Discriminating Significant From Insignificant Model Parameters: The Case of a Dynamic CHO Cell Model

A dissertation
submitted by

Hana Sheikh

In partial fulfillment of the requirements
for the degree of
Master's of Science
in
Chemical Engineering

TUFTS UNIVERSITY
February 2013

Advisers:
Kyongbum Lee
Christos Georgakis

Abstract

Kinetic modeling for metabolic networks, formulated as a set of ordinary differential equations for intracellular species concentrations, provides the ability to simulate the dynamic behavior of cellular metabolism. Such models aim to predict cellular response to various external stimuli, allowing an investigator to develop a detailed fundamental understanding of the phenomena studied. Investigators often choose to include more than the necessary model details rather than risk the error of including less than necessary details. However, this increases the number of parameters that need to be identified from experimental data and introduces a substantial challenge in the identification of the important model parameters. As more details are added to the model, the increased number of parameters implies the necessity of an increased amount of experimental data. However, more experimental data does not imply that the values of the insignificant parameters can be easily determined. Most importantly, it is not clear whether all the model details and the corresponding parameters are necessary for a desired set of model predictions. The present paper presents a computationally efficient methodology to identify the model parameters that are highly significant for the model predictions and thus distinguish them from the insignificant ones.

The proposed approach is inspired by the classic Design of Experiments (DOE) techniques, performed *in silico* using a preliminary model. We start by defining the possible ranges of each of the unknown model parameters, design a set of *in-silico* experiments or, equivalently, a set of selected calculations that are simulated through the preliminary model. Utilizing analysis of variance (ANOVA) and response surface model (RSM) tools we develop a simplified nonlinear meta-model in which only the significant parameters are retained. We applied this methodology to a dynamic model of Chinese hamster ovary (CHO) cell metabolism (Nolan, 2011). This model, comprising 51 parameters and 34 reaction fluxes, was able to provide a reliable preliminary prediction of the effects

of fed-batch process variables such as temperature shift, specific productivity, and nutrient concentrations. A D-optimal design of experiments was used to sample the parameters across their ranges, and a RSM was obtained with antibody flux as the output. Investigating linear, linearly interactive, and quadratic RSMs, we efficiently eliminated approximately 90% of the terms as being not highly significant, shedding light on the importance of each of the 51 original model parameters in the predictions of the metabolic model.

Through this parameter significance methodology, we were able to discriminate the highly significant parameters from the highly insignificant parameters. We demonstrate the utility of parameter significance discrimination as applied to parameter estimation. Refitting the 6 highly significant terms yields a 55% improvement in the objective function from the original model fitting, as compared to refitting the 12 highly insignificant terms which results in just a 6% improvement in the objective function.

Table of Contents

1. Introduction

1.1 Biologics

The use of biologics has increased across different therapeutic areas, and is poised to fuel pharmaceutical revenues and stimulate growth in the biopharmaceutical market. Traditionally, biologics — very target specific and high efficacy drugs —have been attributed with high costs. However, these drugs are in high demand for the treatment of diabetes, cancer and rheumatoid arthritis, prompting investment in biologics (Rossi, 2010). Over the next five years, the biologics industry is anticipated to grow at a yearly rate of 10.5%.

Most therapeutic recombinant proteins are produced in immortalized mammalian cell lines. The most commonly used line of immortalized mammalian cell lines is the Chinese hamster ovary (CHO). Mammalian cells are preferred for therapeutic proteins because they are able to perform all of the post-translational modifications of native human proteins. CHO cells for therapeutic production are most commonly grown in fed-batch cultures (Kim, 2012).

CHO cells are cultured in a nutrient rich medium which contains the necessary components for healthy cell growth and recombinant protein production. The culture medium includes glucose, amino acids, vitamins, minerals, and salts, as well as other components to aid cell viability. For CHO cells, the major substrates required from the media are glucose and amino acids. As a consequence of the rapid glucose uptake which is characteristic of CHO cells, a large fraction of the glucose provided is converted to lactate (Antoniewicz, 2011). Lactate is a toxin that inhibits cell growth and biologics production. An additional potentially toxic byproduct of cell growth is ammonia, from the catabolism of amino acids. Cell culture growth can be optimized in order to reduce the production of the toxic byproducts and extend cell culture growth, and therefore increase biologics production.

1.2 Metabolic Models

Models of cellular metabolism can provide insight into the mechanism behind a process, as well as provide insight into the results of manipulations to a cellular system.

Modeling biological systems through mathematical models has many applications in systems biology and process development. Mathematical models use knowledge of metabolite networks, enzymes and cellular subsystems to analyze and visualize the complex connections of cellular processes. Metabolic flux analysis and flux balance analysis are two commonly used techniques for cellular pathway characterization. Both of these techniques are based on pseudo-steady state mass balances, and knowledge of extracellular metabolite uptake rates, and can be used to determine information about intracellular flux distribution. These techniques can be used to provide an insight into flux distribution under specific environmental techniques. However, models based off of these techniques cannot take into account dynamic cellular responses, and therefore cannot accurately predict the cellular responses to process perturbations (Schilling, 2001).

This deficiency in modeling dynamic cellular response has been partially addressed with the advent of hybrid models. These hybrid models aim to approximate whole-cell metabolic dynamics by utilizing both steady state flux analysis and kinetic rate expressions. This is achieved by segmenting the culture into multiple time-steps. By assuming pseudo-steady rate at each time step, the dynamic nature of the metabolic change can be captured (Schuster, 2000).

Metabolic Flux Analysis

Metabolic flux analysis (MFA) quantifies a set of fluxes through a known metabolic reaction network, or pathways. A pathway is defined as a set of input and outputs connected by stoichiometrically feasible reactions. The rate at which the reactions take place is the pathway

flux. In order to apply MFA to a cellular system, a stoichiometric model needs to be defined for all intracellular reactions. By applying mass balances around the intracellular metabolites and calculating reaction fluxes, a metabolic flux map can be generated. Metabolic reactions generally occur at a much faster rate than cellular regulatory events. Therefore, in the analysis of metabolic networks is often assumed that pathway metabolites are at a pseudo-steady state. MFA can also be used to identify branch points and alternative pathways, calculation of non-measured fluxes, and calculation of maximum theoretical yields (Phair, 2012).

FBA

Flux balance analysis (FBA) is another approach to understanding cellular networks. It involves a linear programming approach, using the stoichiometric coefficients of each reaction as the set of constraints for the optimization. Similar to MFA, FBA assumes that the system is at steady-state, implying that the concentration of a compound in the metabolic network is constant (Schilling, 2001).

1.3 Kinetic Model Motivation

Kinetic models consider the cellular metabolic activity with respect to time. Michaelis-Menten kinetics are the most common way of representing saturation -enzyme kinetics. The model takes the form:

$$v = \frac{d[P]}{dt} = \frac{\mathrm{Vmax[S]}}{Km+[S]} \quad (1)$$

Equation 1: Format of Michaelis-Menten kinetics

The V_{max} represents the maximum rate achieved by the system at the saturating substrate concentrations. The Km is the substrate concentration at which the reaction rate is half of V_{max}. Biochemical reactions are often approximated with Michaelis-Menton kinetics (Chen, 2010).

Kinetic models play an important role in biological investigation. There is much more information in a set of dynamic biological data than can be understood simply by inspection. Building dynamic kinetic models from these data sets furthers the understanding of the underlying biological mechanisms. Furthering the development of kinetic modeling, dynamic experiments reveal information about the inherently complex biological control systems (Phair, 2012).

For a model to be useful from a process engineering standpoint it needs to be able to capture the long term dynamics of a cell culture. Additionally, it is helpful for the model to have a mechanistic foundation which correlates to biochemical pathways, which will allow an engineer to determine the impact of manipulations in process variables on cell growth and biologics production.

Fed-Batch

Fed-batch is an extended batch process in which a nutrient substrate is fed to a culture to sustain growth and production over an extended period, ranging from multiple days to a few weeks. Fed-batch cultures are often used to reach a high cell density for therapeutic protein production. The composition of the feed solution and its schedule of addition strongly affect the growth rate of the culture, and thus need to be optimized for scale-up (Wlaschin, 2006).

Optimization of fed-batch processes involves large sets of process and kinetic variables. Therefore, there is much interest in modeling fed-batch culture dynamics to provide a cost- and time-efficient approach for process optimization.

Fed-batch cultures have been used for decades by the bio-pharmaceutical industry. Despite this, many important cellular metabolic events still remain poorly understood. In the context of fed-batch culture, there are many process perturbations that are used to influence cell growth, although they may not be well understood. The metabolic shift, which results from a reduction in reactor temperature, is used to slow lactate and ammonia production; however, the mechanisms driving this shift are not fully understood and therefore the metabolic shift is often under optimized. Media formulation and feed schedules are also often under-optimized, as there are many unaccounted for metabolic pathways that cannot be considered. Therefore, process optimization is generally limited to the statistical methods of design of experiments (DoE) or high throughput screening (HTS), which result in a minimal gain in mechanistic knowledge (Legmann, 2009).

2 Motivation

2.1 The case of the Dynamic CHO Cell

In this work, we explore a dynamic CHO cell model of a fed batch culture (Nolan), programmed in MATLAB (Mathworks, Natick, MA). This model was derived from a genome-scale model, and is fairly comprehensive in terms of breadth of pathways covered by the model (Nolan et al, 2010b). The model explicitly accounts for redox- and temperature-dependent changes to pathway activities, and directly calculates the measured variables, i.e. metabolite concentration time profiles in the reactor, by defining rate expressions based on extracellular metabolites {Nolan and Lee, 2011).

The form of kinetics utilized in this model is based on a generalized form of Michaelis-Menten kinetics, called convenience kinetics. This generalized form has the flexibility to account for enzyme regulation by activators and inhibitors. In order to facilitate direct calculation of the measureable quantities, the reaction rates were expressed in terms of extracellular metabolite concentrations. The rate expressions were formulated on the basis of dominant extracellular metabolites, which helps foster the appropriate substrate dependences of reactions.

