Use of Artificial Neural Network Techniques to Model Proteomics of Cellular Stress

Andrew James Barnett
BSc

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Master of Philosophy

This research programme was carried out in collaboration with University College London, LGC and Onyvax

May 2011

Abstract

Commercial scale bio processing is a major challenge when developing a whole cell biological vaccine. You must ensure that the end product has a sufficient level of potency to invoke an immune response within the host. Making sure that these key antigens remain on the cell surface during the manufacturing process is essential to make a viable product

Performing cellular stress experiments on an industrial scale would prove far too costly; instead an ultra-scale down model is used here to mimic the effects of a large scale industrial plant in a laboratory, for example to study the effect of hydrodynamic shear on cell membrane and cell surface makers (mimicking shear forces created in the mechanical pumps of the large scale system), by using a rotational shear device.

Here, bioinformatic modelling techniques, such as Artificial Neural Networks have been used to predict the levels of surface markers CD9 CD147 and HLA A-C based on experimental parameters of data obtained from ultra-scaled down experiments. These models where used to predict how changing the parameters would affect the density and abundance of the surface markers on the cell lines outer membrane. The surface markers CD9, CD147 and HLA A-C were chosen because of their relevance to the immune system and because of their strong expression on the surface of the P4E6 cells.

Cell culture conditions where varied systematically to identify an optimum set of conditions to minimise stress and obtain the maximum surface marker coverage. ANN’s where utilised to predict further combinations of different experimental parameters. ANN’s where also to be used to discover novel biomarkers for identifying pre cellular damage stress markers either for a specific cell line or a generic panel of biomarkers for multiple cell lines.

Overall it can be noted that the experiment was a success, from the results of this experiment it can be seen that Artificial Neural Networks can be used to model the proteomics of cellular stress. The mathematical modelling software can be used to predict the impact of changes to the experimental or cellular conditions to predict the MFI surface marker levels
These highly predictive models where then used to predict the levels of each surface marker at varying culture conditions From the data, the models accuracy can be improved dramatically by determining the initial volume of the cells surface markers.

The models indicated that the expression of surface markers CD9 and HLA A-C is not linked to cell stress, but can be linked to the cells passage number / generation time, a cell batch with a higher passage number appears to have a slower decrease in the respective surface marker under shear conditions. Whereas CD147 appears to be directly linked with cellular stress as the expression levels for this surface marker rise when stressed compared to a non-stressed cell culture.

Acknowledgements

With Thanks to:

Prof Graham Ball,	Nottingham Trent University
Prof Bob Rees.	Nottingham Trent University
Dr Jonathan Tepper,	Nottingham Trent University
Dr Christophe Lemete	Nottingham Trent University
Dr James Hind	Nottingham Trent University
Information Technology Technical Staff	Nottingham Trent University
Dr Kate Lawrence,	University College London
Dr Ryan McCoy,	University College London
Dr Mike Hoare	University College London
Dr Mike Delahaye	University College London
Dr Juan Pablo Costa	University College London
Dr Susan Pang	LGC, London
Dr Gavin O’Connor	LGC, London
Dr Steve Ward	Onyvax, London

Table of Contents

- **1. Introduction** 7
  - 1.1. Immune function and the immune system 7
  - 1.2 Commercial bio manufacture and bio processing 8
  - 1.3 Stresses arising from the cellular manufacturing process 9
  - 1.4 Potency Assays 10
  - 1.5 Markers 10
    - 1.5.1. CD9 11
    - 1.5.2. CD147 12
    - 1.5.3. HLA A-C 12
  - 1.6 Bioinformatic Modelling 13
  - 1.7 Artificial Neural Networks (ANN's) 15
    - 1.7.1. ANN Learning 17
    - 1.7.2. Advantages and disadvantages of ANN's 18
  - 1.8 Statistics 19
- **2. Method** 20
  - 2.1. P4E6 General cell culture 20
    - 2.1.1. P4E6 Sample Preparation 20
  - 2.2. Experimental Design for P4E6 20
  - 2.3. Analytical Techniques 20
    - 2.3.1. Analysis of Membrane Integrity by Trypan Blue Exclusion 20
    - 2.3.2. Membrane integrity by Propridium Iodide 21
  - 2.4 Cell Surface Marker Staining 21
  - 2.5. Data Pre Treatment 21
  - 2.6 Artificial Neural Networks 21
    - 2.6.1. Architecture 21
    - 2.6.2. Artificial Neural Network Development 22
      - 2.6.2.1 Initial Model Development – Phase 1 22
      - 2.6.2.2 Initial Model Development – MiTB 22
      - 2.6.2.3 Initial Model Development – Phase 2 22
      - 2.6.2.4 Model Interrogation and Predictions – Phase 1 23
      - 2.6.2.5 Model Interrogation and Predictions – Phase 2 23
- **3. Results Model Development** 25
  - 3.1. Model Development, Phase 1 25
    - 3.1.1. CD9 25
    - 3.1.2. CD147 29
    - 3.1.3. HLA A-C 33
  - 3.2. Model Development, Phase 2 37
    - 3.2.1. CD9 37
    - 3.2.2. CD147 42

- 3.2.3. HLA A-C 46

4. Results Model Predictions 51

- 4.1. Model Predictions, Phase 1 51
  - 4.1.1. CD9 51
  - 4.1.2. CD147 55
  - 4.1.3. HLA A-C 59
- 4.2. Model Predictions, Phase 2 63
  - 4.2.1. CD9 63
  - 4.2.2. CD147 64
  - 4.2.3. HLA A-C 66

5. Discussion 68

- 5.1. Model Development, Discussion 68
  - 5.1.1 CD9 68
  - 5.1.2. CD147 69
  - 5.1.3 HLA A-C 70
- 5.2 Model Predictions, Discussions 71
  - 5.2.1. CD9 71
  - 5.2.2. CD147 72
  - 5.2.3. HLA A-C 73
- 5.3 Secondary analysis and model interpretation discussion 74
  - 5.3.1. Hold Time 75
  - 5.3.2. Passage Number & Generation Time 75
- 5.4 Effect of High / Low initial marker levels on model outcome 76
  - 5.4.1. CD9 76
  - 5.4.2 HLA A-C 76
  - 5.4.3 CD147 76

6. Conclusion 77

7. Future Work 79

8. References 80

1. Introduction

The initial biological data generation was conducted on data from Dr Ryan McCoy (Onyvax) using the Onyvax P4E6 cell line.

1.1 Immune function and the immune system

It is well documented that when a foreign organism gains entry into a host, the host will react with some form of immunological response, in order to destroy the foreign invader before it can do any harm. However, some malignant cancer cells they down regulate the host's immune system causing the immune response to be insufficient, allowing the tumour to grow and spread with little resistance from the host.

There are many different types of therapeutic cancer vaccine in development, either using whole cells, parts of membranes, genetic material or proteins from within cancer cells or peptides to invoke an immune response. Each type of vaccine must have a pre-defined minimum level of potential biological activity to be effective again the tumour it is targeting and this level needs to be constant from batch to batch. Normally this is tested for using potency assays, and this is usually tied in with the vaccines efficacy. For whole cell vaccines this can involve measuring densities of specific antigens on the cells surface.

Dendritic cells have developed intense interest in the cancer vaccine field for their ability to stimulate immunogenic responses. They are extremely specialised in producing molecules that facilitate antigen presentation in T Cells that are tumour antigen specific. This can be beneficial to cancer patients as the tumours may be supressing the hosts immune system. (Mosca *et al*, 2007). The cells can be grown in-vivo and in-vitro, however the way these cells are developed can mean they have different properties, for example, dendritic cells are arranged in a star shape when in situ, but, when spun and analysed, they are seen to display long processes that shrink, grow and bend (Banchereau & Steinman, 1998). Dendritic cells can be cultures in various media, however due to safety concerns related to the use of bovine blood in certain media dendritic cells are being grown in serum free media, to make the cells safe for clinical application (Mosca *et al*, 2007).

Many different molecules for tumour treatment can be loaded onto dendritic cells, not just the tumour antigens but cell lysates, peptides and proteins. Loading the dendritic cells with a tumour specific peptide has its advantages for use as a vaccine, as these can be produced in great quantities, and their sequences can be easily adapted to improve immunogenicity.

Their use can lead to a specific and defined immune response, usually by stimulating production of CD4, (Knight *et al*, 2002), which can be easily monitored by loading the dendritic cells with unfractionated secretions that contain antigenic receptors (Mosca *et al*, 2007)

1.2 Commercial bio manufacture and bio processing

Bio-processing is the production of biological products for a purpose, either for medical treatment, research or commercial activities. Refinement and mass production of these biological products on a mass scale can reduce costs but tight controls must be kept on the manufacturing process otherwise entire batches of the product can be useless and thus worthless. For example Hartmann (*et al*, 2008) for the production of self-assembling peptide based biomaterials; it required their team to be mass produced at low cost to make them viable in potential markets. To achieve this high production low cost a full scale commercial bio processing and bio manufacture system must be set up.

Commercial scale bio processing is a major challenge when developing a whole cell biological vaccine to ensure the end product has a sufficient level of potency to invoke an immune response within the host. This is made even more complex due to the many and somewhat unknown combination of presented antigens on the cell surface needed to trigger the response. Making sure that these antigens remain on the cells surface as well as the desired target antigens of the disease you want to immunise against makes the process more complex.

Small scale devices, graphical representation, modelling tools and regime analysis can be used to predict the quality of the final product, needing only fractions of the resources needed in a full scale manufacturing setup. The use of the smaller quantities of materials means that the initial design process can be carried out quickly and accurately, and allows for adjustments in the experimental procedure with minimal loss of the end product and materials (Chhatre *et al*, 2011).

Performing these experiments on an industrial scale would prove far too costly and wasteful of materials. An ultra-scale down model, in a laboratory, is used to mimic the effects of a large scale industrial plant in a laboratory, for instance to study the effect of hydrodynamic shear on cell membrane integrity and cell surface makers, by using a rotational shear device to mimic shear forces created in the mechanical pumps of the large scale system (McCoy *et al*, 2009).

1.3 Stresses arising from the cellular manufacturing process

Mammalian cells, due to their size and complexity are more likely to sustain damage during the manufacturing process. It is important that the cells are not subjected to hydrodynamic shears and that the cells membrane remains intact allowing it to present the desired surface proteins to invoke an immune response, making the vaccine effective (McCoy *et al*, 2009). If the fluidity of the lipids within the membrane of the cell is altered due to damage, this can have a knock on effect on transport activities through the membrane affecting the cells internal chemical balance causing some enzyme related functions to cease This can lead to an alteration of the cells surface antigen profile (Benkö *et al*, 1986). Further damage can occur to the cells within the bioreactors where the cells are constantly mixed to keep them floating in suspension to ensure they have enough nutrients (McQueen *et al*, 1987). Cellular stress can have an accumulative effect within the manufacturing process. When an already stressed cell is placed into a new environment (for example the movement from one bio-reactor to another) with new stresses, the damage to the cell is likely to be exponentially increased.

From studying the literature there are very few if no papers in the literature on the study of these biomarkers under rotational stress conditions, Motobu (*et al*, 1997) studied the effect of shear stress on Chinese hamster ovary cells, in particular the effect of low shear stress for cell cultivations in bioreactors. It was noted that with increasing the exposure to stress, and the time that the cells are subjected to it, the morphology of the cell was changed significantly. However, cell viability showed a noticeable decrease in viability in cells cultured in a rotational flow bioreactor compared to static cultures.

Rotational shear stresses can also impact on the nutrient balance within a bio-reactor Belfiore (*et al,* 2009*)* suggested that changes in the rotational shear stress applied to mammalian cells can affect the binding proteins used to anchor the cells. If these binding proteins are constantly being replaced, the cell would have to take up an increased amount of nutrients from the surrounding medium, exhausting its available supplies faster. Knowing the cultures media requirements as well as an indication of the condition of the end product would be extremely advantageous during the culture process. Zhang (*et al* 2008) examined the impact of fluid dynamics and shear effect on protein fusion and plasmid DNA for use in drug discovery and development. By using a centrifuge and an Ultra scale down device they were able to mimic the shear damage for mechanical pumps used in a full scale bio manufacture process and what effect this had for downstream recovery.

Under laboratory conditions batch-to-batch variation needs to be taken into account, as within commercial scale bio processing a slight change in the cell line before cultivation could mean a big change in the cells characteristics, which could reduce the vaccines efficacy (Petricciani *et al,* 2006). The technical approaches to therapeutic cancer vaccine development include those based on several different components: whole tumour cells; tumour cell lysates; proteins; peptides; and the genetic sequence that codes for specific tumour cell proteins (Petricciani *et al,* 2006).

1.4 Potency Assays

The World Health Organisation defines potency as tests that "measure biological activity of a vaccine but do not necessarily reflect the mechanism of protection in humans".

Once a whole cell vaccine has been developed, and has shown that the vaccine can evoke an immune response in the host a potency assay must be performed. The potency assay is usually carried out to ascertain the key things that a vaccine must have in order to be effective in a clinical trial; to reach a pre-defined minima of biological potential activity to a given result (such as an increase in a host immune response to a tumour) and to show batch to batch consistency (Petricciani *et al,* 2006). If the potency of a vaccine can be determined in a numerical figure, theoretically a bio-informatic algorithm could potentially be used to predict this value based on factors from the manufacturing process.

1.5 Markers

The rationale behind using CD9, CD147 and HLA a-c was in part because of their relevance to the immune system (as P4E6 was to be part of a vaccine) and because of their strong expression on the surface of the P4E6 cells. Prior to any work or trials being carried out on

the cells Onyvax performed a screen of over 20 surface proteins potentially expressed by the cells and these three were among the highest for expression. These
markers (and three others) were used as indicators of quality control throughout the research and clinical trial process.

1.5.1 CD9

The protein CD9 is a member of the tetraspanin family. This protein responsible for several key cellular functions including adhesion, motility and cell membrane fusion (Hori *et al*, 2004), and is also responsible in part for the regulation of functions such as cell migration, cell differentiation, cell growth and also plays a key role in reproduction. The tetraspanin family is part of the integral membrane proteins group. This group achieves the above processes by changing the way the cell membrane behaves to allow for cell interaction when attaching to adhesion molecules on other cells (Fan *et al*, 2010). CD9 is also required for sperm and egg fusion to produce offspring. This was tested by Miyado (*et al*, 2000), who produced mice lacking in CD9. They discovered that even though the mice were born healthy, the female mice lacking the CD9 protein produced a litter of less than 2% of the size of regular mice that had the CD9 protein intact. This was discovered to be down to the lack of fusion between the sperm and the egg, when artificially fertilised (in vitro) the eggs developed and viable offspring were produced.

One of the prognostic indicators within a number of lymph node metastasis, breast and pancreatic cancers has been the expression levels of the protein CD9, the protein itself is comprised of two 20 – 28 and 76 – 131 extra cellular amino acids that interact with other extra cellular proteins of other cells (McCoy *et al*, 2009). High levels of CD9 expression is often observed in lymph node metastasis and vessel invasion, where the higher the level, the higher the malignancy status in some cancers such as gastric cancer (Hori *et al*, 2003). However Zvieriev (*et al*, 2005) stated that some cancers down regulate the expression of CD9 and that a high expression of CD9 within the body is considered a good sign in the prognosis of head and neck cancers, compared to that of prostrate and gastric cancers. In some cases, the over expression of this protein can indicate a reduction in the malignancy for the cancer being treated, however it can also cause an increase in the malignancy of different types of cancer in other regions of the body if present.

1.5.2 CD147

This protein is a member of the type 1 transmembrane family and has been associated with human pathogen infection, inflammatory diseases and cancerous tissues. CD147 is found in all living tissues, but the elevated levels in tumour tissue means it is now being used as a prognostic indicator (Redzic *et al*, 2011).

The CD147 surface marker is expressed on hemopoietic (blood cell forming) nonhemopoietic cells, blood cells, leukocytes, endothelial and platelets (Chiampanichayakul *et al*, 2006). The marker is used within the cells to regulate multiple protein families, and is unique as it can function as part of a cell but can remain active once released from the cell (Schlegel *et al*, 2009). The marker is often shown to be present in high expression levels in human tumour cells, making this an ideal localised target for a cancer vaccine as the high expression rates would provide a greater chance for the antibodies stimulated by the vaccine to attach to the tumour cells (Chiampanichayakul *et al, 2006*).

CD147 is thought to enhance tumour invasiveness and growth after stimulating the production of extra cellular metalloproteins (Schlegel *et al*, 2009). When the protein is knocked out, the invasiveness of the tumour within the host can be reduced, giving the patient a better chance of survival.

However, the downside of using this marker is that it is also critical in retinal function and foetal development (Iacono *et al*, 2007), so its use in a widely used vaccine may cause pose significant health risks. Both CD9 and CD147 have both been shown to be commonly expressed in many prostate cell lines (McCoy *et al*, 2009). Redzic (*et al*, 2011) stated that CD147 has two primary isoforms related to cancer progression known as the ig0 domain, one that is secreted by all cells and one linked to cancer, these are known as Ig1-Ig2 and Ig0-Ig1-Ig2.

1.5.3 HLA A-C

HLA (or human leukocyte antigen) plays a crucial role within the hosts cell protein degradation pathway, presenting broken down proteins to the hosts' immune cells to invoke an immune response (McCoy *et al*, 2009). HLA markers are crutial in an immune response to tumour cells by the hosts cytotoxic T-Cells. Any tumour cells that have lost their HLA antigens are less susceptible to attack and lysis by the hosts immune T-Cells (Jager *et al*, 2002). In humans, Ramal (*et al* 2000*)*, stated that tumours generally have a show reduced or no expression of the gene that codes for the HLA antigen, allowing the tumour to grow and invade the host and the immune systems T Cells fail to recognise the tumour causing cells.

HLA has also been identified as a major gene in the progression of Type I diabetes. It has been identified that the HLA class I molecules are used by cytotoxic T Lymphocytes as a targeting molecule causing a prolonged immune response resulting in the destruction of the pancreatic cells responsible for insulin production. HLA class II molecules are also thought to provoke the initial immune response (Noble *et al*, 2002)

1.6 Bio-informatic modelling

Here we use bio-informatic modelling as a tool to attempt to understand the effect of multiple cellular stress factors on cell integrity and cell surface marker presentation for use in the creation of a whole cell cancer vaccine. The use of the bio-informatic modelling can be used to examine the outcomes in regards to the surface marker profile of the cell by altering external stresses.

NCBI (National Centre for Biotechnology Information) quotes “Bio-informatics is the field of science in which biology, computer science, and information technology merge to form a single discipline” (NCBI, 2007). Bio-informatics is a rapidly developing in modern biological sciences, a major increase in the use of genomic and proteomic technologies, such as the use of gene micro arrays and mass spectroscopy. The use of this type of analytical equipment can produce highly dimensional and complex data sets that can be up to gigabytes of information in one single output file. For example Mass Spectrometry (MS) utilising Matrix assisted laser/desorption ionization (MALDI) and Time of Flight (TOF) generates a spectrum according to the time of flight of an ion (from its source to a mass analyser, looking at the mass to charge ratio of the ions crated from the sample by the laser), these spectrum can contain hundreds of thousands of points each containing data on a mass per sample analysed (Lancashire *et al*, 2009).

Due to the extra complexities and sheer file size of this new approach to analysis, new computational approaches in both hardware and software have had to be developed to deal with these complexities.

Bio-informatics is used primarily in the novel biomarker and drug discovery market, but is also beginning to play a role in "personalised medicine" to aid in disease prevention by tailoring the treatment to the patients' physique (Gatto, 2003).

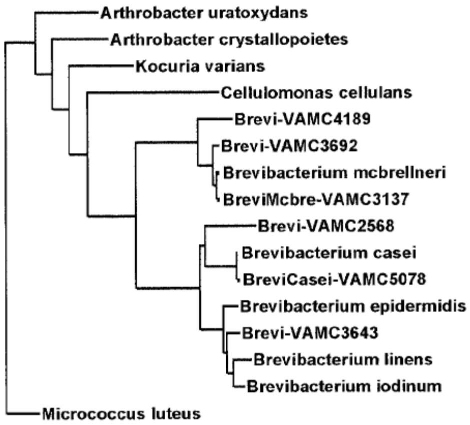
There are several different methods used currently in bio-informatic modelling for analysis of data; Artificial Neural Networks (ANN's), Support Vector Machines (SVM), Random Forest Trees (RFT), Clustering, Principal Component Analysis (PCA) and Decision Trees (DT).

Support Vector Machines (SVM) is based on a principal from computational learning theory called Structural Risk Minimisation which contains a large class of neural networks, radial basis functions and polynomial classifiers and is shown to follow a linear structure when used to analyse highly dimensional data (Huang *et al*, 2004). Working via a binary classification algorithm that separates samples into positive and negative groups, SVM's can separate linear and non-linear data prior to analysis. However SVM analysis is very complex can be very time consuming, especially with large data sets, which can be memory intensive (Ball, Al-Shahib, 2010).

The Random Forest method can produce several classification trees during the analysis process which are then combined into one single tree, this single tree provides a more accurate prediction than the individual components of the model (Peters *et al*, 2007), by using the mode of the output class prediction. Random Forest method is very fast and is particularly useful when dealing with noisy data (Ball, Al-Shahib, 2010).

Hierarchical clustering is often used when analysing gene expression data. This method arranges the data samples in a tree like structure where items with similar characteristics are grouped closely together. Those items with a completely different profile are grouped further away. Cluster analysis is well suited to classifications of micro-organisms and identifying structures in populations (shown in fig 1), however the main drawback with using this analysis is that the algorithm cannot handle highly dimensional data (Ball, Al-Shahib, 2010).

Figure 1: An Example of clustering



Principal Component Analysis (PCA) is a common analytical method used on data generated through Mass Spectrometry to visually represent data and reduce levels of dimensionality. PCA is used to explain levels of variance of the inputs within a data space by assigning an Eigen value. The Eigen value is calculated by transforming the input space into a new space using principal components, expressing these new values as linear combinations of the original data and then drawing a line through the data at a point that explains the most variation (Ball, Al-Shahib, 2010).

Clustering can be further analysed using the Decision Tree methodology in order to sort members of a population into groups after applying either numerical or logical rules to the model and can be applied to both regression and classification models.

1.7 Artificial Neural Networks

Artificial Neural Networks, or ANN's are statistical packages, or programming code either contained within a standalone machine, or over a series of machines linked together to form a network cluster controlled by a server. The design of the ANN is a computational representation of the human brain using a series of hidden layers or 'neurones' that are assigned weights that can influence the output of the neurone

.

Figure 2: example of an ANN network

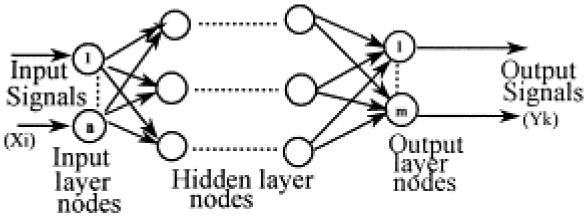
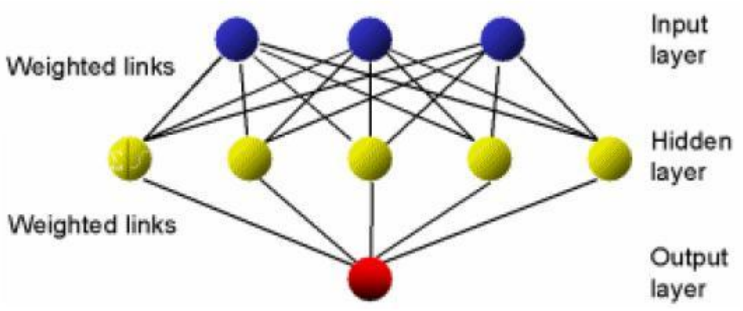


Image taken from Dutta *et al,* Applicability of the modified back-propagation algorithm in tool condition monitoring for faster convergence.

