**Green A, Glassey J. [Multivariate analysis of the effect of operating conditions on hybridoma cell metabolism and glycosylation of produced antibody](#).
Journal of Chemical Technology and Biotechnology 2015, 90(2), 303-313.**

Copyright:

This is the peer reviewed version of the following article: Green A, Glassey J. [Multivariate analysis of the effect of operating conditions on hybridoma cell metabolism and glycosylation of produced antibody](#). *Journal of Chemical Technology and Biotechnology* 2015, 90(2), 303-313. which has been published in final form at <http://dx.doi.org/10.1002/jctb.4481> This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

DOI link to article:

<http://dx.doi.org/10.1002/jctb.4481>

Date deposited:

19/05/2015

Embargo release date:

01 August 2015

This work is licensed under a [Creative Commons Attribution-NonCommercial 3.0 Unported License](#)

**Multivariate analysis of the effect of operating conditions
on hybridoma cell metabolism and glycosylation of produced antibody.**

Short title: Effect of operating conditions on mAb glycosylation

Amy Green and Jarka Glassey

School of Chemical Engineering and Advanced Materials, Newcastle University,
Newcastle-Upon-Tyne, United Kingdom NE1 7RU

Corresponding author: Jarka Glassey, School of Chemical Engineering and
Advanced Materials, Newcastle University, Newcastle-Upon-Tyne, United Kingdom
NE1 7RU.

Email: jarka.glassey@ncl.ac.uk

Key words: Principal Component Analysis; Partial Least Squares; glycosylation;
Chinese hamster ovary cells; bioprocess modelling

Abstract

BACKGROUND: Changes in glycosylation profiles of monoclonal antibodies can have significant impact upon the product quality. Control of critical process parameters in order to ensure consistent product quality is one of the core requirements of the FDA's QbD and PAT initiatives. The effect of operating conditions upon cell metabolism and the glycosylation profile of monoclonal antibody produced using hybridoma cell culture is investigated in this report.

RESULTS: PCA analysis of on-line process data and amino acid concentration profiles reflecting the cell metabolism indicates significant dependence on the operating conditions, particularly DO and pH. PLS models predicting product titre based on amino acid concentration and those predicting the glycosylation profile either based on product titre and glucose concentration or on amino acid concentrations show that both titre and glycosylation can be predicted with satisfactory accuracy for a range of operating conditions.

CONCLUSIONS: Accurate prediction of product titre and glycosylation profile based on amino acid concentration and process variables easily measured in real time opens up the opportunity to control the product quantity and quality during cultivation using critical process parameters, such as pH and DO.

1 **Introduction**

2 In 2004 the FDA (Food and Drug Administration) introduced the Process Analytical
3 Technology (PAT) guidance and Quality by Design (QbD) initiative. They aim for the
4 biopharmaceutical industry to implement a system for designing, analysing and
5 controlling critical quality attributes (CQAs) through critical process parameters
6 (CPPs) to ensure final product quality and increase yields (1). The identification and
7 subsequent control of the parameters which have the greatest influence over the
8 system, to establish robust and well controlled processes, are some of the main
9 challenges faced by industry. QbD and PAT have been applied to laboratory scale
10 fermentations of bacterial expression systems showing the benefits of applying these
11 approaches to the more complex mammalian-cell expression systems (2-4).
12 Historically, development of mammalian cell culture has focused on cell line and
13 culture improvements and the PAT principles have yet to be widely adopted in this
14 area by the biopharmaceutical industry (3, 4).

15 Multivariate data analysis (MVDA, see below) has been shown to be able to extract
16 important process information from data sets and to reduce the complexity of the
17 data set by eliminating co-linearity and noise. (4-7). Given these characteristics,
18 MVDA has the potential to help in the identification of CQAs in line with the QbD and
19 PAT requirements. However, the process/product complexity, changing cultivation
20 environment over the cultivation trajectory and the lack of direct signal to
21 physiological mechanism relevance increases the challenge in applying PAT
22 methods (8).

23 On the other hand, recent reports in the literature indicate increasing levels of
24 understanding the cellular metabolism and the impact of the changes in

1 environmental conditions both on cell metabolism and product quality (9-13). This
2 strengthens the motivation to develop accurate dynamic models combining the
3 proven ability of data-based models to predict important process variables based on
4 process data readily measured during the cultivation with first principles
5 understanding of cell metabolism and product synthesis.

6 Recent studies, e.g. Grainger and James, describe the effect of media composition
7 upon glycosylation (14). The authors demonstrated that through response surface
8 modelling of a design of experiments (DoE), predictions can be made of the
9 response of the glycosylation. This suggests that further information could be gained
10 from adopting a statistical analysis of culture conditions. Currently modelling efforts
11 are focused on the characterisation of the metabolic pathways associated with
12 glycosylation (15). In addition to this a recent study by Amand et al. has shown that
13 whilst there are currently no online control schemes in place, there is the potential for
14 development of online glycosylation control (16).

