating the mechanics and exact metabolism of a carbon source and trace elements. This would be explored in the next section on flux balance analysis. Since it is now known that glycerol and trace elements, to a degree of confidence, were the most significant factors from our screening experiments in chapter 4. This then lead to optimising this specific *E.coli* w3110 fermentation process. We can explore this further in later modelling applications.

5 Advanced Modelling

5.1 Introduction

5.1.1 Purpose

This chapter builds upon the information created from the DoE model initialisation; trace elements and carbon source were deemed statistically relevant parameters upon the output response. Here, we will explore and create growth trajectory predictions and media optimisation designed for maximising microbial biomass. Also, we can then model the dissolved oxygen concentration and substrate utilisation using different techniques and other principles learned from dynamic flux balance analysis, logistic models, and stoichiometric mass balancing. The aim is to help elucidate the metabolism of *E. coli* and determine the effect on the growth. The final goal is to develop soft sensors for better batch control and faster optimisation applied to the parallel bioreactor system.

5.1.2 Principles of Metabolic Modelling

Fermentation modelling is a complex task and, at the same time, imprecise. The process of modelling a microorganism is always dynamic and non-linear. Additional variability between runs exists and identifying the trajectory of microbial growth can be difficult.

The variability between fermentation runs can be affected by many factors including the bioreactor environment and metabolic activity. Microorganisms will break down nutrients that can be utilised for survival or duplication at different phases, meaning the metabolic activities within the cells affects the population dynamically. When performing a fermentation in batch mode, cell growth can be divided in several distinct phases. A well-established

exponential model of population growth is common in modelling. Whereby a population of microorganisms would continue to increase exponentially if the environmental and biological needs of a species of microorganism are met during growth.

Controlling the addition of inorganic compounds, such as water, acids, bases, salts, and organic compounds and carbon-based substrates, becomes more important for modelling and balancing at a stoichiometric level. Glucose can be supplied as a sugar-based substrate or other carbohydrates can be used typically in a batch or fed-batch fermentation and these substrates are key components for several biochemical reactions.

The cell duplication process usually involves many chemical reactions and utilises both the inorganic and organic substrates. Rate determining reactions will influence the growth rate or product formation rate activity. Both genetics and nutrition availability affect the doubling time and metabolic activity of a microorganism population (Stanbury & Whitaker, 1995).

Aeration is one of the more important factors. Controlling aeration can help boost the overall process productivity and help in process optimisation. One mathematical model includes the variation in cell number over time x(t). This differential equation has an inherent constant, where the specific growth rate gives the following form:

$$\frac{dx}{dt} = \mu.x$$

Equation 37

Specific growth rate is a specific metric for individual species of microorganisms and the biochemical environment in which an organism is growing. The specific growth rate is essentially a function of numerous environmental and biochemical factors. Such features include but not limited to moisture, pressure, pH, substrate concentrations (oxygen, nitrogen, and

carbon sources), temperature, physiology, and energy potential (Doran, 1995). These factors are important to identify and control as they are involved in some assumptions for the modelling.

The phases of the growth cycle are what we are trying to describe to ultimately predict the growth trajectory. Therefore, we should review the underlying principles and areas of interest and shortfalls for the modelling (Figure 35 and Figure 36).
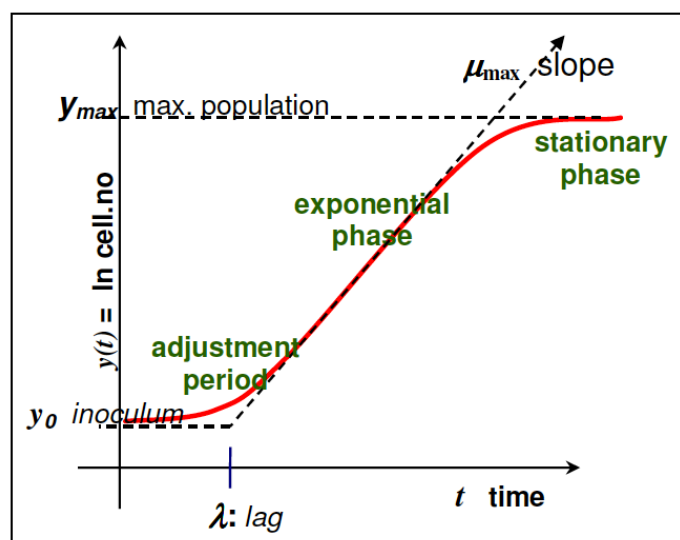

Figure 35 A diagram of an ideal microorganism growth curve with different phases.
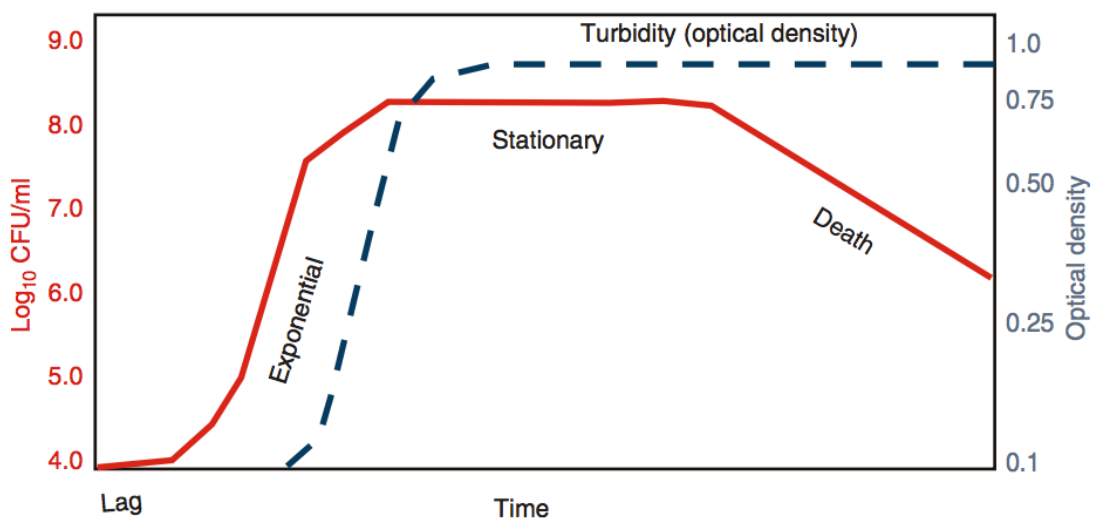

Figure 36 A typical ideal growth cycle with distinct phases and turbidity measurements.

Gene evolution of the cell will normally dictate the length of the lag phase. The regulation of specific "resource-accumulating" enzymes occurs at this stage in the growth cycle at the cost of biomass production. Naturally, the enzymes that are involved in biomass production are not fully activated. The ribosomal genes and rapid amino acid biosynthesis are blocked/locked, preventing cell growth until sufficient quantities/concentrations of primary resources have built up inside the cell (Schultz, 2013). Before optimal biomass production can occur, this enzymatic bottleneck would need to be reduced/shortened through better medium controls as well as by regulating the initial inoculum size. This is the first step in understanding where the modelling could go wrong, where we try to predict a trajectory that would be ultimately bottlenecked by the enzyme regulations. The model must be "feasible".

The strategy to be employed here is that of comparing the factors that can significantly reduce the lag time. Small changes can increase the time in which optimal biomass is reached during exponential growth as well as the consumption of these resources in the process. Lag phase cells usually show a small increase in size but no exponential growth. These cells in their new environment are undergoing adaptation.

Exponential phase (also known as the log phase)

After the accumulation of metabolic intermediates necessary for fast growth, cells start dividing at their maximum specific growth rate. This can be calculated empirically. This maximum specific growth rate is usually of great interest because growth rates can be used to help distinguish the "cause and effect" relationship between varying environmental conditions for rapid growth, especially in time-sensitive productions or optimisation.

The exponential phase also is where the biggest increase of cell numbers occurs. This increase in cell number, and sometimes linked to increases in

biomass, continues until the environment becomes inhibitory (usually by metabolic products accumulating) or a rate-determining substrate is consumed. Many metabolic models include pathways and there are tools that minimise certain pathways from producing products that would cause growth inhibition. Metabolic engineering can be used as another source of species optimisation for growth and biomass.

Equally important is the characterisation, design, and choice of engineering factors affecting oxygen mass transfer during cultivation. If the biological oxygen demand cannot be met for the maximum specific growth rate, then inherently, the growth rate will decrease or become limited to the amount of dissolved oxygen in the culture medium and the number of rate limiting steps of mass transfer into the cells.

As like the schematic in Figure 35 and Figure 36, plots of population growth often show a linear relationship with time from which the specific growth rate and doubling time can be calculated, but the data of the population size must be transformed into the logarithmic scale.

During the stationary phase, net growth is balanced with the number of cells that are dying and lysing. Endogenous metabolism occurs when the external nutrients are consumed, and for maintenance metabolism for as long as possible in the attempt to survive. This type of metabolism occurs throughout the population growth cycle at various levels (Maier, 2000). As the specific growth rate as well as the actual population size starts to slow down, it can be seen as a plateau in the OD measurements. The curve becomes asymptotic.

As the culture enters the death phase, this is specifically where the rate of multiplication of the cells falls below the rate of death of cells. Progression of cell population is towards death is characterised by the decline in the number of viable cells.

5.1.3 Sigmoidal Models of Growth

One of the aims of data fitting and modelling is to minimise the gap between estimated values and actual data. Membership functions are a useful way of doing this efficiently. Mapping the input values where it is not clear as to what they are going to be categorised as is called fuzzy logic. The biomass plateau is set on several "if" statements: if the concentration of carbon source or substrate is limiting, then the degree of membership will show a value between 0 and 1. Since the concentration of the limiting substrate occurs over a continuous period, the membership function is also likely to follow a continuous membership function.

$$f(x, a, c) = \frac{1}{1 + e^{-a(x-c)}}$$

Equation 38

Equation 38 above shows the sigmoidal membership function. This formula was used in the development of the growth model to create a sigmoidal curve. This function has mathematical membership with the Monod equation. This was based on the growth rates of the microbial cell system and the relevant stoichiometry.

Cell growth was measured with variables such as dry cell weight, colony forming units (CFU), OD (although it is useful to represent the data in a biologically relevant manner other than OD), and protein concentration over time.

Sigmoidal patterns under the right conditions relate to the growth phases of microbial cells in the lag, log, and asymptotic, stationary phases. Sigmoidal growth patterns and models are only an approximation at best to the experimental data that we could fit.

We will consider using the log plot of a population size relative to the starting population size. Sigmoidal plots such as the Gompertz and logistic plots see Table 16 which has specific parameters for calculating the lag, log, and asymptote. The models in Table 16 are commonly based on empirical models being fitted to data. There are some mechanistic models available that describe microbial growth also (Yilmaz 2011; Zwietering *et al.* 1990; Fujikawa *et al*, 2004).

The equations in Table 16 have been further developed and modified in a way to accept new terms that include dynamic temperature changes during growth, lag times during the initial cell adjustment to new media at the time of inoculation and various chemical environments with respect to pH and water activity (Zwietering *et al*, 1990). The problem with using unmodified model data is that for example lag time has to be determined from empirical data, instead of the model intrinsically estimating the lag times of new bacterial cultures.

Table 16 adapted from Zwietering *et al* (1990) detailing a number of empirical modelling algorithms for prediction of bacterial cell growth.

Model	Equation
Logistic	$y = \dfrac{a}{[1 + \exp(b - cx)]}$
Gompertz	$y = a \cdot \exp[-\exp(b - cx)]$
Richards	$y = a\{1 + v \cdot \exp[k(\tau - x)]\}^{(-1/v)}$
Stannard	$y = a\left\{1 + \exp\left[-\dfrac{(l + kx)}{p}\right]\right\}^{(-p)}$

All the functions have three or four parameters that need to be calculated to apply the model. What we can do statistically is evaluate the necessity or benefit of having extra (four instead of three) parameters describing the data in the model.

The first parameter of the growth curve models (a) in Table 16 starts at the intersection on the x-axis, specifically the intersection point at which the cell growth is at its steepest. This is best known for its relationship with the maximum specific growth rate.

The second parameter described is the asymptote where the last data value is taken as the asymptote infinity (b). This excludes the fact that the death phase could occur and is not ideal. Unless you have large amounts of data points to choose from for the asymptote parameter, this is usually the last point taken in the plateau.

The third (c) and the lastly the extra parameters (d…) can be calculated by using the smallest difference between experiment and model data. This is calculated by the residual sum of squares analysis between data and model functions. This is best done programmatically to minimise the risk of error.

Kinetic modelling

Compounds are converted within the biochemical reactions. These reactions follow reversible, irreversible, and catalytic reactions.

$$A \rightleftarrows B$$
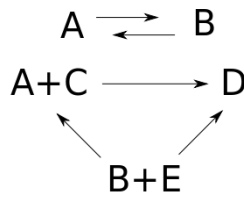$$A+C \longrightarrow D$$
$$B+E$$

Figure 37 Network diagram of reaction pathways

The reactions can be modelled and when rates are predicted, these can be used to optimise production of a specific metabolite or output.

The steps involved in this study were to use and identify biochemical pathways specific for creating biomass and carbon source consumption. Then, assumptions based on kinetic expressions were made and mass balances for the specific species were carried out using a library of pathway information from the EcoCyc and KEGG repository. Next, the parameters needed to mimic and generally fit the empirical data were calculated.

Reactions networks can be broken down into two types:

- Closed networks – these networks reach equilibrium when the reaction rates are zero. This is also the point at which the reaction rates are at a steady state.
- Open networks – This is where the input and output components are transferred continuously. When the metabolite conversion rates within the network system are continuous, the production of metabolites to be transferred out of the system network is dynamically equilibrated with the input substrates. Essentially, the reaction rates are not zero. These networks are the most realistic to the culture. This includes nutrient supply to the network and the secretion of metabolites and waste into the surrounding broth (off gases are usually dissolved components).

Reaction networks are governed by the direction of the reactions in the network. This is where reversible, irreversible, and catalytic reactions are

important for defining rates and applying assumptions/constraints when creating predictive models. This allows to dynamically calculate the fluxes around the network. Identification of reaction rates is the next important component of reaction network and model construction. Reaction rates will determine how one component will be converted into the next component or waste product. Reaction rates are calculated over a period and ordinary differential equations (ODE) are used to calculate the differences between reaction rates. Reaction rates will likely include nonlinear terms and there are solvers that can be used to calculate the expressions. Approximation is a factor to be aware of when creating solutions to the network flux problems in the model. The input values for the reaction expressions and the time scale for the reaction rates are needed for non-linear terms.

ODE equations require all reaction rates to be identified. The production rates, which are deemed additive to the reaction profile, and consumption rates, which are deemed subtractive to the reaction profile, are used to balance each metabolite in a network.

Mathematical models require all possible reactions and rates for production and consumption. Current research is exploring models of larger, whole microbial species networks, but some biochemical or chemical networks are currently unavailable publically owing to the time and investment needed for research and inputting and validating libraries of data. However, smaller network sections or compartments can be explored individually (Schellenberger, *et al.,* 2011).

5.1.4 Mass Action Kinetics

Guldberg and Waage proposed a kinetic scheme for chemical reaction networks in 1864. Their work describes chemical reaction rates to be proportional to original reactant chemical concentrations. It is a widely used generalised assumption. For chemical networks, there are products and

reactants that have kinetic orders. There are exponents that represent these kinetic orders. Kinetic orders represent the question "How does this specific compound effect these reactions?" We can start looking at the effect on reaction rates, other compound modifiers to the reaction, specific compound concentrations, reaction direction, and reactant conversion. These can be described by the power rate law.

$$\frac{dx_i}{dt} = \alpha \prod_{i=1}^{nSubstrates} \left(\frac{S_i}{S_{i_0}}\right)^{g_i} - \beta \prod_{j=1}^{nProducts} \left(\frac{P_j}{P_{j_0}}\right)^{h_i}$$

Equation 39

Alpha represents the forward reaction direction rates interacting with compound S. Similarly, beta represents the reverse reactions; 'h' and 'g' in the power rate law, represent the kinetic effects which are usually specific to the reaction and product stoichiometry. They may have positive or negative integers, representing positive or negative effects, respectively, on the reaction. No effect is represented by zero.

Under the most popular set of mass action assumptions, directed reaction graphs can be used to describe the chemical reaction networks. These graphs are also known as Kirchoff matrices. Usually, the rate of a reaction is determined uniquely by the stoichiometry of the reactant complex:

$$N = (S, C, R)$$

Equation 40

In the Kirchoff matrices (Equation 40), the reactant stoichiometry underpins the rate of a reaction. These are represented in complexes, C, and the reaction sets, R, in the reaction network as a table format. The Kirchoff matrix systematises the reaction connections and kinetic orders (the effects that

compounds have on the reaction). The Kirchoff reaction matrix has connections with the dynamics of a given mass action system. In biological systems, these dynamics correspond to the system behaviour at a local state. Steady state systems can be represented as linear problems that researchers can solve by using linear programming and ordinary differential equations. These are described later. Mass action laws are the generalisation of power law.

5.1.5 Biochemical Networks

Enzymes are the primary basis for all biochemical reactions. They are used in biochemical networks to speed up otherwise lengthy reactions. Factors are limited to the properties of each enzyme, the mechanism of action, and reaction rate (Schmid et al., 2001; Nelson & Cox, 2008). The two classifications of enzymes are for catabolic reactions and anabolic reactions. Anabolic reactions create larger compounds for biomass synthesis and the catabolic reactions metabolise and break down the nutrients into smaller products and generate energy for cell growth and metabolism. Both are needed to construct the entire metabolism of a cell.

Reaction kinetics can show which substrates are efficient or not, and this can be effective in creating an optimised nutrient feed. It is known that the enzymes are not consumed while calculating reaction kinetics.

$$[Substrate] + [Enzyme] \leftrightarrow [ES\ complex] \leftrightarrow [Eenzyme] + [Product]$$
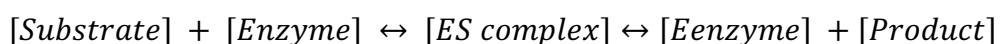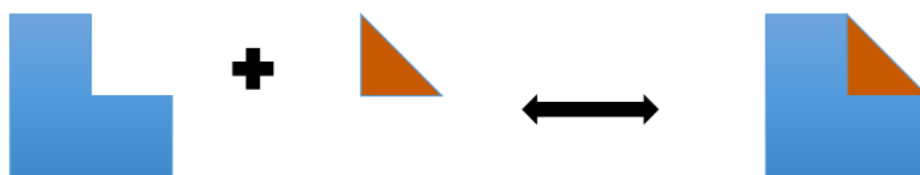
Equation 41

5.1.6 Flux Balance Model

Flux balance analysis (FBA) uses mathematical equations, under a set of limiting constraints to understand metabolite flow for a large-scale model (e.g. genome scale). Its approach is one that focuses on optimisation of an objective. In some cases, this is the maximum cell growth. Once this objective is identified and a large biochemical network is configured to explore the assumed solution space, it would be possible to predict the rate of cell growth. Another way in which this can be applied is with metabolic engineering. In this case, the outcome could lead to the production of a metabolite or e.g. again changes to the growth rate of the organism.

After the representation of the metabolic network, reconstruction is an important step in metabolic calculation and analysis. The widely available tool called the constraint-based reconstruction and analysis (COBRA) in the Matlab toolbox (Orth *et al.*, 2010) enforces these metabolic representations into ordered stoichiometric matrices (S). Metabolites would flow freely (programmatically speaking) through this network, but it is the stoichiometric reactions that enforce the flow constraints.

The stoichiometric calculations are then applied to the objective function (Orth *et al.,* 2010). There needs to be a check on the balance of input and outputs to make the network valid. The stoichiometric constraints fulfil this role by ensuring flux is being balanced. Lower and upper bounds are applied to each reaction to indicate what reaction fluxes are allowed or appropriate. These reaction bounds would essentially balance the input and outputs but the bounds can be set to represent inequalities (Orth *et al.*, 2012). The reaction matrix represents the mass. This means under assumed steady state conditions, the consumption and production levels of a metabolite are equal. For the matrices (S) to be confined, these steady state assumptions are key in FBA, allowing the metabolic fluxes (v) to be constrained and identified as

$$S * v = 0$$
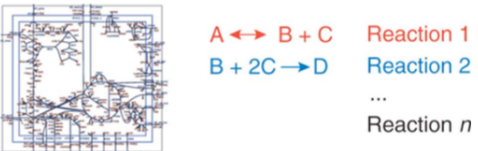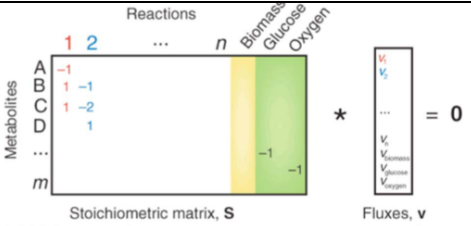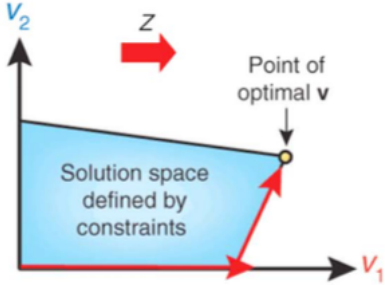
S is the matrix containing a stoichiometry of the internal catabolic reactions that will take place, while v is the reactions rates constructed as a vector. The net metabolite uptake in this steady state case would be 0.

See step 3 of the FBA formulation path below.

Table 17 Steps involved in metabolic pathway analysis and FBA

Step 1	Metabolic reactions are first defined in a genome scale metabolic reconstruction. These are essentially a set of stoichiometric reactions that are specific to the organism of interest. The letters below represent the different reaction substrates and products . e.g.A → B+C; B+D → F; C+E+G → H	
Step 2	Metabolic reactions are then represented into a mathematical matrix. The matrix is set into rows and columns ("I" x "j"). The different metabolites are formulated in rows (j) and the reactions in each column (i).	

Step 3	We use mass balance and stoichiometric boundaries to define the limits of the matrix and reaction schemes. The lower and upper bounds are applied and allow the flux of metabolites through each reaction to be bound by the S * v = 0 equation.	$-v_1 + \quad\quad \ldots = 0$ $v_1 - \quad v_2 + \ldots = 0$ $v_1 - 2v_2 + \ldots = 0$ $\quad\quad v_2 + \ldots = 0$ etc.
Step 4	Then, we apply an objective function (z) that highlights the highest reaction contribution "c", also known as the weight; "v" is the reaction rate in the matrix in step 2: "$Z = c^T v$"	To predict growth, $Z = v_{biomass}$
Step 5	We optimise for or reduce the contribution effect to the final objective function by using linear programming techniques. The constraints can be observed at this stage for better analysis of the mass balance equations and reaction bounds.	
Additional notes	Models can be constructed for prediction but must have the a priori knowledge of a species genome to work out step 1. Programming for automation of the controls for the fermentation conditions will continually be checked and limited by the mass balance assumptions made in step three. Further characterisation of metabolic states can be obtained by comparing the theoretical and experimental data to improve models.	

External metabolic rates (growth rate, substrate uptake, and product accumulation rates) related to the specific organism are additionally constrained (i.e. complimented with the stoichiometric matrices of reactions) to successfully determine the overall (pseudo) steady-state system. To make the estimations relevant to the specific organism, measured metabolic rates (r) are applied to the metabolic fluxes to give:

$$R * v = i$$

Equation 43

Again this time the R is the matrix containing a stoichiometry of the internal catabolic reactions that will take place, while v is the reactions rates constructed as a vector. The net metabolite uptake would be 'i'.

Least squares regression methods are used next, allowing the equations to be solved simultaneously.

$$min\ SSR\ = \sum (r - r_m)^2 / \sigma_r^2$$

Equation 44

The result is the estimation of the metabolic flux for a determined system. The determined system is the system that contains all the necessary external rate experiments (Antoniewicz, 2015).

The objective function is usually defined as the production of biomass, as was the case in this study. It can be used to predict microbial growth and the biomass components (lipids, nucleic acids and proteins). The constraints can be set programmatically to suit our modelling. This makes the whole FBA feasible, matching the real-life fermentation counterpart.

Mass and energy balancing model development is useful for general description of cellular metabolism. Flux balance analysis (FBA) takes this further into modelling the metabolism of the cell (Feist and Palsson, 2008

from Meadows *et al*, 2010)). FBA can be adapted to include several applicable constraints to overcome metabolic engineering problems. Examples of constraints include levels of metabolic transcription, thermodynamics, intracellular crowding problems. Other challenges come in with estimating multiple carbon source consumption and other in situ engineering problems associated with industrial scale fermentations (Meadows *et al*, 2010). The aim in using this FBA is to help elucidate experimental data and understand bioprocess behaviour.

5.1.7 Dynamic Flux Balance Analysis

Dynamic flux balance analysis is an extension of FBA. The aim is to understand metabolite flow and biomass or product formation without masses of kinetic information. The method of DFBA is quite similar to FBA but can be applied to batch or fed-batch applications and does not merely describe dominant measurable metabolites (Hjersted & Henson, 2009; Nikdel & Budman, 2016). Since typical bioprocesses have transient behaviour, dynamic flux balance analysis aims to fix this problem of describing the global system for large- and small-scale operations. This problem is usually set up as an optimisation problem where the goal is to minimise the differences between experimental and *in-silico* data. The concept is to solve the FBA at a set number of times or sampling intervals (Nikdel & Budman, 2016). This problem is solved again, usually by the least squared method and solves the objective function within the boundaries and constraints of the reactions. It is important to note the challenges in finding global solutions to these proposed model problems.

The advantages of DFBA are that fewer input arguments are needed, i.e. just the limiting metabolites, as opposed to the kinetic data needed for determining system flux balances. Then, the rest of the flux problems are solved as constrained stoichiometric relationships.

Unstructured models typically use the growth rate capacity as a function of a limiting substrate, where it is most applicable to batch operations without many changing conditions (Hoffner, 2013). DFBA has a broader approach to the analysis of metabolism and environment interactions. When compared to models that are unstructured, it is difficult to apply predictive approaches.

DFBA can also be used for animal, bacterial, and plant species (Hoffner, 2013). Dynamic metabolic engineering and batch or fed-batch control optimisation at small- and large-scale are the other applications of DFBA (Hoffner 2013).

