



**Manuscript: Enhanced process understanding and multivariate prediction of the relationship between cell culture process and monoclonal antibody quality**

**Authors:** <sup>1</sup>Michael Sokolov , <sup>1</sup>Jonathan Ritscher, <sup>2</sup>Nicola MacKinnon, <sup>2</sup>Jonathan Souquet, <sup>2</sup>Hervé Broly, <sup>1</sup>Massimo Morbidelli , <sup>1</sup>Alessandro Butté

<sup>1</sup>Institute of Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Switzerland

<sup>2</sup>Merck, Biotech Process Sciences, Corsier-sur-Vevey, Switzerland

**Corresponding Author:** Alessandro Butté, Institute of Chemical and Bioengineering, Vladimir-Prelog-Weg 1, 8093 Zürich, Switzerland, Mail: [alessandro.butte@chem.ethz.ch](mailto:alessandro.butte@chem.ethz.ch)

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/btpr.2502

© 2017 American Institute of Chemical Engineers Biotechnol Prog  
Received: Mar 16, 2017; Revised: May 24, 2017; Accepted: Ma 24, 2017

**Abstract:**

This work investigates the insights and understanding which can be deduced from predictive process models for the product quality of a monoclonal antibody based on designed high-throughput cell culture experiments performed at milliliter (ambr-15<sup>®</sup>) scale. The investigated process conditions include various media supplements as well as pH and temperature shifts applied during the process. First, Principal Component Analysis (PCA) is used to show the strong correlation characteristics among the product quality attributes including aggregates, fragments, charge variants and glycans. Then, Partial Least Square Regression (PLS1 and PLS2) is applied to predict the product quality variables based on process information (one by one or simultaneously). The comparison of those two modeling techniques shows that a single (PLS2) model is capable of revealing the interrelationship of the process characteristics to the large set product quality variables. In order to show the dynamic evolution of the process predictability separate models are defined at different time points showing that several product quality attributes are mainly driven by the media composition and, hence, can be decently predicted from early on in the process, while others are strongly affected by process parameter changes during the process. Finally, by coupling the PLS2 models with a genetic algorithm first the model performance can be further improved and, most importantly, the interpretation of the large-dimensioned process-product-interrelationship can be significantly simplified. The generally applicable toolset presented in this case study provides a solid basis for decision making and process optimization throughout process development.

**Keywords:**

(1) Predictive process models, (2) Multivariate data analysis, (3) Product Quality, (4) Partial Least Square Regression, (5) Genetic Algorithm

## 1. Introduction

In the past decades, the biopharmaceutical industry has witnessed a paradigm shift from productivity maximization to quality optimization. Cell culture fed-batch processes yielding a final monoclonal antibody (mAb) concentration as high as 5 g/L have become an established standard in industry<sup>1</sup>. On the other hand, modulation of quality attributes has significantly gained importance, starting from early process development, and not only for biosimilars. The assurance of constant product quality is a key objective for biopharmaceutical processes as the efficacy, potency and safety are dependent on the structural characteristics of the protein<sup>2</sup>. Therefore, the analysis and knowledge about the process parameters affecting product quality has a central role throughout the entire product life cycle from screening, through scale-up and manufacturing to formulation.

The product quality of a therapeutic protein is very complex. It includes information on its glycosylation profile, its charge variant distribution as well as its aggregated and low molecular weight forms. Consequently, the impact of tens of potentially relevant process parameters on numerous quality attributes has to be considered simultaneously to design the process. The utilization of highly automated experimental systems with small working volumes enables an economically attractive and efficient evaluation of the manifold possibilities for process design<sup>3</sup>. In this frame, the major challenge for process model development is the fact that usually both groups, process variables and product attributes, are large-dimensional, highly correlated and are subject to analytical and operational inaccuracies. Moreover, consideration of the process dynamics is essential in order to build adaptive prediction models. Therefore, robust multivariate and time-dependent models are required in order to avoid overfitting, reduce noise, simplify interpretation and reveal important characteristics and trends in the dynamic process-product-interrelationship.

Cell culture process models for the prediction of product quantity have already been presented in several works. For instance, Ündey et al.<sup>4</sup> showed accurate predictions for the titer based on online process data (incorporating information on at least 50 % of the process duration). Sokolov et al.<sup>5</sup> visualized the dynamic change of the interrelationship between the process variables and the produced titer and demonstrated that reliable forecast models can be generated after a few process days. The corresponding models were built using the widely used technique Partial Least Square Regression (PLSR). Charaniya et al.<sup>6</sup> provided an extensive analysis of titer prediction based on Support Vector Regression (SVR) using data from different scales. Le et al.<sup>7</sup> also incorporated the process variability from the inoculation trail and presented comparably accurate prediction models applying PLSR and SVR.

Rouiller et al.<sup>8</sup> showed the effect of process parameters on a few quality attributes (e.g. aggregates and glycoforms) in a process characterization study at 3.5 L lab scale. Similar studies coupling empirical risk assessment to univariate models were presented by Abu-Absi et al.<sup>9</sup> and Nagashima et al.<sup>10</sup> In another work Rouiller et al.<sup>11</sup> analyzed the effect of media supplements on the product quality at microliter scale (deep well plates). All previous analyses were based on design of experiments (DoE) and response surface models (RSM). The first study on multivariate product quality forecasting was presented by Schmidberger et al.<sup>12</sup> using data from 5 L bench-scale bioreactors. In this work, PLSR, SVR, Random Forest and Radial basis function neural networks were compared regarding their capability to predicting various product quality attributes based on different amounts of historic process information. Promising results were achieved for the forecast of some quality attributes early to mid-phase in the process based on PLSR models. Rathore et al.<sup>13</sup> presented an interesting PLSR-based investigation of the relationship of media components and operating conditions to the mAb glycosylation profile in 1.2 L bioreactors using commercially available chemometrical tools. Some quality variables could be predicted decently, whereby

interpretation of the overall relationship was limited due to the large number of variables considered, on the one hand, and a rather small number of quite different cell culture runs, on the other hand.

The novelty for bioprocess analysis presented in this work lies in the generation of a single reliable cell culture model accurately predicting a large set of final product quality attributes (CQAs) based on the overall process information (process attributes, PA) at a certain time point using a large number of designed milliliter-scale cell culture experiments. The combination of such a model with a genetic algorithm (GA) is further improving the model performance and robustness compared to the classical approaches, while significantly simplifying the interpretation and the understanding of the dynamic relation of PAs and CQAs as well as the importance of the different PAs. Consequently, the presented approach provides the basis to efficiently optimize the process simultaneously for multiple objectives.

## **2 Materials and Methods**

### **2.1 Experimental plan and procedure**

The overall data set consists of 91 experimental runs performed using the ambr<sup>®</sup> 15 system (15 mL maximal working volume; TAP Biosystems, UK). The runs were based on four experimental blocks (DoE 1 to 4 in Table 1), which were built by fractional factorial designs as well as some additional one parameter at a time (OPAT) analyses of certain factors. The objective was to test nine supplemented media factors at different concentrations as well as three different process execution strategies. The factors contain the two sugars galactose (Gal) and sucrose, the amino acid asparagine (Asp), the four metals and metal combinations manganese (Mn), iron (Fe), copper-zink (CuZn) and cobalt (Co) as well as spermine and gallic acid. These factors are known to physiologically affect the product quality<sup>1</sup>. The three factors cobalt, spermine and gallic acid were simultaneously tested at a level different from

the basic condition in DoE3 and could, therefore, be considered as a triplet. Table 1 shows the number of different conditions (or levels) tested for each parameter in the four experimental blocks (referred to as DoE1 to DoE4) as well as the corresponding total number of experiments. The three different process execution strategies include two feeding regimes (Feed) testing the volume of the feed at two slightly different levels (in DoE1), two different flow rates of nitrogen (N<sub>2</sub> flow, tested in DoE3), and the temperature shift (Temp Shift) at culture day 6 from 36.5 °C to a smaller level (tested in DoE1 and DoE3).

Prior to the cell culture runs, the cells were expanded in vented spin tubes in a proprietary animal-derived component free expansion media. After at least six dilution steps within 14 days the ambr reactors were inoculated at a seeding cell density of  $0.2 \times 10^6$  cells/mL. The nominal working volumes were within 10 to 15 mL. The culture duration was 14 days. The chemically defined main feed consisted of 30 components and the investigated factors (in Table 1) were added into the main feed, which was added on the days 3, 5, 7 and 10. Glucose was added on a daily basis from day 3. Liquid handling was performed by a robotic system. pH was controlled by the addition of carbon dioxide and sodium carbonate (1M). In many experiments a pH adaptation was performed at culture day 6. However, due to many slightly varying levels before and after day 6, the pH shall be considered as a dynamic and not a fixed process variable.