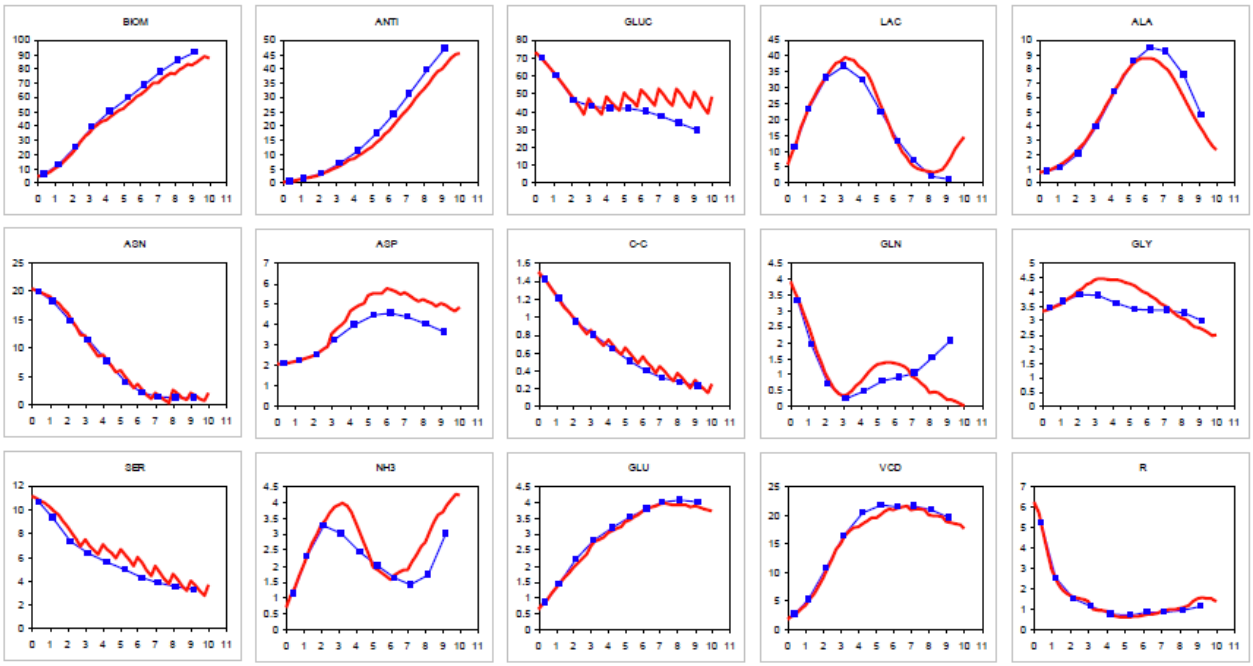
In addition to substrate dependencies, this model also takes into account the impact of the cellular redox state. The redox state-dependent variables allows the model to investigate the effects of potential mismatches between catabolism and oxidative phosphorylation on the long-term dynamics of the culture without the requirement of separate sets of equations for different phases of growth

Not all kinetic parameters needed for the rate expressions in the CHO cell model were available in the published literature. Moreover, many of the available parameters were context dependent.

Therefore, they were derived from the metabolite concentration data. Assuming that the steady state flux of a reaction during balanced growth is an appropriate approximation of an enzyme's activity at near saturation levels, the V_{max} can be calculated. The V_{max} values, as calculated from experimental data, showed general agreement with reported flux and V_{max} values, which indicated that the parameter values obtained are biologically meaningful.

The model is formulated in a structure that the equations and structure could be applied to other types of mammalian cells and recombinant proteins of interested. However, this would require re-calculation of the kinetic parameters used. The model was fitted for two CHO cell lines, and it can be observed that the cell lines have very different kinetic parameters and metabolic profiles, despite both being CHO cells for production of the same antibody (Figure 1). However, both of these cell lines react similarly to process perturbations, demonstrating that the physiological trends are similar for both cell lines. A significant advancement of this model it its ability to accurately predict the timing and magnitude of the metabolic shift, induced by the temperature change.

Cell Line A



Cell Line B

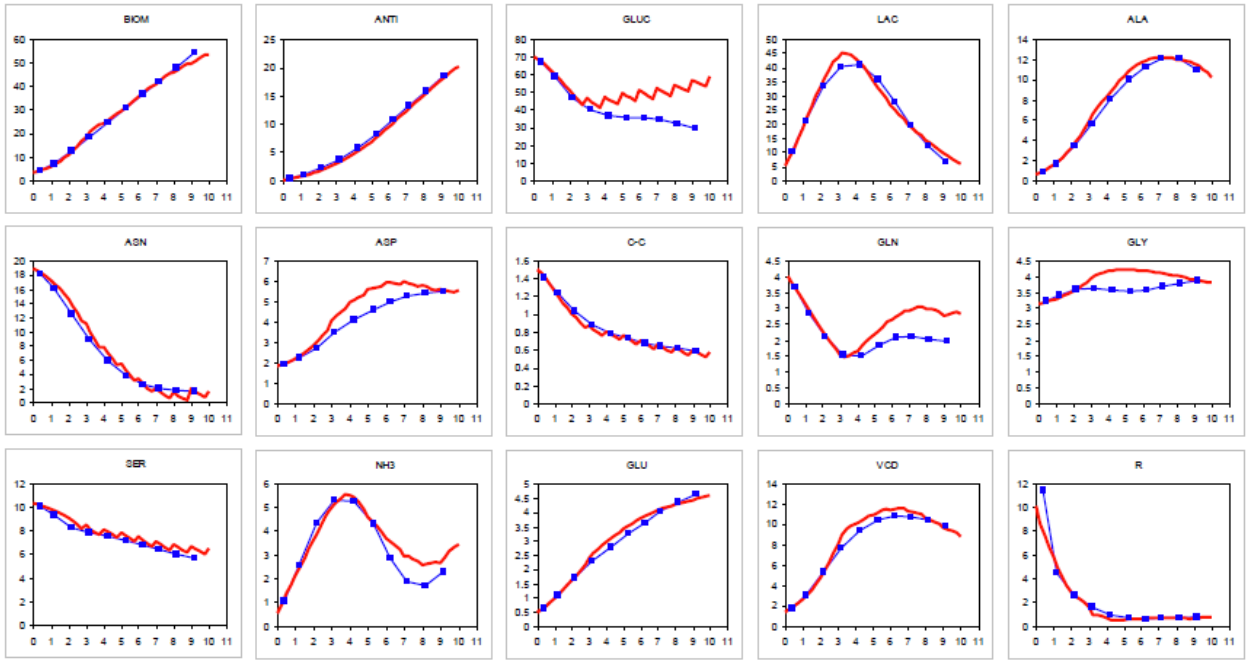


Figure 1: Cell Lines A and B. Their measured profiles are displayed in blue, the model simulation results are in red. The x-axis represents fed-batch culture day, and the y-axis is concentration in mM (Nolan, 2010)

The aim of this model was to not only be able to capture the dynamics metabolic concentration profiles of a fed-batch culture, but to also be able to use this model to investigate optimization of genetic and enzyme activity for increased production of antibody production. A dynamic model holds the advantage of being able to simulate the short term effects of modulations of enzyme activity, particularly in the context of process conditions.

Structure of the CHO Cell Model

The dynamic model of interest describes a CHO-K1 cell line producing a monoclonal antibody. The basis of the model used a network of reactions for a hamster analog mouse, collected from the KEGG database, as a starting point (Figure 2). This network started at approximately 200 reactions and 200 metabolites, and was subsequently grouped and reduced substantially. Care was taken to preserve stoichiometry in the model reduction. The final network model consists of 24 metabolites and 34 reactions (Table 1). The model details are presented in Dynamic Model for CHO Cell Metabolism, Nolan and Lee, 2009.

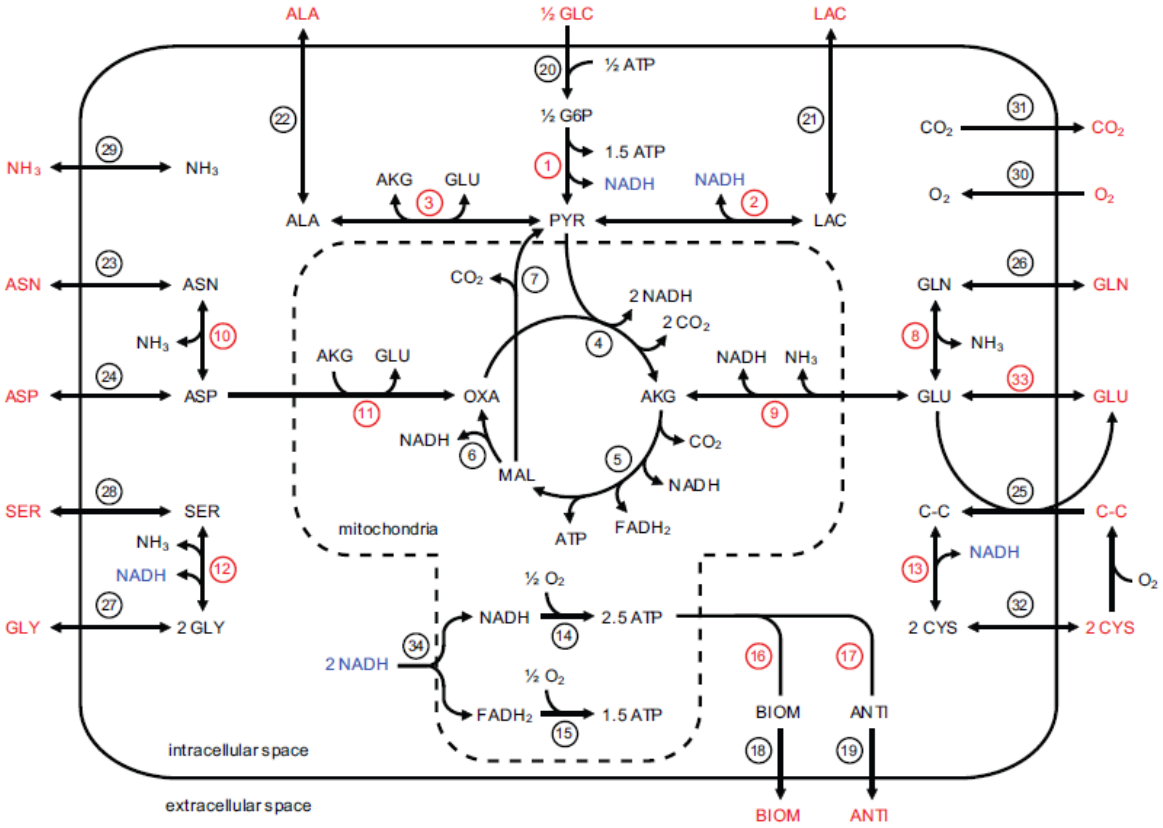


Figure 2 : Network of CHO cell model (Nolan and Lee, 2009)

	Reactions
1	G6P → 2 PYR + 3 ATP + 2 NADHc
2	PYR + NADHc ↔ LAC
3	PYR + GLU ↔ ALA + AKG
4	PYR + OXA → AKG + 2 CO2 + 2 NADH
5	AKG → MAL + CO2 + NADH + FADH2 + ATP
6	MAL → OXA + NADH
7	MAL → PYR + CO2 + NADPH
8	GLN ↔ GLU + NH3 + ATP
9	AKG + NH3 + NADH ↔ GLU
10	ASN → ASP + NH3
11	ASP + AKG → OXA + GLU
12	SER + CO2 + NH3 + NADHc → 2 GLY
13	C-C + NADHc → 2 CYS
14	NADH + 0.5 O2 → 2.5 ATP
15	FADH2 + 0.5 O2 → 1.5 ATP
16	0.084 ALA + 0.041 ASN + 0.080 ASP + 3.755 ATP + 0.026
17	0.061 ALA + 0.034 ASN + 0.039 ASP + 4.000 ATP + 0.024
18	BIOM → BIOM
19	ANTI → ANTI
20	GLC + ATP → G6P
21	LAC ↔ LAC
22	ALA ↔ ALA
23	ASN → ASN
24	ASP ↔ ASP
25	C-C + GLU → C-C + GLU
26	GLN ↔ GLN
27	GLY ↔ GLY
28	SER → SER
29	NH3 ↔ NH3
30	O2 → O2
31	CO2 → CO2
32	2 CYS + O2 → ?
33	GLU → GLU
34	NADHc → 0.5 NADH + 0.5 FADH2

Table 1: Table of reactions used in the flux analysis (Nolan, 2010)

The subsequently reduced network model needs to be adapted to a dynamic system. One approach has been to model the dynamic-evolution of the fluxes with kinetic rate expressions. However, most mammalian metabolic transporters have a large Km, suggesting that control stems from some intracellular enzyme instead. If intracellular metabolites are at pseudo-steady

state, then the source of influence on intracellular enzymes must be extracellular concentrations. By this line of reasoning, many of the kinetic rate expressions were modeled with enzyme-catalyzed reactions dependent on extracellular metabolite concentrations (Nolan, 2010).