Metabolic flux balance analysis assumes that the metabolism is at steady state. This becomes more problematic and tools like COBRA tool box have the ability to determine metabolites that are not at metabolic steady state (Antoniewicz, 2015). A time series order of events is calculated for extracellular concentration and rate:

$$\min \mathrm{SSR} = \sum (c(t) - c_m(t))^2/\sigma_c^2 + \sum (r(t) - r_m(t))^2/\sigma_r^2$$

$$\text{s.t. } R \times v(t) = \mathrm{d}c/\mathrm{d}t$$

$$S \times v(t) = 0$$

(Antoniewicz, 2015)

Another assumption for the DFBA is that the transients are slow in the order of seconds to minutes. This allows the fluxes to be processed at discrete time intervals. Concentration measurements of metabolites are taken from average external rates for each discrete time event. The final step in describing the

culture component concentrations is to take each discrete FBA event and create a timed profile of the transient fluxes (Antoniewicz, 2015).

Data smoothing is an alternative method when deciding which time intervals are to be taken or are required. Assuming that when measurements are taken, the flux of transients is slow, the accuracy and validity may end up being irresolute without the application of data smoothing (Antoniewicz, 2015). Data smoothing can be applied in the form of splines, linear, or polynomial fitting, and smoothed data can be derived. Recent developments have tried to overcome this problem of having to manually select the time intervals or pre-process the data (Leighty & Antoniewicz, 2011). But, this is still a major challenge because of the inherently stoichiometric balancing needed for the model, it has the same limitations as classical FBA. These limitations include parallel, cyclic, and reversible pathways and reactions (Antoniewicz, 2015).

Flux balance analysis is used to understand, map and reconstruct biochemical networks. It is possible to predict several metabolic reactions from a cell and its reconstructed genome (Orth *et al.,* 2010). DFBA is expected to provide better predictive capabilities than a structured cell model. The objective is to obtain values for cell growth and product yields using this systematic approach. There are many reconstructions available on the web today that have been increasing rapidly. This advance is due to improved bioinformatics tools and experimental procedures. The applications related to DFBA can be used in bioreactor design and analysis. It can also help in the simulation of many fermentation and diagnostic problems. The only way to test the validity of the simulations is to perform the fermentation under the same conditions that the simulation is modelling or conditions that are of interest.

Where this fits into the scheme of DFBA and the Multifors is that the parallel bioreactors can be used to evaluate multiple conditions at the same time,

speeding up the model validation process and meeting the need for alternative (albeit small scale) culture technologies for model development and validation. The reconstructed model was based on iY75_1357 version 20 (Available at: http://bigg.ucsd.edu/models/iY75_1357) (King *et al.*, 2015). At the time, the iY75_1357 model metrics contained 1358 genes, 2760 intracellular reactions, and 1953 metabolites that were balanced (King *et al.,* 2015).

5.2 Methods

5.2.1 Analytics

Optical density was measured using a spectrophotometer (Thermo-Fisher scientific, USA) at 600 nm for fermentation samples taken. The OD measurements were performed offline. Samples where pipetted into 1ml cuvettes, and dilutions were used where appropriate to keep the readings below 1.000 on the spectrophotometer. Microsoft Excel was used to determine the maximum specific growth rate using the built-in solver application with the OD experimental data at 600 nm.

5.2.2 Validation of Growth Estimator

The mean square error method (MSE) was used to determine the difference between the simulation model output and the actual experimental value. The MSE was calculated as the sum of the mean of the squared residuals; experimental values were used to evaluate the proposed logistical regression and exponential models during the lag and stationary phases. The smaller the value the closer the model fit to the data.

$$Mean\ squared\ error\ =\ (1/n)\ x\ (sum\ of\ squared\ residuals)$$

Equation 45

5.2.3 Model Data

Predictions of biomass, glycerol, chloride, phosphorus, and sodium which form the components of minimal media M9 used in fermentation were made from the dynamic flux balance analysis.

All simulations were carried out using Matlab 2015b and the COBRA tool box for dynamic flux balance analysis using the GLPK linear programming solver.

SBML and libSBML tool boxes for Matlab are also required elements of the software package. Installation and compiling were carried out on a Windows PC and MacBook pro laptop. The Cobra tool box is a freely available research toolbox available at Opencobra; all Git versions are available at that location.

The DOT% experiments were set up with the medium containing glycerol as the primary carbon source, and contained a defined nutrient medium with salts, phosphorus, sulphur, potassium and nitrogen. The process was operated as a batch culture, dO% was controlled at 40% air saturation. pH was controlled by the automated pump with ammonium hydroxide additions. Broth samples were taken and optical density was measured at 600nm

5.3 Results

5.3.1 GUI Application

The GUI application was developed with Matlab. The building of the GUI included building a mass balance calculator in matlab. Then using the optimisation tool as a built-in optimisation toolbox. This is essentially a solver for stoichiometric mass balance equation and currently requires starting data input from user.

Matlab script used to compute stoichiometric coefficients with the 'Fzero' ordinary differential equation method. In the future scope, there should be further automated integration with other data collecting programs through the use of application programming interface API's.
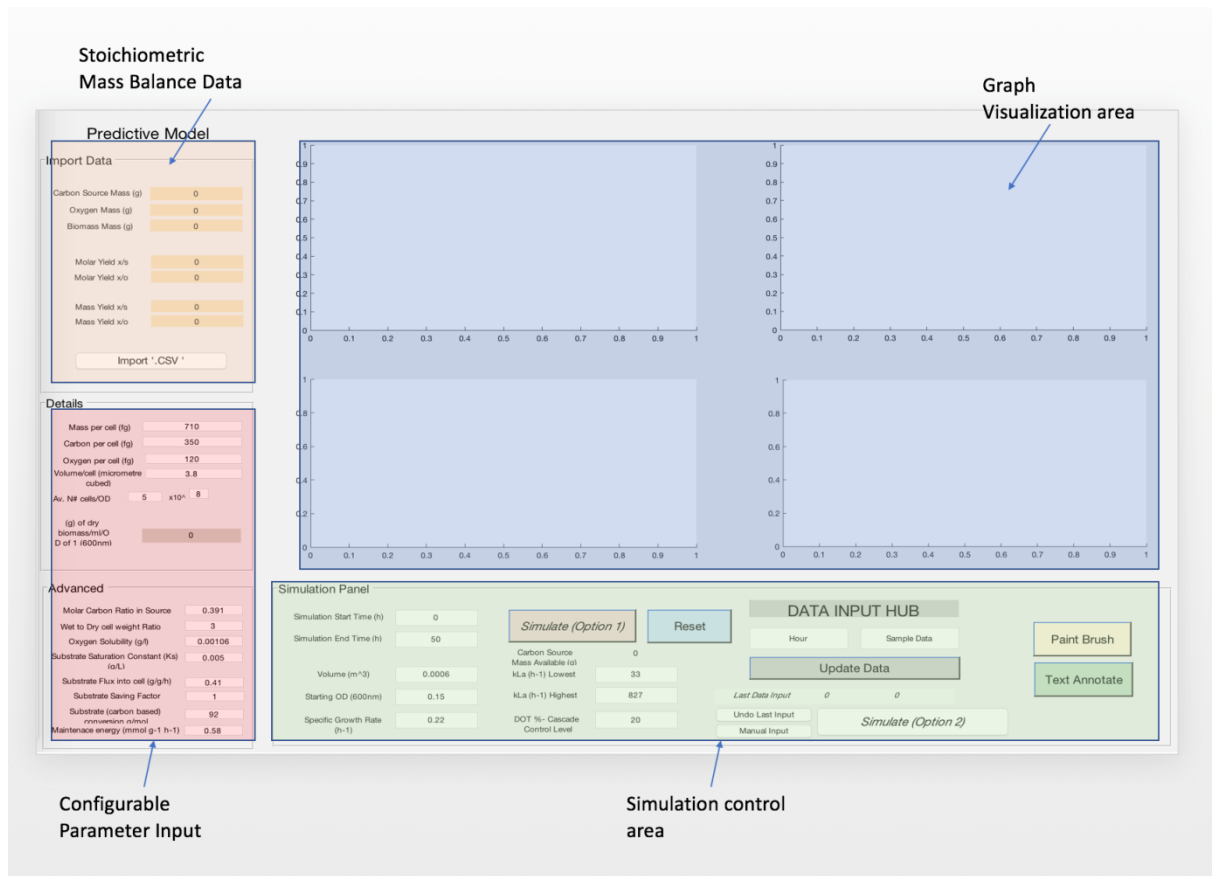
Figure 38 User interface built for growth modelling simulations, data input and DOT% estimation.

The second feature developed was a graphical output/user information input GUI see Figure 38. The information outputs were designed to give information on DOT% profile, which has the ability to show data in the graph area, and microbial growth, using either sigmoidal or exponential growth kinetics.

5.3.2 DOT% Profile Estimator

The DOT% estimator is a dynamic representation of the consumption of oxygen substrate based upon the dry biomass of *E. coli* and the growth kinetics, stirrer speed, and gas velocity parameters.

$$kLa = k.(Pg/V)^{\alpha} \cdot (v_s)^{\beta}$$

$$OTR = kLa \, (C* - C)$$

$$C = \psi.DOT\%/100$$

$$OUR = (\mu/Yx/O_2).X + mX$$

$$dDOT\%/dT = (OTR-OUR)*(100/(kLa* \, \psi))$$

Ψ = fraction of oxygen in air – specific solubility. K_La = volumentric mass transfer coefficient. Alpha and beta are exponents. V_s = gas velocity. OTR = oxygen transfer rate. C = fraction of dissolved oxygen as a percentage in liquid. DOT%= dissolved oxygen tension percentage. X = biomass. mX is biomass maintenance energy requirements. Y(x/O_2) = yield of biomass on oxygen. OUR = oxygen uptake rate. T = time.

5.3.3 Growth Kinetics and curve fitting

We adapted a simple logistical model that takes into consideration an additional exponential curve fitting parameter (Equation 51):

132

$$OD = OD_t = OD_{t=0} + \cfrac{1}{\cfrac{1}{OD_{tf}} - OD_{t=0} + C_{tf}(e^{-D \ x \ tn})}$$

OD at 600 nm, t = time, tf = at final time, n = sample number. C and D were curve fitting parameters that, by using either Matlab or Excel, can solve the model to fit to the experimental data gathered for a simulated future growth trajectory. This was done using the minimising mean squared error (MSE) method by taking maximum and minimum optical density data points, and finding a minimum point where all sample points were as close to the logistic curve as possible using MSE.

This model in Equation 52 is the exponential growth model under non-limiting substrate conditions and non-inhibitive oxidative microbial growth. N is the number of bacteria; μ is the specific growth rate; t is the time in hours (Stanbury & Whitaker, 1995).

The exponential growth equation is the differential equation

$$\frac{dN}{dt} = kN \ \ (k > 0)$$

Its solutions are exponential functions of the form:

$$N = N_0 * e^{kt}$$

where $N_0 = N(0)$ is the initial value of N.

Also can be compared using the base model (Gompertz Equation 54) to the simplified logistical regression model (our work).

Gompertz:

$$N = a * e^{-e^{(b-cx)}}$$

Equation 54

We have separately looked at the initial lag phase and stationary phase to help determine the applicability of a simplified logistical empirical model for speeding up developmental times for a new bioprocess.

Mini-bioreactors that have the function to run in parallel and operate simultaneously may to some degree increase the rate of experimental development. There is potential to reduce times and costs. Characterising the phases is essential in bacterial population growth in a stirred tank reactor to know when, for example, induction and genetically engineered protein expression is optimal after reaching a maximum biomass or specific growth rate during log phase.

The aims of this experiment were:

1) Compare how well the established models (exponential and Gompertz) are able to describe exponential growth trajectories against experimental data.

2) Secondly develop logistic curve models that are better at describing the key growth phases for *Escherichia coli*.

The experiment was set up using defined media and the Infors Multifors 2 built in controller for experimental control. Also, pH, pO_2, and temperature were the controlled parameters in this experiment.

Table 18 Results of the specific growth rate and MSE of the residual fit to three models, exponential growth model, logistic regression for the lag phase, and sigmoidal Gompertz growth model

Parallel mini bioreactor	Maximum Optical Density on Minimal Media (OD 600 nm)	Specific Growth Rate (h-1)	Mean squared error of exponential curve fit (OD 600 nm)	Mean squared error of logistic curve fit (OD 600 nm)	Mean squared error of Gompertz model (OD 600 nm)
A	3.54	0.265	0.064	0.043	0.065
B	3.64	0.300	0.041	0.018	0.020
C	3.60	0.311	0.043	0.016	0.030

According to the maximum optical density of this experiment there was acceptably good reproducibility between the 3 reactors operating in parallel, the average of the three runs was OD = 3.59 +/- 0.047 with standard deviation. This was likely due to the media being defined and good automation that the Infors controller provides.
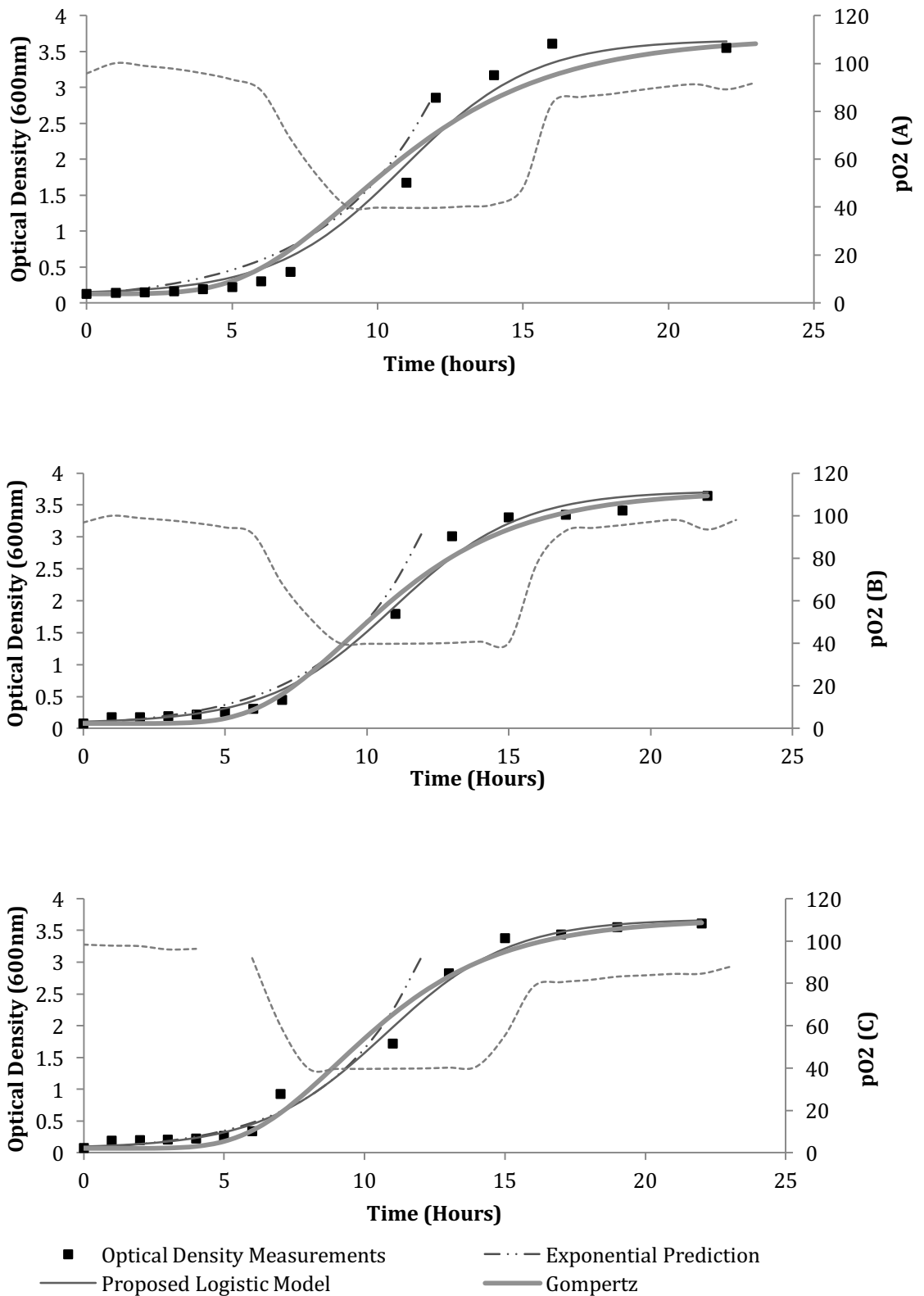
Figure 39 (A-C) *Escherichia coli* **fermentation at 500 mL volume, 2% inoculum. This graph includes prediction models for exponential growth and logistic expression as part of a parallel series of fermentations: reactor A-C. Note pO$_2$ on secondary y-axis is included for information purposes.**

Three separate bioreactor runs were investigated using the similar inoculation start time and initial OD measurements. The pO$_2$ was included to show the increase in oxygen requirements during the log phase of fermentation. Here, it is likely that the minimal medium supplied the nutrients and necessary substrates until hour 15 when primary carbon source, as glycerol in this case a key substrate was not available to the cells. This was likely triggering the stationary phase. Airflow at 1 vvm was supplied to each vessel, but cascade control was initiated when 40% pO$_2$ was reached by increasing the impeller speed from 300 to 1000 rpm. The OD measurements were taken with the aim to validate the proposed modified logistic model. As this was repeated three times, we assumed the n = 3 validation to be adequate under similar operating conditions. Measurements were taken from the lag, log, late exponential and stationary phases of the life cycle of the *E. coli* population. Figure 39 (A-C) show the fermentation data for broth OD at 600 nm; this was used as the growth indicator in these experiments, since the model is designed on the basis of the OD measurement units. Dissolved oxygen values where included as additional data for comparison and reproducibility studies.

The Gompertz model;

$$N = a * e^{-e^{(b-cx)}}$$

Equation 55

Our modified logistic model

$$OD = OD_t = OD_{t=0} + \cfrac{1}{\cfrac{1}{OD_{tf}} - OD_{t=0} + C_{tf}(e^{-D\,x\,tn})}$$

Equation 56

and exponential bacterial growth equation model

$$\frac{dN}{dt} = kN \quad (k > 0)$$

Equation 57

In the form:

$$N = N_0 * e^{kt}$$

Equation 58

were used to validate against experimental data and determine the improvements on which the proposed new model could predict data points. The data points were again used for OD measurements.

From these results in Table 18 other model comparators were used to check how closely fitted the experimental data was across the entire fermentation time. From here, it was determined the model with the best fit by using the mean squared error values (MSE). This showed that the new modified logistic model had the lowest values with best fit. However, noting that these values were just a small advantage on the regular Gompertz model. The proposed model does have potential for use in these specific conditions. But it does require further experimentation for different model creation and curve fittings.

This could further highlight whether this small improvement is akin to just these conditions. It could be a possibility where there is no further improvement than that stated on the Gompertz model. Making the Gompertz equation suitable to model growth also.

Further experiments see Table 19 were taken to establish and validate which model which type of model was most suitable to describe the lag growth phases instead of the entire growth trajectory based again on the smallest mean squared error (MSE). The experiment was set up to further validate any difference between estimated and actual biomass sample readings for a set of fermentations running in parallel in minimal M9 defined media.

Table 19 Results of the specific growth rate and MSE of the residual fit to two models, exponential growth model, and logistic regression for bacteria growth curve fitting.

Parallel mini bioreactor	Final optical density at 23 hours (OD 600 nm)	Specific growth rate (h⁻¹)	Mean squared error of exponential curve fit (OD_600 nm)	Mean squared error of logistic curve fit (OD_600 nm)
A	4.32	0.252	0.0273	0.0072
B	4.46	0.251	0.0014	0.0028
C	4.56	0.263	0.0300	0.0301
D	4.51	0.252	0.0028	0.0035

The results in Table 19 for the mean squared error of exponential curve fit - OD_600nm shows that the exponential model can describe the experimental OD data better in the lag phase. Since its MSE difference was lower than the logistic curve fit. Results from Table 18 show the exponential model failed to describe the changes towards the late stationary phase. For simulation purposes the exponential model was the most inaccurate overall but advantageous in estimating the lag phase.

The proposed modified logistic model has an additional exponential parameter (D) see Equation 56. Due to the MSE method chosen for describing the difference between experimental data and model the extra parameter helps better describe the curvature across multiple growth phases (lag log and stationary phase) instead of favouring fewer growth phases like the other models do. The MSE fit of each of the models in Table 18 to the data favours the modified logistic model (Equation 56).

Overall sigmoidal models tend to better describe later asymptotes. It could be argued that more OD measurements could further refine the model and reduce the risk in the model. However, the approach of repeating simultaneous batch fermentations was adopted to mitigate experimental error/risk across more reactors than just relying on one batch to validate a model. Repeatability is preferred over single-batch data precision. At this stage of experimentation modifying the original logistic model was suitable for simulating a simple growth trajectory of a microbial system.

5.3.4 DOT% Simulation results

The DOT% estimator performance was initially evaluated using Matlab, see materials and methods section 2. This estimation was verified with the experimental results. It is important to note that this estimation incorporates cell metabolism, oxygen solubility, engineering parameters, stirrer speed, superficial gas velocity, and cultivation media (i.e. the glycerol substrate consumption). The low cell density of the OD shows that even for low settings
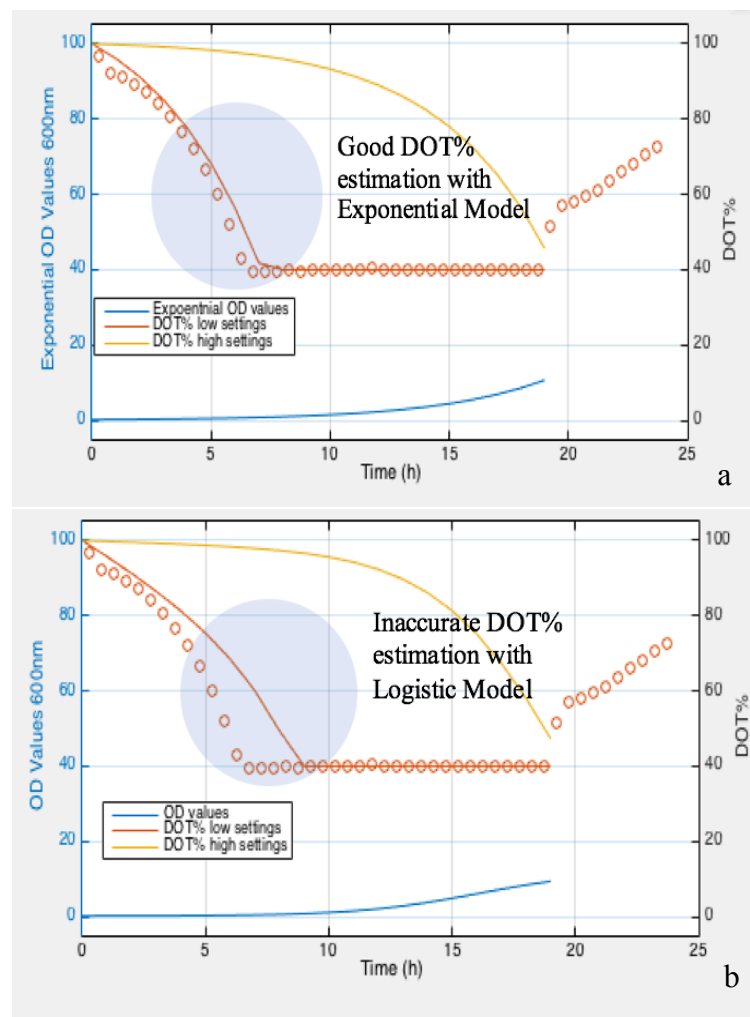
Figure 40 Matlab constructed results with Graph A and Graph B. Graph A (top) includes the DOT% estimator ('----')-calculated on an exponential growth profile. Graph B (bottom) describes the Sigmoidal (logistic) growth profile and its effect of estimating DOT%. The 'o' circles are the experimental data for DOT% based on a glycerol substrate batch concentration of 10g/L. Other details are present below in the simulation panel of the GUI. Low settings are 100 rpm; high settings are 1100 rpm; 0.5 vvm. The 'o' circles are the experimental data for DOT% based on a glycerol substrate batch concentration of 10g/L. Other details are present below in the simulation panel of the GUI. Low settings are 100 rpm; high settings are 1100 rpm; 0.5 vvm

141

i.e. the parameter of the stirrer speed being set to 100 rpm (Figure 40-unbroken red line). It was insufficient to maintain DOT% above 40%, thus cascade control was implemented during fermentation and the high setting of 1100 rpm was sufficient to maintain a critical DOT level before the carbon substrate (glycerol) was assumed to be fully consumed at 19 hours due to a rise in the DOT% data.

Graph A in Figure 40 describes the exponential growth profile and the effect it has on estimating DOT%. Graph B (right) in Figure 40 describes sigmoidal growth profile cellular metabolism and the effect it has on DOT%.

The performance of the DOT% estimator was compared using the sigmoidal growth curve model and exponential growth curve. Each of the models used the first 3 data points (0, 11, 12 h) to work out the specific growth rate as the fermentation was carried out. This enables earlier prediction of the growth trajectory. However, assumptions must initially be made to initiate the profile. Subsequent trajectory patterns must be inputted to re-evaluate during cultivation. The information includes the dry cell biomass composition for a microorganism, RQ, fraction of oxygen in air – solubility in liquid, medium composition, the $k_L a$, OUR and OTR.

The aim of this experiment using the DOT% was to characterise a 'profile' *in silico* and validate it against experimental data. When a profile can be established, controls can be predicted ahead of time. In Figure 40 the exponential growth profile has a good fit (shown by left blue shaded circle region in Figure 40) in the early growth phase when compared to the DOT% in the low and high rpm conditions. This was similar to the results of the section 5.3.2 whereby the exponential model was very good at simulating the lag phase. The sigmoidal growth curve (b, Figure 40) seems to have a lesser fit to the initial DOT% experimental data.

The model starts off by making one steady state solution of the model. Then taking this model solution i.e. output biomass and state variables of interest for the next (subsequent) ordinary differential equation (ODE) calculations. Each step follows in the similar fashion whereby calculation of uptake rates, energy and requirements to maintain the initial objective for biomass production occur. Then these calculations are used as constraints for the next calculation until end of the process. The ODE's included in calculations of interest for the medium were the OD, concentrations of sodium, phosphate, and glycerol. The goal was to mimic cellular metabolism under minimal medium conditions, with the non-organic components in the batch and the glycerol carbon source to support the predicted biomass growth target in our simulation, i.e. not allow the non-organic compounds to become growth limiting and determine the flow of metabolites through the metabolic network. The uptake reactions that were excluded from the model i.e. the reactions metabolites that were deemed not to change during the modelling was the carbon dioxide, oxygen, water and hydrogen.

For uptake flow (i.e. > 0) of metabolites during model setup, the concentration was assumed to be high enough to not be limiting. If the uptake rate for a nutrient was calculated to exceed the maximum uptake rate, then the specified maximum uptake rate 1000 mmol/g dw/h value was given priority instead of the calculated one.

