Characterization of a pDNA Biomanufacturing Fermentation Process Using Definitive Screening Designs and the JMP® 10 Software

JMP Discovery Summit Conference 2012, September 10 -14, Cary, NC

Dogan Ornek, Ph.D
Senior Scientist
Fermentation Development
Lonza Biologics Inc.
97 South Street
Hopkinton, MA 01748
dogan.ornek@lonza.com

Philip J. Ramsey, Ph.D.
North Haven Group &
University of New Hampshire
Durham, NH 03824
pjramsey@cisunix.unh.edu
pjrstats@aol.com

Talk Outline

- Introduction
- Fermentation Technical Details
- Lonza's pDNA Process Description
- Definitive Screening Designs
- The Fermentation Experiments
- Effect Sparsity Principle
- Model Selection Strategies
- pDNA Model Selection
- Comparing a DSD and Augmented Fractional Factorial Design
- Assessing pDNA Process Capability
- Summary and Conclusions
- References

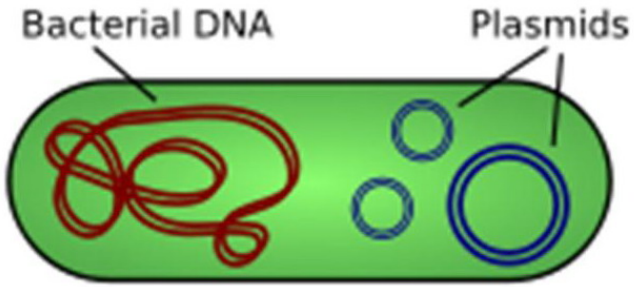
Introduction

Naked plasmid DNA (pDNA) is a circular, nonchromosomal, self-replicating DNA molecule carrying a few useful, but nonessential genes.

e.g. naturally occurring plasmids can encode factors that protect cells from antibiotics or harmful chemicals.

Plasmids are easily moved in and out of the cells and are often used for genetic engineering.

After genes of the protein of interest are added to plasmids, they can be integrated into other genomes, such as plants, protists, and mammals; thereby encoding the protein of interest in other organisms.



Introduction

pDNA is used in gene therapy and vaccine studies. Potential applications are

- 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 approval by U.S. FDA, it is critical to produce high quantity and quality plasmids, and create a well characterized pDNA process

The Fermentation Technical Details

pDNA expressing for therapeutic proteins are transferred and produced in *Escherichia coli (E. coli)*.

The reduced genome *E. coli* host, **MDS42*recA*** (Scarab Genomics, LLC) was used to propagate **pUC19** based pDNA in high cell density fermentation using **ECPM1** based medium (2).

pUC19 plasmids are high copy number *E. coli* plasmids containing portions of the plasmids pBR322 and M13amp19.

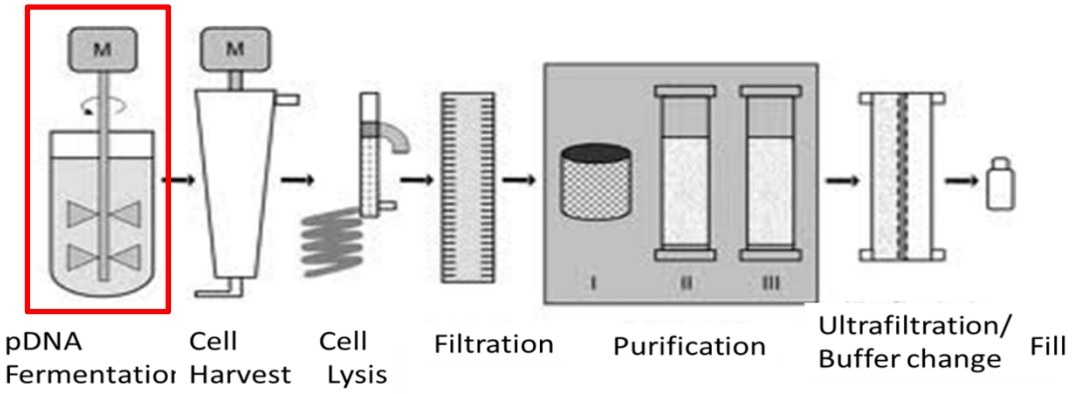
They contain the pMB1 origin of replication from pBR322, but lack the *rop* gene and carry a point mutation in the RNAII transcript.

These together result in a temperature-dependent copy number of 75 per cell at 37^{o}C and >200 per cell at 42^{o}C.

Depending on plasmid size, its production is in the range of 0.5 – 2 g/L in *E.coli* fermentation.

Lonza's pDNA Process Description

Flow diagram of the pDNA production process. This study focuses on the **Fermentation** step.



Fermentation

Clean Genome *E. coli*

Unique Fed-batch HCD

Cell Harvest

Continuous centrifugation

Cell Lysis

Continuous chemical lysis

Clarification & Concentration

Flocculation / RNA Precipitation

Unique Filtration Train

Purification

2 column HIC based

(Lonza Proprietary)

Final Product

ccc-pDNA

Lonza's pDNA Process Description

Upstream and Downstream includes 6 steps

- Fermentation for biomass (*E.coli*) and pDNA production
- Centrifugation for biomass harvest
- Lysis for pDNA extraction from biomass
- Clarification and concentration for pDNA separation from biomass debris (lysed *E.coli*)
- Purification for pDNA separation from other protein, lipids, genomic DNA and endotoxin
- Ultrafiltration/Diafilration for removing salts or other microsolutes from purified pDNA solution (final product)

Lonza's pDNA Process Description

Strain	***E.coli* MDS42*recA***
Plasmid	**pVAX1/*lacZ***
Plasmid Size	**6Kb**

Fermentation Input parameters

- pH: 6.8, 7.0, 7.2
- %DO: 20, 30, 40
- Induction temperature: 39.5, 41, 42.5
- Induction OD_{600}: 20, 30, 40
- Feed rate (mL/h): 1.9, 2.7,3.5

Fermentation output parameters

- pDNA titer
- Optical density
- Wet cell weight

Inoculum [250 mL shake flask]	
Temp.:	**30 ± 1 °C**
OD_{600}:	**6±4**
rpm:	**240**
initial pH:	**7.0 (20 °C)**
Process time:	16 ± 1 h

Batch/Fed Batch Fermentation [0.6 L Fermentor]	
Growth Temp.:	**30 ± 1 °C**
Initial OD_{600}:	**0.03**
pH:	**6.8 - 7.2**
DO:	**20 - 40%**
rpm:	**1200**
Air Flow:	**1 vvm**
Feed start OD	**18±3**
Feed rate	**1.9 - 3.5 mL/h**
Induction OD	**20 - 40**
Induction Temp	**39.5 - 42.5 C**
Process time:	25 ± 0.5 h

Harvest/Storage Bucket centrifuge	
Temp.:	**4-8 °C**
G-Force	**15,000xg**
Time	**30 min**
Cell paste storage	**-80°C**
Process time:	1 h

Characterizing Bio-processes

As stated earlier, characterization and optimization of a pDNA manufacturing process is critical to the FDA approval and economic viability of new therapeutics based on pDNA.

Traditionally, characterization of bio-processes have been performed via a two-stage experimental strategy (FDA, 2011).

1. A Resolution III or IV screening design is performed (often a Plackett-Burman) to identify critical process factors;
2. A response surface design is performed (usually a Box-Behnken) using the critical factors identified with the screening experiment.

The two-stage process is time consuming, costly, and not always effective in identifying the important experimental effects.

Basically, this approach is inefficient in terms of increased R&D cost and longer lead times to approval and commercialization.

Definitive Screening Designs

Recently a new class of screening designs have been developed by Jones and Nachtsheim (2011a, 2011b) and the authors refer to them as **Definitive Screening Designs** (DSD).

The designs have subsequently been enhanced by Xiao (2012).

These designs offer a significant improvement over popular screening designs in couple of ways:

It is possible to estimate main effects, some two-factor interactions and some quadratic effects in a single experiment.

The designs are efficient in terms of the number runs required and at the same time reduce the overall amount of partial aliasing that can occur among potential experimental effects.

Because of these advantages, DSDs are ideal for characterization of bio-processes.

Definitive Screening Designs

The following are some of the advantages of DSDs:

- For k factors the minimum number of runs is 2k+1 for k even and 2k+3 for k odd is recommended.
- All factors are performed at 3 levels so quadratic effects can be estimated.
- Main effects are orthogonal and free of any aliasing.
- No quadratic effects or two-factor interaction effects are completely aliased, so it is possible to estimate both types of effects in a single experiment.
- For $k \geq 6$, the design can estimate a full quadratic model in no more than three factors.

A **full quadratic model** refers to a model containing all main effects, all quadratic effects, and all two-factor interaction effects.

Definitive Screening Designs

Below is an example of a DSD for k = 6.

7/0 Cols	A	B	C	D	E	F
1	0	1	1	1	1	1
2	0	-1	-1	-1	-1	-1
3	1	0	-1	1	1	-1
4	-1	0	1	-1	-1	1
5	1	-1	0	-1	1	1
6	-1	1	0	1	-1	-1
7	1	1	-1	0	-1	1
8	-1	-1	1	0	1	-1
9	1	1	1	-1	0	-1
10	-1	-1	-1	1	0	1
11	1	-1	1	1	-1	0
12	-1	1	-1	-1	1	0
13	0	0	0	0	0	0

The center point (row 13) is added to estimate the intercept of the statistical model fit to the experimental data.

It is recommended that additional center points be added for replication.

The Fermentation Experiments

The talk focuses on an experiment to optimize the pDNA yield of the Fermentation step in a bio-process (see slide 8 for more technical details).

For the experiment k = 5 factors were identified:

1. **pH** (6.8, 7.2) = fermentation solution pH;
2. **Dissolved Oxygen** (%DO) (20%, 40%);
3. **Induction Temperature** (39.5 C, 42.5 C) = Temperature at which the pDNA production is induced in the E. Coli cells.
4. **Induction OD_{600}** (20, 40) = biomass at which the induction is initiated as measured by optical density at 600 nm.
5. **Feed Rate** (1.9, 3.5 mL/hr) = feed rate of a growth media containing 50% glycerol added to the fermentation solution when induction is initiated.

The Fermentation Experiments

Substantial control issues occurred with %DO in the augmented fractional factorial design, possibly masking a %DO effect.

%DO is initially set to 100% and as the biomass grows the %DO gradually decreases to the experimental set point or level.

Then, using agitation and the addition of oxygen, %DO is stabilized in the region of the experimental level.

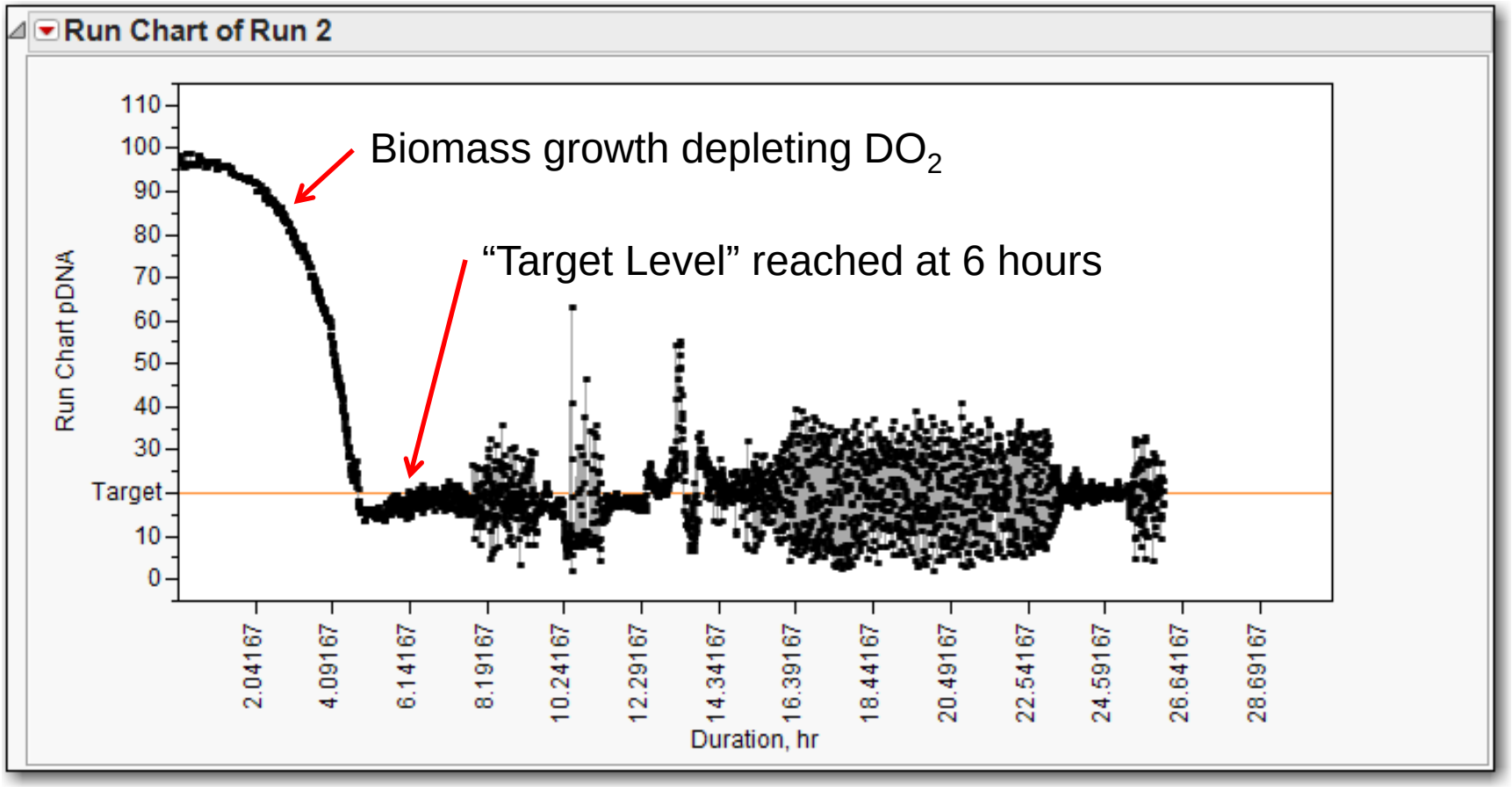
In practice %DO is dynamic over time and hard to control; large deviations and excursions from the target level often occur.