15 The research reported here investigates the link between process operating
16 conditions, cell metabolism and the production of different glycosylated forms of a
17 monoclonal antibody (mAb) produced by a murine hybridoma cell line as a first step
18 towards formulating such combined hybrid models (17). A set of 12 cultivations was
19 selected for model development. Advanced data analysis methods as described
20 below were used initially to provide indications of the influence of operating
21 conditions upon cell growth, product yield and quality. Relationships between the
22 metabolism, as reflected by the amino acid concentrations throughout the cultivation,
23 and the process performance were then explored with the aim of ultimately
24 combining first principle knowledge and data-based models in a hybrid modelling
25 structure providing a more accurate process representation. Within the multivariate

1 techniques Parallel Factor Analysis (PARAFAC) and Multi-way Principal Component
2 Analysis (PCA) were applied to analyse the interactions within the cultivation. Partial
3 Least Squares (PLS) and Multi-way PLS were then applied to investigate the
4 influence of the controlled cultivation conditions upon the product quality
5 (glycosylation profile) and quantity (titre).

6 **Overview of data analysis techniques**

7 Three complementary techniques were used to enable the investigation of the
8 relationships between the different measured variables and subsequently the
9 prediction of the mAb glycosylation pattern.

10 **Parallel Factor Analysis (PARAFAC)**

11 PARAFAC is a generalisation of PCA for higher order arrays (18). With PARAFAC
12 there is no rotation problem, which is beneficial for data sets such as that being
13 investigated in this study where batch length and sampling intervals are irregular.
14 PARAFAC allows for the visualisation of the data in a way that is relative to the raw
15 information, therefore making it useful as a precursor to Principal Component
16 Analysis (18). Details of the methodology for the decomposition of the data into the
17 components can be found in the work of Rasmus Bro (18).

18 **Principal Component Analysis (PCA) and Multi-way PCA (MPCA)**

19 PCA reduces the dimensionality of the data set through turning the noisy data and
20 correlated process measurements into a smaller set of latent variables. These latent
21 variables (*principal components*) are expressions of the linear combinations of the
22 input variables and the corresponding weights (*loadings*). The first few principal
23 components (PCs) capture most of the variance and if the original measurements
24 are highly correlated then the data set can be represented using fewer PCs without a

significant loss of information. In this contribution a multi-way PCA (MPCA) is used to account for the batch character of the process data collected (19).

This study uses time trajectories of various online and offline measurements (as described in the methods section) measured from numerous cultivations (see Table 1). Thus the data represent a three way array with time, variables and cultivation runs representing the three dimensions of the array. There are various ways in which this array can be unfolded to form a two dimensional matrix for classical PCA analysis. Albert and Kinley discuss the advantages of each method in more detail (6).

Partial Least Squares (PLS) and Multi-way PLS (MPLS)

PLS is able to cope with collinear input and output variables and consider the covariance in the data. PLS is closely related to PCA, although as a regression technique, it simultaneously finds latent variables for both the input and the output measurements that best capture the influence of the variability in the input data upon the process output. The main advantage of PLS is that outputs of the system can be calculated from only a few latent variables (LVs), which capture the maximum variance in the system.

The main difference between PLS and PCA is that in PLS every observation in the data sets is represented in both the input and output space. The scores are calculated so that they maximise the approximation of both the input and the output data sets and furthermore so that the correlation between them is maximised. The loadings in PLS provide information about how the variables combine to give the scores values. They are an indication of how important these variables are in the prediction of the modelled process outputs.

Similarly as in the case of MPCA, MPLS decomposes the three way array into a two dimensional matrix. Nomikos et al. and Cunha et al. present the details of the methodology for the decomposition and for correlating the resulting matrices (20, 21).

Methodology

Experimental procedure

A hybridoma cell line (ATCC CRL-1606) was used in this work. The cell line was adapted to chemically defined culture media (TurboDoma TP6, Cell Culture Technologies) and cultivated in controlled parallel bioreactors (DasGip) in batch mode. The culture conditions have been reported earlier (22). Briefly, culture environment was controlled at 37 °C, dissolved oxygen (DO) was set to 50% air saturation and controlled by a constant gas inlet flow rate of 0.05 vvm. pH was equal to 7.2, stirrer speed set to 150 rpm and osmolarity around 320 mOsm/kg. A parameter shift of one of the process control parameters was performed in the early exponential growth phase as described earlier together with the sampling procedure and analytical methods (22). Viable cell concentration, glucose, lactate and ammonia concentrations, amino acid concentrations, mAb concentrations and the glycosylation profile were measured twice a day as offline data. Briefly, glucose and lactate concentrations were determined enzymatically using a Hitado Super GL compact instrument (Hitado, Germany). Ammonium and glutamine concentrations were determined using the L-Glutamine/Ammonia (Rapid) Assay Kit (Megazyme, Ireland). Samples were spun down and supernatants were stored at -20 °C and later used for measurement of antibody concentration (PA ImmunoDetection® Sensor Cartridge, Applied Biosystems, USA) and N-linked glycosylation analysis. For the