Several simulations with various operating parameters were allowed to run, this was to understand the robustness and sensitivity of the predicted biomass output. When the glycerol was allowed to go unfeasibly high/ almost unlimited resource in batch mode, we experienced an impossible biomass solution of 100's of g/L. Under more suitable conditions of the simulation (i.e. glycerol at 800 mmol/L concentrations that reflected a real-life counter-part), experimental biomass data started to follow the simulations at low concentrations at small scale (≤1L). The first set of FBA/exponential model's simulations that were allowed to run, was developed to allow a feedback loop

to tell the simulation to stop if the substrates were. depleted). This end of phase feedback loop, can be seen in the figures below where biomass accumulation is cut off abruptly at the 30-hour mark, to indicate that one or more of the substrates have been depleted during the FBA model simulation. We also included an estimated exponential growth profile. This was to see how similar the two growth trajectories were comparing early lag phase and FBA lag models. There were similar profiles up to hour 24 (see Figure 41) except the exponential growth profile had a slightly raised estimation of how much glycerol was being converted to biomass at maximum specific growth rate between 24-30 hours.

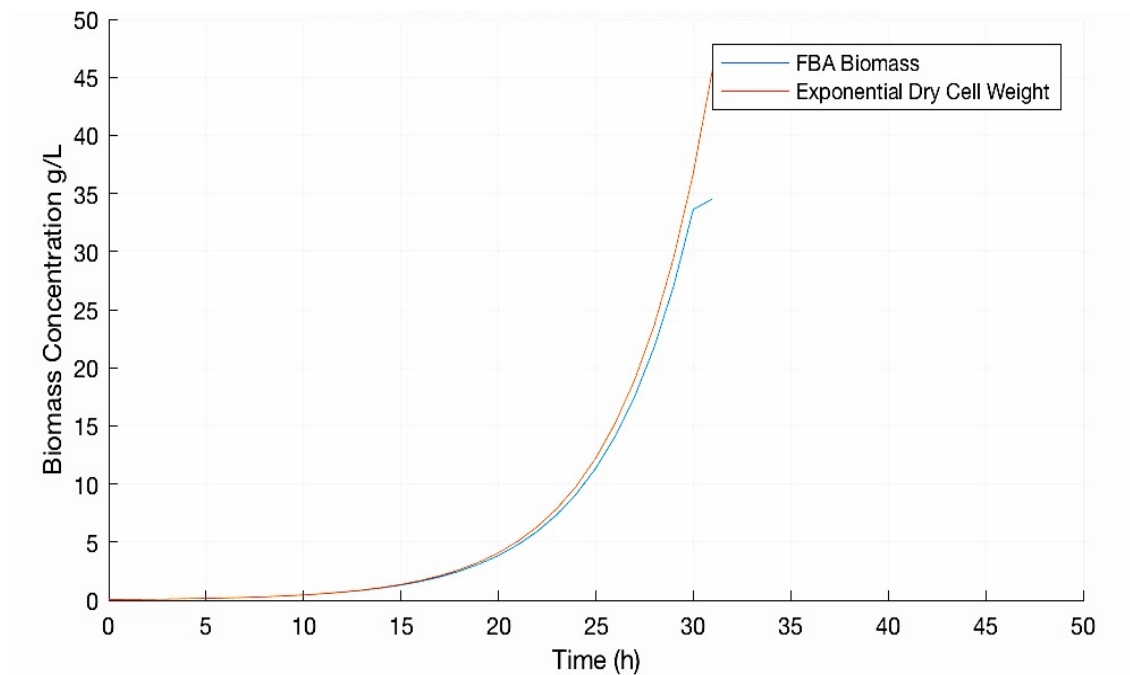**Figure 41 FBA simulation profile example compared with exponential predictive biomass growth profiling (using the typical Exponential growth Equation 55).**

It is important to note that the chosen calculations have two distinct features. An FBA model that includes glycerol metabolism and exponential biomass growth calculation which is not specific to organism biomass production and does not factor in the glycerol consumption.

As expected having the following insightful points leads to better control conditions;

1) How glycerol is being metabolised during culture
2) The duration or how long the lag phase takes at the start of batch
3) When the maximum specific growth rate is obtained,
4) Useful forward- planning for fed-batch process (See chart below at the point of the end of batch simulation phase) (Figure 42).
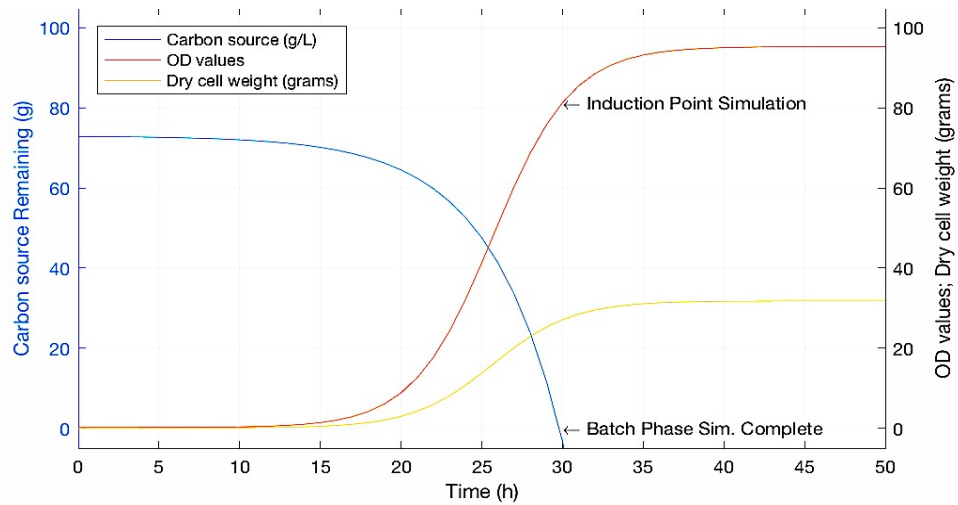
Figure 42 *E. coli* growth simulation using our modified logistic function, including the control strategies that could be used for feeding strategy or harvest. Carbon source consumption was calculated from stoichiometry of the *E.coli* cell and relating the increase of cells to a consumption rate of the carbon source.
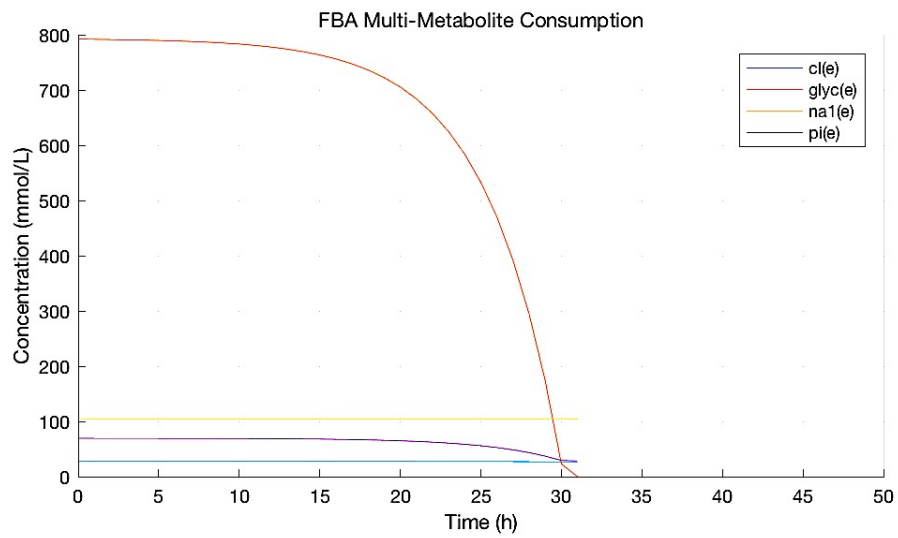
Figure 43 FBA multi-metabolite simulation, showing that the fermentation would end at 30 hours where the total limiting carbon-based substrate would be consumed (cl(e) are chloride compounds available, glyc(e) is the glycerol available, na1(e) is the sodium molecules and pi(e) is the phosphorus molecules available for metabolism.

Figure 44 Part of the development of a simulation software with which the soft sensor and control strategies for fermentation can be visually shown in graph format. This graph is a comparative plot between the exponential growth profile and FBA model.

All of the conditions used for the final simulations are described in the Appendix section 8, which are the simulations conditions shown in Figure 41to Figure 44. Within Figure 44 to Figure 47, initial simulations were started with the focus on use of the DFBA. Biomass growth simulation as well as a comparison of fermentation batch OD data and predictive OD modelling function as described earlier using Equation 56.

Figure 44 is a comparative plot between the exponential growth profile and FBA model. Here the model showed the exponential growth profile that was higher than the estimated biomass concentration using low concentrations of glycerol. Modelling using a non-specific growth calculation, one that doesn't take into consideration complex glycerol metabolism, has limitations leaving the conversion to biomass unregulated.

Exponential modelling is fundamentally susceptible to over estimations as it favours only looking at trajectories that favour maximum growth calculations. Under low glycerol concentrations, the glycerol carbon source in the flux

balance model can be regulated with potential to match the real-life counterpart and has better biomass estimation.
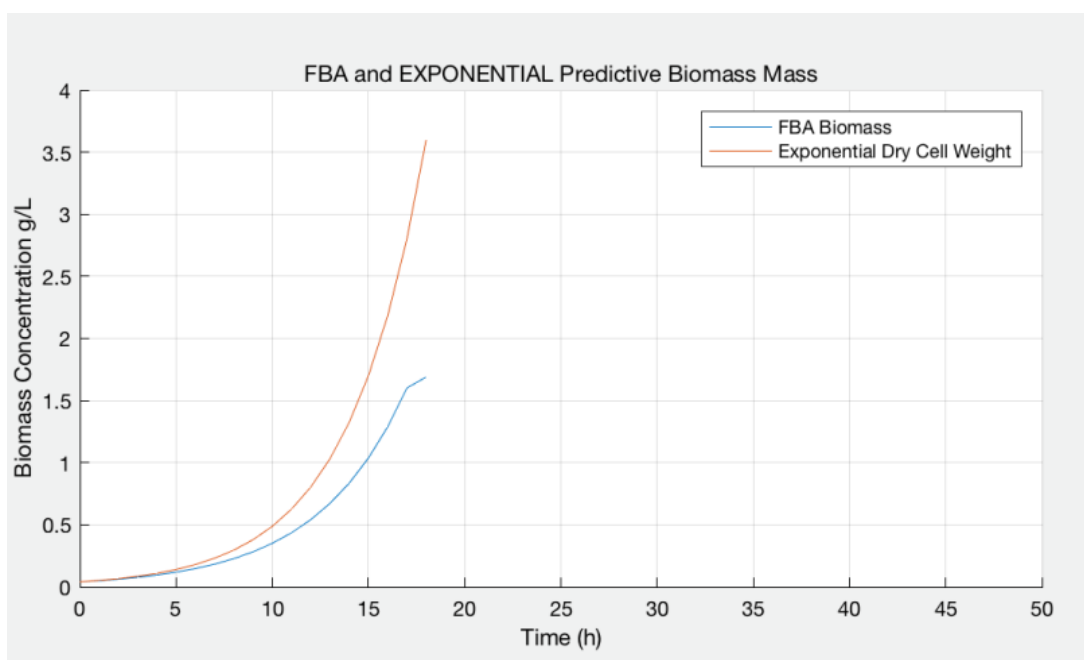

Figure 45 Part of the development of a simulation software with which the soft sensor and control strategies for fermentation can be visually shown in graph format. This graph shows the comparison between our modified logistic function biomass estimation and the experimental values. The ' ○ ' symbols refer to the experimental data points of fermentation carried out from Table 19.

Figure 45 shows the comparison between our modified logistic function biomass estimation and the experimental values from a real-life fermentation performed at 1L scale under the same conditions in the simulation. The results ('o' dots) show average optical density measurements (n=3) taken across a 24-hour period. These measurements show a close resemblance to the predicted OD values of the logistic function.

Figure 46 Part of the development of a simulation software with which the soft sensor and control strategies for fermentation can be visually shown in graph format. Graph using matlab dFBA first described in (Schellenberger *et al*, 2011), looking at the consumption of the main metabolite the glycerol carbon source used in minimal media M9 and the outcome of a simple optical density optimisation as a criterion.
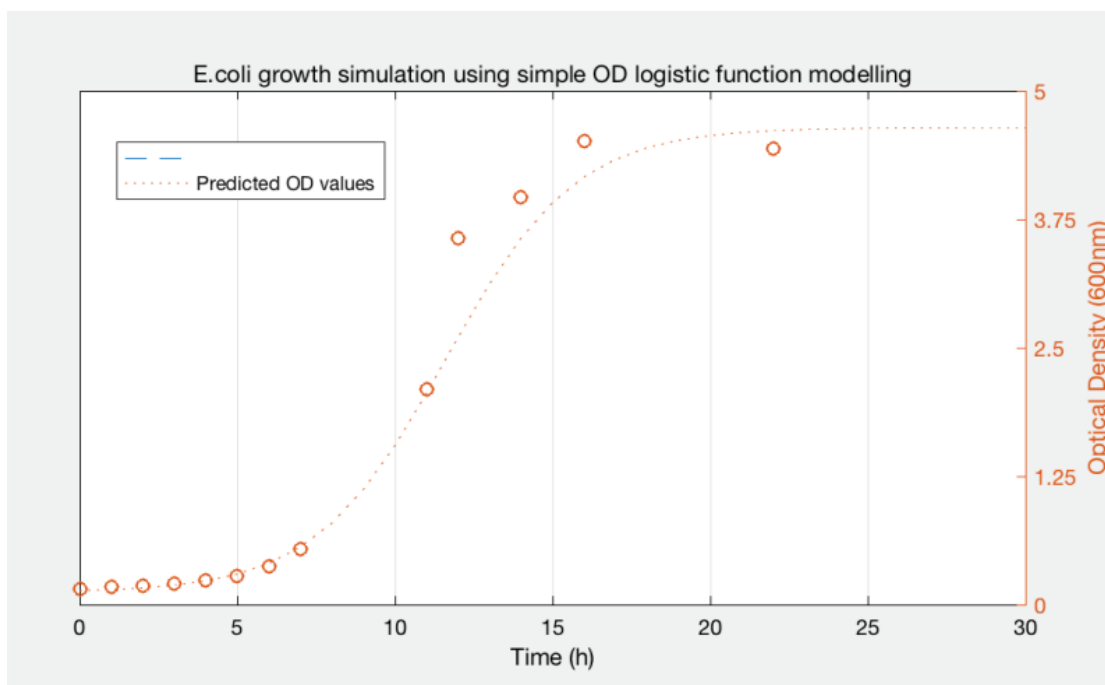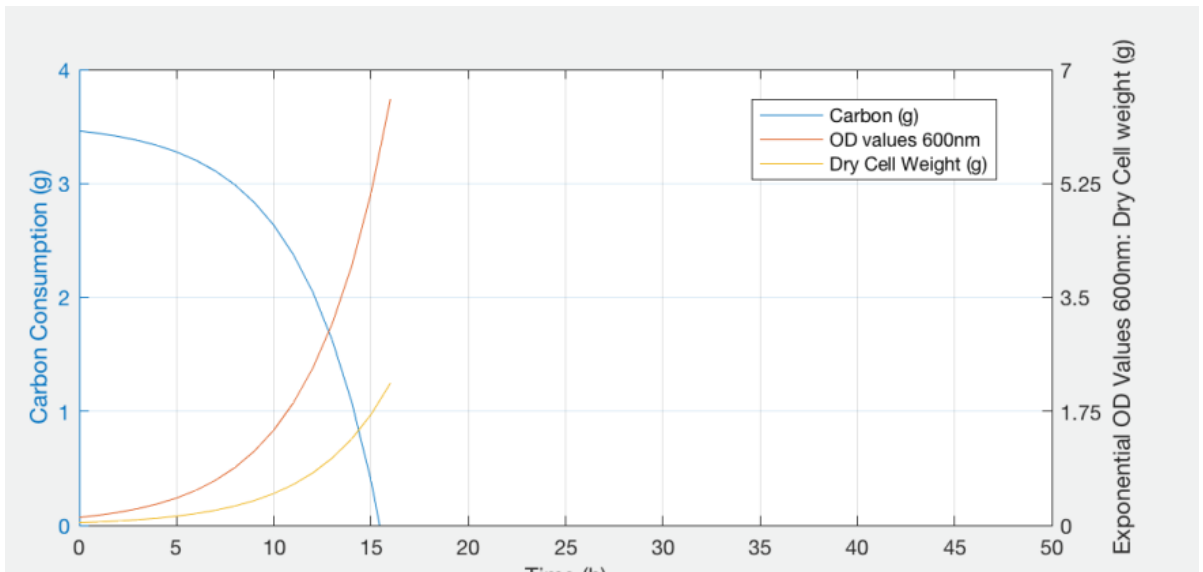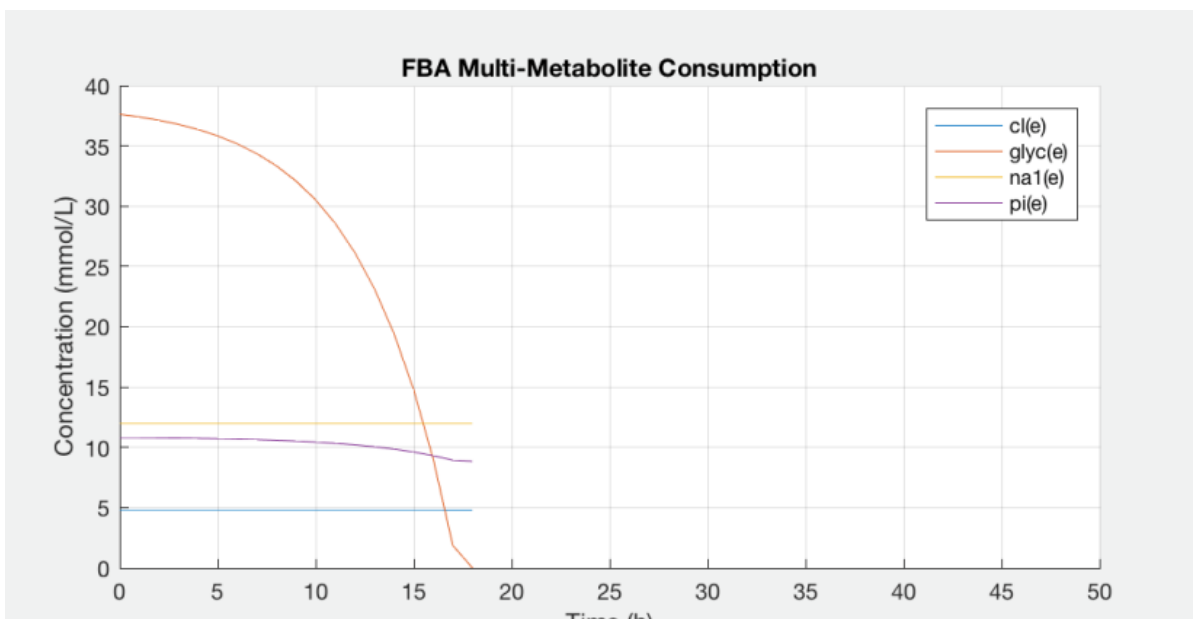

Figure 47 Part of the development of a simulation software with which the soft sensor and control strategies for fermentation can be visually shown in graph format. Graph using matlab dFBA methods first described in (Schellenberger *et al*, 2011) looking at the consumption of the metabolites used in minimal media M9.

Figure 46 and Figure 47 shows the metabolite consumption and optical density of the dFBA model based upon the setting we used for the live fermentation. One application explored was the creation of a well-planned control strategy based on results from:

1) Stoichiometric balancing
2) Metabolic simulations using the COBRA toolbox
3) Real fermentation data

The models created using the COBRA toolbox were based on several constraints or using an objective function to minimise or maximise a desired target. The best way to describe this was from a predefined priority system, with carbon being the most important parameter available.

The difference in priority lies in what the model analytics was trying to achieve – reducing microbial lag time within the fermentation from the start to reach exponential growth quickly. This is another application focus that can be used with these models.

The goals to be achieved from the data and simulations above were as follows:

1) Expected time taken for batch run to complete: The "minimum" focus would be to reach the quickest completion time possible, with the "maximisation focus" being on the shortest time taken to consume key substrates. Indicators in the simulation and real fermentation data are based on consumption of nutrient sources.

2) Value of media components: This is a way of including more expensive material components in the medium or for larger scale modelling and use of mass quantities. This is important when

looking at constraints and planning for budget based or sensitive margins of profit.

3) Biomass or OD targets: The criteria for this constraint are based upon the desired maximisation of biomass or OD alone.

The optimisation criteria that was used in this set up was the singular optical density targets. With this type of application use we obtained model data that could mimic a biochemical network, growth trajectory and metabolite consumption relatively easy based on stoichiometric and dFBA methods.

Finally looking at the overall limitations of the models is of utmost importance. Trust of the models essentially comes down to the inputs, mechanisms and validations. Some of the current limitations are addressed in this section. Further areas that caused concern for the accuracy of the models were the exponential carbon source consumption profile. Calculation of the carbon composition of the cell can be calculated experimentally on a cell by cell basis. Inaccuracy at this stage would cause simulation error in later stages. Other limitations include the need to measure OD and the dry cell weight, and simulation parameters when first establishing model simulation. Lastly, the oxygen requirements to fully convert the carbon source are large and the average air flow may not be reached in high cell-density cultures.

6 Project Conclusion

The parallel Multifors miniature bioreactors were characterised in terms of power input, mixing times, and oxygen mass transfer and checked for consistency across batch fermentations for the parallel bioreactor system. This was key to being able to establish insight into the bioreactors and what physical impact the have, such that the entire software packages and the framework could be built.

In the the project scope described the next layer up of *in silico* or software tool development (shown with connectors in green) i.e. our software management process. The design allowed planning for transfer or input and output of data between different software tools. One of our original issues to overcome with many software packages was the interoperability, which was also one of our research goals. The software tools built and tested for integration purposes are shown between blue connecters and labelled nodes in the main development scope (see Figure 48).

Rapid process optimisation using DoE as a practical tool was useful and designed to handle many complex systems but operate in the parallel bioreactor set up. A statistically robust DoE tool not only covers *E. coli* or yeast culture investigations in a parallel bioreactor format but also provides scope for a full range of applications, including process development, process optimisation, scale-up/down, media development, growth studies, toxicity testing, and high throughput screening. The first instance of our fermentation model was built from basic DoE software tools in Matlab. A model-based DoE tool is thus ready for the Multifors and parallel bioreactor systems that specifically targets designs that would incorporate 4 reactors set up.
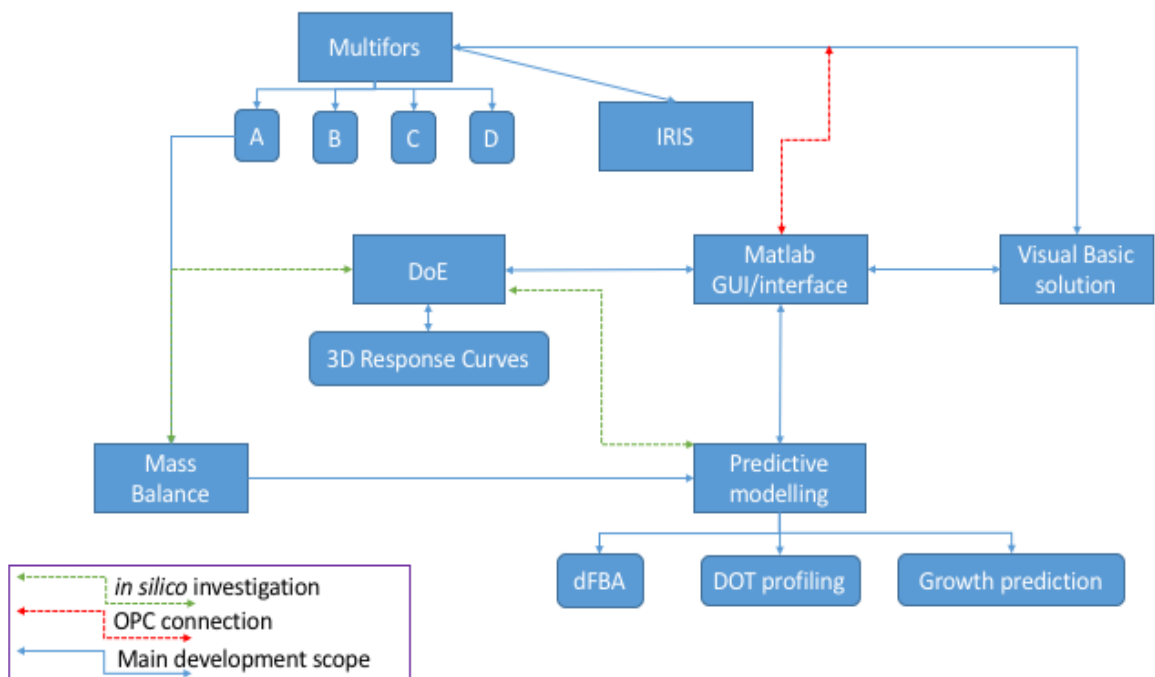
Figure 48 Overview of the Project

A major focus of integration of advanced modelling tools was the integration with bioprocess control software data and metabolic flux analysis software. A key issue was the development of an interface to enable exchange of different sets of data, namely transfer of data from different modelling software and conversion into optimised fermentation recipes and sequences. However, the program and fermentations were still not fully automated, but can at this point give the user the desired feed and strategy for biomass growth. As a proof of concept, the initial fermentation model based on DoE data was to be improved with metabolic flux analysis to develop an advanced batch feed stratagem.

The next objective was to build upon our gained fermentation knowledge and statistical modelling for predicting fermentation set up, operation, and management in one entire software tool. This was one goal of the industrial company, Infors HT who develops the parallel bioreactors. Then with DoE model data and with metabolic flux analysis help improve and develop for bespoke applications.

153

The outcome of this work was the creation of new DoE and DFBA tool/software packages for use with parallel bioreactor systems. We provide techniques for simulation, control and fermentation optimisation.

6.1 Software Validation

Consistency was important for testing modelling assumptions and establishing parameters to be used for building a mathematical model for *E. coli* growth our test application. This development is important in respect to software validation and verification. Software intended for operating control and model building should abide by guidelines set out by the regulatory bodies.

Formal verification and software review in summary should include test procedures, protocols, readiness reviews, and finally qualification. For further review and integrated software development code inspection, consisting of intensive manual error detecting techniques, is vital and was carried out in our project

Minimising risk/errors and ensuring compliance to coding standards set out in the validation documentation during inspection are vital steps. Check-listing is the most preferable technique.