The target %DO level is reached in a range of 3 to 6 hours depending upon the design run.

For each experimental run a %DO profile is stored for later evaluation – readings are taken every 10 seconds.

The Fermentation Experiments

Below is a %DO profile for run 2 of the fractional factorial design. The target %DO level is 20%, one can see the very large deviation about the target after 6 hours; **the range of %DO was 61.1**.



The Fermentation Experiments

The two goals of the experiment were to characterize the fermentation step and to maximize the yield of pDNA.

Keep in mind that the goal was not necessarily to maximize the mass of the E. Coli community, but rather the goal is to maximize the yield of pDNA produced by the E. Coli.

It is possible to substantially increase the mass of a microbial community without maximizing pDNA production.

The three responses of interest are:

1. **Yield** of pDNA titer measured in units of mg/L;
2. **OD600** = measure of biomass by optical density at 600 nm;
3. **WCW** = wet cell weight in units of g/L.

The latter two are a measure of the mass of the microbial community, while pDNA yield is the most important response.

The Fermentation Experiments

Given DSD experiments are new, it was decided to simultaneously perform a separate experiment utilizing a traditional fractional factorial design with the same factors and levels as the DSD.

The goal was to make a direct comparison of the DSD results to the results from a more traditional screening design.

The design selected was a 2^{5-1} resolution V fractional factorial; smaller resolution screening designs are commonly used in bio-process characterization.

The fractional factorial design had 16 factorial runs plus 3 center points for a total of 19 runs.

Subsequently, the 2^{5-1} design had to be **augmented with axial runs** to estimate nonlinear effects detected during analysis, which increased the total number of experimental runs to **31**.

The DSD had a total of **15** runs including 4 center points.

The Fermentation Experiments

Below is the augmented 2^{5-1} design. Note the axial points are set about 7% beyond the upper and lower levels used in the DSD.

14/0 Cols	Block	pH	%DO	Median %DO	Induction Temperature C	Feed Rate, mL/hr	Induction OD600	pDNA, mg/L
1	1	7.2	40	38.8	39.50	3.50	20	581.36
2	1	6.8	20	20.2	39.50	3.50	20	519.80
3	1	6.8	40	39.7	39.50	1.90	20	115.40
4	1	6.8	40	38.2	39.50	3.50	40	407.22
5	1	7.2	40	40.3	39.50	1.90	40	56.18
6	1	7.2	20	20.1	42.50	3.50	20	260.82
7	1	6.8	20	20.0	39.50	1.90	40	94.95
8	1	7.2	20	20.2	39.50	1.90	20	215.03
9	1	6.8	40	40.0	42.50	1.90	40	211.00
10	1	7.0	30	29.9	41.00	2.70	30	321.00
11	1	7.0	30	29.7	41.00	2.70	30	387.35
12	1	6.8	40	39.3	42.50	3.50	20	231.00
13	1	7.2	40	38.8	42.50	3.50	40	351.00
14	1	7.2	20	19.9	39.50	3.50	40	284.00
15	1	6.8	20	19.6	42.50	3.50	40	298.00
16	1	6.8	20	20.0	42.50	1.90	20	191.00
17	1	7.2	20	20.0	42.50	1.90	40	183.02
18	1	7.0	30	29.9	41.00	2.70	30	368.74
19	1	7.2	40	40.5	42.50	1.90	20	111.46
20	2	7.0	30	29.9	41.00	2.70	30	347.74
21	2	7.0	30	29.7	41.00	2.70	30	322.01
22	2	7.3	30	29.8	41.00	2.70	30	213.88
23	2	6.7	30	29.9	41.00	2.70	30	251.68
24	2	7.0	43	42.6	41.00	2.70	30	341.54
25	2	7.0	17	16.9	41.00	2.70	30	327.02
26	2	7.0	30	30.1	42.95	2.70	30	282.70
27	2	7.0	30	29.8	39.05	2.70	30	207.76
28	2	7.0	30	30.0	41.00	3.74	30	307.11
29	2	7.0	30	30.3	41.00	1.66	30	123.39
30	2	7.0	30	30.2	41.00	2.70	43	175.64
31	2	7.0	30	30.0	41.00	2.70	17	195.29

The Fermentation Experiments

Below is the 15 run DSD experiment.

Overall, the %DO control was tighter in the DSD experiment than in the augmented fractional factorial experiment.

18/0 Cols	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

Effect Sparsity

Before proceeding with the analysis of experimental results we need to make a quick diversion to discuss the concept of **Effect Sparsity**.

Effect Sparsity can be thought of as the better-known **Pareto Principle** applied to design of experiments.

For any given experiment the number of active effects is likely to be only some small subset, typically 20% to 30%, of the total number of potential effects.

Over decades of experimentation with physical systems at the macro level, the Effect Sparsity Principle has been well documented.

Without the Effect Sparsity Principle screening designs and to some degree design of experiments in general have little chance of successfully characterizing a physical system.

For a good discussion of Effect Sparsity see Goos and Jones (2011).

Effect Sparsity

In analyzing DSDs one can think of Effect Sparsity in two ways;

- **Absolute Sparsity** = the number of active effects is about 20% to 30% of the total number of possible effects.
- **Relative Sparsity** = the number of active effects is ≤ 50% (more or less) of the number of unique runs in the experiment.

If the number of active effects appears to exceed the 50% level by much, then it may become necessary to augment the DSD.

Although Effect Sparsity is established, it has some weaknesses:

- It lacks a true operational definition;
- It is largely anecdotal with no theoretical underpinning (perhaps the Buckingham Pi Theorem?).

Effect Sparsity is an area of statistics in need of serious research.

Effect Sparsity

A related problem with Effect Sparsity is the lack of a clear operational definition as to what are active or important factors.

The Effect Sparsity Principle does not imply that only 20% to 30% of the potential effects will have **significantly small p-values**.

Rather the Principle implies that only a **small subset of the potential effects are sufficient** to describe the behavior of the response.

Remember, statistically significant effects are not necessarily important effects and vise verse.

In experiments with relatively small amounts of experimental error (small RMSE) a large number of effects appear significant even with relatively small impacts on the response.

With relatively large experimental error few or no effects appear to be significant even with apparent large impacts on the response.

For purposes of this discussion we use the term **important effects**.

Model Selection

Analyzing the DSD experimental results constitutes a special case of the **supersaturated design** problem.

For the case of a DSD with k factors the largest possible model to be considered is the full quadratic model.

As an example, if k = 6 there are potentially 6 main effects, 6 quadratic effects, 15 two-factor interaction effects, and an intercept for a total of 28 model terms to be estimated.

Given the DSD with k = 6 has only 13 unique settings of the experimental factors, we require the Effect Sparsity Principle to hold in order to estimate a useful predictive model for the response.

Since we expect only a small subset of the potential effects to be important, how does one proceed to find a model capturing these important effects?

There exists a substantial model selection problem to be solved.

Model Selection

In analyzing the DSD results we use the following set of principles:

- **Hierarchy** = lower order effects are more likely to be important than higher order effects; e.g., main effects are important more often than two-factor interactions.
- **Heredity** = if a higher order effect is important, then the lower order parent terms of that effect are also important; e.g., if a two-factor interaction is important, then so are the two main effects involved in the interaction.
- **Thee model** does not exist = only in simulations are there correct models; in practice a subset of models will perform well (all models are wrong, some are useful, G.E.P. Box).
- **Subject matter expertise is eminent** in model selection.
- **Parsimony** = the simplest model that adequately predicts the response is the best model.

Model Selection

Given we cannot estimate all of the potential effects with a DSD we must use some type of model selection technique to try and determine the important effects.

The **JMP 10** software has a powerful **Stepwise** regression platform that can easily facilitate the model selection task with a DSD.

Using the Stepwise platform we analyzed the DSD and determined a small set of potential best models (we will also use Stepwise to analyze the augmented fractional factorial design).

In selecting models we generally have two competing issues:

- **Under fitting** the model resulting in biased or inaccurate prediction;
- **Over fitting** the model resulting in inflated prediction error.

The goal is to find the smallest model that adequately predicts the response, this model balances under and over fitting.

Model Selection

Three widely accepted measures of fit for a model are:

- **AICc** = bias corrected Akaike Information Criterion;
- **BIC** = Bayesian Information Criterion;
- **Cp** = Mallow's Cp statistics, where **p** is the model size.

We omit the mathematical details on AICc and BIC, see Burnham and Anderson (2004) or JMP 10 Help for dicussions.

Cp is falling into disuse, but it is included for comparison purposes.

Each statistic punishes under and over fitting, but in a different way so that they may not agree on the best model(s) – they often do not.

There is not agreement in the statistical community as to whether AICc or BIC criterion is preferred; it depends upon the application.

For both the AICc and BIC **smaller values indicate better predictive models**; for Cp the value should be in the vicinity of p.

Model Selection

At present we are experiencing a proliferation of ever more complex model selection algorithms.

We are being overwhelmed with algorithms and starving for direction in terms of which ones to use.

Are the researchers even asking the right questions?

For purposes of this talk the following set of algorithms were examined:

- **Forward Selection** (JMP 10);
- **All Possible Models** (JMP 10);
- **The Lasso** (SAS GLMSelect),
  the **Danzig Selector** is equivalent (Bickel, 2008);
- **Least Angle Regression** (LAR) (SAS GLMSelect);
- **Reversible Jump Models** (Winbugs).

Model Selection

It is beyond the scope of the talk to discuss the pros and cons of these algorithms.

The **Reversible Jump Model** algorithm consistently selected badly under fit models and the technique was abandoned.

The LASSO and LARs are not currently available in JMP 10, are difficult to understand for engineers and scientists, and did not perform any better and perhaps worse than **All Possible Models**.

These findings may not apply to more complex modeling tasks with higher dimensioned data than that found in DSD experiments.

Given we need methods easily understood and used by engineers and scientists from diverse disciplines the simpler algorithms are preferred.

Therefore, **All Possible Models** and **Forward Selection** were used in concert for model selection.