In this format, the uptake rate for a metabolite is controlled by multiple intracellular reactions, which allows the model to simulate the effects of reversible enzymes. This is essential, as reversible enzymes strongly impact the cells dynamics. From this, the convenience kinetic expressions were defined and compared to experimental rate data. An initial set of parameters were derived from literature values, and then additionally modified as needed by using non-linear programming to fit to experimental data as best as possible. If the expression still did not fit the data accurately enough, then additional regulatory factors were introduced (Nolan, 2010).

Temperature and redox dependent factors were treated as variables that could scale the kinetic parameter V_{max}. The metabolite inhibition factors related to lactate were included. The production of antibody is dependent on multiple factors, such as the expression of recombinant gene by the CHO cell. The gene expression impacts the rate of translation, and therefore the effective V_{max}. Antibody production is also dependent on the concentration of amino acids, cofactors, inhibitors, and temperature.

The final simulated profiles of the developed model trend closely with the experimentally measured profiled. For 80% of the simulated metabolite profiles the average deviation from the experimental results was less than 30%. Additionally, the simulations correcting predicted the change in profile for the metabolic shifts (Nolan, 2010).

Fed-Batch Culture Optimization

For a standard mammalian fed batch culture, there are numerous parameters that can be varied: temperature, pH, shift day, and feed conditions. These process control parameters can also be

changed over the course of a fed-batch culture, further introducing time-varying factors which could impact enzyme activity. (Nolan and Lee, 2011)

Ideally, optimization of cellular engineering for a mammalian cell culture also takes into account the process conditions, as the optimal cell modification may vary for different process conditions. Therefore, both process control and cellular parameters need to be considered simultaneously in the design space to optimize cell designs for a dynamic culture.

2.2 Thesis Context

A kinetic model for a metabolic network aims provides the ability to simulate dynamics of cellular metabolism. Such models may have the capacity to predict cellular response to various external stimuli, allowing an investigator to reduce the number of time-consuming biological experiments required to obtain a particular result while still gaining a detailed fundamental understanding of the phenomena studied.

However, one limitation for most kinetic metabolic models in literature is that a significant fraction of the proposed parameters cannot be adequately identified with statistical significance. This is due to the fact that the observed data is typically a very small subset of the biologically feasible design space. In addition, several parameters may be redundant, and may be better expressed by combining the parameters. This problem is further exacerbated by the fact that investigators choose to err on the side of caution, and often retain more terms than necessary rather than risk losing details not represented in the model.

However, retaining an excess number of parameters can have negative consequences. It increases the number of parameters that need to be identified and fitted from experimental data and introduces a substantial mathematical limitation in the identification of the important parameters.

Additionally, it is not always clear if all the model details and parameters are necessary for a desired set of model predictions.

In the context of metabolic models, it is imperative to rigorously identify parameters that significantly affect the model outputs. With a large number of parameters in a highly non-linear model, a great number of the parameters are candidates for parameter estimation and re-fitting. However, if the parameter number is large, it can be difficult to comprehensively perform parameter estimation.

Another region of interest for parameter sensitivity analysis is for model-based optimization, e.g. determining the best conditions for the highest antibody production yield. It is observed that we cannot readily perform a comprehensive model-based optimization on the dynamic CHO cell model. A minimalistic three level exploration of 8 enzyme parameters and 4 process parameters was explored in "Dynamic Model for CHO Cell Engineering" (Nolan and Lee, 2011). This exploration required ~10,000 simulations (Figure 3). Additionally, high producing conditions (red) are found amidst the infeasible conditions (dark blue), which also leads us to question the robustness of this optimization. Lastly, we are unsure if the parameters chosen (enzymes of the network reactions) truly have the greatest impact on the output of the model simulation. For these reasons, model-based global-optimization benefits from the ability to select a few high impact parameters.

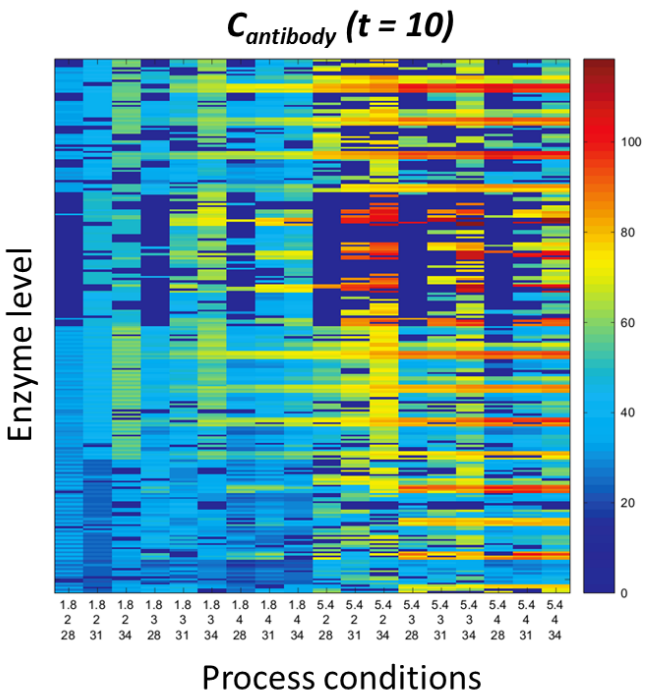


Figure 3: The heat map represents the 4608 potential day 10 antibody concentrations (mM) when subjected to genetic and process manipulations for model based optimization The x-axis labels are: seed density ($\times 10^6$ cells/mL), temperature shift day, and shift temperature (°C). (Nolan and Lee, 2011)

2.3 Thesis Objectives

We aim to develop a methodology for parameter significance determination in the context of kinetic models. This methodology will determine the terms that are appropriate to re-fit for parameter estimation.

3 Literature Review

3.1 Parameter Identifiability

Parameters in a model can be described as identifiable, non-identifiable or non-observable. The value of identifiable parameters can be estimated with some degree of confidence, and their value affects the outputs. The value of non-identifiable parameters cannot be estimated accurately; however, their value affects the outputs. Non-observable parameters have no effect on the outputs (Godfrey, 1984).

Parameters may have non-identifiability issues due to two major sources: the model may have too many parameters or the data may not be appropriately sampled. In the context of parameter estimation, it is important to screen for and detect potential parameter non-identifiability. Once non-identifiability is detected, it can be resolved either by an improved experimental design or a model reduction (Meshram, 2012).

3.2 Parameter Sensitivity

Sensitivity analysis attributes uncertainty in the output of a model to sources in its inputs. This type of analysis has many applications in model development. Sensitivity analysis can be used to test the robustness of a model, by both identifying model inputs that cause uncertainty and should be further studied and also shedding light on unexpected relationships. Additionally, sensitivity analysis leads to an increased understanding of the relationships between input factors and response outputs, which can lead to: an enhanced understanding of the model, model simplification, and an improved set from which to perform model based optimization. (Ravalico, 2005)

Sensitivity analysis is a broad field, and therefore it is important to define a criterion to assess the available methodologies. The first consideration is computation efficiency, which comes into consideration when considering large, complex models with a high number of parameters. The second consideration is parameter interactions. Models need to be able to evaluate and take into account parameter interactions, as they are inevitable in a complex model. The third consideration is data requirements. Often, sensitivity analysis techniques require knowing the parameter probability distributions. However, determining probability distributions of models may prove to be very difficult if not impossible. The fourth consideration is model non-linearities, as complex models rarely behave linearly. The fifth consideration is ability to be applied to decision making. There should be some physical meaning to the sensitivity analysis outputs (Saltelli et al., 2000).

One-at-a-Time

One of the simplest sensitivity analysis approaches is the one-at-a-time method. This method consists of changing one input variable at a time, while the others are kept at their nominal value. The sensitivity can then be measured as a factor of changes in the output, either by partial derivatives or linear regression. This allows direct observation of elementary effects, without concern of confounding, as any change in the output can be directly attributed to the single variable adjusted. This is also useful in the case of unusual model results, as it is immediately evident which input factor is responsible (Montgomery, 2005).

However, this approach has a great weakness in that it cannot fully explore the input space, as it does not take into account parameter interactions.

Regression

Regression analysis is a class of techniques, which quantify how strongly correlated each individual parameter is with the output and the corresponding sensitivity. Regression analysis techniques are traditionally designed for linear models. Non-linear models can be analyzed through a rank transformation of the outputs. However, some information is lost in rank transformation, as the analysis is now being performed on a surrogate model. Additionally, rank transformations lose their efficacy for non-monotonic models.

Regression techniques have weaknesses in their ability to account for parameter interactions, as well as their ability to account for non-linear models. (Saltelli, 2008)

FAST and eFAST

Fourier amplitude sensitivity testing is a variance based global sensitivity analysis method. The sensitivity value is determined based on conditional variances which indicate the individual or interaction effects of inputs on the output, in the form of the total sensitivity indices. The parameter space is sampled by a transformation function. For FAST, the transformation function is independent from the model structure. However, the estimation of the total sensitivity results in the grouping of results. Additionally, FAST assumes a uniform parameter distribution. While this is advantageous if the distributions are unknown, it can be an incorrect assumption. eFAST is very similar to FAST, the main difference being the transformation function used to sample the parameter space (Cukier, 1973).

Sobol

The Sobol sensitivity methodologies are based on emulators. The emulator is trained on sampled data set, and fit with a simpler approximation to form a meta-model within an appropriate margin of error. Then, the sensitivity measurements are calculated from the emulator. These

methodologies have the advantage of requiring a fewer number of runs. For Sobol, the method uses a decomposition of the model into summands of increasing dimensionality via the Sobol sequence. A disadvantage of Sobol is its requirement of knowledge of the parameter distributions (Sobol, 1993).