glycosylation analyses, the mAB samples were purified using the Vivapure® miniprepG purification kit (Satorius Stedim Biotech, France) and enzymatically deglycosylated by N-glycosidase F (PNGase F glycerol free) (New England Biolabs, USA). N-linked oligosaccharides were fluorescently labeled, separated using Agilent 1200 HPLC system (Agilent Technologies, USA) equipped with a fluorescence detector for analysis by hydrophilic interaction chromatography (HILIC) using a GlykoSep N-Plus column (4.6 mm x 150 mm, Prozyme, USA) and analysed by Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight (MALDI-TOF) Mass Spectrometry (MS) as detailed previously (22). The profiles of temperature, dissolved oxygen, pH, stirring speed and gas flow rate, together with the gas inlet fractions of oxygen (O₂) for DO control and carbon dioxide (CO₂) and the base addition volume used for pH control were recorded online over the whole culture.

Multivariate methodology

The multivariate statistical analysis carried out in this contribution was performed using the program MATLAB (Math Works, Inc.) with the software PLS toolbox (Eigenvector Research, Inc.).

To perform the PCA analysis 12 experimental cultivations were used. The conditions of the cultivations can be seen in Table 1. The online process data used in the analyses included dissolved oxygen (DO), gas inlet fraction of oxygen (O₂) and carbon dioxide (CO₂), volume of base addition, pH, temperature, stirrer speed, and gas inlet flow rate. The offline variables included viable cell concentration, total cell concentration and titre. The amino acid data included the concentrations of the twenty amino acids and ammonia at 11 sampling time points during the cultivation. Finally the glycosylation was recorded at the end of the cultivation and the concentrations of nine glycan forms were expressed as a percentage of peak area

(see Experimental procedure). The number of PCs and LVs retained in the models was chosen based on the Kaiser-Guttman method (23) to capture the greatest variation in the data whilst maintaining model robustness. According to this method, PCs or LVs with an eigenvalue greater than 1 were retained in the model (see Table 3 for details on the number of PCs and LVs retained in the models). For PCA a comparison of two analyses was performed. The analysis of the whole cultivation data from inoculation to the end of the cultivation was compared with the trends observed from data collected after the shift in the relevant process operating parameter until the end of the cultivation (referred to as 'post shift' hereafter).

The PLS models were developed using the same experimental data with the exception that one of the runs used for PCA model development was instead used as a validation run. Therefore 11 runs were used to build the PLS models (Table 2: runs 1-4 and 6-12) and 2 runs to validate (Table 2: runs 5 and 13). This was caused by the unavailability of glycosylation data for one of the validation runs used in the MPCA analysis. The SIMPLS algorithm (24) within the PLS toolbox with 'leave one out' cross validation method minimising the Root Mean Square Error of Cross Validation (RMSECV) was used. The data was first autoscaled (the mean value of each column removed from each value and divided by the standard deviation of the column). Model performance is reported in terms of average Root Mean Square Error (RMSE) over the duration of the cultivation post shift for the validation batches (Table 2) for the prediction of titre. In the case of the glycosylation prediction, the average RMSE over the two validation runs for each glycan form concentration at the end of the cultivation is reported in Table 3.

RMSE is defined as the measure of the differences between the values predicted by a model and those observed from experimentation.

$$RMSE(\hat{y}) = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (1)$$

where $\hat{y}_i$ is the predicted value of y_i and n is the number of measurements.

This error criterion is frequently used in modelling to report model performance as it has the same units as the quantity being estimated and thus allows comparison of models predicting the same output variable as they all relate to the same y_i .

Results and Discussion

The cultivation conditions for the experimental cultivations used in this report were varied as summarised in Table 1 and 2. There are four groups, each consisting of three runs which investigate the influence of individual process parameters upon the quality and the quantity of the produced mAb.

Parallel Factor Analysis (PARAFAC)

The parallel factor analysis was carried out first to assess the interactions in the raw data using data from runs 1 to 12 described in Table 1. The data was arranged in a three way array with time, batches and variables representing the dimensions, respectively. Where batch lengths were unequal, 'not a number' was entered. To construct the PARAFAC analysis, the PARAFAC function within the PLS toolbox was used with three components. This was selected as the analysis was shown to explain the effects of interest i.e. the constructed model was shown to explain variability in cultivation duration (Figure 1S plate A in the supplementary material), variables (Figure 1S plate B) and batches (Figure 1S plate C). Using more components did not yield any further useful observations.