The Artificial Neural Network system can be applied to find influences within environments on different species or predict if a cell will turn cancerous under different conditions. ANNs have been used to predict specific outcomes for specific subjects by learning patterns of outcomes taken from the results of tests performed on the subjects under similar conditions. Several parameterisations have been developed for the use of ANNs, the main use is the identification within the models of the strength of certain influences of the inputs on the trained model (Lancashire *et al,* 2009).

ANNs are based loosely on the learning patterns, processes and neuron structure of the brain. Within the ANN there are many different types of mathematical analysis such as radial base function and multi-layered perceptron (MLP). The MLP is organised into layers where each layer has a manually assigned number of 'neurons' to form an input layer where details are added into the model, several hidden layers within the model and the output layer where the predictions of the model are displayed. The number of layers and neurons within the "hidden layer" can be used to aid in predictions by assigning each neuron of each layer a 'weight', to influence the predictions of the model as the data is passed through each of these hidden layers. (Rumelhart *et al,* 1986).

Figure 3: Example diagram of the structure of a multi-layered perceptions ANN



A basic version of an Artificial Neural Network is the Back Propagation Algorithm (BPA), which works solely on the relationship between the input data and the output value and has the ability to model very complex data and the relationships between the inputs. The BPA works on a layered feed forward network where information can move in a unidirectional way between successive layers until the output layer is reached. The weighted links within a layered network are key to the ability of the model to create predictions, as they create the link between an input and the output value of the model and an alteration of one of the weighted links within the model can drastically alter the output value. Once the weights within the trained model where determined the relative importance of the input variables (and how they can influence the model) can be identified (Ball *et al*, 1999).

1.7.1 ANN Learning

As with human neurological development, the ANN does not know initially how to analyse the data presented to it, therefore would need to be trained to analyse the data to produce an output of the desired accuracy. In order for the model to be trained the model must undergo either one of two major learning paradigms, supervised or unsupervised learning. In the case of supervised learning the model is presented with known input and output data. For example to differentiate between cases of different subgroups a Boolean value is assigned to each case and then the model will use the inputs to predict the Boolean value. The model then corrects the weights (which are initially randomised) within its hidden layers by feeding back into the system an error value (such as the mean squared error) by comparing the models output to the true output of the data so after each epoch the error decreases until there is no significant reduction in the error value, or the error value starts to increase, known as a back propagation algorithm (Rumelhart *et al, 1986*).

Unsupervised learning is the reverse of supervised learning where the model must make its own decisions on how to initially group the data and then analyse it, for example Principal Component Analysis (PCA) (Lancashire *et al*, 2009).

1.7.2 Advantages and disadvantages of ANN's

Due to the ways that the ANN's analyse data, they are capable of handling large complex data sets that can contain missing sections of data, or data points that contain a lot of noise. Unlike linear models, they can also analyse data that contains non linear interactions between inputs. Increasingly ANNs are being used to predict "real world" trends basing future cases on patterns that have been seen within the training and test data sets (Lancashire *et al*, 2009).

However, if the model is trained excessively without stopping parameters, the ANN models can be over-fitted to the data, becoming very good at predicting results for the individual data set, but loosing the ability to predict accurately any future cases within the same study. However, the problem of over fitting can be solved by using a regularization method called "early stopping". This method can halt the training algorithm once either a set number of epochs have been reached, or when there is no further improvement in the error of the model within the test and validation sections of the data (Plumb *et al*, 2005).

The amount of time taken to process the data can vary depending on the hardware and level of complexity of the data, to counter the time issue, hardware must be upgraded and this can be costly. Also the quality and reproducibility can be a negative factor with ANN's, for example if the data has high background noise the ANN may have issues in distinguishing markers of interest from erroneous results (Lancashire *et al*, 2009). Also using a cross validation approach where each of the inputs of the training, test and validation sub sets of data are randomly assigned for each sub-sample, this is known as the Monte Carlo resampling method and can allow comparisons to be drawn between models. Bootstrapping is also very useful for estimating the error of models, but this method creates many replicates from the data set rather than re-analysing subsets (Lancashire *et al*, 2009).

1.8 Statistics

Sensitivity analysis can aid in determining which parameters are the key drivers of a mathematical model's results. Saltelli (*et al*, 2002) explains a sensitivity analysis looks at the relationships between a series of inputs into a model and its. Such an analysis is relevant for a number of practices, including quality assurance of models and codes, and is useful in the identification of crucial regions within mathematical parameters.

R Squared is a statistical term saying how good one set of inputs are at predicting another. An R Squared value can range between 0 and 1, with 1 being a perfect prediction and 100% accurate.

2. Method

Please note method steps 2.1 to 2.4 was performed by researchers at University College London data from these experiments were collated and tabulated then presented to Nottingham Trent University for analysis.

2.1 P4E6 General Cell Culture

P4E6 (Onyvax propriety cell line) cells were grown in keratinocyte serum free medium. Sub-culturing was carried out by the removal of growth medium followed by one wash in Dulbecco's phosphate buffer. Cells were centrifuged at 500 g for 3 minutes, supernatant discarded and the pellet was re-suspended in fresh growth medium. Cell counts where performed and passage number noted.

2.1.1 P4E6 Sample Preparation

Once the desired number of cells for experimentation was obtained, the cells were harvested, the supernatant was removed. Holding times for cell suspensions pre-exposure to shear were between 2.5 and 3.5 hours.

2.2. Experimental Design for P4E6

P4E6 cells were sheared in the rotating disc shear device at either low 4000 rpm, medium 5000 rpm or high 6000 rpm. Shear rates for 180 min or held statically as a non-sheared control. Samples were taken every 45 minutes and analysed using light microscopy techniques or stained using propidium iodide (PI) or frozen at -80C for cell surface marker staining. Sample analysis for every time point and each assay was conducted in triplicate

2.3. Analytical Techniques

2.3.1. Analysis of Membrane Integrity by Trypan Blue Exclusion

Intact and non-intact cells were counted using the trypan blue exclusion method, applied to a haemocytometer and counted under a standard light microscope.

2.3.2. Membrane Integrity by Propridium Iodide

P4E6 Cell line was dispensed into 96 well plates, a binding buffer was added to each well and the plate was left on ice, in the dark for 20 to 25 minutes. Samples were then re-suspended in binding buffer for analysis.

2.4 Cell Surface Marker Staining

Samples were removed from storage at –80^{o}C and defrosted in a water bath set at 37^{o}C for two min. Cells were washed and left on ice, in the dark for 20 min

2.5. Data pre-treatment

The data was tabulated into an excel file, an initial marker value was then calculated from the average of the time point zero Mean Fluorescing Index maker values for each surface marker per sample

$$\frac{\text{(Marker MFI Experiment x, triplicate A)} + \text{(Marker MFI Experiment x, triplicate B)} + \text{(Marker MFI Experiment x, triplicate C)}}{3}$$

2.6. Artificial Neural Networks

2.6.1. Architecture

A custom network model was created in Statistica (Statsoft, www.statsoft.com) where the ANN MLP was used to predict the MFI surface markers based on a series of inputs from physical parameters of the experiment, and an initial MFI marker value.

The multi-layer perceptron of the model was instructed to train for 1000 epochs, with a cycle value of 3000 events, learning rate of 0.1 and a momentum of 0.01, with a total of 5 hidden neurons. The model was randomly sub sampled 50 times, with 60% of the data used for training, 20% for test and the remaining 20% used for validation. This process is termed Monte Carlo Cross validation or Random Sample Cross validation. This was used to create each model in the initial model development phases.

2.6.2 Artificial Neural Network Model Development

2.6.2.1 Initial Model Development – Phase 1

The initial model was developed using the criteria above. Utilizing the stepwise technique a model starting with the inputs of shear time and shear speed was created. In this stepwise technique markers were added in a sequential fashion and the improvement in predictive monitored. All inputs were assessed at each step and a panel containing the optimal set of markers developed

Each ANN model of each step was created, prediction tables (the values predicted by the ANN) where calculated for the 50 random sample cross validation models, the average prediction value over the 50 models was taken and plotted against the observed value from the flat file. An R squared value was then calculated to indicate the accuracy of the model and the amount of variance explained. The model was then introduced to an additional input; size of cell, once the results from this model was generated another input was added to the model, Mean Fluorescing Index (MFI) IgG. The next input to be added to the model was “experiment” referring to the experiment letter. This was repeated for each of the surface markers CD9, CD147 and HLA A-C.

2.6.2.2 Initial Model Development – MiTB

A model was created with the inputs of shear time, shear speed, size of cell, MFI IgG and experiment. Once the ANN model was created, prediction tables (the values predicted by the ANN) where calculated for the 50 random sample cross validation models, the average prediction value over the 50 models was taken and plotted against the observed value from the flat file. An R squared value was then calculated to indicate the accuracy of the model

2.6.2.3 Initial Model Development – Phase 2

The model was developed by initially starting with the inputs shear speed, time and the initial marker value for the surface marker being studied, again the initial surface marker value was used as a baseline correction. The stepwise technique was utilized again to add further inputs of hold time, passage number, generation time and doubling time. An R squared value was then calculated to indicate the accuracy of the model.

A sensitivity analysis was performed for each model in phase 2 and the results tabulated to indicate which input had the greatest influence on each model per marker.

After the results from model development phase 1 were tabulated, the model with the highest r squared value was then used to predict levels of each surface marker at varying time points, shear device speeds, hold times, generation time and passage numbers. This data was then graphed using a scatter plot to show the variation in the marker level over the time period

2.6.2.4 Model Interrogation and Prediction – Phase 1

The optimized model containing the inputs shear speed, shear time, passage number, generation time, initial surface marker MFI value and hold time was interrogated to predict marker values at varying input values and culture conditions, to see if changes in experimental conditions would have an impact on the overall surface marker profile of the cell after being subjected to shear stress. The model was asked to predict values for each surface marker at 0, 4000, 5000 and 6000 RPM at fixed time points 0, 60 and 120 minutes for a hold time of 1 and 3 hours with a fixed value for passage number (59), generation time (13) and doubling time (60 hours) over 50 random sample cross validation models, the average prediction value over the 50 models was taken and plotted over a time line

Next the model was asked to produce a series of MFI value predictions for each surface marker at 0, 4000, 5000 and 6000rpm at o, 60 and 120 minutes, with fixed values for hold time (2 hours) and doubling time (60 hours). In this model the passage number was changed to 66 and the generation time to 18. 50 random sample cross validation models were created, the average prediction value over the 50 models was taken and plotted over a time line. Another model was then created with the same fixed values as above but with a passage number of 63 and a generation time of 15. 50 random sample cross validation models were created, the average prediction value over the 50 models was taken and plotted over a time line.

Finally another over 50 random sample cross validation models were created with the variables for generation time set to 11 and the passage number set to 58.

2.6.2.5 Model Interrogation and Predictions – Phase 2

Phase 2 was designed to develop and interrogate models that indicatd how the initial surface marker profile of the cell prior to stress could affect the surface marker profile after stress. 50 random cross validation models were created for each of the surface markers with a fixed value for hold time (2 hours), generation time (13), passage number (59) and doubling time

(60 hours). The initial surface marker value was replaces with the highest or lowest value for each surface marker and the results averaged and plotted over a time line. The data was then interrogated to determine any differences between the values produced by each model

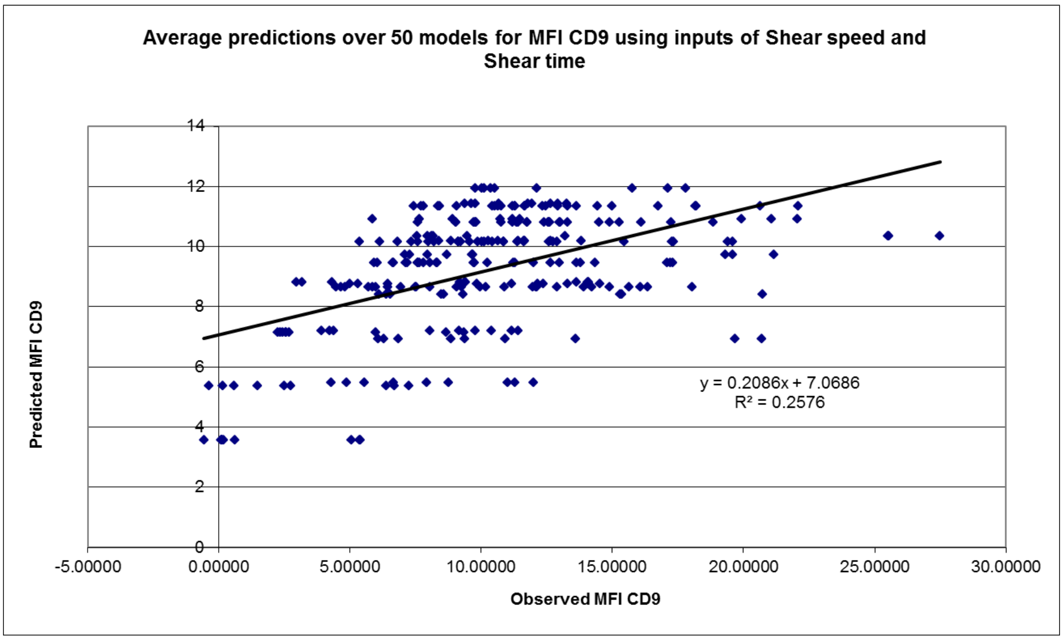
3. Results Model Development

3.1. Model Development phase 1

3.1.1. CD9

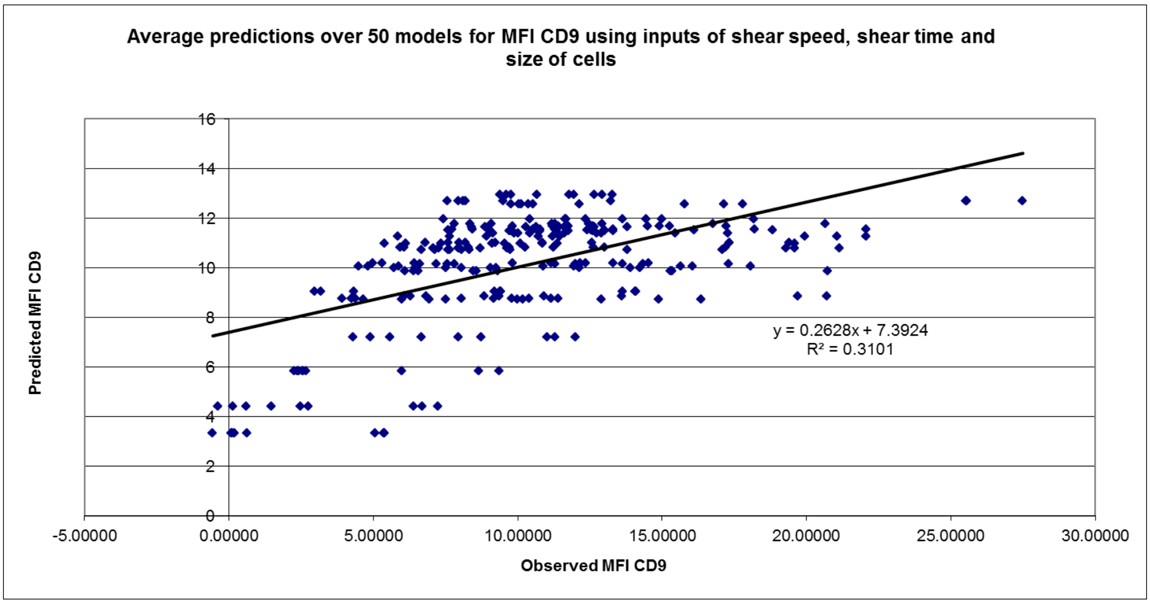
After the initial model using the two inputs of speed and time, an average was taken over the 50 random sub sampled models (including test, training and validation data) and the average value was plotted against the actual value from the data, shown in Graph 1. An R squared value was taken from the plotted values.

Graph 1: Graph showing the actual v averaged predicted MFI CD9 using the inputs of "Shear speed (rpm)" and "shear time (min)"



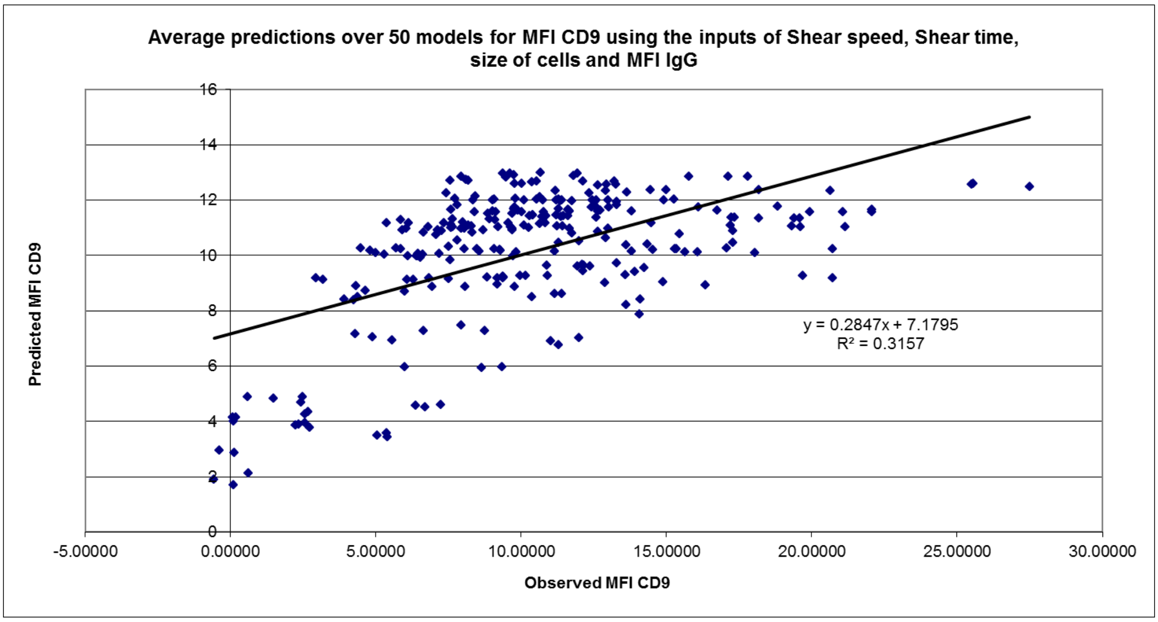
Initially model step 1 for CD9 utilised only inputs for shear speed and time. This model did not perform well showing a weak relationship between actual and predicted Graph 1 having an R squared value of 0.257. The equation of this graph indicated a gradient of 0.21, a strong deviation away from the expected gradient of 1. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 7.06 indicated predictions were systematically higher than they should be. This model was further modified by the addition of the size of cell parameter.

Graph 2: *Graph showing the actual v averaged predicted MFI CD9 using the inputs of "Shear speed (rpm)","size of cell" and "shear time (min)"*



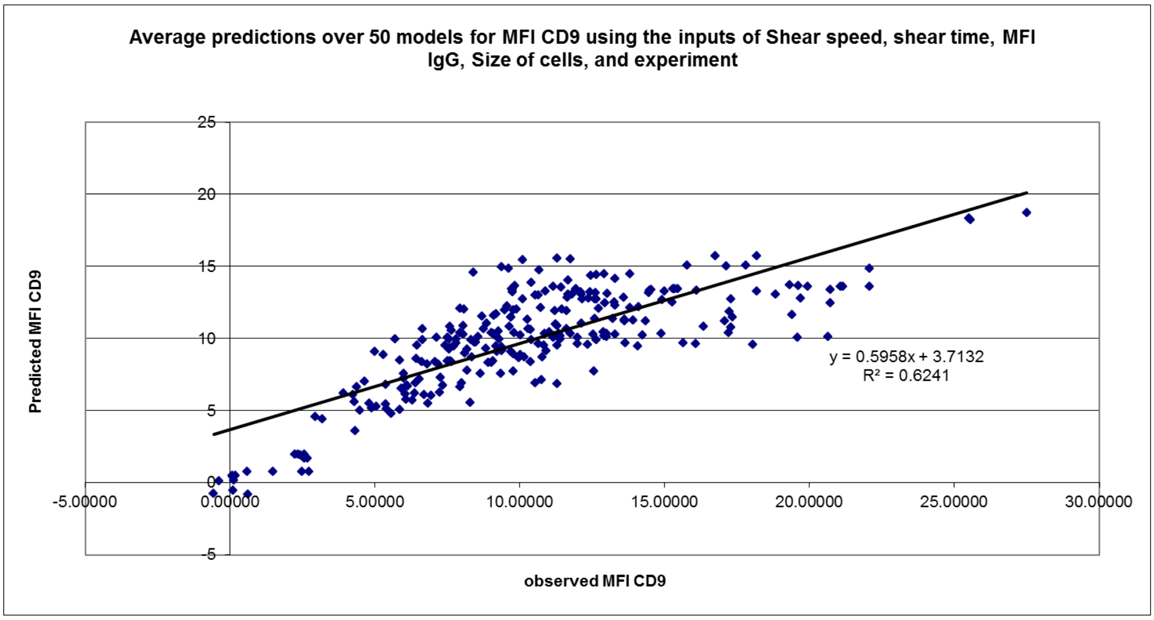
Model step 2 for CD9 utilised only inputs for shear speed, shear time and size of cells. This model also showed a poor performance with a weak relationship between actual and predicted. This resulted in the line of best fit having an R squared value of 0.3101 (Graph 2). The equation of this graph indicated a gradient of 0.26, a strong deviation away from the expected gradient of 1. This again showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 7.39 indicated predictions were systematically higher than they should be. This model was further modified by the addition of the MFI IgG parameter.

Graph 3: *Graph showing the actual v averaged predicted MFI CD9 using the inputs of "Shear speed (rpm)","size of cell", "MFI IgG" and "shear time (min)"*



Model step 3 for CD9 was created using the inputs of shear speed, shear time, size of cells and MFI IgG. This model did not perform well showing a weak relationship between actual and predicted Graph 3 having an R squared value of 0.3157. The plots on this graph indicated a gradient of 0.28, again a strong deviation away from the expected gradient of 1. This model also showed an under prediction at the high end of the range and an over prediction of values at the low end of the range. The plots gave an intercept of 7.17; this indicated predictions were systematically higher than they should be. To try and improve the predictions of this model it would be further modified by the addition of the experiment parameter.