3.3 DoE Background

Design of experiments (DoE) is a well-established class of methodology for experimentation in which factors are varied for the purpose of observing the change in outputs. DoE formulates an experimental space and schema in a fashion that will provide the most information with a limited number of experiments. Additionally, DoE allows for the measurement of how interactions between different factors influence the model. (Montgomery, 2005)

The DoE methodology maps experiments in such a fashion to study interactions. It also allows for a statistical model to be developed from the results of factor variation on the outputs of the experiment. The resultant statistical model can be used to better study the process of interest, which in our case is fed-batch culture. In this fashion, DoE aids in the understanding of the process without requiring knowledge of the underlying process.

The statistical models created from the DoE are often in the form of response surface models (RSM). RSMs approximate a functional relationship between a response output and the factors. In general, this relationship is approximated by a low-degree polynomial, negating the need to understand the true relationship between the factors and outputs. RSMs are generally used for three purposes: to establish an approximate relationship between outputs and factors that can be used to predict response values, to determine the significance of factors, to determine the optimum values of factors that result in the desired output over a certain design space (Khuri, 2010).

Factorial Designs

There are several types of design considered in DoE methodology. Factorial designs are the simplest, testing every possible combination of the factors at all of the levels. For designs with a larger number of factors, this can get quite unwieldy in experimental numbers, as an *l*-level design with *n* factors has l^n corresponding experimental points (Box, 2005).

Fractional Factorial Designs

Fractional factorial designs utilize a subset of the full factorial design, substantially reducing the number of experimental points to be run. This subset is chosen so as maximize the sparsity of effects, in which main and low-order effects take precedent. An *l* level design with *k* factors and *p* fraction has l^{k-p} experimental points. Traditionally, fractional factorial designs are only used for 2-level designs, as response surface models are more experimentally efficient at higher levels. Using a factorial design at 2-levels means that any curvature of the response will not be observed, as quadratic terms cannot be accounted for at two levels. This limits the applications of fractional factorial designs (Box, 2005).

Central Composite Designs

Central composite designs (CCD) are an augmentation of a 2-level factorial design, in which additional center and 'star' points which allow for the modeling of second order models with only a 2-level design. CCDs offer an efficient alternative to 3-level factorial designs. However, as the number of design variables increases, CCD loses its appeal, as the number of experiments required increases too quickly.

Optimal Designs

Optimal designs are a class of experimental designs in DoE, which are optimized with respect to some statistical criterion. A sub-optimal design will require additional runs to estimate

parameters with the same degree of precision as an optimal design. Optimal designs offer a few concrete advantages over sub-optimal designs. Because optimal designs require fewer experimental runs than sub-optimal designs, they reduce the cost of experimentation. Optimal designs are also capable of handling multiple types of factors. Lastly, optimal designs can be optimized for constrained design-spaces (Montgomery, 2005).

There are many types of optimal designs. We discuss two of the most commonly used types of optimal designs, D- and I-optimal.

D-optimal designs are computer algorithm generated designs, the D- abbreviating 'determinant'. They are designed with the minimization of the determinant for the variance-covariance matrix of the parameter estimates for the specified RSM. The motivations for using a D-optimal design instead of classical designs are that classical designs require too many runs for the amount of resources, or the design space is constrained (Unal 1998). D-optimal designs are the most appropriate design for screening experiments, as the optimality criterion delivers precise estimates of the coefficients. Precise estimates of the factor effects are essential for discriminating factor impact. Therefore, D-optimal designs are most commonly used for model inference and decision making.

I-optimal designs minimize the lowest integrated or average prediction variance over the design space, with the I- abbreviating 'integrated variance'. This makes I-optimal designs the choice for models with a focus on prediction (Montgomery, 2005).

3.4 Response Surface Models

Response surface models (RSM) are often used in the design of experiments methodology to provide a meta-model for the more complex process being studied. They are often constructed as a set of polynomial bases. The most common bases used for RSMs are linear, quadratic, quadratic with interactions, and cubic, although higher powers can also be used (Equation 2). The inputs to the model are often scaled to be within the range of [-1,1] in order to improve model robustness. (Box, 2005)

$$y = \beta_0 + \sum_{j=1}^{51} \beta_j x_j + \sum_{k=j+1}^{51} \sum_{j=1}^{51} \beta_{jk} x_j x_k + \sum_{j=1}^{51} \beta_{jj} x_j^2 \quad (2)$$

Equation 2: Format of quadratic RSM with interactions

4 Our Methodology

We aim to develop a methodology for parameter significance determination in the context of kinetic models. This methodology will determine the terms that are appropriate to re-fit for parameter estimation. For the development of this methodology, we explore the dynamic CHO cell model.

4.1 The Rationale

For the development of this parameter sensitivity methodology, we use the dynamic CHO cell model referenced earlier. Our interest in parameter sensitivity methodology for the dynamic CHO cell model stems from the need to perform parameter estimation globally.

In the dynamic CHO cell model, all of the parameter estimation was performed locally, not globally, as it has too many parameters to solve for at once. The kinetic parameters that had a literature basis were fitted through parameter estimation. Many of the kinetic parameters used in the model development had no literature value to be referenced. As such, many kinetic parameters were found through parameter estimation, with the goal of minimizing the difference between the simulated and experimental metabolic profiles. Especially with the lack of literature values, there is a need to appropriately perform parameter estimation and fit the kinetic parameters to the experimental data.

Because the parameter estimation was performed locally, it is unlikely that the values selected for the dynamic CHO cell model provide the best fit. Global parameter estimation greatly increases that likelihood. Using parameter sensitivity methods, we can reduce the number of terms to be considered, making global parameter estimation feasible, and thus improve the confidence of the model fit.

Based on the literature review, we determined that there is not an appropriate sensitivity analysis technique applicable to the dynamic CHO cell model. The sensitivity analysis methods considered were: one-at-a-time, regression, SOBOL, and FAST. We want to explore the interactions between parameters, which one-at-a-time factor analysis does not provide. Regression analysis is not applicable because of its weaknesses in accounting for parameter interactions and for non-linear models. SOBOL is not applicable, as we do not have prior knowledge of the parameter distributions. And lastly, we cannot implement FAST because it groups the parameters and we are interested in the results for the individual parameters. Due to these considerations, it is our objective to develop a new methodology for parameter sensitivity.

4.2 Sensitivity Analysis Methodology

In general, most sensitivity analysis methods follow a similar format. We base our methodology format on the work "Development of a Combined Mathematical and Experimental Framework for the Control and Optimisation of Mammalian Cell Culture Systems." by Alexandros Kyparissidis (Kyparissidis, 2012).

The steps we performed were as follows:

1. Identify the inputs to be varied
2. Quantify the range to be explored for each input
3. Identify the outputs to be analyzed
4. Establish how to sample the design space, and run the model
5. Using the model outputs, create a meta-model
6. Determine the sensitivity of the inputs based on the meta-model

Defining the Inputs

We first consider the potential inputs of the process. There are two types of parameters that can be varied: process and kinetic parameters. The four process variables to be considered are: shift temperature, shift day, seed density and harvest day. The design space for process variables is constrained by feasibility, as certain process variable values will not allow for cell viability. The nominal value for the process variable is considered to be the default values for much of the experimental work. The design space for the process variables is thus dictated by logical feasible conditions, which are:

Process Variable	Lower Bound	Upper Bound
Shift Temperature	28°C	34°C
Shift Day	2	4
Seed Density (x 10^6)	1.8	5.4
Harvest Day	8	10

Table 2: Process variables and their lower and upper bounds

The 47 kinetic parameters to be considered are V_{max}, Km, Ki, and TCs, as defined for convenience kinetics. The nominal values for the kinetic parameters are set to the values defined in the original model, and scaled to be in the same order of magnitude (Table 3).

Kinetic Parameter	Scaled Value
Vmax 1	3.8
Ki1	2
Km1	1
Exp1a	3
Exp1b	1
TB1b	5
Vmax2	2.2
Km2	6
Vmax3f	3.5
Vmax3r	1.5
Km3a	4
Km3b	2.5
Km3c	5
TC3	2
Vmax8f	2.2
Vmax8r	2
Km8a	2.5
Km8b	1
Km8c	1
TC8b	5
Vmax9f	1
Vmax9r	1
Km9	7
Vmax10f	4.75
Vmax10r	2
Km10z	3
Km10b	1
Km10c	2
TC10b	1.5
Vmax11	5.5
Vmax12f	0.9
Vmax12r	2.5
Km12a	1
Km12b	3
Vmax13	3
Km13	1
Vmax16	2.5
Km16a	4
Km16b	3
Km16c	2
TC16b	3
Vmax17	5.25
Ki17	3
Exp17a	5
Exp17b	1
Vmax33a	2
Vmax33b	2

Table 3: Center values for the parameters, scaled to be in the same order of magnitude

To determine a starting range for the kinetic parameter space we need to consider the range over which the kinetic parameters may influence the model output. The parameter space is highly non-linear, and so we first need to understand the parameter space of the kinetic terms independent of each other. For this task, we utilized the linear sensitivity analysis performed for the dynamic CHO cell model by Nolan et al (Nolan, 2010). The value of each kinetic parameter was varied until a 5% change in the collective outputs was observed. The percent change in normalized kinetic parameter value required to induce a 5% change is plotted (Figure 4).

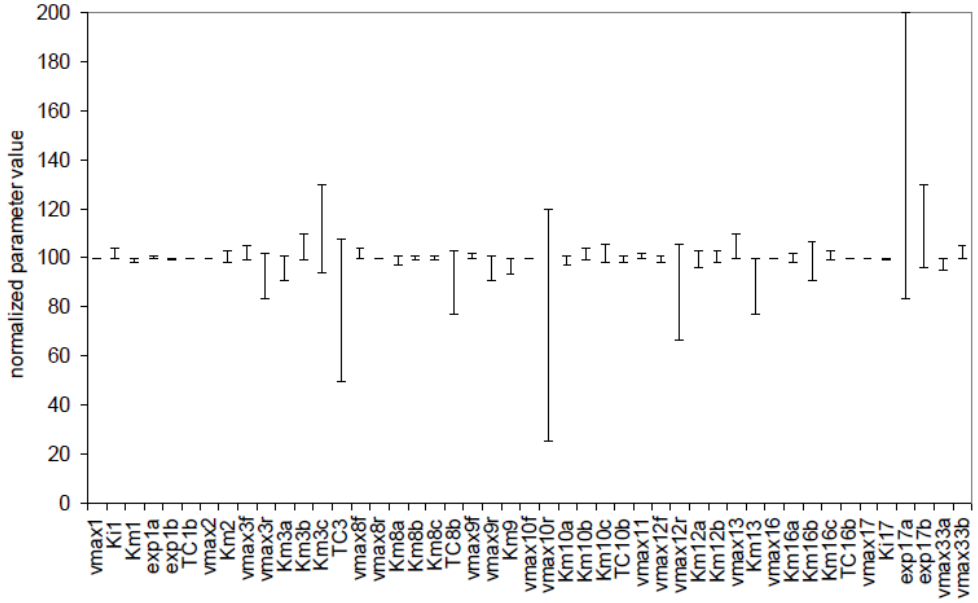


Figure 4: The parameter values were normalized to percentages, such that 100 represents the control parameter set and 150 represents a 50% increase from the control parameter set. The value of each parameter was varied (both up and down) independently until the simulated metabolic profiles collectively deviated from the control set by >5%, this value was defined as the lower or upper bound. (Nolan and Lee, 2010)

This figure demonstrates that the majority of the kinetic parameters can influence a 5% change in the outputs with less than a 50% change in the nominal value, with many kinetic parameters

requiring an even smaller change. We therefore set ±50% the nominal value as the largest possible design space to be explored, and will later investigate reducing the design space.

