JMP Discovery Americas 2023

CMC, SVEM, Neural Networks, DoE, & Complexity: It's All About Prediction

Patricia Dianne McNeill, Lundbeck Seattle Biopharmaceuticals
Philip J. Ramsey, University of New Hampshire and Predictum LLC

Acknowledgement

Thanks to Dr. Tiffany D. Rau, for her numerous insights and knowledge of bio-processes, CMC development, and regulatory affairs that inspired a good deal of the thinking and methodology discussed in this presentation regarding applications to biotechnology.

Tiffany D. Rau, Ph.D.

Global Biotechnology Consultant

Six Sigma Master Blackbelt

Bio Pharma Technical Consulting (Ireland)

Rau Consulting, LLC (USA)

tiffany@raubiotech.com

tiffany@bptc.ie

www.bptc.ie and www.raubiotech.com

Overview

- Introduce Lundbeck
- Process Development for a Biologic Drug
- Case Study
  - Phase I DSD
  - Phase II Augment with Space Filling Design

THIS IS LUNDBECK

Lundbeck

WE ARE

A global pharmaceutical company specialized in discovering and developing innovative treatments for brain diseases

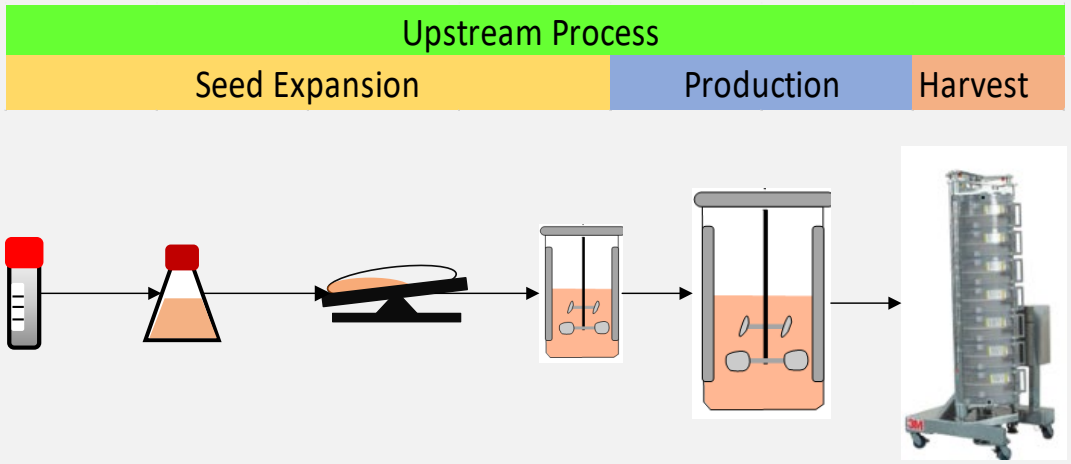
OUR BELIEFS

Patient-driven
Courageous
Ambitious
Passionate
Responsible

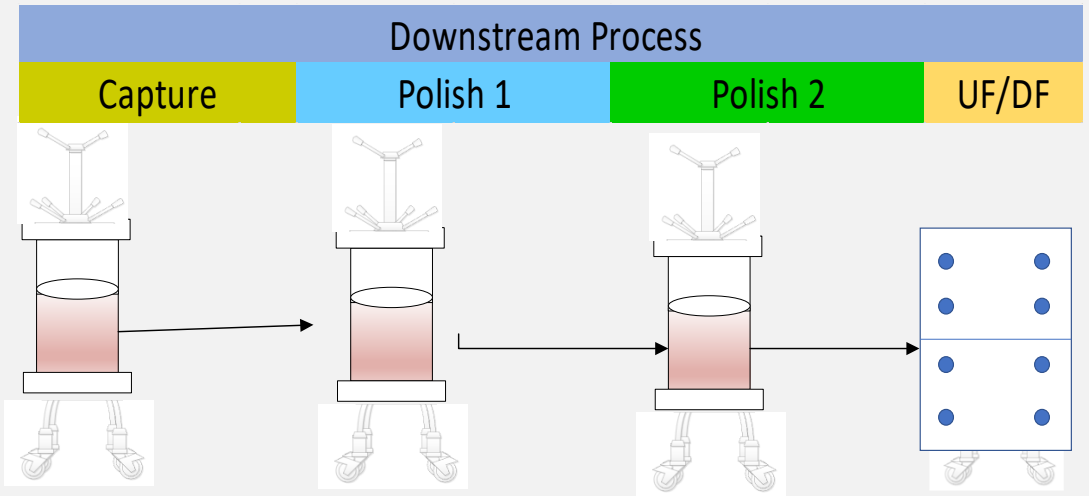
OUR LONG-TERM AMBITION

To be #1 in Brain Health

The Process we want to Develop for Drug Substance

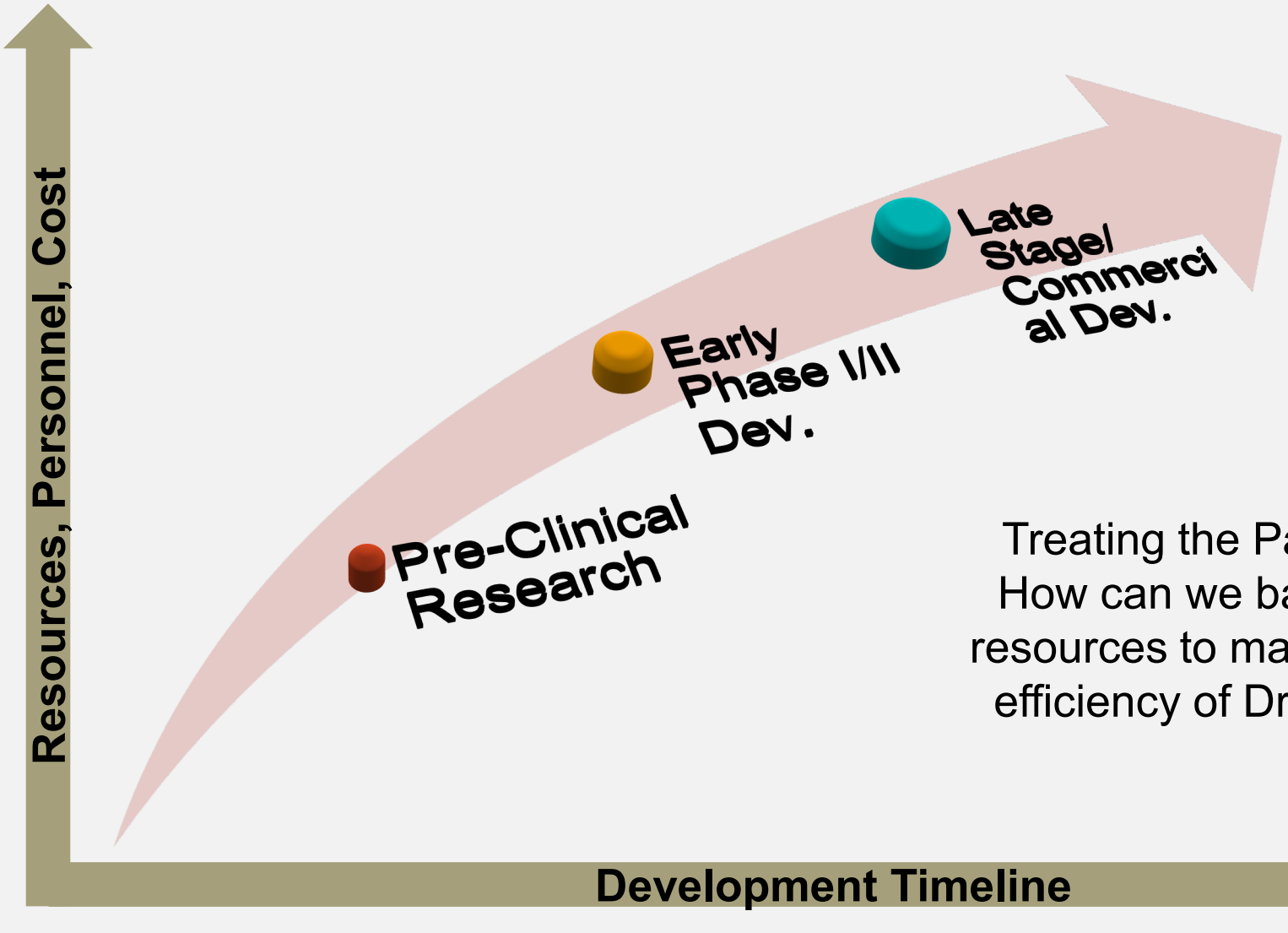


- Upstream Process Parameters
  - Seed Expansion
    - Seed Density, Temperature, Volume, Duration, CO_2
  - Production
    - pH, Seed Density, Temperature, Feeds (composition, %), Oxygen control, Duration



- Downstream Process Parameters
  - Capture and Polish Steps
    - Resins, Wash Buffer compositions, Elution Buffer compositions, Binding capacity, Temperature, Flow Rates, pH
  - Ultra Filtration/Diafiltration (UF/DF)
    - Load ratio, Buffer composition, pH, Temperature, Pressure, Flow Rates, Concentration

Development Timeline



Treating the Patient is the goal
How can we balance effort and resources to maximize speed and efficiency of Drug Development

Steps for Process Development

- Perform Risk Assessment
  - Identify key parameters that need screening – include stake holders in the discussion