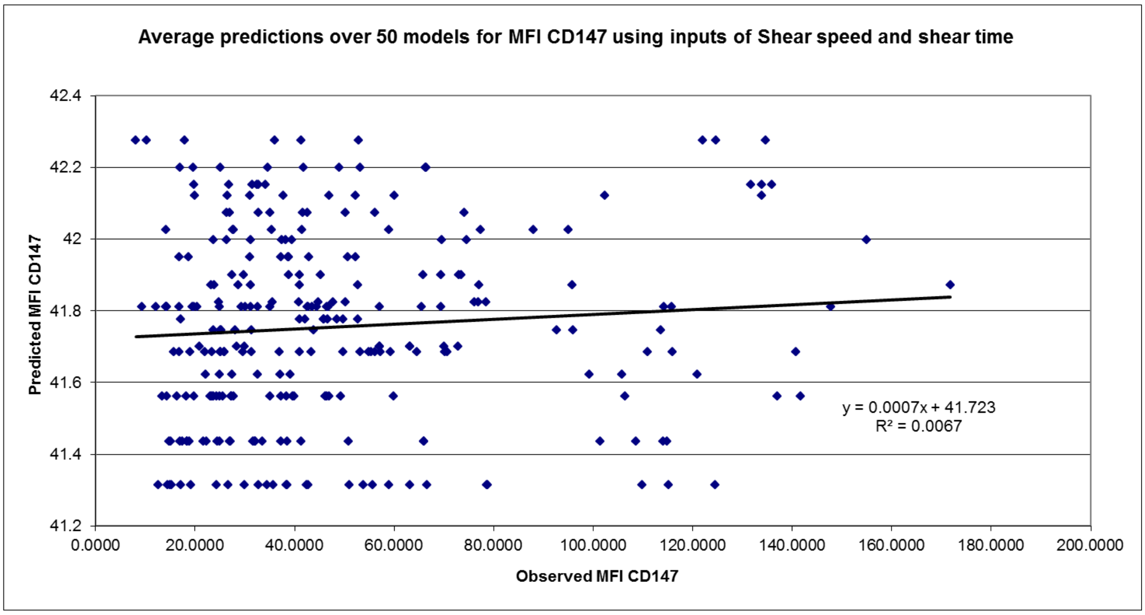
Graph 4: *Graph showing the actual v averaged predicted MFI CD9 using the inputs of "Shear speed (rpm)","size of cell", "MFI IgG", "experiment" and "shear time (min)"*



Model step 4 for CD9 utilised only inputs for shear speed, shear time, size of cells, MFI IgG and experiment. This model performed better showing a stronger relationship between actual and predicted Graph 4 having an R squared value of 0.6241. The equation of this graph indicated a gradient of 0.60, showing a smaller deviation away from the expected gradient of 1 than the previous step. This showed less of an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 3.71 indicated predictions were higher than they should be. At this stage the addition of this input resulted in a better explanation of variation.

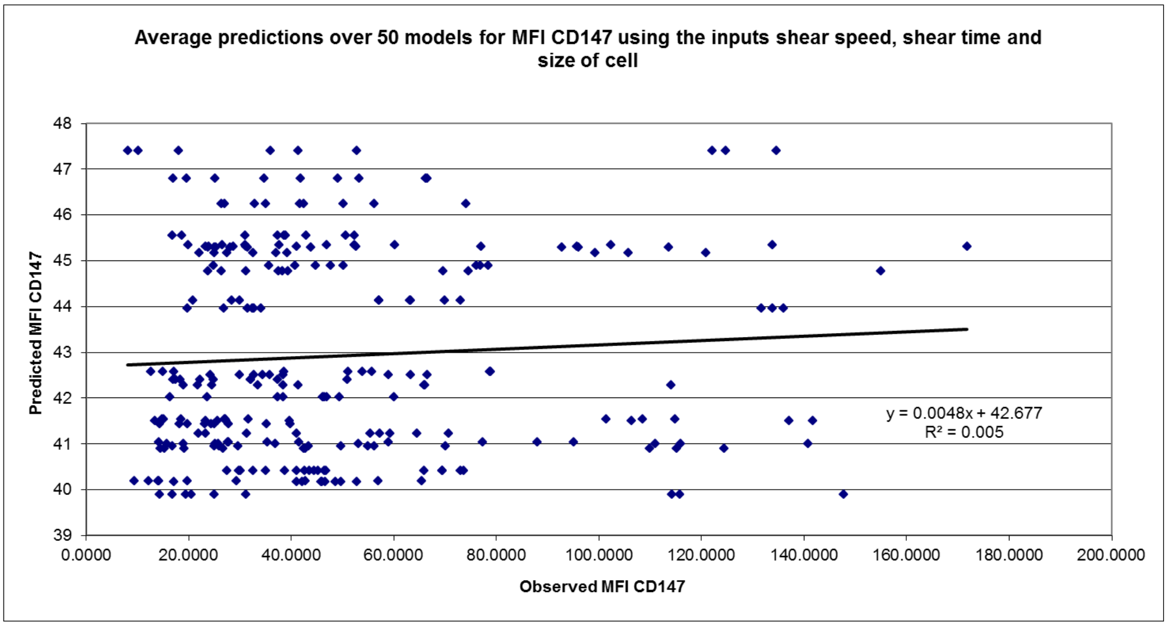
3.1.2 CD147

Graph 5: *Graph showing the actual v averaged predicted MFI CD147 using the inputs of "Shear speed (rpm)" and "shear time (min)"*



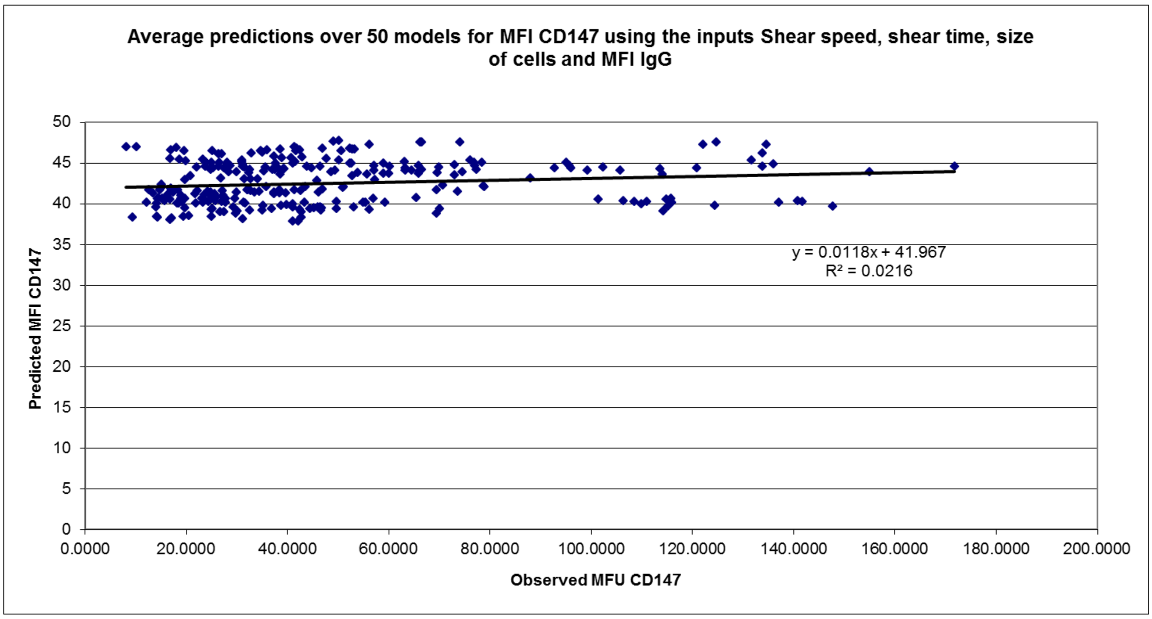
Initially the model created for step 1 CD147 utilised only inputs for shear speed and time. This model showed an extremely poor relationship between actual and predicted values (Graph 5) having an R squared value of 0.0067, approaching randomness. The equation of this graph indicated a gradient of 0.0007, which is a large deviation away from the expected gradient of 1. The model once again showed an extreme over prediction of values at the low end of the range and an under prediction at the high end of the range. The intercept value showed predictions were systematically higher than they should be. To try and improve the model the next step was to add the size of cell parameter as an input.

Graph 6: *Graph showing the actual v averaged predicted MFI CD147 using the inputs of "Shear speed (rpm)","size of cell" and "shear time (min)"*



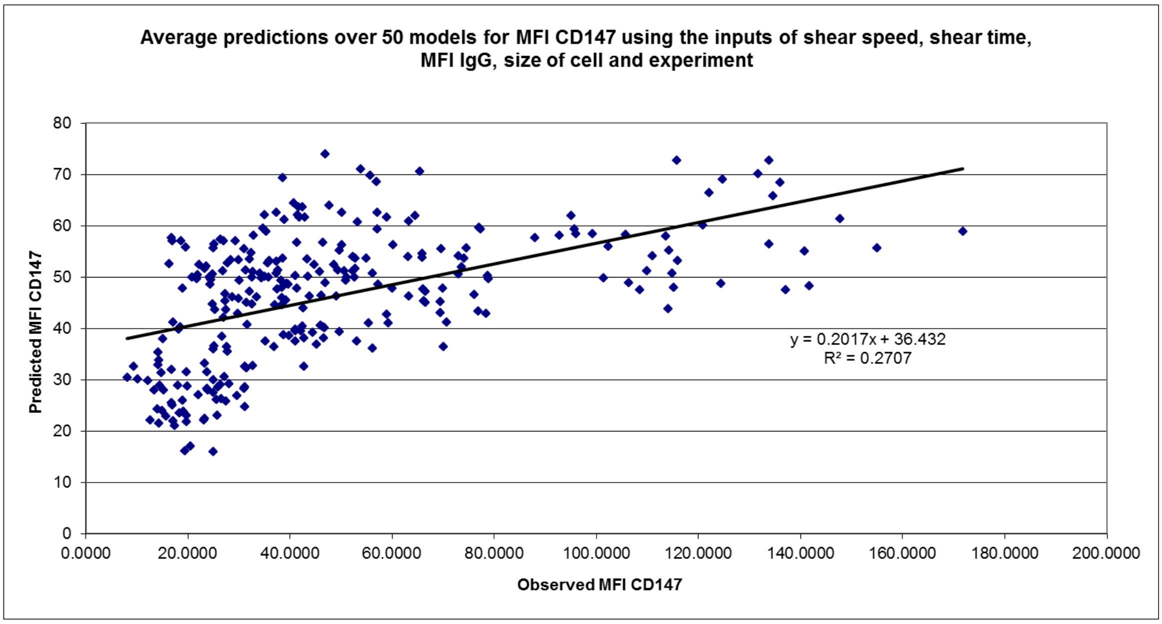
The model developed for step 2 predicting the CD147 surface marker (graph 6), utilised only inputs for shear speed, shear time and size of cells. This model produced an R squared value of 0.005, a lower 5 squared than those show in the previous model (Graph 5). The equation of this graph indicated a gradient of 0.0048, a strong deviation away from the expected gradient of 1. An intercept of 42.67 indicated predictions were systematically higher than they should be. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. This model was further modified by the addition of the MFI IgG parameter.

Graph 7: *Graph showing the actual v averaged predicted MFI CD147 using the inputs of “Shear speed (rpm)”, ”size of cell”, “MFI IgG” and “shear time (min)”*



Step 3 for this marker used the inputs shear speed, shear time, size of cells and MFI IgG. This model performed poorly. The calculated R squared value of 0.0216 again indicated a weak relationship between the actual and predicted values (Graph 7). The equation of this graph indicated a gradient of 0.0118, when compared to an expected gradient of 1, showed once again an over prediction of values at the low end of the range and an under prediction at the high end of the range. The intercept of 41.97 indicated higher prediction values than expected, as this should be 0. The next model would have an extra input of MFI IgG to see if this input would explain further variance in the models

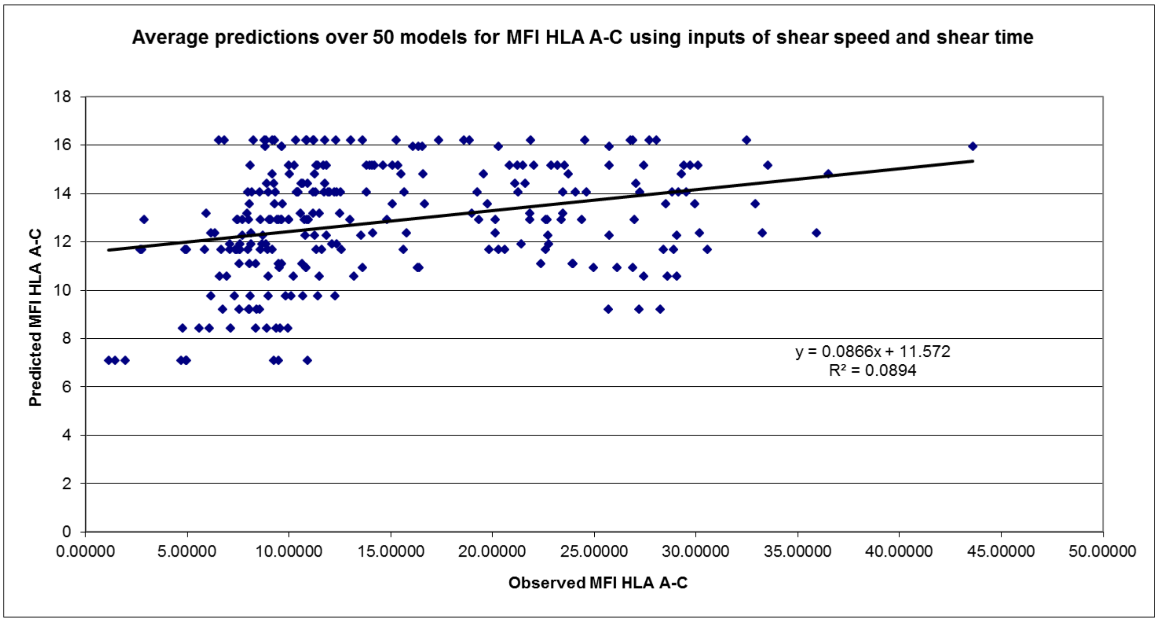
Graph 8: *Graph showing the actual v averaged predicted MFI CD147 using the inputs of “Shear speed (rpm)”, ”size of cell”, “MFI IgG”, “experiment” and “shear time (min)”*



The model developed for step 4 predicting CD147 utilised only the inputs for shear speed, shear time, size of cells, MFI IgG and experiment. This model showed a weak relationship between actual and predicted Graph 6 having an R squared value of 0.27. Again the equation of this graph indicated a gradient of 0.2, a large deviation away from the expected gradient of 1. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 36.43 indicated predictions were higher than they should be.

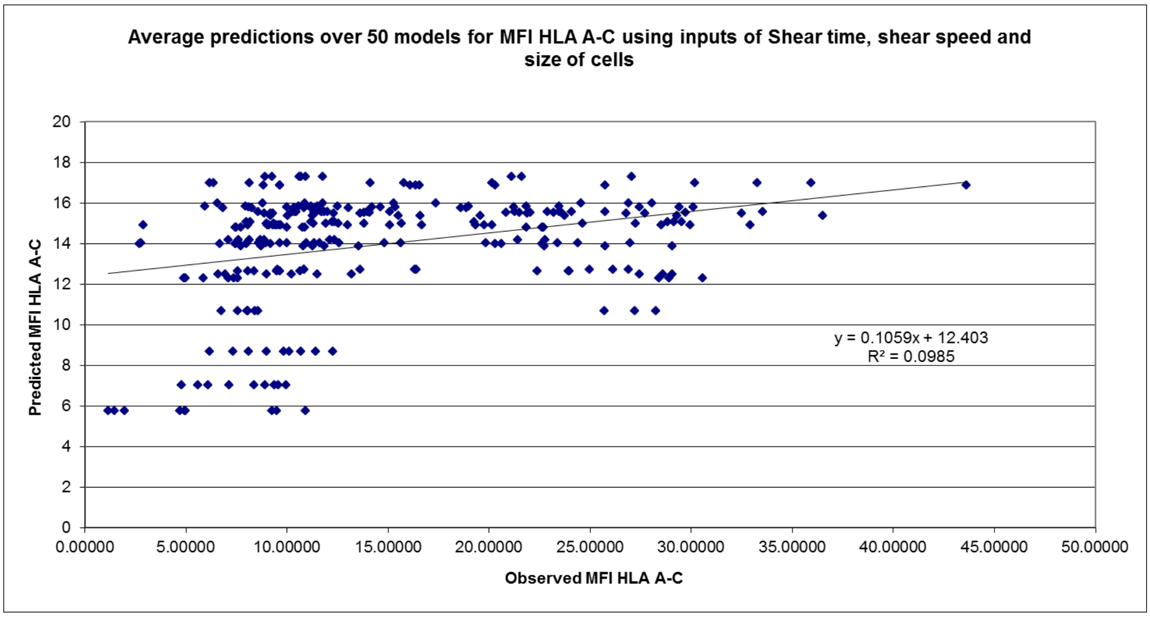
3.1.3. HLA A-C

Graph 9: *Graph showing the actual v averaged predicted MFI HLA A-C using the inputs of "Shear speed (rpm)" and "shear time (min)"*



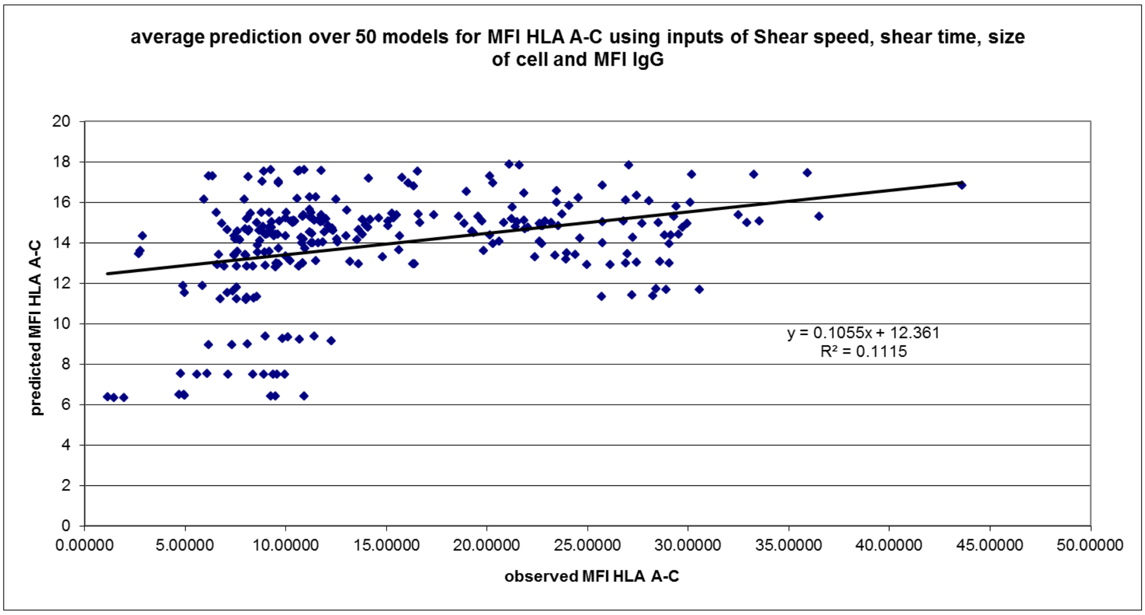
Model step 1 for HLA A-C was created using just the inputs for shear speed and time. This model did not perform well showing a weak relationship between actual and predicted Graph 9 having an R squared value of 0.089, closer to being random values. The equation of this graph indicated a gradient of 0.087, a strong deviation away from the expected gradient of 1. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 11.57 indicated predictions were systematically higher than they should be. This model was further modified by the addition of the size of cell parameter.

Graph 10: *Graph showing the actual v averaged predicted MFI HLA A-C using the inputs of "Shear speed (rpm)", "Size of cell" and "shear time (min)"*



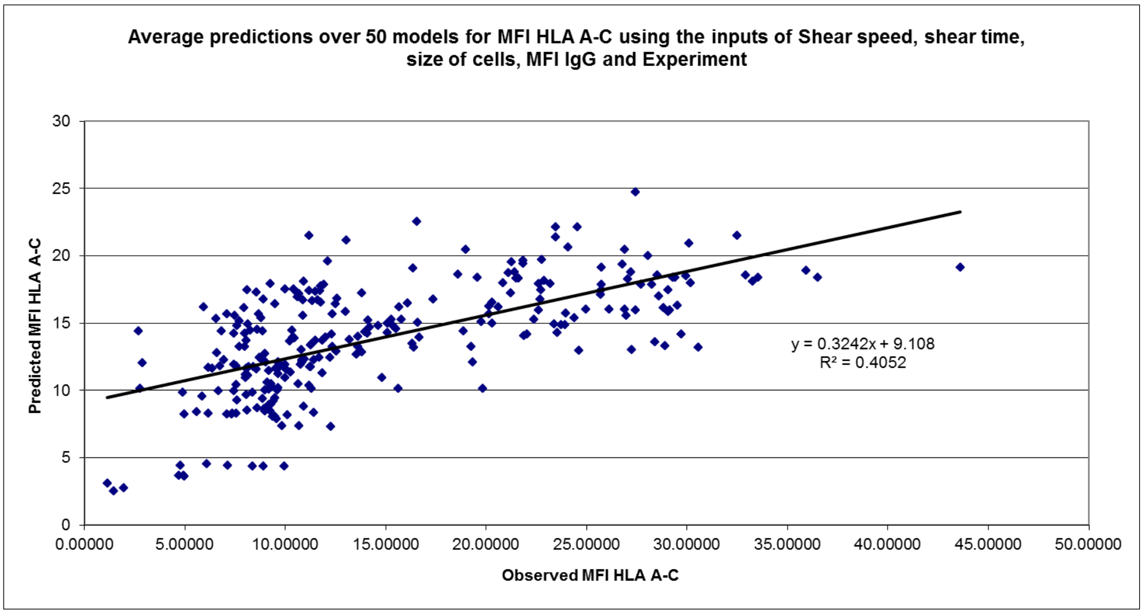
The inputs of shear speed, shear time and size of cells was used to create the model step 2. This model performed poorly indicating a weak relationship between actual and predicted (Graph 10) having an R squared value of 0.098. The equation of this graph indicated a strong deviation away from the expected gradient of 1 with a gradient of 0.11. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 12.4 indicated predictions were systematically higher than they should be. This model was further modified by the addition of the MFI IgG parameter.

Graph 11: *Graph showing the actual v averaged predicted MFI HLA A-C using the inputs of "Shear speed (rpm)", "Size of cell", "MFI IgG" and "shear time (min)"*



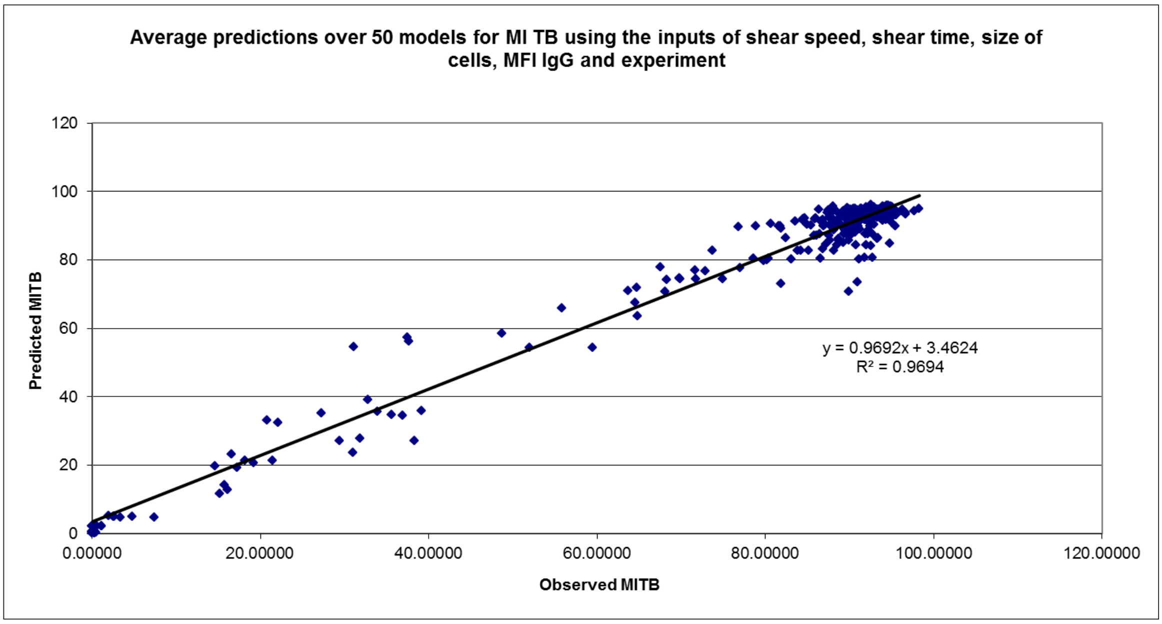
Model step 3 for HLA A-C was created using the inputs of shear speed, shear time, size of cells and the additional input MFI IgG. This model did not perform well showing a weak relationship between actual and predicted Graph 11 having an R squared value of 0.1115.. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 12.36 indicated predictions were higher than they should be. This model was further modified by the addition of the experiment parameter.