Although we are ultimately interested in the interactions of parameter space and its impact on the model, this simplistic linear approach helps in providing an adequate starting point for the parameter space.

Outputs of Interest

Having thus determined our inputs and their corresponding ranges, we then look to define our outputs of interest. The model simulates 50 outputs: 34 fluxes and 16 external metabolite concentrations over the time-course of the fed-batch. The model simulates the profiles at 4 points a day for the duration of the 8-10 day batch, resulting in 33-41 time points (including t=0). In order to properly capture the dynamic nature of the model, we consider multiple time points. We therefore determined the sensitivities of the parameters for the 50 outputs, for each time point independent of the other (Nolan, 2010).

Each simulation took approximately 1 minute to complete (when run on an Intel Core 2 Duo CPU at 2.70 GHz with 2.0 GB of RAM).

Sampling the Parameter Space

The next step is to determine the experimental points to be simulated. We considered four types of sampling methods, from most comprehensive to least: factorial, fractional factorial, CCD and optimal designs.

The most comprehensive method for parameter space sampling is a factorial design. A two-level design does not provide details of quadratic effects, and with the 51 parameters would require 2^{51} (about 2.2×10^{15}) experimental points to be simulated. This is a prohibitively high number of experimental points to simulate (requiring ~31 million years to run), and still would not provide

the level of detail we are interested in (e.g. quadratic effects), and was thus factorial designs were not considered further.

For fractional factorial designs, even the most conservative options once again are in the range of 10^{10} simulations, without giving us information about quadratic level interactions, and so we determine that fractional factorial designs are also not an option.

CCD is often the 'go-to' classic DoE design, developed by Box-Wilson. However, for higher numbers of parameters, CCD cannot be applied, as neither JMP nor Matlab is capable of creating a design with 51 parameters. Additionally, using CCD for just 25 factors (half the number of factors we are investigating) yields a design with over 5 million runs, still an impractical number of runs.

D-optimal designs allow for user specification of higher order RSMs. A full quadratic RSM (linear, interaction, and quadratic terms) specified by D-optimal design for 51 parameters yields a design space with ~2000 runs, which is a much more realistic experimental plan.

Therefore the D-optimal design was selected for this project. We used JMP to create a full quadratic D-optimal design for 51 factors. The execution of the full set of experimental points was performed in MATLAB, and the outputs were returned to JMP for analysis. The statistical values tracked were p-value and beta coefficient for each term in the RSM. For our initial analysis, we consider a p-value of ≤ 0.05 to be significant.

Having marked a ±50% bound as the upper limit of the design space of interest, via a linear sensitivity analysis, we then tested this and more narrow parameter spaces to determine the appropriate range when considering parameter interactions. A parameter space that is too large may induce parameter effects that are disproportionately large and therefore overshadow other

high impact parameter effects. Conversely, a parameter space that is too small may not create noticeable impacts.

We performed the full set of experimental points for bounds of ±50, 30, 20, 10, and 5%. We tabulated the instances of significance for the parameters over the model outputs (figure). In comparing the instances of significance for the outputs over the various sized design spaces, we find that there is a prominent outlier. Within the design space of ±20%, more terms were found to be significant for each output than the other design space sizes. This high number of significant terms suggests that we have found an appropriate balance in choosing the size of the parameter space to be explored. We thus choose our parameter space to be ±20%.

A second consideration for the simulation is error. The simulation returns near identical values for repeat runs of the center points. Statistically, this indicates that the observed variance of error is 0, which will always return a poor lack-of-fit. In order to defray the impact of this, we add an error to the simulation outputs. The experimentally determined results which the model is based on are concentrations, and so we look the error of concentration measurements. In general, a ±5% error is acceptable for metabolite concentration measurements by HPLC-MS. Therefore, we add an error of ~5% to the outputs of the model simulation before analysis. With or without error, we find that the general trend of term significance remains the same; however, it brings the lack-of-fit into a reasonable range, which will be useful for later analysis.

Response Surface Model

Our next task is to determine the order of RSM necessary to study our model. We look at both linear and full quadratic models, with the factors normalized to vary between -1 and +1. While we would like to also study cubic models, the number of experimental runs was much too high to be practical. When we compare the lack-of-fit values for linear versus quadratic RSMs, we find

that the lack-of-fit for linear RSMs is significant. However, we find that the lack-of-fit for quadratic RSMs is more appropriate (Table 4).

RSM Type	BIOM	ANTI	ASP	GLY	SER	NH3	GLU	VCD	TCD	%
Full Quadratic Model	0.96	0.81	0.45	0.27	0.07	0.00	0.95	0.86	0.20	0.99
Linear Model	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

Table 4: Lack of fit for concentration outputs

4.3 Parameter Discrimination Criterion

Having thus satisfactorily determined both the design space of interest and the RSM design of interest, we then move forward with the task of parameter sensitivity analysis. We treat each time-point independently. We first look at the p-value for each term in the RSM. We consider a value of $p \leq 0.05$ to correspond to significance. We calculate for each of the 50 RSM outputs, which of the terms are significant. We then repeat this process for each time-point. Comparing the time-points, we observe that significance is similar across all of the time-points. We now have p-value significance of the 51 parameters for each of the 50 y-outputs for each time point. We then sum the number of instances of significance across the y-outputs for each parameter term. The most possible instances of significance is 50, the lowest is 0. We use these values to get an overall sense of parameter significance for the specific time point (Figure 5).

We observe that the parameter significance is not straightforward. Parameter terms that directly appear in a flux equation are not necessarily significant for that flux output (Figure 6). For example, Vmax9f, Vmax9r, and Km9 appear in the kinetic expression of the flux for the y-output 'AKG + NH3 + NADH ↔ GLU' but all three terms are not significant for that output.

Parameter	BIOM	ANTI	GLC	LAC	ALA	ASN	ASP	C-C	GLN	GLY	SER	NH3	GLU	VCD	TCD	%
Intercept	<0.01	<0.01	<0.01	0.01	<0.01	0.00	<0.01	<0.01	0.52	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Shift Temp	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.77	<0.01	<0.01	0.83	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Shift Day	<0.01	<0.01	0.13	<0.01	0.79	<0.01	<0.01	0.18	<0.01	<0.01	<0.01	0.02	0.06	<0.01	<0.01	<0.01
Seed Density	<0.01	<0.01	<0.01	<0.01	0.00	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.00	<0.01	<0.01	<0.01
Harvest Day	<0.01	<0.01	<0.01	0.50	0.00	<0.01	0.17	<0.01	0.10	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Vmax 1	0.01	<0.01	0.75	<0.01	<0.01	0.02	<0.01	0.11	<0.01	<0.01	<0.01	<0.01	<0.01	0.01	0.01	0.01
Ki1	0.36	<0.01	0.34	<0.01	<0.01	0.00	<0.01	<0.01	0.01	<0.01	0.00	<0.01	<0.01	0.32	0.36	0.44
Km1	0.03	<0.01	0.07	<0.01	<0.01	0.09	0.08	0.18	0.08	0.00	0.00	<0.01	0.00	0.05	0.02	0.01
Exp1a	0.20	0.10	0.24	0.57	0.29	0.50	0.54	0.34	0.06	0.01	0.13	0.29	0.01	0.25	0.20	0.17
Exp1b	0.00	<0.01	0.37	<0.01	<0.01	0.40	<0.01	0.10	0.69	<0.01	0.00	0.00	<0.01	0.00	0.00	0.00
TB1b	0.14	<0.01	0.16	<0.01	<0.01	0.24	<0.01	0.98	<0.01	<0.01	0.00	<0.01	<0.01	0.16	0.14	0.12
Vmax2	0.05	<0.01	0.86	<0.01	<0.01	0.69	<0.01	0.35	0.12	<0.01	0.01	<0.01	0.00	0.08	0.05	0.03
Km2	0.61	0.27	0.12	0.31	0.05	0.32	0.22	0.29	0.62	0.60	0.23	0.16	0.19	0.73	0.61	0.49
Vmax3f	<0.01	<0.01	0.02	0.64	<0.01	0.95	<0.01	0.04	<0.01	0.00	0.00	<0.01	<0.01	<0.01	<0.01	<0.01
Vmax3r	0.66	0.80	0.65	0.10	<0.01	0.66	0.15	0.92	0.54	0.16	0.40	<0.01	<0.01	0.66	0.66	0.67
Km3a	0.64	0.36	0.86	0.72	<0.01	1.00	0.56	0.96	0.03	0.94	0.90	0.04	0.00	0.80	0.63	0.45
Km3b	0.79	0.27	0.68	0.68	<0.01	0.08	0.09	0.58	0.21	0.84	0.96	0.00	<0.01	0.71	0.79	0.89
Km3c	0.37	0.22	0.67	0.02	0.01	0.69	0.32	0.53	0.53	0.60	0.34	0.17	0.08	0.39	0.37	0.38
TC3	0.60	0.95	0.85	0.64	0.14	0.32	0.83	0.24	0.55	0.47	0.55	0.39	0.07	0.43	0.60	0.87
Vmax8f	<0.01	<0.01	0.00	0.16	0.02	0.02	<0.01	0.15	<0.01	<0.01	0.00	0.02	<0.01	<0.01	<0.01	<0.01
Vmax8r	<0.01	<0.01	0.06	0.00	<0.01	0.41	<0.01	0.86	<0.01	<0.01	0.08	0.11	<0.01	<0.01	<0.01	<0.01
Km8a	0.01	0.00	0.25	0.01	0.01	0.36	0.00	0.89	<0.01	0.00	0.11	0.13	<0.01	0.01	0.01	0.00
Km8b	<0.01	<0.01	0.00	0.02	<0.01	0.01	<0.01	0.21	<0.01	<0.01	0.00	0.02	<0.01	<0.01	<0.01	<0.01
Km8c	0.01	0.00	0.59	0.55	0.01	0.40	0.03	0.74	<0.01	0.07	0.28	0.98	<0.01	0.02	0.00	0.00
TC8b	0.77	0.96	0.96	0.98	0.73	0.69	0.15	0.96	0.00	0.42	0.80	0.29	0.10	0.75	0.77	0.82
Vmax9f	0.02	0.22	0.08	0.52	<0.01	0.78	<0.01	0.19	<0.01	0.33	0.11	0.08	<0.01	0.03	0.02	0.02
Vmax9r	0.55	0.58	0.27	0.56	0.22	0.32	0.00	0.57	0.02	0.86	0.27	0.07	0.02	0.52	0.54	0.60
Km9	0.35	0.25	0.17	0.67	0.00	0.14	<0.01	0.04	0.52	0.68	0.05	0.14	<0.01	0.26	0.35	0.50
Vmax10f	0.00	<0.01	0.21	0.04	<0.01	<0.01	<0.01	0.79	<0.01	<0.01	0.22	<0.01	0.04	0.00	0.00	0.00
Vmax10r	0.21	0.16	0.27	0.90	0.16	0.88	0.07	0.20	0.99	0.54	0.29	0.37	0.89	0.35	0.20	0.10
Km10z	0.16	0.17	0.35	0.63	0.00	0.00	<0.01	0.20	<0.01	0.36	0.52	<0.01	0.05	0.16	0.16	0.18
Km10b	0.51	0.93	0.55	0.47	0.01	0.00	<0.01	0.49	0.00	0.94	0.41	<0.01	0.33	0.65	0.51	0.38
Km10c	0.51	0.39	0.37	0.50	0.26	0.03	<0.01	0.41	0.02	0.95	0.58	<0.01	0.56	0.57	0.51	0.47
TC10b	0.76	0.25	0.73	0.51	0.00	<0.01	<0.01	0.53	<0.01	0.50	0.81	<0.01	0.09	0.83	0.76	0.70
Vmax11	<0.01	<0.01	0.02	0.67	<0.01	0.00	<0.01	0.39	<0.01	0.00	0.13	<0.01	<0.01	<0.01	<0.01	<0.01
Vmax12f	0.18	0.83	0.28	0.15	0.00	0.76	0.49	0.34	0.42	<0.01	<0.01	0.51	0.93	0.22	0.17	0.14
Vmax12r	0.92	0.70	0.55	0.18	0.21	0.54	0.52	0.47	0.61	0.00	0.86	0.47	0.15	0.84	0.92	0.97
Km12a	0.17	0.18	0.46	0.76	0.32	0.94	0.00	0.89	0.05	<0.01	0.03	0.00	0.00	0.23	0.17	0.13
Km12b	0.86	0.95	0.49	0.98	0.79	0.78	0.32	0.70	0.75	<0.01	0.01	0.02	0.29	0.87	0.86	0.85
Vmax13	0.28	0.02	0.51	0.30	0.07	0.86	0.01	<0.01	0.30	0.09	0.36	0.52	<0.01	0.42	0.28	0.17
Km13	0.40	0.45	0.49	0.82	0.90	0.15	0.42	<0.01	0.96	0.72	0.16	0.10	0.00	0.40	0.40	0.43
Vmax16	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Km16a	<0.01	<0.01	<0.01	0.35	0.02	0.00	<0.01	0.02	0.82	<0.01	<0.01	0.00	0.60	<0.01	<0.01	<0.01
Km16b	0.00	0.01	0.24	0.11	0.00	0.93	<0.01	0.87	0.60	0.06	0.42	0.00	0.67	0.00	0.00	0.00
Km16c	<0.01	<0.01	<0.01	0.00	<0.01	0.00	<0.01	0.00	0.03	<0.01	<0.01	0.00	0.99	<0.01	<0.01	<0.01
TC16b	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.00	<0.01	<0.01	<0.01
Vmax17	0.14	<0.01	<0.01	<0.01	<0.01	0.08	<0.01	0.01	<0.01	<0.01	<0.01	<0.01	0.02	0.23	0.14	0.08
Ki17	0.00	<0.01	<0.01	0.00	<0.01	0.43	<0.01	0.03	<0.01	<0.01	<0.01	<0.01	<0.01	0.00	0.00	0.00
Exp17a	0.27	0.12	0.61	0.29	0.14	0.70	0.35	0.99	0.92	0.00	0.37	0.90	0.84	0.42	0.27	0.15
Exp17b	0.57	0.28	0.67	0.66	0.52	0.99	0.23	0.91	0.65	0.21	0.84	0.15	0.48	0.55	0.57	0.62
Vmax33a	0.37	0.75	0.90	0.76	0.95	0.48	0.07	0.80	0.10	0.90	0.89	0.46	<0.01	0.38	0.37	0.40
Vmax33b	0.80	0.38	0.62	0.40	0.00	0.73	0.00	0.66	0.02	0.82	0.79	0.13	<0.01	0.78	0.80	0.84