The analysis indicated that PARAFAC can capture the variation in the system and the controlled parameter shift at 30 hours. Figure 1S plate A demonstrates this shift

in the Mode 1 of the model for each of the three components. PARAFAC was especially beneficial in assessing the quality of the experimental runs as it highlighted specific runs deviating from the typical observed behaviour (as illustrated by Figure 1S plate C). However this deviation is not excessive and can be explained by the different operational conditions each experiment operated under. Figure 1S (C) shows that of the 12 batches used in the analysis (Table 1), batch 9 is highlighted as exhibiting different behaviour. This batch was carried out with a pH shift to a high value at 30 hours. Subsequent analyses also highlighted the different behaviour of this batch (see results of PCA below and a possible hypothesis for this behaviour).

Multi-way Principal Component Analysis (MPCA)

Online data analysis

Figure 1 shows the bivariate scores plots of MPCA analysis based on on-line data only, with each batch represented by one symbol, depending on the operating parameter shift in that batch (see Table 1 for conditions and Figure 1 legend for symbols used for each parameter shift). Figure 1 demonstrates that there is distinct clustering of data recorded from batches carried out under various cultivation conditions. By comparing graphs A-C and D-F it can be seen that the PCA carried out with the data from the whole cultivation and the PCA carried out only using the post shift data present different results due to the variability introduced by pre-shift data. This variability in pre-shift data is caused by slight differences during the inoculation procedure in the bioreactor and results in a growth lag-phase observed in the first 24 hours of the culture. As only the post-shift data is of interest in this case since this study is investigating solely the impact of the change in operating variables upon the process performance, all subsequent discussions will concentrate on the

analysis of post-shift results. The highest DO concentration (90%) is indicated to display a different behaviour compared to the two cultivations carried out at lower DO concentrations (50 and 10%) as it is placed in the opposite quadrant to these (see square symbols in Figure 1 D-F). This appears to indicate a negative influence of high DO values on the process, which is in line with observations of Tummer *et al.* (11). Similar observations can be made in terms of the effect of pH with batches carried out at different pH post shift being placed in opposite quadrants (see circle symbols in Figure 1 D-F). Ivarsson *et al* (22) discuss in detail particularly the negative impact of high pH upon the process performance. PC4-PC6 (not shown) do not show any unexpected results with the cultivations equally distributed and grouped by cultivation condition.

Amino acid data

In order to explore the metabolic activity of the cells during the cultivation, samples were analysed to establish the concentrations of all twenty amino acids and ammonia. In the initial analyses ammonia was indicated as an extreme outlier, limiting the possibility to explore the relationships between the amino acid concentrations. Given the analytical measurement inaccuracies for this variable, it was removed from further analyses.

The PCA analysis of the amino acid data showed that 4 principal components captured over 90% of the variation in the data (PC1 explaining 53.79%, PC2 20.52%, PC3 10.49% and PC4 6.97%). Figure 2 shows the bivariate score plots for the amino acid data. It is interesting to note that experimental runs with DO variations are very tightly clustered showing very similar amino acids patterns independent of DO concentration. Similarly, the cultivations carried out under varying sparger settings are also clustered together which shows this operating parameter

1 does not appear to influence the metabolic state as reflected by the amino acid
2 composition. In contrast the pH cultivations show a spread indicating that the amino
3 acids are sensitive to the operating pH, as reported in the literature previously, e.g.
4 (25). This spread is particularly dominant in the direction of the PC2, indicating that
5 this PC is influenced mainly by the pH. As PC1-3 are all highly influenced by pH, it
6 suggests that pH is a critical factor in amino acid metabolism. Higher PCs (not
7 shown) did not show further significant effects on the analysis.

8 An analysis was then performed to consider individual amino acid profiles per batch
9 to explore the contributions of these towards the patterns described above. Figure 3
10 shows the bivariate plots of coefficients (loadings) of the first three PCs, colour
11 coded according to the experimental run to which they relate (Table 1). The
12 coefficients indicate that the individual amino acids did not show any particular
13 patterning implying that all amino acids exhibit similar influence on the principal
14 components. It is the higher principal components which are of more interest. For
15 PCs 2 and 3 experimental runs 1-3 and 9-12 (DO shift to 50%, 10%, 90% and
16 sparger settings 0.005 vvm, 0.1 vvm and 0.2 vvm, respectively) show distinct
17 grouping, suggesting distinct patterns in amino acid profiles in these batches that are
18 different compared to batches 4-8 (osmolarity shift to 350, 380 and 450 mOsm/kg
19 and pH shift to 7.0 and 7.5, respectively) showing no distinct patterning. Interestingly,
20 cultivation 9 in which pH was shifted to 8.0 behaves differently to the other two pH
21 shift cultivations and this is in line with the observations reported earlier (22). Since
22 PCs 2 and 3 account for approximately 21% and 10% of the variation, respectively,
23 the cause of this variation can still be said to be significant. Therefore, since
24 osmolarity and pH experiments are shown to result in such spread of amino acid
25 data, it suggests that these two process operating parameters need to be optimised

to achieve better control of the metabolic state of the culture. Furthermore, previous studies have shown that DO and sparger do not have a significant impact on growth rate within the ranges investigated in this research (22).

