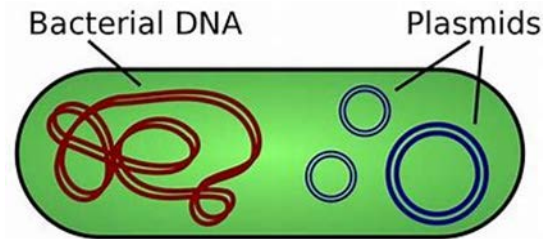


Characterizing Bio-processes With Augmented Full Quadratic Models and Fractionally Weighted Bootstrapping + Autovalidation: a pDNA Case Study



Philip J. Ramsey, Ph.D.
Principal Lecturer in Statistics
University of New Hampshire
Durham, NH, USA 03824
philip.ramsey@unh.edu

Predictum,
Senior Data Scientist and Statistical Consultant
Philip.ramsey@predictum.com
pjrstats@gmail.com



Tiffany D. Rau, Ph.D.
Global Biotechnology Consultant
Rau Consulting, LLC.
Vanderbilt University Advisory Board
tiffany@raubiotech.com

Agenda

- Chemistry, Manufacturing and Control (CMC) Journey
 - Opportunities and challenges in developing a next generation medications
 - Why DoE, Predictive Modeling, and Characterization is critical?
- Characterizing Bio-processes With Augmented Full Quadratic Models
- Fractionally Weighted Bootstrapping + Autovalidation:
- Case Study pDNA Case Study
 - Putting it all together

CMC Pathway – General

Clinical Development Phases



Product and Process Development Stages



QbD Risk Assessments and Milestones



- | | | |
|---|------------------------------------|-------------------------------------|
| 1. Target Product Profile Identified | 4. Initial Process Risk Assessment | 7. Control Strategy Risk Assessment |
| 2. Quality Target Product Profile Defined | 5. Process Risk Assessment 2 | 8. Control Strategy Defined |
| 3. Critical Quality Attribute Risk Assessment | 6. Design Space Defined | 9. Ongoing Improvement and Support |

Example Cell and Gene Therapies

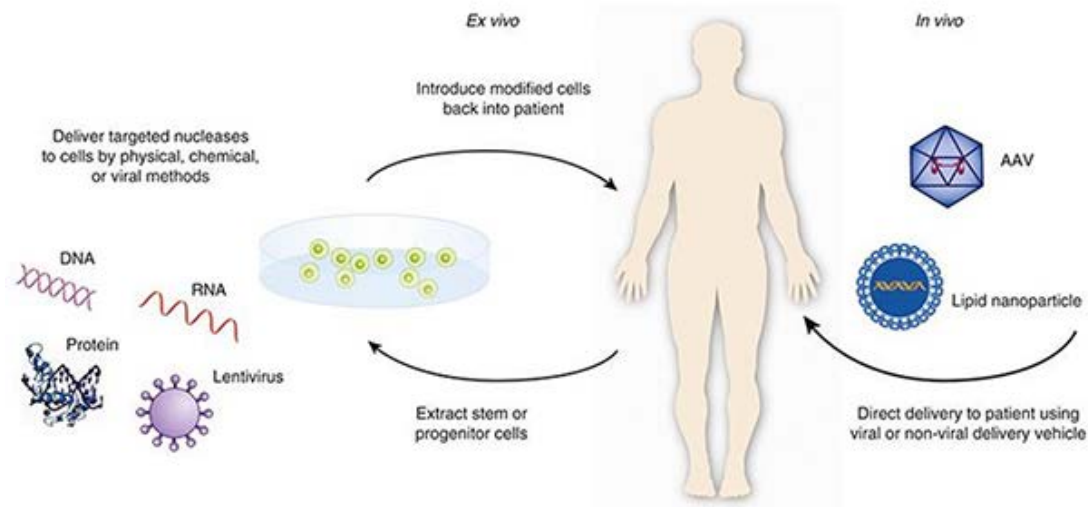
Very Diverse!

Cell Therapies

Cellular immunotherapies
Cancer vaccines
Stem cells & stem cell-derived
Therapeutics for multiple
indications
Regenerative Medicine

Gene Therapies

Plasmid DNA
Viral Vectors
Bacterial vectors
Human
Regenerative Medicine



pDNA and Applications in Human Health

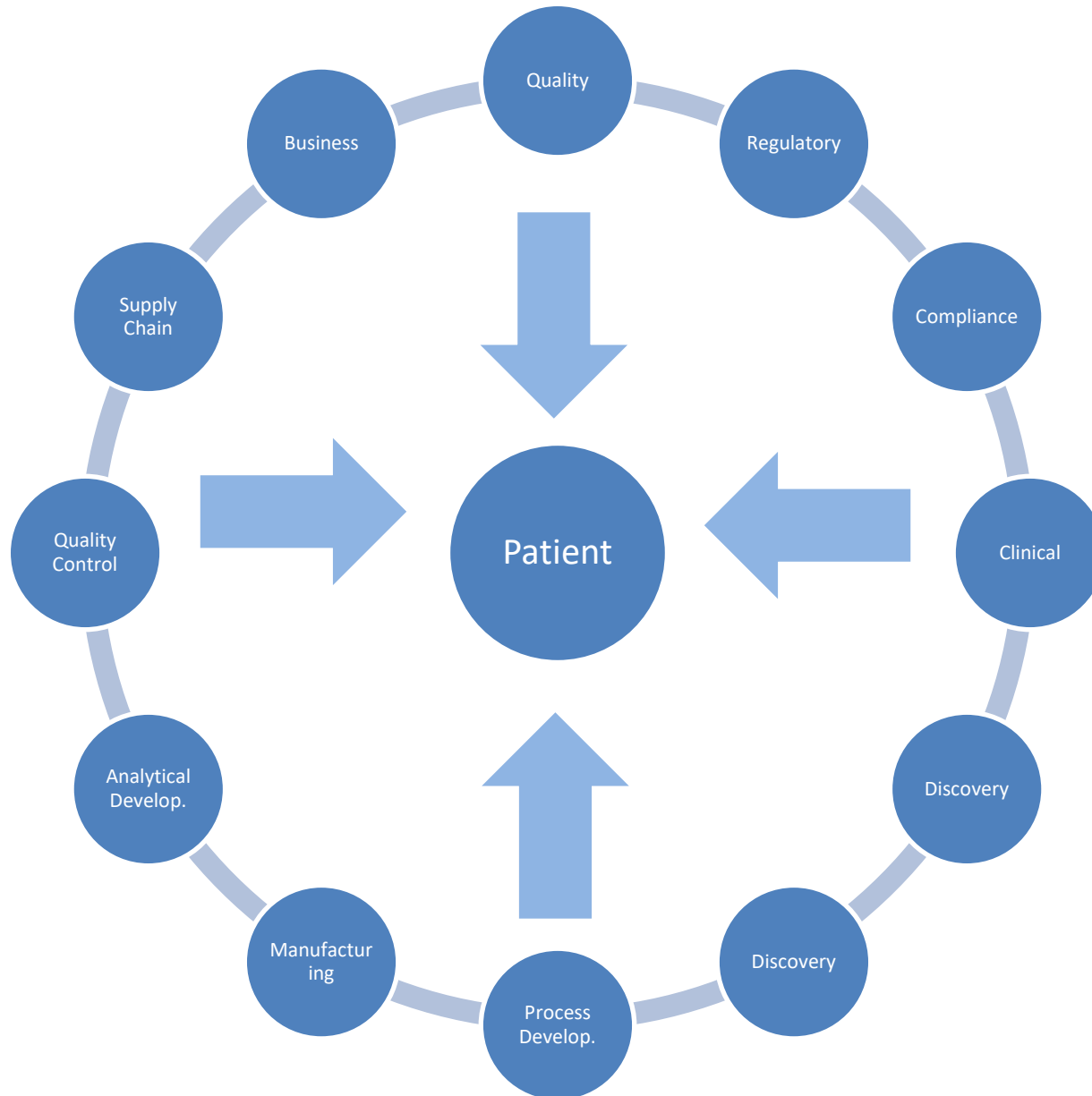
- Potential applications for pDNA
 - Preventive vaccines for viral, bacterial or parasitic diseases;
 - Immunizing agents for the preparation of hyper immune globulin products;
 - Therapeutic vaccines for infectious diseases;
 - Cancer vaccines;
 - Gene replacement application wherein the desired gene product is expressed from the plasmid after administration to the patient.
- As gene therapy and DNA vaccines advance towards regulatory approval, it is critical to produce pDNA in a compliant manner at the appropriate quality and volume levels. Processes need to be well characterized.