Figure 5: Heat map of parameter significance of linear terms for the concentration RSMs at t=harvest day. The cell value represents the p-value. Red represents insignificant terms for the RSM. Green represents significant terms for the RSM.

Flux Equation	G6P → 2 PYR + 3 ATP + 2 NADHc	PYR + NADHc ↔ LAC	PYR + GLU ↔ ALA + AKG	PYR + OXA → AKG + 2 CO2 + 2 NADH	AKG → MAL + CO2 + NADH + FADH2 + ATP	MAL → OXA + NADH	MAL → PYR + CO2 + NADPH	GLN ↔ GLU + NH3 + ATP	AKG + NH3 + NADH ↔ GLU	ASN → ASP + NH3	ASP + AKG → OXA + GLU	SER + CO2 + NH3 + NADHc → 2 GLY	C-C + NADHc → 2 CYS	NADH + 0.5 O2 → 2.5 ATP	FADH2 + 0.5 O2 → 1.5 ATP	0.084 ALA + 0.041 ASN + 0.080 ASP + 3.755 ATP + 0.026 CYS + 0.452 G6P + 0.087 GLN + 0.056 GLY + 0.427 OXA + 0.096 SER → BIOM + 0.004 FADH2 + 0.008 GLU + 0.445 MAL + 0.639 NADH + 0.209 PYR	0.061 ALA + 0.034 ASN + 0.039 ASP + 4.000 ATP + 0.024 CYS + 0.048 GLU + 0.045 GLN + 0.072 GLY + 0.126 SER → Antibody
Term	**Y**	**Y2**	**Y3**	**Y4**	**Y5**	**Y6**	**Y7**	**Y8**	**Y9**	**Y10**	**Y11**	**Y12**	**Y13**	**Y14**	**Y15**	**Y16**	**Y17**
Intercept	***	***	***	***	***	***		***		***	***	***	***	***	***	***	***
Shift Temp	***	***	***	***	***	***	***			***	***	***		***	***	***	***
Shift Day	***	***	***	***			***	***		***		***	***			***	***
Seed Density	***		***	***	***	***	***	***		***	***	***	***	***	***	***	***
Harvest Day		***	***		***	***	***	***		***	***		***	***	***	***	***
Vmax 1		***	***	***	***	***	***			***	***	***	***	***	***	***	***
Ki1	***			***	***	***	***			***	***		***	***	***	***	***
Km1			***	***	***	***				***	***			***	***		***
Exp1a	***									***							***
Exp1b			***		***	***	***	***		***	***			***	***		***
TB1b		***	***		***	***	***			***	***			***	***		***
Vmax2			***		***	***	***	***		***	***		***	***	***	***	***
Km2	***	***											***				***
Vmax3f			***				***			***	***						
Vmax3r			***						***	***	***						
Km3a			***				***			***	***						
Km3b			***														
Km3c	***																
TC3			***														
Vmax8f			***										***				
Vmax8r			***					***	***							***	***
Km8a			***														***
Km8b			***									***					
Km8c			***					***								***	
TC8b																	
Vmax9f								***		***	***						
Vmax9r																	
Km9																	
Vmax10f							***	***		***	***						
Vmax10r																	
Km10z			***					***		***	***						
Km10b				***				***		***	***						
Km10c										***	***						
TC10b	***	***					***	***		***	***						
Vmax11	***		***		***	***	***	***	***	***	***						
Vmax12f				***		***		***				***		***			
Vmax12r																	
Km12a				***			***			***		***					
Km12b										***	***	***					
Vmax13				***	***	***							***	***	***		
Km13				***									***				
Vmax16	***	***	***	***	***	***	***	***		***	***	***	***	***	***	***	***
Km16a			***	***						***							
Km16b	***			***	***	***		***				***	***	***	***	***	
Km16c		***			***	***		***				***	***	***	***	***	
TC16b	***	***	***		***	***	***	***		***	***	***	***	***	***	***	***
Vmax17	***				***	***	***	***					***	***	***		***
Ki17	***				***	***							***	***	***	***	***
Exp17a				***		***									***	***	
Exp17b				***												***	***
Vmax33a				***			***				***						
Vmax33b			***	***			***				***						

Figure 6: A value of "***" represents term significance. A colored cell represents a terms presence in the respective flux equation.

We then repeat this analysis for all of the time-points of the simulation, and then sum to the parameters accordingly. We now have an understanding of the parameter significance for the terms across the fed-batch culture.

Criterion A

Our first criterion is a rule for discriminating highly insignificant parameters. The goal is to apply a criterion to the parameter terms, such that we can sort the terms by their results, and determine appropriate parameter groupings.

We consider a value of $p \leq 0.05$ to correspond to significance, and a p-value of >0.05 to be non-significant, and sort instances of significance for the terms accordingly. We find that for 6 of the 51 parameter, their linear terms are never significant. Additionally, we find that for 12 of the 51 parameters, their linear and cross terms are significant in less than 5% of the outputs over time. We note that these 12 parameters are all kinetic parameters. We categorize these parameters as highly insignificant.

We define parameter insignificance as a parameter's ability to take on any value in the parameter space with minimal impact on the model outputs.

With this criterion, we are now able to discriminate highly insignificant terms. Before we apply this information to the actual model, we first test this on the RSMs. We observe the impact of term removal on lack-of-fit for the RSMs. If term removal has a large impact on the RSMs, it can be anticipated that it will also have a large impact on the actual model. We test this by observing the change in lack-of-fit statistic, when the RSMs is re-fitted without the 12 insignificant terms. We find that the removal of all 12 of the insignificant terms has a minimal impact on the lack-of-

fit statistic (Table 5). We therefore feel confident moving on to the next step, and making further decisions.

RSM Type	BIOM	ANTI	ASP	GLY	SER	NH3	GLU	VCD	TCD	%
Full Quadratic Model	0.96	0.81	0.45	0.27	0.07	0.00	0.95	0.86	0.20	0.99
Reduced Model 12 Insignificant Terms Removed	0.96	0.79	0.43	0.26	0.07	0.00	0.72	0.85	0.20	0.99

Table 5: A comparison of lack-of-fit for the full quadratic model versus the reduced model

Looking at the results across the different y-outputs, we can see that while we have great resolution within the insignificant terms, but not a lot of resolution through the significant terms (Figure 7). We find that there is a gradient and range of degrees for parameter insignificance. However, there is not much distinction between terms that are significant. From this, we decide we need another criterion to further discriminate the highly insignificant from highly significant terms.

