BioProcessing JOURNAL
TRENDS & DEVELOPMENTS IN BIOPROCESS TECHNOLOGY

# Optimization of a Conventional Glycosylation Analytical Method Using Machine Learning and Experimental Design

By Eliza Yeung and Philip Ramsey

## Abstract

Glycosylation is one of the most common post-translational modifications in mammalian-expressed biologics, and is considered to be a critical quality attribute of therapeutic glycoproteins. Due to its biological relevance, physiochemical assessment on the glycosylation profile is always important to the success of a drug development initiative. This article describes the combination of experimental design and machine learning techniques applied to characterize and optimize a conventional, non-derivatized glycoprofiling method on glycans derived from a human immunoglobulin using high-performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD). Two independent experimental designs, a 16-run definitive screening design (DSD) and a 28-run central composite design (CCD), were incorporated with a machine learning technique known as "self-validating ensemble modeling (SVEM)" and used to build predictive models for four chromatographic responses. We show that the predictive models created using SVEM on the DSD data reliably predicted the behavior of the chosen responses when applied to CCD validation data. This demonstrates that the DSD is an efficient alternative to the larger, traditional CCD in which the combination of experimental design and machine learning can effectively characterize and optimize analytical methods.

## Introduction

Glycosylation is a common post-translational modification in a glycoprotein produced by mammalian cell expression. It is commonly thought that glycosylation structures impact drug product quality in the areas of safety, efficacy, stability, and immunogenicity.[1] Glycosylation is a complex biosynthesis process dependent on many factors that result in structural variations of a glycoprotein. The formation of various glycoforms (glycosylation variants) is controlled by a combination of enzymatic activities and regulation processes of glycosyltransferases and glycohydrolases that utilize monosaccharide substrates of structural variation, but close chemical similarity. Moreover, these enzymatic pathways are influenced by their expression levels, subcellular localization, as well as the glycoprotein production's processes and conditions. Consequently, glycosylation is a non-template-driven and highly process-specific biosynthesis process that may lead to considerable possible combinations of heterogeneous glycoforms of different glycan sizes.

Due to the profound contributions of therapeutic glycoproteins to current and future drug markets[2], and their complexity and structural diversity of glycans, controlling protein glycosylation is yet another important measure in a product lifecycle that can assure product quality. However, achieving consistent glycosylation profiles and reliable structural analysis for drug development still remains a challenge. To date, a non-mass spectrometric analytical platform using high-performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) has been widely used.[3] The applications of HPAEC-PAD are described in three USP general chapters for the analysis of unlabeled glycans or monosaccharides.[4–6] The method has demonstrated excellent coverage of *N*-glycan species, and the peak assignment is identified by either spiking or retention time comparison with glycan standards. Moreover, the HPAEC-PAD method has been routinely used to analyze neutral and sialylated *N*-glycans of monoclonal antibodies.[7]

In this article, a procedure of glycoprofiling assessment on glycans derived from a human immunoglobulin, featuring a simple non-derivatized analytical procedure of HPAEC-PAD integrated with design of experiment (DOE), was demonstrated using two experimental designs, a definitive screening design (DSD) and a central composite design (CCD). The DOE constitutes a large class of efficient, information-centric data collection structures for scientific inquiry. For a given set of experimental factors of size K and associated levels, one can create a design to estimate a set (possibly large) of potential experimental effects with a relatively small number of trials (runs).

DSDs were developed by Jones and Nachtsheim[8] and have become popular in biopharma due to the relatively small number of trials required to estimate experimental effects.[9] For process characterization and optimization, an empirical "full quadratic" (FQ) model is utilized[10] or a larger and more complex variant of the FQ we refer to as the "partial cubic" (PC)[11] (also known as the "interaction") model is employed. The FQ model includes all main effects (first-order terms), all quadratic effects (second-order terms), and two-way interaction effects (cross-product terms). Experimental designs capable of estimating terms in the FQ or PC models are often referred to as "response surface designs" (RSD). The CCD is a well-known example of an RSD. The smaller DSD is a very efficient RSD (compared to the CCD) in terms of number of runs, as DSD is capable of estimating both two-way interaction and quadratic terms. Since the DSD is relatively new, it was decided to independently perform a DSD plus a CCD to validate and compare each design's effectiveness in glycoprofiling. Both the experimental designs and subsequent analyses were performed with JMP® statistical software (SAS Institute, Inc.).

In summary, the results demonstrated that both DSD (16 trials) and CCD (28 trials) experiments resulted in the same predictive models and conclusions where, in this case study, the number of experimental runs used for DSD were only 60% of those for CCD.

## Materials

Analytical, reagent-grade chemicals and water purified using the Nanopure™ system (Thermo Scientific) to 18 MΩcm (or above) were used. Mobile phases were prepared from sodium acetate (Sigma S7545) and sodium hydroxide (J.T. Baker 3727). The in-house glucose ladder was prepared (non-derivatized) from a partial acidic hydrolysis of dextran (Sigma D9260). Glycans were derived from RNase B (Sigma R7884) and a purified human immunoglobulin G (IgG) (Sigma I4506). Briefly, PNGase F (Sigma P7367) digestion was performed following standard procedures to release *N*-glycans and high mannose glycans. All glycan analyses, the glucose ladder, released *N*-glycans of a human IgG, and released high mannose glycans of RNase B, were performed on a HPAEC-PAD (Thermo Scientific) system.

## Methods

### High-Performance Anion-Exchange Chromatography with Pulsed Amperometric Detection

The Thermo Scientific Dionex™ CarboPac™ PA100 (2×250 mm) column with a 2×50 mm guard column was used (maintained at 30°C). Waveform A was applied to the electrochemical detector for glycan analysis. Detection was carried out by pulsed amperometry with a gold electrode and Ag/AgCl reference electrode. The glycans were eluted with gradients of sodium acetate (NaOAc) in sodium hydroxide (NaOH) at a flow rate of 0.25 mL/min. A partial-loop mode of "5 µL injection with a 3 µL cut volume" was used. The Chromeleon™ data system (Thermo Scientific) provided necessary integration.