pDNA and what is next?

- Cell Therapies, Gene therapies and DNA vaccines are advancing along the CMC pathway towards commercialization
- Critical to produce pDNA in a compliant manner at the appropriate quality and volume levels.
- Processes need to be well characterized.
 - Small Sample sizes.
 - Patients health is often critical – last resort.
- Manufacturing facilities need to be available
 - Capacity is increasing but is limited



Path to Commercialization is Integrated



Augmenting the Full Quadratic Model

Process Development (PD) involves three important activities (especially true for **Quality by Design or QBD**):

- **Creating (Developing)** the Process
- **Characterization** of the entire operating region;
- **Optimization** of the process KPIs including quality attributes.

PD requires the development of valid models which **accurately predict future process performance**.

Box and Wilson (1951) pioneered the **full quadratic model (FQM)** as a basis for predictive process models.

FQM for 2 factors: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2$

FQM approximates the response surface well in the vicinity of an optimum, it is often a poor approximation to a response surface over the process region, so FQM is not suitable for characterization

Augmenting the Full Quadratic Model

Cornell and Montgomery (1996) discuss augmenting the FQM with additional higher order interaction terms such that the model better approximates the entire process region.

- An augmented FQM using their approach adds terms such as

$$Y = FQM + \beta_{112}X_1^2X_2 + \beta_{122}X_1X_2^2 + \beta_{1122}X_1^2X_2^2$$

- A drawback to the approach is that the number of additional terms becomes very large for more than three factors; even CCDs become supersaturated.
- For example with 5 experimental factors there are **20 Linear x Quadratic** interactions to be added to the full FQM.
- Such models however are far more likely to be generalizable in complex biological systems; e.g., bioreactors and fermenters.
- The **FWB+AV** method combined with **Model Averaging** can estimate these augmented FQM models..

Augmenting the Full Quadratic Model

The *FQM* for K factors has the following number of terms

$$N_{FQM} = \frac{(K+1)(K+2)}{2}$$

The number of additional linear by quadratic interaction terms is

$$N_{L*Q} = K(K-1)$$

The number of additional quadratic by quadratic interaction terms is

$$N_{Q*Q} = \frac{K(K-1)}{2}$$

The total number of terms in the complete augmented *FQM* is then

$$N_{Total} = 2K^2 + 1$$

For $K = 5$ the total number of terms including the intercept = 51.

FWB + Autovalidation

Requirement: A training (to fit the model) and a validation (to test the model) data set is important to build accurate predictive models

Challenge: DoE typically does not have enough trials to form a validation set for predictive modeling

Solution: Gotwalt and Ramsey (2017) proposed a method of validation referred to as autovalidation

- Training set can be used for both purposes.
- **How?:** The original data is the test set and a copy of the original data is a validation set,
 - Random gamma weights are applied to both datasets such that the training and validation copy are anti-correlated.
 - Autovalidation is then combined with FWB to generate thousands of iterations of modeling.

FWB + Autovalidation

Fractionally weighted bootstrapping (FWB) randomly assigns new gamma weights to the data over thousands of iterations; all of the data is used on every iteration.

- This has the effect of generating thousands of bootstrap samples of the original training and autovalidation sets.

A predictive model is fit to the data on each iteration and the coefficient estimates and validation error tracked.

- The end result is a table with possibly thousands of coefficient estimates for the model of interest; any term not entering a model on a single iteration is assigned a 0 coefficient value.

A null factor is added to each model as a calibration check.

The FWB table of results provides the user with a set of coefficient estimates and a table containing the proportion of times each possible term entered a model over the FWB runs.

Ensemble Modeling for Prediction

Traditional statistical modeling focuses on a single best model, while machine learning often employs ensembles of models to make predictions or classifications; e.g., Bootstrap Forest or XGBoost.

One ensemble approach is **Model Averaging** where the coefficients in the model are averages of coefficient estimates derived from fitting large numbers of models; e.g., Best Subsets Regression.

The averaging is a form of coefficient regularization that also mitigates any over fitting impacts on prediction.

FWB provides an excellent source of individual coefficient estimates for model averaging, where the model averages can be based on 100s or even 1,000s of estimates from the FWB runs.

Model averaging allows the estimation of a supersaturated model for a design, which is not possible with traditional statistical modeling; the number of parameters $p > N$ the number of runs.