Model Selection

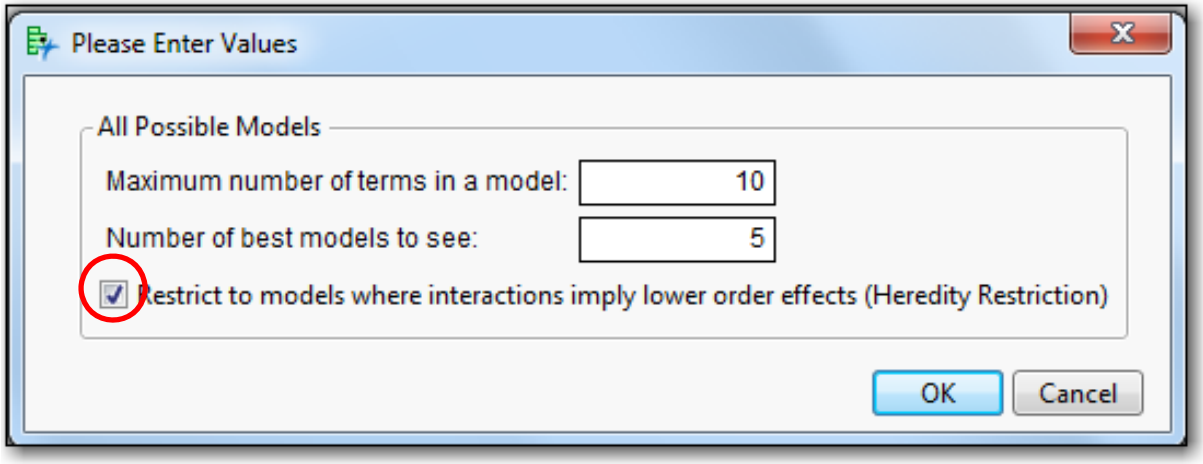
The model selection heuristic used is as follows:

1. Specify a full quadratic model and use the **Stepwise** platform.
2. Use **Forward Selection** with the **Minimum BIC or AICc** criterion to see how many effects might be important; these models are also candidates for final models.
3. Use **All Possible Models** with the maximum model size for a DSD set to a maximum value for which AICc can be estimated.
4. Sort the All Possible Models report in ascending order by AICc (or BIC) and make the report into a data table.
5. Create an **Overlay plot of AICc and BIC** (and **Cp** if it is used) by model size (Number).
6. Interpret the graph to select a candidate model size or sizes.
7. Examine this set of models and select one or more models for further investigation.

DSD: Model Selection for pDNA

For the DSD experiment we use **All Possible Models** with the maximum model size set to 10 and specify that only 5 models for each model size be displayed in the output.

Also select the option to impose the **Heredity Restriction**.

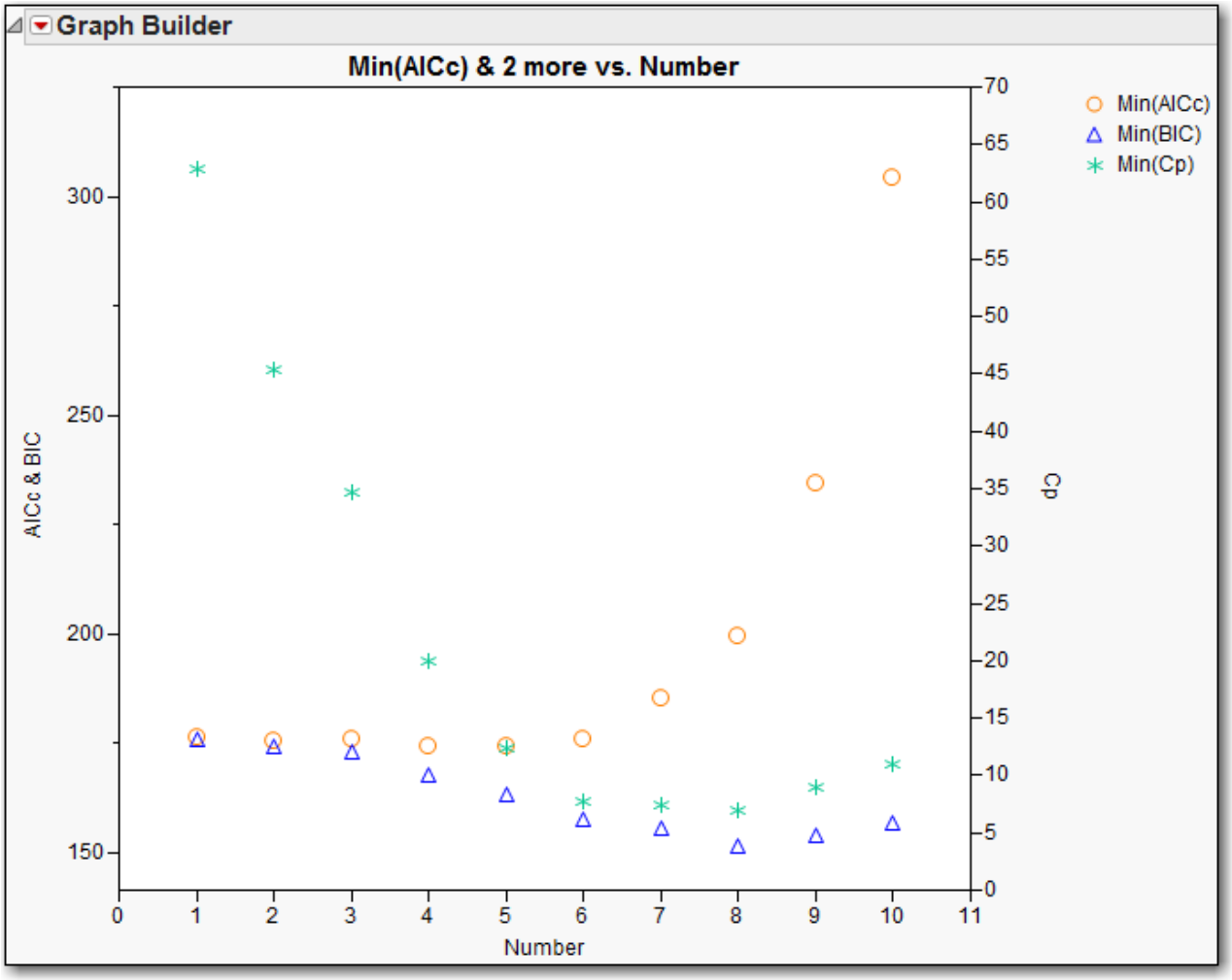


For this case 10 is the largest model that can be specified for AICc.

Finally, click on **OK** to generate the All Possible Models **Report**, which is then sorted in ascending order by AICc.

DSD: Model Selection for pDNA

From the **Graph Builder** overlay plot, AICc indicates a model with 4 or 5 terms, while BIC and Cp indicate a model with 7 or 8 terms.



DSD: Model Selection for pDNA

Based on AICc models with 4 and 5 effects were compared and a 5 effects model was selected for further analysis based on the relatively low RMSE in addition to the small AICc

	Model	Number	RSquare	RMSE	AICc	BIC	Cp
1	Induction Temperature C(39.5,42.5),Feed rate, mL/hr(1.9,3.5),I	4	0.8401	46.6822	174.2871	168.0354	19.9804
2	%DO(20,40),Induction Temperature C(39.5,42.5),Feed rate, m	5	0.9019	38.5471	174.4622	163.4186	12.3293
3	pH(6.8,7.2),%DO(20,40),Feed rate, mL/hr(1.9,3.5),pH*%DO	4	0.8339	47.5824	174.8601	168.6084	20.9531
4	pH(6.8,7.2),%DO(20,40),Feed rate, mL/hr(1.9,3.5),pH*%DO,%	5	0.8935	40.1616	175.6931	164.6495	13.6403
5	%DO(20,40),Induction Temperature C(39.5,42.5),Feed rate, m	5	0.8926	40.3362	175.8233	164.7796	13.7853
6	pH(6.8,7.2),%DO(20,40),Feed rate, mL/hr(1.9,3.5),pH*%DO,%	5	0.8798	42.6599	177.5035	166.4599	15.7749
7	pH(6.8,7.2),%DO(20,40),Feed rate, mL/hr(1.9,3.5),Feed rate,	4	0.7946	52.9066	178.0420	171.7903	27.0860
8	pH(6.8,7.2),%DO(20,40),Feed rate, mL/hr(1.9,3.5),pH*%DO,p	5	0.8733	43.8013	178.2957	167.2520	16.7931
9	%DO(20,40),Induction Temperature C(39.5,42.5),Feed rate, m	4	0.7588	57.3320	180.4519	174.2002	32.6782
10	%DO(20,40),Feed rate, mL/hr(1.9,3.5),%DO*Feed rate, mL/hr,	4	0.7582	57.4128	180.4942	174.2425	32.7845

DSD: Model Selection for pDNA

Based on BIC an 8 effects model was examined having the smallest RMSE.

Also a 7 effects model having the smallest BIC was further examined.

	Model	Number	RSquare	RMSE	AICc	BIC	Cp
1	pH(6.8,7.2),%DO(20,40),Induction Temperature C(39.5,42.5),I	7	0.9590	28.2550	185.3741	155.7466	7.4060
2	pH(6.8,7.2),%DO(20,40),Induction Temperature C(39.5,42.5),	7	0.9576	28.7331	185.8775	156.2500	7.6246
3	pH(6.8,7.2),%DO(20,40),Induction OD600(20,40),Feed rate, m	7	0.9568	28.9987	186.1536	156.5260	7.7476
4	pH(6.8,7.2),%DO(20,40),Induction Temperature C(39.5,42.5),	7	0.9545	29.7656	186.9367	157.3091	8.1093
5	pH(6.8,7.2),%DO(20,40),Induction OD600(20,40),Feed rate, m	7	0.9535	30.0893	187.2611	157.6336	8.2647
6	pH(6.8,7.2),%DO(20,40),Induction Temperature C(39.5,42.5),I	8	0.9742	24.2010	199.4156	151.4961	7.0282
7	pH(6.8,7.2),%DO(20,40),Induction Temperature C(39.5,42.5),I	8	0.9714	25.4981	200.9819	153.0624	7.4716
8	pH(6.8,7.2),%DO(20,40),Induction Temperature C(39.5,42.5),I	8	0.9711	25.6156	201.1199	153.2004	7.5129
9	pH(6.8,7.2),%DO(20,40),Induction Temperature C(39.5,42.5),I	8	0.9700	26.1052	201.6878	153.7683	7.6871
10	pH(6.8,7.2),%DO(20,40),Induction Temperature C(39.5,42.5),I	8	0.9699	26.1527	201.7424	153.8229	7.7041

DSD: Model Selection for pDNA

First, we examine the n = 5 effects model based on AICc

Lack Of Fit

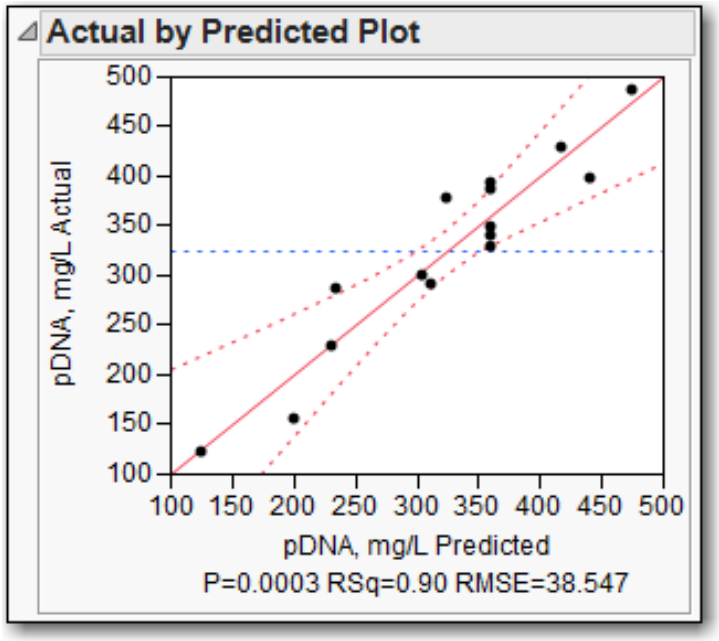
Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	5	9883.408	1976.68	2.2659
Pure Error	4	3489.505	872.38	Prob > F
Total Error	9	13372.912		0.2241
				Max RSq
				0.9744

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	360.56286	14.56944	24.75	<.0001*
%DO(20,40)	-34.19583	14.36566	-2.38	0.0412*
Induction Temperature C(39.5,42.5)	-21.92917	14.36566	-1.53	0.1612
Feed rate, mL/hr(1.9,3.5)	80.7625	14.36566	5.62	0.0003*
Induction Temperature C*Feed rate, mL/hr	-59.53864	16.43654	-3.62	0.0056*
Feed rate, mL/hr*Feed rate, mL/hr	-82.20377	20.36881	-4.04	0.0029*

Press

Press	Press RMSE
71537.287366	69.0590508



DSD: Model Selection for pDNA

Next, we examine the n = 7 effects model based on BIC.