The four DoEs were performed in a consecutive manner and interpreted independently from the analysis carried out in this work. While some conditions were sequentially eliminated (e.g., the levels of Gal from DoE2 to DoE4), new conditions were added throughout the time course of the experiments (e.g., CuZn and N<sub>2</sub> flow in DoE3). The small number of experiments in DoE1 is due to the fact that it also aimed at testing different cell lines, which are not considered in this work.

The offline measurement of pH, pCO<sub>2</sub> and pO<sub>2</sub> were performed by ABL5 (Radiometer, Denmark). Viable cell density (VCD) and viability were measured with ViCell® (Beckman Coulter, USA). The concentrations of glucose, ammonium and lactate were measured with Bioprofile 100+ (Nova Biomedical, USA). The remaining culture volumes at day 14 were purified using small-scale Protein A columns (Phytip®, PhyNexus, San Jose, CA). The final product quality was analyzed from the Phytip eluates. Glycosylation was quantified by capillary gel electrophoresis with laser induced fluorescence detection (CGE-LIF, DNA genetic analyzer 3130XL, Life Technologies, Darmstadt, Germany), charge variants by imaged capillary isoelectric focusing (iCE280 analyzer, ProteinSimple, Santa Clara, CA), aggregation by size exclusion high performance liquid chromatography (SE-HPLC, Water, Baden-Wättwil, Switzerland) and low molecular weight forms by non-reduced capillary SDS-PAGE (Labchip GX II, Perkin Elmer, Schwerzenbach, Switzerland).

## 2.2 Data set and multivariate analysis

For the analyses the data are separated in three characteristic blocks, referring to the works of Kourti<sup>14</sup> and MacGregor<sup>15</sup>, which are visualized in Figure 1. The block Z is spanned by the experimental runs (rows) and the different process conditions, which are fixed by the experimental design (columns). The block X is built by the experimental runs (rows) and different process variables (columns), which are measured at different time points throughout the cell culture (third dimension). For the analysis this 3-dimensional X block is rearranged (unfolded<sup>14</sup>) in several two-dimensional blocks, each representing a time point of measurement. Consequently, the rearranged overall X block distinguishes process variables at different time points as different columns. The Y block represents the different final quality attributes (columns) for the experimental runs (rows). Both, the process variables as well as the quality attributes can feature missing data (shown as black elements in Figure 1).

In this work, based on the experimental plan shown in Table 1, the 12 Z variables contain the 9 factors (3 are coupled as they are always simultaneously varied) as well as the conditions for the temperature shift, the nitrogen flow and the feeding regime. As pointed out before, pH adaptation will cause pH to vary before and after day 6. Therefore, in the subsequent analysis, the variable pH shall be quantified based on its time profile (in the X block). Moreover, the X block consists of the process variables VCD, viability (Via), the concentrations of ammonium (NH<sub>4</sub>), lactate (LAC) and glucose (GLU) as well as the partial pressures of carbon dioxide (pCO<sub>2</sub>) and oxygen (pO<sub>2</sub>). Overall, eight X variables were quantified about six times during the culture (on days 3, 5, 7, 10, 12, 14). The overall percentage of missing values in the X data was 8%. Only DoE 2 features a relatively higher amount of missing values (21%). The Y block incorporates the variables product quantity (titer), fragments (LMW), aggregates (Agg), 5 charge variants (C1 to C5) and 6 grouped or 11 specific glycoforms, with a total of 14 or 19 quality attributes, respectively. The considered 11 specific glycoforms partitioned into 6 characteristic groups: forms containing zero to two galactose molecules (G0, G1, G2), afucosylated forms (AF), sialylated forms (Sia) and high mannose forms (HM). The missing data ratio for the Y variables was 2%.

Figure 2 visualizes the widely used multivariate techniques Principal Component Analysis<sup>16</sup> (PCA) and Partial Least Square Regression<sup>17</sup> (PLSR). PCA is a visualization tool, which provides a ‘low-dimensional window’, revealing groups among the experiments as well as their relation to the correlation structure of the variables. Thereby, the original large-dimensional correlated data set is represented in a new coordinate system of latent variables (LVs, in this case also called Principal Components PCs) defined along the directions of the largest correlations in the data. The relation of the new system to the original one is given by the scores T (projections of the experiments) and the loadings P (directions of the original variables). Such an analysis can be applied to both, correlated process data<sup>5</sup> (dashed square in



Figure 2) and correlated product quality data<sup>18</sup> (solid square in Figure 2). PLSR is a regression tool, linking the process data (Z and X) to the quality data (Y). Again, a lower-dimensional set of LVs is defined, in this case along the directions of the largest covariance of Z and X with Y. The relations in this new system are represented by the scores T (projections of the experiments), the W\*-loadings (contribution of Z and X variables to explain Z along certain LV) and C-loadings (directions of Y variables in LV space). One can distinguish between the classically used PLS1 models, where a single Y variable is regressed, and PLS2 models<sup>19,20</sup> for a simultaneous regression of a set of Y variables.

In order to avoid over-fitting, K-fold cross-validation<sup>21</sup> (with K = 5 groups) was used for the generation of the PLSR models. While leaving out one fifth of the data set, a model is generated on the remaining (training) data and the left-out (test) data are used to test the prediction capabilities of the model. This procedure is repeated for all the five groups and the model performance is evaluated by the root mean square error in cross-validation (RMSECV, an averaged absolute deviation of the predicted to the observed values) and the relative variance explained in cross-validation ( $Q^2$ , ratio of explained to total variance in the predicted data). The latter is independent of the scale of the Y variable and reaches a value of 1 for a perfect prediction model, while 0 reflects the performance of the model using the mean values of the predicted (Y) variables. In order to compare the RMSECVs for different Y variables in different ranges, the RMSECVs are scaled with the standard deviation of the respective Y variable. In this case, a value of 0 corresponds to a perfect model, while a value of 1 represents the model always predicting the mean. The optimal number of LVs was determined by using the minimum of RMSECV for PLS1 or the minimum of the Y-averaged scaled RMSECV for PLS2 models.

PLS-based regression models offer numerous ways to access the variable relation from the X and Y space<sup>17,22</sup>. In PLS1 models,  $\beta$ -coefficients<sup>5,7,23–25</sup>, Variable Importance in the

Projection (VIP)<sup>5,15,23,25–27</sup> as well as  $W^*$ -loadings (and  $C$ -loadings)<sup>5,15,17,28</sup> have been widely used measures for variable importance in bioprocesses.  $\beta$ -coefficients are the regression coefficients of the auto-scaled  $X$  variables corresponding to a certain  $Y$  variable. Therefore, in both, PLS1 and PLS2 models, a set of  $\beta$ -coefficients for each  $Y$  variable needs to be separately compared. As for the  $W^*$ -loadings, if for instance the first LV accounts for a significant portion of the explained variance, the  $X$  variables featuring an absolute, high  $W^*$ -loading on this LV, are likely to be influential. The VIP values of the  $X$  variables represent their relative importance to explaining both, the entire  $X$  and the entire  $Y$  space, in the latent variable model. However, when several product quality variables ( $Y$ ) are simultaneously analyzed in PLS2 models, the interpretation based on  $\beta$ -coefficients is likely to become cumbersome with an increasing number of  $Y$  variables to be evaluated. The advantage of the VIP values to this regard is the capability to link information from the entire  $X$  and  $Y$  spaces. If large portions of the variance are explained by the first few LVs, a joint interpretation of the  $C$ - and  $W^*$ -loadings in a single plot is an alternative<sup>13,17</sup>, which again tends to become difficult if many  $X$  and  $Y$  variables are present.

In this work, a genetic algorithm (GA) is presented as a powerful variable selection tool for PLS2 models. In general, GAs<sup>29</sup> compare different models with a defined binary selection pattern (chromosome) for the predictor set (for each variable either 1 = leave in model, or 0 = out). Based on a guided procedure, chromosomes can be found ensuring the best prediction, which eventually define the optimal selection of the predictor variables. In this work, 100 different populations of experiments<sup>30</sup> were created by randomizing the order of the experiments and, for each of them, the GA was used to select predictor variables. Since a 5-fold cross-validation scheme was used, GA selection depends on the experimental order. At the end of the optimization, the best chromosomes from each population were tested on all other populations, thus obtaining a distribution of objective functions for each chromosome.

After sorting the different chromosomes based on the mean values of their objective function, a two-sample t-test<sup>31</sup> was performed to eliminate chromosomes having a mean value statistically different from that of the best chromosome. The corresponding averaged Akaike information criterion<sup>32</sup> (AIC) was used as objective function, whereby the number of predictor variables is penalized, so that models with less variables are favored compared to similarly accurate models with more variables to simplify the interpretation.