Partial Least Squares (PLS) and Multi-way PLS

The model structures and performance of each of the PLS models developed in this study are summarised in Table 4 showing the input and output data for each model, average RMSE error for the two validation runs over the duration of cultivation post shift and the number of latent variables used in the model. For glycosylation pattern prediction, average RMSE values over the two validation runs are shown individually for each glycan to enable the relative accuracy of the model to be evaluated.

Titre

Partial Least squares (PLS) method was used to investigate whether the amino acid data could be used to predict product titre. Whilst the standard approach of using on-line data to predict the titre was also used (data not reported here), by using amino acid data as model inputs, the relationship between the metabolic state and the product quantity can be explored. The model consisted of four latent variables (LVs), which accounted for 95.83% of the cumulative variance in the input data and 99.41% of the cumulative variance in the output data. Cross-validation results indicated that lower numbers of LVs resulted in a significantly reduced accuracy of prediction.

The model was constructed to predict the titre at various time points throughout the cultivation based on amino acid data only. Figure 4 shows the titre predictions of the resulting models for the two validation runs (circles and squares) together with the experimental error bars. Except for the time points around 10 and 20 hours post shift (sample points 3-5), the model provides satisfactory prediction of the titre based only

on the amino acid concentration data collected post shift until the end of the cultivation. The prediction inaccuracy following the shift in operating parameters may be related to the significant changes in the amino acid profiles post shift, as indicated by the PCA analysis (Figures 2 and 3). In particular the depletion of glutamine at this stage in the cultivation will have a significant impact upon the analysis.

Glycosylation

MPLS was used to investigate the relationship between a matrix consisting of titre (Model 1) and titre and glucose measurements (Model 2) over the duration of the cultivation post-shift and the glycosylation patterns of the product at the end of the cultivation. A model was also constructed considering amino acid data used to predict the concentration of the nine glycan forms measured (Model 3). Finally, a model using on-line data only to predict the concentration of the glycan forms (Model 4) was also developed. In each case the cross-validation determined the optimum number of latent variables to be four and the percentage of the variance in the input and output data is listed in Table 3 for each of these models.

Figures 5 and 6 show the predictions of the glycosylation prediction models (Model 2 and Model 3), respectively, for the two validation runs (open symbols) along with the predictions for the 11 training runs (black symbols) for individual glycan forms (A – I). Model 3 resulted in the lowest RMSE over all the glycan forms for the two validation runs (0.75) compared to Model 2 (1.43), Model 1 (3.82) and Model 4 (13.78). Clearly with only two validation batches it is more difficult to select the ‘best’ performing model as the predictions for one of the validation batches is frequently more accurate than for the other batch. Model 3 and Model 2 predictions were selected here to demonstrate the ability of the Model 3 to predict some of the glycan forms more accurately than models based on titre (with or without glucose) did. Figure 5D

1 illustrates this issue with prediction accuracy in terms of the concentration of
2 glycoprotein G2F, although when amino acid data is used as input instead of the titre
3 and glucose, this inaccuracy is significantly reduced (Figure 6D). This indicates that
4 both titre together with glucose concentration and amino acid concentrations show
5 satisfactory ability to predict the concentrations of the majority of glycan forms of the
6 product at the end of the cultivation. Since glycosylation is directly correlated to the
7 titre and the availability of glucose is a prerequisite for the availability of nucleotide
8 sugar donors, one would expect significantly better prediction accuracy for Model 2
9 compared to the other models. The investigated batch runs, however, are not limited
10 by glucose and mAb productivity is rather low. Therefore, these two parameters do
11 not seem be better suited than the amino acid data to predict the glycosylation data.
12 Although using amino acid data to predict glycosylation would initially appear counter
13 intuitive, the relatively rapid analytical technique compared to the system dynamics
14 would still enable in-time control action, which may prove to be more accurate as it is
15 based on a measurement reflecting a metabolic state of the culture.

16 Ideally, Model 4, based on on-line data only would be most valuable in a PAT based
17 on-line control scheme. However, as Table 3 and Figure 2S (supplementary
18 material) indicate, the on-line data is not sufficient to provide an accurate estimate of
19 the glycosylation profiles. It can be argued that in order to develop models that
20 enable control of the glycosylation profiles using the standard control variables, an
21 in-depth understanding of the mechanisms involved in the cell metabolism is
22 required. First principles models can be used to elucidate the metabolic processes
23 and their impact upon the final glycoprotein production. When combined with
24 multivariate data analysis models, such as PLS demonstrated in this study, the

1 resulting hybrid models have the potential to offer better prediction and control of
2 mammalian cell cultivations (17).