Graph 12: *Graph showing the actual v averaged predicted MFI HLA A-C using the inputs of "Shear speed (rpm)", "Size of cell", "MFI IgG", "Experiment" and "shear time (min)"*



Model step 4 (Graph 12) for HLA A-C utilised only inputs for shear speed, shear time, size of cells, MFI IgG and experiment. The model performed well showing a weak relationship between the actual and predicted showing an R squared value of 0.4. The equation of this graph indicated a gradient of 0.32, a large deviation away from the expected gradient of 1. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 9.1 indicated predictions were higher than they should be.

Graph 13: : *Graph showing the actual v averaged predicted MI TB using the inputs of "Shear speed (rpm)", "Size of cell", "MFI IgG", "Experiment" and "shear time (min)"*



Model step 1 for MiTB (Graph 13) utilised only inputs for shear speed, shear time, size of cells, MFI IgG and experiment. This model performed well showing a strong relationship between the actual and predicted with an R squared value of 0.97. The equation of this graph indicated a gradient of 0.96, a marginal deviation away from the expected gradient of 1. This showed an over prediction of values at both the low end of the range and at the high end of the range are predicting well. An intercept of 3.4 indicated predictions were higher than they should be.

3.2 Model Development phase 2

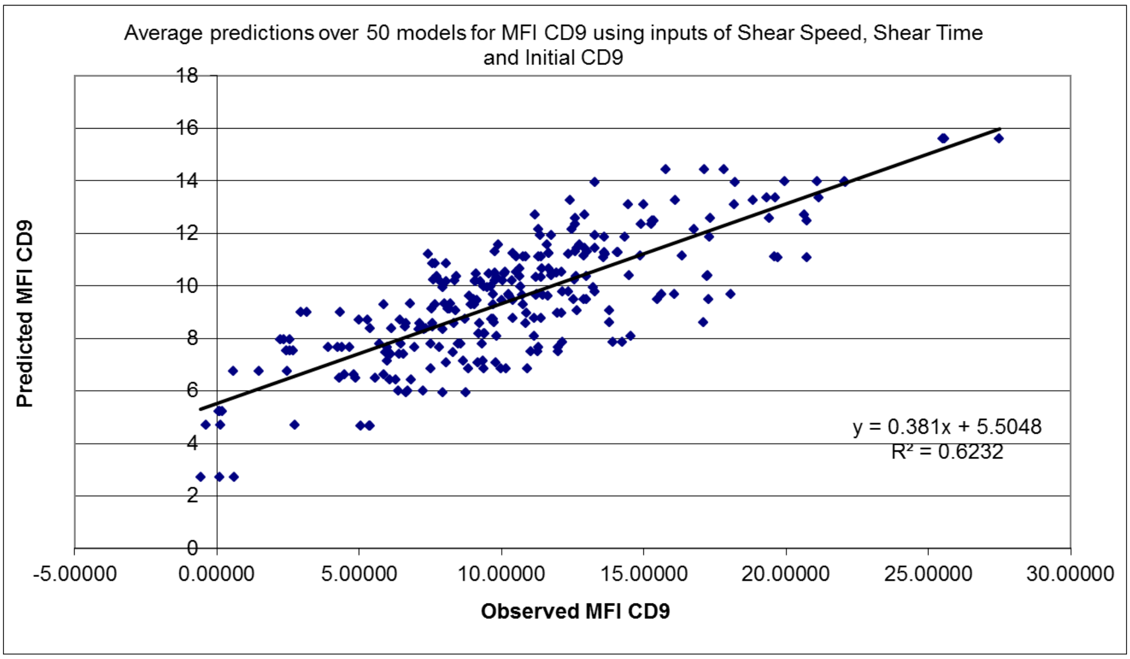
From the high sensitivity results for the experiment letter input from the previous experiments, it was noted that another factor was affecting the cells before being stressed in the shear device, in this section, the model was developed using the stepwise system, including a surface marker profile of the cell at time point zero.

3.2.1 CD9

The results below where built from ANN models including the "initial marker" input to act as a base line correction for any experimental variance between the cell batches.

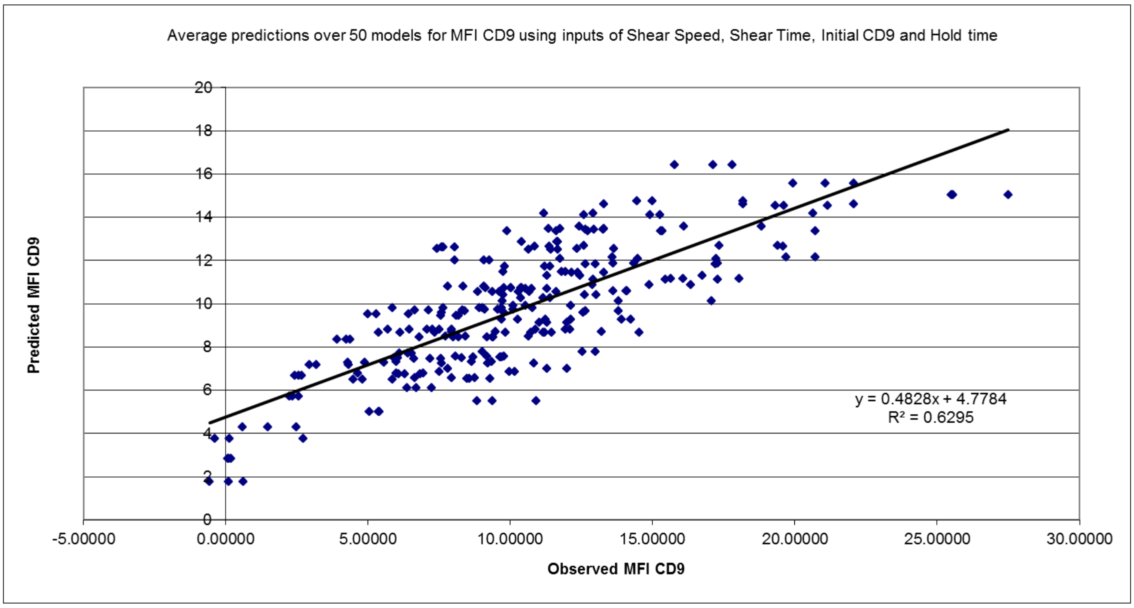
After the initial model using three inputs consisting of Shear Speed, Initial CD9 and Shear Time, an average was taken over the 50 random sub sampled models (including test, training and validation data) and the average value was plotted against the actual value from the data.

Graph 14*: Graph showing the actual v averaged predicted MFI CD9 using the inputs of "Shear speed (rpm)", "initial CD9" and "shear time (min)"*



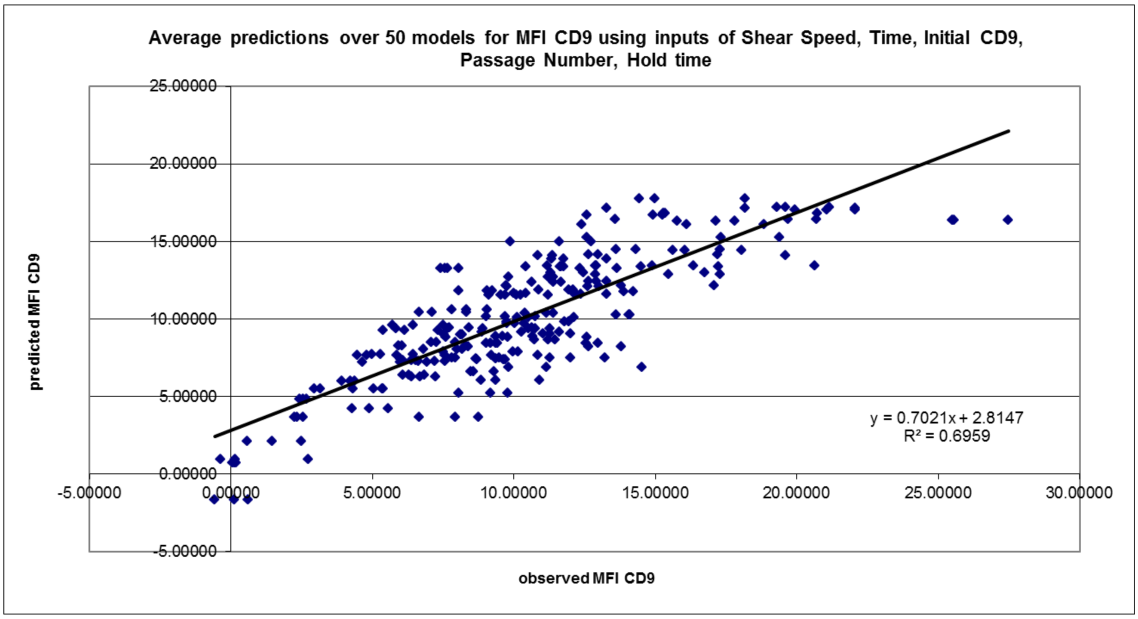
The phase 2 model step 1 for CD9 utilised only inputs for shear speed, time and this time the initial CD9 marker value. This model (Graph 14) performed poorly showing a weak relationship between the actual and predicted having an R squared value of 0.62. The equation of this graph indicated a gradient of 0.38, still showing a deviation away from the expected gradient of 1. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 5.50 indicated predictions were systematically higher than they should be. This model was further modified by the addition of the hold time parameter.

Graph 15: *Graph showing the actual v averaged predicted MFI CD9 using the inputs of "Shear speed (rpm)", "initial CD9", "Hold time" and "shear time (min)"*



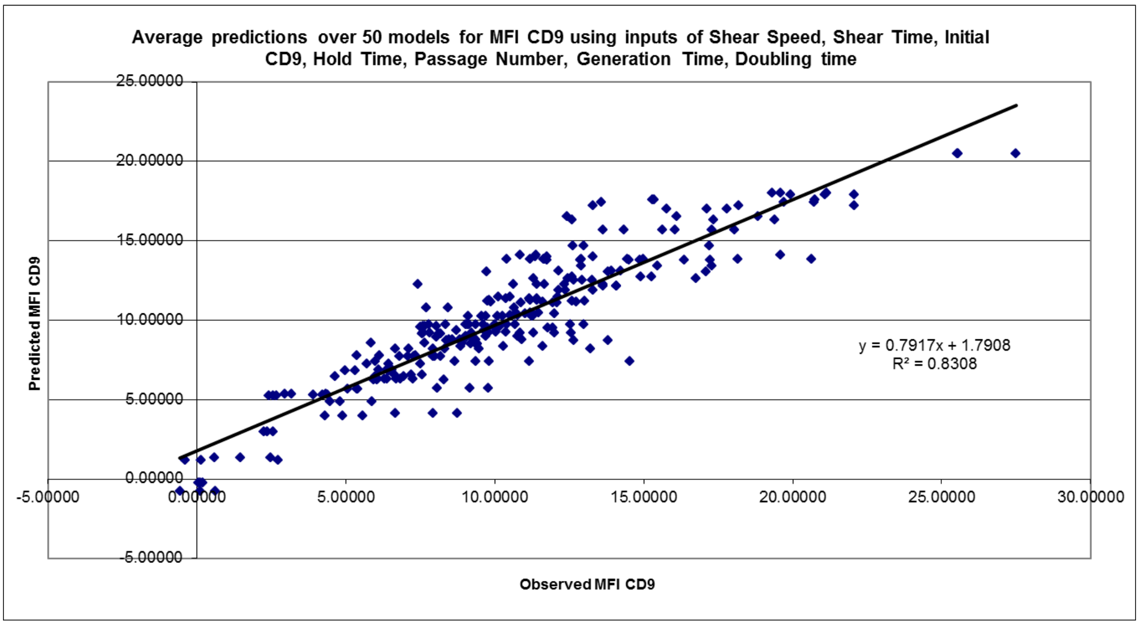
Model step 2 for CD9 was created using the inputs of shear speed, shear time, hold time and the initial CD9 marker value. This model did perform well showing a weak relationship between actual and predicted Graph 15 having an R squared value of 0.62. The equation of this graph indicated a gradient of 0.48, still showing a deviation away from the expected gradient of 1. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 4.7 indicated predictions were systematically higher than they should be. This model was further modified by the addition of the passage number parameter.

Graph 16: *Graph showing the actual v averaged predicted MFI CD9 using the inputs of "Shear speed (rpm)", "initial CD9", "Hold time", "Passage Number" and "shear time (min)*



Model step 3 for CD9 utilised only inputs for shear speed, time, hold time, passage number, doubling time and the initial CD9 marker value. This model performed ok showing a slight relationship between actual and predicted. Graph 16 showed an R squared value of 0.69. The equation of this graph indicated a gradient of 0.70, still showing a deviation away from the expected gradient of 1, but less than that of the previous model. The model showed an improved set of prediction values at both the low end high end of the range. An intercept of 1.9 indicated predictions were systematically higher than they should be, however greatly reduced from that in previous models. This model was further modified by the addition of the generation time and doubling time parameters.

Graph 17: *Graph showing the actual v averaged predicted MFI CD9 using the inputs of "Shear speed (rpm)", "initial CD9", "Hold time", "Passage Number", "Generation tine", "doubling time" and "shear time (min)"*



Model step 4 for CD9 (Graph 17) utilised only inputs for shear speed, time, hold time, doubling time, passage number, generation time and the initial CD9 marker value. This model showed a strong relationship between the actual and predicted having an R squared value of 0.83. The equation of this graph indicated a gradient of 0.79, still showing a deviation away from the expected gradient of 1. An intercept of 1.8 indicated predictions were systematically higher than they should be, however greatly reduced from that in previous models.

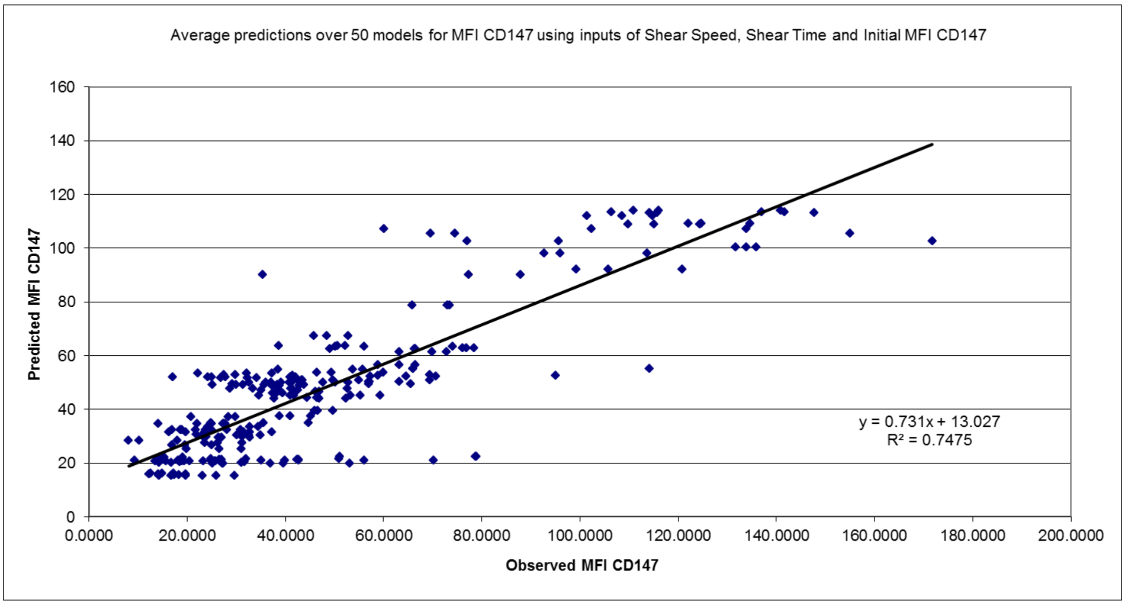
Table 1: Global Sensitivity analysis for MFI CD9 models

Speed	Initial CD9	Time (min)	Hold Time	Passage Number	Generation Time	Doubling Time
2.148030	1.616707	1.353921				
1.463865	1.776238	1.432283	1.670767			
1.521067	2.132421	1.746174	1.584531	1.633817		
3.020321	3.508796	2.387951	3.20364	6.11595	5.72908	3.100002

Table 1 shows the global sensitivity analysis for each of the 50 sub models utilising the various combinations of inputs. For the model using the speed, time and initial MFI CD9 inputs speed had the largest influence on the model. For both the models using the inputs speed, time, initial CD9, hold time, and the model containing the previously mentioned and passage number the initial MFI CD9 input has the largest sensitivity value indicating this marker had the biggest influence of the predictive output of the model. The model using all the inputs except for Passage Number showed that the input with the highest sensitivity within the model was hold time, compared to that of passage number in the final model using all seven inputs.

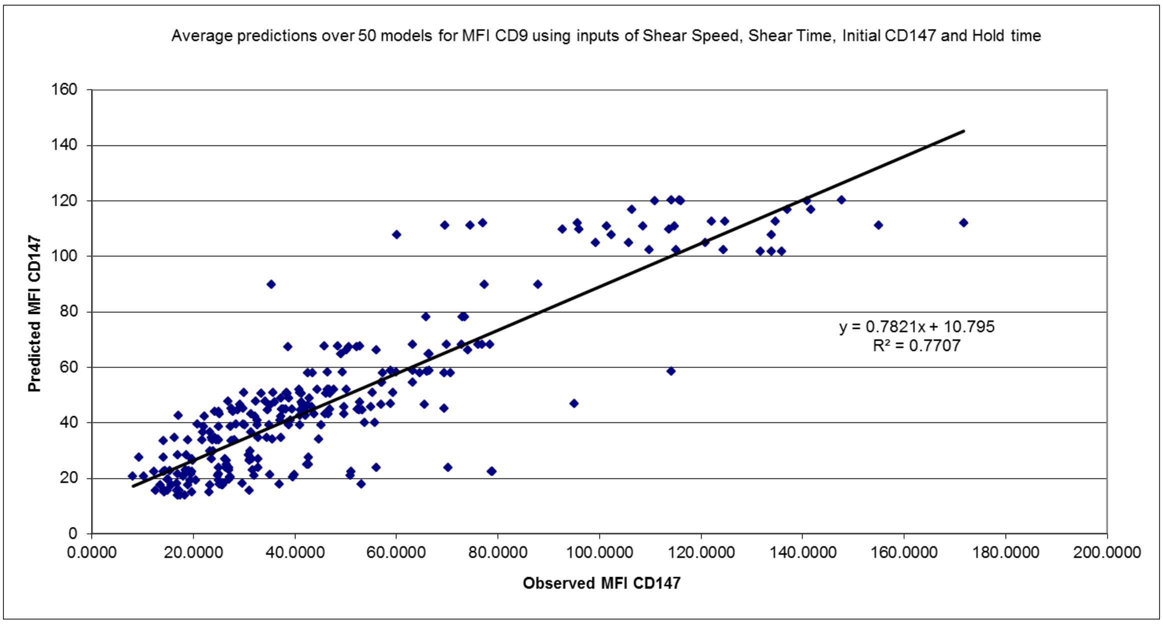
3.2.2 CD147

Graph 18: *Graph showing the actual v averaged predicted MFI CD147 using the inputs of "Shear speed (rpm)", "initial MFI CD147" and "shear time (min)"*



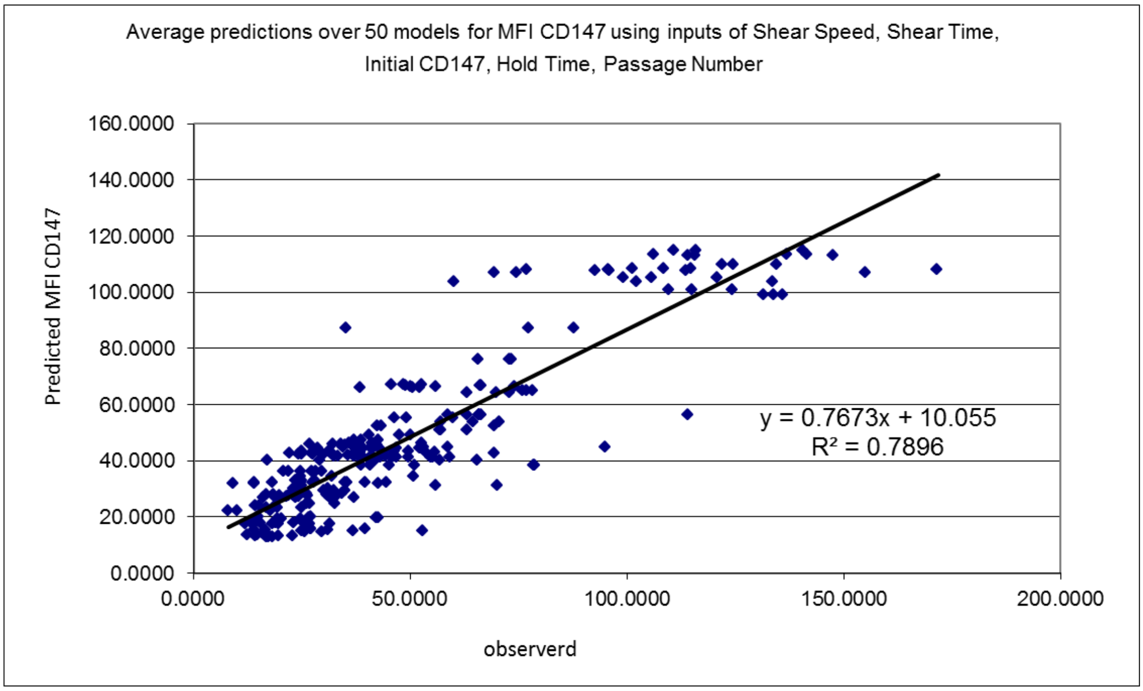
Initially model step 1 for CD147 utilised only inputs for shear speed, time and the initial CD9 marker value. This model performed well showing a relationship between the actual and predicted values. Graph 18 shows an R squared value of 0.74. The equation of this graph indicated a gradient of 0.73. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 13 indicated predictions were higher than observed. This model was then enhanced with the hold time parameter.