### HPAEC-PAD Mobile Phases

- **A:** 500 mM NaOAc
- **B:** 200 mM NaOH with 10 mM NaOAc
- **C:** Deionized water

Method gradient for a standard run was 60 min at:

- T = 0 min, 2% A, 40% B, and 58% C
  (% C was calculated as the remaining % of A and B)
- T = 12 min, linear Gradient_01
  (1.25 mM NaOAc/min) to reach final 5% A and constant 40 % B
- T = 24 min, linear Gradient_02
  (2.085 mM NaOAc/min) to reach final 10% A and constant 40% B
- T = 42 min, linear Gradient_03
  (5.555 mM NaOAc/min) to final 30% A and constant 40% B

Equilibration to next initial condition was applied during:

- T = 42.1–60 min

### Design of Experiment for Glycoprofiling

JMP statistical software was used for experimental designs and subsequent analyses. Optimization of the glycoprofiling method was performed on four responses using DOEs on five factors of three levels. Two DOEs, DSD and CCD, consisted of 16 and 28 trials respectively, and were performed with JMP 14.2 (or JMP Pro 14) software. The SVEM analyses of the experimental data were performed using the JMP Pro 15 program.

## Results and Discussion

### Glycoprofiling Approach Integrated with Experimental Design

The utilization of DOE in analytical method development has become an upward trend under the quality by design

(QbD) paradigm.[12,13] In this study, DOE was applied to characterize and optimize the HPAEC-PAD method for direct analysis of glycan without sample derivatization.[7] In general, glycans are separated on the column and each individual glycan is assigned in terms of the retention time. To achieve a reliable structural analysis that supports physiochemical assessment of a proposed glycoprotein, a good peak resolution is critical. Moreover, optimizing DOE studies of multiple runs requires a significant amount of glycan sample released from a drug. Instead, this glycoprofiling procedure includes a surrogate sample, a glucose ladder, in place of a glycan sample in the DOE runs.

The glycan peaks are expressed in terms of relative retention time (RRT). RRTs are expressed as glucose units (GUs) aligned with glucose oligomer peaks of an external glucose ladder (*e.g.,* GU=1 for glucose, GU=2 for glucose dimer, and so on). Any glycan peak can be expressed as a GU, whereby the peak located between two GUs is calculated as a fraction of GU, assuming a linear interpolation. In fact, a labeled GU library was applied to reduce experimental variation and give high reproducible GU values for the assignment of 2-aminobenzamide (2-AB) labeled *N*-linked glycans.[14] Moreover, GU values of an external malto-oligosaccharide ladder had been used to assign peaks of non-derivatized galacto-oligosaccharides.[15]

Typically, a glycoprotein is treated with PNGase F to release glycans in which GU values are assigned to peaks according to an external glucose ladder. The glucose ladder serves as both a surrogate sample and a system suitability in the glycoprofiling method. Subsequent optimization DOE studies are performed using the surrogate sample (the glucose ladder) with selected DOE factors and responses of interest. Predictive models are chosen for each response. Consequently, the analytical method is optimized with all responses of interest. In the case study, a human IgG is used. **Figure 1** shows an overlay of the representative IgG unlabeled glycans with the glucose ladder. Prior to the case study, a good precision of peak retention times (GUs) for the major glycans of IgG (A–F in **Figure 1**) and the high mannose glycoforms of RNase B was demonstrated (data not shown). Hence, the glucose ladder is deemed to be a suitable surrogate sample that can reliably represent the glycans released from a glycoprotein.

In the study, the DSD had 16 trials, including three additional replicate center points, and the CCD had 28 trials, including four replicate center points. The DSD was used as a training set for empirical model development, and the CCD was used as a validation set to test the accuracy of the DSD-based models. **Table 1** contains the details of the experimental (five chromatographic) factors each and levels (three) for both the DSD and CCD. For each run in the experiments, four peak characteristics of retention time (RT), resolution (Resol), peak area, and tailing factor were tracked to evaluate the desirable method performance for each of eleven glycans, G01 to G11, where "G#" denotes a peak corresponding to "GU#" as shown in **Figure 1**. Since there are 44 potential responses to analyze, it was decided to select four representative responses for discussion in this article. The four responses chosen for the purpose of glycoprofiling method characterization and optimization are the retention time for peak G03 (RT_G03), the resolution of peaks between G03–G04 (Resol_G03), G05–G06 (Resol_G05), and G10–G11 (Resol_G10). Retention time for G03 is targeted for 8.5 min. The goal for Resol_G03 and Resol_G05, corresponding to the region of neutral glycans (elute before the peak of G06), is to maximize resolution. The goal for Resol_G10, corresponding to the region of charge glycans (charged sialylated and highly charged glycans are expected to elute after the peak of G08 and G11,