Missing values can be imputed by the general correlation structure of the overall data set, if their distribution is rather random. In this work, the Trimmed Scores Regression algorithm by Ferrer et al.<sup>33</sup> was applied separately to the unfolded X and Y data in order to avoid any information exchange prior to the regression analysis. Moreover, before the analyses the data was auto-scaled, i.e. each variable was first centered by its mean and the scaled by division by its standard deviation. The analysis was executed with MATLAB (R2015b, The MathWorks, Inc.). PCA was performed using *princomp*, the PLSR functions from the *libPLS package* in the Mathworks Central were adapted and extended to PLS2. The GA was adapted from Leardi et al.<sup>29</sup> and extended to PLS2.

### 3 Results and Discussion

#### 3.1 PCA on Quality Attributes

Figure 3 presents the results obtained from performing PCA on the 14 product attributes (titer, LMW, aggregates, 5 charge isoforms C1 to C5, 6 glycan species HM, G0, G1, G2, AF, Sia).

The loading plot in Figure 3A visualizes the correlation structure of these Y variables. In general, one can observe that several Y variables are strongly correlated and that 61 % of the total variance can be explained by two Principal Components (PCs) only. For instance, along PC1 the lower processed and intermediate glycoforms (HM, AF and G0; negative loadings on PC1 and PC2) can be distinguished from the higher processed ones (G1, G2, Sia; positive

loadings on PC1 and PC2). Also for the charge variants a trend can be observed along PC1 and particularly along PC2 (separating the positively charged forms C1 and C2 from the neutral and negatively charged forms C3 to C5).

The score plot corresponding to the first two PCs in Figure 3 B visualizes the 91 experiments distinguishing the four DoEs (by color and shape). In particular, DoE2 (deep gray circles; mostly positive PC1 and PC2 values) and DoE3 (light grey diamonds; mostly negative PC1 and PC2 values) can be clearly separated. This indicates that the factors tested in those experiments are likely to have affected different groups of Y variables. Therefore, PCA of the CQAs is an important first step to evaluate the degree of correlation within the product quality as well as to provide the basis for first analysis hypotheses. To test such hypotheses, a detailed regression analysis is required, which shall be presented in the following sections.

Figure 3 C visualizes the variance explained in the PCA for two distinct cases: the first one (red dashed curve), where the glycoforms within the Y variables are composed of each specific measured form (19 Y variables); and the second one (solid black curve), where they are composed of six characteristic groups (cf. Figure 3 A, 14 Y variables). For the latter case, it is evident that four PCs are capable of explaining 80 % of the variance in the Y data, while six PCs explain 90 %. A similar trend can be observed in the other case, where 5 and 8 PCs are needed to explain 80 and 90% of the variance, respectively. This motivates to consider the product quality data as one characteristic, highly correlated information block, rather than considering each quality attribute separately and justifies the use of PLS2 models later in this work.

### 3.2 Separate Quality Attribute Prediction with PLS1

Following the current standard procedure for multivariate prediction modeling, PLS1 regression models were separately built for each quality attribute as a model benchmark to be

later compared with the PLS models. In order to investigate the predictability of the quality of the produced mAb, increasing portions of the process history were used to model final product quality. Figure 4 shows the predictability (based on the variance explained in cross-validation,  $Q^2$ ) of different quality attributes (columns) from the process information until a certain point of time (rows). For instance, in the first row, the final quality attributes are separately predicted using the fixed process conditions (Z) only, i.e. this model corresponds to the classical black box models. The further rows are labelled with respect to the time, until when information on the process variables (shown as X blocks in Figure 2) was incorporated in the model. Accordingly, the second element in the first column represents the model for titer, which was built using Z (12 variables) and the information of process day 3 (T3), for a total of 17 predictor variables. The predictability of each CQA was determined with a different PLS1 model for each novel portion of process information added to the predictors (number shown in brackets). For each row, the missing X data was imputed based on the correlation of the PA variables available until then.

In Figure 4 values close to 100 % (black elements) indicate a perfect predictability, while values close to 0 % (white elements) indicate that prediction is not possible. The CQAs show different characteristic predictability patterns along the process time. The variable LMW features a high predictability (71 % variance explained) singly based on the fixed process conditions. The predictability of LMW does not further improve by incorporating the process history. For most of the other quality variables, namely titer, aggregates, C1 and C3 to C5, G0 to G2, HM and Sia, the predictability does not significantly change after the process information up to day 5 have been included. In particular, titer and the charge variants C1 and C3 to C5 show a further improvement when also adding the information from day 7. Only C2 and AF seem to require information from later stages of the process. In general, most of the variables can be decently predicted by such models. Five variables (titer, C1, C3, C4 and G2)

reach a good prediction accuracy above 80 %, six further variables (aggregates, LMW, C2, G0, G1 and HM) have a sufficient accuracy close to or above 70 %, while the three variables (C5, AF and Sia) feature the weakest predictability among the CQAs. Given the correlation structure shown in Figure 3 A, one can also observe that for (positively or negatively) correlated variable groups such as C1 and C3 as well as G0, G1 and G2, the general trends in the process time dependent prediction are quite similar.

The fact that the fixed Z conditions are usually insufficient to model the final product quality, and that, consequently, information on the process variables tend to be very important, can have two explanations. Some CQAs might be affected by changes in the process variables. In particular, the pH adaptation on day 6, which is singly incorporated in the pH profile (pH at six time points), is likely to be an important source of additional information for variables such as titer and the charge variants. Moreover, it is important to point out that besides the nitrogen flow rate, the fixed process conditions affect the process starting from day 3 (supplemented factors and feeding regime) and on day 6 for the temperature shift. Those conditions result in a dynamic response of both, the quality attributes and the process variables. As pointed out in Sokolov et al.<sup>5</sup> this response can be non-linear so that this behavior can be better reproduced by several additional time dependent process variables (X) in the model compared to the purely linear relation to the process driver (Z). Regarding this particular process, it seems that the manipulations on day 3 are not sufficient to drive the process into distinct directions, so that a clearer or more consistent effect on the quality attributes can be achieved by the information from day 5 (and even 7). Moreover, the adaptation of pH and the temperature shift (both at day 6) are likely to be further important process levers, yielding different behavior from day 7 on, so that the CQAs affected by those conditions can be better predicted using also the process response (from day 7) to those changes.

In order to obtain more specific interpretations on the relationship of process and quality variables, it is required to consider the variable importance of each single PLS1 model. Different techniques for such an analysis have been presented, for example, by Le et al.<sup>7</sup> and Sokolov et al.<sup>5</sup> for predicting titer. Due the large number of correlated CQAs (as shown in the PCA pre-analysis in section 3.1) the first objective of this work is to simplify the overall interpretation by connecting the process to the large-dimensional product quality information by a single model, which shall then be evaluated.

### 3.3 Joint Quality Attribute Prediction with PLS2

Figure 5 presents the results obtained from PLS2 models. Thereby, the basis for the prediction models being the process data (X) in each row and the quality data (Y) in each column is equivalent to the models presented in Figure 4. The only significant difference is that the large-dimensional product quality is predicted with a single PLS2 model, as opposed to modeling each single CQA separately with PLS1 models. Consequently, in Figure 5 each row corresponds to one PLS2 model, each using a different amount of process history for the prediction.

Comparing the results shown in Figure 5 to the ones from Figure 4, one can observe that the model quality of the black box model linking the fixed conditions Z to the final quality attributes is 2 to 9 percentage units worse for most of the CQAs, while the variables HM, AF and Sia feature a slightly better predictability with PLS2. Figure S1 in the supplementary information provides an element-wise comparison of the two modeling approaches. After adding the information from day 3, the general trends for improvement of predictability are consistent with the ones in Figure 4. The predictability of the variables Agg, LMW, G0, G1, G2 and Sia does not significantly change compared to the black box models, while the one of titer, C1, C3 to C5, HM and AF improves. After adding the information from day 5, the

results of the two approaches are very similar. The PLS2 model performs only slightly worse for most of the variables (featuring less than five per cent units deviation), while only for C1 it lacks 8 per cent units. Eventually, the PLS2 model incorporating the entire process information until day 14 (last row) features almost equivalent results for all quality attributes to the corresponding models in Figure 4.

The PLS2 regression analysis is driven by the correlation structures of the process variables (maximally 57) and the 14 quality variables as well as the covariance of the two domains. The strong covariance is revealed by the relatively small number of latent variables (LVs) used in each of the models. For instance, for the model incorporating the entire history (last row in Figure 5), the relationship between the process and the product can be optimally modeled by using 11 uncorrelated LVs. This means, that 11 independent effects have to be considered to understand the interrelationship of the 57 product and 14 quality variables. Given such strong interactions (correlation among Y shown in Figure 3 and covariance among X and Y based on the low number of LVs), it is very attractive to build PLS2 models. While as discussed before such a model features a very similar prediction accuracy to the classically used PLS1 model, PLS2 results in a single model for the large set of CQAs, so that it can be interpreted in a much simpler manner simultaneously accounting for the interrelationship of all the considered PAs and CQAs.