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	3	2098.9003	699.633	0.8020
Pure Error	4	3489.5047	872.376	Prob > F
Total Error	7	5588.4050		0.5542

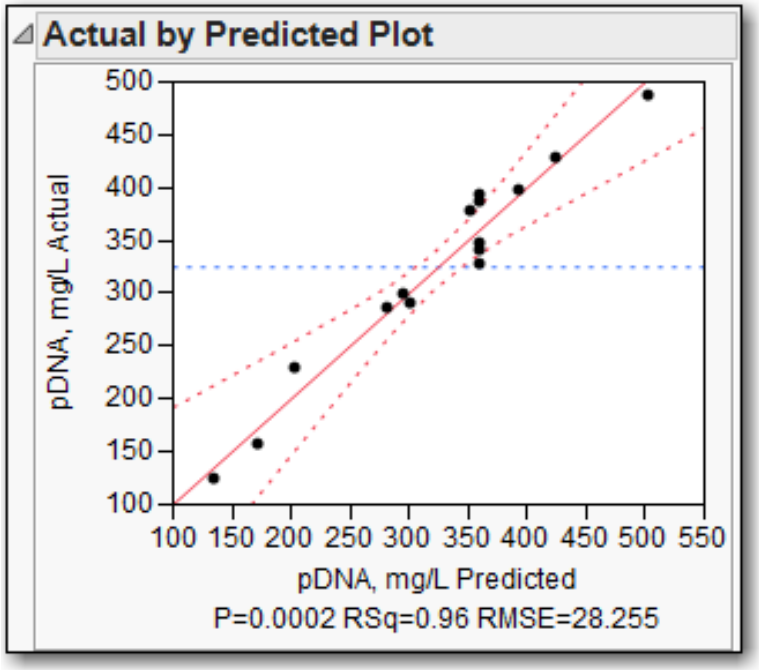
Max RSq
0.9744

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	360.56286	10.67938	33.76	<.0001*
pH(6.8,7.2)	-33.98295	11.66539	-2.91	0.0226*
%DO(20,40)	-35.99432	11.66539	-3.09	0.0177*
Induction Temperature C(39.5,42.5)	-14.17159	11.66539	-1.21	0.2638
Induction OD600(20,40)	19.648864	11.66539	1.68	0.1360
Feed rate, mL/hr(1.9,3.5)	95.660227	11.66539	8.20	<.0001*
Induction Temperature C*Feed rate, mL/hr	-59.53864	12.04796	-4.94	0.0017*
Feed rate, mL/hr*Feed rate, mL/hr	-82.20377	14.93031	-5.51	0.0009*

Press

Press	Press RMSE
58297.820337	62.3419711



DSD: Model Selection for pDNA

Finally, we examine the n = 8 effects model based on BIC.

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	2	24.6321	12.316	0.0141
Pure Error	4	3489.5047	872.376	Prob > F
Total Error	6	3514.1368		0.9860

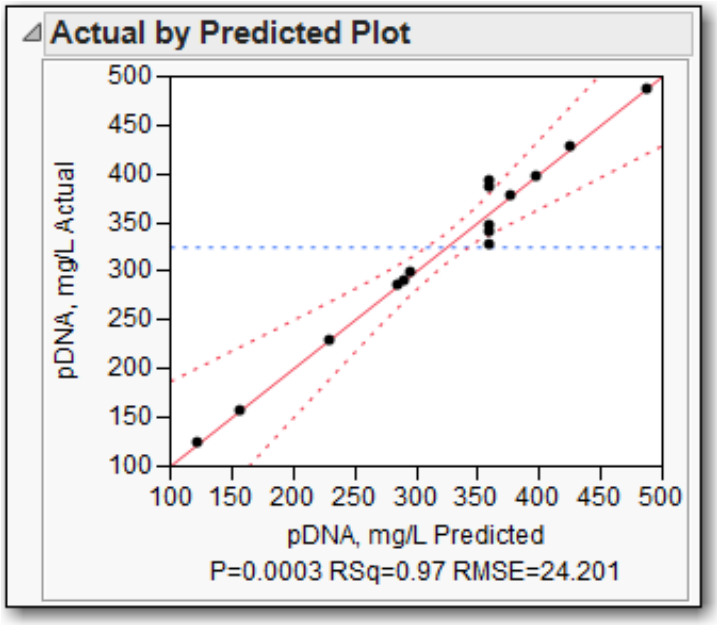
Max RSq
0.9744

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	360.56286	9.147127	39.42	<.0001*
pH(6.8,7.2)	-33.98295	9.991668	-3.40	0.0145*
%DO(20,40)	-35.99432	9.991668	-3.60	0.0113*
Induction Temperature C(39.5,42.5)	-14.17159	9.991668	-1.42	0.2059
Induction OD600(20,40)	19.648864	9.991668	1.97	0.0968
Feed rate, mL/hr(1.9,3.5)	95.660227	9.991668	9.57	<.0001*
%DO*Feed rate, mL/hr	19.500833	10.36226	1.88	0.1089
Induction Temperature C*Feed rate, mL/hr	-57.76583	10.36226	-5.57	0.0014*
Feed rate, mL/hr*Feed rate, mL/hr	-76.88536	13.09669	-5.87	0.0011*

Press

Press	Press RMSE
4975.7601978	18.2131092



The Actual by Predicted plot and very small Press indicate that this model may be over fit.

DSD: Model Selection for pDNA

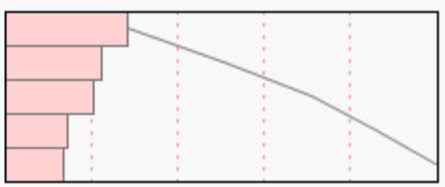
Pareto plots of the estimated coefficients can also be used to see how the three models compared in terms of the relative sizes of the estimated effects.

All three models agree on the important effects.

pH is a smaller effect in two models and did not appear in the 5 effect model; this was expected.

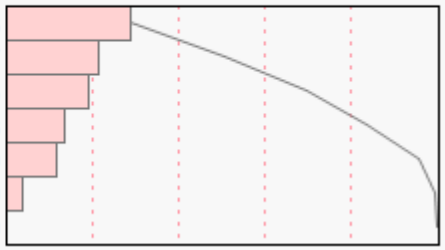
Pareto Plot of Transformed Estimates

Term	Orthog Estimate
Feed rate, mL/hr(1.9,3.5)	55.95390
%DO(20,40)	-43.72195
Feed rate, mL/hr*Feed rate, mL/hr	-40.16726
Induction Temperature C*Feed rate, mL/hr	-28.53593
Induction Temperature C(39.5,42.5)	-26.92780



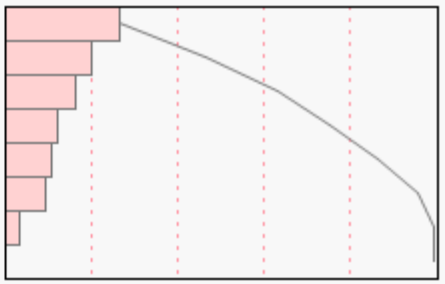
Pareto Plot of Transformed Estimates

Term	Orthog Estimate
Feed rate, mL/hr(1.9,3.5)	59.82483
%DO(20,40)	-45.09927
Feed rate, mL/hr*Feed rate, mL/hr	-40.16726
Induction Temperature C*Feed rate, mL/hr	-28.53593
Induction Temperature C(39.5,42.5)	-24.65000
Induction OD600(20,40)	-8.11769
pH(6.8,7.2)	0.21909



Pareto Plot of Transformed Estimates

Term	Orthog Estimate
Feed rate, mL/hr(1.9,3.5)	59.82483
%DO(20,40)	-45.09927
Feed rate, mL/hr*Feed rate, mL/hr	-36.68344
Induction Temperature C*Feed rate, mL/hr	-27.56784
Induction Temperature C(39.5,42.5)	-24.65000
%DO*Feed rate, mL/hr	21.45534
Induction OD600(20,40)	-8.11769
pH(6.8,7.2)	0.21909



DSD: Model Selection for pDNA

Four confirmation runs were performed; the measured pDNA titer for run 3 appears to be an outlier and was excluded from the analysis.

To the right are the overall mean and standard deviations of the residuals for the three model predictions of the 3 confirmation run pDNA titers.

Overall, the 8 effect model exhibited the smallest estimated mean bias and second smallest standard deviation.

Based on the confirmation runs **the 8 effects model is selected as best**, however the other two models are close in performance.

Residuals, Confirmation Runs n=5

Summary Statistics

Mean	21.716674
Std Dev	12.737003
Std Err Mean	7.3537123
Upper 95% Mean	53.357144
Lower 95% Mean	-9.923797
N	3

Residuals, Confirmation Runs n=7

Summary Statistics

Mean	29.483214
Std Dev	23.640785
Std Err Mean	13.649014
Upper 95% Mean	88.210181
Lower 95% Mean	-29.24375
N	3

Residuals, Confirmation Runs n=8

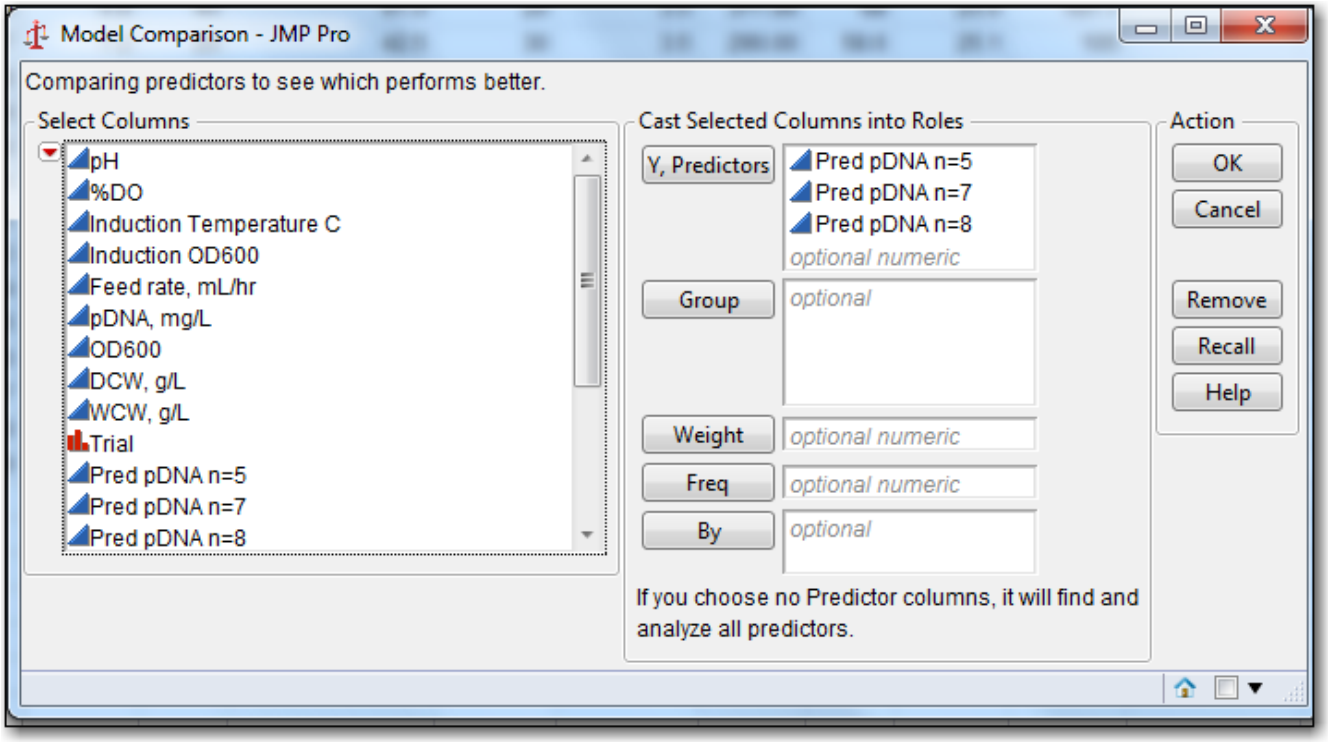
Summary Statistics

Mean	14.118921
Std Dev	15.094682
Std Err Mean	8.7149185
Upper 95% Mean	51.616189
Lower 95% Mean	-23.37835
N	3

DSD: Model Comparison With JMP Pro

If using the JMP Pro version one can use the **Model Comparison** platform (**Analyze → Modeling → Model Comparison**) to make comparisons of the final models.

Model Comparison uses the prediction formulas saved to the data table to make the comparisons.