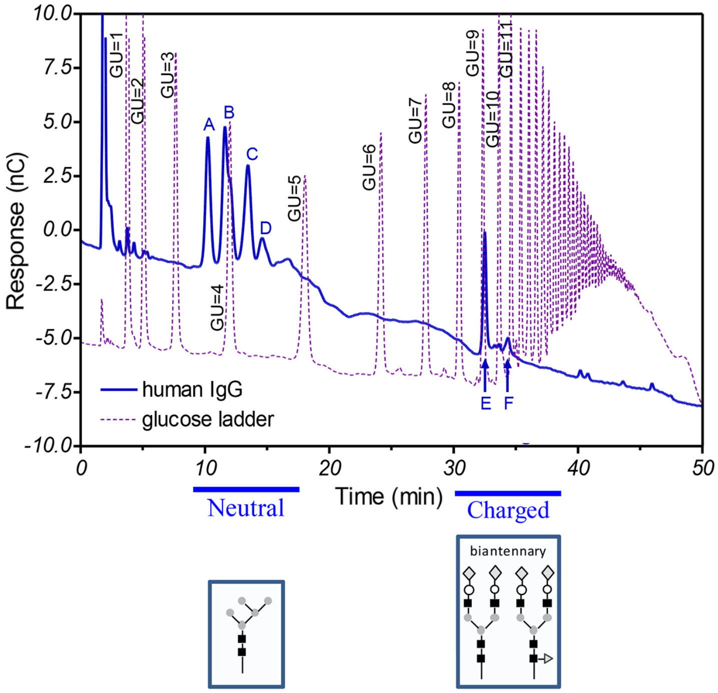


**FIGURE 1.** Glycoprofile assignment of a human IgG.

**TABLE 1. DOE levels for HPAEC-PAD method.**

| **DOE Factor (level)** | **Concentration (mM)** | | |
|---|---|---|---|
| | **Low** | **Middle** | **High** |
| Initial NaOAc Concentration | 3 | 14 | 25 |
| Initial NaOH Concentration | 60 | 80 | 100 |
| **DOE Factor (level)** | **Method Gradient (mM/min)** | | |
| | **Low** | **Middle** | **High** |
| Linear Gradient_01 (0–12 min) | 0.415 | 1.250 | 2.085 |
| Linear Gradient_02 (12–24 min) | 1.250 | 2.085 | 2.915 |
| Linear Gradient_03 (24–42 min) | 4.720 | 5.555 | 6.390 |

respectively) is also to be maximized. Moreover, the peak shapes of G04 and G05 tend to be relatively broad and less symmetric compared to those of early and late eluted peaks, but the actual shapes are dependent upon experimental conditions. When assessing the method performance of tailing factor, G04 or G05 (or both) are representative of the worst-case. Hence, tailing factor was monitored to remain within the range of 0.8–1.2 based on the standard requirement of USP tailing, but not formally analyzed.

## Strategy for Analysis of the Experimental Data

The goal of the analyses is to find settings of the experimental factors that optimize the performance of the four responses selected for the glycoprofiling. In the optimization process with experimental data, it is essential to estimate an empirical model that accurately predicts future method performance. We simply refer to these as "predictive models." If an empirical model cannot accurately predict future performance, then optimized factor settings based on the experimental results will not accurately characterize future performance. The optimized results are not useful in this case. With regard to assessing the future accuracy of a predictive model, in machine and deep learning, the data are partitioned into a training set used to estimate the model and a separate validation or test set used to determine the predictive capability of that estimated model. Predictive models that accurately and precisely predict the responses in the validation set are preferred; we say they generalize. In the present work, the DSD is defined as the training set and is used to estimate predictive models for the four chosen responses. Meanwhile, the CCD serves as a validation set used to evaluate the generalizability of the DSD-based predictive models. If the predictive models validate on the CCD data, then this indicates that the smaller DSD design effectively captured the relationships between the responses and the experimental factors, demonstrating that the DSD is an efficient alternative to the larger CCD.

Predictive model building is the primary focus of machine learning and deep learning while traditional statistical modeling, with an emphasis on hypothesis tests of model components, is more focused on explanation or attribution.[16,17] Unfortunately, explanatory models tend to predict poorly compared to predictive models.[16] Explanatory models are not constructed to predict and therefore, are usually not generalizable where they predict poorly on new data. For the glycoprofiling data, we use machine learning methods to build predictive models (reference[18] provides a discussion of common machine learning methods used to build predictive models) on the DSD data, validate the models on the CCD data, and subsequently, use the validated models to optimize the HPAEC-PAD glycoprofiling method.

Commonly used predictive model-building algorithms are unstable, a challenge not well-recognized by scientists.[19] Small perturbations in the response values (*e.g.,* the response noise changes slightly) result in substantial changes in the models selected by the algorithm. Therefore, considerable variation in prediction performance exists among selected models. In fact, a predictive model generated by a single pass through a selection algorithm is in part an artifact of the noise in the response. Thus, considerable predictive model uncertainty exists.

A solution to the instability problem[19] is to perform a set of predictive model-building simulation trials of some size, possibly thousands of iterations. On each simulation trial, the original responses are perturbed by new random noise values and a predictive model is selected. Thousands of predictive models so obtained are then averaged together to create an average (ensemble) predictive model. Ensemble modeling is commonly used in machine learning[20] in part, to overcome the instability issue. Ramsey *et al*.[21] and Lemkus *et al*.[22] have shown that the ensemble approach to predictive model-building results in more stable models and superior predictive performance as compared to traditional methods based on a single model-building pass through the data. This strategy[21,22] is known as "self-validating ensemble-modeling (SVEM)." The SVEM method was used in the current work to fit predictive models to each of the four responses in the HPAEC-PAD experiment. However, a full discussion of SVEM is beyond the scope of this paper.

## Analysis of the HPAEC-PAD Experimental Data

Although the SVEM algorithm was used to fit predictive models, the analyst must first define the model. As discussed earlier, two candidate models for optimization tasks are the FQ (equation 1) and PC (equation 2):

$$Y = \beta_o + \sum_{i=1}^{K} \beta_i X_i + \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \beta_{ij} X_i X_j + \sum_{i=1}^{K} \beta_{ii} {X_i}^2$$

**Equation 1**

$$Y = \beta_o + \sum_{i=1}^{K} \beta_i X_i + \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \beta_{ij} X_i X_j + \sum_{i=1}^{K} \beta_{ii} {X_i}^2 + \sum_{i=1}^{K} \sum_{j \neq i}^{K} \beta_{iij} {X_i}^2 X_j + \sum_{i=1}^{K} \sum_{j \neq i}^{K} \beta_{ijj} X_i {X_j}^2 + \sum_{i=1}^{K} \sum_{j \neq i}^{K} \beta_{iijj} {X_i}^2 {X_j}^2$$

**Equation 2**

The FQ model includes main effects, and all two-way interaction and quadratic effects to account for curvature in the response behavior. The larger PC model builds upon the FQ model by adding in all possible linear-by-quadratic and quadratic-by-quadratic interaction terms. The FQ is often insufficient to model the complex kinetics found in chemical and biological systems.[11] The addition of the linear-by-quadratic effects to the PC model provides more flexibility to accommodate complex kinetic behavior. Similarly, the PC model is expected to accommodate complicated performance behavior observed in analytical procedures.