Having provided this attractive model basis, the second objective of this work is to show how a genetic algorithm (GA) can be used to improve the accuracy and robustness of those models and to provide understanding on the interrelationship of process and product quality variables.

### **3.4 Improvement of Precision and Robustness with GA**

Three different approaches are used to build PLS2 models. The first uses all the existing information until a certain point of time such as described in section 3.3 and shown in Figure



5. The second one uses only the variables until the given time point, which feature a VIP value equal to or larger than 1 (i.e., above the average importance level of a variable).

Similarly to the second approach, the third utilizes singly those variables identified as relevant by the GA. In order to also evaluate the stability of those prediction models, each approach is repeated 100 times with different sets for calibration and testing in the cross-validation.

Figure 6 shows the mean of the scaled RMSECVs and their standard deviation for two exemplary quality variables, LMW and C3, which are predicted based on different lengths of process history. The models with all available Z (and X) variables are shown in white, while the ones using only those selected by the VIP method or the GA are visualized in gray and black, respectively. The models predicting C3 (Figure 6 B) show that the scaled prediction accuracy generally improves from around 0.4 to almost 0.2 scaled units for all three procedures. As discussed before, this suggests that the process manipulations throughout the process (such as the feed and the pH shift) are likely to provide important information to predict the charge variant. Thereby, the model using all initial conditions Z (white bar) has a very large variability. This indicates that the corresponding black box models cannot accurately distinguish valuable information from bias, thus often resulting in poor prediction models. For the models using all available variables, the accuracy substantially improves when the information from day 3 on (i.e., when the first process response to the process settings) is added. Moreover, the robustness of those models improves as the variability (standard deviation of RMSECV) is almost as small as the one for the models based on singly the variables selected by the GA (black bars). For instance (as shown in Figure 7), the GA selects in average only 6 out of the 12 Z variables to build the black box models preserving singly the most relevant information. In particular, for C3 the selection based on the VIP criterion results in less accurate (larger average RMSECV) and less robust predictions (larger standard deviation of RMSECV). For the prediction of LMW (shown in Figure 6 A), one can

observe that the GA features a stable prediction of LMW with a RMSECV of around 0.25 for all time points, while the other two approaches are less precise (in particular when using singly Z) and less robust. This interesting prediction pattern suggests that the LMW is influenced by the process settings Z only, i.e. the process history does not provide important information. In other words, the formation of LMW forms is likely to be directly induced by the media supplements and operating conditions. While the focus of this analysis was put on the error comparison of the different analysis methods, it shall be pointed out that the scaled error of 0.25 units for LMW corresponds to 0.1 % in a range from 1.7 to 3.8 %, which can be considered as an acceptable error.

It is very important to highlight the fact that, for both CQAs, the GA based models are not only more precise but also more robust compared to the other two procedures (i.e., with smaller standard deviation of the RMSECV) and are based on a significantly smaller amount of process variables (up to factor of 5 compared to all available variables, and factor of 3 compared to VIP-selected variables). In particular, where only Z is used, an important noise reduction of the RMSECV can be achieved by GA. This clear identification of relevant process drivers is very helpful for process understanding, so to integrate the derived knowledge into hybrid<sup>34</sup> or deterministic models for such a process. The combination of PLS2 models with GA provide a robust and simple-to-interpret basis for process understanding. As for the VIP based method, although this is extremely faster than the GA method in terms of computational time, it has been shown that this method might result in less precise and non-robust predictions. In other words, those models are likely not to reduce all sources of noise and, hence, provide a less intuitive and biased basis for process understanding. In the next section the actual variables selected by the GA are revealed to provide a deeper understanding of the interrelationship of the process and the product.

### 3.5 Evaluation of Process-Product-Interrelationship with GA

The GA was applied to 100 sets of experiments, each of them obtained by randomizing the experimental order, to select the most informative variables at increasing process history, as shown in Figure 6. Figure 7 shows the frequency of selection of the available variables for each case. In general, outlined boxes indicate that a certain variable was available for selection to the model, whereas the corresponding black bar quantifies the frequency of the selection. For instance, for the model until process day 3 (T3), besides the 12 Z variables, there was information from day 3 for VCD, Via, NH<sub>4</sub>, pCO<sub>2</sub> and pO<sub>2</sub> available (17 variables). Out of those, the 5 variables, namely Gal, Fe, Asn, Temp Shift, N<sub>2</sub> Flow and pCO<sub>2</sub> at day 3, were always selected, while Cobalt, Spermine and Gallic Acid were selected approximately in one third of the cases each, Via and VCD at day 3 in two thirds of the cases. Note that the total number of selected variables in the chromosomes that do not have a statistically different mean value of the RMSECV is not always the same and varies between 6 and 7.

Comparing the models along the time course of the process one can observe several trends. The average number of selected variables (indicated in brackets) for each model is substantially smaller compared to the available ones (cf. Figure 5 in brackets) and decreases from one half for the Z-based model to below one fifth for the model based on the full process history (T14). Thereby, the average total number of significant variables increases only slightly from 6 to 11 selections. This indicates that there is irrelevance, noise and, especially, redundancy in the process data. As for the selections, the Z variables Gal, Fe and Temp shift are almost always selected in each of the seven cases, while Mn, CuZn, Sucrose and Feed are usually discarded. Also the Z variable Asn is often selected. However, the analysis of the selected variables indicates that, from day 5, either of the two variables Asn and NH<sub>4</sub> at day 5 is selected, i.e. according to the GA these two variables contain the same information and can be considered as redundant. As Asn can decompose to NH<sub>4</sub>, this observation is biologically

sound, i.e. the process responds with production of  $\text{NH}_4$  to an Asn feed, and in some cases this process response is slightly more informative (possibly due to non-linearity) than the actual Asn feeding level. The Z variables Cobalt, Spermine and Gallic Acid appear to be selected in equal ratio for each model, which is not surprising given their coupled design (cf. Triplet in Table 1). From this perspective, this result validates the capability to handling redundancy by the GA. While one of them is always selected in the model with only Z variables, their effect is increasingly represented by other process variables at the later time points as those variables are selected less often. In order to obtain a deeper understanding it would be important to decouple those in the design of the subsequent experiment.  $\text{N}_2$  flow is important in the first two models and appears less frequently in the models from day 5 on, which suggests that its effect is later on better reflected by other process responses (presumably by lactate and pH due to stripping of  $\text{CO}_2$ ), although it should also be considered as a potentially critical process parameter, able to influence the product quality.

For a better understanding of the selection patterns, it is important to point out that the X variables represent characteristic process information at different points of time. In other words, those variables reflect the entire process behavior, which of course is affected by the manipulations of process settings throughout the process. For this particular process those are the concentrations of 9 components in the feed (cf. Table 1), which is added on the days 3, 5, 7, 10, the volume of the feed, the nitrogen flow rate as well as the pH and temperature shifts (both on day 6). If an X variable is selected instead of a Z variable, this indicates that it captures the process behavior (or the differences among the runs) better compared to the corresponding Z variable. An example can be a different behavior at an intermediate level for a process parameter (e.g., an optimal pH), compared to lower and higher levels. Therefore, the pH itself cannot capture these opposing trends towards the tested limits, whereby a variable such as viability is likely to show a clear behavior (high at optimum, low towards both limits).

If an X variable is selected additionally to the set of Z variables, this is likely to show that the process response to a certain setting is non-linear so that (besides the linear relation to Z) additional information is required to better explain the behavior of the different runs. Such a non-linear behavior can be either an enhanced effect at high levels of a process parameter (amplification) or a reduced effect at high levels compared to a purely linear model (saturation). Considering the X variables in detail, VCD appears to be mostly informative early in the process (at day 3 or, if available, at day 5). The same is true for Via at day 3, whereby even more than VCD it is likely to provide some further information at the final day 14. The first observation suggests that the behavior of the cells (concentration, viability) directly after the first feed provides additional characteristic information. This means, that neither the pH nor the temperature shift on day 6 tend to introduce a further non-linear effect, which is best explained by the behavior of the cells (compared to other variables). The second selection is related to a rather sudden viability drop of some runs at the process end. The variable GLC is mostly selected at day 5 and at day 12, i.e. further characteristic (non-linear) effects can be captured by the dynamic behavior of this variable. The first selection is likely to represent the transition from batch to fed-batch mode on day 3, and the second selection occurs when the GLC consumption starts rapidly to decrease due to cell death. Interestingly, LAC seems to track a dynamic trend in the process as it is usually selected at the latest possible time point. Besides the prior mentioned coupled effect to Asn at day 5, NH<sub>4</sub> provides further important information on day 10, which for instance could be attributed to a (non-linear) response to the final Asn feed on day 10. The variable pH is selected always at day 7. This means that besides the effects already captured by the other variables, a further characteristic effect, namely the pH shift, can be described measuring pH at day 7. Both variables, pCO<sub>2</sub> and pO<sub>2</sub>, are likely not to provide any further information. The important effect described by pCO<sub>2</sub> at day 3 is then likely to be represented (in a more distinct manner) in the information provided by the other X variables.