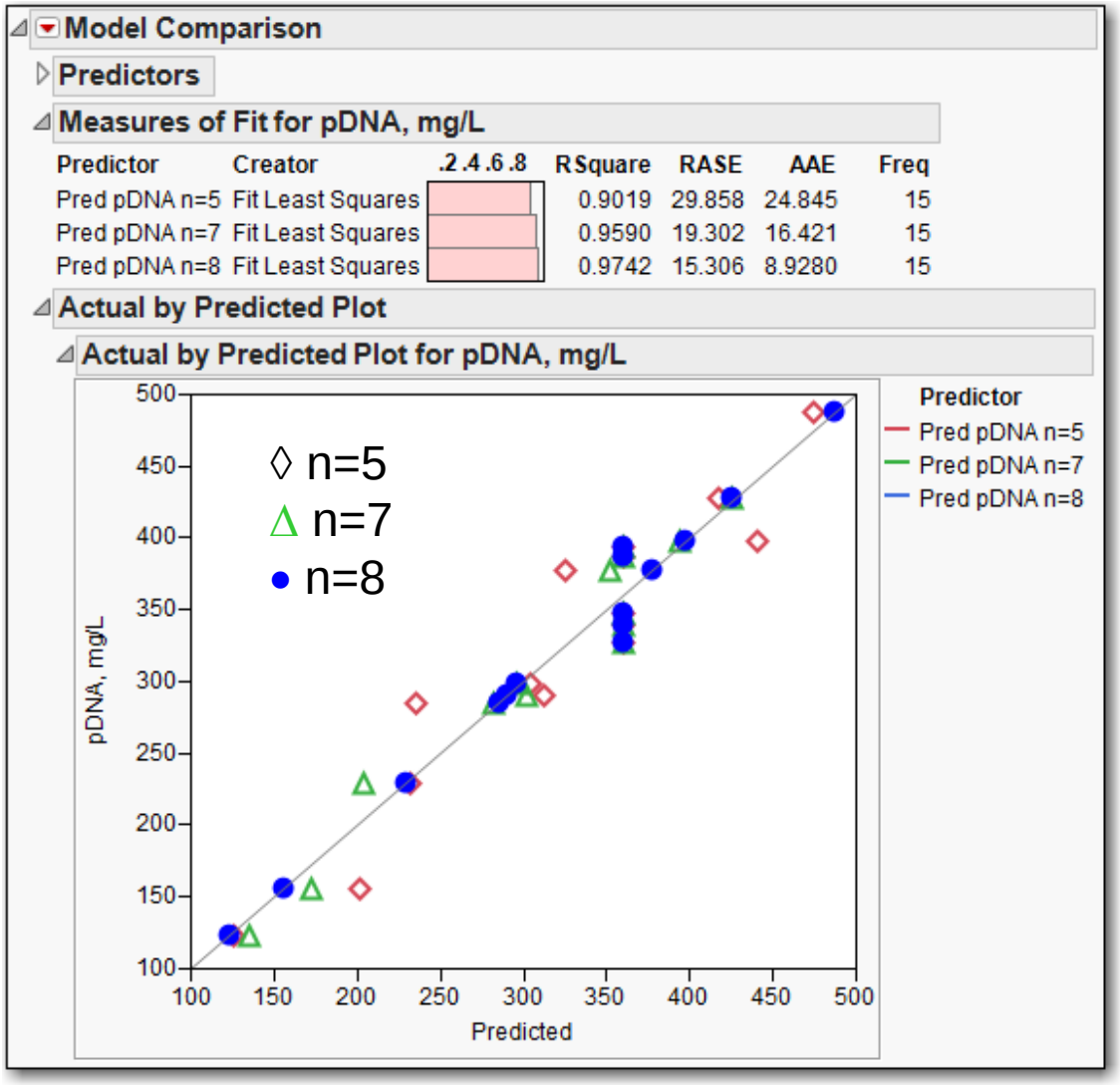
DSD: Model Comparison With JMP Pro

One can use fitting measures to compare models.

Actual by Predicted and Residual Plots are available.

A new column can be created in the data table which is an average of the prediction formulas.

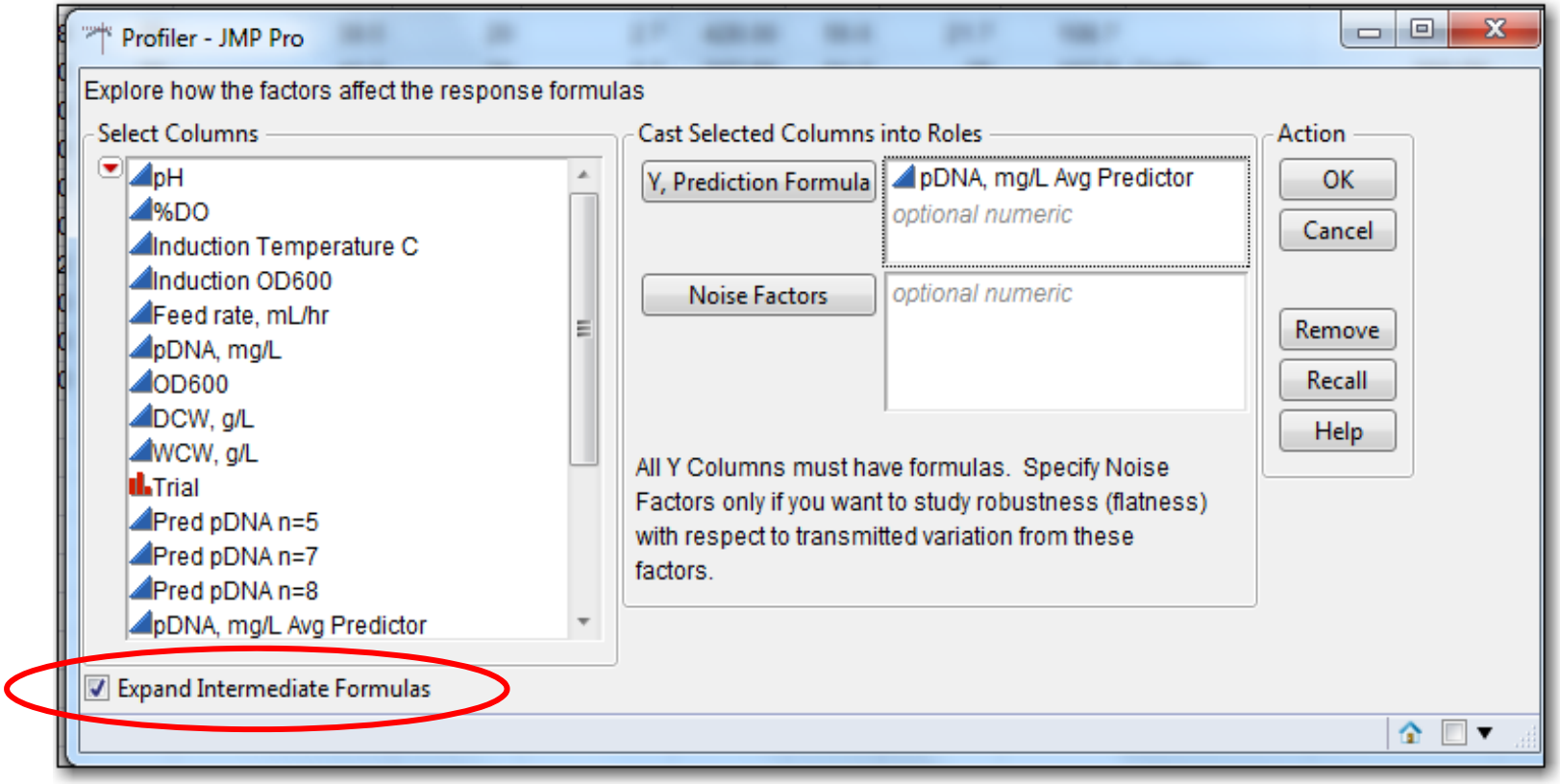
Perhaps an average of the model predictions will supply more precise predictions.



DSD: Model Comparison With JMP Pro

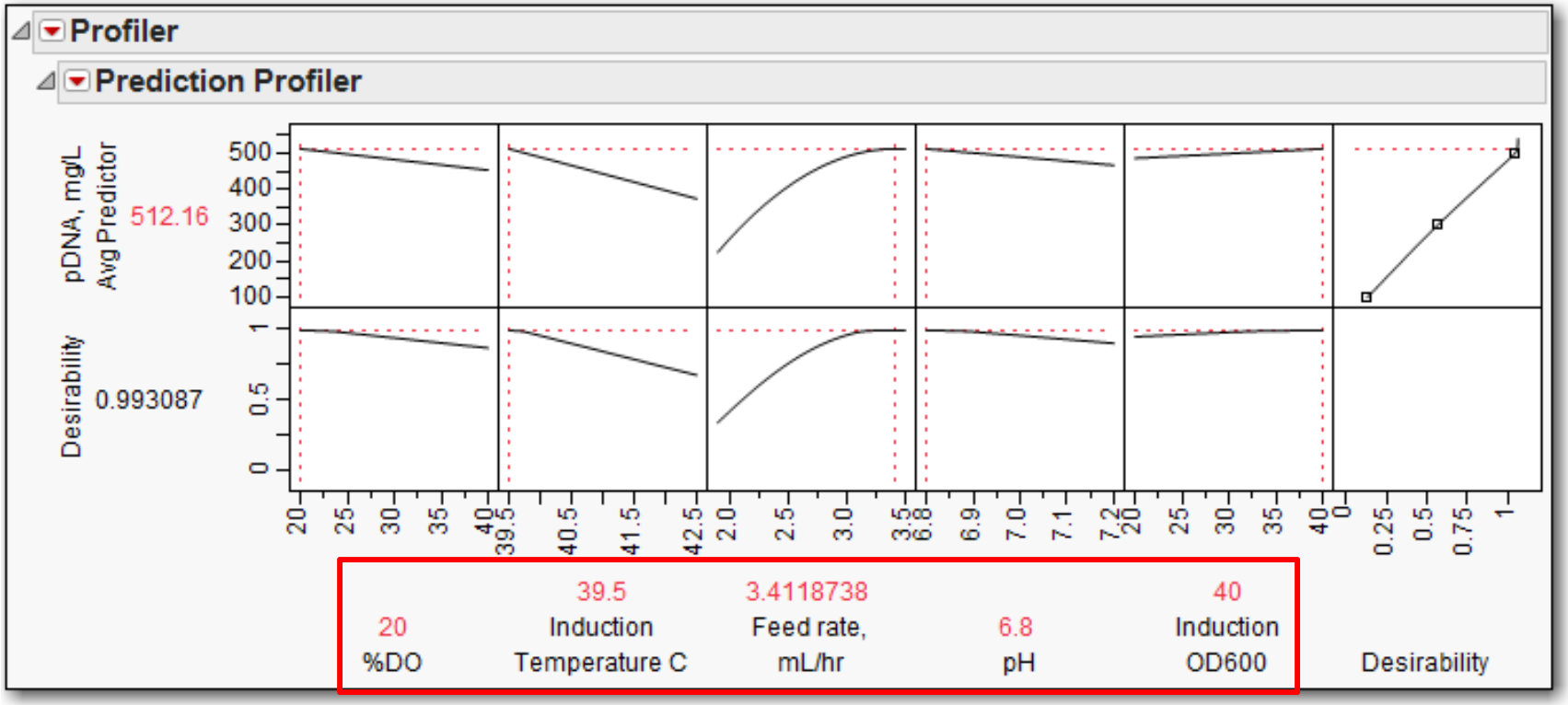
For the three DSD models an average prediction column was created and evaluated in the Profiler (**Graph → Profiler**).

Select the **Expand Intermediate Formulas** option.