**Figure 2** displays an interaction graph for Resol_G10 vs Initial NaOAc and Initial NaOH. Notice that the curvature in the relationship between Resol_G10 and Initial NaOAc changes with the levels of Initial NaOH. We say that the quadratic effect of Initial NaOAc interacts with the linear effect of Initial NaOH. The FQ model cannot adapt to linear-by-quadratic interaction behavior. However, the PC model does fit linear-by-quadratic interactions and has more flexibility to fit response surface models where the underlying kinetics are complex. There are scenarios where linear-by-quadratic interactive behavior is commonly observed.

Using the SVEM algorithm, an ensemble predictive model was fit to each of the four chosen responses. Both the FQ and PC models were fit to the responses, and the model with the smallest prediction error on the CCD validation data was selected as the final model. All models fit using SVEM were based on the forward selection method.[18] **Table 2** summarizes the predictive modeling results. As shown, the root average square error (RASE) column is the standard deviation of prediction error for the CCD validation data, and a smaller error indicates a more accurate prediction. The $R^2$ validation column is the observed $R^2$ for the predictive model applied to the CCD validation data and provides a measure of goodness of fit for the model. The percent relative standard deviation (% RSD) column is based on dividing the RASE by the average of each response. Overall, the noise or variation in each response appears to be consistent (neither increasing nor decreasing) over the range of observed values, so for comparison purposes it was decided to calculate % RSD at the average of each response. The % RSD value for Resol_G10, based on the FQ model (bottom row in the table), is not meaningful because no relationship exists between the observed values and the model predictions, $R^2 = 0.065$.

A common machine learning method to visualize the quality of the fit for a predictive model is an actual by predicted plot. The actual observed responses are plotted on the Y-axis and the corresponding predicted values are plotted on the X-axis. If the predictive model is a good fit to the observed values, then the plot should exhibit a linear trend with a slope of ∼1.0 and an intercept ∼0.0. **Figure 3** displays the actual by predicted plots of the four responses in **Table 2**. In all four cases, the observed slopes and intercepts are not significantly different from 1.0 and 0.0 respectively. This provides evidence that the four predictive models, estimated using the DSD training data, provide accurate prediction of the corresponding responses in the CCD validation data. Notice too that in the actual by predicted plots, the level of variation or noise in the responses is consistent over the response ranges. This provides visual evidence that the level of noise or variation in the responses is constant over the entire range of observed values.

In order to illustrate the instability problem with traditional model fitting, which makes a single pass through a model-fitting algorithm, three predictive models were fit using the SVEM algorithm, but only a single pass through

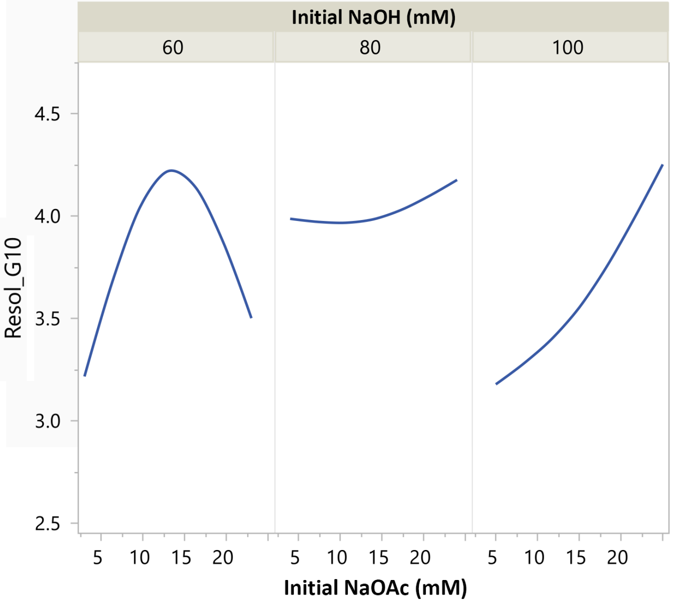


**FIGURE 2.** Interaction graph of Resol_G10 vs Initial NaOAc and Initial NaOH.

**TABLE 2. Predictive modeling results for four responses using SVEM.**

| Response* | Model | Number of Predictors | RASE Validation | $R^2$ Validation | % RSD Validation** |
|---|---|---|---|---|---|
| RT_G03 | PC† | 26 | 0.186 | 0.996 | 2.80 |
| Resol_G03 | PC | 26 | 0.721 | 0.925 | 12.67 |
| Resol_G05 | PC | 40 | 0.406 | 0.893 | 7.01 |
| Resol_G10 | PC | 40 | 0.246 | 0.778 | 7.08 |
| Resol_G10 | FQ | 20 | 0.478 | 0.065 | 13.76 |

*SVEM was used with $N$=1,000 models in each ensemble.
**Based on RASE divided by the average of each response.
†RT_G03 cannot be impacted by Gradient_02 and Gradient_03 due to shorter retention time.