Case Study: Fitting Models to Experimental Data

- **Plasmid DNA Case Study:**
 - Demonstrates the technique of autovalidation, FWB, and Model Averaging.
 - Illustrates how to address the inherently nonlinear and interactive behavior of bioprocessing

Optimize a pDNA Fermentation

Problem

- Develop a reliable Fermentation process for pDNA production
- Current Fermentation strategies do not produce enough product

Challenges

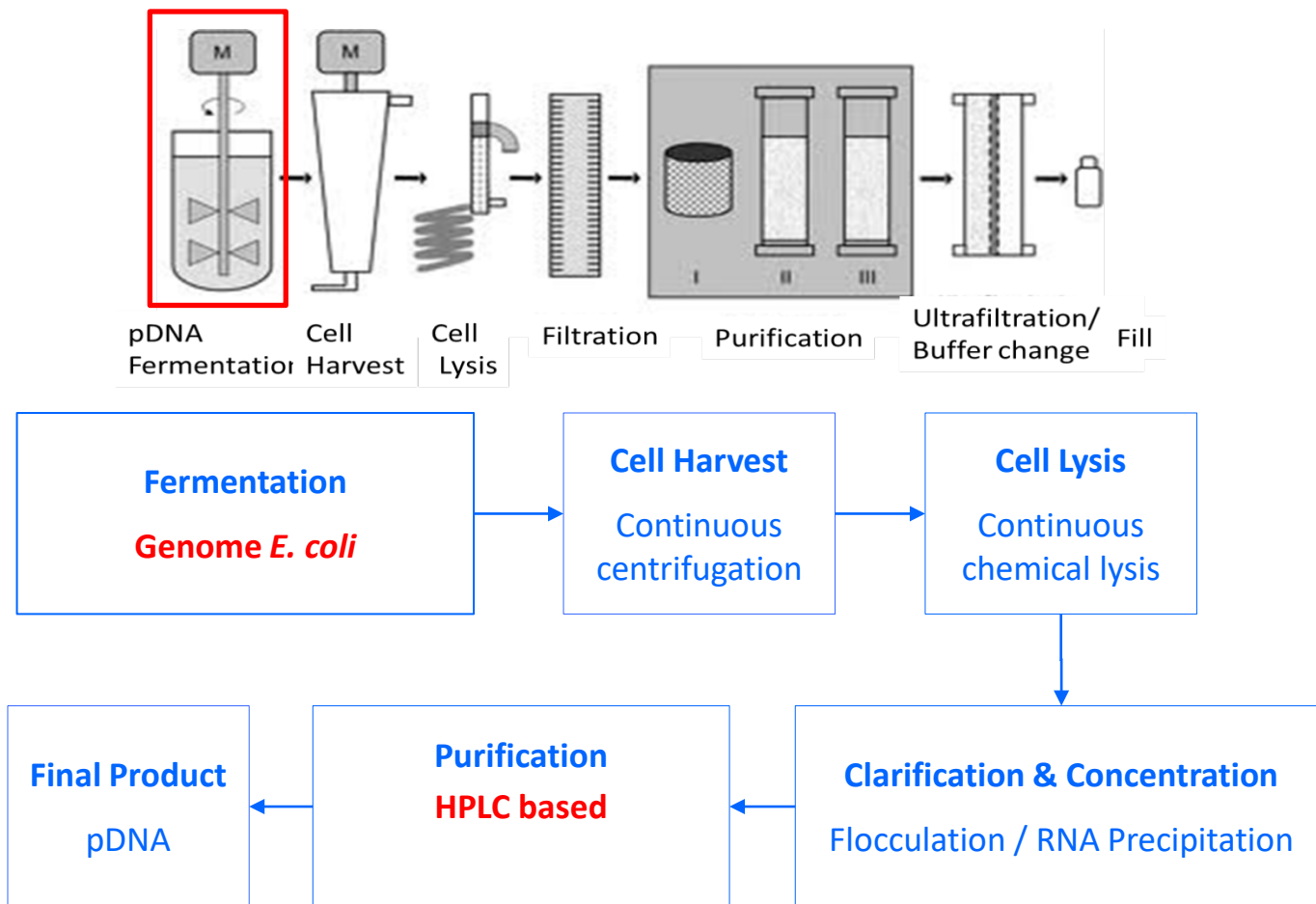
- No current data exists that is useful for optimization and for characterization of the fermentation process
- Need high quality data to produce predictive models for optimization

Solution

- Conduct an efficient Definitive Screening Design experiment (15 runs)
- Use DoE data to build a predictive model for optimization and characterization.
- Predictive model includes interactions and polynomial terms and closely approximates the fermenter performance.

Case Study: Characterizing pDNA Manufacturing

Flow diagram of the pDNA production process. This study focuses on the **Fermentation** step using E-Coli.

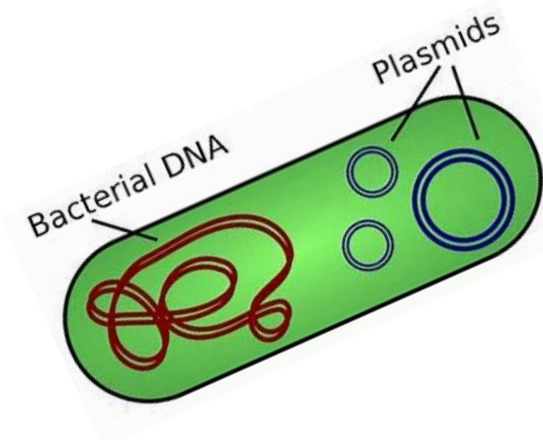


Case Study: Characterizing pDNA Manufacturing

- Using a novel method referred to as fractionally weighted bootstrapping with autovalidation (FWB+AV), a predictive response surface model was fit to the DoE data.
- Experimental design and novel analysis was performed using the JMP Pro version 15 statistical software.
- Model was subsequently used to characterize and optimization.
- JMP Pro Prediction Profiler tool performs the optimization based upon the defined predictive model.
- Profiler may also be used to run large simulation experiments over the entire process region in order to fully characterize performance.

Case Study: Characterizing pDNA Manufacturing

- 5 factor, 15 run Definitive Screening Design (DSD) to characterize and optimize a fermentation process to manufacture pDNA.
- 31 run Central Composite Design (CCD) was done separately for comparison to the DSD results.
 - The CCD acts as a true validation set for comparison to the auto-validation method using the DSD as the training data.
- Experimental Response - **pDNA titer in mg/L.**