  - Avoid prior assumptions - this is new cell line, new protein
- Initial Development
  - Narrow down categorical factors (cell line, resin, media, buffer)
- Early Development
  - Definitive Screening Design (DSD) – gain understanding of important process parameters
  - Be bold in level setting – widen your design space to get meaningful information
- Late Development
  - Space-filling designs build on early development studies for prediction of performance
  - Design experiments across Upstream and Downstream rather than silo investigations
- References:
  - Jones, Bradley. (2011). A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects. Journal of Quality Technology. 43. 1.
  - Effective Design-Based Model Selection for Definitive Screening Designs Jones B., Nachtsheim C.J. (2017) *Technometrics*, 59 (3) , pp. 319-329.
  - https://community.jmp.com/t5/JMP-Blog/Proper-and-improper-use-of-Definitive-Screening-Designs-DSDs/ba-p/30703

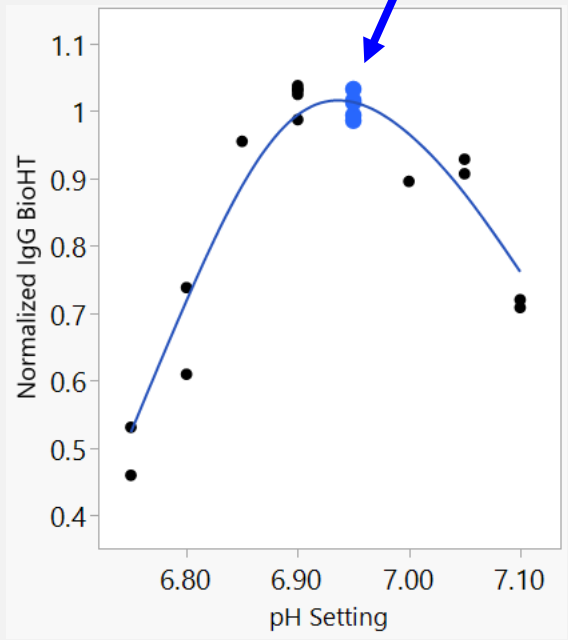
Steps for Process Development

- Perform Risk Assessment
  - Identify key parameters that need screening – include stake holders in the discussion
  - Avoid prior assumptions - this is new cell line, new protein
- Initial Development
  - Narrow down categorical factors (cell line, resin, media, buffer)
- **Early Development**
  - **Definitive Screening Design (DSD) – gain understanding of important process parameters**
  - **Investigates low, middle and high setting for continuous factors – allows fitting of a curve**
  - **Be bold in level setting – widen your design space to get meaningful information**
- **Late Development**
  - **Space-filling designs build on early development studies for prediction of performance**

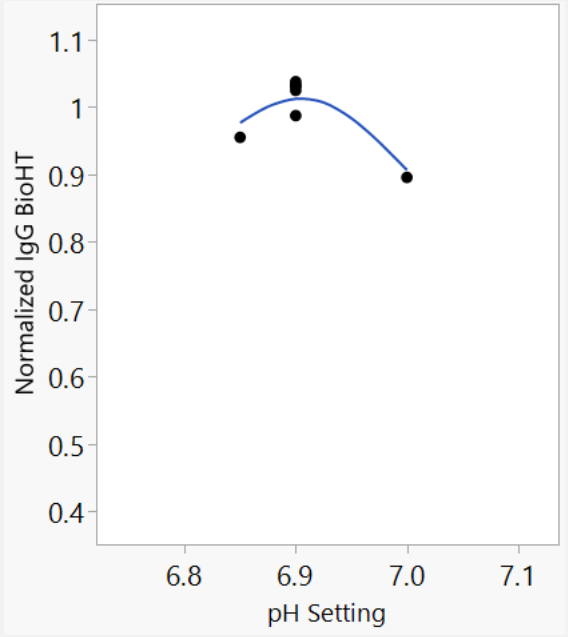
Analyze with SVEM

- References:
  - Jones, Bradley. (2011). A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects. Journal of Quality Technology. 43. 1.
  - Effective Design-Based Model Selection for Definitive Screening Designs Jones B., Nachtsheim C.J. (2017) *Technometrics*, 59 (3) , pp. 319-329.
  - https://community.jmp.com/t5/JMP-Blog/Proper-and-improper-use-of-Definitive-Screening-Designs-DSDs/ba-p/30703