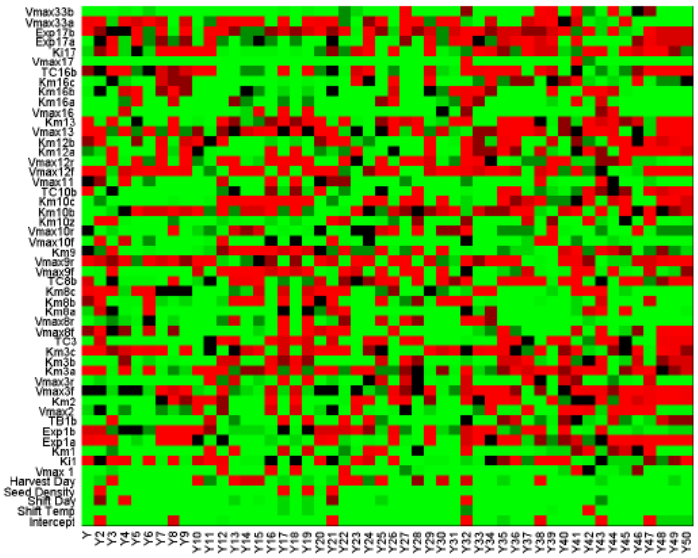


Figure 7: A heat map of the overall term significance from criterion A. Green represents significant, red is insignificant. There is little discrimination between significant terms

Criterion B

We next look to the coefficients of the RSMs. The magnitude of the coefficients, relative to the constant term, determines the weight of the parameter in each RSM. It therefore may have telling information that can be translated to further discriminate parameter significance. We look to the magnitude of the beta coefficient of a parameter over the beta coefficient of the intercept of the RSM.

We use the magnitude of the beta coefficient as a second screening criterion. Therefore, for a term to be considered significant, it must have a p-value of ≤ 0.05 and a coefficient ratio, $\frac{B_k}{B_o}$, with a magnitude of ≥ 0.1. Using this criterion, we can thus sort to find the terms that are considered highly significant. Applying this criterion, we find that there is much more discrimination now in significance (Figure 8). We observe that a very small sub-population of the terms appear as significant, as compared to criterion A (Figure 7).

We find that there are 10 terms which are considered significant in at least 10% of the RSMs with these criteria. We note that all 4 of the process parameters are considered highly significant. This is not surprising, as process parameters are the go-to choice of engineers for affecting process outputs.

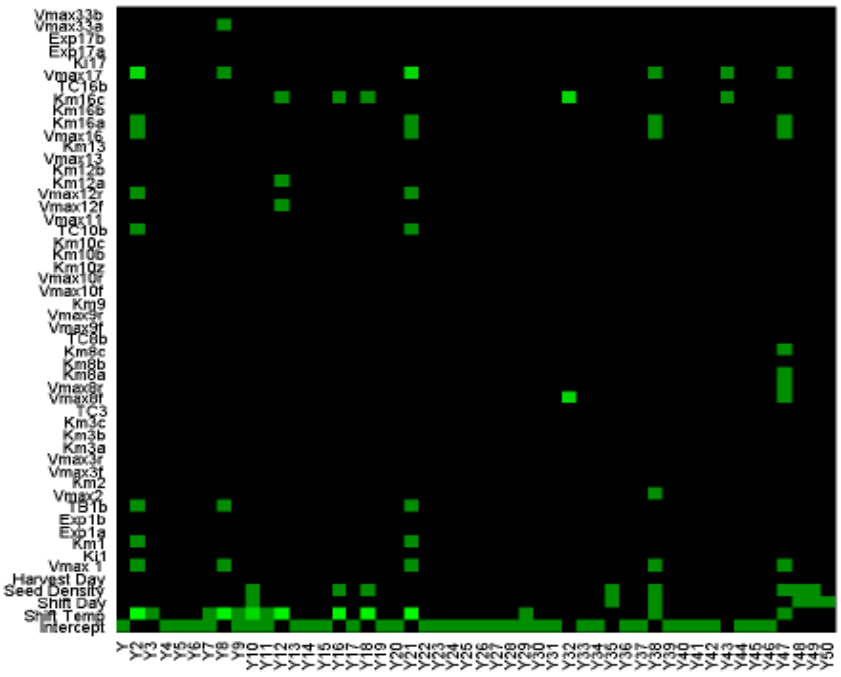


Figure 8: A heat map of the overall term significance from criterion B. Green represents highly significant.

Having now created two categorizations of significance, we now wish to test their relevance in the simulation. We find that the highly insignificant terms are:

Exp17b

Exp1b

Km10b

Km13

Km3a

Km3c

TC8b

V_{max} 12f

V_{max} 13

V_{max} 2r

V_{max} 3f

V_{max} 9r

And the highly significant terms are:

Harvest Day

Km16a

Km16b

Seed Density

Shift Day

Shift Temp

V_{max} 1

V_{max} 8r

V_{max} 16

V_{max} 17

We wish to simultaneously vary the values of the parameters for each grouping, and observe their collective impact on the output of the model. If the parameters labeled insignificant were selected correctly, we predict that varying their values will have a minimal impact on the model output. Accordingly, if the parameters labeled as highly significant were sorted correctly, we predict that varying their values will have a large impact on the model output.

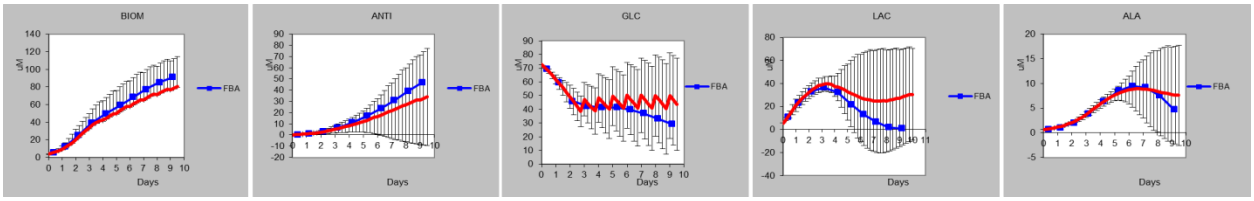
We vary groupings of the model simultaneously, as we have predicted that these parameters have highly significant or insignificant interactions as well. The values of the parameters are varied uniformly within the design space of ±20% of the nominal value. 1000 runs of the varied

parameters are performed. We measure the sum square error and the deviation from the original simulation.

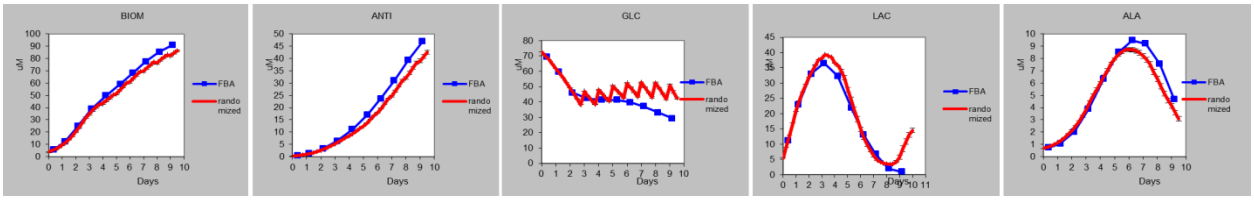
In order to make a fair comparison, we vary only 6 kinetic parameters at a time within each grouping. We vary only the kinetic parameters in this grouping, once again on the basis of being able to compare like-types of parameters, as there are no process parameters in any of the other groupings. We consider this parameter variation to be a Monte-Carlo type simulation. Additionally, in order to have a basis for inherent variation in the system, we vary the values of 6 parameters that have been randomly selected from the full set, without concern to its potential significance, to calculate the result for a 'random' Monte-Carlo simulation.

Even from just visual observation, it can be observed that the high significance parameter groupings have a very large deviation from the mean. The insignificant terms have almost no deviation from the mean. The terms with random significance are observed to have an intermediary deviation from the mean. This visually confirms our results (Figure 9).

Highly Significant Terms



Insignificant Terms



Random Significance Terms

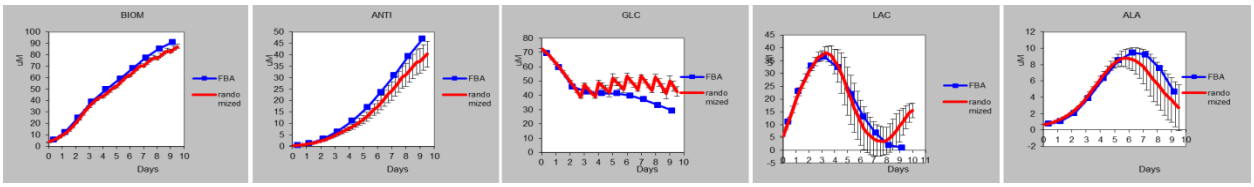


Figure 9: Plots of the Monte-Carlo simulations for the highly significant, insignificant, and random significance groupings. The error bars represent the 2-sigma of the Monte-Carlo simulation, the red line is mean of Monte-Carlo simulations, the blue line is experimental data.

In order to quantitatively confirm our results, we look to the sum square error from the experimental data. We find that variation of six of the 12 insignificant parameters has a SSE of 0.2. If we vary all 12 of the insignificant parameters, we find that the SSE error increases to 1. Comparing this to the SSE of the very significant terms, we find that varying 6 of the highly significant kinetic parameters yields a SSE of 5. This is 25 times higher than the SSE for varying the same number of insignificant terms, and 5 times higher than the SSE for varying twice the number of insignificant terms. This is quantitative proof that the groupings are correct.

Parameter Group	No. of Parameters in Overall Group	No. of Parameters Sampled	Average SSE
Very Significant	6	6	5
Insignificant	12	6	0.2
Insignificant	12	12	1
Original Model	-	-	0.2

Table 6: A comparison of sum square error (SSE) across the different groupings from the Monte-Carlo simulations.

Additional Notes

While we are satisfied with our initial groupings, we still have 29 parameters which have not been labeled. We look to see if we can further discriminate and label these remaining parameters. We look to do this by looking at another discrimination criterion. We initially used the

magnitude of the beta coefficient of the RSMs as a discrimination tool for high significance. Therefore, by adjusting the beta-coefficient and the p-value, we attempt to further discriminate the terms into “intermediate significance” and “intermediate insignificance”. We find that we are unable to divide the remaining parameters into meaningful groups.

For this particular model, we find that we are unable to further divide the intermediate significance terms into meaningful groupings.

5 Parameter Estimation

With the knowledge of which groups are significant, we are now able to more comprehensively look at parameter estimation. We begin parameter estimation by first looking at the very significant parameters. Because these terms are highly significant and have the largest impact on the outputs, they are the parameters of choice to manipulate first.

We want to re-estimate the kinetic parameters, and thus we are currently not interested in the process parameters. There are six highly significant kinetic parameters. We set the objective function as a minimization of the sum squares error of the simulation and the experimental data.