DSD: Model Comparison With JMP Pro

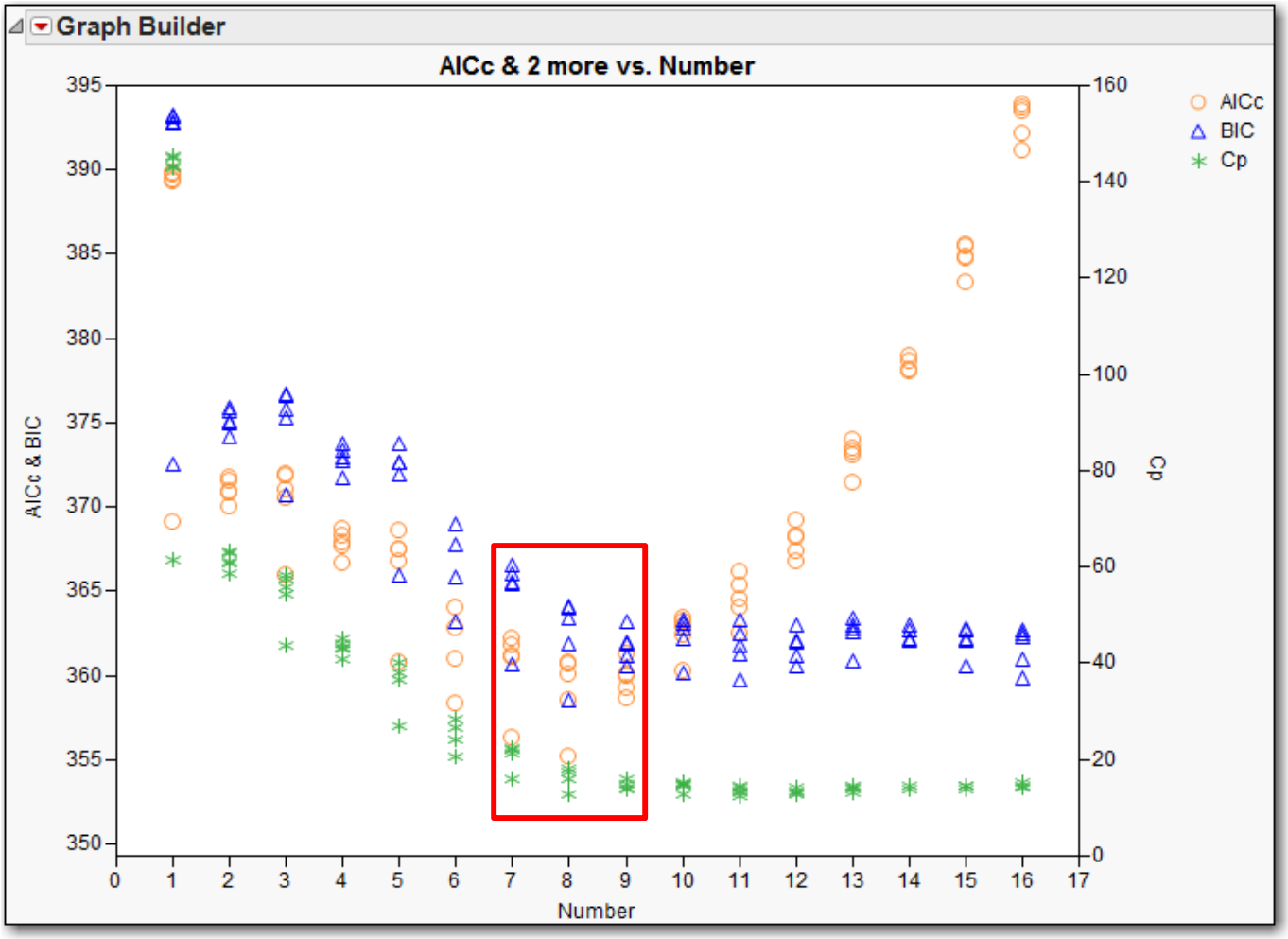
Below, we show the optimized settings of the factors to maximize pDNA production; this is based on a combination of the three predictive models.



Fractional Factorial: Model Selection for pDNA

Next we select models for the augmented fractional factorial experiment following the same heuristic as for the DSD results.

All three criteria indicate models in the range of 7 to 9 terms are best.



Fractional Factorial: Model Selection for pDNA

Also, given the axial points were run in a separate block from the fraction factorial runs, a blocking variable was included in the model.

Below are three selected models and unfortunately the blocking variable appears to be important in all three.

The 10 effects model was selected due to the inclusion of %DO in that model and the lowest RMSE.

	Model	Number	RSquare	RMSE	AICc	BIC	Cp
1	Block{2-1},Induction Temperature C(39.5,42.5),Feed Rate, mL	8	0.8497	53.6291	355.2326	358.5724	12.9630
2	Block{2-1},Induction Temperature C(39.5,42.5),Feed Rate, mL	8	0.8328	56.5535	358.5245	361.8643	15.8718
3	Block{2-1},Induction Temperature C(39.5,42.5),Feed Rate, mL	9	0.8566	53.6099	358.6631	360.5422	13.7652
4	Block{2-1},Induction Temperature C(39.5,42.5),Feed Rate, mL	9	0.8537	54.1478	359.2820	361.1611	14.2647
5	Block{2-1},pH(6.8,7.2),Induction Temperature C(39.5,42.5),Fe	9	0.8504	54.7663	359.9861	361.8653	14.8451
6	Block{2-1},Induction Temperature C(39.5,42.5),Feed Rate, mL	8	0.8242	57.9907	360.0804	363.4203	17.3579
7	Block{2-1},%DO(20,40),Induction Temperature C(39.5,42.5),F	9	0.8498	54.8607	360.0929	361.9720	14.9342
8	Block{2-1},%DO(20,40),Induction Temperature C(39.5,42.5),F	10	0.8734	51.6271	360.2526	360.1271	12.8735
9	Block{2-1},pH(6.8,7.2),Induction Temperature C(39.5,42.5),Fe	8	0.8209	58.5424	360.6675	364.0074	17.9383
10	Block{2-1},%DO(20,40),Induction Temperature C(39.5,42.5),F	8	0.8203	58.6267	360.7567	364.0966	18.0275
11	Block{2-1},%DO(20,40),Induction Temperature C(39.5,42.5),F	9	0.8439	55.9421	361.3031	363.1823	15.9668
12	Block{2-1},%DO(20,40),Induction Temperature C(39.5,42.5),F	10	0.8644	53.4256	362.3756	362.2501	14.4240
13	Block{2-1},pH(6.8,7.2),Induction Temperature C(39.5,42.5),Fe	10	0.8621	53.8751	362.8951	362.7696	14.8199
14	Block{2-1},Induction Temperature C(39.5,42.5),Feed Rate, mL	10	0.8607	54.1537	363.2149	363.0894	15.0669
15	Block{2-1},pH(6.8,7.2),Induction Temperature C(39.5,42.5),Fe	10	0.8595	54.3730	363.4655	363.3400	15.2622

Fractional Factorial: Model Selection for pDNA

Below are the Fit Model report details for the 8 effects model. Notice there is no evidence of Lack of Fit.

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	7	25358.026	3622.58	1.4331
Pure Error	15	37915.656	2527.71	Prob > F
Total Error	22	63273.682		0.2634

Max RSq
0.9099

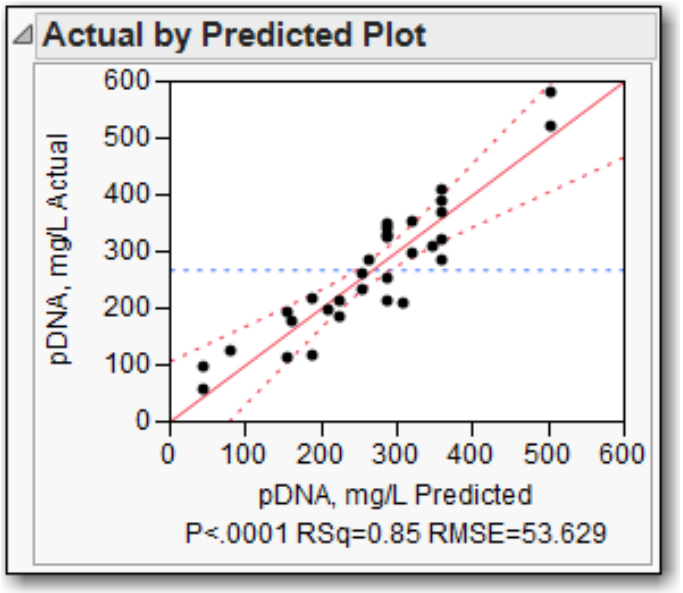
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	322.93186	17.39871	18.56	<.0001*
Block[1]	36.174492	12.19574	2.97	0.0071*
Induction Temperature C(39.5,42.5)	-17.50351	12.18213	-1.44	0.1648
Feed Rate, mL/hr(1.9,3.5)	102.88937	12.18213	8.45	<.0001*
Induction OD600(20,40)	-18.88777	12.18213	-1.55	0.1353
Induction Temperature C*Feed Rate, mL/hr	-54.155	13.40727	-4.04	0.0005*
Feed Rate, mL/hr*Feed Rate, mL/hr	-42.27195	20.34801	-2.08	0.0496*
Induction Temperature C*Induction OD600	52.37375	13.40727	3.91	0.0008*
Induction OD600*Induction OD600	-59.89621	20.34801	-2.94	0.0075*

Press

Press	Press RMSE
124115.6214	63.2750324

Actual by Predicted Plot



Fractional Factorial: Model Selection for pDNA

Below are the Fit Model report details for the 9 effects model and again no evidence of Lack of Fit.

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	6	22438.872	3739.81	1.4795
Pure Error	15	37915.656	2527.71	Prob > F
Total Error	21	60354.528		0.2508

Max RSq
0.9099

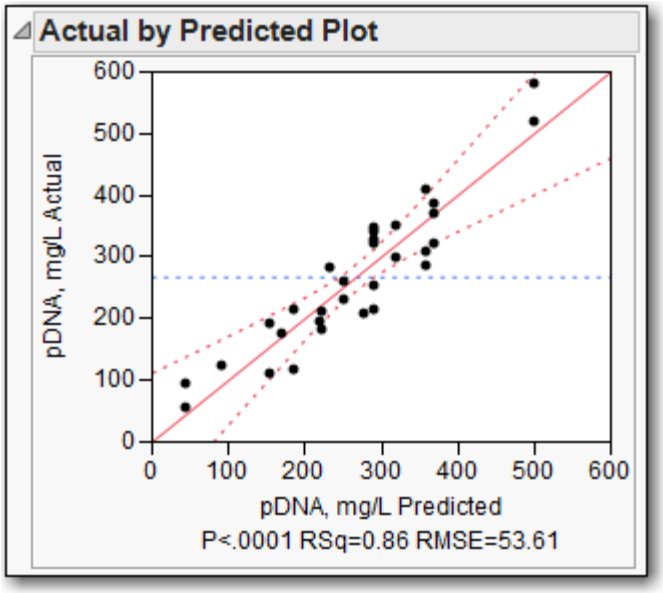
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	330.47749	18.93556	17.45	<.0001*
Block[1]	39.937589	12.75037	3.13	0.0050*
Induction Temperature C(39.5,42.5)	-17.50351	12.17779	-1.44	0.1654
Feed Rate, mL/hr(1.9,3.5)	102.88937	12.17779	8.45	<.0001*
Induction OD600(20,40)	-18.88777	12.17779	-1.55	0.1358
Induction Temperature C*Induction Temperature C	-20.83124	20.6696	-1.01	0.3250
Induction Temperature C*Feed Rate, mL/hr	-54.155	13.40248	-4.04	0.0006*
Feed Rate, mL/hr*Feed Rate, mL/hr	-38.57089	20.6696	-1.87	0.0761
Induction Temperature C*Induction OD600	52.37375	13.40248	3.91	0.0008*
Induction OD600*Induction OD600	-56.19515	20.6696	-2.72	0.0129*

Press

Press	Press RMSE
146479.11776	68.7396016

Actual by Predicted Plot



Fractional Factorial: Model Selection for pDNA

Below are the Fit Model report details for the 10 effects model with the %DO effect.