In summary, at this stage one can highlight three supplements (Gal, Fe and Asn) and three process controls (temperature shift, pH shift and N<sub>2</sub> flow levels) with a potentially large impact on one or several CQAs. Similarly, previous analysis suggests that Mn, CuZn, Sucrose and Feed at the tested levels tend to not have affected the product quality. Moreover, one can localize further characteristic (non-linear) effects based on the selection patterns of the informative process variables VCD, Via, GLC, LAC and NH<sub>4</sub>. While providing a general dynamic importance pattern of PAs to describe the CQAs, those conclusions do not enable to distinguish the different effect on the various CQAs. For this purpose, we want to present two methods.

In the first, two PLS2 models were built in combination with the GA, one singly based on the Z variables and one for the last process day. Figure 8 shows the W\*-loadings of the PAs selected by the GA (black markers) and the C-loadings of the CQAs (white markers) for initial conditions and process settings (A) and the full history until day 14 (B), respectively. Compared to the loading plot in Figure 3 A, here one can observe the correlation among both CQAs and PAs, as well as their interrelationship. 48 % of this interrelationship (i.e. the corresponding covariance) is explained singly by 6 selected Z variables, and 23% is explained by the first LV, as shown in Figure 8 A. For instance, one can observe a positive impact of Gal on G1, G2 and Sia and a negative one on G0, which represents the galactosylation process in the glycosylation network<sup>35</sup>. Moreover, one can identify a positive effect of Temp Shift and N<sub>2</sub> flow on HM, of AF as well as of Fe on LMW and C3 as well as of Spermine on C5. Such conclusions provide a basis for mechanistic hypotheses and further investigations. In fact, the increase of HM forms as a result of the temperature downshift has already been reported<sup>3</sup> and could be related to a temperature dependent activity of the different enzymes along the glycosylation network. The beneficial role of the GA becomes even more evident in Figure 8 B, where the W\*-loadings of singly the 11 informative (out of 57) PAs are visualized

together with the C-loadings of the Y variables. This simplification enables to observe that, on the one hand, some direct effects observed in Figure 8 A are also valid using the information until the end of the process, e.g. Fe on LMW, Temp Shift on HMW and AF as well as Gal on G0, G1, G2 and Sia. On the other hand, one can observe that the effects prior described by Spermine and N2 flow are now reflected and extended to further effects by process variables at characteristic time points. For instance, Via at day 14 seems to provide important information to describe the final aggregate content, while a high value of LAC at day 10 is likely to have a negative effect on the formation of titer and higher processed glycoforms. All those observations enable in a single plot (or a few plots if the further LVs are considered) to identify and hypothesize on potential interrelationships of certain informative PAs and certain CQAs.

However, for a detailed analysis of a certain CQA one could either carry out a PLS1 regression coupled to a GA as shown in Sokolov et al.<sup>5</sup> or consider the  $\beta$ -coefficients of the PLS2 model related to the given CQA. For the latter method, seven PLS2 models were built for the different process times (cf. rows in Figure 5). To exemplify this analysis, the corresponding sets of beta-coefficients for two selected Y variables, LMW and C4, shall be considered here. Figure 9 A shows the  $\beta$ -coefficients of the PAs (columns), which were selected at least in one of the models (rows, labeled with number of variables selected and number of LVs in corresponding PLS2 model) to predict LMW. The  $\beta$ -coefficients are standardized, such that -100 refers to a strong negative and 100 to a strong positive effect. One can observe that, for modeling LMW, particularly the Z variables Fe and Temp Shift are important, whereas less information is generally provided by the process variables. This is in line with the hypothesis on the effect of Fe on LMW observed in Figure 8, as well as with the trend observed in Figures 5 and 6, where it was evident that the information of the Z variables was mostly sufficient to predict the final LMW (although the influential Z variables were not



yet specified). The prediction of C4 in Figure 9 B shows an opposite trend. While very little of its variability can be generally explained by the initial Z variables, several important responses of process variables (VCD and pCO<sub>2</sub> at day 3, NH<sub>4</sub> at day 5, Via and pCO<sub>2</sub> at day 7, LAC at day 10 and Via at day 12) tend to provide important pieces of information to model C4. This reflects the trend observed in Figure 5, where the predictability of C4 clearly improves with more process history added to the Z information and suggests that this CQA is far more affected by the dynamic behavior of the process variables (i.e., by non-linear effects). It is also worth observing that the number of selected variables almost equals the number of LVs, which means that each selected variable represents one independent (uncorrelated) effect in the process.

#### 4 Summary and Conclusion

This work presents a multivariate analysis enabling to predict and to understand the relation of the cell culture process settings and variables (PAs) with the product quality attributes (CQAs) of the monoclonal antibody produced at milliliter scale. First, PCA shows that the 14 CQAs are strongly correlated and can be represented in a significantly smaller space of variables. Then, the classically used PLS1 models were built for each CQA using different amounts of process history to investigate the predictability of the product quality along the process time. In this analysis, the CQAs showed several characteristic predictability patterns. Motivated by the observations from the PCA, this analysis was repeated with PLS2 models simultaneously predicting the large set of 14 CQAs. Both, prediction accuracy as well as the characteristic predictability patterns remained comparable, while the number of models to be interpreted was substantially decreased enabling a much simpler basis for interpretation. Then, those PLS2 models were coupled to variable selection by a genetic algorithm (GA). This combination showed more accurate as well as more robust prediction results for all CQAs compared to using all process variables or the ones selected by the VIP criterion. Also



in this complex case, the GA was capable of eliminating noise, inconsistency and redundancy in the data and provided a characteristic set of very informative predictors for each time point during the process. These selections enabled to conclude on potentially influential media supplements (galactose, asparagine and iron) and process controls (temperature shift, pH shift and nitrogen flow). Moreover, those selection revealed the characteristic time points when the process variables such as viable cell density as well as lactate provide valuable information on (non-linear) effects in the process. Eventually, a predictive process model built by combining PLS2 with the GA was used to find and interpret the direct relation of all the PAs and the CQAs. First, the process-product-interrelationship was visualized for two time points in a joint loading plot, where constant interactions and dynamic changes could be found. Then, based on standardized  $\beta$ -coefficients for two CQAs one could visualize that some quality attributes are directly and constantly affected by media supplement additions and modifications of the process settings (e.g., LMW by iron and temperature shift level), while others are highly dependent on the non-linear process response to such additions and process manipulations (e.g., C4).

The presented methodology, showing the combination of PLS2 models with a GA, is generally applicable to process prediction and variable selection problems. In particular, given the large-dimensional process and product quality information in bioprocesses, the methodology provides a manageable basis for interpretation and understanding on the interrelationship of process attributes and CQAs. Although the experimental design did not specifically target this kind of analysis, the obtained results enabled to make decisions on the potentially influential process factors and their levels (or concentrations) and the design of the subsequent experiments. Moreover, the gained understanding allows to tune the process simultaneously towards certain goals (e.g., a defined product quality in biosimilar development) as well as to simplify or intensify the sampling scheme with regards to singly

the relevant process information. Such models are likely to serve as a robust basis for process monitoring, control and scale-up. The fact that all those results were obtained at an experimental scale of 15 mL highlight that reliable multivariate models can be generated based on automated high throughput experiments, which shall motivate the early process development at such economically and practically more attractive scale. Given all the provided benefits we strongly believe that the implementation of the presented methodology will significantly accelerate and simplify process development, so that both costs and time to market can be reduced. Therefore, we encourage the commercial software package developers to fine-tune their procedures towards the suggested level of complexity in order to automate and facilitate the data analysis procedure for bioprocess engineers and laboratory experts.