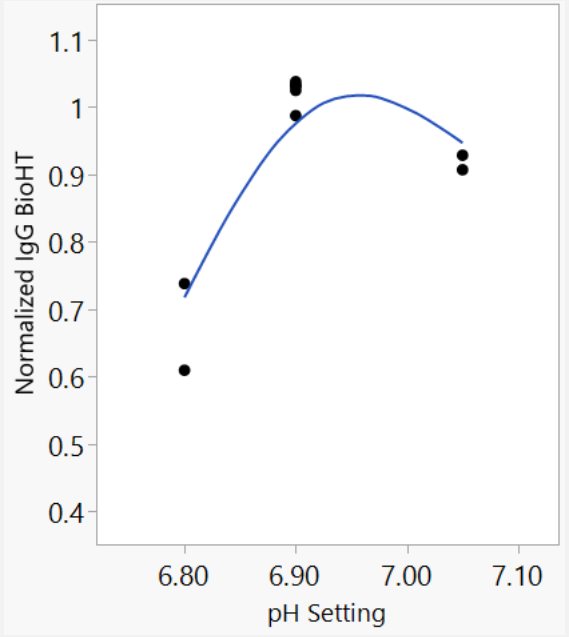
Setting Factor Ranges



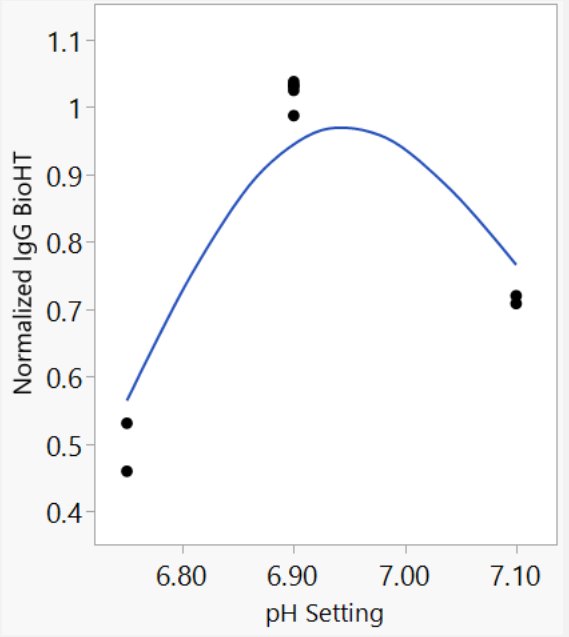
Full Data Set



Conservative Range Setting



Wider Range Setting



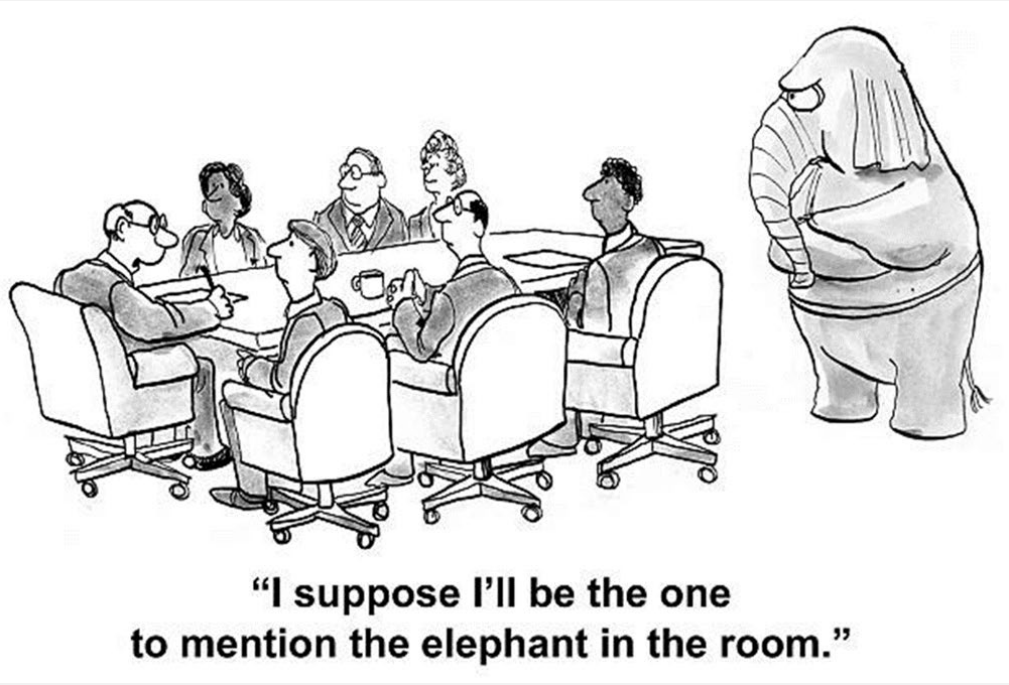
Bold Range Setting

The Complexity Problem

The Elephant in the room for science and statistics is the inherent complexity of physical systems…
Dr. Daniele Fanelli, LSE, *"The Neglected Importance of Complexity in Statistics and Metasciences"*, September 2022.

The implications for the future of statistical practice are profound requiring new thinking and innovative methods.

Example, the long-held DoE concepts of **Effect Sparsity** and **Hierarchy** are reductionist and upside down to the physical realities of complex systems.



Complex systems are defined by the interactivity of their elements.

The response surfaces are complex in shape due to the interactivity.

The Prediction Problem

Development work along the CMC pathway is inherently about predicting future performance.

Traditional DoE and analysis are focused on explanation not prediction; knowing "things" about a complex system is insufficient to predict the system behavior.

Prediction is **a measure of how well a model interpolates** over a design region.

If a predictive model adequately approximates the shape of the response surface, then the model will predict accurately at locations in the design space not used in the estimation of the model.

Predictive modeling requires a **training set** to estimate the model and a **test set** to evaluate the performance of the model at new locations in the design region.

Interpolation cannot be evaluated on a training set; a test set is required.

The Prediction Problem

The kinetics of chemical and biological systems are indeed complex and vary substantially throughout the design space due to the interactivity.

Kinetic behavior in the boundary region is often different from kinetic behavior observed within interior regions.

Both classical and optimal response surface designs are primarily boundary designs having most of the design points located at or near the boundaries; e.g., central composite designs (terrible designs for prediction).

Models fit to boundary designs can provide only speculation about kinetic behavior within the design region; little or no data exists to evaluate the actual behavior.

In order to model complex systems, designs must have substantial numbers of interior design points; space filling designs are the best option available today.

The Prediction Problem

Beside the problem of boundary designs is the assumption that full quadratic models are sufficient to approximate complex response surfaces.

The full quadratic model is too stiff in general to accommodate the shapes of response surfaces in complex systems; linear models in general are too stiff.

If one performs experiments with sufficient interior points, then the inadequacy of the full quadratic model frequently becomes apparent.

Cornell and Montgomery (1998) pointed out the problem but were ignored.

With the **SVEM** algorithm we can employ machine learning methods that allow more flexible models to approximate response surfaces.

We do not discuss SVEM directly as there are quite a few talks on SVEM available from past *Discovery;* there are **references at the end of the presentation**.

The Bio-process Characterization Experiment

A hybrid experimental design was employed to study the behavior of a unit operation in a bio-process. The experimental design incorporated 7 factors and there are 13 responses; due to time constraints we focus on only 4 responses.

Due to the highly proprietary nature of the experiment and a need to protect Lundbeck intellectual property the design has been partially anonymized. Therefore, the factor names, settings, and response names are anonymous.

The design structure and relationships between the factors and the responses have been preserved in the anonymization.

The hybrid design is a combination of a 19-run Definitive Screening Design(**DSD**) including 4 replicate center points and a 16-run Space Filling Design (**SFD**).

The DSD incorporates 3 center points and the SFD 2 center points.

The Bio-process Characterization Experiment

We will take several approaches to the analysis of the experiment.

1. A more typical traditional response surface approach. The DSD serves as the training set for modeling and the SFD as a test set to evaluate prediction.

2. The SFD is the training set for modeling and the DSD serves as a test set..

3. Next, we will construct a partition of the original 35 observations into a training set and test set since we have sufficient data to do so; a **holdback** test set is used

4. Finally, models are fit to the full data set for comparison of the predictive modeling strategies. In DoE it is common to fit models to the full training data set and attempt to judge prediction performance solely on the fit to the training data; without a validation or text set this is a poor, high-risk strategy for predictive modeling

The Bio-process Characterization Experiment

The following model building algorithms and platforms are employed:

1. **Fit Definitive (FD),** Fit Definitive Screening*;
2. **Forward Selection (FWS) with AICc**, Generalized Regression;
3. **SVEM FWS**, Generalized Regression;
4. **Moving Average (MA)**; Stepwise;
5. **SVEM Neural Network** (NN); Predictum SVEM Addin.
6. **K-Fold cross validation** for FWS and NN (full data set only).