3 Furthermore in comparing the RMSE values shown in Table 4 and the predictions of
4 the models shown in Figures 5, 6, and 2S it can be seen that there might be benefit
5 in constructing models to individually predict each glycoprotein. This would allow for
6 the exploration of issues in prediction accuracy and thus aid in elucidating the
7 underlining mechanisms of formation and the correlations between the input and
8 output variables.

9 **Concluding remarks**

10 This study has presented an analysis and demonstrated the need for monitoring and
11 control of glycosylation in mammalian cell culture where culture conditions might be
12 varied. The exploratory analysis using PARAFAC and (M)PCA techniques indicated
13 that the response of the culture to the shift in individual operating parameters can be
14 captured using these techniques and the emerging multivariate patterns can be
15 correlated to the most influential process parameters. In the online data patterns the
16 shifts in pH and DO resulted in the most distinctive patterns. As far as the impact
17 upon the cell metabolism, as reflected by the amino acid concentration data, is
18 concerned, the clear influence of pH and partially that of osmolarity shift was
19 highlighted, whereas the influence of the dissolved oxygen and sparger setting shifts
20 was not detectable over the range of the settings investigated in this research. PLS
21 and MPLS models have demonstrated the ability to predict the mAb titre from the
22 amino acid data as well as the ability to predict the concentration of various
23 glycosylation forms of the product at the end of the cultivation either based on the
24 titre and glucose concentration data or using the amino acid data. However, the

glycosylation prediction based on on-line data only resulted in a poor model prediction performance. This indicates that relying purely on on-line data may not be sufficient when product quality is to be controlled and additional knowledge is required.

The next stage of this research will explore the incorporation of first principles models via an agent based system (26) to investigate whether better understanding of cellular metabolism can aid in the prediction of glycoproteins. The intention is to use these first principles models to provide the link between the standard online monitoring system and the offline measurements which can be used to control glycosylation. The overall aim of the project is to then take this framework and apply it to both the upstream and the downstream in order to optimise the entire process.

Acknowledgements

Authors are indebted to Dr Miroslav Soos and Marija Ivarsson, ETH Zurich, Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, for the provision of the data used in this study and for invaluable comments and input to the manuscript. Financial support for this project was provided by the Engineering and Physical Sciences Research Council (EPSRC) through the project EP/G037620/1, through the sponsor company Fujifilm Diosynth Biotechnologies.

References

1. FDA. Guidance for industry: PAT - A framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. 2004.
2. Gnath S, Jenzsch M, Simutis R, Lübbert A. Process Analytical Technology (PAT): batch-to-batch reproducibility of fermentation processes by robust process operational design and control. *Journal of biotechnology*. 2007;132(2):180-6.
3. Carrondo M, Alves P, Carinhas N, Glassey J, Hesse F, Merten O-W, et al. How can measurement, monitoring, modeling and control advance cell culture in industrial biotechnology? *Biotechnology journal*. 2012;7(12):1522-9.
4. Mercier S, Diepenbroek B, Dalm M, Wijffels R, Streefland M. Multivariate data analysis as a PAT tool for early bioprocess development data. *Journal of biotechnology*. 2013;167(3):262-70.
5. Kirdar A, Green K, Rathore A. Application of multivariate data analysis for identification and successful resolution of a root cause for a bioprocessing application. *Biotechnology progress*. 2008;24(3):720-6.
6. Albert S, Kinley R. Multivariate statistical monitoring of batch processes: an industrial case study of fermentation supervision. *Trends in biotechnology*. 2001;19(2):53-62.
7. Glassey J. Multivariate data analysis for advancing the interpretation of bioprocess measurement and monitoring data. *Advances in biochemical engineering/biotechnology*. 2013;132:167-91.
8. Teixeira A, Oliveira R, Alves P, Carrondo M. Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. *Biotechnology advances*. 2009;27(6):726-32.
9. Ozturk S, Riley M, Palsson B. Effects of ammonia and lactate on hybridoma growth, metabolism, and antibody production. *Biotechnology and bioengineering*. 1992;39(4):418-31.
10. Li J, Wong C, Vijayasankaran N, Hudson T, Amanullah A. Feeding lactate for CHO cell culture processes: impact on culture metabolism and performance. *Biotechnology and bioengineering*. 2012;109(5):1173-86.
11. Trummer E, Fauland K, Seidinger S, Schriebl K, Lattenmayer C, Kunert R, et al. Process parameter shifting: Part I. Effect of DOT, pH, and temperature on the performance of Epo-Fc expressing CHO cells cultivated in controlled batch bioreactors. *Biotechnology and bioengineering*. 2006;94(6):1033-44.
12. Nolan R, Lee K. Dynamic model of CHO cell metabolism. *Metabolic engineering*. 2011;13(1):108-24.
13. Ozturk SS, Palsso BO. Growth, metabolic, and antibody production kinetics of hybridoma cell culture: 1. Analysis of data from controlled batch reactors. *Biotechnology progress*. 1991;7:471-80.
14. Grainger R, James D. CHO cell line specific prediction and control of recombinant monoclonal antibody N-glycosylation. *Biotechnology and bioengineering*. 2013;110(11):2970-83.
15. Niu H, Amribt Z, Fickers P, Tan W, Bogaerts P. Metabolic Pathway Analysis and Reduction for Mammalian Cell Cultures - Towards Macroscopic Modeling. *Chemical Engineering Science*. 2013.
16. Amand M, Tran K, Radhakrishnan D, Robinson A, Ogunnaike B. Controllability analysis of protein glycosylation in CHO cells. *PloS one*. 2014;9(2).
17. Von-Stosch M, Davy S, Francois K, Galvanauskas V, Hamelink J-M, Luebbert A, et al. Hybrid modeling for quality by design and PAT – benefits and