Software elements are documents that should be produced while software development is in progress. This ensures traceability and assurance of software verification and validation. Face validity assessments are used to give the developer or user a measurable qualitative benchmark. It gives the user an idea of how well suited a variable is when used to test out a problem or, as in this case, a software programme. A proposed variable or test must be subjective and have a degree of measurability and do what it claims to do. In this case, DoE's must be statistically robust. The DFBA should describe the

flux of metabolites through each of the reactions, subject to constraints. When a test can accurately measure what it is supposed to measure, it is deemed valid. Examples of our code elements are provided in the appendix.

Some of the key aims of software validation in this project was identifying what management is needed at various stages of the software development cycle. These were set as management tasks that include planning and resources required to meet both the biological development of the process but also the technological development of the software controlling and tools being used. Prospective validation was used as an approach for the integrated software. Protocols would ideally include test parameters, product characteristics, equipment to be used, and acceptable test points. This was especially important for encompassing smaller software parts into a larger software package.

This prospective validation approach and its implementation was specific to our project for:

- Developing software models for fermentation simulations
- Face-validating the models with real-life data sets
- Using software models for predicting product quality
- Integrating statistical software into what is already existing (enhancing the existing software)
- Programming the logic controllers for controlling the reactor equipment during fermentation.

The input and outputs from each task, and anomaly reporting was documented (see appendix section 8.3 for an example). Off-the-shelf code auditing and reviewing software like those used in coding software development (e.g. Microsoft visual studio) could have been used further in identifying the learning outcomes set by this project. Risks and assumptions needed to be identified and allocated to each task set out in the software

process cycle were prepared and implemented throughout. Independent code review was carried out by Infors HT software team in this case.

6.1.1 Quality by Design

Quailty by design (QbD) approach could allow for an additional layer of guidance for future work in modelling and software development for parallel bioreactors. The trade-off for higher levels of regulation and employment of quality-by-design in a process which could lead to, better product quality, reduced process costs, and better time management (McCurdy, 2011).

QbD descriptions are normally quite broad and not directed to a specific process or type of manufacturing. QbD is the concept behind predefining and planning for new breakthroughs scientifically, yet ultimately planning quality before manufacturing starts (Juran, 1992). The QbD practice is useful when users build, plan, validate and deploy measures to enhance quality assurance, which is an overarching aim for the software development in this research study. QbD is the concept behind planning from relevant guidance manuals like cGMP, training, PAT, or experimental design. It would subjectively improve and move away from out-dated practices into a more desired up-to-date state (McCurdy, 2011).

An example of successfully employing QbD is that of "process control" where the results of introducing and building upon PAT in real-time feedback provide bespoke predictive controls for a given biological process. A typical process control element could be the automatic adjustment of the pump-feeding rate (the control element). This could be achieved by predicting the nutrient uptake rates i.e. feedback element from soft-sensing models and relevant PATs used at given points during microbial fermentations.

Scientific understanding is the key to having a well-defined process, requiring all variables that are critical to a given process to be characterised and

explained. Management of these variables is addressed and can be predicted reliably (McCurdy, 2011).

A successful QbD program could be like the one described below. (FDA, 2003; Rathore & Winkle, 2009)

1) Start by defining an objective for the given process. This can be broad like increasing or capitalising on new process capacity or based on business benefits, product safety, and systematic development.

2) Predefine and explain all the variables pertaining to the process inputs, e.g. broth purity, delivery rates, acid and base potency, etc.

3) Explain and manage output quality of variables like cell viability at n-time or other product or metabolite yields, stability, hydrophobicity, impurities, etc.

4) Complete risk assessment where necessary, considering factors (temp, pH) that may affect the outlined product quality variables and the original objective(s). Engineering theory such as scale-up and translation implications and using scientific theory is a requirement.

5) Understanding the manufacturing and culture process to show significant parameters that interact with and affect identified critical quality attributes (CQA) and other critical process parameters (CPP).

   a. Select PAT-approved equipment and practices as best as possible (in line/online) that would monitor a time-based response to the risk of changes to any CQA.

   b. DoE experimental plan setup: Establish the initial design space to map dependency of multiple variables identified from the risk assessment.

   c. Highlight the parameters that significantly interact with the CQA.

6) Use reaction kinetics, mass balance, and other deterministic approaches, like stoichiometry, for quantifying CPPs.

7) Highlight significant risks/threats.

8) Simplify process into a mathematical equation or model (quite tricky in some complex systems).

9) Check and validate this model.

10) Implement novel process controls.

11) Assess and continually improve for desired future QbD programmes and process states.

(FDA, 2003; Rathore & Winkle, 2009)

A QbD approach and implementation with respect to its software was a research objective.

6.2 Future work: Process Automation

Process automation was a key focus from the beginning of the project and one that is highly relevant to the future work of this project. Process automation with respect to the study was based on how to simplify the fermentation procedure for a user. Also, how could we describe what is required to reach an optimal bioprocess solution for the user. Is there a need to query a complex bioprocess upstream before starting a fermentation process? It certainly would be useful. How could we do this? Can I simply push a button and optimise the bioprocess? To do this, should we have started with one target, i.e. just the *E.coli* fermentation optimisation or multiple objectives, i.e. control software development and fermentation modelling. It was with such a query that was focused on multiple objectives to reach this research initial goal towards process automation. Process automation is ideal for further work.

Process automation would typically aim to integrate computer-based knowledge, such as applications and relevant products, and then assess it in order to reallocate resources continually throughout a process. Applying process automation aims with relevant goals that are specific to bioreactors, bioprocess control software, shakers, and fermentation applications can help with overall process optimisation and the goals set out in this project.

If qualitative oriented objectives could be applied, to a bioprocess or bio-products, then this could provide benefits like cost containment or an improved risk management. Risk management would help in reducing the time taken for a fermentation campaign or planning by simulating or preventing many "out of our control" risk factors.

There has been advancement with improved software and hardware within the competitive bio-industry in recent years. The execution of relevant process automation has resulted in enhanced quality of trust and robustness

in a bioprocess. This is evident in the dairy production industry where automation and machinery are commonly applied to produce enhanced products, consistency, scale, and improved knowledge. The successful application of automation has resulted in other industries, including the chemical and biochemical industries, adopting this approach (Junker and Wang, 2006).

Successful experimental microbial culture requires the identification of many complex input and output variables. These complexities include various physical and physiological properties, the biochemical environment, and engineering principles. Achieving a desired outcome from process input variables can be difficult to control and automate in practical terms. To address such difficulties, analytical and statistical strategies can be employed. Most statistical packages and analytical software used for control purposes are in fact isolated; the statistical programmes are not currently integrated sufficiently into bioprocess control software. This is usually disadvantageous considering that the standalone software is not specific to a unit step or process. The licensing of such software products is costly and additional training is required. This is because the software is often complex and necessitates more than general knowledge to use and draw conclusions from results.

Advancing analytical technologies are important as user tools. Requiring more understanding and elucidating a process where there are unknowns in the bioprocess. Most technologies and statistical software aim to provide models and allow for simulations with a good degree of satisfaction. Typically, software and the analytics are becoming as complex as the chemical and biochemical processes themselves (flux balance analysis, pathway analysis, metabolic level modelling, genomic characterisation etc.). These could provide detailed knowledge about the process and better control at various steps in a bioprocess. However, as many companies have been operating the

same equipment for decades, their automation, control software, and algorithms will be out of date.

This research includes the characterisation of the limitations faced using miniature or bench-top scale bioreactors. Further characterisation and successful high cell-density fermentation demonstrations are needed to improve the time taken between early and late bioprocess development stages. This allows the potential to then actually scale a process up to pilot plant scale and commercially larger plants. This understanding would help limit the costs to getting scale-up controlled in an acceptable way and have outputs that meet the needs of the company or department (stakeholders).

Process development, process analysis, and process control guidelines outlined by the FDA encourage innovative development work, and this research could provide the innovation needed for process automation (FDA, 2003). The risk-based approach would typically allow implementation of up-to-date control strategies to mitigate the risk of poor product quality. In the context of this research, this could mean implementing the scientific bases and engineering principles for fermentation culture at lab bench scale to scale-up for pilot plant processes or manufacturing. The aim would be to prevent the risk of unknown factors interfering with consistency and robustness for a product going into further development. As such, bioprocess modelling, simulations, and improved control strategies are continuously growing areas of research. The importance, in summary, is to mitigate the risks associated with these complex biological processes using these innovative solutions.

In practical terms, the data obtained from experiments would give a better understanding when conclusions are drawn rapidly. Critical quality attributes, which are usually predefined or identified before experimentation, are the typical benchmarks for product quality or performance. Limitations in innovation come from the laboratory methods themselves. For example, the

approach used for offline samples can be limited to only historical and data-driven models from previous experiments (Wechselberger, Seifert & Herwig, 2010). In contrast, live real-world, data-driven methods can provide better control strategies to be implemented while in motion or in progress. The result could hopefully provide the quality and performance that were originally outlined by the critical quality attributes. However, research like this would still need to be verified and assured.

Operating boundaries and operating parameters are the fundamental bases for process control. An unusual but relevant advantage of adopting PAT and quality-by-design (QbD) framework is that it can allow for improved robustness while operating at a higher risk. The risk involved is that the operating boundaries can be broader, capitalising on the wealth of live data to improve while in progress. This would require robust validated algorithms and potentially result in, for example, fewer rejected batches. Up-to-date soft sensors and hardware technology need to be adopted. The PAT framework is a welcome move, despite the discrepancies involved.

7 Bibliography

Aiba, S., Humphrey, A.E. & Millis, N.F., 1973. *Biochemical Engineering.* 2nd ed., pp.195-217. New York: Academic Press.

Antoniewicz, M.R., 2015. Methods and advances in metabolic flux analysis: a mini-review. *Journal of Industrial Microbiology and Biotechnology,* 42(3), pp.317-325.

Betts, J. & Baganz, F. year. Miniature bioreactors: Current practices and future opportunities. *Microbial Cell Factories*, 5(2006), p.21.

Bryant, J., 1977. The characterisation of mixing in bioreactors. *AdV. Biochem. Eng.,* 5, pp.101-123.

Cabaret, F., Fradette, L. & Tanguy, P.A., 2008. Gas-liquid mass transfer in unbaffled dual-impeller mixers. *Chem Eng Sci*, 63, pp.1636-1647.

Cants A.H.G., et al., 2005. Validation of the Danckwerts-plot technique by simultaneous chemical absorption of CO_2 and physical desorption of O_2. *Chem Eng Sci*, 60, pp.5809-5818.

Chang, L., Liu, X. & Henson, M.A., 2016. Nonlinear model predictive control of fed-batch fermentations using dynamic flux balance models. *Journal of Process Control*, 42, pp.137-149.

Chen, H-C. & Hu, Y-C., 2006. Bioreactors for tissue engineering. *Biotechnology Letters*, 28(18), pp.1415-1423

Betts, J.I. & Baganz, F., 2006. Miniature bioreactors: current practices and future opportunities. *Microbial Cell Factories,* 5, p.21.

Doig, S.D., Pickering, S.R.C., Lye, G.L. & Baganz, F., 2005. Modelling surface aeration rates in shaken microtitre plates using dimensionless groups. *Chem. Eng. Science,* 60, pp. 2741-2750.

Doran, P.M., 1995. *Bioprocess Engineering Principles.* 2nd edn. UK, Elsevier Ltd., pp. 8, 274-275.

Engenharia, E. D., 2016. Novel approaches for dynamic modelling of *E. coli* and their application in Metabolic Engineering.

Elmahdi, I., et al., 2003 pH control in microwell fermentations of *S. erythraea* CA340: Influence on biomass growth kinetics and erythromycin biosynthesis. *Biochem. Eng. J.,* 16, pp.299-310.

Fagerbakke, K., Heldal, M. & Norland, S., 1996. Content of carbon, nitrogen, oxygen, sulfur and phosphorus in native aquatic and cultured bacteria. *Aquatic Microbial Ecology.*Vol 10; 15-27
FDA, 2002. *General Principles of Software Validation; Final Guidance for Industry and FDA Staff.* Place: Health San Francisco, p.47.

FDA, 2003. *Guidance for industry PAT-A framework for innovative pharmaceutical development, manufacturing, and quality assurance.* Place: publisher.

Ferreira-Torres, C., Micheletti, M. & Lye, G.J., 2005. Microscale process evaluation of recombinant biocatalyst libraries: Application to Baeyer-Villiger monooxygenase catalysed lactone synthesis. *Bioprocess Biosystems Eng.,* 28: 83-93.

Garcia-Ochoa, F. & Gomez, E., 2009. Bioreactor scale-up and oxygen transfer rate in microbial processes: An overview. *Biotechnology Advances,* 27(2), pp.153-76.

Gill, N.K., Appleton, M., Baganz, F. & Lye, G.J., 2008a. Design and characterisation of a novel miniature bioreactor system for parallel microbial fermentation. *Biochem. Eng. J.,* 39, pp.164-176.

Gill, N.K., Appleton, M., Baganz, F. & Lye, G.J., 2008b. Quantification of power consumption and oxygen transfer characteristics of a stirred miniature bioreactor for predictive fermentation scale-up. *Biotechnol. Bioeng.,* 100, pp.1144-1155.

Hanly, T.J., Tiernan, A.R. & Henson, M. A., 2013. Validation and optimization of a yeast dynamic flux balance model using a parallel bioreactor system. *IFAC Proceedings Volumes* (IFAC-PapersOnline), 12, pp.113-118.

Henzler, H.J., 2000. Particle stress in bioreactors. *Advances in Biochemical Engineering Biotechnology,* 67, pp.35-82.

Hodgson BJ, et al., 2004. Intelligent modelling of bioprocesses: a comparison of structured and unstructured approaches. *Bioprocess Biosyst. Eng.*, 26, pp.353-359.

Hoffner, et al., 2009. A reliable Simulator for Dynamic Flux balance Analysis. *Biotechnology and Bioengineering*, 110(3), pp.792-802.

Hjersted, J.L. & Henson, M.A., 2009. Steady-state and dynamic flux balance analysis of ethanol production by *Saccharomyces cerevisiae. IET Syst Biol*, 3(3), pp.167–179.

Hughmark, G.A., 1980. Power requirements and interfacial area in gas-liquid turbine agitated systems. *Ind Eng Chem Process Des Dev,* 19, pp.638-641.

Islam, R.S., Tisi, D., Levy, M.S. & Lye, G.J., 2007. Framework for the rapid optimisation of soluble protein expression in Escherichia coli combining microscale experiments and statistical experimental design. *Biotechnol. Progr.,* 23, pp.785-793.

Islam, R.S., Tisi, D., Levy, M.S. & Lye, G.J., 2008. Scale-up of E. coli growth and recombinant protein expression conditions from microwell to laboratory and pilot scale based on matched $k_L a$. *Biotechnol. Bioeng.,* 99, pp.1128-1139.

Juran, J.M., 1992. *Juran on Quality by Design – The New Steps for Planning Quality into Goods and Services*. Place: The Free Press.

Junker, B.H. & Wang, H.Y., 2006. Bioprocess monitoring and computer control: Key roots of the current PAT initiative. *Biotechnology and Bioengineering,* 95(2), pp.226-261.

Kadlec, Gabrys & Strandt, 2008. *Data driven soft sensors in the process industry*. Place: Elsevier.

King, Z.A., et al., 2015. BiGG Models: A platform for integrating, standardizing, and sharing genome-scale models. *Nucl Acids Res.* 44 (D1): D515-D522

Kramer, M.A., Calo, J.M. & Rabitz, H., 1981. An improved computational method for sensitivity analysis: Green's function method with "AIM". *Applied Mathematical Modelling*, 5(6), pp.432-441.

Leighty, R.W. & Antoniewicz, M.R., 2011. Dynamic metabolic flux analysis (DMFA): A framework for determining fluxes at metabolic non-steady state. *Metab Eng,* 13(6), pp.745-755.

Löffelholz, C. et al., 2013. *Chem. Ing. Tech.* 85, pp.1-2, 40-56.

Luong, H.T. & Volesky, B., 1979. Mechanical power requirements of gas–liquid agitated systems. *AIChE J,* 25, pp.893-895.

Löffelholz, C., et al., 2011. Single-use Technology in Biopharmaceutical Manufacture. In: R. Eibl & D. Eibl (eds.). year. *Title.* Hoboken: John Wiley & Sons.

Maier, R.M., 2000. *Environmental Microbiology: Bacterial Growth – chapter 3.* Place: Elsevier Store.

McCurdy, V., 2011. Process Understanding. In: I. Houson, ed. Year. *Title.* 1st ed. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA.

Miller, D.N., 1974. Scale-up of agitated vessels gas-liquid mass transfer. *AIChE Journal*, 20(3), pp.445-453.

Moresi, M. & Patete, M., 1988. Prediction of kLa in conventional stirred fermenters. *J Chem Tech Biotechnol*, 42, pp.197-210.

Meusel, W., et al., 2016. *Recommendations for process engineering Characterisation of single use bioreactors and mixing systems by using experimental methods.* Frankfurt: Dechema, pp.4-51.

Micheletti, M., et al., 2006. Fluid mixing in shaken bioreactors: Implications for scale-up predictions from microlitre scale microbial and mammalian cell cultures. *Chem. Eng. Sci.*, 61, pp.2939-2949.

Micheletti, M. & Lye, G.J., 2006. Microscale bioprocess optimisation. *Curr Opin Biotechnol*, 17, pp. 611-618.

Nienow, A.W., Wisdom, D.J. & Middleton, J.C., 1977. Effect of scale and geometry on flooding, recirculation and power in stirred vessels. *Proc Eur Conf Mix,* 2, pp.1-16.

Nienow, A.W, Hunt, G. & Buckland, B.C., 1994. A fluid dynamic study of the retrofitting of large agitated bioreactors: Turbulent flow. *Biotechnol Bioeng*, 44, pp. 1177-1185.

Nikdel, A. & Budman, H., 2016. Identification of active constraints in dynamic flux balance analysis. *Biotechnology Progress*, pp.1–11.

Nelson, D.L. and Cox, M.M., 2008. Principles of biochemistry. 5th ed. New York: WHFreeman and Company.

Orth, J.D., Thiele, I. & Palsson, B.Ø., 2010. What is flux balance analysis? *Nature Biotechnology*, 28(3), pp.245-248.

Rathore, A. & Winkle, H., 2009. Quality by design for biopharmaceuticals. *Nature Biotechnology,* 27(1).

Rutherford, K., Mahmoudi, S.M.S., Lee, K.C. & Yianneskis, M., 1996. The influence of Rushton impeller blade thickness on the mixing characteristics of stirred vessels. *Trans IChemE,* 74(Part A), pp.369-378.

Salehi-Nik, N., et al., 2013. Engineering Parameters in Bioreactor's Design: A Critical Aspect in Tissue Engineering. *BioMed Research International*, 2013.

Saltelli, A. & Commission, E., 2004. Global Sensitivity Analysis An introduction. *Science*.

Saltelli, A., et al., 2008. *Global sensitivity analysis: The primer*. John Wiley & Sons.

Santek, B., Novak, S., Horvat, P., & Maric, V., 2004. Scale-up criteria of bioreactor. *Kem Ind,* 53(1), pp.7-24, 41.

Schmid, A., et al., 2001. Review article; Industrial biocatalysis today and tomorrow. *Nature*, 409, pp.258-266.

Schultz, D. & Kishony, R., 2013. Optimization and control in bacterial Lag phase. *BMC Biology*, 11, p.120.

Schellenberger, J., et al*.,* 2011. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox v2.0. *Nature Protocols*, 6, pp.1290-1307.

Stanbury, P.F., Whitaker, A. & Hall, S. J., 1995. *Principles of Fermentation Technology*. 2nd ed. Butterworth-Heinemann: Oxford; pp.248-249.

Villegas, R.M., 2016. Identification of Dynamic Metabolic Flux Balance Models Based on Parametric Sensitivity Analysis; Thesis.

Van't Riet, K., 1979. Review of measuring methods and results in non-viscous gas-liquid mass transfer in stirred vessels, *Ind. Engng Chem. Process Des. Dev.,* 18, pp.357-364.

Wechselberger, P., Seifert, A. & Herwig, C., 2010. PAT method to gather bioprocess parameters in real-time using simple input variables and first principle relationships. *Chemical Engineering Science,* 65(21), pp.5734-5746.

Zhuang, K., et al., 2011. Genome-scale dynamic modelling of the competition between rhodoferax and geobacter in anoxic subsurface environments. *ISME J*, 5, pp.305-316.

8 Appendix

Parameter	Value	units	Comments
Length of vessel (total)	22	cm	Measured value
Volume of vessel (total)	1.4	L	Measured value
Gas velocity	1	vvm	Variable operating parameter (to make FBA model assumptions valid
Temperature	37	Degrees Celsius	Controlled operating parameter
Stationary liquid pressure at base of Vessel (15.7 cm)	15.4	mbar	Calculated value (engineering basis)
Stationary liquid pressure at lowest point of the upper impeller (8.1 cm)	7.94	mbar	Calculated value (engineering basis)
Carbon source mass	3.461	grams	Calculated (stoichiometry)
Oxygen mass	1.872	grams	Calculated (stoichiometric)
Biomass	1.795	grams	Expected result for stoichiometric model.
R quotient	0.7		Expected basal metabolic rate quotient assumption.
Approximate biomass target	2	g/L	Target value (low for minimal medium chosen).
Ash content	5.5	%	(http://systemsbiology.ucs

			d.edu/)
E. coli Stoichiometry	(1,1.77,0.49,0.24) [C:H: O::N]	mol	
Molar yield	1.91138	mol/mol (Glycerol)	Calculated value (stoichiometry)
Molar yield	1.22905	Biomass/oxygen	Calculated value (stoichiometry)
Mass yield	0.548967	Biomass/Substrate [g/g]	Calculated value (stoichiometry)
Mass yield	1.01486	Biomass/oxygen	Calculated value (stoichiometry)
Average mass per cell	710	Femtograms	Fagerbakke, K., Heldal, M. & Norland, S., (1996)
Average carbon Per cell	350	Femtograms	Fagerbakke, K., Heldal, M. & Norland, S., (1996)
Average oxygen per cell	120	Femtograms	Fagerbakke, K., Heldal, M. & Norland, S., (1996)
Volume per cell	3.8	μm^3	Fagerbakke, K., Heldal, M. & Norland, S., (1996)
Average number of cells/OD	5×10^8	cells	
Molar carbon ratio in carbon source	0.391		Calculated (Stoichiometry)
Wet to dry cell weight ratio	3		
Oxygen solubility	0.00106	g/L	
Substrate saturation constant (Ks)	0.005	g/L	
Substrate (carbon based) Conversion Molecular weight	92	g/mol	Molar Mass value of carbon substrate

Maintenance energy	0.58	mmol/g/h	(http://systemsbiology.ucsd.edu/ 2013, Accessed; 16/04/2017)
Simulation Start time	0	h	Simulation parameter
Simulation End time	<50	h	Simulation parameter
Volume of liquid	0.001	m^3	Operating variable
Starting OD measured	0.15		Operating variable, measured at 600nm
Specific growth rate	0.22	h^{-1}	Simulation parameter/experimental data calculation
Lowest specific Kla achievable by reactor	33	h^{-1}	Engineering parameter
Highest Specific Kla achievable by reactor	827	h^{-1}	Engineering parameter
DOT cascade control level	20	%	Operating parameter/simulation parameter

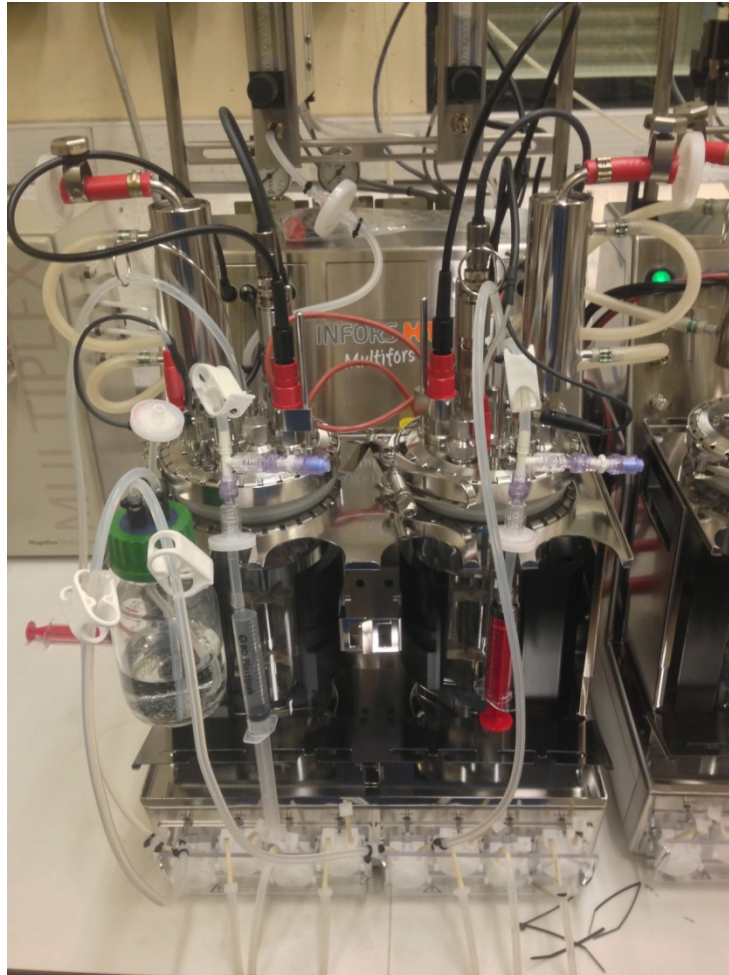# 8.1 Standard Operating Procedure

Multifors 2

Figure 49 Picture of the Multifors Bioreactor vessels

Preparation

SWITCH VESSEL ON.

Is the vessel clean? Vessel O ring ok?

Check walls for material sticking esp. around the working volume level. Remove the clamping ring holding the top plate to the vessel and lift the top plate clear check the O-ring on the underside of the top plate, replace if

necessary. Ensure when reassembling that the drive shaft sits vertically in the ceramic bearing and turns freely. Check for large cracks in the glass and replace if needed.

Port fittings?

Check each port fittings with O-rings and must not be flattened split of kinked. Check the glass vessel is seated properly and earthed with the metal prong. Secure the vessel by unhinging the clamps. Inoculation membrane (septum) fitted place the bottom 13.5Pg port fitting on top of the septum and replace the closure port fitting on top ready for sterilisation. Blank off any ports not used. Any unused pipe fittings not used not be tied off and looped together if needed.

pH calibration

Calibrate before autoclaving and each time the reactor is used. Do a two-point calibration with pH 4 and 7and temperature probe. Select bioreactor and pH calibrate on the bioreactor tab. When the reading steadies select confirm. Repeat for the second point calibration standard. Rinse with distilled water between standards.