We explore multiple optimization routines in Matlab: fmincon, simulannealbnd, and lsqnonlin, and found the most success with simulated annealing.Simulated annealing is a method for solving unconstrained and bound optimization problems. At each iteration, a new point is randomly generated. The algorithm accepts all points that lower the objective, but also accepts oints that raise the objective within a certain probability. In this way, the algorithm avoids being trapped in a shallow local minimum (Mathworks).

We first re-fitted the 6 highly significant parameters simultaneously, to perform global parameter estimation. We find that the fit is greatly improved, as many model dynamics are now properly captured (Figure 10). The overall objective function is improved by 55% (Table 7).

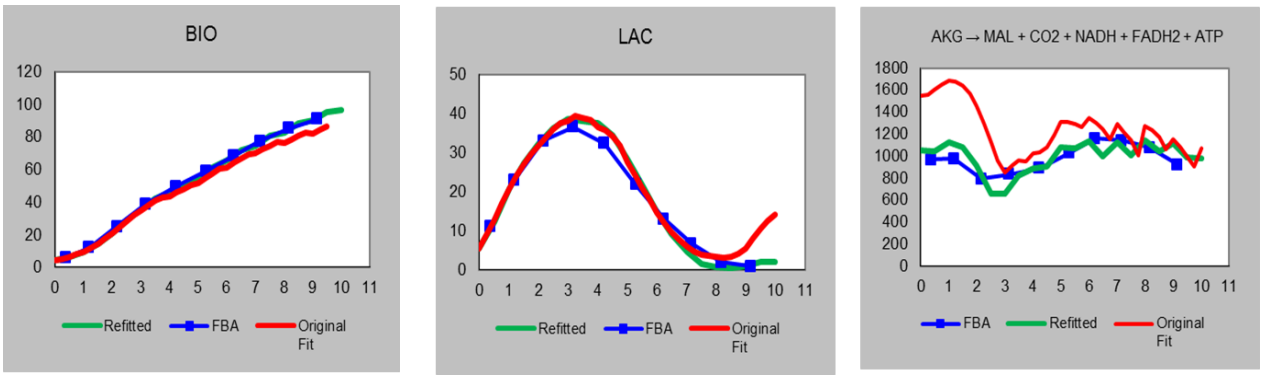


Figure 10: The newly fitted model better captures the model dynamics.

In order to demonstrate that we have picked the correct parameters, we repeat the parameter estimation starting with the 12 insignificant parameters. We find that starting with the insignificant parameters for parameter estimation only yields a 6% improvement in the objective function.

Using the established groups, we look to further demonstrate the impact of the correct groupings on parameter estimation, in the context of the order of parameter estimation. We perform the parameter estimation on the highly significant terms first. We refit the simulation with the results, and then continue the parameter estimation with the 12 insignificant terms. The objective function considered is the sum square of error. We consider this to be a grouped parameter estimation effort, in the 'correct' order, estimating the high impact terms first and the low impact terms last-as a fine tuning mechanism. The overall objective function is improved by 61% (Table 7).

We contrast these results to the parameter estimation performed in the 'inverse' order, with the insignificant terms fit first and the highly significant terms fit last. We find that performing the optimization in the reverse order yields a worse overall fit, and requires a higher number of iterations. Furthermore, the objective function from fitting in the reverse order is less improved compared to the fitting of just the high significant terms.

Terms Re-Estimated (number of terms)	# Iterations Required	Objective Function Value	% Improvement
Insignificant Terms Only (12)	4328	0.41	6%
Significant Terms Only (6)	5672	0.2	55%
Complete Re-Estimation Correct Order (6+12)	7986	0.17	61%
Complete Re-Estimation Inverse Order (12+6)	13792	0.22	50%
Starting Value	-	0.44	-

Table 7: Results from parameter estimation

Through this result, we demonstrate the importance of correct parameter grouping on parameter estimation. Parameter estimation of the highly significant terms yields a large improvement in the overall objective function. Additionally, grouping the highly significant terms allows global parameter estimation for the model, as all terms can be re-estimated simultaneously.

6 Conclusion

Kinetic models of metabolic networks of cells provide insight into the mechanism behind cellular processes, providing a detailed fundamental understanding of the phenomena studied. These kinetic models are generally formulated as set of ordinary differential equations, balancing and calculating intracellular species concentrations. Detailed models of this form provide the ability to simulate the dynamic behavior and responses of cellular metabolism.

These highly detailed models often contain more than the necessary model details, as investigators would rather include too many rather than too few details and parameters. This means that there are often model details and parameters that are not necessary for a desired set of model predictions. However, this causes in an increase in the number of parameters that need to be identified through the experimental data. More importantly, this introduces a substantial mathematical limitation in the identification of parameter significance.

In the framework of kinetic cellular models, there are many reasons for an interest in correctly identifying parameter importance. In a large and highly non-linear model, a great deal of the parameters may require parameter estimation and re-fitting. However, in a large non-linear problem, it can be difficult to efficiently perform parameter estimation.

We find that for the case of the Dynamic CHO Cell (Nolan), that the literature techniques for parameter sensitivity analysis and subsequent parameter re-estimation are in-adequate. We therefore have developed a novel parameter sensitivity discrimination methodology.

Our methodology is inspired by Design of Experiments techniques, and performed on the in-silico model. We demonstrated a technique to appropriately define the appropriate design space

and sampling for the model. We utilized a D-optimal design of experiments to obtain quadratic RSMs. Using the results from the RSMs we were able to group insignificant, intermediate significance, and highly significant terms.

In testing these groupings for the Dynamic CHO cell model using a Monte-Carlo simulation, we find that the groupings were performed correctly.

Additionally, we find that using these parameter groupings, we are able to efficiently perform parameter re-estimation.

For future work we consider the performance of model-based optimization on the dynamic CHO cell model. A comprehensive model-based optimization with all 51 parameters considered becomes an unwieldy and large problem. Minimalistic model-based optimization is demonstrably incomplete, as high producing solutions are found in the midst of infeasible solutions. With a better understanding of parameter significance, a more complete model-based optimization can be performed.

7 References

"Simulated Annealing." MATLAB. MathWorks, n.d. Web. 15 Jan. 2013.

Ahn, Woo Suk, and Maciek Antoniewicz. "Metabolic Flux Analysis of CHO Cells at Growth and Non-growth Phases Using Isotopic Tracers and Mass Spectrometry." Metabolic Engineering 13 (2011): 598-609. Print.

Arfken, G. B., and H. J. Weber. Mathematical Methods for Physicists. San Diego ; London: Academic, 2001. Print.

Chatterjee, Samprit, and Ali S. Hadi. Sensitivity Analysis in Linear Regression. New York: Wiley, 1988. Print.

Chen, W.W.; Neipel, M.; Sorger, P.K. (2010). "Classic and contemporary approaches to modeling biochemical reactions". Genes Dev 24 (17): 1861–1875.

Covert; Schilling, C.; Palsson, B. (2001). "Regulation of gene expression in flux balance models of metabolism". Journal of Theoretical Biology 213 (1): 73–88.

Cukier, R.I., C.M. Fortuin, K.E. Shuler, A.G. Petschek and J.H. Schaibly (1973). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. Journal of Chemical Physics, 59, 3873–3878.

Edgar, Thomas F., and David Mautner Himmelblau. Optimization of Chemical Processes. New York: McGraw-Hill, 1988. Print.

Godfrey, K.R. and Fitch, W.R. 1984 “The deterministic identifiability of nonlinear pharmacokinetic models”, J. Pharmacokin. Biopharm., 12, 177-191

Khuri, André I., and Siuli Mukhopadhyay. "Response Surface Methodology." Wiley Interdisciplinary Reviews: Computational Statistics 2.2 (2010): 128-49. Print.

Kim, JY, YG Kim, and GM Lee. "CHO Cells in Biotechnology for Production of Recombinant Proteins: Current State and Further Potential." Applied Microbiology and Biotechnology 93 (2012): 917-30.

Kyparissidis, Alexandros. "Development of a Combined Mathematical and Experimental Framework for the Control and Optimisation of Mammalian Cell Culture Systems." Thesis. Imperial College London, 2012.

Legmann, R., Schreyer, H. B., Combs, R. G., McCormick, E. L., Russo, A. P., “A predictive high-throughput scale-down model of monoclonal antibody production in CHO cells.” Biotechnol Bioeng. 2009 Dec 15;104(6):1107-20.

Meshram, Mukesh, Saideh Naderi, and Brendan McConkey. "Population-Based Modeling of the Progression of Apoptosis in Mammalian Cell Culture." Biotechnology and Bioengineering 109-5 (2012)

Montgomery, Douglas C. Design and Analysis of Experiments. Hoboken, NJ: John Wiley & Sons, 2005. Print.

Myers, Raymond H. Response Surface Methodology. Boston: Allyn and Bacon, Inc., 1971

Nolan, Ryan P., and Kyongbum Lee. "Dynamic Model of CHO Cell Metabolism." Metabolic Engineering 13 (2011) 108–124

Nolan, Ryan P., and Kyongbum Lee. "Dynamic Model of CHO Cell Engineering." Journal of Biotechnology 158 (2012) 24– 33

Nolan, Ryan. "Dynamic Model of Chinese Hamster Ovarian Cell Metabolism in Fed-Batch Culture." Thesis. Tufts University, 2011.

Phair, Robert. "Chapter 1: The Rationale for Kinetic Modeling." BioInformatics Services, n.d. Web. 26 Dec. 2012.

Ravalico, J. K., 1H. R. Maier, 1G. C. Dandy, 2J. P. Norton and 2B. F. W. Croke, 2005 "A comparison of sensitivity analysis techniques for complex models for environment management" International Congress on Modelling and Simulation : advances and applications for management and decision making, Melbourne, 12-15 pp.2533-2539

Rodgers, S. T., 2009. A predictive high-throughput scale-down model of monoclonal antibody production in CHO cells. Biotechnol Bioeng. 104, 1107-20.

Rossi, Amy. "With Injectable Biologic Therapies on the Rise, Payers Face Tough Reimbursement Issues" Biotechnol Healthc. 2010 Summer; 7(2): 31.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D. Saisana, M., and Tarantola, S., 2008, Global Sensitivity Analysis. The Primer, John Wiley & Sons.

Schuster, S., Fell, D. A., Dandekar, T., (2000), "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks". Nat Biotechnol. 18, 326-32.

Sobol, I. M., Sensitivity Estivmates for Nonlinear Mathematical Models, Mathematical Modeling & Computational Experiment, 1, 407-414, 1993.

Unal, Resit, and Roger Lepsch. "Response Surface Model Building and Multidisciplinary Optimization Using D-Optimal Designs (1998)." American Institute of Aeronautics and Astronautics (1998).

Venkataramanan, Madhumita. "Nature.com." Spoonful of Medicine. Nature Medicine, n.d. Web. 26 Dec. 2012.

Wlaschin KF, Hu WS: Fedbatch culture and dynamic nutrient feeding. Cell Culture Engineering 2006, 101:43-74.