- challenges of applications in biopharmaceutical industry. *Biotechnology Journal*. 2014;9.
18. Bro R. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*. 1997;38:149-71.
19. Nomikos P, MacGregor JF. Monitoring batch processes using multiway principal component analysis. *AIChE Journal*. 1994;40.
20. Nomikos P, MacGregor JF. Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems*. 1995;30.
21. Cunha C, Glassey J, Montague G, Albert S, Mohan P. An assessment of seed quality and its influence on productivity estimation in an industrial antibiotic fermentation. *Biotechnology and bioengineering*. 2002;78(6):658-69.
22. Ivarsson M, Villiger TK, Morbidelli M, Soos M. Evaluating the Impact of Cell Culture Process Parameters on Monoclonal Antibody N-Glycosylation. *Journal of biotechnology*. 2014.
23. Guttman L. Some necessary conditions for common-factor analysis. *Psychometrika*. 1954;19(2):149-61.
24. De Jong S. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*. 1993.
25. Ozturk SS, Palsson BO. Growth, Metabolic, and Antibody Production Kinetics of Hybridoma Cell Culture: 2. Effects of Serum Concentration, Dissolved Oxygen Concentration, and Medium pH in a Batch Reactor. *Biotechnology progress*. 1991;7(6):481-94.
26. Gao Y, Kipling K, Glassey J, Willis M, Montague G, Zhou Y, et al. Application of agent-based system for bioprocess description and process improvement. *Biotechnology progress*. 2010;26(3):706-16.

Table 1: Experimental conditions of each of the 12 experimental runs used in the PCA analysis. Numbers highlighted in bold show the parameter and value to which that parameter was shifted at 30 hours into the cultivation. Prior to the parameter shift the cultivations were operated at a standard setting (50% dissolved oxygen; 380 mOsm/kg osmolality; 7.2 pH; 0.05 vvm sparging).

Run identifier	Dissolved oxygen (%)	Osmolality (mOsm/kg)	pH	Sparging (vvm)
Run 1	50	380	7.2	0.05
Run 2	10	380	7.2	0.05
Run 3	90	380	7.2	0.05
Run 4	50	350	7.2	0.05
Run 5	50	380	7.2	0.05
Run 6	50	450	7.2	0.05
Run 7	50	380	7.0	0.05
Run 8	50	380	7.5	0.05
Run 9	50	380	8.0	0.05
Run 10	50	380	7.2	0.05
Run 11	50	380	7.2	0.10
Run 12	50	380	7.2	0.20

1 Table 2: Experimental conditions of each of the 13 experimental runs used to construct and
2 validate the PLS models. **Numbers highlighted in bold show the parameter and value to**
3 **which that parameter was shifted at 30 hours into the cultivation. Prior to the**
4 **parameter shift the cultivations were operated at a standard setting (50% dissolved**
5 **oxygen; 380 mOsm/kg osmolality; 7.2 pH; 0.05 vvm sparging).**

Run identifier	Dissolved oxygen (%)	Osmolality (mOsm/kg)	pH	Sparging (vvm)
Run 1	50	380	7.2	0.05
Run 2	10	380	7.2	0.05
Run 3	90	380	7.2	0.05
Run 4	50	350	7.2	0.05
Run 6	50	450	7.2	0.05
Run 7	50	380	7.0	0.05
Run 8	50	380	7.5	0.05
Run 9	50	380	8.0	0.05
Run 10	50	380	7.2	0.05
Run 11	50	380	7.2	0.10
Run 12	50	380	7.2	0.20
Run 13 (Validation)	50	420	7.2	0.05
Run 5 (Validation)	50	380	7.2	0.05

6

Table 3: Parameters of multivariate analysis carried out on process data. For each analysis (column 1) the input data (column 2), the MVDA technique (column 3), the number of PCs/LVs (column 4) and the cumulative variance explained (column 5) are shown.