Lack Of Fit

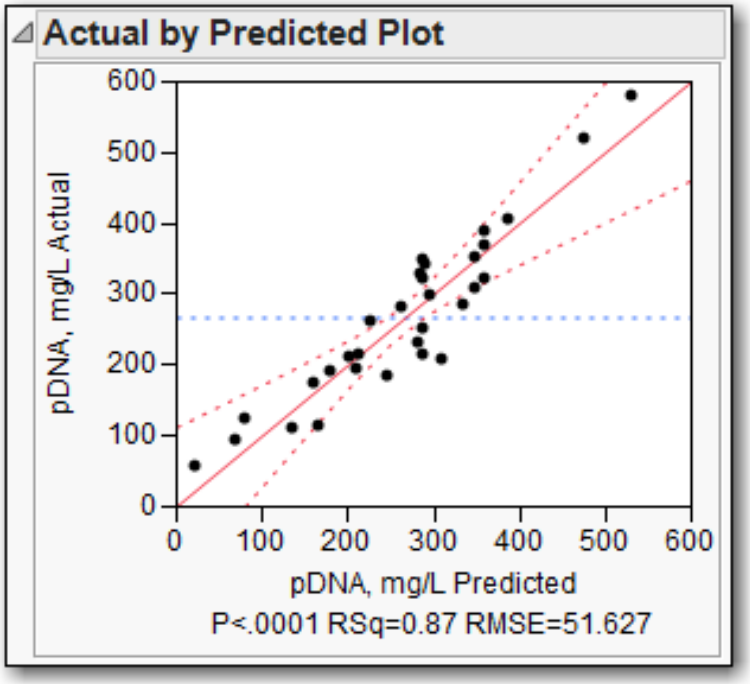
Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	15	39495.831	2633.06	0.9532
Pure Error	5	13811.413	2762.28	Prob > F
Total Error	20	53307.244		0.5744
				Max RSq
				0.9672

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	322.93186	16.74923	19.28	<.0001*
Block[1]	36.174492	11.74049	3.08	0.0059*
%DO(20,40)	1.9027864	11.72739	0.16	0.8727
Induction Temperature C(39.5,42.5)	-17.50351	11.72739	-1.49	0.1512
Feed Rate, mL/hr(1.9,3.5)	102.88937	11.72739	8.77	<.0001*
Induction OD600(20,40)	-18.88777	11.72739	-1.61	0.1229
%DO*Feed Rate, mL/hr	24.87	12.90679	1.93	0.0683
Induction Temperature C*Feed Rate, mL/hr	-54.155	12.90679	-4.20	0.0004*
Feed Rate, mL/hr*Feed Rate, mL/hr	-42.27195	19.58844	-2.16	0.0433*
Induction Temperature C*Induction OD600	52.37375	12.90679	4.06	0.0006*
Induction OD600*Induction OD600	-59.89621	19.58844	-3.06	0.0062*

Press

Press	Press RMSE
123967.81018	63.2373435



Fractional Factorial: Model Selection for pDNA

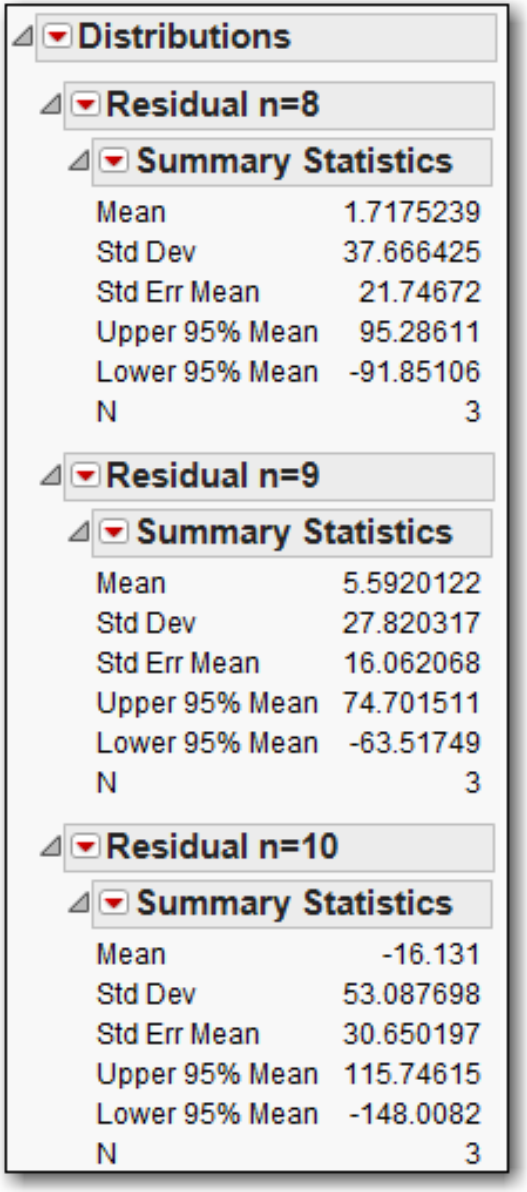
Finally, we compare the three models on the confirmation runs.

The 10 effects model exhibits the largest estimated bias and standard deviation of the residuals and was removed from further consideration.

Between the 8 and 9 effects models there is no clearly superior model.

Based on the smaller Press for 8 effects model it was selected as the “best” model.

Also, invoking the Principle of Parsimony the 8 effects model is preferred.



Fractional Factorial: Model Selection for pDNA

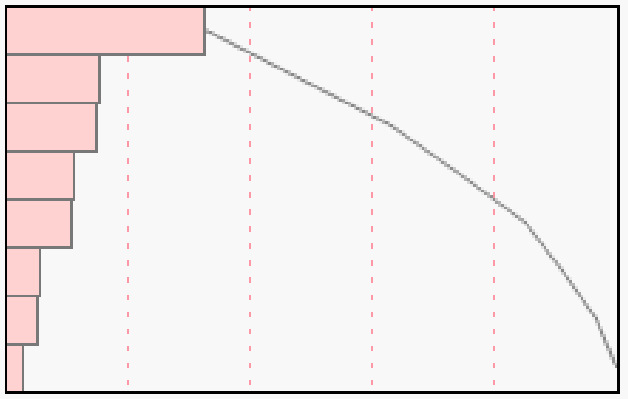
Below is a Pareto Plot of the coefficients for the 8 effects model.

It is interesting that %DO does not appear to be an important factor.

Is this a result of the control issues?

Pareto Plot of Transformed Estimates

Term	Orthog Estimate
Feed Rate, mL/hr(1.9,3.5)	81.35168
Induction Temperature C*Feed Rate, mL/hr	-38.90610
Induction Temperature C*Induction OD600	37.62641
Induction OD600*Induction OD600	-28.35286
Feed Rate, mL/hr*Feed Rate, mL/hr	-26.76814
Induction OD600(20,40)	-14.93402
Induction Temperature C(39.5,42.5)	-13.83952
Block[1]	7.35013



Model Selection for WCW

We omit the details, but the same heuristic was used to select best models for the WCW (wet cell weight) a measure of biomass.

A model was selected from the DSD experimental results and from the augmented fractional factorial results.

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	5	112.25486	22.4510	0.7508
Pure Error	4	119.61200	29.9030	Prob > F
Total Error	9	231.86686		0.6266

Max RSq
0.9728

Parameter Estimates WCW DSD Experiment

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	104.5814	1.340678	78.01	<.0001*
%DO(20,40)	-3.754545	2.053232	-1.83	0.1007
Induction Temperature C(39.5,42.5)	-11.13485	1.975723	-5.64	0.0003*
Induction OD600(20,40)	11.645455	2.053232	5.67	0.0003*
Feed rate, mL/hr(1.9,3.5)	14.701515	1.975723	7.44	<.0001*
Induction Temperature C*Feed rate, mL/hr	-6.439535	2.119798	-3.04	0.0141*

Press

Press	Press RMSE
679.92273097	6.73262074

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	18	562.74361	31.2635	1.3142
Pure Error	3	71.36500	23.7883	Prob > F
Total Error	21	634.10861		0.4694

Max RSq
0.9909

Parameter Estimates WCW Augmented FF Experiment

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	103.37638	1.527984	67.66	<.0001*
Block[1]	2.664958	1.162665	2.29	0.0323*
pH(6.8,7.2)	-2.616099	1.248232	-2.10	0.0484*
%DO(20,40)	-2.173375	1.248232	-1.74	0.0963
Induction Temperature C(39.5,42.5)	-9.275026	1.248232	-7.43	<.0001*
Feed Rate, mL/hr(1.9,3.5)	11.52322	1.248232	9.23	<.0001*
Induction OD600(20,40)	10.194014	1.248232	8.17	<.0001*
%DO*Induction Temperature C	-3.7125	1.373764	-2.70	0.0133*
Induction Temperature C*Induction Temperature C	-7.555336	2.035697	-3.71	0.0013*
Induction Temperature C*Induction OD600	2.6625	1.373764	1.94	0.0662

Press

Press	Press RMSE
1225.9090155	6.28851748

Comparing the Two Experiments

An important goal of the study was to compare the performance of a DSD against a more traditional screening design.

A direct comparison is complicated by several issues.

1. The relatively **looser control for %DO** in the augmented fractional factorial experiment compared to the DSD experiment.
2. The blocking factor in the augmented fractional factorial is an important effect with no known cause.
3. The axial run levels for the factors in the augmented fractional factorial design are farther from the design center than the settings for the factors in the DSD; this may allow the augmented fractional factorial design to better estimate quadratic effects.
4. The augmented fractional factorial design contains twice as many runs as the DSD.

Comparing the Two Experiments

We can make a comparison on the basis of how well the models from the two designs predict the response and on the factors found to be important in each of the two designs.

The goal of the experiments is to find factor settings that maximize pDNA production.

The table below compares four models and the factor settings to maximize pDNA using **Desirability Functions** in the **Prediction Profiler**. The two emboldened models are the selected models.

Model	pH	%DO	Ind. Temp.	Ind. OD600	Feed Rate	Pred. pDNA	LCL 95%	UCL 95%	Press RMSE
DSD n=8	6.8	20.0	39.5	40.0	3.40	**522.7**	462.9	582.5	18.2
DSD n=7	6.8	20.0	39.5	40.0	3.46	**537.6**	471.5	603.7	62.3
Aug. FF n=9			39.5	23.7	3.50	**468.2**	406.8	529.7	68.7
Aug. FF n=8			39.5	24.5	3.50	**476.4**	417.6	535.2	63.3

Comparing the Two Experiments

From the table of results from the optimization, the 95% confidence intervals about the predicted maximum pDNA overlap, indicating that all four models could predict the same theoretical mean pDNA production.

The primary difference between the DSD models and the augmented fractional factorial models is in the **roles of %DO and Induction OD600**.

Due to the %DO control issues in the augmented fractional factorial design this may explain why it does not appear to be an important factor in that experiment.

In the DSD experiment %DO is an important factor and this **result is consistent with the bio-process literature** on %DO.

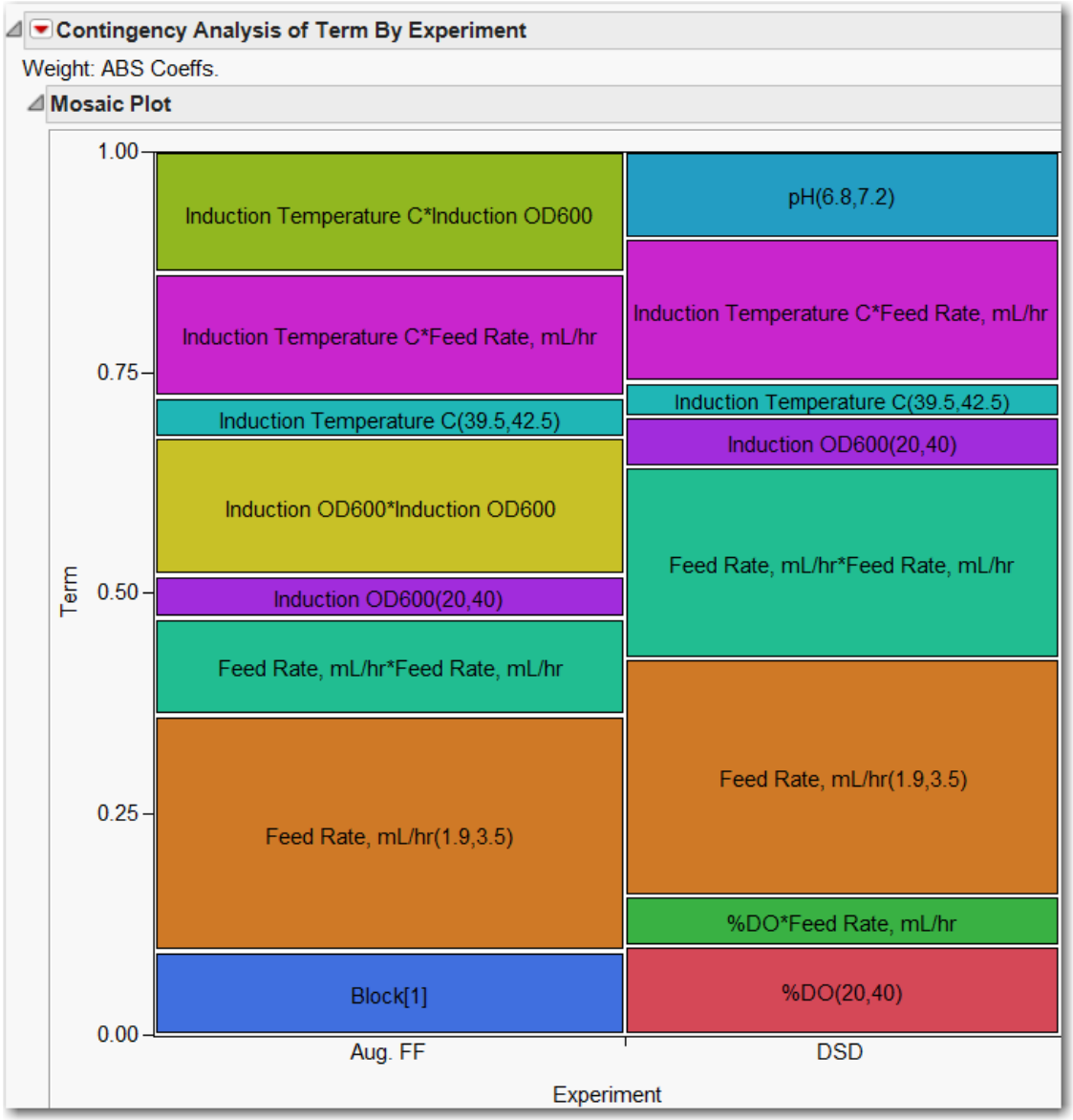
Induction OD600 is an important factor in both experiments, but has a stronger quadratic effect in the augmented fractional factorial.