The **full quadratic model (FQM)** is used for the FD, FWS, and MA methods.

*Fit Definitive Screening requires a **DSD;** it is not a predictive modeling algorithm.

The Bio-process Characterization Experiment

The four responses were selected based upon the observed complexity of the observed kinetic behavior; **a subjective assessment** by the authors.

The lower the level of complexity in the response surface the more likely a traditional response surface approach will be satisfactory; i.e., a **boundary design** combined with an assumed **full quadratic model** (FQM).

The table shown here lists the rated kinetic complexities (**ratings are important**)

Response	Complexity
Y2	Low
Y1	Moderate
Y3	High
Y4	Very High

The Bio-process Characterization Experiment

Before getting into the modeling results, we need to define criteria to assess the prediction capability of the fitted models; again, prediction is a measure of interpolation as assessed on the test set.

The most common measure is **Root Average Square Error (RASE)** of prediction. This is the standard deviation of prediction error, usually calculated on the test set (when one is available); JMP often refers to this as a validation RASE.

Although the most common measure of prediction performance, RASE is misleading in many cases as models with **substantial prediction bias** have smaller RASE values.

One must consider the **bias vs. variance tradeoff** in assessing the performance of predictive models.

The Bio-process Characterization Experiment

An effective measure of prediction bias is the slope of a simple linear model fit to the actual by predicted plot, again on the test set when available.

A **Slope = 1** and **Intercept = 0** indicate no detectable prediction bias.

The farther the fitted slope is from 1, the greater the prediction bias.

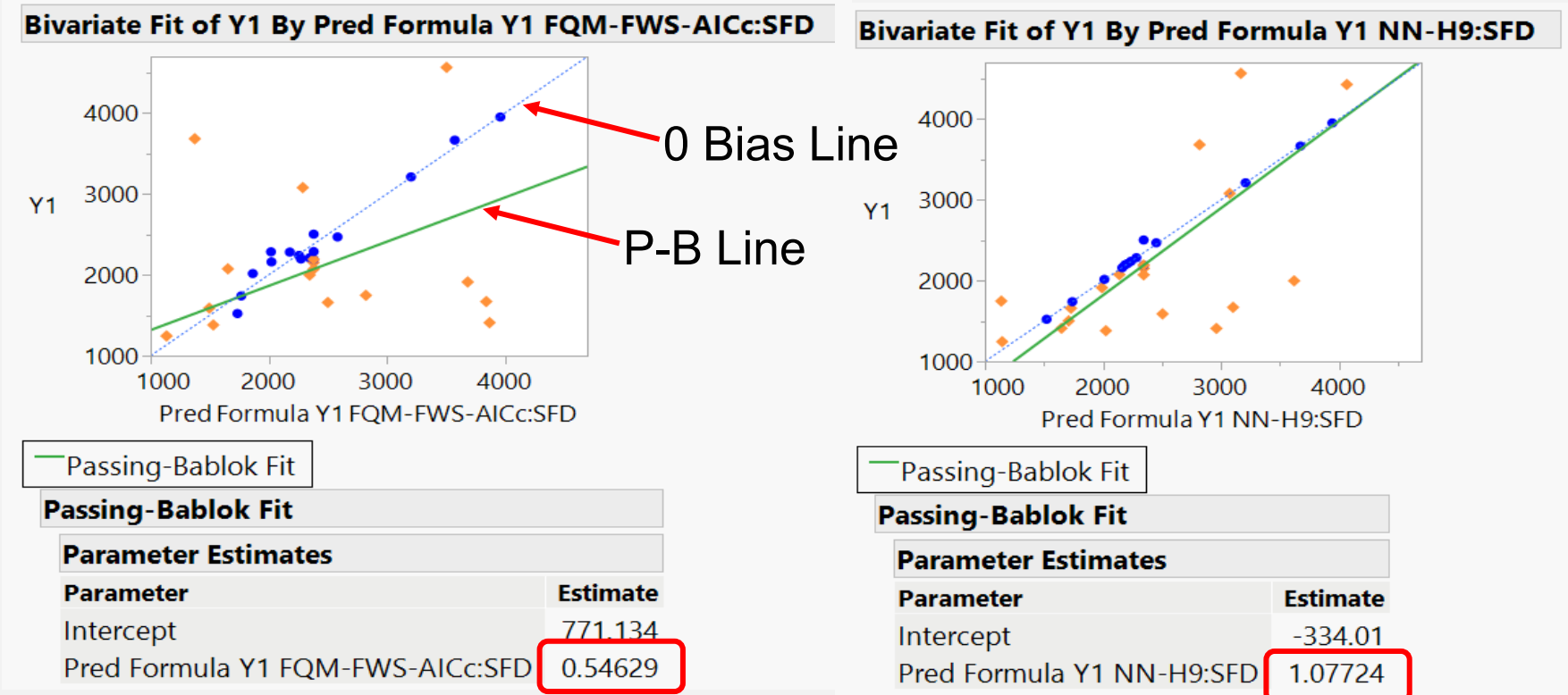
Predictive models with relatively low RASE values and slopes near to 1 are preferred; we use a slope range of **0.85 - 1.15** to define acceptable slope values.

The fitted slope of the actual by predicted plot typically is the usual ordinary least squares fit.

However, since the X axis consists of the predicted values there exist considerable error and the fitting is better viewed as an **errors in variables** problem.

The Bio-process Characterization Experiment

The **Passing-Bablok (P-B)** algorithm is a straightforward approach to fitting slopes with errors in the X values. The option is available in the JMP 17 Fit Y by X. Below are shown two actual by predicted plots with the P-B slopes, the model to the left has high prediction bias and the model to the right little or no bias.



The Bio-process Experiment: DSD as Training Set

Current DoE practice for response surface methods (RSM) is to use a boundary design (DSD in this case) and assume a full quadratic model or subset of the model is sufficient to approximate the underlying response surface.

Therefore, we use the DSD in the role of the **training set** to fit predictive models using the 5 approaches mentioned earlier.

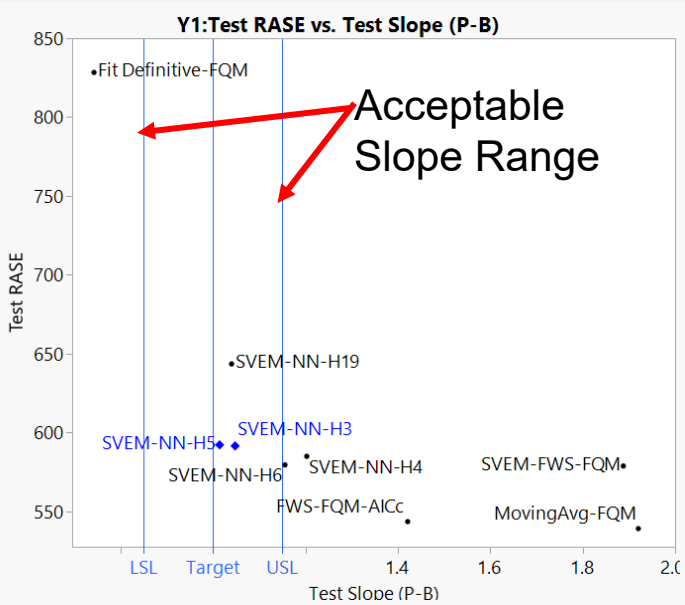
The DSD models are **interpolating** to the interior SFD points.