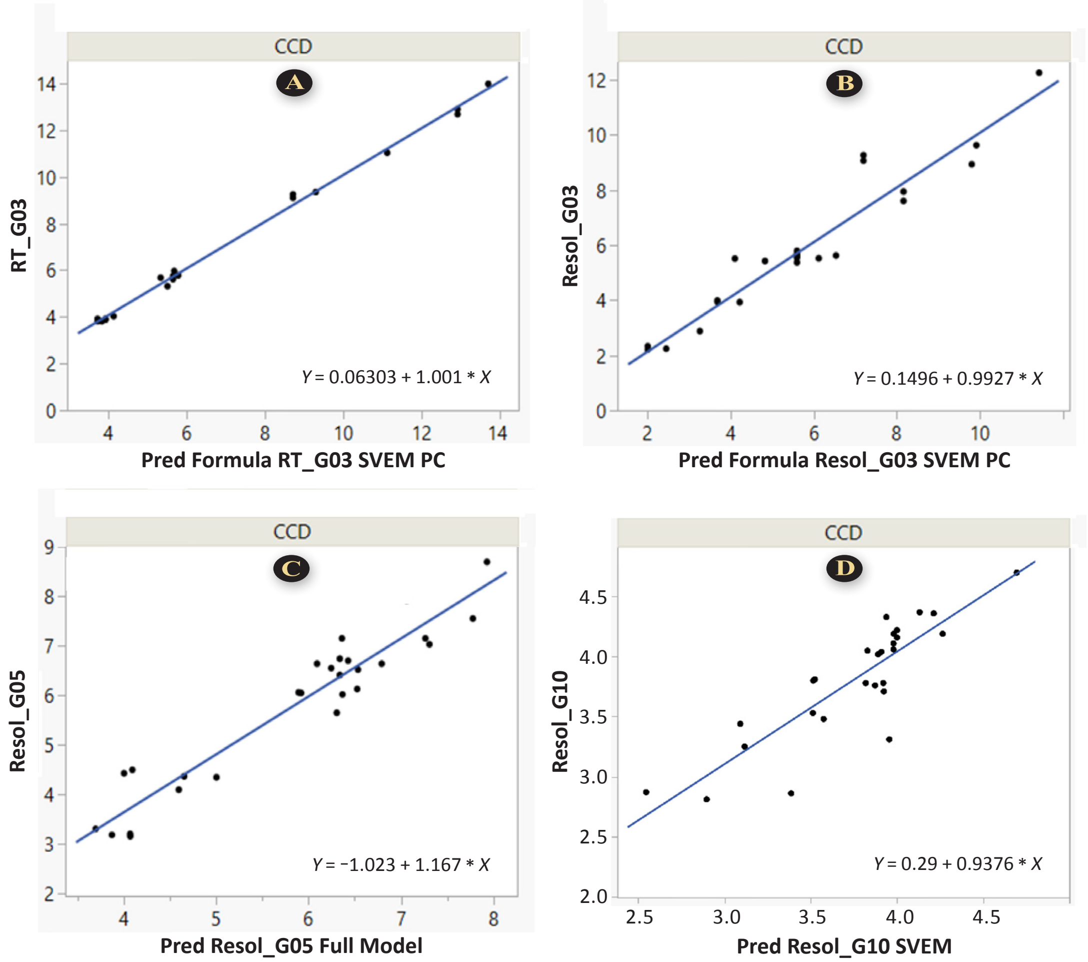


**FIGURE 3.** Actual by predicted plots for the four responses: **(A)** RT_03; **(B)** Resol_G03; **(C)** Resol_G05; and **(D)** Resol_G10.

the forward selection algorithm was used. The exercise was limited to the Resol_G10 response, and **Table 3** contains the summary of the predictive modeling results for the CCD validation data. The first row in **Table 3** presents the results for the best SVEM model from **Table 2**, and the other three rows are the results from three individual models fit to the DSD data using a single pass through the forward selection method. The first row in **Table 3** shows that the original SVEM model, based upon an ensemble average of 1,000 models, performs best in terms of RASE, $R^2$, and slope. Models PC(1) and PC(2) exhibit very poor prediction performance while PC(3) performs well, although inferior to the original SVEM model in the first row. Unfortunately in practice, if one generates a single predictive model from a single pass through a model-fitting algorithm (*e.g.,* forward selection), then that fitted model may perform very poorly since substantial model uncertainty exists. The SVEM algorithm overcomes the uncertainty by fitting many models and using the ensemble average as the final predictive model. In the end, the results presented in **Table 3** demonstrate the effectiveness of the SVEM algorithm.

**TABLE 3. Validation results for four models fit to Resol_G10 using the DSD data.**

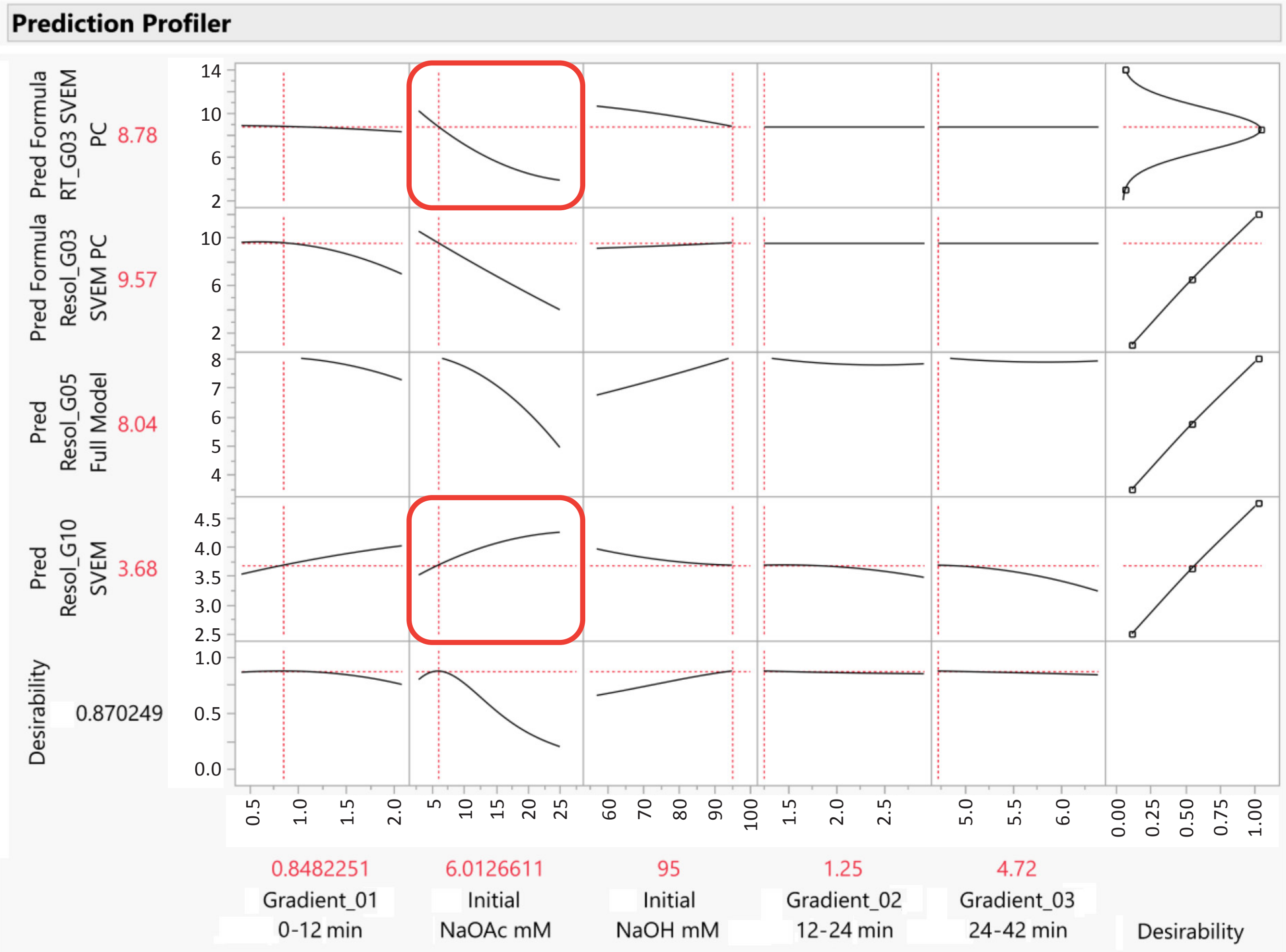
| Response | Model | SVEM Iterations | RASE Validation | $R^2$ Validation | Slope* |
|---|---|---|---|---|---|
| Resol_G10 | PC | 1,000 | 0.246 | 0.778 | 0.939 |
| Resol_G10 | PC(1) | 1 | 0.431 | 0.285 | 0.499 |
| Resol_G10 | PC(2) | 1 | 0.470 | 0.165 | 1.356 |
| Resol_G10 | PC(3) | 1 | 0.272 | 0.720 | 0.814 |