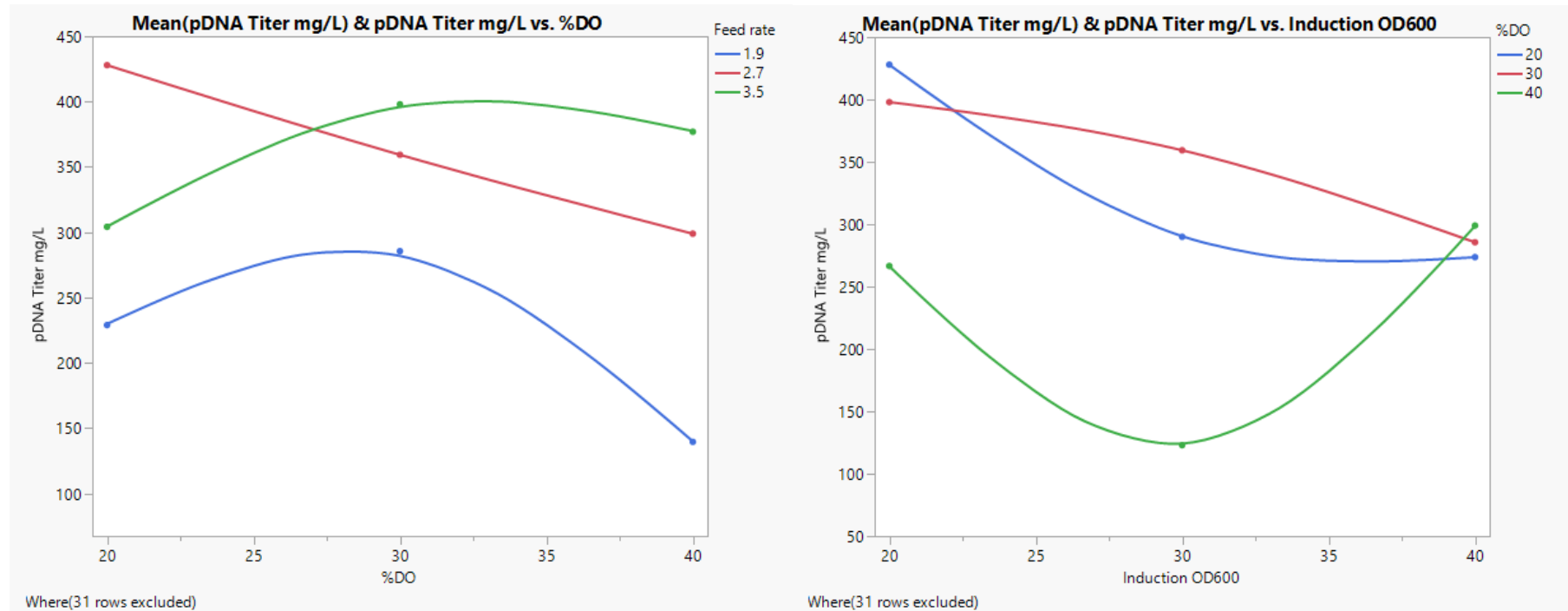


	pH	%DO	Induction Temperature C	Induction OD600	Feed rate, mL/hr	pDNA, mg/L
1	7.0	40	42.5	20	1.9	156.20
2	7.0	20	39.5	40	3.5	487.15
3	7.2	30	39.5	20	3.5	398.00
4	6.8	30	42.5	40	1.9	285.60
5	7.2	20	41.0	40	1.9	229.00
6	6.8	40	41.0	20	3.5	377.00
7	7.2	20	42.5	30	3.5	290.00
8	6.8	40	39.5	30	1.9	123.00
9	7.2	40	42.5	40	2.7	299.00
10	6.8	20	39.5	20	2.7	428.00
11	7.0	30	41.0	30	2.7	327.80
12	7.0	30	41.0	30	2.7	339.74
13	7.0	30	41.0	30	2.7	387.35
14	7.0	30	41.0	30	2.7	393.97
15	7.0	30	41.0	30	2.7	348.08

Case Study: Characterizing pDNA Manufacturing

Interaction plots created in Graph Builder suggests the need for augmented FQM terms.

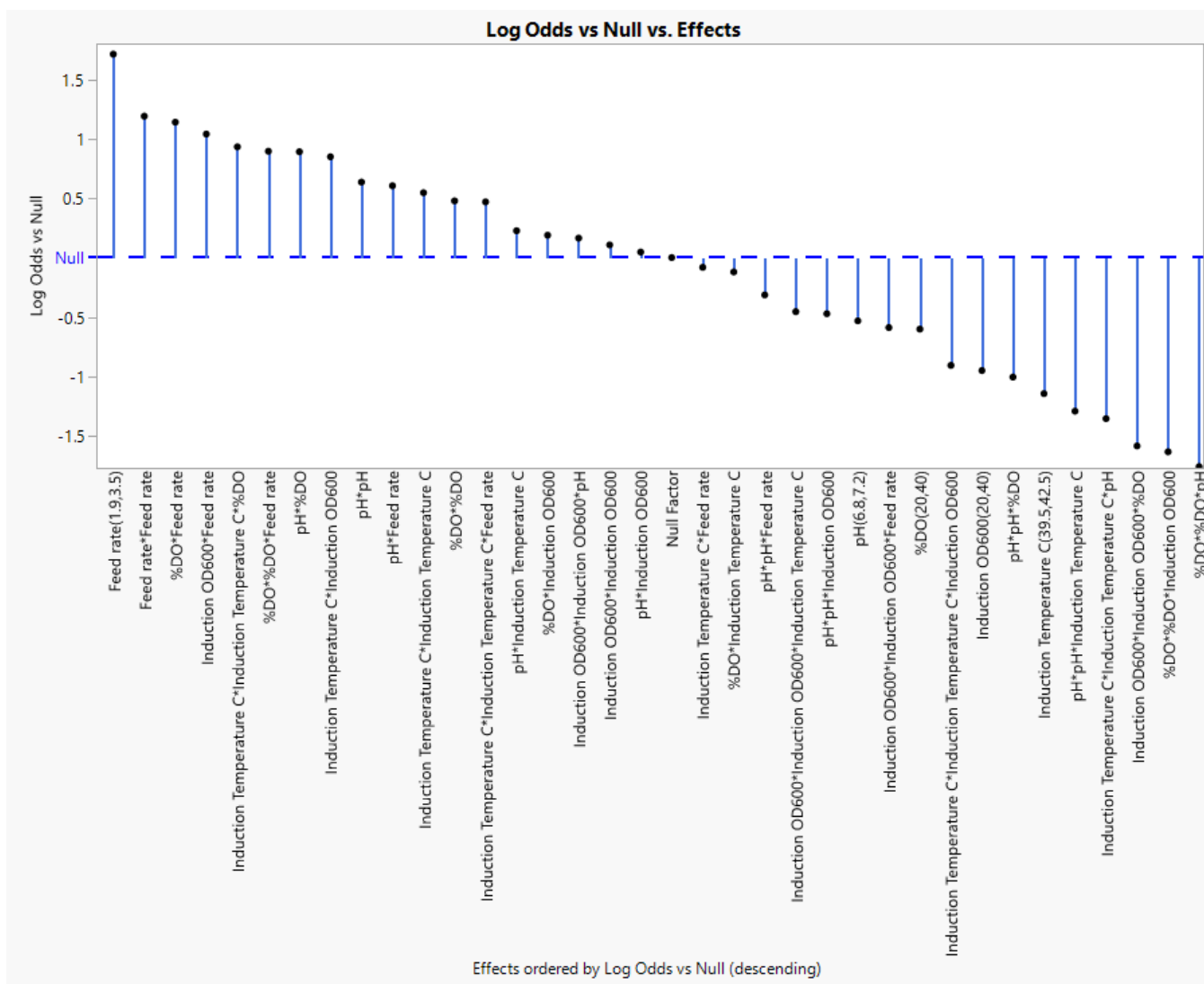
Notice the curvilinear effect of **%DO** differs across **Feed rate** (plot on left) similarly the curvilinear effect of **%DO** differs over **Induction OD600** (plot to the right).