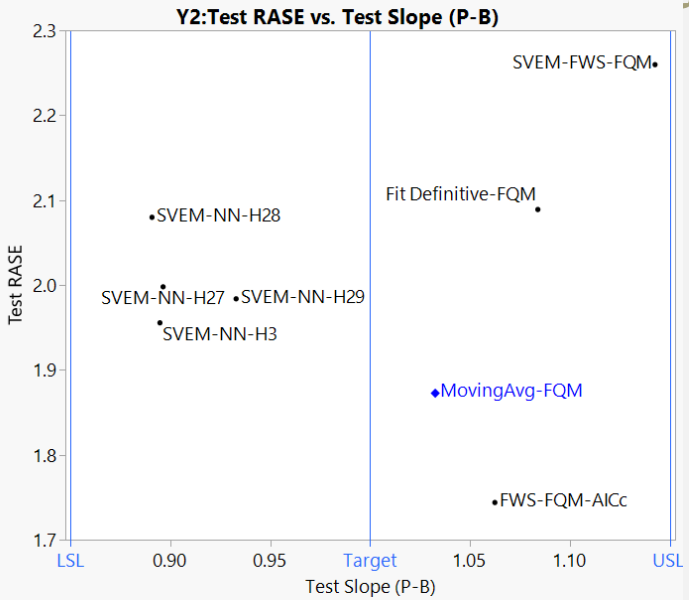
The predictive models are subsequently applied to the SFD data in the role of a **test set** to evaluate prediction performance; the ability to interpolate over the design space.

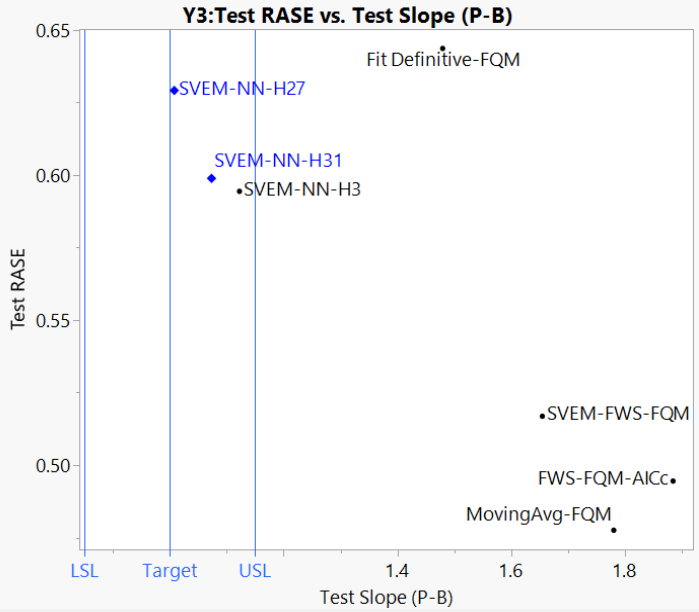
Since our four responses range from low complexity to very high complexity, we expect the algorithm prediction performances too vary substantially across those responses.

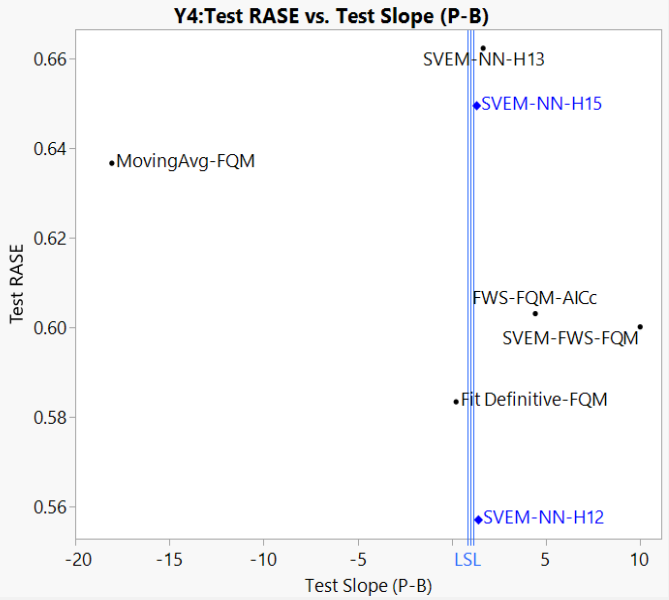
The Bio-process Experiment: DSD as Training Set

Model Algorithm	Response	Test RASE	Test Slope (P-B)
Fit Definitive-FQM	Y2	2.09	1.08
FWS-FQM-AICc	Y2	1.74	1.06
SVEM-FWS-FQM	Y2	2.26	1.14
MovingAvg-FQM	Y2	1.87	1.03
SVEM-NN-H3	Y2	1.96	0.89
SVEM-NN-H27	Y2	2.00	0.90
SVEM-NN-H28	Y2	2.08	0.89
SVEM-NN-H29	Y2	1.98	0.93
Fit Definitive-FQM	Y1	828.36	0.74
FWS-FQM-AICc	Y1	543.48	1.42
SVEM-FWS-FQM	Y1	578.76	1.89
MovingAvg-FQM	Y1	539.06	1.92
SVEM-NN-H3	Y1	591.47	1.05
SVEM-NN-H4	Y1	584.93	1.20
SVEM-NN-H5	Y1	592.17	1.01
SVEM-NN-H6	Y1	579.49	1.16
SVEM-NN-H19	Y1	643.48	1.04
Fit Definitive-FQM	Y3	0.64	1.48
FWS-FQM-AICc	Y3	0.49	1.88
SVEM-FWS-FQM	Y3	0.52	1.65
MovingAvg-FQM	Y3	0.48	1.78
SVEM-NN-H3	Y3	0.59	1.12
SVEM-NN-H27	Y3	0.63	1.01
SVEM-NN-H31	Y3	0.60	1.07
Fit Definitive-FQM	Y4	0.58	0.22
FWS-FQM-AICc	Y4	0.60	4.43
SVEM-FWS-FQM	Y4	0.60	9.99
MovingAvg-FQM	Y4	0.64	-18.06
SVEM-NN-H13	Y4	0.66	1.66
SVEM-NN-H12	Y4	0.56	1.41
SVEM-NN-H15	Y4	0.65	1.32









FQM = full quadratic model.

FWS = forward selection.

MA = moving average.

NN = neural network

Hxx = number of hidden nodes in the NN.

The Bio-process Experiment: DSD as Training Set

For the low complexity response **Y2** all the model algorithms fit to the DSD data predicted the SFD responses well with the **Moving Average** with **FQM** model having the best slope.

For the moderate complexity response **Y1** only the **NN** models fit to the DSD data predicted the SFD responses well with **H=5** hidden nodes having the best slope.

For the high complexity response **Y3** again only the **NN** models fit to the DSD data predicted the SFD responses well with **H=27** hidden nodes having the best slope.

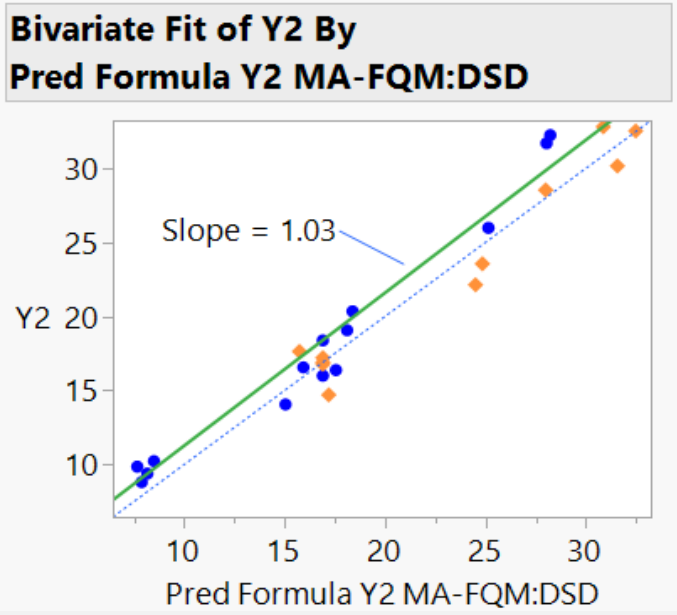
For the very high complexity response **Y4** none of the models fit to the DSD data predicted the SFD responses well; an **NN** with **H=15** hidden nodes having the best slope = 1.37.