## Reference List

1. Brühlman D, Jordan M, Hemberger J, Sauer M, Stettler M, Broly H. Tailoring recombinant protein quality by rational media design. *Biotechnol Prog*. 2015;31(3):615-629. doi:10.1002/btpr.2089.
2. Bumbaca D, Boswell CA, Fielder PJ, Khawli LA. Physiochemical and Biochemical Factors Influencing the Pharmacokinetics of Antibody Therapeutics. *AAPS J*. 2012;14(3):554-558. doi:10.1208/s12248-012-9369-y.
3. Bruehlmann D, Sokolov M, Butte A, Sauer M, Hemberger J, Souquet J, Broly H, Jordan M. Parallel experimental design and multivariate analysis provides efficient screening of cell culture media supplements to improve Biosimilar product quality. *Biotechnol Bioeng*. 2017. doi:10.1002/bit.26269.
4. Ündey C, Ertunç S, Mistretta T, Looze B. Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control. *J Process Control*. 2010;20(9):1009-1018. doi:10.1016/j.jprocont.2010.05.008.
5. Sokolov M, Soos M, Neunstoecklin B, Morbidelli M, Butte A, Leardi R, Solacroup T, Stettler M, Broly H. Fingerprint detection and process prediction by multivariate analysis of fed-batch monoclonal antibody cell culture data. *Biotechnol Prog*. 2015;31(6):1633-1644. doi:10.1002/btpr.2174.
6. Charaniya S, Le H, Rangwala H, Mills K, Johnson K, Karypis G, Hu W-S. Mining manufacturing data for discovery of high productivity process characteristics. *J Biotechnol*. 2010;147(3-4):186-197. doi:10.1016/j.jbiotec.2010.04.005.
7. Le H, Kabbur S, Pollastrini L, Sun Z, Mills K, Johnson K, Karypis G, Hu W-S. Multivariate analysis of cell culture bioprocess data--lactate consumption as process indicator. *J Biotechnol*. 2012;162(2-3):210-223. doi:10.1016/j.jbiotec.2012.08.021.
8. Rouiller Y, Solacroup T, Deparis V, Barbafieri M, Gleixner R, Broly H, Eon-duval A. Application of Quality by Design to the characterization of the cell culture process of an Fc-Fusion protein. *Eur J Pharm Biopharm*. 2012;81(2):426-437. doi:10.1016/j.ejpb.2012.02.018.
9. Abu-Absi SF, Yang L, Thompson P, Jiang C, Kandula S, Schilling B, Shukla A a. Defining process design space for monoclonal antibody cell culture. *Biotechnol Bioeng*. 2010;106(6):894-905. doi:10.1002/bit.22764.
10. Nagashima H, Watari A, Shinoda Y, Okamoto H, Takuma S. Application of a quality by design approach to the cell culture process of monoclonal antibody production, resulting in the establishment of a design space. *J Pharm Sci*. 2013;102(12):4274-4283. doi:10.1002/jps.23744.
11. Rouiller Y, Périlleux A, Vesin M-N, Stettler M, Jordan M, Broly H. Modulation of mAb quality attributes using microliter scale fed-batch cultures. *Biotechnol Prog*. 2014;30(3):571-583. doi:10.1002/btpr.1921.
12. Schmidberger T, Posch C, Sasse A, Gülch C, Huber R. Progress toward forecasting product quality and quantity of mammalian cell culture processes by performance-based modeling. *Biotechnol Prog*. 2015;31(4):1119-1127. doi:10.1002/btpr.2105.
13. Rathore AS, Pathak M, Singh SK, Read EK, Agarabi CD, Khan M, Brorson KA, Kumar Singh S, Pathak M, Read EK, Brorson KA, Agarabi CD, Khan M. Fermentanomics: Relating Quality Attributes of a Monoclonal Antibody to Cell

- Culture Process Variables and Raw Materials using Multivariate Data Analysis  
Journal: *Biotechnol Prog*. 2015;n/a-n/a. doi:10.1002/btpr.2155.
14. Kourti T. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *J Chemom*. 2003;17(1):93-109. doi:10.1002/cem.778.
  15. García-Muñoz S, Kourti T, MacGregor JF, Mateos AG, Murphy G. Troubleshooting of an industrial batch process using multivariate methods. *Ind Eng Chem Res*. 2003;42(15):3592-3601. doi:10.1021/ie0300023.
  16. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;2(August):433-459. doi:10.1002/wics.101.
  17. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst*. 2001;58(2):109-130. doi:10.1016/S0169-7439(01)00155-1.
  18. Sokolov M, Ritscher J, MacKinnon N, Bielser JM, Bruehlmann D, Rothenhausler D, Thanei G, Soos M, Stettler M, Souquet J, Broly H, Morbidelli M, Butte A. Robust factor selection in early cell culture process development for the production of a biosimilar monoclonal antibody. *Biotechnol Prog*. 2017;33:181-191. doi:10.1002/btpr.2374.
  19. Martens H, Næs T. *Multivariate Calibration*. Wiley; 1989.
  20. Mercier SM, Rouel PM, Lebrun P, Diepenbroek B, Wijffels RH, Streefland M. Process analytical technology tools for perfusion cell culture. *Eng Life Sci*. 2016;16(1):25-35. doi:10.1002/elsc.201500035.
  21. Triba MN, Moyec L Le, Amathieu R, Goossens C, Bouchemal N, Nahon P, Rutledge DN, Savarin P. PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters. *Mol Biosyst*. 2014;11(1):13-19. doi:10.1039/C4MB00414K.
  22. Gauchi JP, Chagnon P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemom Intell Lab Syst*. 2001;58(2):171-193. doi:10.1016/S0169-7439(01)00158-7.
  23. Kirdar AO, Green KD, Rathore AS. Application of multivariate data analysis for identification and successful resolution of a root cause for a bioprocessing application. *Biotechnol Prog*. 2008;24(3):720-726. doi:10.1021/bp0704384.
  24. Mears L, Norregard R, Sin G, Gernaey K V, Stocks SM, Albaek MO, Villez K. Functional unfold principal component regression methodology for analysis of industrial batch process data. *AIChE J*. 2016;62(6):1986-1994. doi:10.1002/aic.15172.
  25. Ahuja S, Jain S, Ram K. Application of multivariate analysis and mass transfer principles for refinement of a 3-L bioreactor scale-down model-when shake flasks mimic 15,000-L bioreactors better. *Biotechnol Prog*. 2015;31(5):1370-1380. doi:10.1002/btpr.2134.
  26. Kirdar AO, Conner JS, Baclaski J, Rathore AS. Application of multivariate analysis toward biotech processes: case study of a cell-culture unit operation. *Biotechnol Prog*. 2007;23(1):61-67. doi:10.1021/bp060377u.
  27. Procopio S, Krause D, Hofmann T, Becker T. Significant amino acids in aroma compound profiling during yeast fermentation analyzed by PLS regression. *LWT - Food Sci Technol*. 2013;51(2):423-432. doi:10.1016/j.lwt.2012.11.022.
  28. Tsang VL, Wang AX, Yusuf-Makagiansar H, Ryll T. Development of a scale down

- cell culture model using multivariate analysis as a qualification tool. *Biotechnol Prog.* 2014;30(1):152-160. doi:10.1002/btpr.1819.
29. Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. *J Chemom.* 1992;6(5):267-281. doi:10.1002/cem.1180060506.
30. Deng B-C, Yun Y-H, Liang Y-Z. Model population analysis in chemometrics. *Chemom Intell Lab Syst.* 2015;149(B):166-176. doi:10.1016/j.chemolab.2015.08.018.
31. Duncan DB. t Tests and Intervals for Comparisons Suggested by the Data. *Biometrics.* 1975;31:339-359. <http://www.jstor.org/stable/2529425>.
32. Akaike H. Akaike's Information Criterion. In: *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011:25. doi:10.1007/978-3-642-04898-2\_110.
33. Folch-Fortuny A, Arteaga F, Ferrer A. PCA model building with missing data: new proposals and a comparative study. *Chemom Intell Lab Syst.* 2015;146:77-88. doi:10.1016/j.chemolab.2015.05.006.
34. von Stosch M, Hamelink J-M, Oliveira R. Hybrid modeling as a QbD/PAT tool in process development: an industrial E. coli case study. *Bioprocess Biosyst Eng.* 2016;39(5):773-784. doi:10.1007/s00449-016-1557-1.
35. del Val IJ, Kontoravdi C, Nagy JM. Towards the implementation of quality by design to the production of therapeutic monoclonal antibodies with desired glycosylation patterns. *Biotechnol Prog.* 2010;26(6):1505-1527. doi:10.1002/btpr.470.

## Figure Captions

**Figure 1** Definition of three characteristic data blocks: fixed process conditions by DoE (Z), time evolution of process variables which can be re-structured to a sequence of data blocks for each point of time (X) and final product quality attributes (Y). Black matrix elements indicate missing values.

**Figure 2** Building latent variable models for process data. PCA on process data (dashed line), PCA on product quality data (solid line) and PLSR of process data to product quality data (overall graphic).

**Figure 3** PCA on 14 quality attributes using grouped glycovariables: (A) Loading plot for first two PCs; (B) Score plot for first two PCs where experiments are coded according to underlying DoE (1 = hexagon, 2 = circle, 3 = diamond, 4 = triangle), the axes indicate the variance explained by the corresponding PC; (C) Cumulative variance explained by PCA with grouped (solid line) and specific glycovariables (dashed line).

**Figure 4** Variance explained in cross-validation ( $Q^2$  in %) for PLS1 models for different CQAs at day 14 (columns) based on different lengths of process history utilized (rows, where the first one represents the controlled process conditions and the further ones refer to last incorporated process day). The number of process variables utilized in each model is shown in brackets. Each table element corresponds to the result of one PLS1 model.