		Multivariate technique	Number of PCs/LVs	Cumulative variance explained
Exploratory Analysis	Online process data	MPCA	6	83.42
	Amino acid data	MPCA	4	91.77
	Glycosylation data	MPCA	3	90.25
Glycosylation regression	Model one <i>(Predicted from titre)</i>	MPLS	4	X (87.48) Y (75.57)
	Model two <i>(Predicted from titre and glucose)</i>	MPLS	4	X (95.39) Y (70.36)
	Model three <i>(Predicted from Amino Acids)</i>	MPLS	4	X (88.74) Y (76.95)
	Model four <i>(Predicted from Online data)</i>	MPLS	4	X (80.23) Y (71.73)

Table 4: Root mean squared error (RMSE) of the (M)PLS regression models. For each model the input and output data are shown with the RMSE calculated either as an average value of the whole cultivation for both validation runs (titre prediction model) or as an average value at the end of the cultivation for both validation runs for each glycan form (glycosylation prediction). The number of LVs in each model is also shown.

	X data	Y data	Root Mean Squared Error (RMSE)									LV
	Amino Acids	Titre (<i>max.value=100.93 mg/ml</i>)	6.92 (average over whole cultivation)									4
			Glycan forms									
			A	B	C	D	E	F	G	H	I	
Model 1	Titre	Glycosylation	1.67	7.82	10.2	9.34	0.33	0.55	2.41	0.86	1.22	4
Model 2	Titre, glucose	Glycosylation	1.09	3.90	3.37	2.57	0.23	0.68	0.77	0.21	0.06	4
Model 3	Amino Acid	Glycosylation	0.77	2.77	0.67	0.78	0.15	0.40	0.48	0.27	0.50	4
Model 4	Online data	Glycosylation	2.53	34.5	38.0	21.8	0.59	0.51	8.65	7.51	9.93	4

1
2
3
4

Figure 1: Bivariate scores plots for online process data showing the first three principal components. A-C whole cultivation data, D-F data post shift in operating variable (see symbols and Table 1)

Figure 2: Bivariate scores plot for amino acid data collected following the shift in operating conditions showing the first three principal components

Figure 3: Coefficients (loadings) for PCA analysis of amino acid data. Loadings are colour coded to show experimental runs

1
2

Figure 4: PLS predictions for titre from amino acid data for the two validation cultivations. (A) run 13 (Table 2) and (B) run 5 (Table 2). Measured values (black symbols) and predicted values (open symbols) are shown for each run as their absolute values in mg/L, together with experimental error bars.

1
2
3
4
5
6

Figure 5: PLS predictions for glycosylation profile from titre and glucose post shift (Model 2). Predictions for training runs (black circle) and validation runs (open circle) are shown on x-axis vs the actual values on y-axis for each of the nine glycan forms measured as % peak area (A-I). (A) G0; (B) G0F; (C) G1F; (D) G2F; (E) G0-GalS; (F) G2F-Gal; (G) G2F-S; (H) G2F-GalS; (I) G2F-S2. The line represents the parity line (prediction = measure value).

Figure 6: PLS predictions for glycosylation profile from amino acid concentrations post shift (Model 3). Predictions for training runs (black circle) and validation runs (open circle) are shown on x-axis vs the actual values on y-axis for each of the nine glycan forms measured as % peak area (A-I). (A) G0; (B) G0F; (C) G1F; (D) G2F; (E) G0-GalS; (F) G2F-Gal; (G) G2F-S; (H) G2F-GalS; (I) G2F-S2. The line represents the parity line (prediction = measure value).

Figure 1S: PARAFAC analysis of a 3D online data matrix. (A) Mode one, representing time of cultivation, with the x-axis showing dimensionless time; (B) Mode two, representing online variable recorded, where (1) DO, (2) O2 (3) CO2 (4) pH (5) Base (6) Stirrer Speed (7) Temperature; (C) Mode three, representing batch number on x axis (see Table 1). Component (1) = red; Component (2) = yellow; Component (3) = blue.

Figure 2S: PLS predictions for glycosylation profile from online data post shift (Model 4). Predictions for training runs (black circle) and validation runs (open circle) are shown on x-axis vs the actual values on y-axis for each of the nine glycan forms measured as % peak area (A-I). (A) G0; (B) G0F; (C) G1F; (D) G2F; (E) G0-GalS; (F) G2F-Gal; (G) G2F-S; (H) G2F-GalS; (I) G2F-S2. The line represents the parity line (prediction = measure value).