Electrodes in place.

Fit p02 electrode and with O-ring and hand tighten only. Same for pH electrode. (these must be stored in 3M KCL so as not to dry out the electrodes.

Medium

Add medium to the vessel through an open port and hand tighten the port back. Make sure the fill volume is 90% of the WORKING volume, remaining space is for inoculum. (can be sterilised with 20 ml of water if heat labile media is used).

Air filters

Inlet air filter is fitted with a red filter, and clamped off for autoclaving. Exit gas cooler (green ribbon) connected and left open for autoclaving. All filters should be covered with a little cotton wool and aluminium foil.

Sample device

Ensure the sample device is attached to the top plate with silicone tubing and a cable tie and clamped off for autoclaving.

Reagents

Ensure feed and antifoam bottles are correctly filled and labelled and attached to the pumps. DO NOT autoclave acid or base. For the autoclaving remove the transparent cover shield and place the pump heads on the metal support under the bottle holder frame. Ensure the free silicone is attached to the vessel (one line into the pump head and one line to the vessel) CLAMP OFF.

Antifoam probe

Loosen the collar to fit the electrode careful not to damage the transparent sheath. Seat correctly at the required height and tighten the clamp

Autoclaving

- All electrodes capped (with foil)

The pH electrode can use the red cap to protect from moisture. The pO2 electrode can be seated with some aluminium foil.

- Filters covered with cotton wool and aluminium foil.
- Remove the temp sensor from its pocket
- Reagent and lines are clamped off
- Use autoclave tape

Sterilise at 121°C and allow to cool use gloves to transport the vessel over to a trolley.

CHECKLIST FOR VESSEL READY FOR USE

Vessel

Check that the stirrer is seated properly still. Fit the vessel into its holder and make sure its earthed properly. Fit the holder back to the Multifors base unit and secure the vessel.

Top plate connections

Fit the exit gas cooler connection (one for supply and return). pH electrode connects with its black cable (the one with the red cap). pO2 electrode connected. Place the temperature probe fully inserted into its pocket. Fit antifoam electrode (red into the electrode and black onto the top plate to make a circuit). Remove foil from air inlet and connect the silicone tubing on the rotameter (on the base unit).

Air and water supply switched on to 0.5vvm

Remove the clamp on the air inlet. Set the air inlet to 0.5vvm.

Replace peristaltic pumps

Fit the removable pump heads to the pump motor shafts beneath the vessel. Replace the transparent safety cover.

Tubing unclamped and prime the lines.

Use the switches on the front of the base unit to fill the tubing until the metal inlet pipe on the top of the vessel. Keep the sample line remained closed off until needed.

Set points limits set correctly.

For *E.coli* set the temperature to 37 degrees and a set point for 300rpm make sure the rest of them are set to OFF. Check the stirrer turns freely.

Minimum 2 hours elapsed

3-4 hours would be needed for electrode polarization could leave this overnight to inoculate in the morning.

Calibrate oxygen electrode 100% value

Adjust 1250rpm and 1.5vvm for 10-15 mins and saturate with nitrogen (for 0% calibration) and then oxygen (for 100% calibration) go to p02 calibrate and set high ref value (for 100%) and low ref point (for 0%) once readings are steady.

Decide which controllers are on or off

Temperature, stirrer, pH control, p02 control, AF, FEED off initially. Cascade controller for stirrer speed set. Select pO2 parameter select cascade, stirrer speed and max 1200 and min 300. Move to ON.

START EXPERIMENT;

With the parameters switched on manually IRIS can be started just before inoculation, check the link to the software is OK. SET feed pump rate after 4 hours if doing a fed batch process. This is set between 2-100% set to 2 percent for a low feed rate.

INOCULATION

This should have been setup the previous afternoon. The medium in the flasks should be cloudy compared to the negative controls. Ensure the consumables for the first inoculation are ready (this is for aseptic conditions a syringe and needle). Make sure that for the first sample OD is to be measured with pipette tips and cuvettes). The inoculum is usually 5-10% of the working volume. (but usually going to choose 2.5%) for example a 600ml working volume this would be 15ml. remove inoculation port closure and add ethanol to sterilise septum. Pierce membrane with needle and inject. Spray with ethanol again and close the septum with port closure.

Sampling

The clamp is removed from the sample tubing. A syringe is fitted to the air filter connected to the short pipe on the head plate of the sample device. Another syringe to the draw out solution side. Transfer sample to a bottle if required. Or cuvette for OD sampling.

T2, T4, T6, T8 ready all the way to the end of experiment.

Checklist for the end of the experiment.

Set all parameters to OFF before handling pump tubing.

Pump tubing emptied and washed

For each reagent bottle empty out and fill with distilled water. Use the all tab to empty reagent tubing and wash out with water into the reactor vessel. Re-clamp the lines for autoclaving filled with water.

Culture decontamination

A more usual procedure is to heat kill with steam in an autoclave.

The reagent tubing is prepared as above and filled with water.

The pH and pO2 electrodes are re-capped the exit gas cooler is disconnected from the water supply. the temperature probe is removed from its pocket in the vessel.

Latch the side shields together to free the vessels. Cover the filters with foil again.

Place the vessel, holder and bottles in the autoclave at sterilize at 121°c for one hour of other time and temperature as directed.

Vessel disassembled and cleaned

Latch the side shields to free the vessel for removal. Remove the H and pO2 electrodes carefully, rinse them in distilled water and replace their end caps) ensure some electrolyte is in the pH sheath. Remove the sample bottle and

clean. The top plate, drive shaft and sparger should be rinsed clean under a tap (a deep sink is required). The vessel glass can be removed from the holder CARE WITH HANDLING and the heat killed culture tipped out into the sink. Do not lose the bottom support with the ceramic bearing.

Post decontamination clean/check

Reagent bottles emptied but tubing filled with distilled water

pH and pO2 electrodes stored properly

vessel dry and re-assembled onto top plate

Vessel checked for marks, cracks and fouling

Areas around the bioreactor cleaned and any spills decontaminated and wiped up.

SERVICES (COOLER AIR and NITROGEN) are switched off.

No waste material left in the sink.

8.2 Appendix for Chapter 5 Design of Experiments analysis: Factor Input vs. Response Output Interactions

Below are the typical factor response box plots that are made from a DoE experiment. These indicate the data spread across the design space and the larger the box plot the larger the degree of interaction between the data
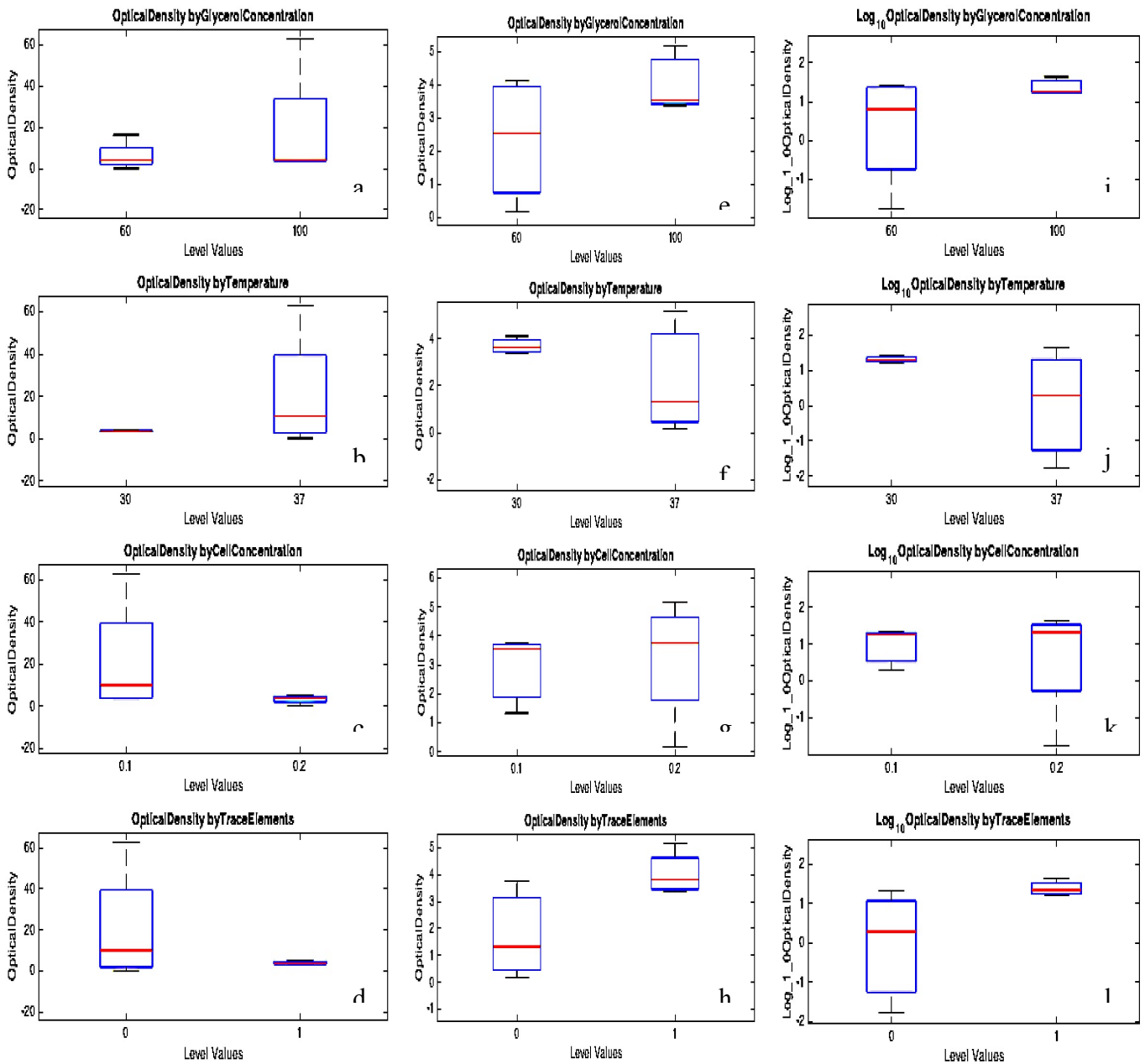

Figure 50 Optical density box plots showing the interaction between the variable inputs and the response value outputs. The graphs are organised vertically per model refinement. Left column (unrefined model), central column (data-refined model), right column (data-transformed model).

collected from the responses to the input parameters.

Here, we look again at dividing the plots in Figure 50 into three columns, in order of the models we tried to build. To summarise, the first column of interaction box plots (a,b,c,d), refers to the raw unrefined data model creation before removal of the outliers. Second middle column set of graphs (e,f,g,h) show the model with data refinement (i.e. outliers removed). and the third column in Figure 50 the data transformed using the (log y) transformation.

Several factors appeared to have changed the average response level between the first and second models. In the first model (the unrefined model), only half of the interaction plots have a large spread across each of the levels of the design space that we investigated. On refining the model (i.e. data refining and initialising the model), the data were found to have a greater spread across both levels than with the first unrefined model.

At the final attempt, we found that the factors appeared to have changed the average response level, whereas previously, only half had a large spread at each of the levels, this time, the data had an almost original spread across both levels.

8.2.1 Model Initialisation with no Data Refinement or Transformations (Model 1) and ANOVA

```
Generalized Linear regression model:
    OpticalDensity ~ [Linear formula with 5 terms in 4 predictors]
    Distribution = Normal

Estimated Coefficients:
                             Estimate      SE        tStat      pValue

                             _____   _____   _____   _____

    (Intercept)               -45.113     84.828    -0.53181    0.63172
    GlycerolConcentration     0.57213     0.50966     1.1226    0.34336
    Temperature                  1.02      2.9124     0.35023    0.74931
    CellConcentration          -81.25      203.87    -0.39855     0.7169
    TraceElements_1             -20.5      28.831    -0.71104    0.52836


8 observations, 3 error degrees of freedom
Estimated Dispersion: 416
F-statistic vs. constant model: 1.09, p-value = 0.492
```

Factor (unbalanced 2-level factor design) with all data.	Probability of confidence (%)
'GlycerolConcentration'	65.66
'Temperature'	25.06
'CellConcentration'	28.31
'TraceElements'	47.16

On examining the p-values of the 5 model coefficients in Table 21, no terms appear significant at our desirable 0.1/90% confidence level. However, obviously being cautious, these terms fell at only 51% confidence or above using the GLRM method. The stepwise method produced results

Further analysis of for our first model 1 but instead using stepwise regression methodology yielded:

Method:

We used the stepwise general linear modelling method to create our model from the actual data. In this instance, the non-significant terms were removed see (steps above) and allowed to refit our model. It was found that the R-squared and R-squared adjusted values were still not ideal. The ANOVA table shows us that the model was not a significant fit, and the lack-of-fit test further shows this as significant (<0.05). In ideal cases, we would be looking for no lack of fit in our model.

We used stepwise linear modelling to calculate a model and to fit parameters to the model. Stepwise method started from one model, the constant, and subtracted terms one at a time (although addition of terms can also be applied), choosing an optimal term each time in a greedy fashion, until it cannot improve further. Using stepwise fitting to find a good model, which is one that has only relevant terms.

The result depends on the starting model. Usually, starting with a constant model leads to a small model where a constant model has been used by default. Starting with more terms can lead to a more complex model but one that has lower mean squared error. You cannot use robust options along with stepwise fitting. Robust fitting saves you the trouble of manually discarding outliers so after a stepwise fit, examine your model for outliers.

The results of fitting for model 1 yielded steps that tested all main factors using a "stepwise" general linear fitting method. The resultant steps taken were as follows;

Table 22 Stepwise regression method individual step calculations

Order	Action	Factor Interaction	F Stat	P value
1	Removing	GlycerolConcentration:Temperature	NaN	NaN
2	Removing	GlycerolConcentration:CellConcentration	NaN	NaN
3	Removing	GlycerolConcentration:TraceElements	NaN	NaN
4	Removing	Temperature:CellConcentration	Inf	NaN
5	Removing	GlycerolConcentration^2	-Inf	NaN
6	Removing	Temperature^2	Inf	NaN
7	Removing	CellConcentration^2	NaN	NaN
8	Removing	Temperature:TraceElements	0.7037	0.5557
9	Removing	Temperature	0.1072	0.7744
10	Removing	CellConcentration:TraceElements	0.8863	0.4160
11	Removing	CellConcentration	0.0853	0.7848

Estimated Coefficients include the results as follows:

Table 23 Statistical analysis and estimated coefficients using the stepwise method for model fitting to raw data as model 1

	Factor Interaction	Estimate	Standard error	tStat	pValue
1	Intercept	-28.2183	24.6478	-1.11449	0.3041
2	Glycerol Concentration	0.6993	0.3324	2.1042	0.0893
3	Trace Elements	-30.6767	13.2940	-2.3076	0.0691

	Values
Number of observations	8
Error degrees of freedom	5
Root Mean Squared Error	16.3
R-squared	0.566
Adjusted R-Squared	0.393
F-statistic vs. constant model	3.26
p-value	0.124

The model had a weak R-squared value (0.566) but the fact that R-squared adjusted is considerably smaller indicates that there was undoubtedly previously had some terms in this model that were not significant (see stepwise removal steps above). Scanning the column of p-values showed these glycerol concentration and trace elements at the 0.10 confidence level. The combination of estimate values and the weak correlation of this model suggests that the model is still not suitable to accurately predict a response from our experimental data.

Recommendations at this stage were to include checking how significant the lack of fit was to the experimental data. After this percentage of confidence was calculated.

After removing the non-significant terms from the model and refitting the following analysis of variance table was produced;

	SumSq	DF	MeanSq	F	pValue
Total	3056.4	7	436.62		
Model	1730.9	2	865.44	3.2646	0.12386
Residual	1325.5	5	265.1		
. Lack of fit	1181	1	1181	32.707	0.0046253
. Pure error	144.44	4	36.109		

Using coded variables for our models

A = Glycerol Concentration

B = Temperature

C = Initial Cell concentration

D = Trace Elements

General linear regression method results (model 1)

Optical Density $\sim$ -45.113 + 0.572*A + 1.02*B – 81.25*C - 20.5*D

~51% degree of confidence

Equation 59

Stepwise Linear Regression results (model 1)

Optical Density $\sim$ -28.2183 + 0.699*A - 30.6767*D

~88% degree of confidence

Equation 60

8.2.2 Model Refinement with Data Refining only (model 2) and ANOVA

With outliers removed and the model refitted to the experimental data, the results of fitting the trial model that included all the main factors using a "stepwise" general linear fitting method involved the following steps:

Table 24 Generalised linear regression model (GLRM) model initialisation and significance of main factor effects on Optical density.

```
Generalized Linear regression model:
    OpticalDensity ~ [Linear formula with 5 terms in 4 predictors]
    Distribution = Normal

Estimated Coefficients:
                            Estimate        SE        tStat       pValue
                          _____   _____   _____    _____

    (Intercept)              4.5273      8.3821      0.54011     0.64322
    GlycerolConcentration   -0.0050417   0.063144   -0.079844    0.94363
    Temperature             -0.045714    0.27949    -0.16356     0.88511
    CellConcentration        -6.65       19.565     -0.3399      0.76631
    TraceElements_1          2.5867      3.1949      0.80963     0.50316


7 observations, 2 error degrees of freedom
Estimated Dispersion: 3.83
F-statistic vs. constant model: 0.654, p-value = 0.679
```

Table 25 Confidence levels for each chosen factor using GLRM

Factor (unbalanced 2-level factor design) with all data	Probability of confidence (%)
'GlycerolConcentration'	5.64
'Temperature'	11.49
'CellConcentration'	23.37
'TraceElements'	49.68

The results of fitting for model 2 yielded steps that tested all main factors using a "stepwise" general linear fitting method. The resultant steps taken were as follows;

Order	Action	Factor	Fstat	pValue
1	Removing	GlycerolConcentration:Temperature	NaN	NaN
2	Removing	GlycerolConcentration:CellConcentration	NaN	NaN
3	Removing	GlycerolConcentration:TraceElements	NaN	NaN
4	Removing	Temperature:CellConcentration	NaN	NaN
5	Removing	Temperature:TraceElements	NaN	NaN
6	Removing	GlycerolConcentration^2	Inf	NaN
7	Removing	Temperature^2	Inf	NaN
8	Removing	CellConcentration^2	NaN	NaN
9	Removing	GlycerolConcentration	0.0127	0.9285
10	Removing	Temperature	0.0274	0.8838
11	Removing	CellConcentration:TraceElements	2.1527	0.2386
12	Removing	CellConcentration	0.4182	0.5531

Table 26 Linear Regression model and formula

```
Linear regression model:
    OpticalDensity ~ 1 + TraceElements

Estimated Coefficients:
                    Estimate       SE       tStat       pValue
                    _____    _____    _____    _____

    (Intercept)      1.7533     0.76031     2.3061    0.069244
    TraceElements_1  2.2917      1.0058     2.2785    0.071675


Number of observations: 7, Error degrees of freedom: 5
Root Mean Squared Error: 1.32
R-squared: 0.509,  Adjusted R-Squared 0.411
F-statistic vs. constant model: 5.19, p-value = 0.0717
```

	SumSq	DF	MeanSq	F	pValue
Total	17.674	6	2.9457		
Model	9.003	1	9.003	5.1914	0.071675
Residual	8.671	5	1.7342		

As with the stepwise method chosen, the obvious outliers where excluded from fitting the model to the experimental data. In some cases, it was difficult to ascertain a good general linear model fit, as evident in this case. This was realised when the ANOVA analysis had no significant terms left over after data fitting, i.e. all the p values were above 0.1. We used the stepwise general linear modelling method to create our model from the actual data; in this instance, the non-significant terms were removed and allowed to refit our model. It was found that the R^2 and R^2 adjusted values were still not ideal, i.e. 0.509 and the adjusted 0.411. However, the model generally shows an improvement over the previous model and will be accepted as possibly the best that could be done without conducting a new experiment designed to fit a quadratic model at this stage of model processing. The ANOVA table showed that the model had a significant fit of over 90% confidence, i.e. the lack-of-fit test was not a problem. As with ideal cases, we would be looking for no lack of fit in our model; we have one ideal case for our model fit to data.

Using coded variables for our models

A = Glycerol Concentration

B = Temperature

C = Initial Cell concentration

D = Trace Elements

General linear regression method results (model 2)

$$Optical\ Density \sim 4.527 - 0.005*A - 0.046*B - 6.65*C + 2.58*D$$

~32% degree of confidence

Equation 61

Stepwise Linear Regression results (model 2)

:

$$Optical\ Density \sim 1.7533 + 2.2917*D$$

~92% degree of confidence

Equation 62

8.2.3 Further Analysis of Model Refinements

Testing the model assumptions using residual graphs for model 1

To examine the assumption that the residuals are approximately normally distributed, are independent, and have equal variances, we generated several plots.
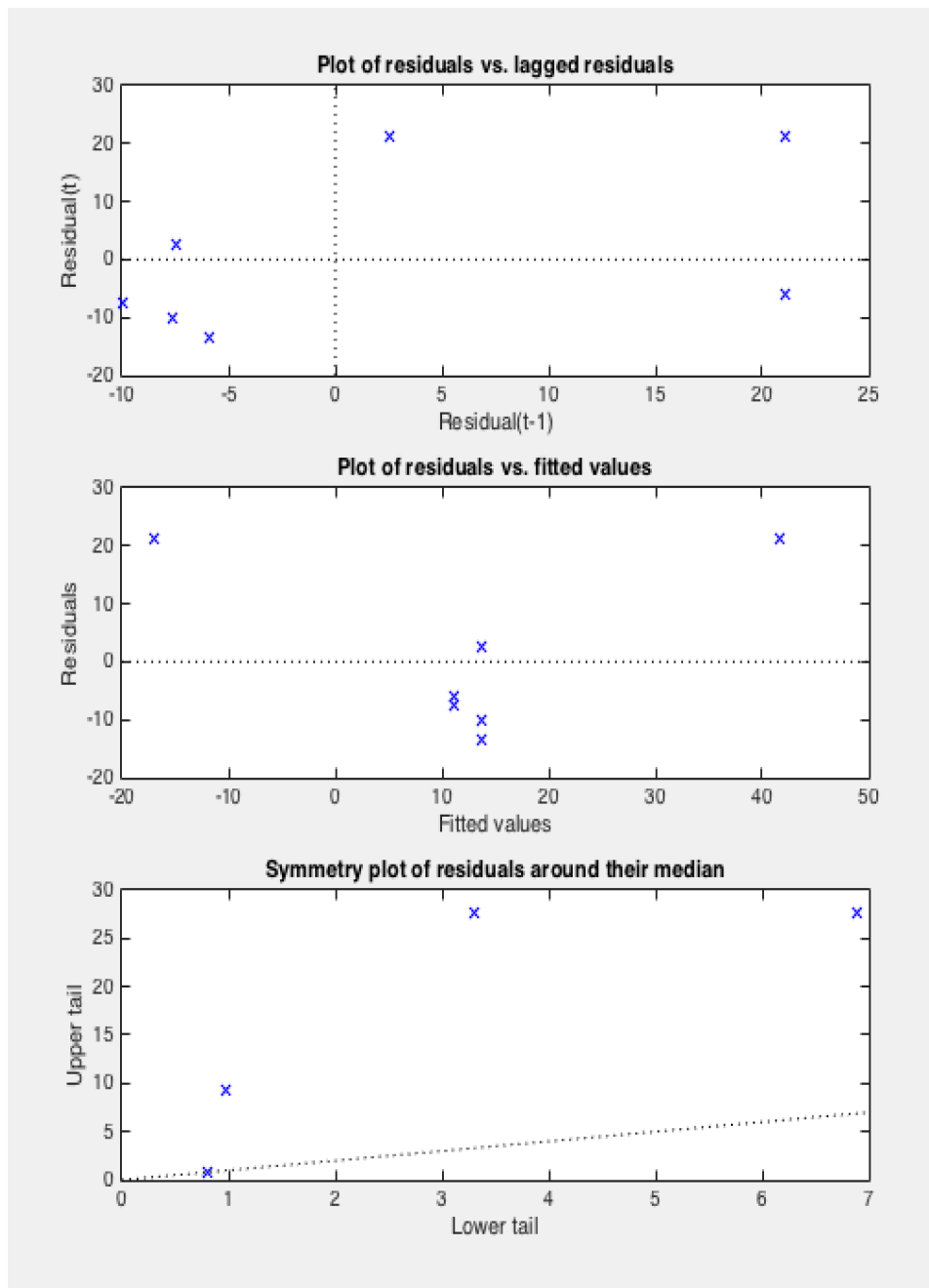
Figure 51 Diagnostic information for the model development (round 1)

Here in this analysis, the aim was to determine whether there was serial correlation among the residuals. The scatter plot shows more crosses in the upper-right and lower-left quadrants than in the other two quadrants, indicating positive serial correlation among the residuals. Another potential issue is when residuals are large for large observations. There is some tendency for larger fitted values to have larger residuals. Perhaps, the model errors were proportional to the measured values.