*Slope of the actual by predicted plot for each model.

## Optimization

Once acceptable predictive models have been estimated for each of the responses of interest, those models can be subsequently used for characterization and optimization of the method in study. Characterization and optimization are demonstrably about prediction where one is attempting to predict future performance of the procedure. We use the four predictive models identified in **Table 2** to simultaneously optimize the HPAEC-PAD method. Recall that the goal is to hit a target of 8.5 min for RT_G03, and the resolutions of Resol_G03, Resol_G05, and Resol_G10 are to be maximized. Optimization by employing empirical models, predictive models in the present case, is typically performed using desirability functions.[23] Desirability functions are mappings of each original response to a dimensionless 0.00–1.00 scale. The actual functional form depends upon the goal of the optimization. The most desirable level would be 1.00 with 0.00 being a least desirable level of performance. The set of responses are then optimized on the dimensionless desirability scale, removing any potential impacts on optimization of very different measurement scales among the responses. Optimization using desirability functions as the responses and the predictive models for each response search, the design space for settings of the experimental factors, provides the overall most desirable levels for all responses. The desirability optimization method is implemented in the Prediction Profiler application of the JMP software.

We illustrate desirability optimization using JMP and the four predictive models applied to the CCD validation data. **Figure 4** displays the optimized results using the JMP Prediction Profiler. The contours in each cell depict the nature of the relationship between each experimental factor and the responses. The two highlighted cells in **Figure 4** show the relationships between RT_G03 and Resol_G10 to Initial NaOH. Notice that Initial NaOH has a negative and nonlinear effect on RT_G03 while Initial NaOH has a positive and nonlinear impact on Resol_G10. In other words, increasing Initial NaOH results in reduced retention time for G03 while increasing the resolution of Resol_G10. The optimization routine searches for settings of the experimental factors that predict the most desirable levels that



Prediction Profiler

Pred Formula RT_G03 SVEM PC 8.78

Pred Formula Resol_G03 SVEM PC 9.57

Pred Resol_G05 Full Model 8.04

Pred Resol_G10 SVEM 3.68

Desirability 0.870249

Gradient_01 0-12 min 0.8482251

Initial NaOAc mM 6.0126611

Initial NaOH mM 95

Gradient_02 12-24 min 1.25

Gradient_03 24-42 min 4.72

**FIGURE 4.** Optimization results for the four responses using desirability functions.

**TABLE 4. Simultaneously optimized settings of experimental factors for all responses.**

| Factor | Initial NaOAc | Initial NaOH | Gradient_01 | Gradient_02 | Gradient_03 |
|---|---|---|---|---|---|
| Setting | 6.013 | 95 | 0.848 | 1.25 | 4.72 |

**TABLE 5. Optimized predicted values for four responses.**

| Response | RT_G03 | Resol_G03 | Resol_G05 | Resol_G10 |
|---|---|---|---|---|
| Predicted Value | 8.78 | 9.57 | 8.04 | 3.68 |

can be achieved across all four responses. **Table 4** lists the optimized settings for each of the experimental factors. **Table 5** lists the optimized predicted values for each of the four responses.