Analyzing the pDNA Case Study with JMP

1. **Add all 20 linear*quadratic interaction terms** to the FQM – 40 candidate predictors total; there would be 51 terms if we included quadratic by quadratic interaction terms.
2. **Use the Best Subset** for model selection – Generalized Regression; Advanced option set largest model size to 5 (depends upon host computer resources).
3. Use the FWB+AV procedure for $N = 2500$ repetitions (JMP Pro Simulate function);
4. **Track the coefficient estimates on each trial** and set the value to 0 if that predictor is not selected in a model;
5. **Track the RASE** (Validation) on each trial and create weights;
6. **Use model averaging to create a prediction equation** where a weighted average is used with $\text{Weights} = 1/RASE^2$. The weights are normalized (0, 1) and worse 5% of models excluded.

Case Study: Characterizing pDNA Manufacturing



Graph shows how often effects entered a model relative to the null factor.

Case Study: Characterizing pDNA Manufacturing

- Screenshot of the full 41 term model fit using FWB+AV and model averaging.
 - Remember, the original DSD had $N = 15$ trials.
- Model was fit using a JMP Pro Addin that performs FWB+AV, Model Averaging.
 - Uses RASE validation weights, and saves the formula to the data table.
- Contact Predictum: Wayne@predictum.com if you are interested in the Addin.

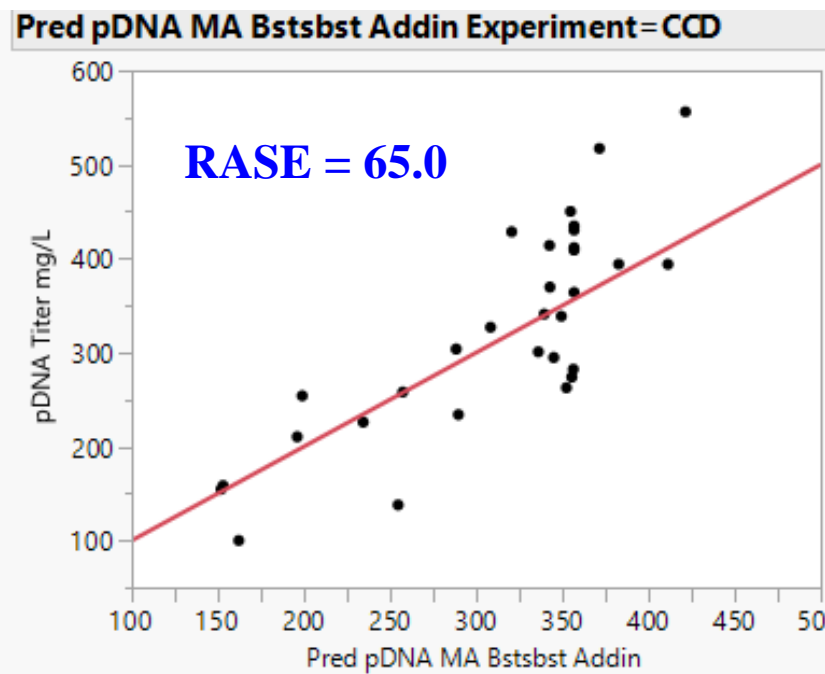
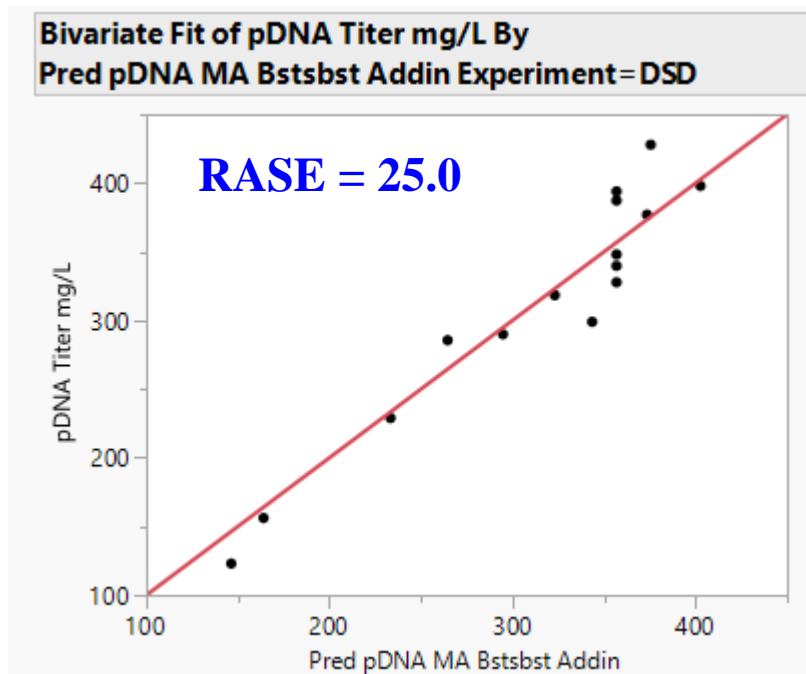
-74487.34492			
+ -0.297655735	• %DO		
+ 30454.681788	• Feed rate		
+ -844.2057653	• Induction OD600		
+ 1917.22625	• Induction Temperature C		
+ 6941.1564734	• pH		
+ -129.8101653	• Feed rate	²	
+ 5.2899690199	• Induction OD600	²	
+ -17.87695455	• Induction Temperature C	²	
+ -775.608523	• pH	²	
+ -4.725793996	• Feed rate		• Induction OD600
+ -1096.894156	• Feed rate		• Induction Temperature C
+ -1941.602944	• Feed rate		• pH
+ 0.0329167175	• Feed rate		• Induction OD600 ²
+ 13.183059476	• Feed rate		• Induction Temperature C ²
+ 137.02317889	• Feed rate		• pH ²
+ 4.1446560177	• Induction OD600		• Induction Temperature C
+ 158.88733581	• Induction OD600		• pH
+ 0.0896450412	• Induction OD600		• Feed rate ²
+ 0.0029509506	• Induction OD600		• Induction Temperature C ²
+ -8.768716199	• Induction OD600		• pH ²
+ 141.45845958	• Induction Temperature C		• pH
+ 0.4422432733	• Induction Temperature C		• Feed rate ²
+ -0.034520492	• Induction Temperature C		• Induction OD600 ²
+ 4.6379279921	• Induction Temperature C		• pH ²
+ -0.365310128	• pH		• Feed rate ²
+ -0.568947009	• pH		• Induction OD600 ²
+ -3.012675762	• pH		• Induction Temperature C ²

Case Study: Characterizing pDNA Manufacturing

Model Averaging used with FWB to fit the full 41 term model on the DSD data.