**Testing the model assumptions using residual graphs for model 2**
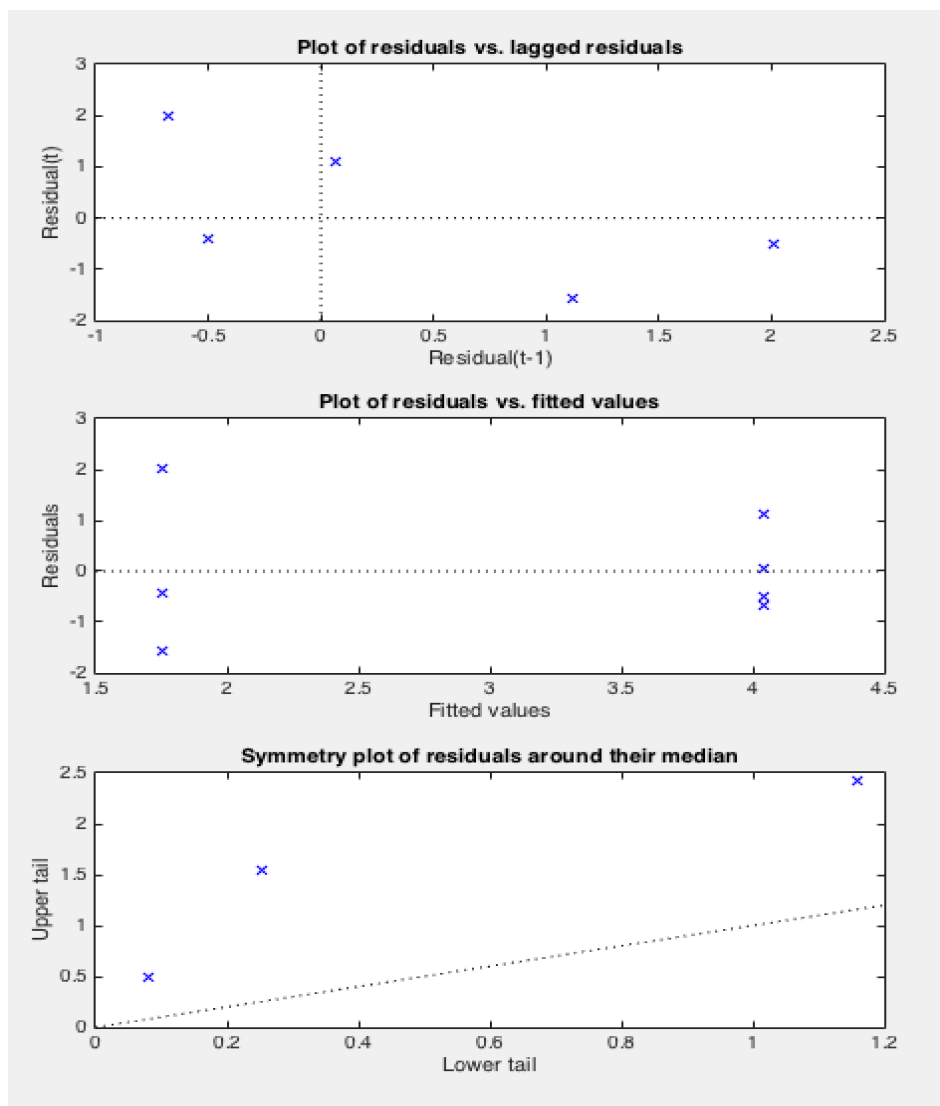
Figure 52 Diagnostic information for the model development (round 2)

Again, analysis of the residuals was conducted to determine whether there was serial correlation among the next model (2) residuals. The scatter plot shows equal crosses in the upper-right and lower-left quadrants than in the other two quadrants, indicating neutral serial correlation among the residuals. Another potential issue is when residuals are large for large observations. There is now a very low tendency to overestimate lower values. Again, perhaps, the model errors were proportional to the measured values.

Testing the model assumptions using residual graphs for model 3
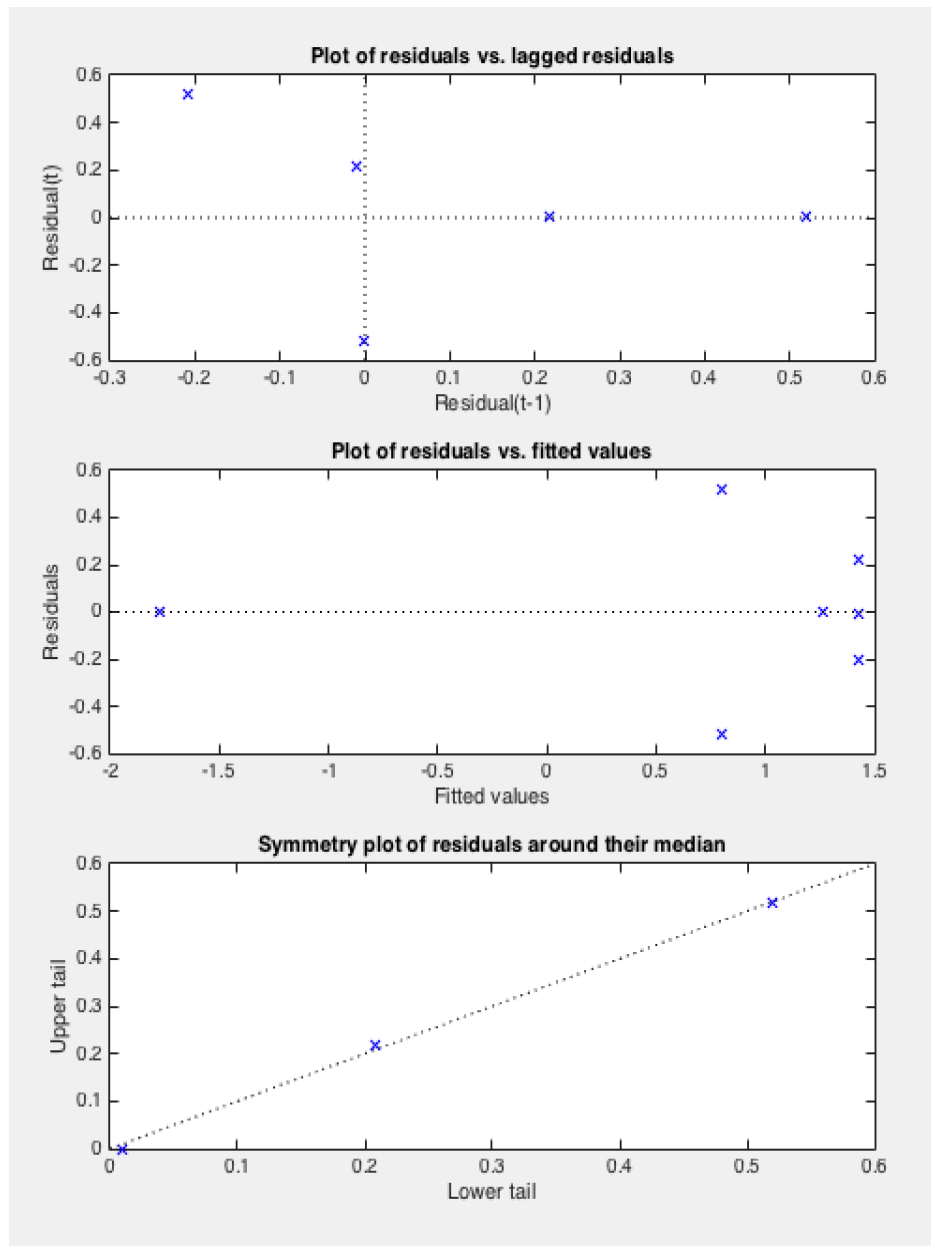
Figure 53 Diagnostic information for the model development (round 3)

Lastly this set of residual analysis once again aims to determine whether there was serial correlation among the model (3) residuals. The scatter plot shows equal crosses in the upper-right and lower-right quadrants than in the other two quadrants, indicating neutral serial correlation among the residuals. There was now a very low tendency to overestimate around the middle

(centre points) of these values. Again, perhaps, the model errors were proportional to the measured values. There is excellent symmetry of the residuals around their median – this was as linear as possible.

Testing the model assumptions for distribution, probability and case order (model 1)
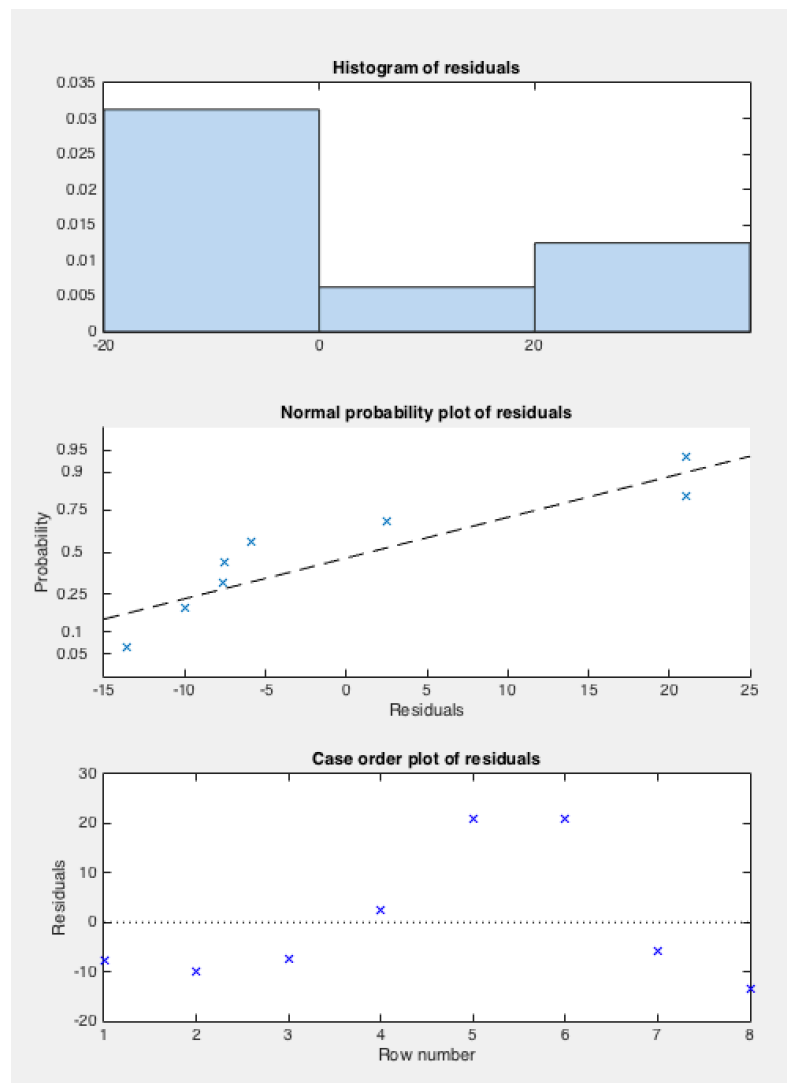
Figure 54 Diagnostic information for the model development (round 1) involving testing model assumptions

The values above 20 in Figure 54 histogram of residuals are potential outliers. It is difficult to understand from the normal probability plot the data distribution

194

and its effect on the linearity of the probability, implying a suboptimal fit to normally distributed residuals. From the histogram of residuals, the residuals are not normally distributed.

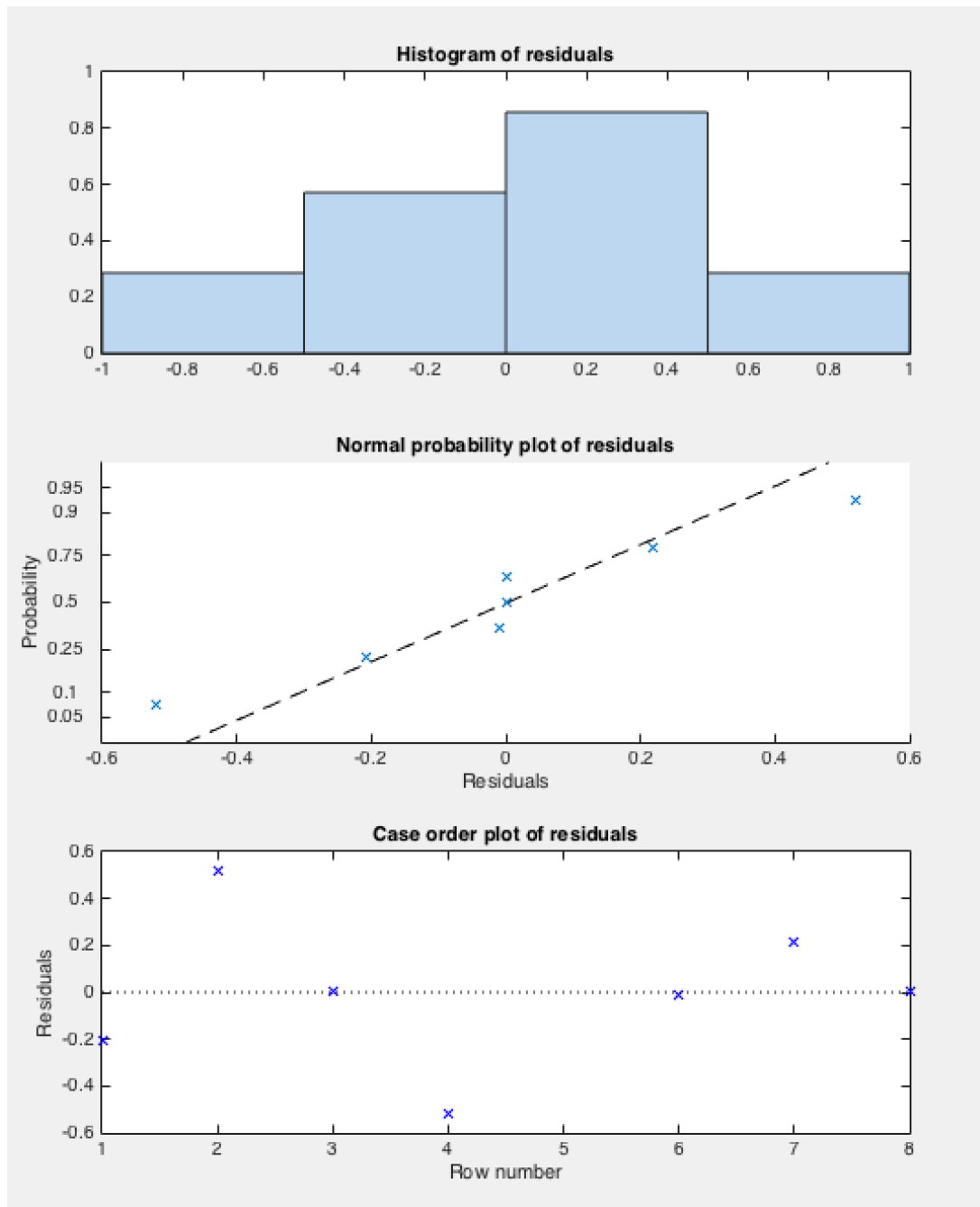## Testing the model assumptions for distribution, probability and case order (model 2)
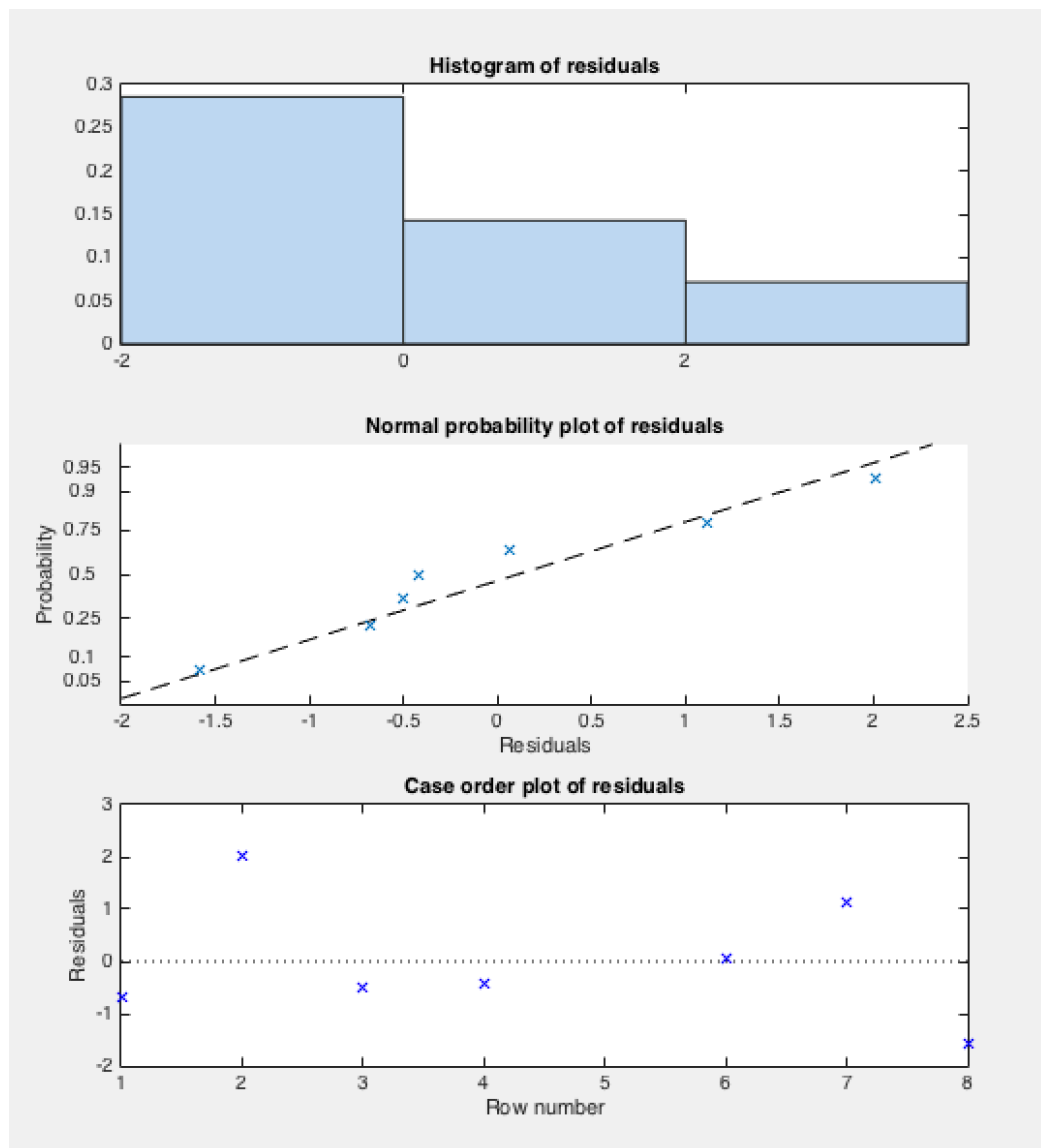
Figure 55 Diagnostic information for the model development (round 2) involving testing model assumptions

The histogram of the residuals for model 2 helped in identifying the symmetry of the data around the zero point. The new residuals plot looks weakly symmetric, without the obvious problems from before. However, the normal probability plot of residuals showed a much better linear correlation. This was now a lot better than the previous model. However, there might be a degree of serial correlation among the residuals, resulting in the need to create another set of data before using similar methods in this model fitting (going from model 2 to model 3).

Testing the model assumptions for distribution, probability and case order (model 3)

Figure 56 Diagnostic information for the model development (round 3) involving testing model assumptions

Our histogram of the residuals for model 3 helped us identify the symmetry of the data around the zero point. The new residuals plot looks a lot better, the best in fact, without the obvious problems from before in trial fitting model 1. Moreover, the normal probability plot of residuals shows a much better linear correlation. The residuals were also evenly distributed across each of the row numbers without any obvious pattern.
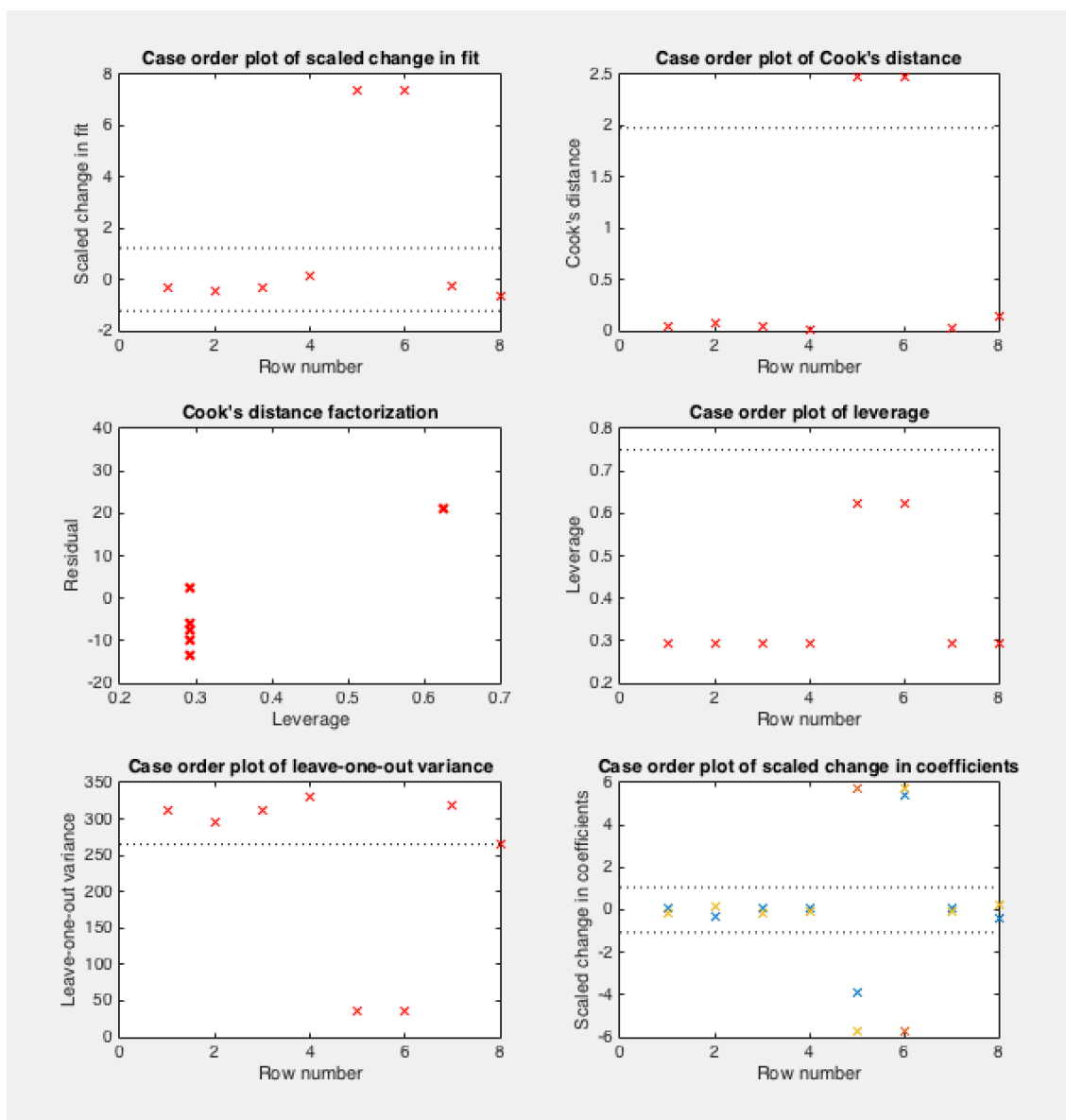
Testing for other model assumptions (model 1)

Figure 57 Diagnostic plots for model development (round 1) involving testing model assumptions

There were a few points with high leverage for this further residual analysis for model 1, but this did not reveal whether the high-leverage points were outliers. A case-order plot of Cook's distance and Cook's distance factorisation helped us identify one point with a large Cook's distance. We would be removing these data, and this is the method with which we could identify outliers against the model assumptions.

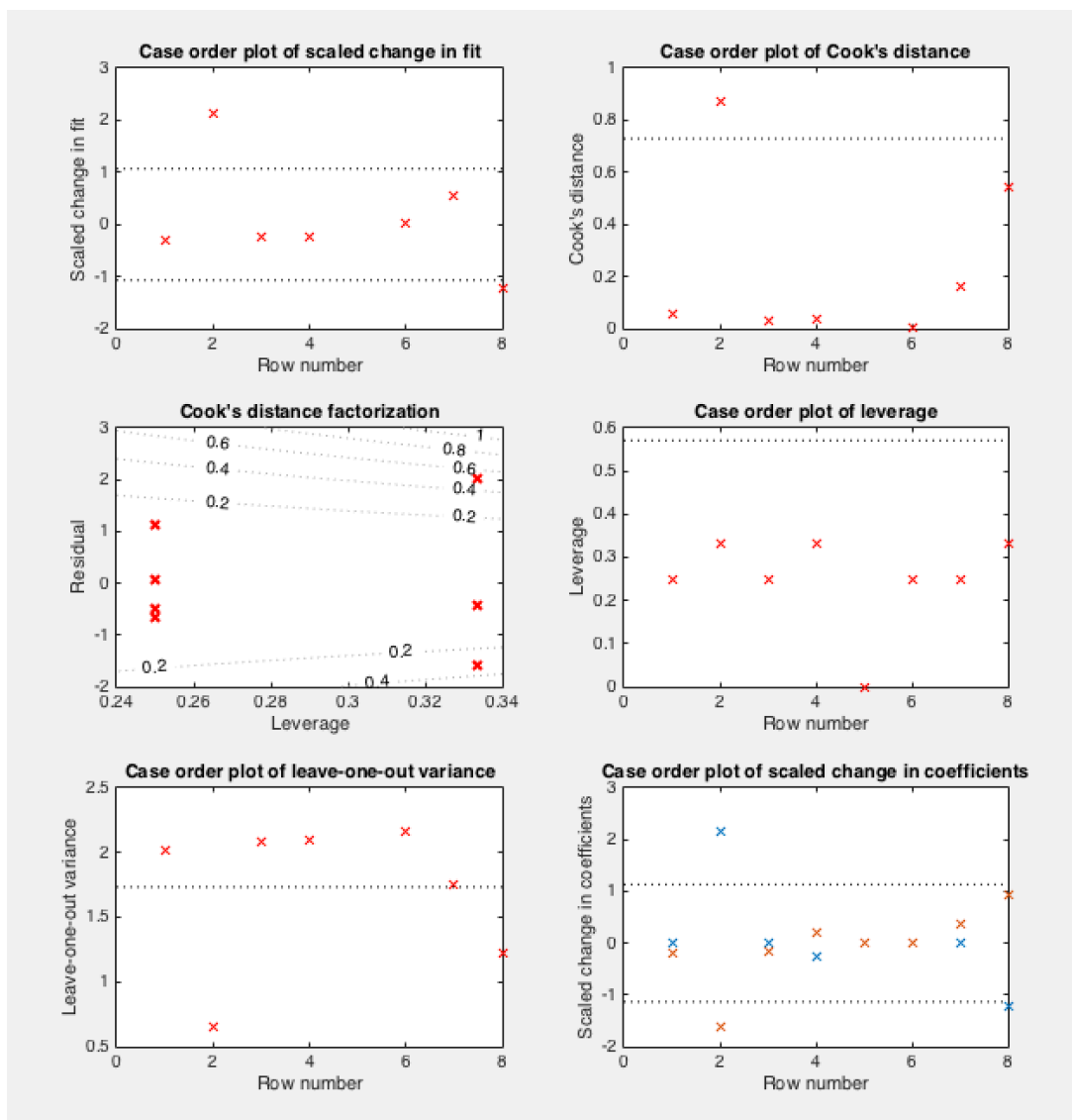## Testing for other model assumptions (model 2)

Figure 58 Diagnostic plots for the model development (round 2) involving testing model assumptions.

Again, as mentioned before, this residual analysis for model 2, the residual estimates tended to overestimate the scaled change in the fit to the model – this is seen at row 2 of the x-axis row number in the graphs above. For model 1 the leverage identified obvious outliers, and this time in model 2, the leverage was more evenly distributed. Of note, the outlier was excluded from the model this time. This showed us a better model fit with the estimate residuals.