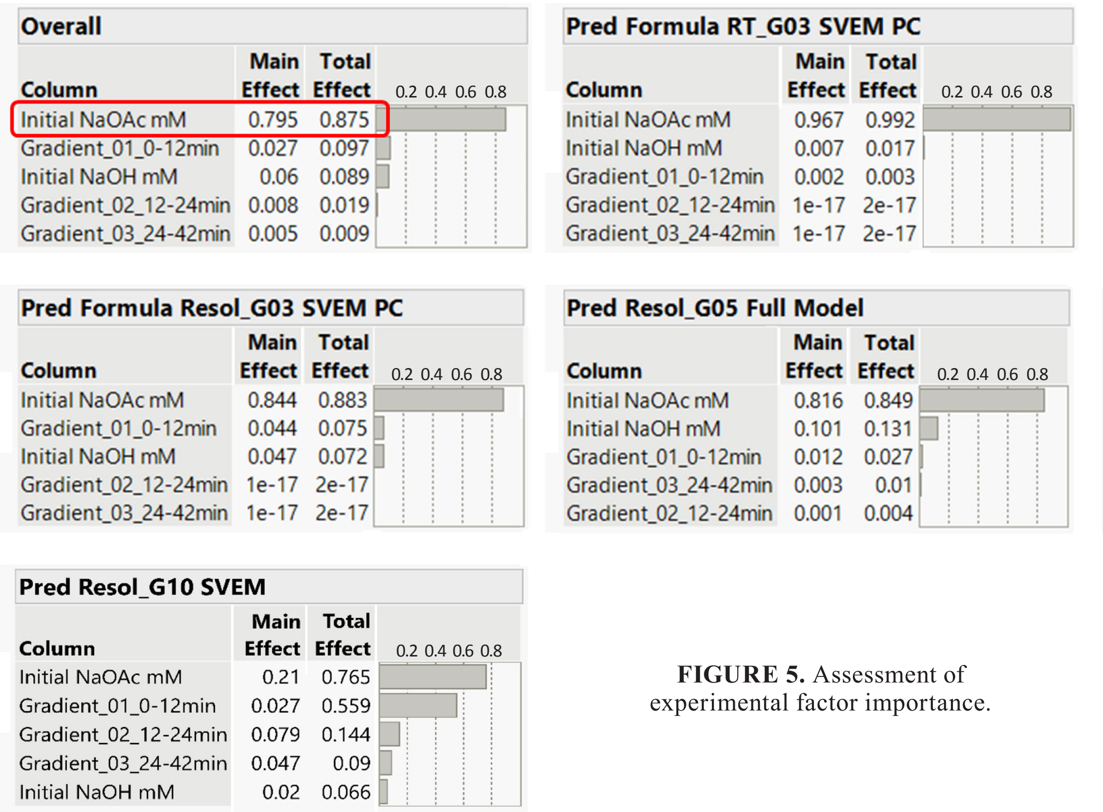
### Variable Importance

QbD strategies require that the importance of the process factors, in terms of long-term process stability, must be assessed (*i.e.,* which factors require tight control). Assessing variable importance is also a prediction problem. In the long term, one is required to predict the impact of the process or method factors (parameters). The problem is complicated by the reality that interactions among the factors account for a large portion of the observed variation in responses. However, using our predictive models, which include interaction terms, makes it possible to assess the overall impact of the factors using those predictive models. The JMP Prediction Profiler contains a variable importance assessment tool based on the work of Sobol.[24] (The considerable mathematical detail will not be discussed in this article.)

**Figure 5** contains the JMP Variable Importance analysis results computed over all four responses on the CCD validation data. This assessment is based upon the four predictive models and therefore captures all the impacts of each factor, both individually and in combination with other factors. Based upon the report in **Figure 5**, the overall impact of Initial NaOAc explains 87.5% of total variation across all four responses. Clearly, Initial NaOAc is an important variable that must be tightly controlled to maintain stable and acceptable performance of the HPAEC-PAD method.

## Conclusions

Implementing quality by design (QbD) principles to the analytical procedure development is no different from those used in the drug development process. The same goals are to identify critical attributes and parameters, define the control strategy, and establish robustness in the method or the process. Certainly, these goals are realized throughout the procedure and the product lifecycles. In this article, we described a glycoprofiling procedure that integrates DSD into a simple, conventional chromatography method, HPAEC-PAD, using the QbD approach. This procedure also uses an inexpensive glucose ladder to achieve two-folds of benefits, serving as a



**Overall**

| Column | Main Effect | Total Effect |
|---|---|---|
| Initial NaOAc mM | 0.795 | 0.875 |
| Gradient_01_0-12min | 0.027 | 0.097 |
| Initial NaOH mM | 0.06 | 0.089 |
| Gradient_02_12-24min | 0.008 | 0.019 |
| Gradient_03_24-42min | 0.005 | 0.009 |

**Pred Formula RT_G03 SVEM PC**

| Column | Main Effect | Total Effect |
|---|---|---|
| Initial NaOAc mM | 0.967 | 0.992 |
| Initial NaOH mM | 0.007 | 0.017 |
| Gradient_01_0-12min | 0.002 | 0.003 |
| Gradient_02_12-24min | 1e-17 | 2e-17 |
| Gradient_03_24-42min | 1e-17 | 2e-17 |

**Pred Formula Resol_G03 SVEM PC**

| Column | Main Effect | Total Effect |
|---|---|---|
| Initial NaOAc mM | 0.844 | 0.883 |
| Gradient_01_0-12min | 0.044 | 0.075 |
| Initial NaOH mM | 0.047 | 0.072 |
| Gradient_02_12-24min | 1e-17 | 2e-17 |
| Gradient_03_24-42min | 1e-17 | 2e-17 |

**Pred Resol_G05 Full Model**

| Column | Main Effect | Total Effect |
|---|---|---|
| Initial NaOAc mM | 0.816 | 0.849 |
| Initial NaOH mM | 0.101 | 0.131 |
| Gradient_01_0-12min | 0.012 | 0.027 |
| Gradient_03_24-42min | 0.003 | 0.01 |
| Gradient_02_12-24min | 0.001 | 0.004 |

**Pred Resol_G10 SVEM**

| Column | Main Effect | Total Effect |
|---|---|---|
| Initial NaOAc mM | 0.21 | 0.765 |
| Gradient_01_0-12min | 0.027 | 0.559 |
| Gradient_02_12-24min | 0.079 | 0.144 |
| Gradient_03_24-42min | 0.047 | 0.09 |
| Initial NaOH mM | 0.02 | 0.066 |

**FIGURE 5.** Assessment of experimental factor importance.

sample surrogate to optimize glycoprofiling and as a system suitability testing sample for the analytical procedure. The SVEM machine learning technique was applied to the DSD data to build predictive models for four chromatographic responses. The results were verified using a traditional experimental design (CCD) in which the same predictive models and conclusions were obtained, demonstrating that the DSD is an efficient alternative to the larger CCD.