**Figure 5** Variance explained in cross-validation ( $Q^2$  in %) for PLS2 models for different CQAs at day 14 (columns) based on different lengths of process history utilized (rows, where the first one represents the controlled process conditions and the further ones refer to last incorporated process day). The number of process variables utilized in each model is shown in brackets and the corresponding optimal number of latent variables (opt LV) is presented in the last column. Each table row corresponds to the result of one PLS2 model.

**Figure 6** Scaled RMSECV and corresponding standard deviation for PLS2 models using all (white), VIP-based (grey) and GA-based (black) selected variables to predict Titer (A) and LMW (B) at different process times.

**Figure 7** GA selections for PLS2 models at different process times (for each case 100 models cross-validated with different calibration and test sets). Encircled boxes indicate available variables at the given process time, the height of the black filling indicates the frequency of selection. The average number of variables in each model is provided in brackets.

**Figure 8** Joint plot of  $W^*$ -loadings (black symbols) representing process attributes and  $C$ -loadings (white symbols) representing quality attributes in the PLS2 model. The axes quantify amount of variance explained by the respective LV out of the overall variance explained ( $Q^2$ ) in the corresponding PLS2 model: (A) using singly GA-selected Z information; (B) using GA-selected information throughout entire process duration (T14).

**Figure 9** Standardized  $\beta$ -coefficients (value 100 representing very strong positive effect and value -100 very strong negative effect, dark color indicating large absolute effect) in PLS2 models at different time points for LMW (A) and C4 (B). The first number in brackets indicates the number of variables selected by the GA and the second represents the number of LVs used in the corresponding model.



Accepted Article

**Table 1:** Overview of the investigated parameters in the four designed experiments (numbers indicate amount of different conditions tested, superscripts indicate different conditions).

DoE No.	1	2	3	4
No. of runs	7	24	24	36
Manganese (Mn)	1	1	1	2
Galactose (Gal)	1	6	3	2
Iron (Fe) <sup>1</sup>	1 <sup>*</sup>	1 <sup>#</sup>	1 <sup>#</sup>	1 <sup>#</sup>
Copper Zink (CuZn)	1	1	2	1
Asparagine (Asp)	1	7	2	7
Cobalt / Spermine / Gallic acid <sup>2</sup>	1	2	1	1
Sucrose	1	1	1	5
Temperature shift (Temp Shift)	2	1	1	3
Feeding regime (Feed)	2	1	1	1
Nitrogen flow (N2 Flow)	1	1	2	1

<sup>1</sup> Iron was tested on a different level in DoE1 compared to the other DoEs.

<sup>2</sup> The three parameters are coupled as they were always varied simultaneously.

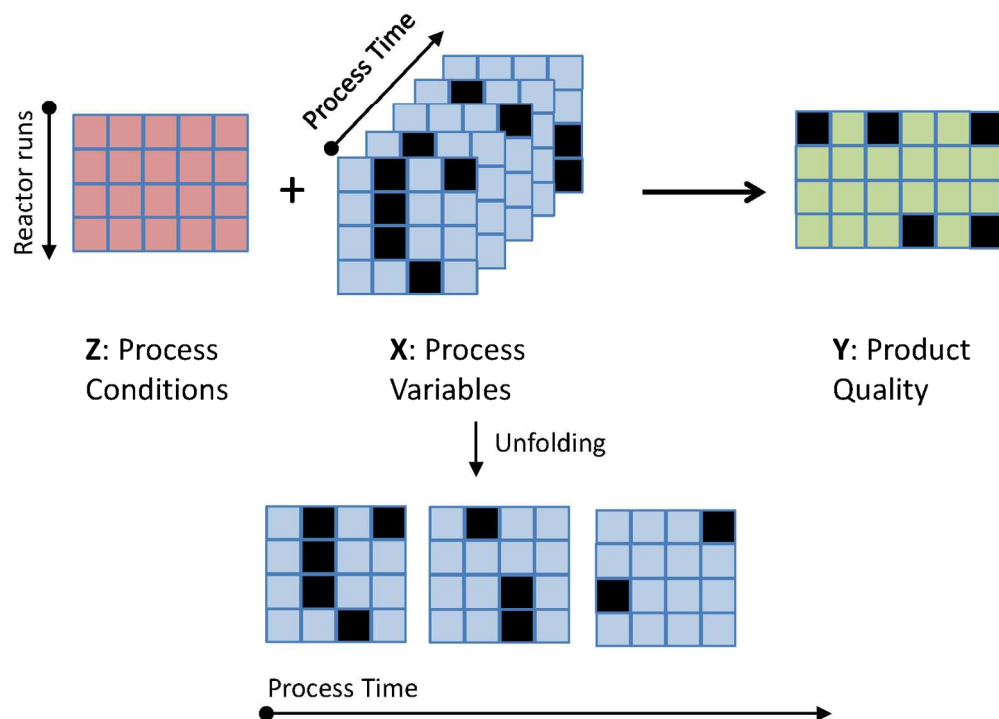


Figure 1 Definition of three characteristic data blocks: fixed process conditions by DoE (Z), time evolution of process variables which can be re-structured to a sequence of data blocks for each point of time (X) and final product quality attributes (Y). Black matrix elements indicate missing values.

149x110mm (300 x 300 DPI)

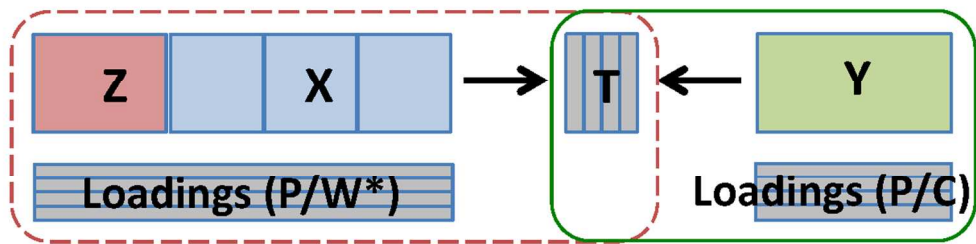


Figure 2 Building latent variable models for process data. PCA on process data (dashed line), PCA on product quality data (solid line) and PLSR of process data to product quality data (overall graphic).

105x30mm (300 x 300 DPI)

Accepted A

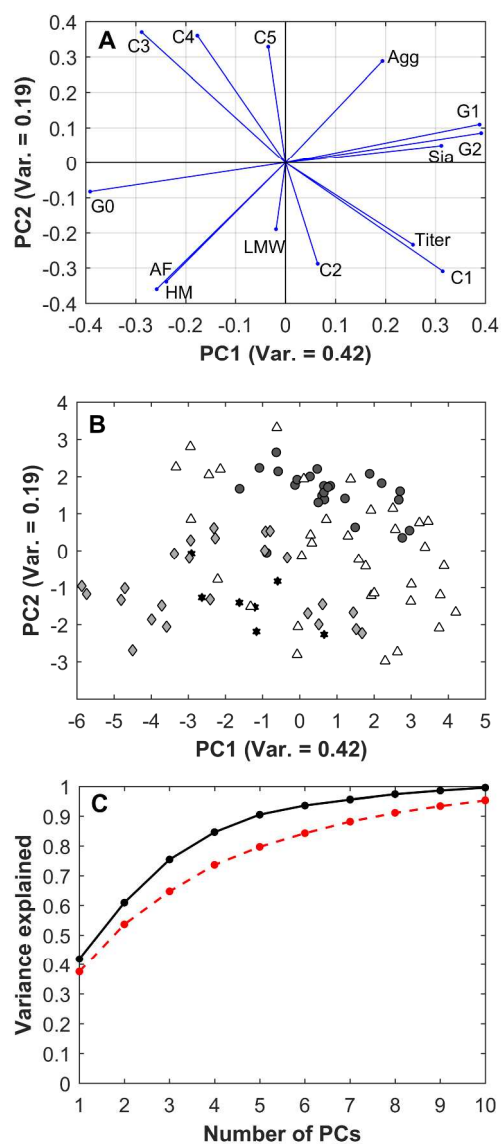


Figure 3 PCA on 14 quality attributes using grouped glycovariables: (A) Loading plot for first two PCs; (B) Score plot for first two PCs where experiments are coded according to underlying DoE (1 = hexagon, 2 = circle, 3 = diamond, 4 = triangle), the axes indicate the variance explained by the corresponding PC; (C) Cumulative variance explained by PCA with grouped (solid line) and specific glycovariables (dashed line).