Graph 19: *Graph showing the actual v averaged predicted MFI CD147 using the inputs of "Shear speed (rpm)", "initial MFI CD147", Hold time and "shear time (min)"*



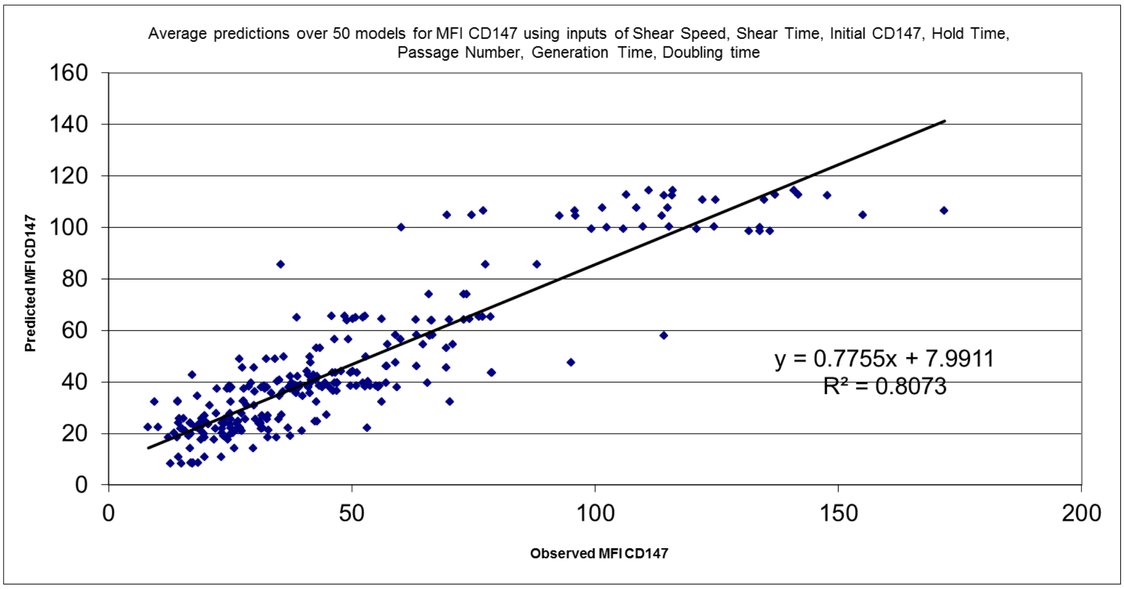
Model step 2 for CD147 utilised only inputs for shear speed, time, hold time and the initial CD147 marker value. Again this model performed well with an R squared value of 0.77. The equation of this graph indicated a gradient of 0.78, still showing a deviation away from the expected gradient of 1. An intercept of 10.8 indicated predictions were higher than observed in the experiments.

Graph 20: *Graph showing the actual v averaged predicted MFI CD147 using the inputs of "Shear speed (rpm)", "initial MFI CD147", "Hold time", "Passage Number" and "shear time (min)"*



This model used inputs for shear speed, time, hold time, passage number, doubling time and the initial CD147 marker value. Again this model performed well showing a relationship between the observed and predicted values shown in Graph 20. An R squared value of 0.78 was calculated. The equation of this graph indicated a gradient of 0.76, still showing a deviation away from the expected gradient of 1. The model showed a slight over prediction of values at both ends of the range. The graph also shows an intercept of 10, which indicates predictions were systematically higher than they should be, however greatly reduced from that in previous models.

Graph 21: *Graph showing the actual v averaged predicted MFI CD147 using the inputs of "Shear speed (rpm)", "initial MFI CD147", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)"*



Model step 4 for CD147 was created using the inputs of shear speed, time, hold time, doubling time, passage number, generation time and the initial CD147 marker value. The model did perform well indicating a strong relationship between the actual and predicted values. Graph 21 shows an R squared value of 0.80. The equation of this graph indicated a gradient of 0.77, still showing deviation. The intercept value of 7.99 still indicates that the models predictions were higher than the observed values, however greatly reduced from that in previous models.

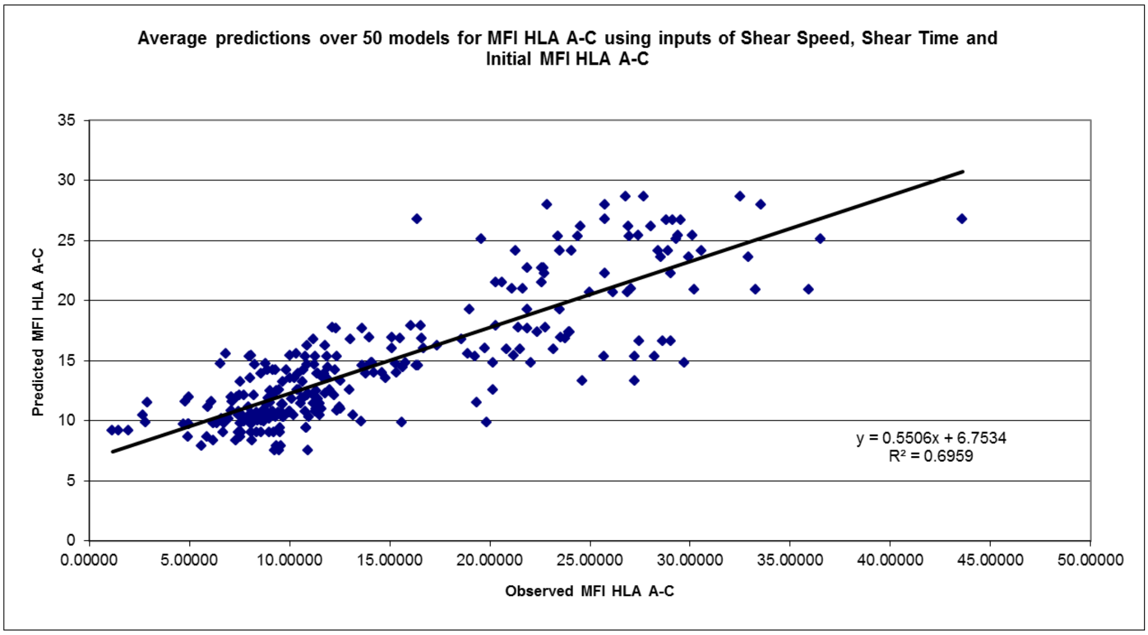
Table 2: Global Sensitivity Analysis for MFI CD147 Models

Initial CD147	Doubling Time (Hours)	Time (min)	Hold Time	Generation Time	Speed	Passage Number
3.649293		1.119729			1.388021	
4.303928		1.188441	1.408585		1.124629	
3.916817		1.170919	1.443956		1.169597	1.236082
3.285769	1.225134	1.230822	1.273689	2.608524	1.183491	2.147692

Table 2 showed that in the global sensitivity analysis for the 50 models used to predict MFI CD147 was the initial MFI CD147 marker input for all models generated

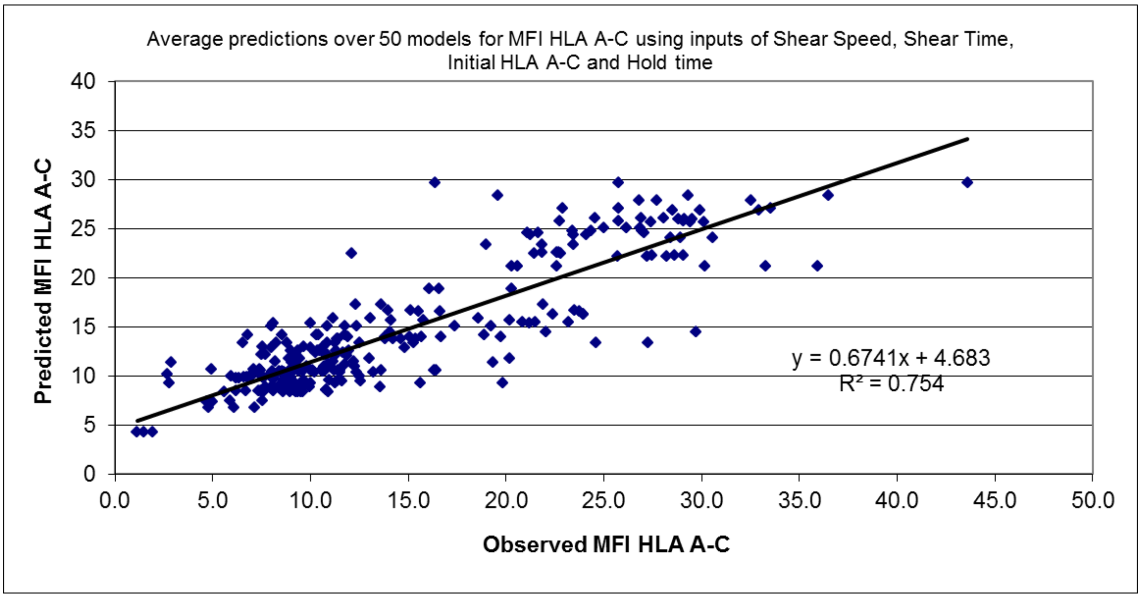
3.2.3 HLA A-C

Graph 22: *Graph showing the actual v averaged predicted MFI HLA A-C using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C" and "shear time (min)"*



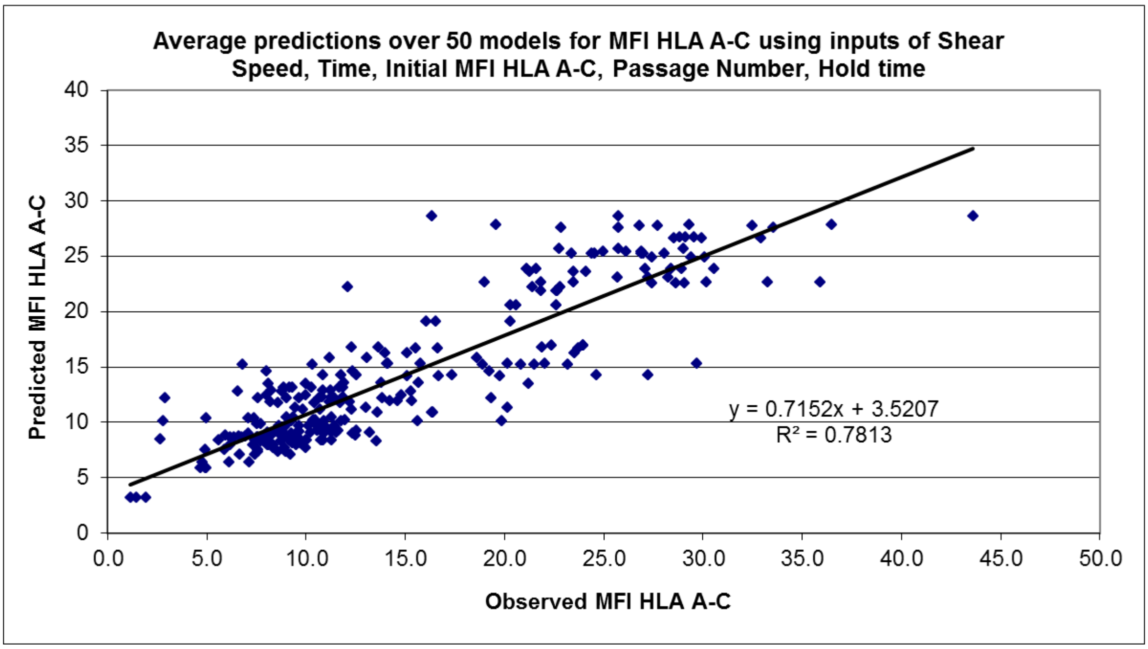
Initially the model for step 1 predicting the HLA A-C surface marker (graph 22) was created using the inputs of shear speed, shear time and the initial HLA A-C marker value. Again this model performed well with an R squared value of 0.7, showing a strong relationship between observed and predicted. There is still some deviation away from the expected as the graph indicated a gradient of 0.55. Again an intercept of 6.6 indicated predictions were systematically higher than they should be. This model was further modified by the addition of the hold time parameter.

Graph 23: *Graph showing the actual v averaged predicted MFI HLA A-C using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C", Hold time and "shear time (min)"*



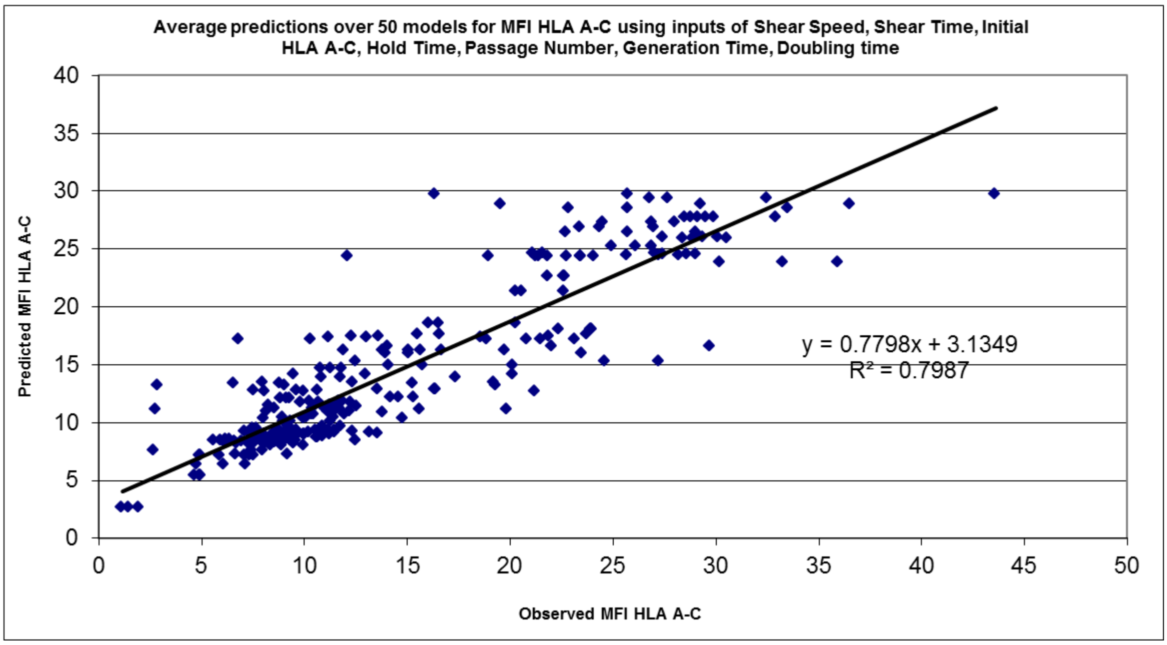
Model step 2 for HLA A-C utilised only inputs for shear speed, time, hold time and the initial HLA A-C marker value. This model did perform well showing a relationship between actual and predicted Graph 23 having an R squared value of 0.75. The equation of this graph indicated a gradient of 0.68, still showing a deviation away from the expected gradient of 1. This showed an over prediction of values at the low end of the range and an under prediction at the high end of the range. An intercept of 4.6 indicated predictions were systematically higher than they should be. This model was further modified by the addition of the passage number parameter.

Graph 24: *Graph showing the actual v averaged predicted MFI HLA A-C using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C", "Hold time", "Passage Number" and "shear time (min)"*



Model step 3 for HLA A-C utilised only inputs for shear speed, time, hold time, passage number, doubling time and the initial HLA A-C marker value. This model did perform well showing a relationship between actual and predicted Graph 24 having an R squared value of 0.78. The equation of this graph indicated a gradient of 0.71, still showing a deviation away from the expected gradient of 1. This showed a slight over prediction of values at the low end of the range and an under prediction at the high end of the range still. An intercept of 3.5 indicated predictions were systematically higher than they should be, however greatly reduced from that in previous models. This model was further modified by the addition of the generation time and doubling time parameters.

Graph 25: *Graph showing the actual v averaged predicted MFI HLA A-C using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)"*



Model step 4 for HLA A-C utilised only inputs for shear speed, time, hold time, doubling time, passage number, generation time and the initial HLA A-C marker value. This model did perform well showing a relationship between actual and predicted Graph 25 having an R squared value of 0.80. The equation of this graph indicated a gradient of 0.78, still showing a deviation away from the expected gradient of 1. This showed a slight over prediction of values at the low end of the range and an under prediction at the high end of the range still. An intercept of 3.13 indicated predictions were systematically higher than they should be, however greatly reduced from that in previous models.

Table 3: Global Sensitivity Analysis for MFI HLA A-C Models

initial HLA	Hold Time	Generation Time	Speed	Passage Number	Doubling Time (Hours)	Time (min)
4.088185	2.390197	2.279790	2.200413	2.051198	1.779703	1.543568
3.431842	1.639568		2.051747	1.314822		1.479011
3.127535	1.380311		1.618807			1.431905
2.221384			1.834105			1.202420

Table 3 showed that in the global sensitivity analysis for the 50 models used to predict MFI HLA A-C was the initial MFI HLA A-C marker input for all models.

From the results gathered in the initial analysis of this experiment showed that the model could be used to predict MFI values on the P4E6 whole cell vaccine for the surface markers CD9, CD147 and HLA A-C, using an Artificial Neural Network Algorithm

The accuracy of the models produced was determined using an R Squared value generated when the model prediction is compared against the observed biological data. The R squared value or coefficient of determination function when applied to this kind of model will range in value from 0 to 1, as the value increases towards 1 the likelihood of the predicted event increases.

4. Results Model Prediction

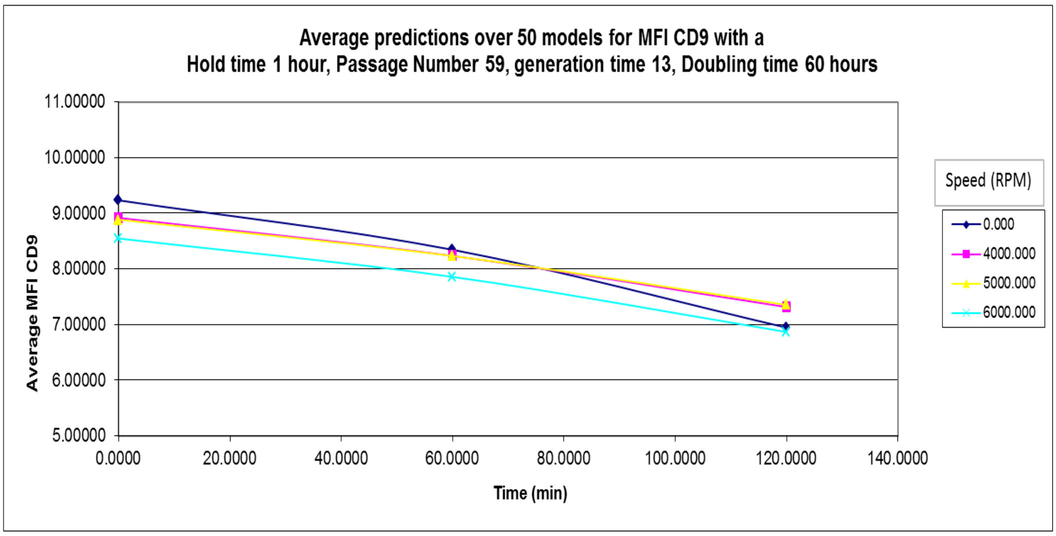
4.1. Model Prediction Phase 1

In this section of the results, the model will be tested to create a hypothetical set of results for MFI marker levels based on a series of experimental parameters and to look at what effect the changes in experimental parameters have on the MFI surface marker value for each of the biological markers.

These models where created using the inputs from the model that gave the highest r squared value when the average predicted value over the 50 models was plotted against the observed value. These models where created using the inputs *"Shear speed (rpm)", "initial MFI marker value* (taken from the raw data)*", "Hold time", "Passage Number ", "Generation tine", "doubling time" and "shear time (min)".*

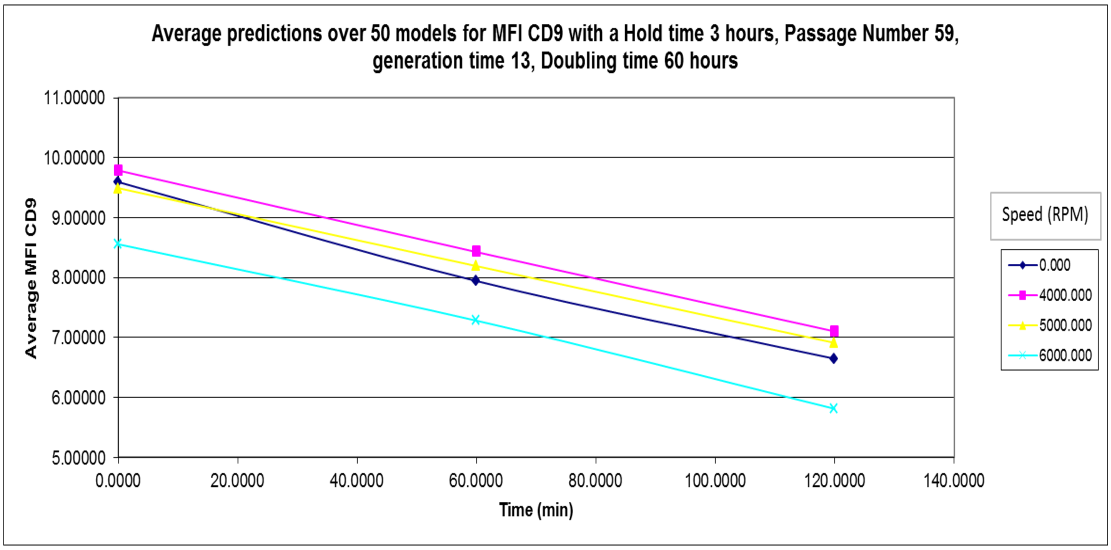
4.1.1. CD9

Graph 26: *Graph showing the predicted MFI CD9 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD9", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 1 hour*



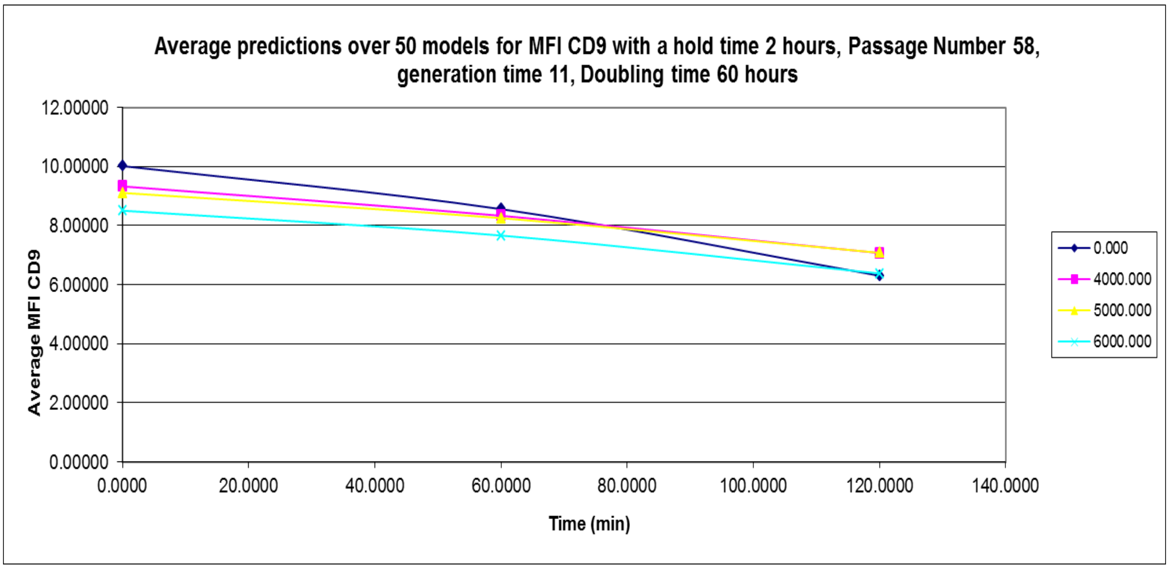
Graph 26 shows that, the ANN over the 50 models has predicted a steady decline in the average MFI CD9 marker as time increases for all shear speeds, with the sample subjected to no rotational shear having the largest decrease.