## References

**[1]** Zhang P, Woen S, Wang T, Liau B, Zhao S, Chen C, Yang Y, Song Z, Wormald MR, Yu C, Rudd PM. Challenges of glycosylation analysis and control: an integrated approach to producing optimal and consistent therapeutic drugs. *Drug Discov Today*, 2016; 21(5): 740–65. https://doi.org/10.1016/j.drudis.2016.01.006 PMid: 26821133

**[2]** Philippidis A. Top 15 best-selling drugs launched in 2020. *Gen Eng Biotechn*, 2020 Dec 23. https://www.genengnews.com/a-lists/top-15-best-selling-drugs-launched-in-2020/

**[3]** Reusch D, Haberger M, Maier B, Maier M, Kloseck R, Zimmermann B, Hook M, Szabo Z, Tep S, Wegstein J, Alt N, Bulau P, Wuhrer M. Comparison of methods for the analysis of therapeutic immunoglobulin G Fc-glycosylation profiles – part 1: separation-based methods. *mAbs*, 2015; 7(1): 167–79. https://doi.org/10.4161/19420862.2014.986000 PMid: 25524468

**[4]** United States Pharmacopeia (USP). <1084> *Glycoprotein and glycan analysis – general consideration*. 2011.

**[5]** United States Pharmacopeia (USP). <212> *Oligosaccharide analysis – general consideration*. 2015.

**[6]** United States Pharmacopeia (USP). <210> *Monosaccharide analysis – general consideration*. 2016.

**[7]** Grey C, Edebrink P, Krook M, Jacobsson SP. Development of a high performance anion exchange chromatography analysis for mapping of oligosaccharides. *J Chromatogr B*, 2009; 877(20-21): 1827–32. https://doi.org/10.1016/j.jchromb.2009.05.003 PMid: 19482525

**[8]** Jones B, Nachtsheim CJ. A class of three-level designs for definitive screening in the presence of second-order effects. *J Qual Technol*, 2011; 43(1): 1–15. https://doi.org/10.1080/00224065.2011.11917841

**[9]** Erler A, de Mas N, Ramsey P, Henderson G. Efficient biological process characterization by definitive-screening designs: the formaldehyde treatment of a therapeutic protein as a case study. *Biotechnol Lett*, 2013; 35(3): 323–9. https://doi.org/10.1007/s10529-012-1089-y PMid: 23160743

**[10]** Box GEP, Wilson KB. On the experimental attainment of optimum conditions. *J R Stat Soc*, Series B, 1951; 13(1): 1–45. https://www.jstor.org/stable/2983966

**[11]** Cornell JA, Montgomery DC. Interaction models as alternatives to low-order polynomials. *J Qual Technol*, 1996; 28(2): 163–76. https://doi.org/10.1080/00224065.1996.11979657

**[12]** Martin GP *et al*. Stimuli to the revision process. Proposed new United States Pharmacopeia (USP) general chapter: <1220> *The analytical procedure lifecycle*. 2016 Oct. https://www.uspnf.com/sites/default/files/usp_pdf/EN/USPNF/revisions/s201784.pdf

**[13]** Sangshetti JN, Deshpande M, Zaheer Z, Shinde DB, Arote R. Quality by design approach: regulatory need. *Arabian J Chem*, 2017; 10(S2): S3412–25. https://doi.org/10.1016/j.arabjc.2014.01.025

**[14]** Yu YQ. Waters Corporation Application Note. *N-linked glycan characterization and profiling: combining the power of accurate mass, reference glucose units, and UNIFI software for confident glycan assignments*. 2016 Jan. https://www.waters.com/nextgen/is/en/library/application-notes/2016/n-linked-glycan-accurate-mass-reference-glucose-units-glycan-assignments0.html

**[15]** van Leeuwen SS, Kuipers BJH, Dijkhuizen L, Kamerling JP. Comparative structural characterization of 7 commercial galacto-oligosaccharide (GOS) products. *Carbohydr Res*, 2016; 425: 48–58. https://doi.org/10.1016/j.carres.2016.03.006 PMid: 27035911

**[16]** Shmueli G. To explain or to predict? *Statist Sci*, 2010; 25(3): 289–310. https://doi.org/10.1214/10-STS330

**[17]** Efron B. Prediction, estimation, and attribution. *J Am Stat Assoc*, 2019; 115(530): 636–55. https://doi.org/10.1080/01621459.2020.1762613

**[18]** James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*, 6th ed. Springer, New York; 2013. https://www.springer.com/gp/book/9781461471387

**[19]** Breiman L. Heuristics of instability and stabilization in model selection. *Ann Statist*, 1996; 24(6): 2350–83. https://doi.org/10.1214/AOS/1032181158

**[20]** Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. *J Wildlife Manage*, 2003; 67(3): 655–6. https://doi.org/10.2307/3802723

**[21]** Ramsey P, Levin W, Lemkus T, Gotwalt C. *SVEM: a paradigm shift in design and analysis of experiments* (2021-EU-45MP-779). JMP Discovery Summit Europe 2021. https://community.jmp.com/t5/Discovery-Summit-Europe-2021/SVEM-A-Paradigm-Shift-in-Design-and-Analysis-of-Experiments-2021/ta-p/349244

**[22]** Lemkus T, Ramsey P, Gotwalt C, Weese M. Self-validated ensemble models for design of experiments. *Chemometr Intell Lab*, 2021 (unpublished, manuscript submitted).

**[23]** Goos P, Jones B. *Optimal design of experiments: a case study approach*, 1st ed. Wiley & Sons, New York; 2011. https://www.wiley.com/en-us/Optimal-Design-of-Experiments%3A-A-Case-Study-Approach-p-9780470744611

**[24]** Sobol IM. Sensitivity estimates for nonlinear mathematical models. *Math Mod Comp Exp*, 1993; 4: 407–14.

## About the Authors

**Eliza Yeung, PhD***, Associate Director of Process Characterization, R&D Services,
Cytovance Biologics, 800 Research Parkway, Oklahoma City, Oklahoma 73104 USA

Philip Ramsey, PhD, Mathematics and Statistics,
University of New Hampshire, Durham, New Hampshire 03824 USA

***Corresponding Author:***
Email: eyeung@cytovance.com; Website: www.cytovance.com; Phone: 405-319-8310