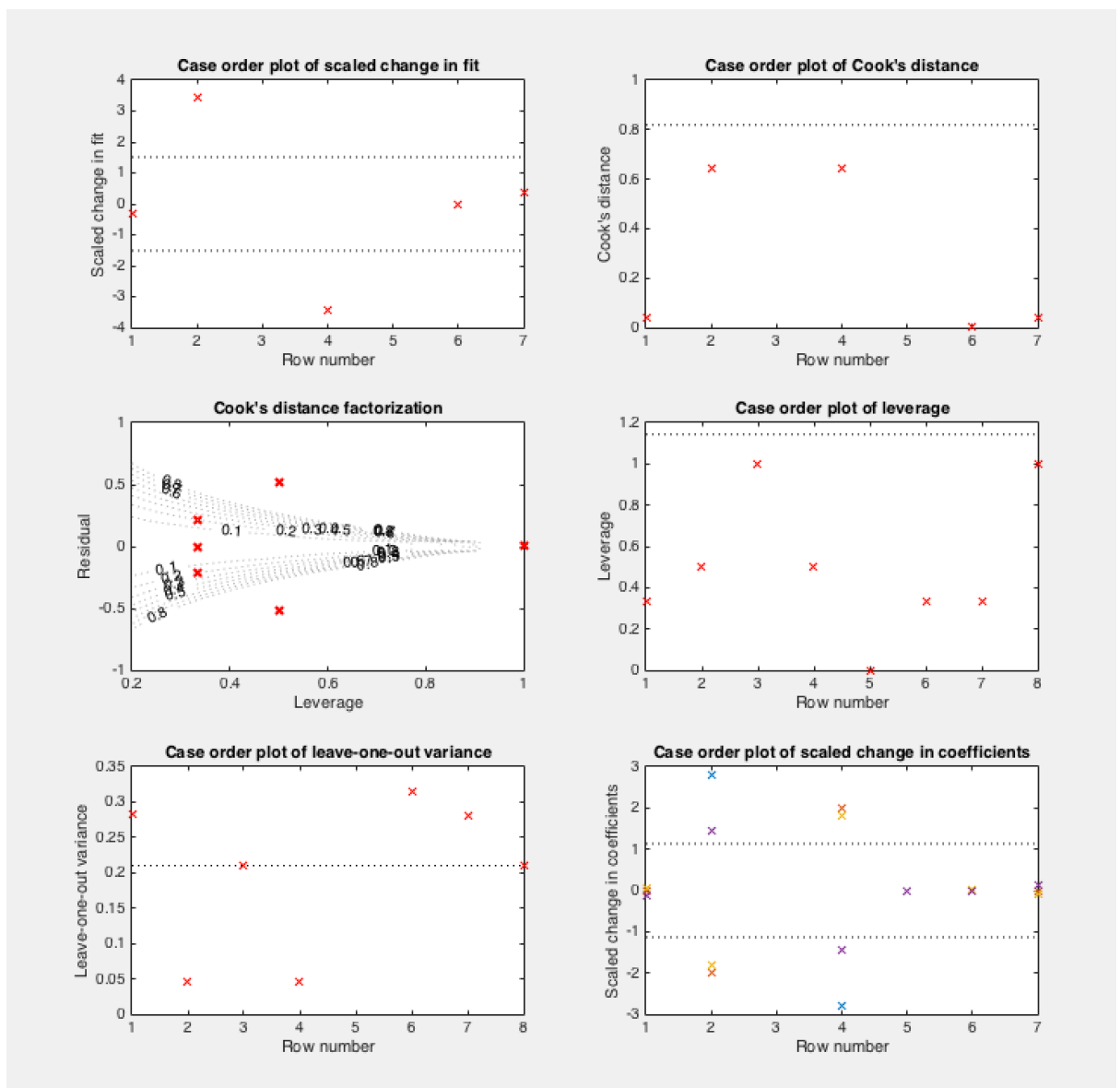## Testing for other model assumptions (model 3)

Figure 59 Diagnostic plots for the model development (round 3) involving testing model assumptions

As mentioned previously, there will always be a tendency of residual estimates overestimating a scaled change in fit to the model, evident at row 2 of the x-axis when the same data were transformed. In the previous model fitting (i.e. model 1), the leverage identified obvious outliers, this time, the leverage was more evenly distributed. Again, data of row 5 were excluded from the model this time, as in the case of model 2. This model 3 showed us the best model fit with the estimate residuals.

8.2.4 Appendix for chapter 5 model 1 supplementary results: Testing the Model Assumptions using Residual Graphs

To examine the assumption that the residuals are approximately normally distributed, are independent, and have equal variances, we generated several plots.
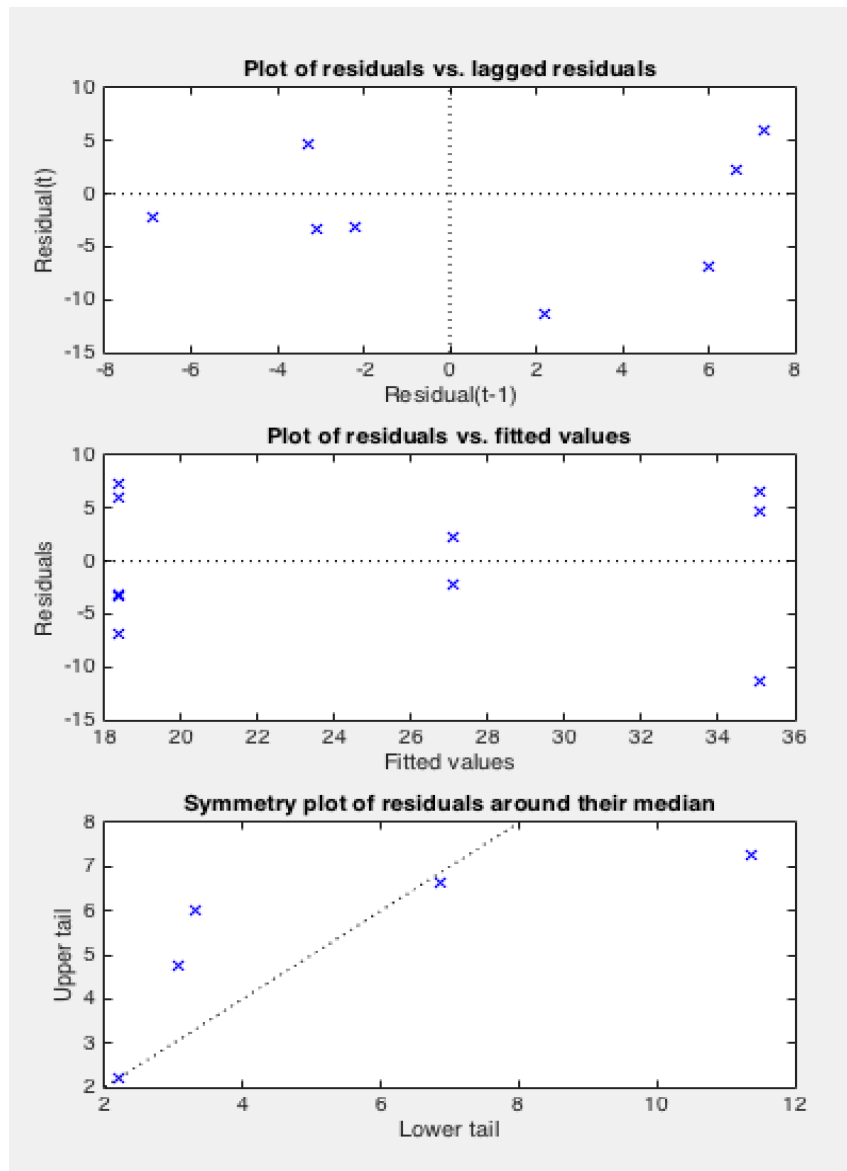

Figure 60 Residual diagnostic plots

Here this residual analysis for the DoE experiments aimed to determine whether there was serial correlation among the residuals. The scatter plot

shows an almost even distribution of residuals in all of the quadrants, indicating neutral or no serial correlation among the residuals. Another potential issue is when residuals are large for large observations. There is some tendency for larger fitted values to have larger residuals. This is probable as shown here in the above fitted vs. residuals diagnostic plot. The symmetry plot was also not linear, so our data maybe non-uniform and non-normally distributed. The two key assumptions for our model fitting were undertaken. The next set of diagnostic plots would aim to check if this was true.
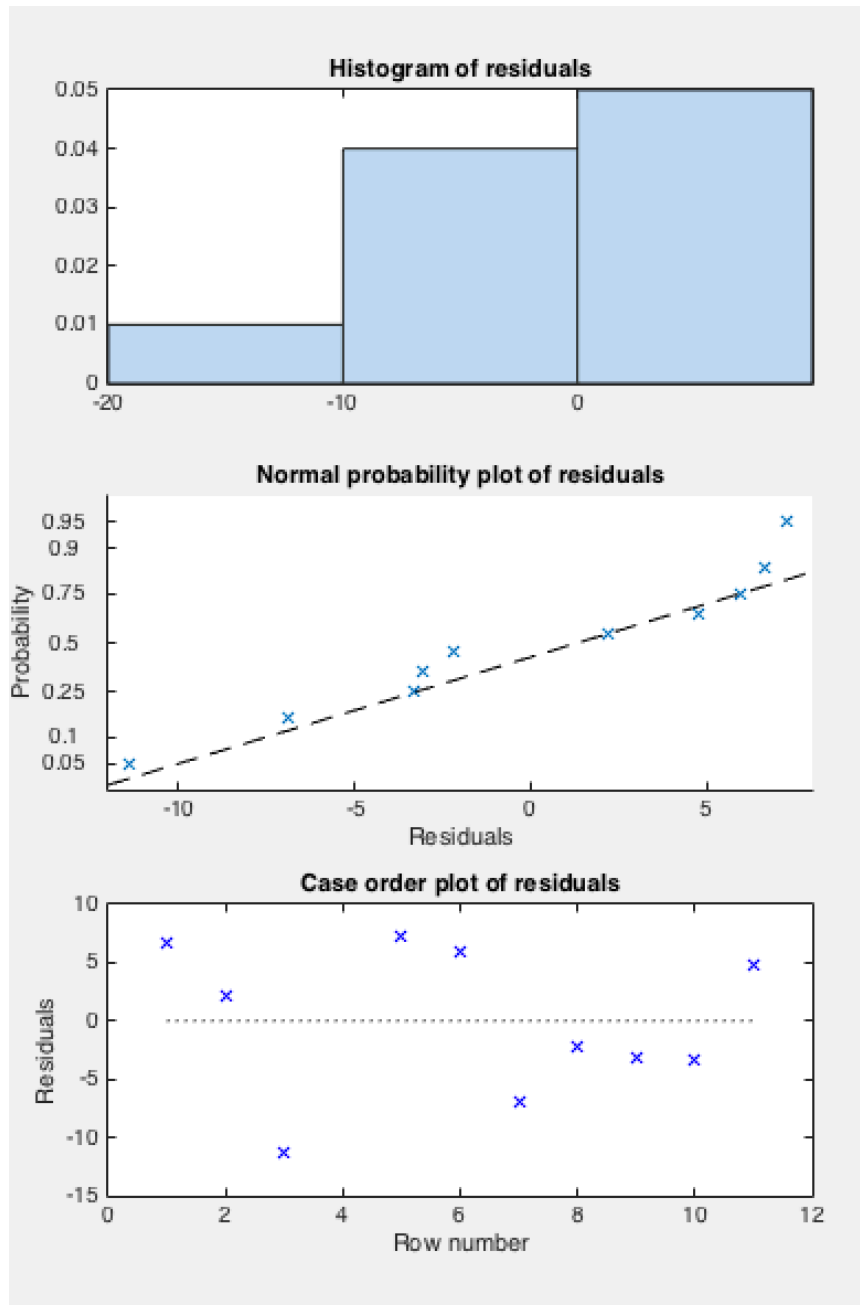
Figure 61 Residual diagnostic plots for DoE optimisation experiment for OD responses.

There might be possible outliers below 10 for these residual analysis. From the normal probability plot, it is difficult to understand the data distribution and its effect on the linearity of the probability; however, it looks skewed below the zero point, implying a probably suboptimal fit to normally distributed residuals. From the histogram of the residuals, the residuals are not normally distributed.
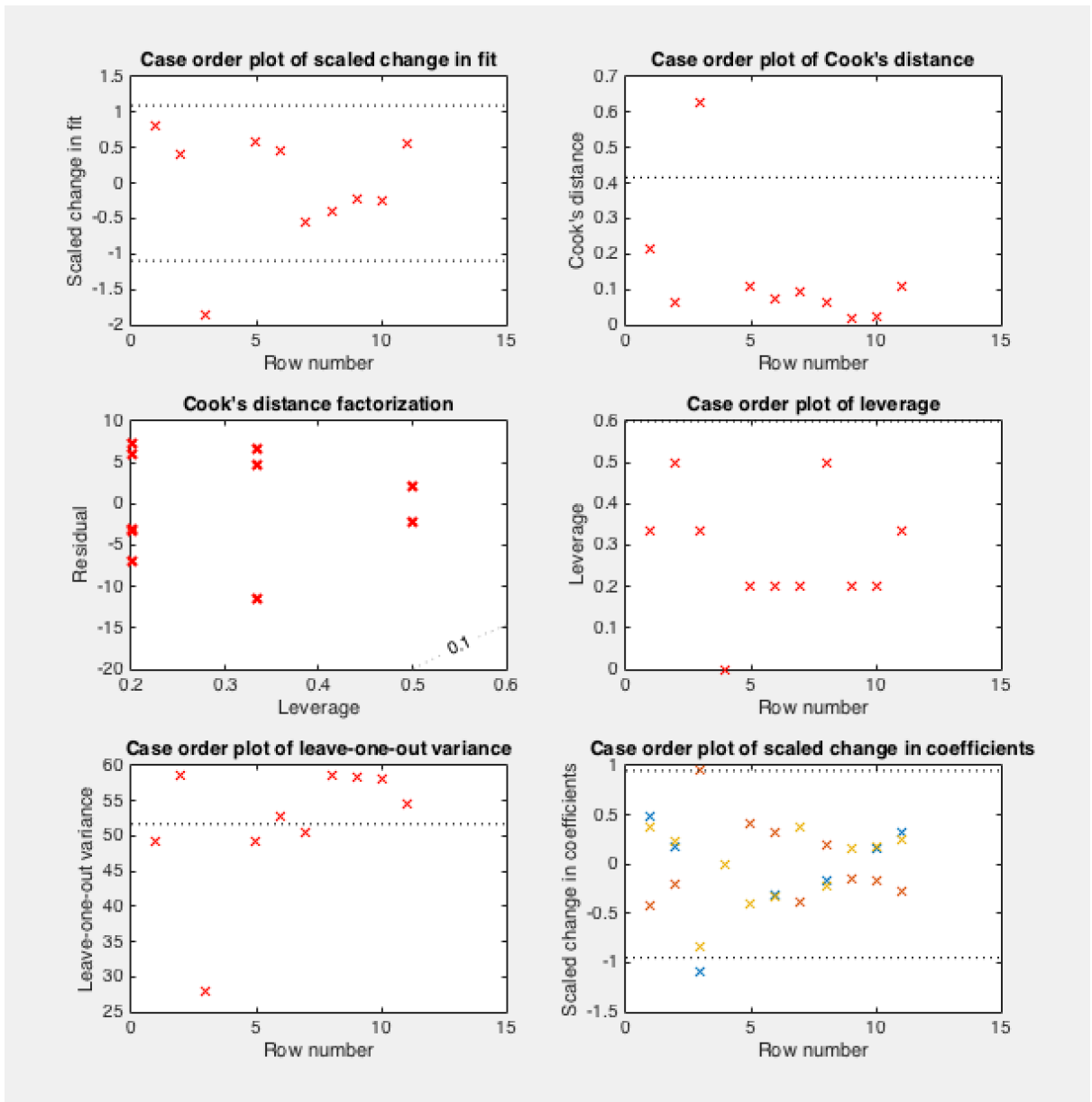
Figure 62 Residual diagnostic plots

There was one point with high leverage.in Figure 62. However, this did not reveal whether the high-leverage points were outliers. A case-order plot of Cook's Distance and Cook's Distance Factorisation helped us identify one point with a large Cook's distance. We removed these data, as there was no obvious pattern to the data during the experiments while in the designated run order.

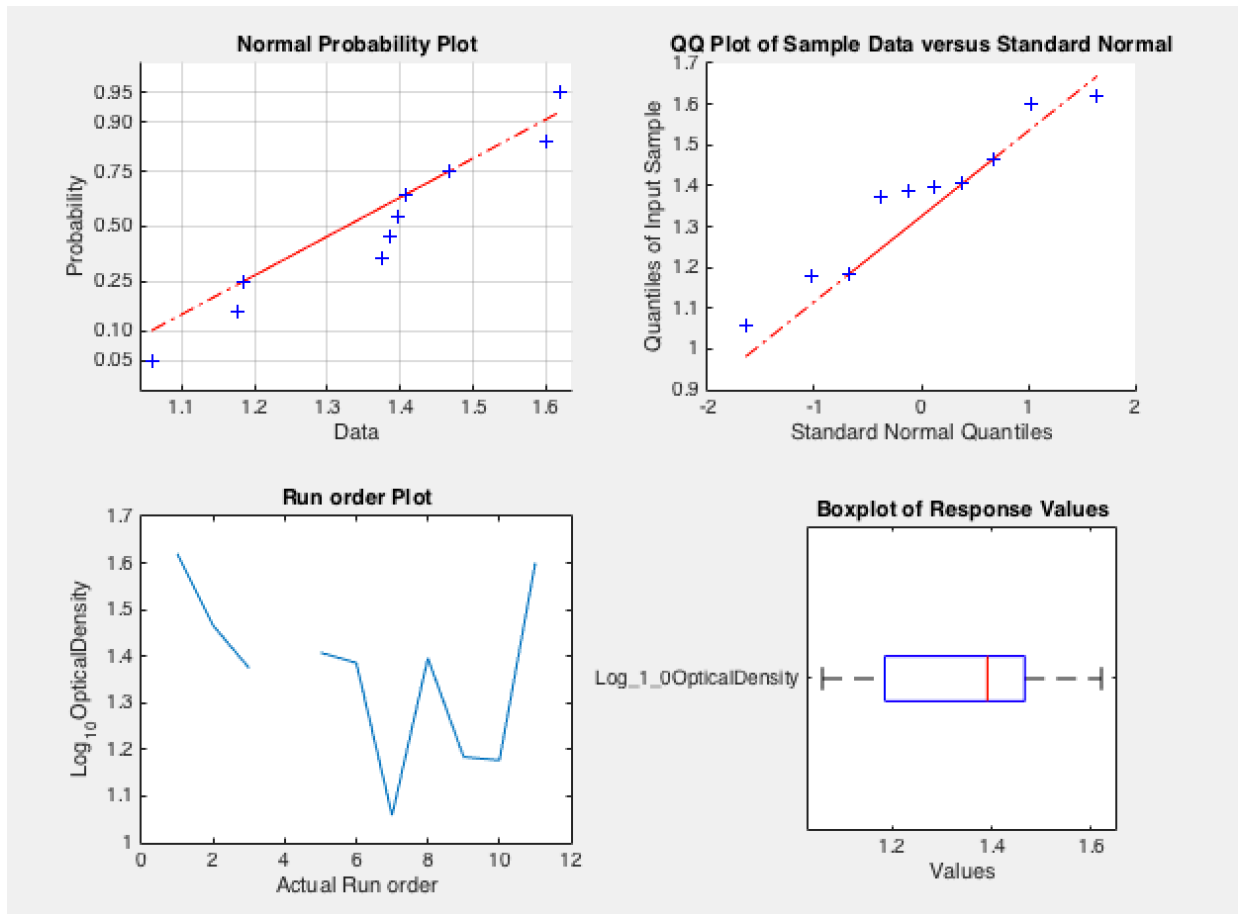## 8.2.5 Appendix for chapter 5 model 2 supplementary results.

Figure 63 Data plots and diagnostics for log transformed data responses for chapter five first response (model 2)

There was good distribution of data and no obvious trends and patterns in the data in the run order that was designed. This was confirmed by the histogram and normal probability plots (not shown for simplicity).

Again, stepwise regression method was used to model the experimental fermentation data. The non-significant factor interactions were eliminated and were left with the same significant regression model terms, specifically including a quadratic term.

Table 27 Linear regression model and formulation

```
Linear regression model:
    Log_1_0OpticalDensity ~ 1 + GlycerolConcentration + GlycerolConcentration^2

Estimated Coefficients:
                               Estimate        SE        tStat      pValue
                              _____   _____   _____   _____

    (Intercept)                  5.2607       1.3315     3.9508    0.0055262
    GlycerolConcentration       -0.09793     0.034371   -2.8492    0.024716
    GlycerolConcentration^2    0.00059633   0.00021673    2.7515    0.028441


Number of observations: 10, Error degrees of freedom: 7
Root Mean Squared Error: 0.136
R-squared: 0.564,  Adjusted R-Squared 0.439
F-statistic vs. constant model: 4.53, p-value = 0.0547
```

Table 28 Confidence level testing using ANOVA statistical methods

	SumSq	DF	MeanSq	F	pValue
Total	0.2955	9	0.032833		
Model	0.16667	2	0.083333	4.5278	0.054721
. Linear	0.027323	1	0.027323	1.4846	0.26253
. Nonlinear	0.13934	1	0.13934	7.571	0.028441
Residual	0.12883	7	0.018405		

Table 27 and Table 28 show here the statistical evidence of our model 2 fit using p values as our degree of confidence in the model fitting as per usual. If a linear model had been chosen to describe our process control, then it would have an even worse-off estimation effect of our model fit than when compared to the previous model fitting. The non-linear, in fact, quadratic interaction term included in the model would result in curvature of the 3D plot and contour plots.

8.2.6 Appendix for chapter 5 model 3 supplementary results.

8.2.6.1 Statistical analysis of model fitting our second variable packed cell volume with experimental results.
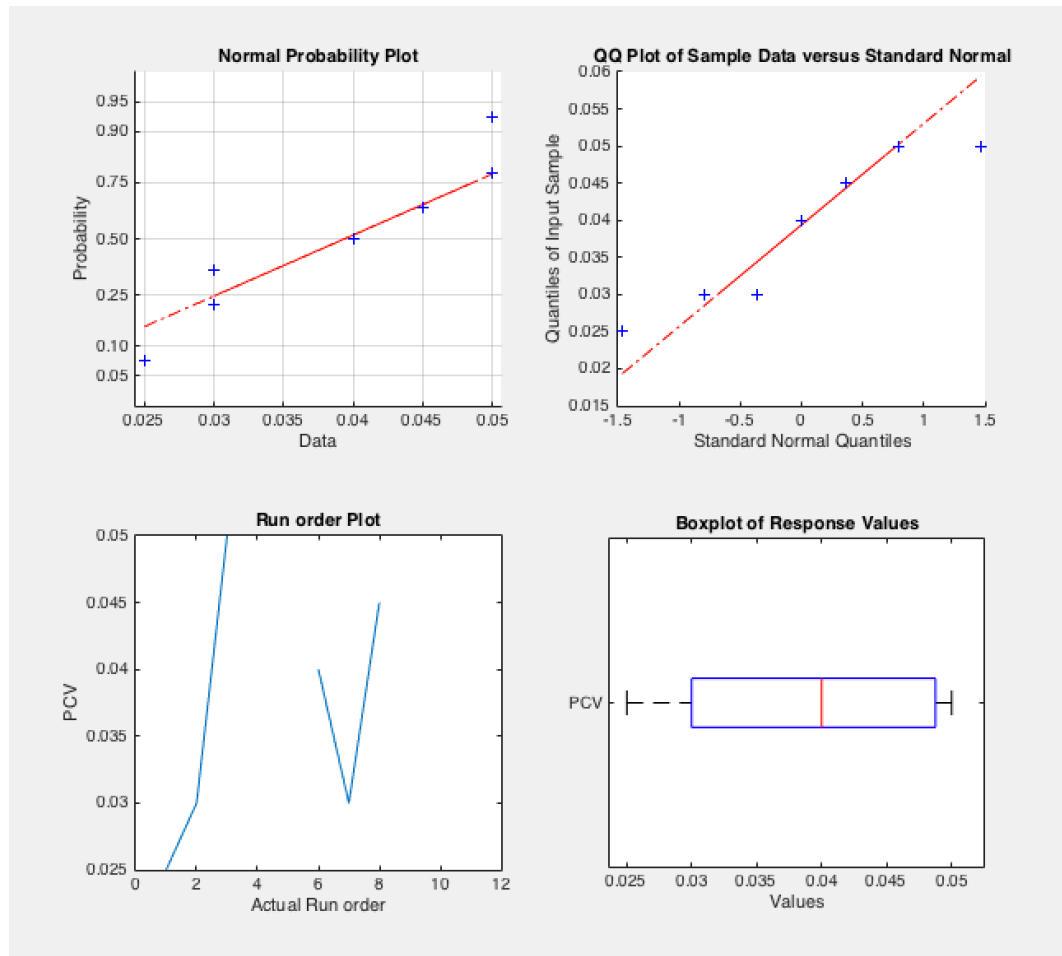

Figure 64 Data plot diagnostics for second response packed cell volume for the optimisation DoE experiment.

Potentially, there was normality within both our Q-Q plot and the normal probability plot as they both display an almost linear fitting. The run order plot clearly confirms very weak data quality as 4 of the data observations were failures (run 4, 5, 9, and 10); also, with the data included in the model, there were no obvious trends. There was a wide range of data which could likely indicate that the data has some distribution problems or lack of fit issues.

Next there was a statistical fit to the initial full model using all types of interactions and main effects from the second response data. All the presented terms and interactions presented were not significant. This could be seen with a poor model fitting, i.e. it was difficult to know the response output to the actual controllable factors. However, while proceeding with the modelling, the steps involved with the stepwise regression was repeated.