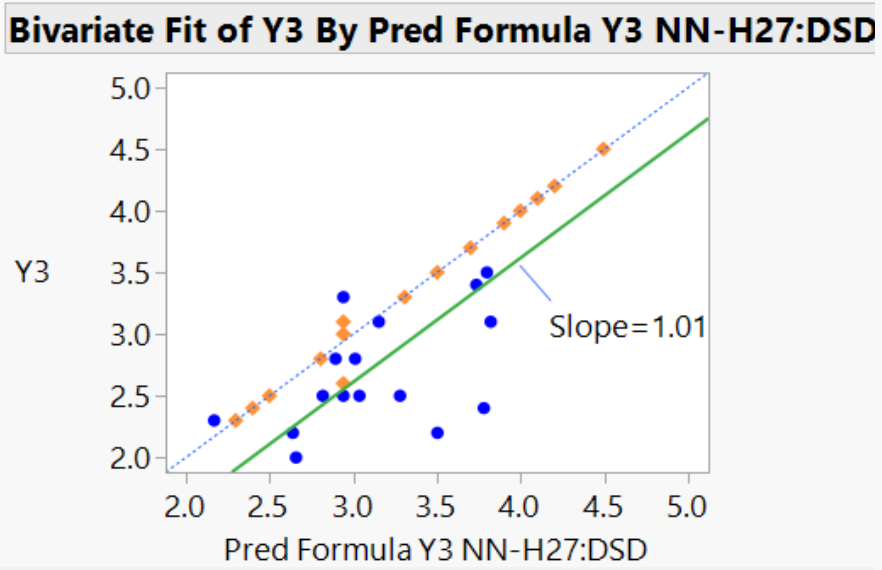
When the responses exhibit considerable complexity the boundary designs are not a sufficient basis for fitting models that predict the interior behavior.

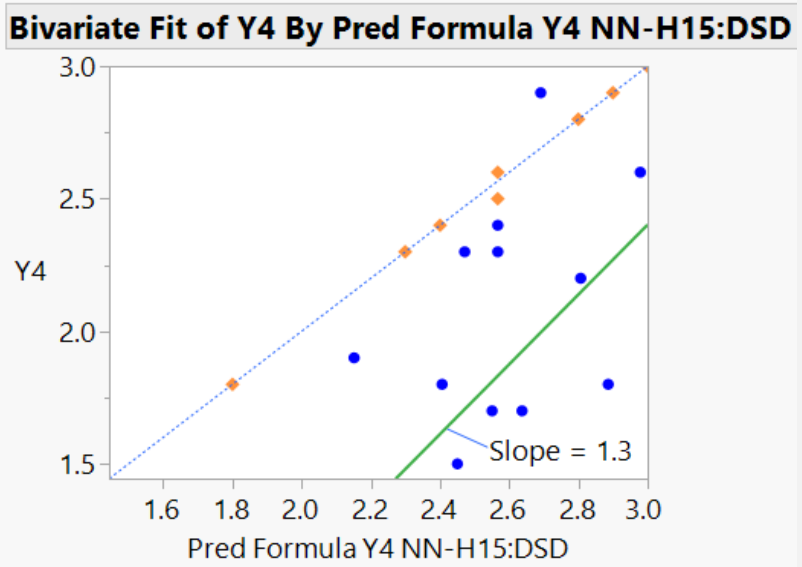
The Bio-process Experiment: DSD as Training Set

Here are shown actual by predicted plots for Y2, Y3, and Y4. For the low complexity Y2 response the DSD and SFD observations align well indicating the boundary behavior is predictive of interior behavior; the best models in slope are displayed.

For the high and very high complexity responses Y3, Y4 the DSD and SFD observations do not align. The boundary region is not predictive of the interior region







● SFD ♦ DSD

The Bio-process Experiment: SFD as Training Set

We repeat the analyses for the case where the DSD was used as a training set and the SFD as a test set except here we reverse the roles.

Here predictive models are fit to the **SFD** data as the **training set**, using the same five algorithms and then those models are applied to the **DSD** data as a test **set.**

The SFD models on the interior are **extrapolating** to the DSD boundary points.

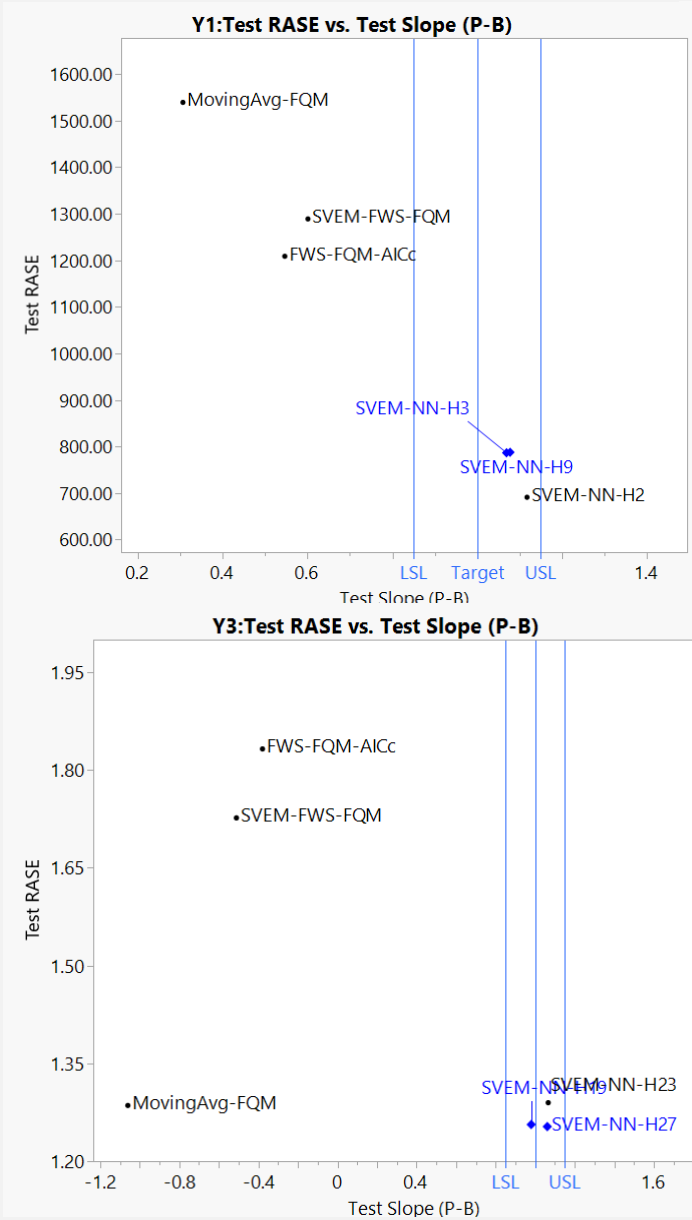
The goal is to see if predictive models fit using the interior SFD points can sufficiently predict the boundary settings of the observations in the DSD.