190x424mm (300 x 300 DPI)

Z (12)	39	31	31	63	16	43	46	18	34	61	59	63	40	19	42
T3 (17)	51	35	33	72	56	52	68	31	37	64	63	66	50	41	41
T5 (25)	64	67	67	76	75	63	74	51	57	72	69	78	49	43	47
T7 (33)	72	67	66	77	77	67	75	65	63	75	73	81	49	40	51
T10 (41)	75	74	73	80	78	69	75	72	66	75	73	81	49	40	53
T12 (49)	82	75	74	77	78	66	74	74	67	77	74	82	52	41	53
T14 (57)	81	73	73	80	79	74	75	73	66	77	73	82	54	43	52
	Titer Monomers	Agg	LMW	C1	C2	C3	C4	C5	G0	G1	G2	HM	AF	Sia	

Figure 4 Variance explained in cross-validation (Q2 in %) for PLS1 models for different CQAs at day 14 (columns) based on different lengths of process history utilized (rows, where the first one represents the controlled process conditions and the further ones refer to last incorporated process day). The number of process variables utilized in each model is shown in brackets. Each table element corresponds to the result of one PLS1 model.

70x44mm (300 x 300 DPI)

Accept

Z (12)	24	30	30	62	21	42	45	18	32	59	58	61	41	21	41	6
T3 (17)	30	32	30	65	47	41	57	22	35	60	59	62	50	40	41	5
T5 (25)	62	57	56	73	66	60	68	49	50	67	65	72	46	44	44	8
T7 (33)	72	63	62	75	71	64	73	62	58	74	72	77	50	42	50	10
T10 (41)	75	69	68	74	76	64	74	68	65	75	72	79	47	38	51	12
T12 (49)	80	73	73	76	77	66	71	74	68	77	74	83	47	37	47	15
T14 (57)	80	74	74	77	78	70	73	73	67	77	73	82	50	39	50	15
	Titer Monomers	Agg	LMW	C1	C2	C3	C4	C5	G0	G1	G2	HM	AF	Sia	opt LV	

Figure 5 Variance explained in cross-validation (Q2 in %) for PLS2 models for different CQAs at day 14 (columns) based on different lengths of process history utilized (rows, where the first one represents the controlled process conditions and the further ones refer to last incorporated process day). The number of process variables utilized in each model is shown in brackets and the corresponding optimal number of latent variables (opt LV) is presented in the last column. Each table row corresponds to the result of one PLS2 model.

70x44mm (300 x 300 DPI)

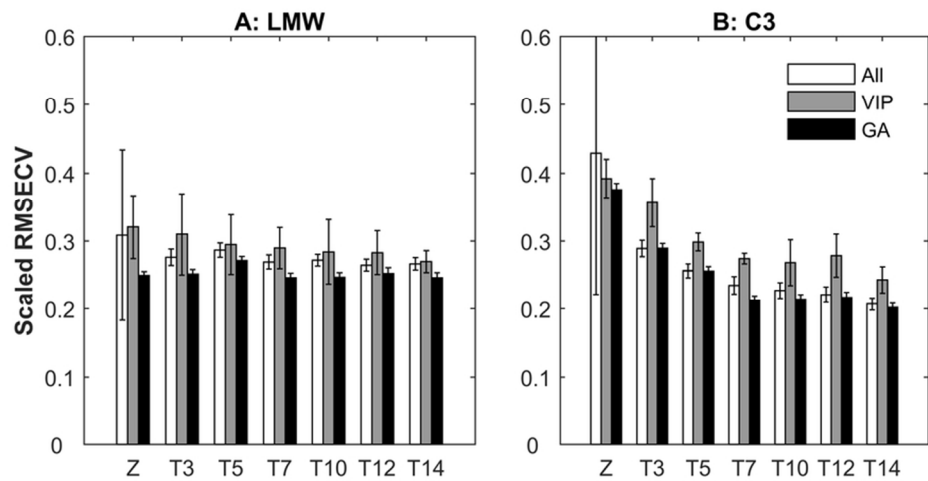


Figure 6 Scaled RMSECV and corresponding standard deviation for PLS2 models using all (white), VIP-based (grey) and GA-based (black) selected variables to predict Titer (A) and LMW (B) at different process times.

80x40mm (300 x 300 DPI)

Accepted

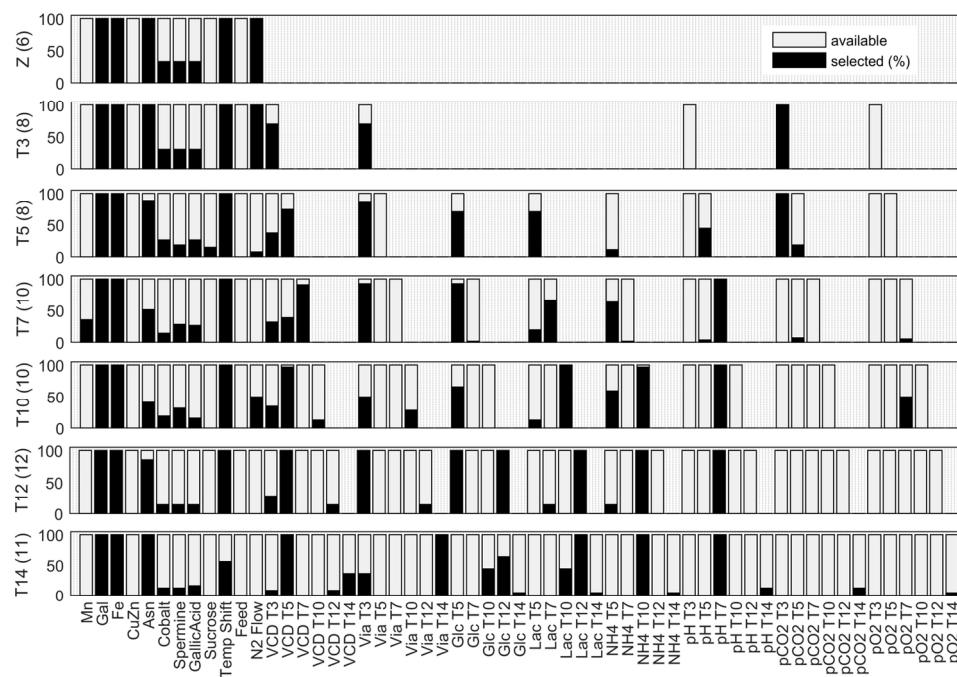


Figure 7 GA selections for PLS2 models at different process times (for each case 100 models cross-validated with different calibration and test sets). Encircled boxes indicate available variables at the given process time, the height of the black filling indicates the frequency of selection. The average number of variables in each model is provided in brackets.

139x96mm (300 x 300 DPI)



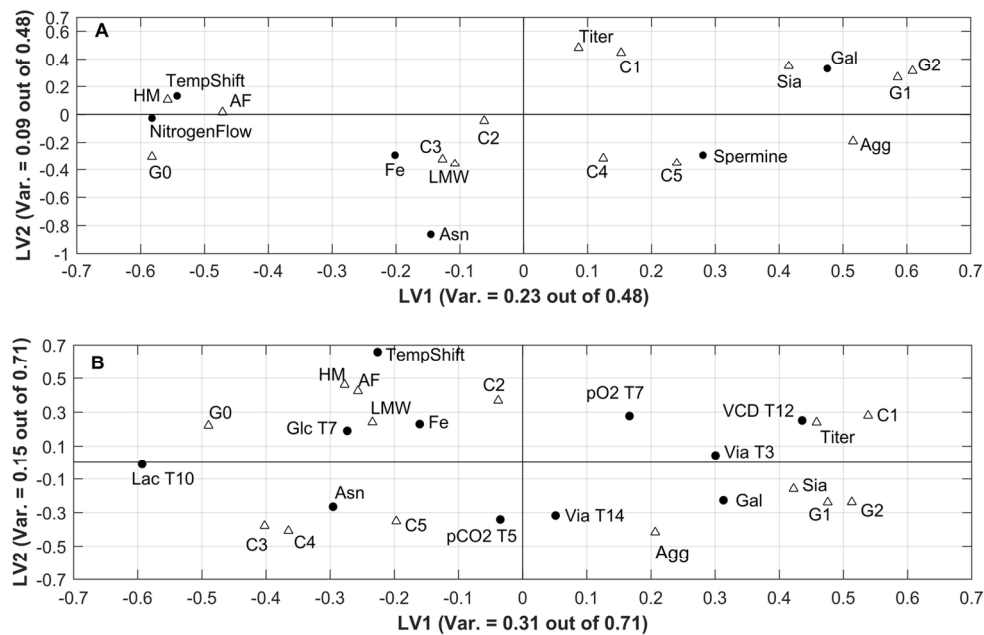


Figure 8 Joint plot of W\*-loadings (black symbols) representing process attributes and C-loadings (white symbols) representing quality attributes in the PLS2 model. The axes quantify amount of variance explained by the respective LV out of the overall variance explained (Q2) in the corresponding PLS2 model: (A) using singly GA-selected Z information; (B) using GA-selected information throughout entire process duration (T14).

139x96mm (300 x 300 DPI)



162x147mm (300 x 300 DPI)