8.2.6.2 Test the Model Assumptions using Residual Graphs for the Second Response

To examine the assumption that the residuals are approximately normally distributed, are independent, and have equal variances, we generated several plots.
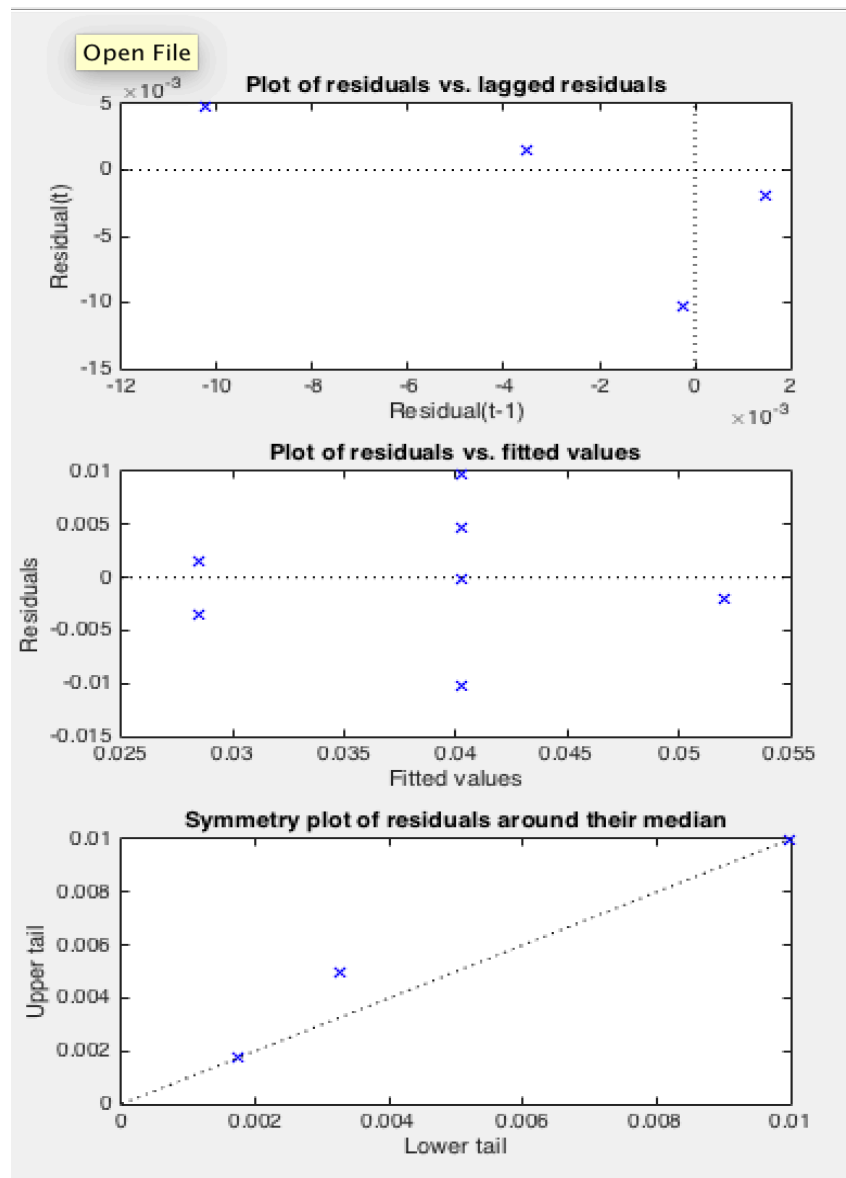

Figure 65 Residual plot diagnostics for second response packed cell volume for optimisation DoE experiment.

This analysis again looks at the serial correlation among the residuals in Figure 65. Here there was serial correlation among the residuals. The symmetry plot is also linear, and so the data was deemed uniform and normally distributed. The two assumptions for this model fitting were undertaken.
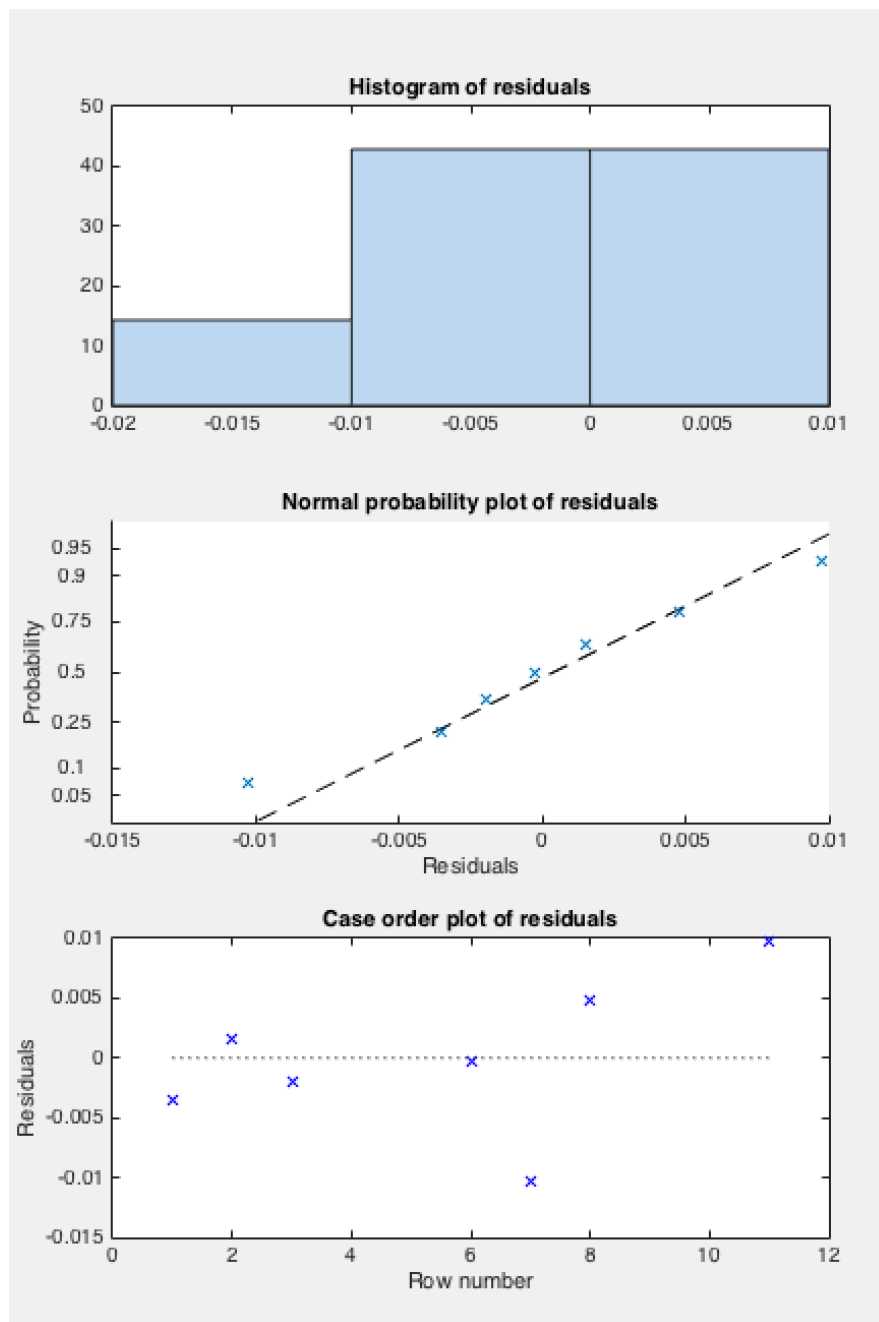

Figure 66 Additional residual diagnostic plots for second response packed cell volume for optimisation DoE

There might be possible outliers below -0.01 for the residuals (see Figure 66). The linearity of the residuals in the normal probability looks good. The underlying assumptions were not violated in these diagnostic plots.
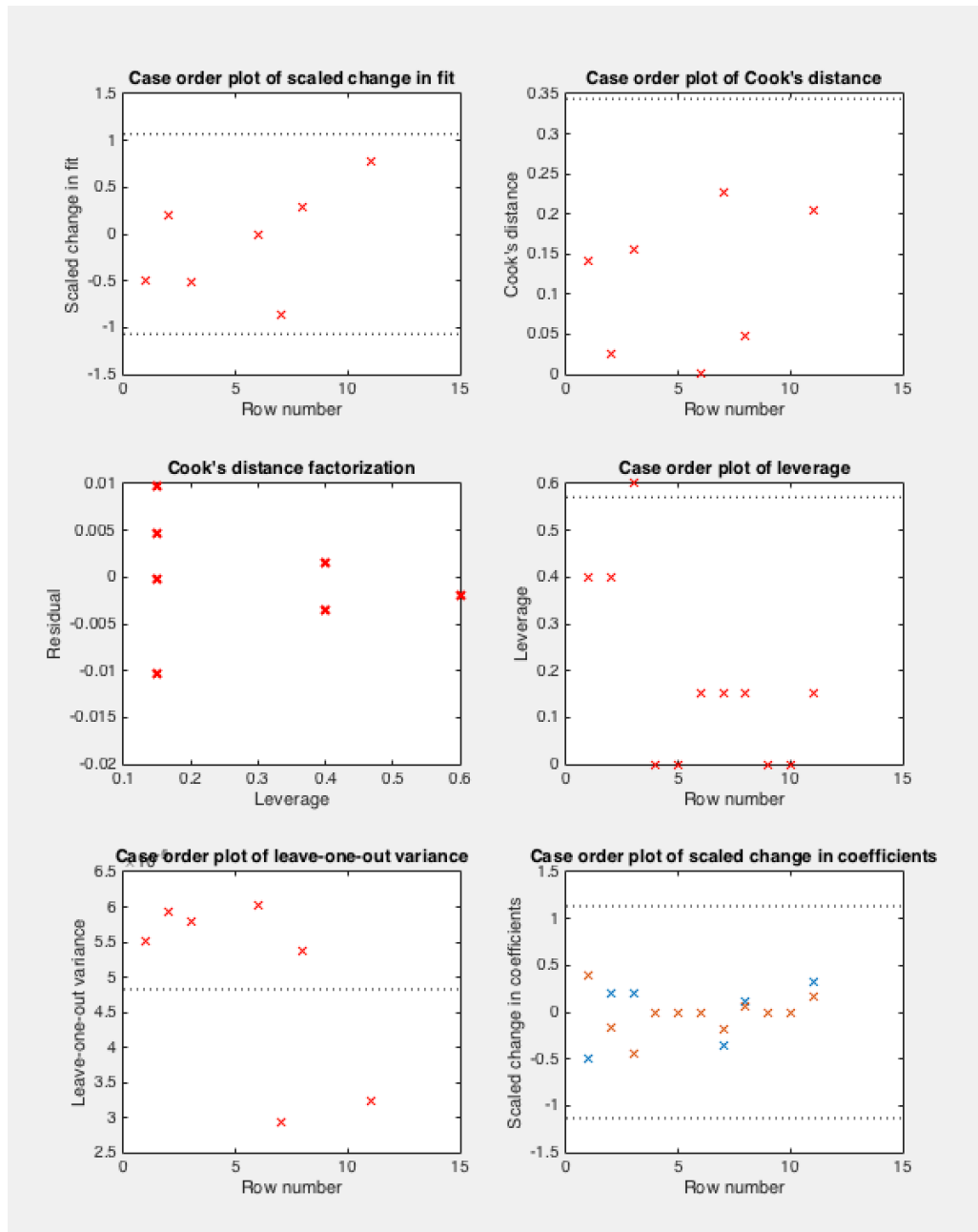

Figure 67 Additional residual diagnostic plots for second response

There was one residual with high leverage (in Figure 67). However, the data in the Cook's Distance did not necessarily reveal any outliers. It was difficult to determine a trend in the residual diagnostic plots above, as it is highly likely that the data was too poor of a quality leading to low precision.

8.3 Appendix for Chapter 6 – example code review, validation and documentation for software development.

<div style="border: 1px solid">

Function Help for EasyDoE

F1,F2,F3,AutoDiagnosis,F4,F5,F6 classes

F1 class ChooseDesign

function name in ChooseDeisgn **F1 Examples**

[OutputHiLo,RecipeDesignList,NameOfFactors] =
ChooseDesign(2,2,4,'GlycerolConcentration',60,100,'Temperature',30,37)
[OutputHiLo,RecipeDesignList,NameOfFactors] =
ChooseDesign(3,3,4,'GlycerolConcentration',60,100,'Temperature',30,37,'CellConcentration',2.5,5)

F1 Input (3+varargin)

NumberOfFactors = NumberOfRuns = DesignResolution =

(pick 1-10)(2, 3 or more) (note: it is 2^K) (3-8)

---

varargin steps (3 at a time) Factorname1 = (e.g.'factor1' or 'temp' or 'feedrate') (note: string) Factorname1Lo = (e.g. 40) (note: double) Factorname1Hi = (e.g. 80) (note: double)

Repeat varargin steps for each factor (must be a minimum of 2+)

F1 output (3 output 'var's)

OutputHiLo = [M by N] double array which are Hi (1) and Lo (-1) coded factor points. where M = length of NumberOfFactors, and N = 2^K where K = Number of Runs.

RecipeDesignList = [M by N] double array which consists of theparameterised recipe set point, equivalent to the Hi and Lo codes above.

NameOfFactors = The output var (MWCellArray) type of the varargin factor name steps.

F2 class known as GetAnova

function name in GetAnovaAndProbability% UPDATE: this has now been fixed for rank Defiency in missing data and/or small sample sets for fractional factorial deisgns.(02/11/2015)

F2 usage

[prob2,tbl,stats,ResponseResults] =
GetAnovaAndProbability(NameOfFactors,RecipeDesignList,ResponseResults) Example [prob2,tbl,stats, ResponseResults] = GetAnovaAndProbability(NameOfFactors,RecipeDesignList,[1,2,3,4]')

F2 Input (3 inputs)

NameOfFactors = output from F1 (note: MWCellArray of Strings)

</div>

RecipeDesignList = smae as F1 outputResponseResults = ie. a Numeric array matching the same order as the randomised

runs.

F2 output (4 output 'var's)

prob2 =

tbl = Stats =

[M by N] (MWCellArray containing String in the first coloumn probability numbers from 5 to '95%' in the second column.) MWCellArray containing strings of the ANOVA results information MWStructure (type) (of the ANOVA results, used primarily for compatibility of future diganostics and Advanced user section.)

ResponseResults = [1 by N] (double array results entered from earlier.)

F3 class known as DiagnosticsF3

function name is diagnosticsAndOutput

F3 usage

[mdlstep,tblAnova4,coefficientsNames1,coefficients4model,modelFormula, mdl] = ...
diagnosticsAndOutput(OutputHiLo, RecipeDesignList,NameOfFactors,logicalVarsToInclude,...
ResponseName,ResponseResults,CatergoricalBoolean,CatergoricalLogicalArray)

Example: for 3 factors with 2 levels (Hi Lo) and one of these are Catorgorical Factor(s) we want to show graphical output for variable 1 and 3 (see logicalVarsToinclude).

[mdlstep,tblAnova4,coefficientsNames1,coefficients4model,modelFormula, mdl] = ...
diagnosticsAndOutput(OutputHiLo, RecipeDesignList,NameOfFactors,[1 0 1],{'OpticalDensity'},...
ResponseResults,true,[0 1 0 0])

Example: for 3 factors with 2 levels (Hi Lo) and NONE of these are Catorgorical Factor(s)we want to show graphically factors 1 against 2 this time.

[mdlstep,tblAnova4,coefficientsNames1,coefficients4model,modelFormula, mdl] = ...
diagnosticsAndOutput(OutputHiLo, RecipeDesignList,NameOfFactors,[1 1 0],{'OpticalDensity'},ResponseResults,'','')

F3 Input (6 - 8 Input 'var's Depending on catorgories)

OutputHiLo = RecipeDesignList = NameOfFactors = LogicalVarsToInclude =

ResponseName = ResponseResults =

Note: CatergoricalBoolean =

CatergoricalLogicalArray=

double array as similar to output from F1same as F1 outputsame as F1 outputEssentially going to be a logical array (M by1) numerical array which is filled with 1 (true:include) or 0 (false:exclude), please see examples.

MWCellArray of (1 by 1) length and must be astring with no trailing or white spaces.MUST be the output of F2 in a [M

by 1]numerical array (this is automatic if going from F2 to F3).

either written as just ~ true or '' dependingon whether to include which logical catorgorical factors in the diagnostics and statistics.Essentially a logical array of [1 by M]numeric array consisting on 1 (true:isCatorgical:include) or 0 (false:notCatorgorical:exclude)BE SURE TO ADD AN EXTRA 0ontop of the number of factors.AT THE END TO MAKE THIS WORK.

F3 output (6 output 'var's and possibly some Graphical pop-ups please be aware of error's)

mdlstep = tblAnova4 =

coefficientsNames1 =

coefficients4model =

modelFormula = mdl =

Matlab LinearModel:Type:Class structure thingy not sure what this will be in c sharp language. New Anova Summary table type array, thismight be a MWCellArray or just a MWArray

type in the .Net Assembly.[1 by N] cell array of strings. Thesestrings are the terms or Matlab quick 'syms' which are used in thenew Anova model (second Anova that we discussed about)that is generated from calling F3 function[M by 1] 'double's numeric array type ofthe coeffecients generated from the model calledfrom the Anova private method of this functionMatlab LinearModel:Type:Class structure thingy again not sure what this will bei n c sharp language. Matlab LinearModel:Type:Class structurethingy but this time its included in this .Net Assembly for futurecompatibility of advanced user section for Diagnostics.

Autodiagnosis class for Screening

function name in AutoDiagnosisAndOutput **AutoDiagnosis Examples**

[mdlstep,tblAnova4,coefficientsNames1,coefficients4model,modelFormula, mdl] = ...
AutoDiagnosisAndOutput(OutputHiLo,
RecipeDesignList,NameOfFactors,logicalVarsToInclude,ResponseName,ResponseResults,Catergorical
Boolean,CatergoricalLogicalArray)

PLEASE REFER TO DiagnosticsAndOutput function Above, (it does exactly the same input and output but has a different mechanism in the function.

AutoDiagnosis Input (8)

F3 Input (6 - 8 Input 'var's OutputHiLo = RecipeDesignList = NameOfFactors = LogicalVarsToInclude =

ResponseName = ResponseResults =

Depending on catorgories)double array as similar to output from F1same as F1 outputsame as F1 outputEssentially going to be a logical array (M by1) numerical array which is filled with 1 (true:include) or 0 (false:exclude), please see examples.MWCellArray of (1 by 1) length and must be astring with no trailing or white spaces.MUST be the output of F2 in a [M by 1]

Note: CatergoricalBoolean =

CatergoricalLogicalArray=

either written as just ~ true or '' dependingon whether to include which logical catorgorical factors in the diagnostics and statistics.Essentially a logical array of [1 by M]numeric array consisting on 1 (true:isCatorgical:include) or 0 (false:notCatorgorical:exclude)BE SURE TO ADD AN EXTRA 0ontop of the number of factors.AT THE END TO MAKE THIS WORK.

numerical array (this is automatic if going from F2 to F3).

AutoDiagnosis output (6 output 'var's and possibly some Graphical pop-ups please be aware of error's)

---

mdlstep = tblAnova4 =

coefficientsNames1 =

coefficients4model =

modelFormula = mdl =

F4 class

Matlab LinearModel:Type:Class structure thingy not sure what this will be in c sharp language. New Anova Summary table type array, thismight be a MWCellArray or just a MWArray

type in the .Net Assembly.[1 by N] cell array of strings. Thesestrings are the terms or Matlab quick 'syms' which are used in thenew Anova model (second Anova that we discussed about)that is generated from calling F3 function[M by 1] 'double's numeric array type ofthe coeffecients generated from the model calledfrom the Anova private method of this functionMatlab LinearModel:Type:Class structure thingy again not sure what this will bei n c sharp language. Matlab LinearModel:Type:Class structurethingy but this time its included in this .Net Assembly for futurecompatibility of advanced user section for Diagnostics.

---

function name in ChooseOptimisationDesign **F4 Examples**

[NameOfCCDFactorsConcat, CodedDesignListBlockRandomised, CCDRecipeDesignBlockRandomised] = ChooseOptimisationDesign ... (NumberOfCCDFactors,CenterPointsNo,CCDtype,blocksize, FactorNameCellArray, FactorLowCellArray, FactorHighCellArray)

l[NameOfCCDFactorsConcat, CodedDesignListBlockRandomised, CCDRecipeDesignBlockRandomised] = F4Coded (3,6,'faced',8, {'factor1','factor2','factor3'},[9,19,29],[11,21,31]l

---

F4 Input(7)

NumberOfCCDFactors = CenterPointsNo =

CCDtype =

Blocksize = FactorCellArray = FactorLowCellArray =

FactorHighCellArray =

F4 output (3)

NameOfCCDFactorsConcat =

This is the This is the the design This is the 'inscribed'

Number of Factors to be looked at in the designthe number of center points that need to be added to

Design type ('circumscribed','faced' or

This is a number of runs (number of batches) expected to be within each block Normally 4 or 8MWCellArray of string's with no trailing or white spacesof the factor names.

Numerical array of "Lo" values for the design space. Thismust be the same number of factors chosen in

NumberOfCCDFactors

Numerical array of "Hi" values for the design space. Thismust be the same number of factors chosen in NumberOfCCDFactors

These are the Names of the Optimisation Design Factors with Factor 1, Factor 2, ...Factor n, 'Blocks' 'Original Run Order'

CodedDesignListBlockRandomised =

CCDRecipeDesignBlockRandomised =

F5 class

Is the coded design list of hi and lows. (then which block it is in and then The original run order.

Is the parameter (recipe set points) design list of hi and lows. (then which block it is in and then The original run order.

---

function name is OptimisationDiagnosticsAndOutput **F5 Examples**

[mdlstep,tblAnova4,coefficientsNames1,coefficients4model,modelFormula, mdl] = ...
OptimisationDiagnosticsAndOutput(CodedDesignListBlockRandomised,
CCDRecipeDesignBlockRandomised, NameOfCCDFactorsConcat,[1 1 0 0
0],{'OpticalDensity'},[1,1,2,3,4,5,1,2,3,4,5,1,2,4,3,5,5,5,5,5]','','')

F5 Input (8)

CodedDesignListBlockRandomised = CCDRecipeDesignBlockRandomised = NameOfCCDFactorsConcat =

LogicalVarsToInclude =

ResponseName = ResponseResults =

This is the output from F4, (example; -1.64,-1,0,1,1.64)This is the output from F4, (not coded but recipe like properties).This is the output from F4, (this is a cell basedarray of strings containing the factor names, and the block and original run order. Essentially going to be a logical array (M by1) numerical array which is filled with 1 (true:include)or 0 (false:exclude), please see examples.MWCellArray of (1 by 1) length and must be astring with no trailing or white spaces.[M by 1]numerical array of in a downward 0 dimension = M length direction.

Note:

CatergoricalBoolean = CatergoricalLogicalArray=

either written as just ~ true or '' dependingon whether to include which logical catorgorical factors in the diagnostics and statistics.

Essentially a logical array of [1 by M]numeric array consisting on 1 (true:isCatorgical:include) or 0 (false:notCatorgorical:exclude)

F5 output (6 output 'var's and possibly some Graphical pop-ups please be aware of error's)

---

mdlstep = tblAnova4 =

coefficientsNames1 =

coefficients4model =

modelFormula = mdl =

F6 class

Matlab LinearModel:Type:Class structure thingy not sure what this will be in c sharp language.

New Anova Summary table type array, this might be a MWCellArray or just a MWArray type in the .Net Assembly. Im not entirely sure.

[1 by N] cell array of strings. Thesestrings are the terms or Matlab quick 'syms' which are used in the new Anova model (second Anova that we discussed about)that is generated from calling F3 function

[M by 1] 'double's numeric array type ofthe coeffecients generated from the model called from the Anova private method of this function

Matlab LinearModel:Type:Class structure thingy again not sure what this will bei n c sharp language.

Matlab LinearModel:Type:Class structurethingy but this time its included in this .Net Assembly for future compatibility of advanced user section for Diagnostics.

function name in DualOptimisation **F6 Examples**

DualOptimisation(CCDRecipeDesignBlockRandomised, NameOfCCDFactorsConcat,[1 1 0 0 0],
{'OpticalDensity'}, [1,1,2,3,4,5,1,2,3,4,5,1,2,4,3,3,6,6,6,6]','true',2.5, {'OpticalDensity2'},
[1,4,3,2,1,1,5,4,3,2,1,5,4,2,1,3,3,3,3,4]','true',3) DualOptimisation(CCDRecipeDesignBlockRandomised,
NameOfCCDFactorsConcat,[1 1 0 0 0],
{'OpticalDensity'},[1,1,2,3,4,5,1,2,3,4,5,1,2,4,3,3,6,6,6,6]','false',2,
{'OpticalDensity2'},[1,4,3,2,1,1,5,4,3,2,1,5,4,2,1,3,3,3,3,4]','false','')
DualOptimisation(CCDRecipeDesignBlockRandomised, NameOfCCDFactorsConcat,[1 1 0 0 0],
{'OpticalDensity'},[1,1,2,3,4,5,1,2,3,4,5,1,2,4,3,3,6,6,6,6]','true','',
{'OpticalDensity2'},[1,4,3,2,1,1,5,4,3,2,1,5,4,2,1,3,3,3,3,4]','false',2)
DualOptimisation(CCDRecipeDesignBlockRandomised, NameOfCCDFactorsConcat,[1 1 0 0 0],
{'OpticalDensity'},[1,1,2,3,4,5,1,2,3,4,5,1,2,4,3,3,6,6,6,6]','false','',
{'OpticalDensity2'},[1,4,3,2,1,1,5,4,3,2,1,5,4,2,1,3,3,3,3,4]','true','')

F6 Input (11)

CCDRecipeDesignBlockRandomised =

This is the output from F4, (not coded but recipe like properties).

This is the output from F4, (this is a cell basedarray of strings containing the factor names, and the block and original run order.

Essentially going to be a logical array (M by1) numerical array which is filled with 1 (true:include)

NameOfCCDFactorsConcat LogicalVarsToInclude =

ResponseName = ResponseResults = R1maxBoolean =

R1Target =

ResponseName2 = ResponseResults2 = R2maxBoolean =

R2Target

0 (false:exclude), please see examples. string with no trailing or white spaces.

[M by 1]numerical array of in a downward 0 dimension = M length direction.

Must be either 'true' or 'false' to help the user find the target (+see below) or mean data criteria (default if target value is not entered by the user) for optimisation

Must be either a specified double type numeric or '' blank if blank then the mean value of the experiemental data will be used. if target value is entered then this will be used for optimisation criteria.

MWCellArray of (1 by 1) length and must be a string with no trailing or white spaces.

[M by 1]numerical array of in a downward 0 dimension = M length direction.

Must be either 'true' or 'false' to help the user find the target (+see below) or mean data criteria (default if target value is not entered by the user) for optimisation

Must be either a specified double type numeric or '' blank if blank then the mean value of the experiemental data will be used. if target value is entered then this will be used for optimisation criteria.

orMWCellArray of (1 by 1) length and must be a

F6Optimisation output (0) just graphical data. %intentionally left blank

Acknowledgements and Notes Thankyou for reading F3 is a little buggy.This manual is a temporary manual for Alex Stucker, INFORS HT. Signed and written by *Aarron Erbas*, UCL.Time of Writing this report manual 20/08/2015.Time of Updating(1) report 02/11/2015