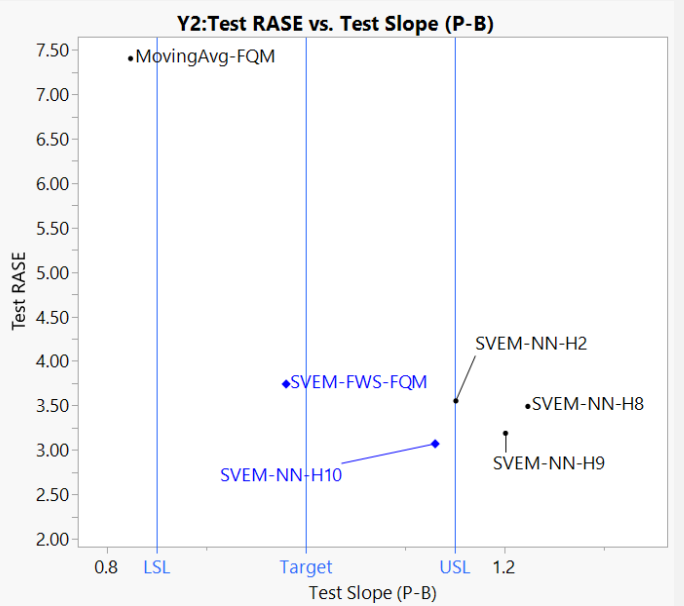
Since the Fit Definitive algorithm only works with the DSD structure it was not included in the analyses performed on the SFD training set.

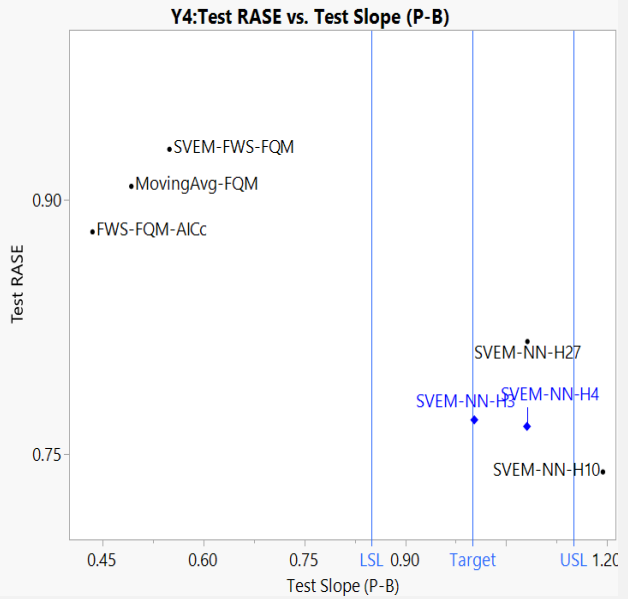
Again, all four responses are studied to understand the impact of complexity on the prediction performance.

The Bio-process Experiment: SFD as Training Set

Model Algorithm	Response	Test RASE	Test Slope (P-B)
FWS-FQM-AICc	Y2	7.88	0.76
SVEM-FWS-FQM	Y2	3.74	0.98
MovingAvg-FQM	Y2	7.40	0.82
SVEM-NN-H2	Y2	3.55	1.15
SVEM-NN-H8	Y2	3.49	1.22
SVEM-NN-H9	Y2	3.19	1.20
SVEM-NN-H10	Y2	3.07	1.13
FWS-FQM-AICc	Y1	1208.90	0.55
SVEM-FWS-FQM	Y1	1289.16	0.60
MovingAvg-FQM	Y1	1539.64	0.31
SVEM-NN-H2	Y1	690.95	1.12
SVEM-NN-H3	Y1	786.59	1.07
SVEM-NN-H9	Y1	787.65	1.08
FWS-FQM-AICc	Y3	1.83	-0.38
SVEM-FWS-FQM	Y3	1.73	-0.51
MovingAvg-FQM	Y3	1.29	-1.06
SVEM-NN-H19	Y3	1.26	0.98
SVEM-NN-H23	Y3	1.29	1.07
SVEM-NN-H27	Y3	1.25	1.06
FWS-FQM-AICc	Y4	0.88	0.43
SVEM-FWS-FQM	Y4	0.93	0.55
MovingAvg-FQM	Y4	0.91	0.49
SVEM-NN-H3	Y4	0.77	1.00
SVEM-NN-H4	Y4	0.77	1.08
SVEM-NN-H10	Y4	0.74	1.19
SVEM-NN-H27	Y4	0.82	1.08