Graph 27: *Graph showing the predicted MFI CD9 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD9", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 3 hours*



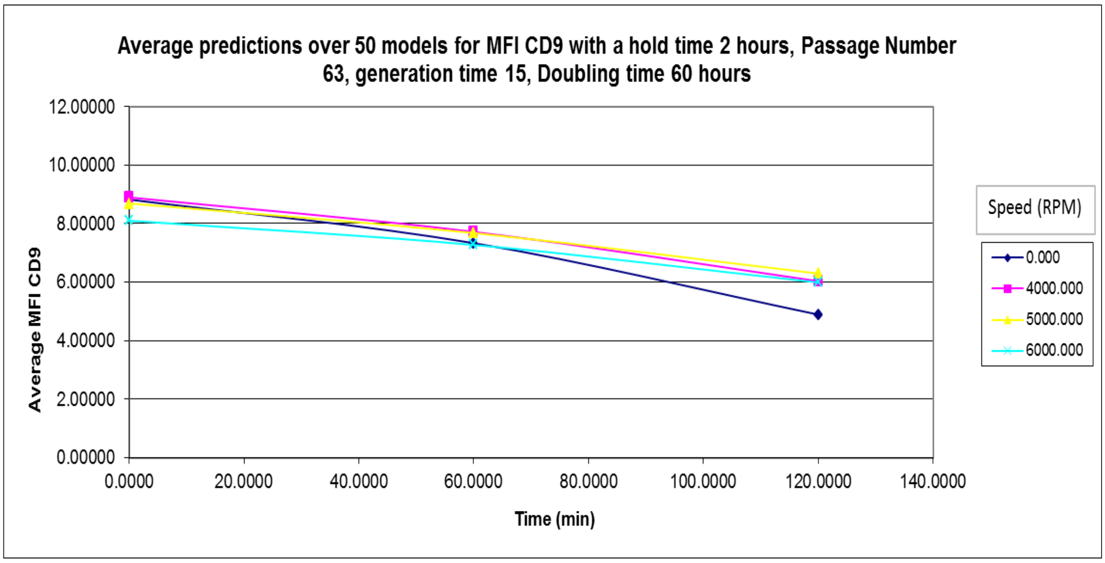
Graph 27 shows that the ANN predicts the 50 models has predicted a steady decline in the average MFI CD9 marker as time increases over all shear speeds, and a similar decline for the cells not subjected to any shear

Graph 28: *Graph showing the predicted MFI CD9 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD9", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" with a passage number of 58 and generation time of 11*



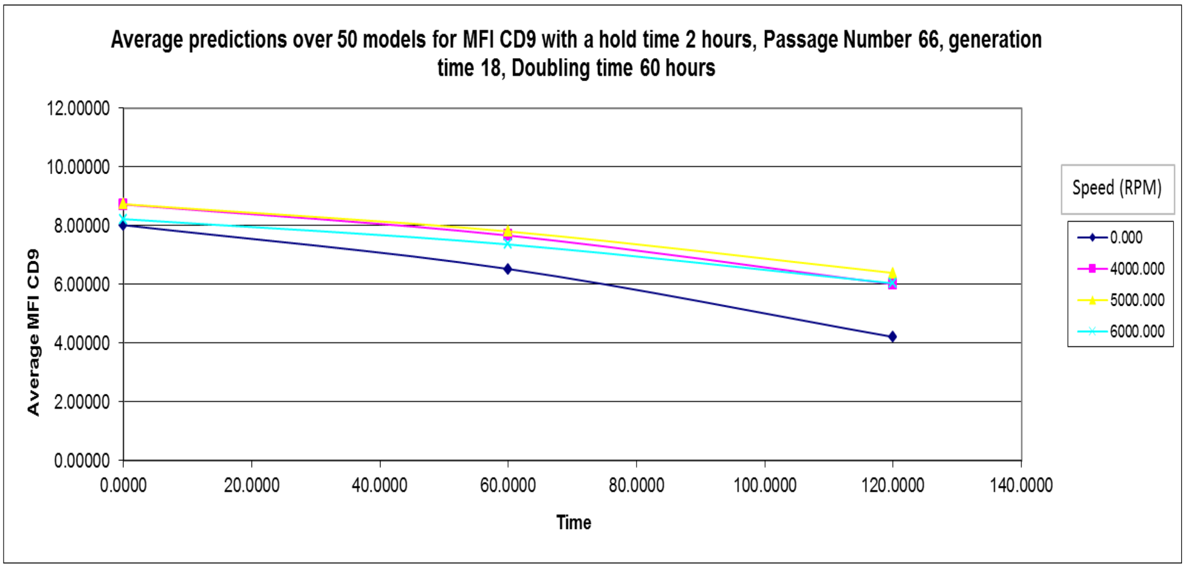
Graph 28 shows that, the ANN over the 50 models has predicted a steady decline in the average MFI CD9 marker as time increases for all shear speeds, with the sample subjected to no rotational shear having the largest decrease, where the samples subjected to rotational shear of 4000, 5000 and 6000 rpm shows a reduced decrease in the CD9 MFI value predicted by the model.

Graph 29: *Graph showing the predicted MFI CD9 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD9", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" with a passage number of 63 and generation time of 15*



Graph 29 shows that, the ANN has predicted a decline in the average MFI CD9 marker as time increases for all shear speeds, with the sample subjected to no rotational shear again having the largest decrease in the CD9 MFI value

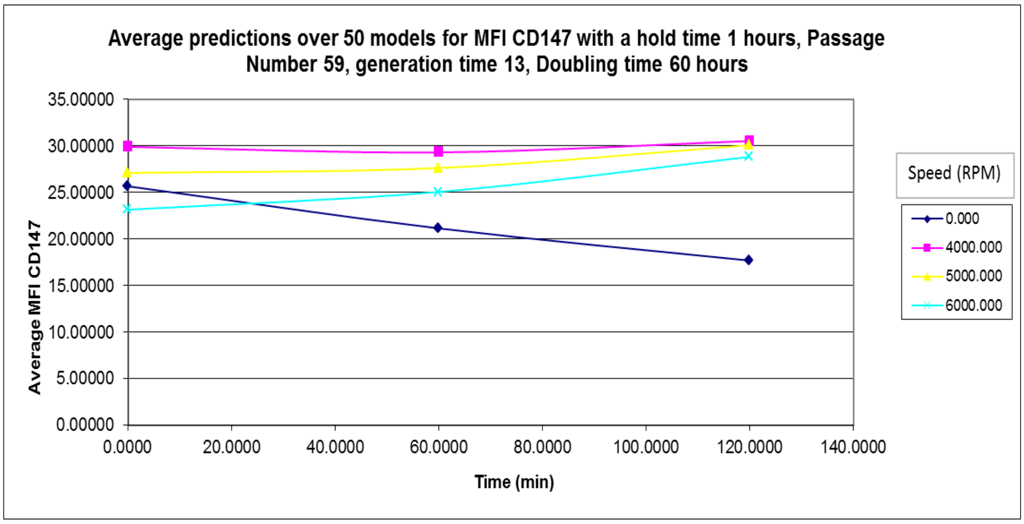
Graph 30: Graph showing the predicted MFI CD9 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD9", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" with a passage number of 66 and generation time of 18



Graph 30 shows that, the ANN again has predicted a decline in the average MFI CD9 marker as time increases from an MFI value of over 8 to a value below 7 for all shear speeds, with the sample subjected to no rotational shear having the largest decrease.

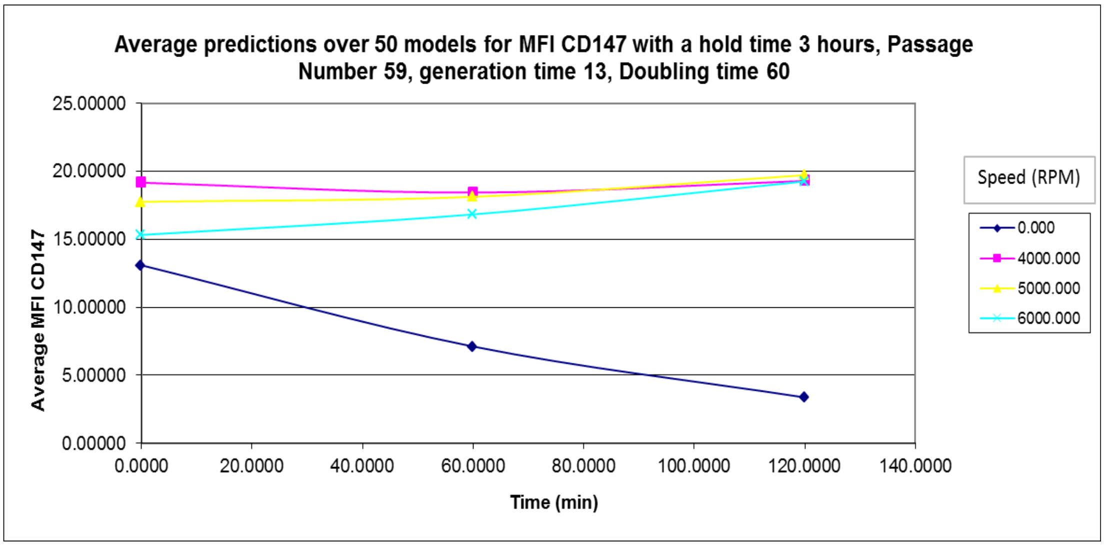
4.1.2. CD147

Graph 31: *Graph showing the predicted MFI CD147 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD147", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 1 hour*



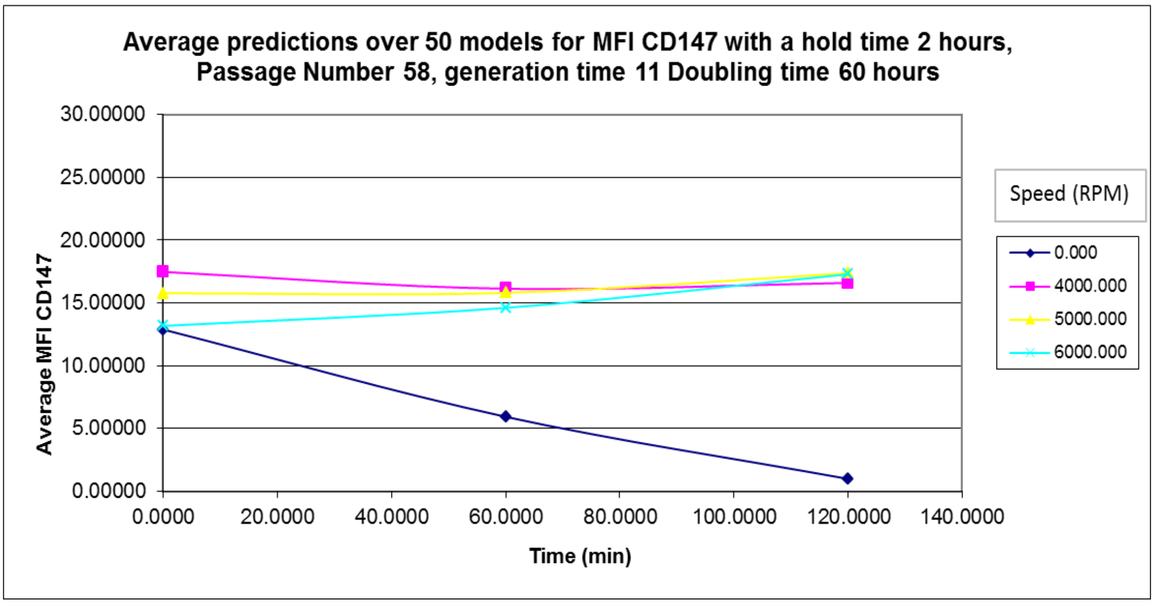
Graph 31 shows that, the ANN over the 50 models has predicted an increase in the average MFI CD147 marker as time increases for shear speeds of 4000, 5000 and 6000rpm, with the sample subjected to no rotational shear showing a decrease in the MFI value. The predictions for the rotational shear speed of 4000 rpm showing the smallest increase with the 6000 rpm showing the largest MFI value increase.

Graph 32: *Graph showing the predicted MFI CD147 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD147", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 3 hours*



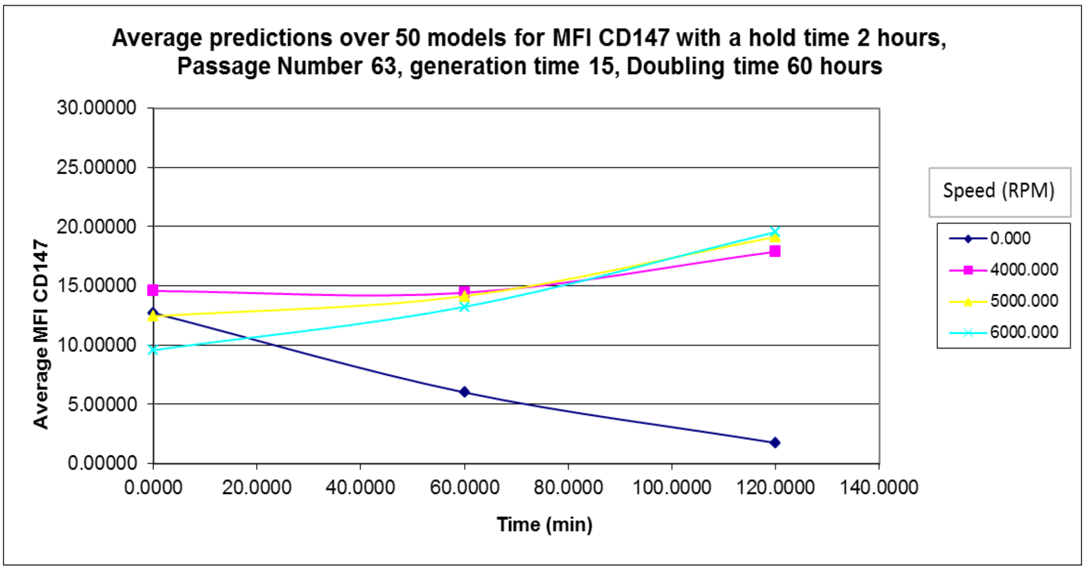
Graph 32 shows an increase in the average MFI CD147 marker as time increases for shear speeds of 4000, 5000 and 6000 rpm, with the sample subjected to no rotational shear showing a decrease in the MFI value. Again the predictions for the rotational shear speed of 4000 rpm showing the smallest increase with the 6000 rpm showing the largest MFI value increase. However, the MFI values for all the RPM's is less than that of the samples in graph 31.

Graph 33: *Graph showing the predicted MFI CD147 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD147", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" with a passage number of 58 and generation time of 11*



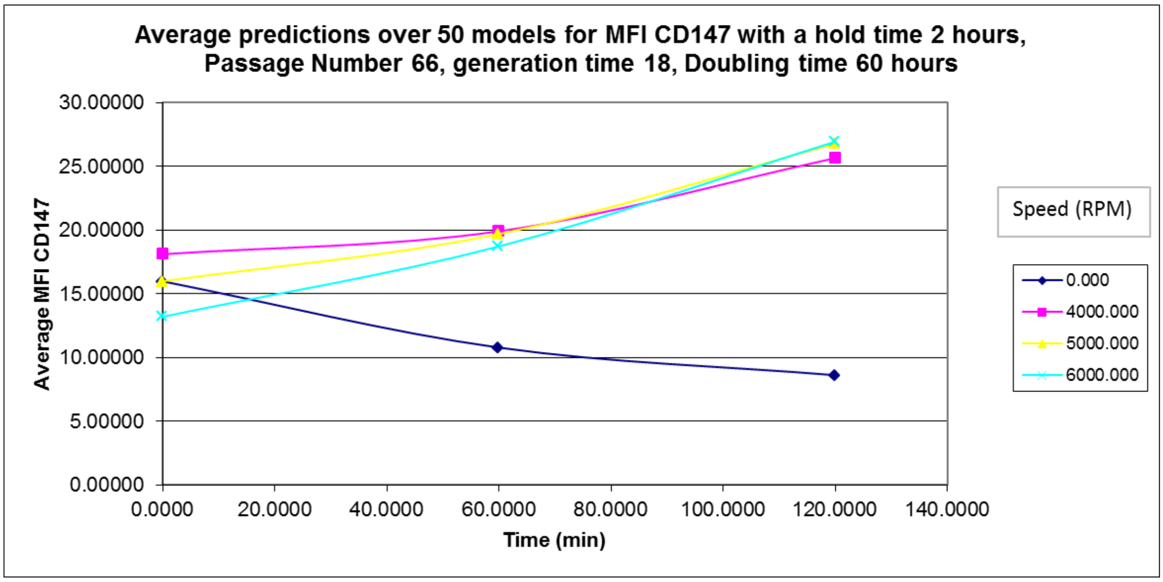
Graph 33 shows that, the ANN over the 50 models has predicted a slight increase in the average MFI CD147 marker as time increases for shear speeds of 4000, 5000 and 6000 rpm, The predictions for the rotational shear speed of 4000 rpm showing the smallest increase with the 6000 rpm showing the largest MFI value increase Once again, the sample subjected to no rotational shear showing a decrease in the MFI value..

Graph 34: *Graph showing the predicted MFI CD147 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD147", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" with a passage number of 63 and generation time of 15*



In Graph 34 the ANN has predicted an increase in the average MFI CD147 marker as time increases for shear speeds of 4000, 5000 and 6000 rpm, with the sample subjected to no rotational shear showing a decrease in the MFI value. The predictions for the rotational shear speed of 4000 rpm showing the smallest increase with the 6000 rpm showing the largest MFI value increase. The MFI values for all the RPM's is less than that of the samples in graph 33.

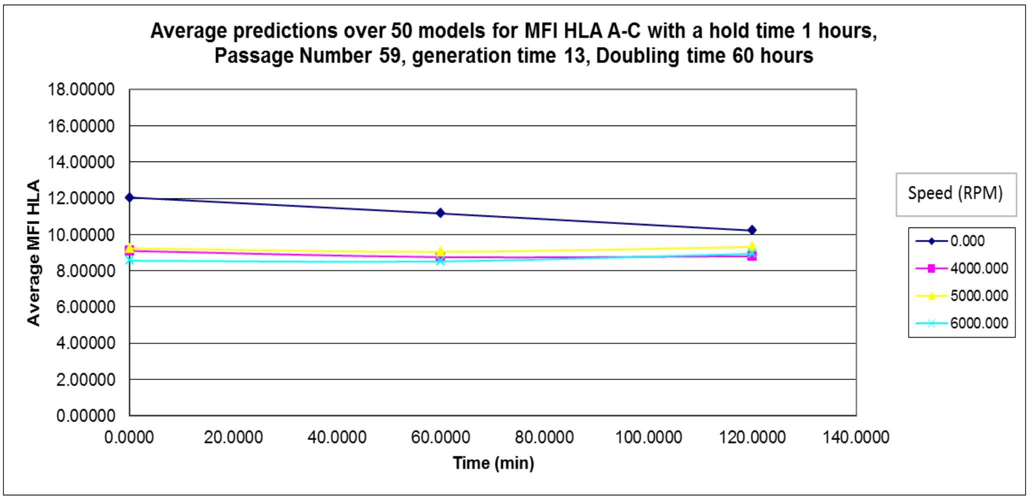
Graph 35: Graph showing the predicted MFI CD147 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD147", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" with a passage number of 66 and generation time of 18



Graph 35 shows that, the ANN over the 50 models has predicted an increase in the average MFI CD147 marker as time increases for shear speeds of 4000, 5000 and 6000 rpm. The non-rotational shear sample again showing a decrease in the MFI value. The predictions for the rotational shear speed of 4000 rpm showing the smallest increase with the 6000 rpm showing the largest MFI value increase. The MFI values for all the RPM's are higher than that of the samples in graph 36.

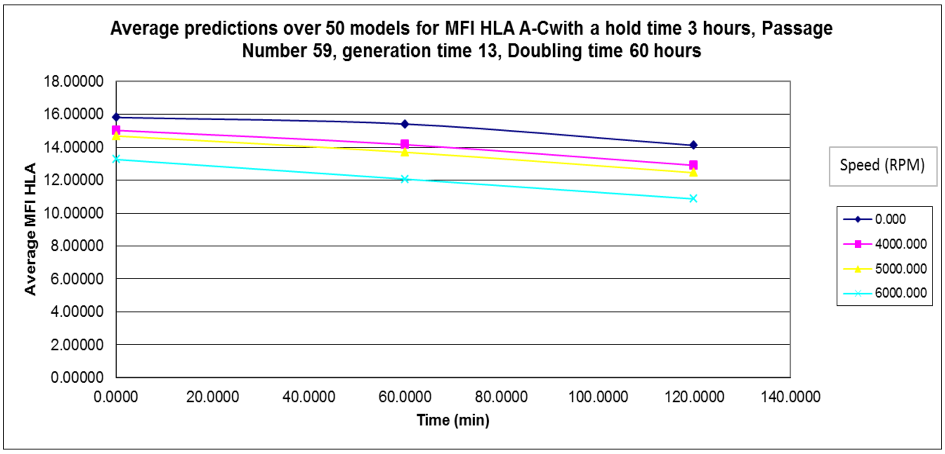
4.1.3 HLA A-C

Graph 36: *Graph showing the predicted MFI HLA A-C over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 1 hour*



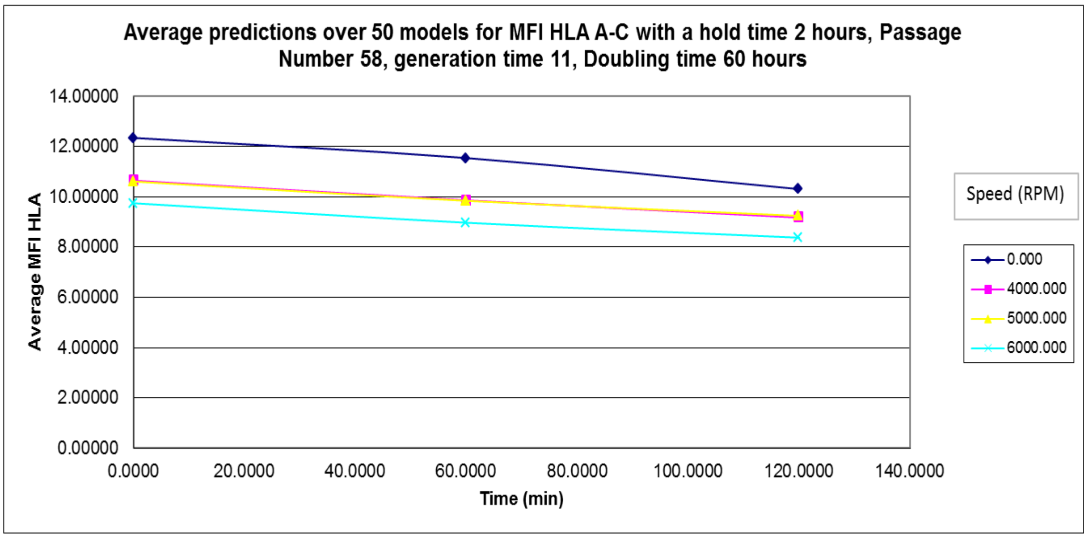
Graph 36 shows that the model has predicted a minor increase in the average MFI HLA A-C over the predicted shear speeds range over time for those samples subjected to rotational shearing. The predictions for no shear sample indicate a decrease in the MFI marker value over 120 minutes.