Comparing the Two Experiments

To the right is a Mosaic Plot weighted by the absolute values of the coefficients.

Overall the two designs found the same important effects.

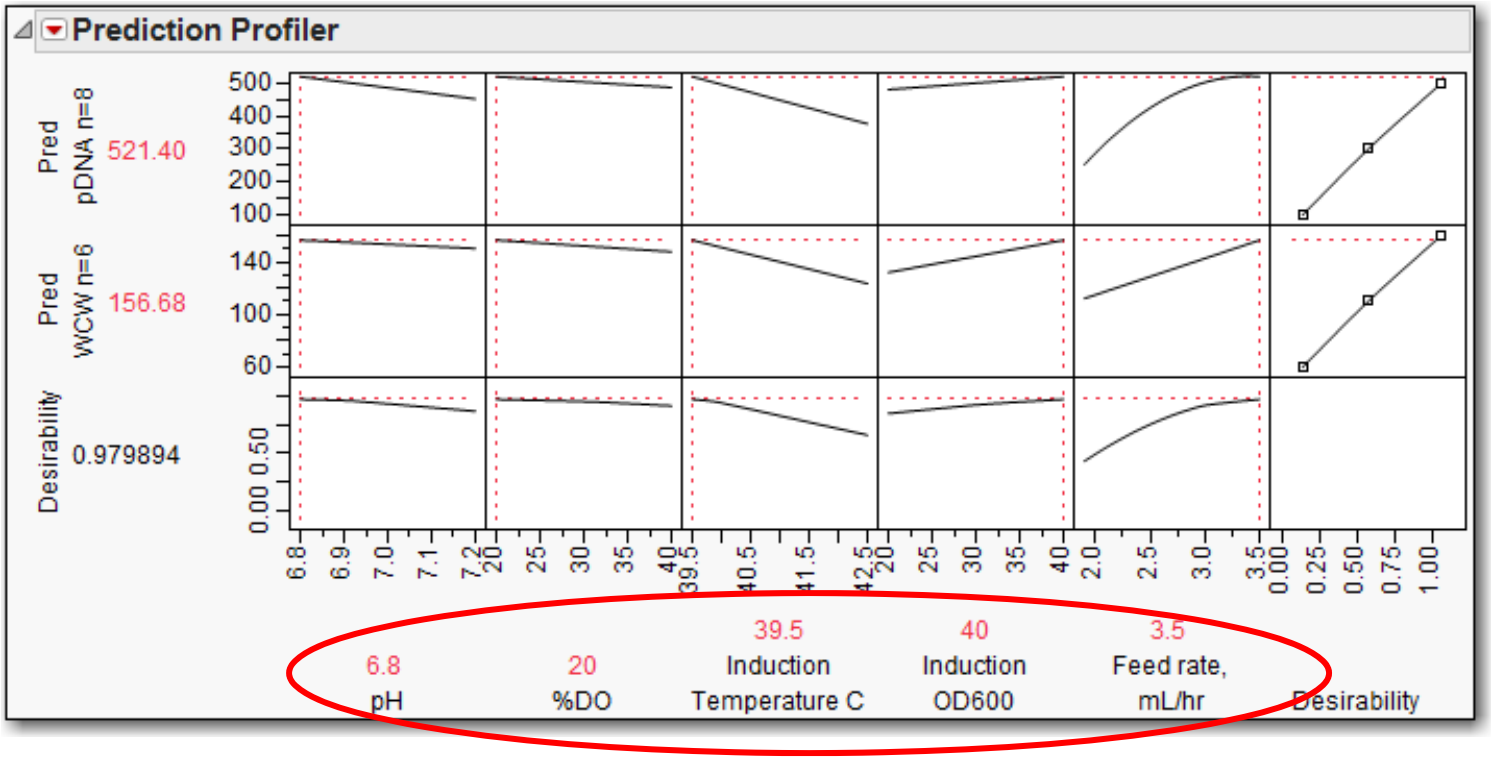
The primary discrepancy is in the importance of %DO as discussed earlier.



Comparing the Two Experiments

The secondary goal of the experiment was to maximize the E. Coli biomass as measured by **Wet Cell Weight** (WCW).

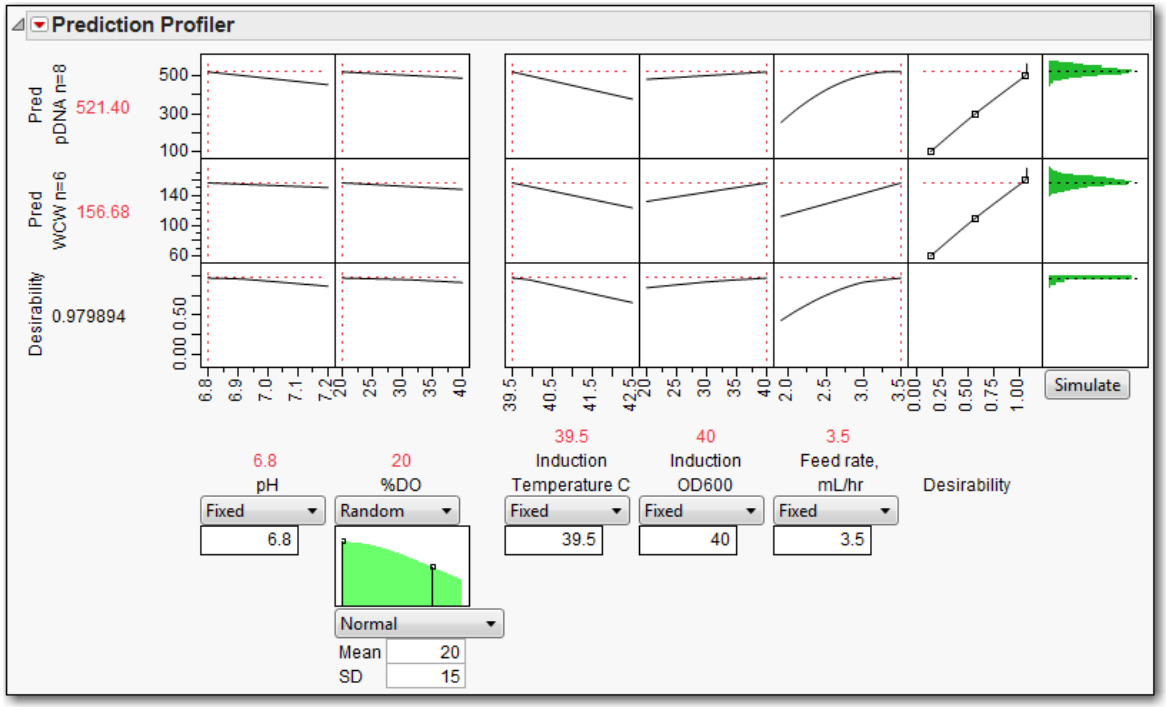
Using the **Profiler** in JMP and the predictive models for pDNA and WCW from the DSD experiment, we can perform a dual optimization.



Assessing Process Capability for pDNA

Using the pDNA predictive model from the DSD experiment and the Simulator in the JMP Prediction Profiler, one can examine the sensitivity of the pDNA response to variation in the inputs.

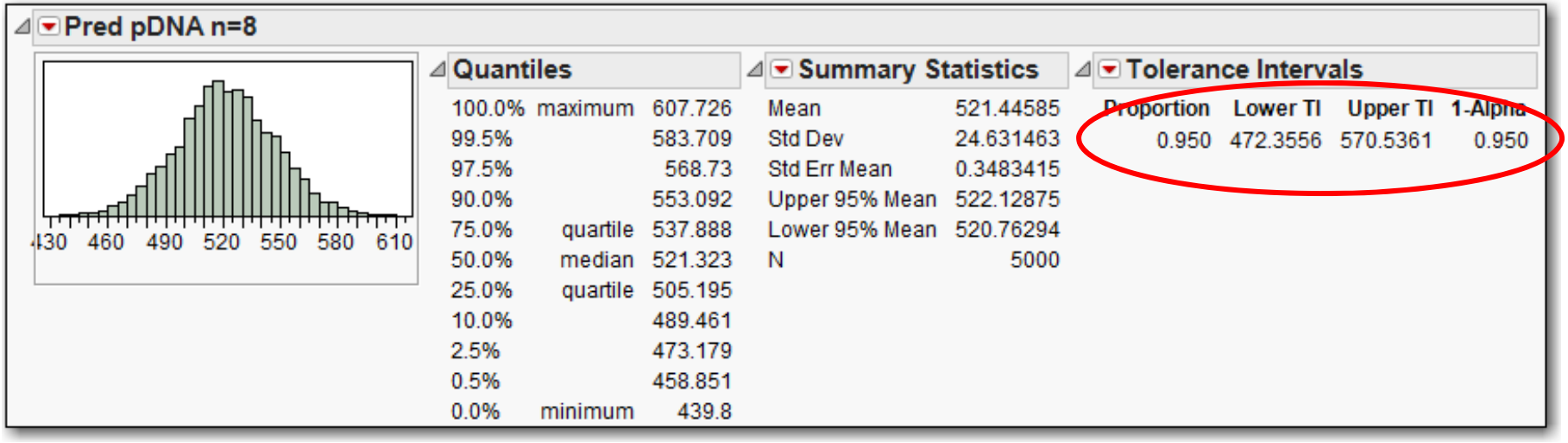
Our primary concern is %DO, which is known to vary substantially from batch to batch, we will use a %DO standard deviation of 15% to assess the impact on pDNA production.



Assessing Process Capability for pDNA

Using the results of the simulation, a 95% - 95% **tolerance interval** was estimated using the Distribution platform to assess the variation in pDNA production caused by poor %DO control.

The tolerance interval indicates with 95% confidence that at least 95% of the batches will have a pDNA titer in the range of **472.4 mg/L to 570.5 mg/L** at the optimized settings for the process factors.



Summary and Conclusions

A study was preformed to compare the **Definitive Screening Designs** to more traditional designs to characterize and optimize a fermentation step in a bio-manufacturing process.

Overall the DSD results **were as good and possibly better** than the results from the more traditional augmented fractional factorial experiment.

The subject matter expert found the DSD results to be superior.

Possibly due to control issues with %DO in the augmented fractional factorial experiment, %DO did not appear to be an important factor, which is counter to published results for %DO in fermentation.

In the DSD results, %DO was found to be an important factor.

Apart from %DO both experiments found the same process factors to be important; pH was found to have a minor effect in the DSD experiment and did not show up in the traditional design.

Summary and Conclusions

The DSD predictive model performed well in confirmation trials.

Given that the augmented fractional factorial design required 31 runs, while the DSD experiment had 15 runs, the DSD offers the potential to characterize and optimize a bio-process with far less experimentation, development time, and cost.

As a result, Definitive Screening Designs are recommended as a viable, cost effective, way to characterize bio-processes.

The end result is lower R&D cost and shorter time to commercialization for a new pDNA manufacturing process.

The case study also demonstrated that **All Possible Models** is a viable approach to model selection for a DSD experiment provided that **Effect Sparsity** holds to a reasonable degree – this is an area where future research is needed.

References

Bickel, P. J., Ritov,Y. and Tsybakov,A. (2008) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist*.

Burnham, K P, and D R Anderson. *Multimodal Inference: Understanding AIC and BIC in model selection*. Sociological Methods and Research 33, Nr. 2 (2004): 261-304.

FDA, CDER, CBER, and CVM (2011). Process Validation: General Principles and Practices. Guidance for industry.

Goos, P, and Jones, B. *Optimal Design of Experiments: A Case Study Approach*. John Wiley & Sons, LTD, 2011.

Jones, B. and Nachtsheim, C. (2011a). *Efficient Designs With Minimal Aliasing*. Technometrics, 53, 1, 62 – 71.

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.

References

Xiao, L, Lin, D, and Bai, F. *Constructing Definitive Screening Designs Using Conference Matrices*. Journal of Quality Technology 44, Nr. 1 (2012): 2-8.