The average model was then applied to the 31 run CCD completed separately (New lots of Raw Materials including E.coli strain)

Actual by Predicted plots on the DSD training & CCD validation data are shown below







Case Study: Characterizing pDNA Manufacturing

For comparison purposes, the model averaging/FWB process was repeated using the traditional 21 term full quadratic model.

Screenshot of the Model Comparison report shows on both the Training DSD data and Validation CCD data the augmented 41 term model had a lower RASE than the traditional full quadratic model.

Model Comparison

Measures of Fit for pDNA Titer mg/L

Experiment	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
CCD	Pred pDNA MA_ FQM			0.5892	69.958	59.508	31
CCD	Pred pDNA MA Bstbst Addin			0.6420	65.315	50.795	31
DSD	Pred pDNA MA_ FQM			0.7353	43.665	30.814	15
DSD	Pred pDNA MA Bstbst Addin			0.9129	25.049	19.586	15

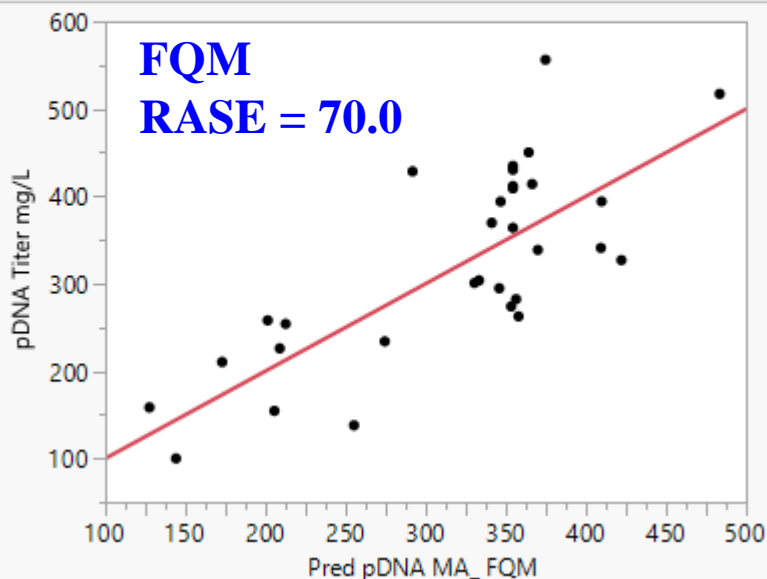
Case Study: Characterizing pDNA Manufacturing

Actual by Predicted plots on the validation CCD data for the FQM and augmented FQM models are shown.

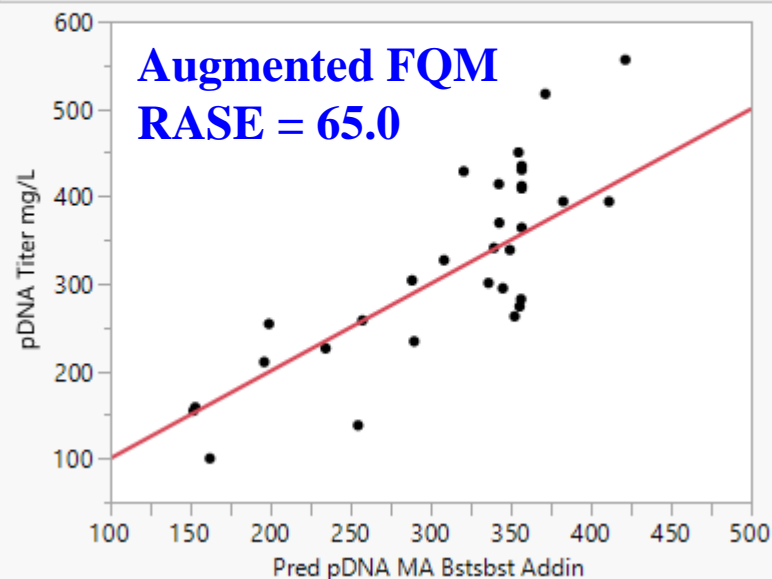
The augmented FQM with $p = 40$ predictors has lower prediction error (RASE) on the CCD validation data than the $p = 21$ term FQM.

Fit Group Experiment= CCD

Bivariate Fit of pDNA Titer mg/L
By Pred pDNA MA_ FQM Experiment= CCD



Bivariate Fit of pDNA Titer mg/L By
Pred pDNA MA Bstsbst Addin Experiment= CCD



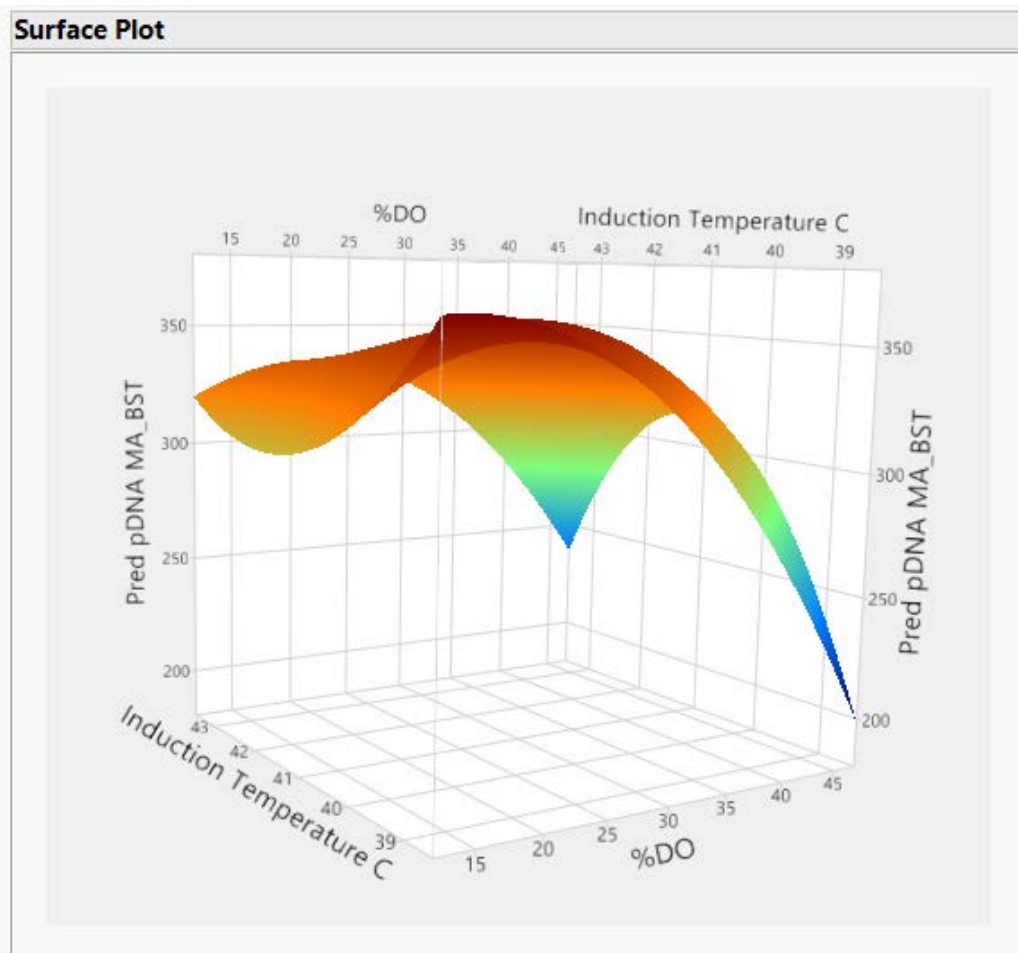
Case Study: Characterizing pDNA Manufacturing