Graph 37: *Graph showing the predicted MFI HLA A-C over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 3 hours*



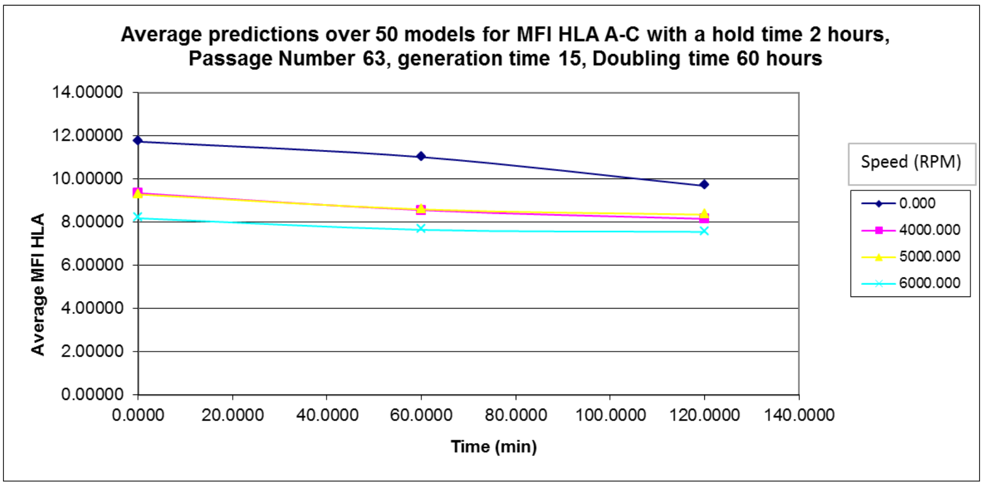
Graph 37 shows that, the ANN over the 50 models has predicted a decline in the average MFI HLA A-C over all of the predicted shear speeds range over time, with the MFI value for all decreasing by an MFI value of at least 2

Graph 38: *Graph showing the predicted MFI HLA A-C over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" with a passage number of 58 and generation time of 11*



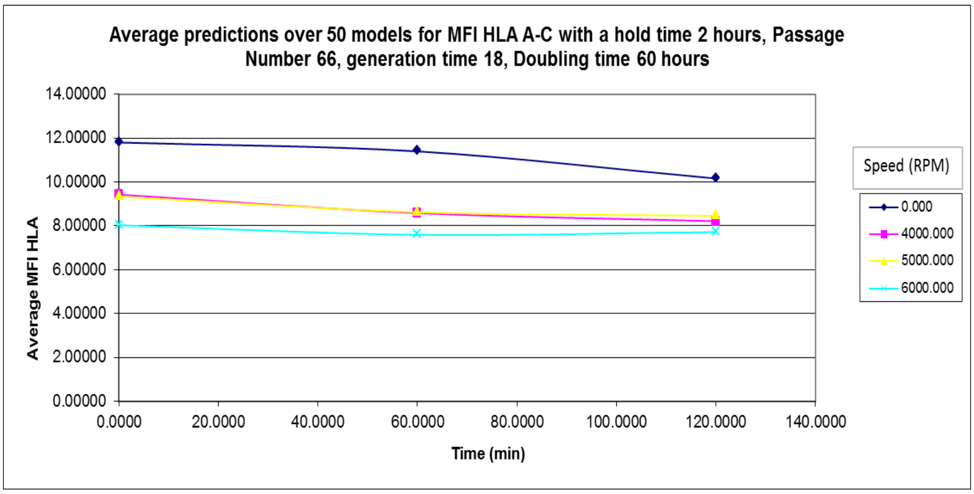
The ANN predictions shown in Graph 38 predict a decline in the average MFI HLA A-C over all of the predicted shear speeds range over time, with the no shear sample showing the biggest decrease and the 6000 rpm prediction showing the smallest decrease over the 120 minute period.

Graph 39: *Graph showing the predicted MFI HLA A-C over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" with a passage number of 63 and generation time of 15*



Graph 39 shows that, the ANN over the 50 models has predicted a decline in the average MFI HLA A-C over all of the predicted shear speeds range over time, with the no shear sample showing the biggest decrease and the 6000 rpm prediction showing the smallest decrease over the 120 minute period.

Graph 40: Graph showing the predicted MFI HLA A-C over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" with a passage number of 66 and generation time of 18



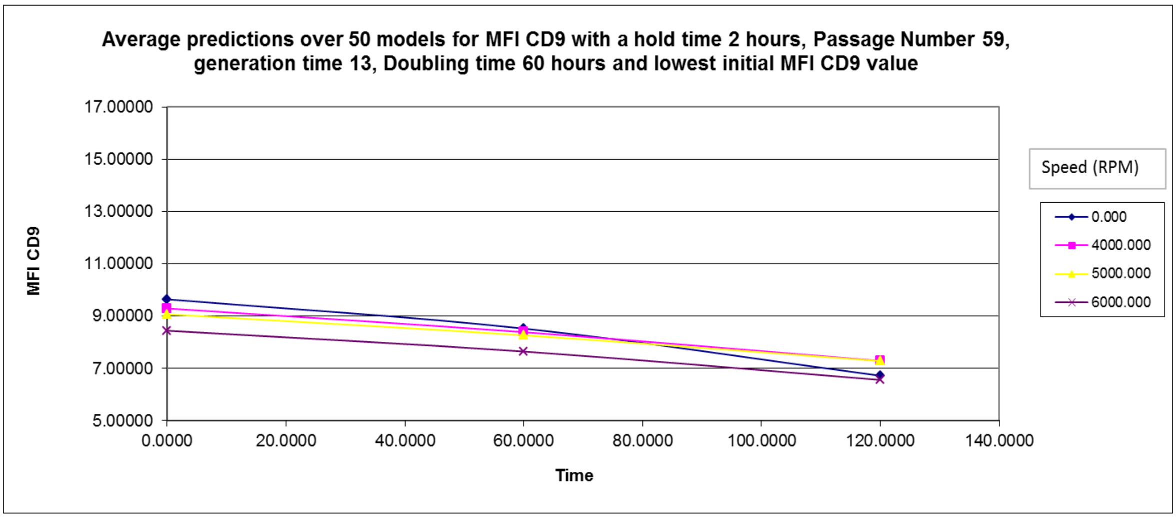
Graph 40 predicts a decline in the average MFI HLA A-C over all of the predicted shear speeds range over time, with the no shear sample showing the biggest decrease and the 6000 rpm prediction showing the smallest decrease over the 120 minute period with 4000 and 5000 rpm having a similar decline.

4.2. Model Predictions Phase 2

4.2.1. CD9

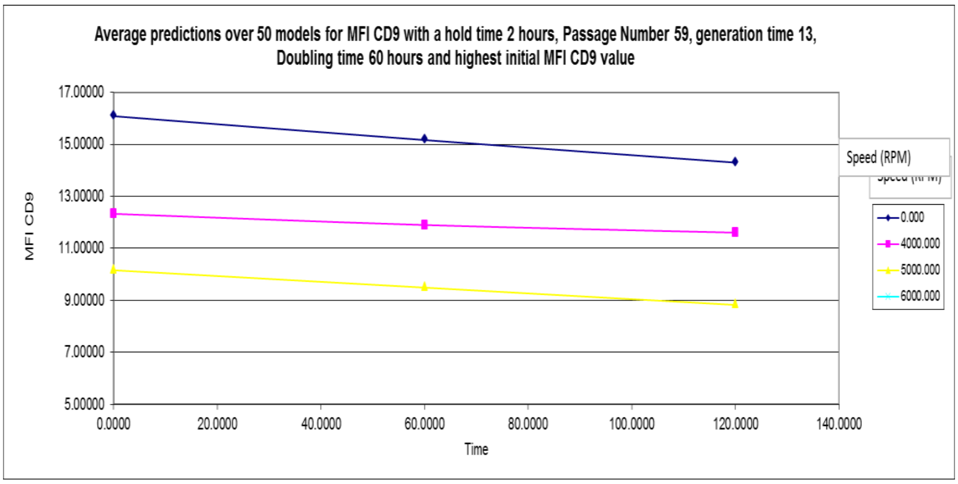
In this section of the results, the model will be tested to create a hypothetical set of results for MFI marker levels based on a series of experimental parameters and to look at what effect the quality of the cell or the initial quantity of the markers on the surface of the culture cells had on the outcome of the experiment.

Graph 41: *Graph showing the predicted MFI CD9 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD9", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 2 hours with a low initial marker level*



Graph 41 predicts the average MFI CD9 samples with a low staring initial marker level that are not subjected to shear forces have a greater decline than those subjected to the shear.

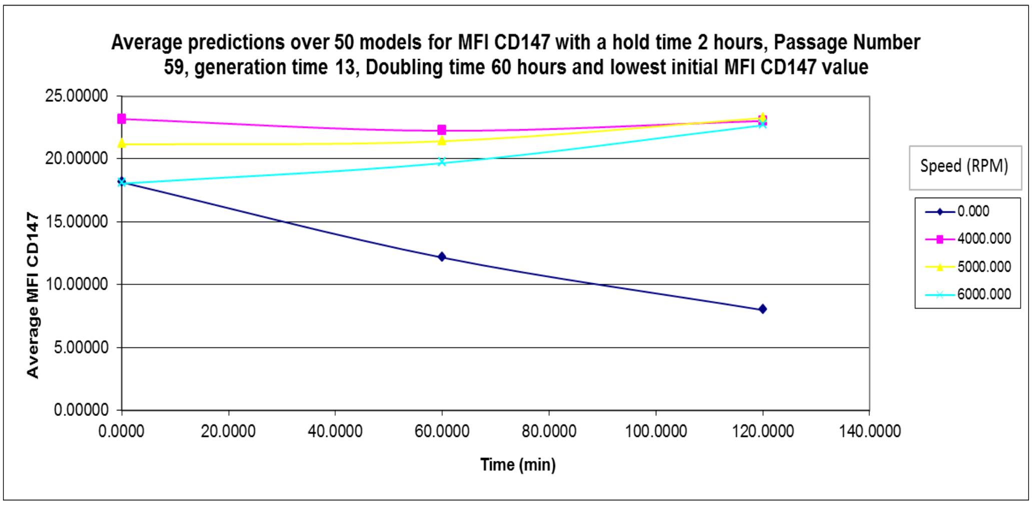
Graph 42: *Graph showing the predicted MFI CD9 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD9", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 2 hours with a high initial marker level*



Graph 42 indicates the average results over 50 models from the ANN predicting MFI CD9 with a high initial marker level before the culture is subjected to shear forces. The model also indicates displays a steady decline for the MFI CD9 marker over all shear speeds over the time period, similar to that of the low initial marker level.

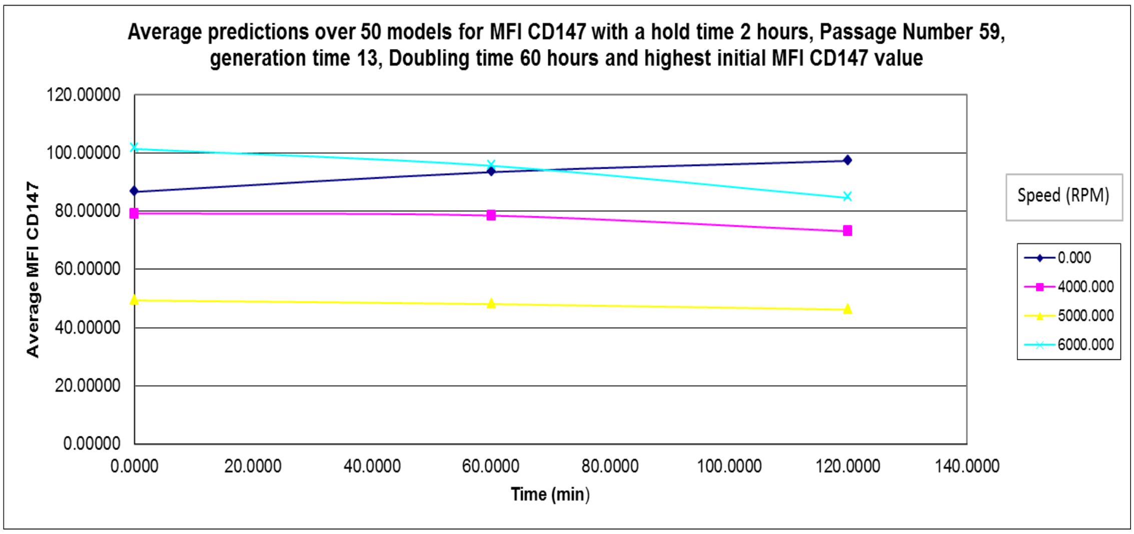
4.2.2. CD147

Graph 43: *Graph showing the predicted MFI CD147 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD147", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 2 hours with a low initial marker level*



Graph 43 shows the average results over 50 models from the ANN predicting MFI CD147 with a low initial marker level before the culture is subjected to shear forces. The graph displays a steady decline for the MFI CD147 marker sample that was not subjected to any rotational shear. For all other shear speeds the samples where predicted to increase over the time period with the 6000 rpm predicted to have the largest increase in the MFI CD147 marker level.

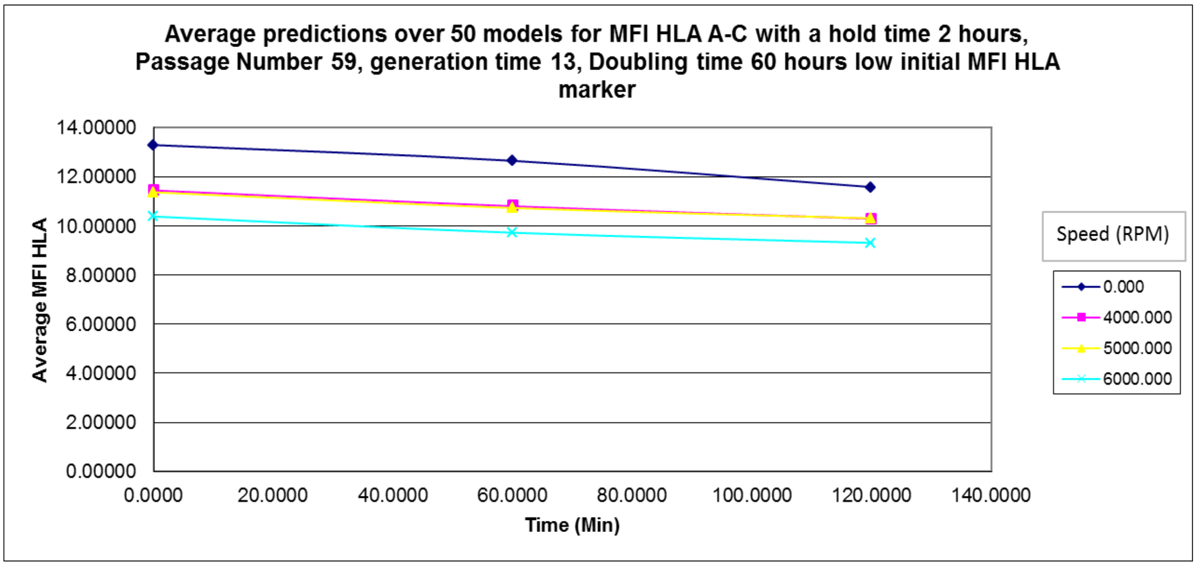
Graph 44: *Graph showing the predicted MFI CD147 over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI CD147", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 2 hours with a high initial marker level*



Graph 44 shows the average results over 50 models from the ANN predicting MFI CD147 with a high initial marker level before the culture is subjected to shear forces displays a steady increase for the MFI CD147 marker that was not subjected to any rotational shear speeds with the 4000, 5000 and 6000 showing a decline in the predicted MFI CD147, with the 6000 rpm prediction showing the greatest reduction with the 5000 rpm sample showing the smallest reduction.

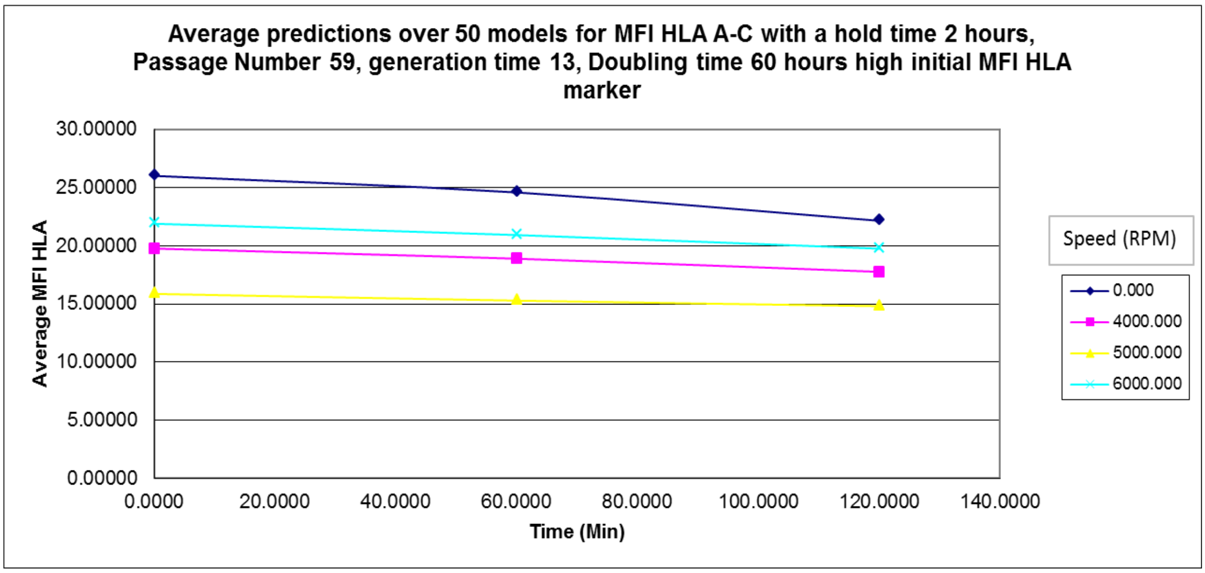
4.2.3. HLA A-C

Graph 45: *Graph showing the predicted MFI HLA A-C over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 2 hours with a low initial marker level*



Graph 45 plots the average results over 50 models from the ANN predicting MFI HLA A-C with a low initial marker level. The graph displays a steady decrease for all the MFI CD147 markers over each shear speed over the time, with the 6000 rpm prediction showing the smallest reduction over the time period

Graph 46: *Graph showing the predicted MFI HLA A-C over varying shear speeds, using the inputs of "Shear speed (rpm)", "initial MFI HLA A-C", "Hold time", "Passage Number", "Generation Time", "Doubling time" and "shear time (min)" for a hold time of 2 hours with a high initial marker level*



Graph 46 shows the average results over 50 models from the ANN predicting MFI HLA A-C with a high initial marker level before the culture was subjected to shear forces. The graph displays a steady decrease for the MFI HLA A-C marker that was not subjected to any rotational shear speeds with the 4000, 5000 and 6000 showing a decline in the predicted MFI CD147, with the 6000 rpm prediction showing the greatest reduction with the 5000 rpm sample showing the smallest reduction.

5. Discussion

5.1. Model Development Discussion

5.1.1. CD9

The first model created to predict the MFI for surface marker CD9 only used two inputs; “shear speed” and “shear time” produced an R squared value of 0.26, indicating that the likelihood of this outcome happening to be quite low. This was due to a large amount of variance between the data sets over the same speed and time values.

The input “size of cell” increased the R Squared value of the model to 0.31 an increase of 0.05, the addition of this input into the model aided in explaining some of the variance in the data. This small increase of the R squared value would indicate that the size of the cell has little impact on surface marker presentation, however according to Olcay (*et al,* 1999) other surface markers such as CD3, CD7 and CD10 increased and decreased as to cell size. This could indicate that whilst the size of the cell may have an influence on surface markers, other factor(s) have a bigger influence in this range of experiments.

Adding of the input “MFI IgG” to the previous MFI CD9 model increased the R Squared value by 0.0057 to 0.3157, indicating that this input had little impact on the predictive power of the model, however, this input did aid in explaining some of the variance in the model. This could indicate that during the stressing procedure, the cells surface could be presenting proteins this antibody isotype can latch onto, but this surface protein is not directly linked to CD9 expression levels

The “experiment” input, when added to the other inputs for the MFI CD9 model increased the R Squared value produced by the model to 0.6241, an increase of 0.3084. Compared to the previous inputs this single input doubled the accuracy of the model, indicating that a substantial portion of the variance in the data is related to this one input. The experiment inputs explanation of the large amount of variance in the model would indicate that there is a further parameter within the experimental setup that would need to be investigated further.

From the initial analysis which used the following inputs used to build the network, “Shear speed (rpm)”,”size of cell”, “MFI IgG”, “experiment” and “shear time (min)”, as inputs where added to each model, the accuracy of the average predictions over the 50 models increased, indicated by the increase in the R squared value in each model, when this predicted value was compared to the observed value from the biological data when predicting MFI CD9

5.1.2. CD147

Initially this ANN model was created using the inputs of Shear Speed and Shear time produced an R squared value of 0.067, indicating that the likelihood of this outcome happening to be very low. This was due to a large amount of variance between the data sets over the same speed and time values.

The “size of cell” input increased the R Squared value of the model to 0.005 a decrease of 0.062, the addition of this input into the model lowered the R Squared value for this model, compared to the previous model, this input added noise to the model, again indicating that the size of the cell does not have a direct impact on CD147 surface marker expression

Adding the input “MFI IgG” to the previous MFI CD147 model increased the R Squared value by 0.0116 to 0.0216, indicating that this input had very little impact on the predictive power of the model, however, this input did aid in explaining some of the variance in the model. Again, the results would indicate that IgG attachment levels on the cells surface is not directly linked to the CD147 expression levels.

The addition of the input “experiment” to the previous MFI CD147 model increased the R Squared value produced by the model to 0.2017, an increase of 0.1801. Compared to the previous inputs this single input increased the accuracy of the model significantly, indicating that a substantial portion of the variance in the data is related to this one input, this is in a similar trend to the addition of this input to the CD9 model.

From the initial analysis which used the following inputs used to build the network, “Shear speed (rpm)”,”size of cell”, “MFI IgG”, “experiment” and “shear time (min)”, as inputs where added to each model, the accuracy of the average predictions over the 50 models increased, indicated by the increase in the R squared value in each model, when this predicted value was compared to the observed value from the biological data when predicting MFI CD147

5.1.3. HLA A-C

The initial ANN model created using the inputs of Shear Speed and Shear time produced an R Squared value of 0.0894, indicating that the likelihood of this outcome happening to be very low. This was due to a large amount of variance between the data sets over the same speed and time values.

The addition of the input "size of cell" increased the R Squared value of the model to 0.0985 an increase of 0.0091, the addition of this input into the model the R Squared value for this model, compared to the previous model, this input added noise to the model. Similarly to CD9 and CD147, the size of the cell is not a direct factor to the expression levels for these surface markers.

Adding the input "MFI IgG" to the previous MFI HLA A-C model increased the R Squared value by 0.013 to 0.1115, this input had very little impact on the predictive power of the model, however, this input aided in explaining some of the variance in the model.

The "experiment" input when added to the previous MFI HLA A-C model, increased the R Squared value produced by the model to 0.4062, an increase of 0.2947. Again the addition of this one single input increased the accuracy of the model significantly, this is a good indication that a large amount of the variance in the data is related to this one input, again this is in a similar trend to the addition of this input to the CD9 and CD147 models.

The results above, with the addition of extra inputs from the base shear speed and shear time model, there was still a large amount of variance in the model that could not be explained by the inputs. However, it was noted that the "experiment" input which categorised the experiments into groups A, B, C and D in the biological data did explain a significant amount of the variance, indicating there was some inter experimental variations that needed to be investigated further.

From the initial analysis which used the following inputs used to build the network, “Shear speed (rpm)”,”size of cell”, “MFI IgG”, “experiment” and “shear time (min)”, as inputs where added to each model, the accuracy of the average predictions over the 50 models increased, indicated by the increase in the R squared value in each model, when this predicted value was compared to the observed value from the biological data when predicting MFI HLA A-C

5.2 Model Prediction Discussion

5.2.1 CD9

The first model in the stepwise process created for surface marker CD9 using only 3 inputs of “shear speed”, “Shear time” and “initial MFI CD9” produced an r squared value of 0.6232, indicating that the likelihood of this outcome to be significantly higher than that of the original model for CD9 which only used the two inputs of “Shear time” and “Shear Speed”. The initial marker input has already explained almost all of the variance in the data set as the model containing the inputs “Shear speed (rpm)”,”size of cell”, “MFI IgG”, “experiment” and “shear time (min)”. These results indicate that the surface marker profile of the cell before stress is applied can influence how the cells surface reacts to the stress.

The input “hold time” when added as an input to the model, decreased the R Squared value of the model to 0.6295, by 0.0033 which when compared to the previous model, it would appear that this input has only added noise to the model. From this it would indicate that the hold time for this cell culture has little impact on the surface marker profile during a stress event. Wymer (*et al*, 2010) noted that a shorted hold time provided a small but statistically significant in qPCR estimated densities of their samples when compared to a long hold time process

Adding the input “Passage Number” to the previous MFI CD9 model increased the R Squared value by 0.0664 to 0.6959, indicating that this input had a small impact on the predictive power of the model. Siissalo (*et al*, 2007) stated that in CACO-2 cells passage number has been shown to influence the expression levels of MDR1, with its expression being high in low passage numbers and the dropping as the passage number is increased.

The addition of the inputs “generation time” and “doubling time to the previous MFI CD9 model increased the R Squared value to 0.8308, indicating that this input had a large impact on the predictive power of the model, as the R Squared value increased by 0.1349. This would indicate that these two factors have a high influence on the surface profile of the cell. Generation time is stated as the interval between the divisions of a cell

5.2.2. CD147

Again the initial model created for surface marker CD147 using only 3 inputs of "shear speed", "Shear time" and "initial MFI CD147" produced an r squared value of 0.7475. The initial surface marker value once again explained almost all of the variance in the data set as the model containing the inputs "Shear speed (rpm)","size of cell", "MFI IgG", "experiment" and "shear time (min)". Again this result indicates that the initial surface marker expression levels largely govern the surface marker profile after stress.

Adding the input "hold time" to the model increased the R Squared value of the model to 0.7707, the addition of this input into the model increased the R-Squared value by 0.0232, a small increase, again indicating that the hold time has little influence over surface marker expression

The input "Passage Number" when added as an input to the previous MFI CD147 model increased the R Squared value to 0.7896. This input had an impact on the predictive power of the model, raised the R squared value by 0.0189. This would indicate that, again, the passage number of the culture has an impact of the ability of the cell to present surface markers, however the passage number has more of an influence over the CD147 marker then CD9

The addition of the inputs "generation time" and "doubling time to the previous MFI CD147 model increased the R Squared value to 0.8073, indicating that this input had a positive effect on the predictive power of the model, raising the R Squared value by 0.177

The significance of the initial MFI CD147 marker in the predictive accuracy of the model is shown in the sensitivity analysis from the models; the marker "initial MFI CD147" had the highest sensitivity value in each of the models predicting the CD147 marker. This indicated that the initial concentration of this marker on the cells surface before subjecting the cell to rotational shear stress can be an indicator as to how the surface marker profile would change during the stress period.

5.2.3 HLA A-C

The initial model created for surface marker HLA A-C using only 3 inputs of "shear speed", "Shear time" and "initial MFI HLA A-C" produced an r squared value of 0.6959, indicating that the likelihood of this outcome to be significantly higher than that of the original model for HLA A-C which only used the two inputs of "Shear time" and "Shear Speed" that had an R Squared value of 0.0894. The initial marker input has already explained almost all of the variance in the data set as the model containing the inputs "Shear speed (rpm)","size of cell", "MFI IgG", "experiment" and "shear time (min)". Again this noticeable increase in the R Squared value indicates that the initial surface marker expression level is a key factor over HLA A-C expression levels during stress

The addition of the "hold time" input raised the R Squared value of the model to 0.754, the addition of this input into the model the R Squared value for this model. Again, indicating that the hold time has some influence over the HLA surface marker expression during stress

When the input "Passage Number" was added to the previous MFI HLA A-C model it increased the R Squared value to 0.7813, this input had a positive impact on the predictive power and accuracy of the model, showing that passage number can also influence the surface marker profile for HLA during stress

Adding the inputs "generation time" and "doubling time" to the model increased the R Squared value to 0.7987, this would indicate that the input had a positive impact on the predictive power and therefore accuracy of the model. The generation and doubling time of the cells in this instance do have a slight influence of the surface marker profile for HLA A-C

The sensitivity analysis values for the MFI HLA A-C marker shows the marker "initial MFI HLA A-C" had the highest sensitivity value in each of the models predicting the HLA marker. This indicated that the initial concentration of this marker on the cells surface before subjecting the cell to rotational shear stress can be an indicator as to how the surface marker profile would change during the stress period.

5.3. Secondary analysis and model interpretation discussion (prediction of the influence of cellular stress)

The models created in this section of the results used the marker set that gave the greatest R Squared value (and therefore highest accuracy of the models) from the previous set of experiments.

These models where created to examine the effect of changing the experimental parameters would have on the surface marker profile of the whole cell vaccine, the graphs indicate the predicted levels of the surface marker for each shear speed over the time course of the experiment.

5.3.1 Hold time

These models were created using the markers *“Shear speed (rpm)”, “initial MFI marker”, “Hold time”, “Passage Number”, “Generation Time”, “Doubling time” and “shear time (min)”*. All of the experimental parameters remained the same; the ANN was then asked to predict the MFI surface marker values for hold times of 1 and 3 hours.

From the results shown in Graphs 26 and 27, when increasing the hold time for the cell culture the model predicted that the surface marker CD9 has a marginally faster decline over time for all shear speeds; however, as this decline also happens in cells that are not subjected to any shear stresses, it would indicate that the CD9 level of expressions is not directly linked to cell stress factors, The model also predicted that the cells have a similar starting concentration for each hold time.

For CD147, the model showed an increase in the hold time for the experiment does not appear to impact the surface marker levels over the shear time period for each of the shear speeds. From these results it would indicate that CD147 surface marker production is stimulated when the cells were subjected to rotational shear stress. This can be seen in the difference between the non-shear and sheared cells having a clear difference in marker profile over the time period, the cells subjected to no shear stress showed a decline in the CD147 marker levels, whereas the cells subjected to the shear stress’s marker levels stayed constant or increased marginally. The cells that have a longer hold time appear to have a higher starting concentration of the surface maker. This would indicate that under stress factors the

CD147 marker is produced by the cell, possibly to help maintain vital protein families within the cell (Schlegel *et al*, 2009), possibly related to membrane integrity. During shear stresses it is likely the cell membrane will become damaged and will need repairing, it is possible that the CD147 marker is liked with a cellular repair system

The surface marker HLA A-C, from the predictions shown in graphs 36 and 37 indicate that a longer hold time causes the HLA A-C surface marker level to fall slightly compared to a shorter hold time when exposed to rotational stresses remained almost stationary. However the model did indicate that longer hold time increased the starting concentration of the HLA surface marker at time point zero minutes. The cell culture that was not subjected to shear stress showed a decline of the HLA A-C surface marker. This would indicate that the HLA A-C marker is not directly linked with any cellular processes that are linked to stress as there is virtually no change between the stressed and unstressed cells.

From the predictions this could indicate that the HLA A-C surface marker could be linked with a cells ability to adapt to its surrounding conditions, as the marker level was maintained in the cell culture that had a short hold time, but showed a decline for each shear speed when the hold time was extended to 3 hours. The cell culture that was not exposed to any shear stress showed a similar decline for each hold time over the 120 minutes.

5.3.2 Passage number and generation time

From the models generated for the CD9 surface marker shown in graphs 28, 29 and 30 as the passage number and generation time was increased the cell cultured showed little variation in the decline of the surface marker over time caused by the change over each of the shear speeds. The model indicated again that the CD9 surface marker is not directly linked to cell stress responses. As the generation time and passage number increases the starting values of this marker drop marginally.

For CD147 for each of the samples subjected to shear in graphs 33, 34 and 35 as the generation time and passage number was increased, the model predicted an increase of the surface marker over the time. With the predictions for passage number 66 and a generation time of 18 having the largest increase. The model predicted that the culture not subjected to the shear stress still showed a decline over the time period. These results could indicate that the marker is again linked to a stress response, and to the physical condition of the cell.

In the model predicting the change for HLA A-C, as the passage number and generation time is increased the model showed that there would be no visible change in the reduction of the surface markers concentration over time for any of the shear stresses, including the samples that were not subjected to any rotational shear stress.

5.4 Effect of High / Low initial marker levels on model outcome

Within the data sets it was noted that some of the cell cultures / samples had varying MFI levels per batch going into the experiment, this series of predictions where to see what the model would predict under the highest and lowest values of these markers, and to see if these levels would have an effect on the cells surface marker levels during the experiments.

5.4.1 CD9

The model predicted that there would be no change in the reduction of the CD9 marker level on the cells surface over the time period for each of the rotational shear stress speeds, just an increase in the initial starting value of the marker at time point zero. This would indicate that the production of this marker is not linked to any cell regulation cycles. This could be useful in a whole cell vaccine as the culture process would not alter the levels of the marker on the cell.

5.4.2 HLA A-C

The model predictions for HLA A-C are similar to that of the CD9 marker levels, the marker levels continue to reduce over the time course, including the cells not subjected to shear. This could be useful in a whole cell vaccine as the culture process would not alter the levels of the marker on the cell.

5.4.3 CD147

For the CD147 marker the model predicted a noticeable change in the behaviour of the surface marker levels over the time period. For the low initial starting value, the ANN predicted that the surface marker would increase marginally over the 120 minutes, were as the cells not subjected to shear would show a steady decline in their CD147 surface marker levels, however for the cell culture with a high starting CD147 cell surface marker level, those cells NOT subjected to any shear stresses is now predicted to increase.

6. Conclusion

Overall it can be noted that the experiment was a success, from the results of this experiment it can be seen that using Artificial Neural Networks to model the proteomics of cellular stress. The mathematical modelling software can be used to predict the impact of changes to the experimental or cellular conditions to predict the MFI surface marker levels for CD9, CD147, and HLA A-C. Whilst the addition of extra inputs from the base shear speed and shear time model, there was still a large amount of variance in the model that could not be explained by the inputs used to create the model.

However, it was noted that the “experiment” input which categorised the experiments into groups A, B, C and D in the biological data did explain a significant amount of the variance in the initial model set, indicating there was some inter experimental variations that needed to be investigated further. From the data, the models accuracy can be improved dramatically by determining the initial quantity of the cells surface markers, this theory was proven in the sensitivity analysis were the initial value marker for all of the surface markers came out with the highest sensitivity value. This would indicate that the ability of the cell to withstand rotational stresses can be directly linked with the way the cells are cultured and must be harvested at an optimum time to be fully effective with the correct levels of surface markers desired

For the second phase of experiments, it can also be noted that the experiment was a success. The model was able to predict what any changes to the given experimental parameters would do to the surface marker profile of the cells using the markers from the previous experiment. Changes in the physical parameters of the experiment had varying effects on the cells, depending on their initial surface marker profile at the start. From the predictive model results it indicates that CD9 and HLA A-C expression is not linked to cell stress pathways, but can be linked to the cells passage number / generation time, a cell batch with a higher passage number appears to have a slower decrease in the respective surface marker under shear conditions. Whereas CD147 appears to be directly linked with cellular stress as the expression levels for this surface marker rise when stressed compared to a non-stressed cell culture. CD147 expression can also be linked to the passage number / generation, as the model predicted that as the passage number is increased, the rate of expression increases under stress.

This could indicate that this profile can also be used to indicate the quality of the cell and therefore its viability, with a possible link to its potency (i.e. how effective) if this whole cell vaccine was to be administered into a patient.

7. Future Work

The first key piece of any work regarding this series of experiments would be to examine if the predictions made by the model in graphs 26 to 46 are correct and their level of accuracy. This would involve having to create lab experiments to match the model criteria exactly. I would especially recommend further investigation into the CD147 low / high initial marker model show in graphs 43 and 44, as the results seem to move dramatically away from each other for the cells not subjected to any shear stresses. If this is the case with the expression levels of the CD147 marker, another series of experiments could be run to determine the point at which the surface marker starts to increase by itself rather than decrease. The accuracy of the model could be examined and further changes to the inputs for the model can be added to increase accuracy if needed.

The experiment can be enhanced further by increasing the number of inputs used in the model, incorporating other factors such as temperature, cell concentration and nutrient content of the media. Also increase the doubling time, passage and generation times to see what impact this would have on the outcomes of the experiment. These extra inputs could explain any outstanding variance produced in this experiment's results.

This would need to concentrate on improving the ANN model by testing the model on other biological data sets from several different cell lines noting the culture conditions. This can be used to determine if the surface markers behave in the same way for all cell lines and how these markers react to cellular stress factors. From this work a panel of stress markers could be produced and standardised for use across industry for rotational shear stresses

During the commercial bio manufacturing process, cells are not subjected to just rotational shear, as the cultures are passed into or between bio-reactors, flasks etc, the cells can be subjected to many different types of stress. Another set of experiments that the ANN models can be used for is to study the effect of elongation and compression stress using flow cytometry. This can also be used to look at the flow stresses around an entry point into a continuous centrifuge, as Boychyn (*et al,* 2001) stated that there is great difficulty in predicting the consequences and impact of flow stresses. If the impact of this can be examined and modelled on a cellular level, the centrifuges can be altered to elevate extra stress on the cells.

8. References

Ball G, Al-Sahib A, *Data Mining for Predictive Proteomics,* In: Mass spectrometry for microbial proteomics. EDS Shah H. N. and Gharbia S.E. John Wiley and Sons Ltd., West Sussex United Kingdom. Pp 411 -422.

Ball G, Mian S, Holding F, Allibone O, Lowe J, Ali S, Li G, McCardle S, Ellis I. O, Creaser C, Rees R. C, *An integrated approach utilising artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers,* bioinformatics vol. 18 no 3, pages 395-404, 2002

Ball, G; Palmer-Brown, D; Mills G; "*A Comparison of Artificial Neuronal Networks and controversial statistical techniques for analysing environmental data*", 1999

Banchereau J, Steinman R, M. Dendritic cells and the control of Immunity. Nature, Volume 392, 1998.

Belfiore L A, Bonani W, Leoni M, Beliore C J, *Stress-sensitive nutrient consumption via steady and non-reversing dynamic shear in continuous-flow rotational bioreactors,* Biophysical Chemistry Volume 141, Issues 2-3,Pages 140-152, 2009

Benkö S, Hilkmann H, Vigh L and van Blitterswijk W. J, *Catalytic hydrogenation of fatty acyl chains in plasma membranes; effect on membrane lipid fluidity and expression of cell surface antigens,* Biochimica et Biophysica Acta 896 (1987) 129-135

Boychyn M, Yim S. S. S, Shamlou P. Bulmer A. M, More J, Hoare M. *Characterization of flow intensity in continuous centrifuges for the development of laboratory mimics*. Chemical Engineering Science 56 4759–4770, 2001

Chiampanichayakul S, Peng-in P, Khunkaewla P, Stockinger H, Kasinrerk W, *CD147 contains different bioactive epitopes involving the regulation of cell adhesion and lymphocyte activation,* Immunobiology 211 167–178, 2006

Chhatre S, Titchener-Hooker N.J. *Micro-Biochemical Engineering: Using Small-Scale Devices to Predict Industry-Scale Downstream Performance.* Comprehensive Biotechnology (Second Edition) Volume 2, Pages 891–903, 2011

Dutta K A, Paul S, Chattopadhyay A B, *Applicability of the modified back-propagation algorithm in tool condition monitoring for faster convergence*, Journal of Materials Processing Technology 98 299-309, 2000

Fan J, Zhu G. Z, Niles R. M, *Expression and Function of CD9 in Melanoma Cells.* Mol Carcinog; 49(1): 85–93, 2010

Gatto J. G*, The changing face of bioinformatics,* DDT Vol. 8, No. 9 2003

Hartmann B.M, Kaar W, Falconer R.J, Middelberg A.P.J. *Chromatography-free peptide bio manufacture (2008).* Centre for Biomolecular Engineering, Australian Institute for Bioengineering and Nanotechnology, University of Queensland, Brisbane, Australia

Hori H, Yano S, Koufuji K, Takeda J and Shirouzu K, *CD9 expression in gastric cancer and its significance*, Journal of Surgical Research, Volume 117, Issue 2, Pages 208-215, 2004

Huang Z, Chen H, Hs C, Chen W, and Wu S, *Credit rating analysis with support vector machines and neural networks: a market comparative study*, Decision Support Systems Volume 37, Issue 4, Pages 543-558, 2004

Iacono K, Brown A L, Greene M I, Saouaf S J, *CD147 immunoglobulin superfamily receptor function and role in pathology*, Experimental and Molecular Pathology 83 283–295, 2007

Jager M. J, Monique H, Hurks H, Levitskaya J, and Kiessling R, *HLA Expression in Uveal Melanoma: There is No Rule Without Some Exception,* Human Immunology 63, 444–451, 2002

Knight S. C, Burke F, Bedford P. A, *Dendritic cells, antigen distribution and the initiation of primary immune responses to self and non-self-antigens.* Cancer Biology, Vol. 12, pp. 301–308 2002

Lancashire L J, Lemetre C, Ball G R, *An introduction to artificial neural networks in bioinformatics – application to complex micro array and mass spectrometry datasets in cancer studies*, Briefings in bioinformatics, 2009

McCoy R, Hoare M, Ward S, *Ultra Scale-Down studies of the effect of shear on cell quality; processing of a human cell line for cancer vaccine therapy,* American Institute of Chemical Engineers DOI 10.1002/btpr.229 (24th July 2009

McQueen A, Meilhoc E, and Bailey J. E, *Flow effects on the viability and lysis of suspended mammalian cells*, Biotechnology Letters Vol 9 No 12 831-836, 1987

Miyado K, Yamada G, Yamada S, Hasuwa H, Nakamura Y, Ryu F, Suzuki K2, Kosai K, Inoue K, Ogura A, Okabe M, Mekada E. *Requirement of CD9 on the Egg Plasma Membrane for Fertilization*. Science; Vol. 287 no. 5451 pp. 321-324

Mosca, P. J, Hobeika A. C, Clay T. M, Morse M. A, Lyerly H. K. *Dendritic cell vaccines*. Frontiers in Bioscience 12, 4050-4060, 2007

Motobu M, Wang P, Matsumura M, *Effect of Shear Stress on Recombinant Chinese Hamster Ovary Cells*, Journal of Fermentation and Bioengineering Vol. 85, No. 2, 190-195. 1998

Noble J. A, Valdes A. M, Bugawan T. L, Apple R, J, Thomson G, and Erlich H. A. *The HLA Class I A Locus Affects Susceptibility to Type 1 Diabetes*. Human Immunology; Volume 63, Issue 8, August 2002, Pages 657–664

Olcay L, Ertem U, Okur H, Tuncer A. M. *The importance of cell size and surface marker analysis in childhood acute myeloblastic leukemia*. Leukemia Research 23, 701–707, 1999

Peters J, De Baets B, Verhoest N. E. C, Samson R, Degroeve S, De Becker P and Huybrechts W, *Random forests as a tool for ecohydrological distribution modelling*, Ecological Modelling Volume 207, Issues 2-4, 10 2007, Pages 304-318

Petricciani J, Egan W, Vicari G, Furesz J, Schild G, *Potency assays for therapeutic live whole cell cancer vaccines*, Biologicals 35, 107e113, 2007

Plumb P A, Rowe R C, York P, Brown M, *Optimisation of the predictive ability of artificial neural network (ANN) models: A comparison of three ANN programs and four classes of training algorithm*, European Journal of Pharmaceutical Sciences Volume 25, Issues 4-5, Pages 395-405, 2005

Ramal L M, Maleno I, Cabrera T, Collado A, Ferron A, Lopez-Nevot M A, Garrido F, *Molecular strategies to define HLA haplotype loss in micro-dissected tumour cells,* Human Immunology Volume 61, Issue 10, Pages 1001-1012, 2000

Redzic J. S, Armstrong G. S, Isern N. G, Jones D. N. M, Kieft J. S, Elsenmesser E Z. *The Retinal Specific CD147 Ig0 Domain: From Molecular Structure to Biological Activity*. Journal of Molecular Biology, 411, 68–82, 2011

Rumelhart D E, Hinton G.E, Williams R. J, *Learning representations by back propagating errors*. Nature 323: 533 – 536, 1986

Saltelli A, Andres T.H, Homma T. *Sensitivity analysis of model output: An investigation of new techniques (2002).* Japan Atomic Energy Research Institute of Tokai-Mura

Schlegel J, Redzic J. S, Porter C. C, Yurchenko V, Bukrinsky M, Labeikovsky W, Armstrong G. S, Zhang F, Isern N. G, DeGregori1 J, Hodges R and Eisenmesser E. Z, *Solution Characterization of the Extracellular Region of CD147 and Its Interaction with Its Enzyme Ligand Cyclophilin A*, Journal of Molecular Biology Volume 391, Issue 3, 21, Pages 518-535 2009

Siissalo S, Laitinen L, Koljonen M, Vellonen K. S,Kortejärvi H, Urtti A, Hirvonen J, Kaukonen A. M. *Effect of cell differentiation and passage number on the expression of efflux proteins in wild type and vinblastine-induced Caco-2 cell line.* European Journal of Pharmaceutics and Biopharmaceutics 67, 548–554, 2007

Wymer L, Oshima K, Paar J, Doolittle M, Lavender J, Varma M, Haugland R. *Effects of Holding Time, Storage, and the Preservation of Samples on Sample Integrity for the Detection of Fecal Indicator Bacteria by Quantitative Polymerase Chain Reaction (qPCR)-based Assays.* US Environmental Protection Agency, 600/R-10/150, 2010

Zhang H, Hoare M. *Characterization of gas–liquid interface in an ultra-scale-down device and its implications for protein formulation.* Journal of Biotechnology 136S, S460–S495 2008

Zvierieva V, Wangb J, Chevrettec M. *Over-expression of CD9 does not affect in vivo tumorigenic or metastatic properties of human prostate cancer cells*. Biochemical and Biophysical Research Communications, Volume 337, Issue 2, Pages 498–504, 2005