Below is the Profiler display on the CCD (Validation) data with the optimized setting for titer and Variable Importance Report.



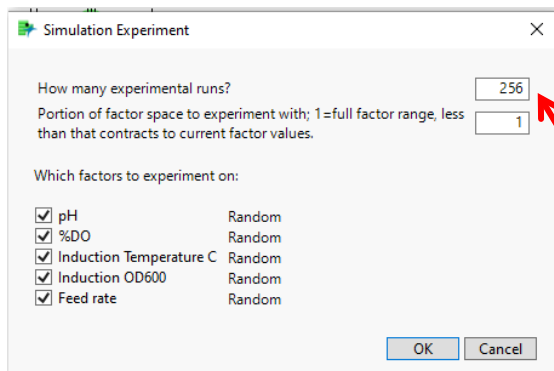
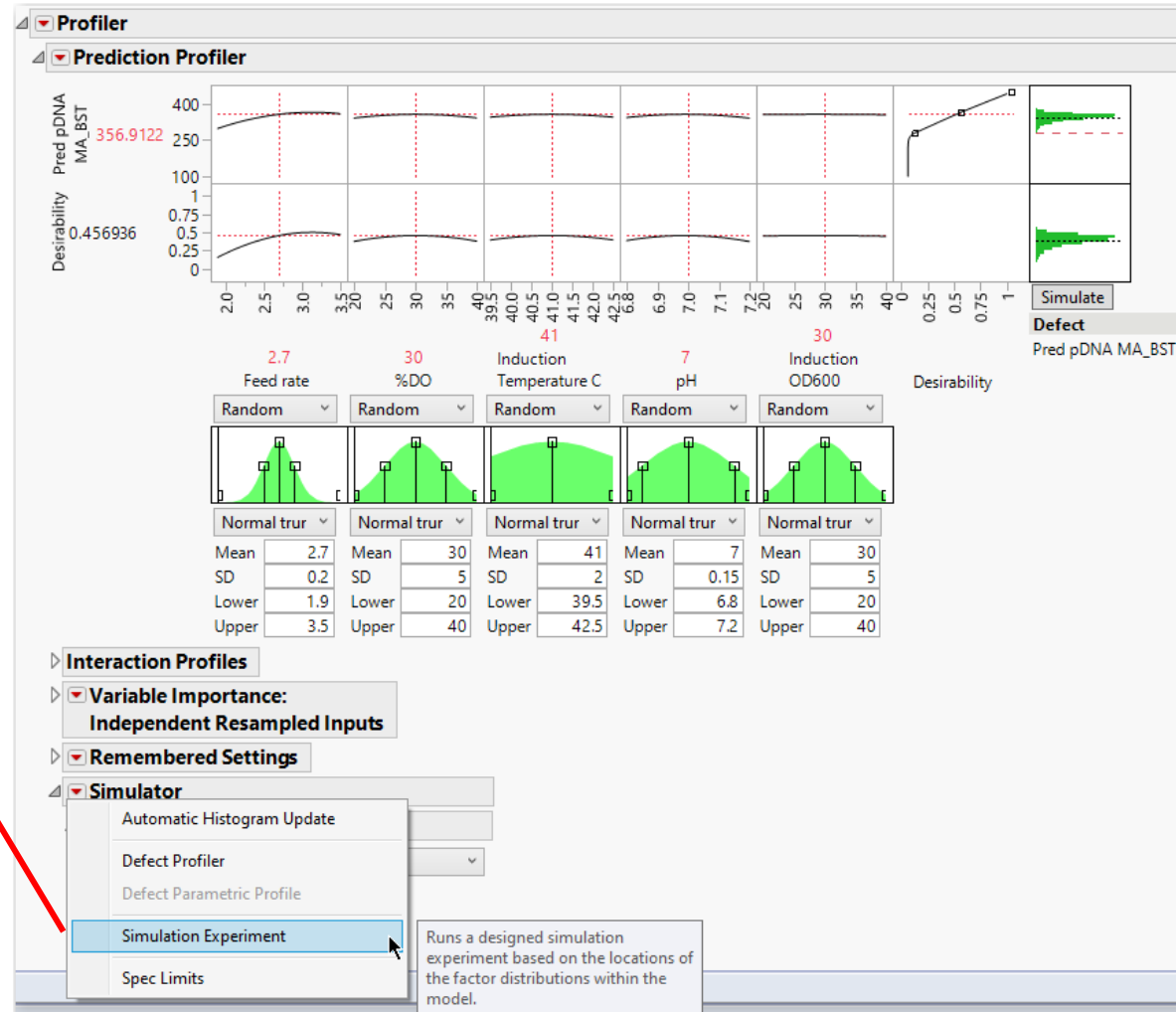
Case Study: Characterizing pDNA Manufacturing

- 3D view of the Titer response surface as a function of Induction Temperature and %DO is shown.
- Illustrates Highly nonlinear relationship between Titer and the experimental factors.



Assessing: Manufacturing Risk due to Variation in Process Factors

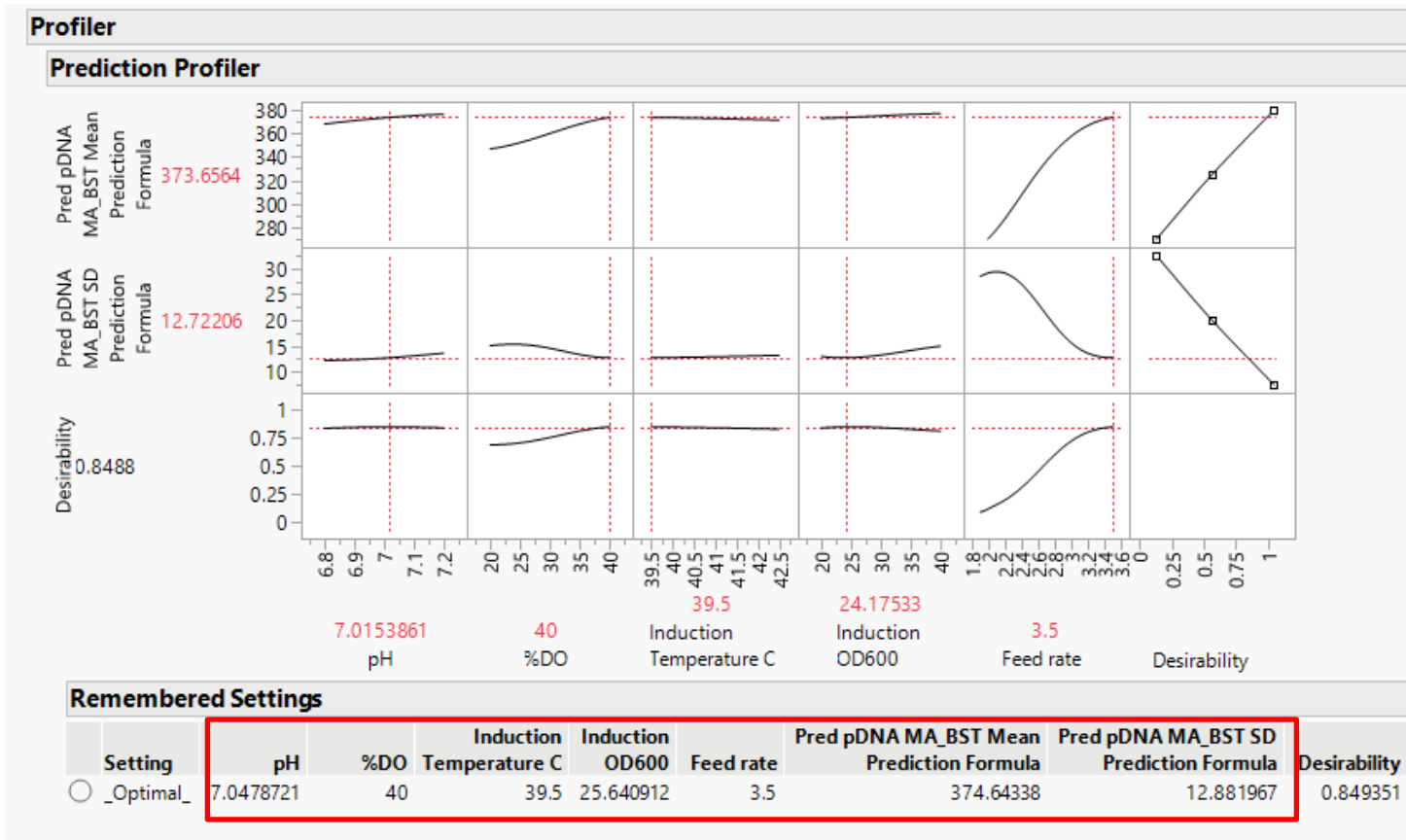
- This can be accomplished using the fitted model and the **Simulator** in the JMP Prediction Profiler.
- Assuming variation in the factors, one can perform a study of process variation.
- Experimental region explored using a **space filling design**.



Case Study: Characterizing pDNA Manufacturing

The simulation experiment provides a mean and standard deviation of the response at each location in the space filling design.

These two responses can be modeled with Gaussian Process models and then an optimization performed to maximize yield and minimize variation in yield. Below are the optimization results.



Executive Summary

- Journey from Discovery to Commercial Pharma/Biotech product is complex
 - System thinking approach is critical
 - Out of the box and holistic approaches are needed to effectively deliver
- Process Design and Development work is inherently about Prediction.
- Fractionally Weighted Bootstrapping combined with autovalidation allows one to build predictive models from designed experiments.
- Biologic systems are highly interactive and nonlinear, the full quadratic model is not sufficient to fully characterize such systems.
 - The interaction models of Cornell and Montgomery are more capable of characterizing biological systems.
 - The linear by quadratic interaction terms are especially important.
- Model averaging combined with FWB provides a way to fit these often large interaction models.
- pDNA case study demonstrates the ability of FWB combined with model averaging, and the interaction model to characterize the entire design space as Quality by Design requires.

BIBLIOGRAPHY

- Bose, A and Chatterjee, S (2018). U-Statistics, M_m -Estimators and Resampling. *Hindustan Book Agency and Springer*.
- Box, G.E.P., Wilson, K.B., “On the Experimental Attainment of Optimum Conditions”, J. of Royal Statistical Society, Ser B, 1951, 13, p1-45.
- Cornell, J. and Montgomery, D. (1996). Interactions as Alternatives to Low-Order Polynomials. *JQT, Vol. 28, No. 2*.
- FDA Cell and Gene Therapy Products - <https://www.fda.gov/vaccines-blood-biologics/cellular-gene-therapy-products/what-gene-therapy> Accessed 08 August 2020
- Gotwalt, C., and Ramsey, P. “Model Validation Strategies for Designed Experiments”, JMP, Discovery Summit 2018 Presentations, Frankfurt, DEU
- Rubin, D. (1981), The Bayesian Bootstrap, *Annals of Statistics*.
- Wu, Y., Boos, D., and Stefanski, L. (2007), Controlling Variable Selection by the Addition of Pseudovariables, *Journal of the American Statistical Association*.
- Jones, B. and Nachtsheim, C. (2011b). A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects. *Journal of Quality Technology*, 43, 1, 1 – 15.
- Smucker, B., Edwards, D, & Weese, M. (2020) Response Surface Models: To Reduce or Not To Reduce. *Journal of Quality Technology*.

JMP Pro Add-in for FWB+AV

Michael D. Anderson of JMP has developed an excellent Add-In to facilitate fractionally weighted bootstrapping with autovalidation.

Highly recommend you download and install the Add-In if you wish to perform the analysis in JMP Pro. See the link below.

<https://community.jmp.com/t5/JMP-Add-Ins/Add-in-To-Support-Auto-Validation-Workflow/ta-p/189991>