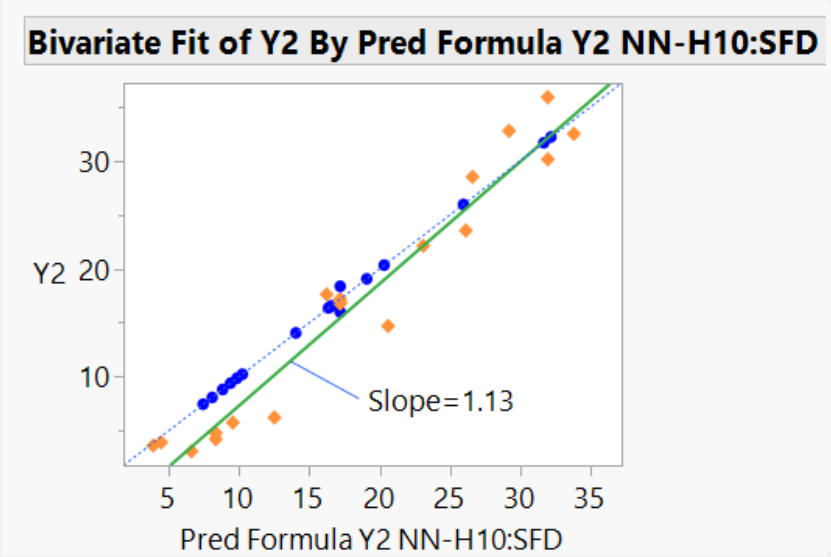


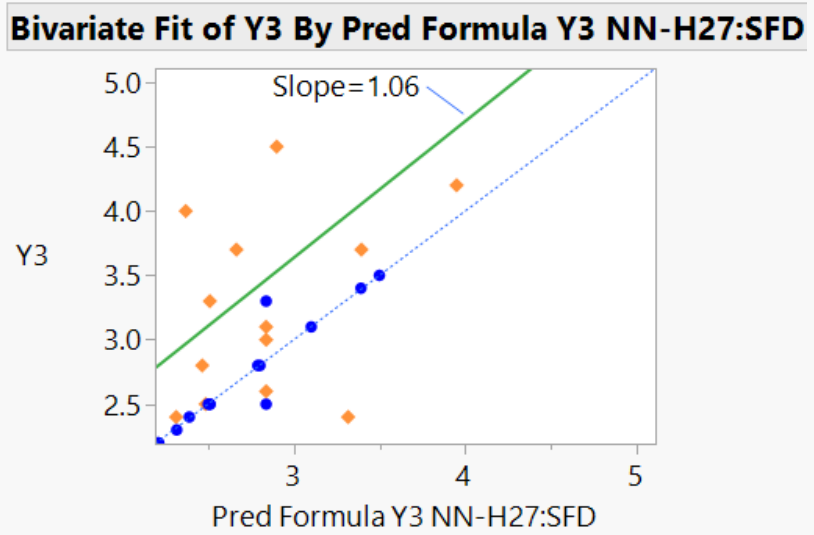


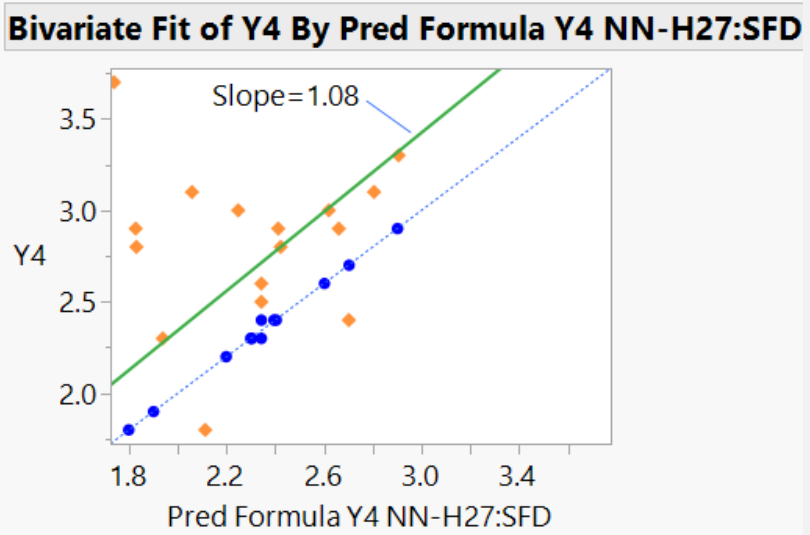
The Bio-process Experiment: SFD as Training Set

Here are shown actual by predicted plots for Y2, Y3, and Y4. For the low complexity Y2 response the DSD and SFD observations align fairly well. Interior behavior is predictive of boundary behavior.

For the high and very high complexity responses Y3, Y4 the DSD and SFD observations do not align. The interior region is not predictive of the boundary region







● SFD ♦ DSD

The Bio-process Experiment: Holdback Test Set

Given there are 35 total observations to work with an attempt was made to use a typical holdback test set approach.

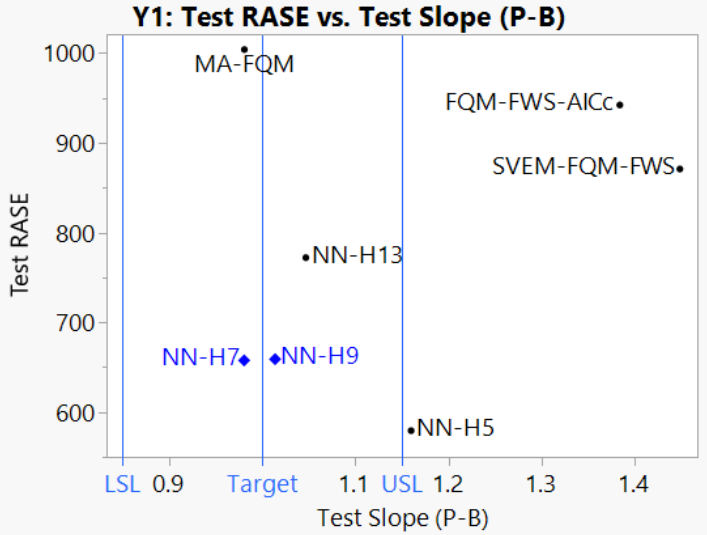
The data were partitioned into an 18-run **training set** and 17-run **test set**. The partitions were created such that the training and test sets had similar portions of DSD and SFD runs, and both sets equivalently covered the design space.

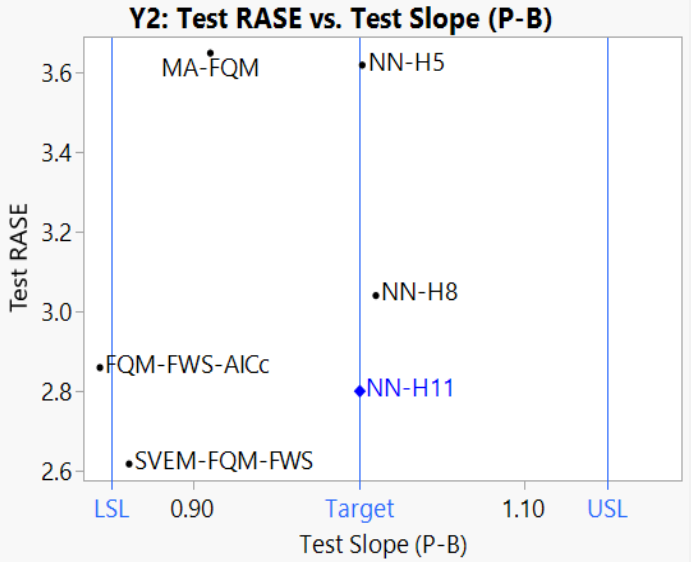
Predictive models were fit to the training partition and then applied to the test partition to evaluate prediction performance.

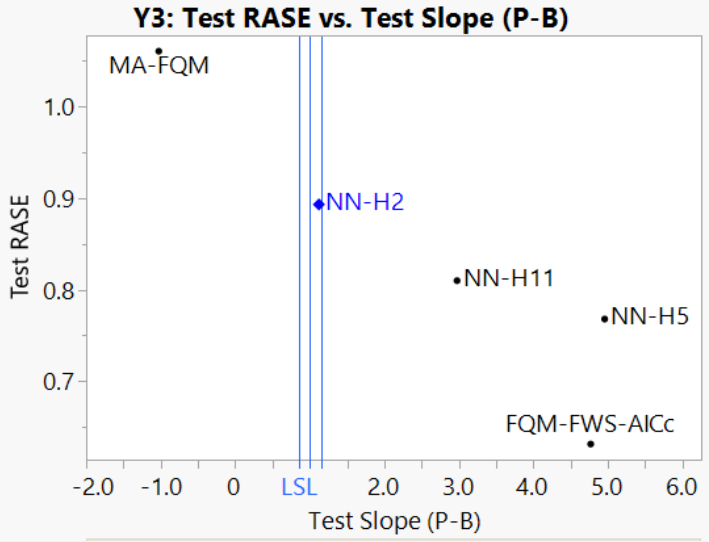
Again, the Fit Definitive method cannot be used as the training set does not have a DSD structure.

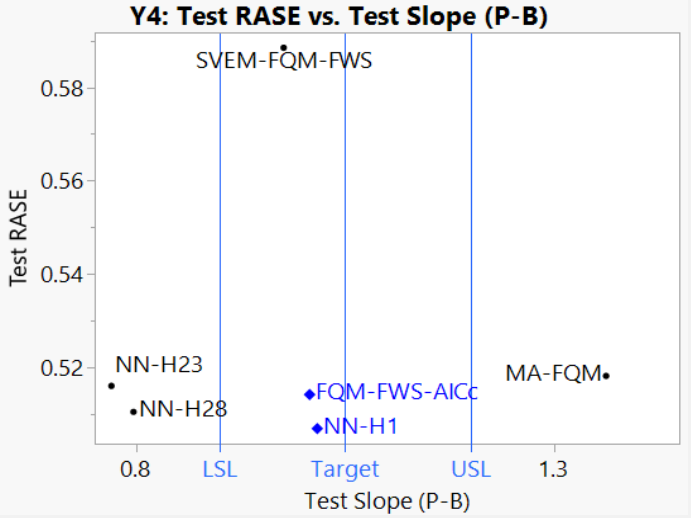
The Bio-process Experiment: Holdback Test Set

Model Algorithm	Response	Test RASE	Test Slope (P-B)
FQM-FWS-AICc	Y1	942.3498	1.3837
SVEM-FQM-FWS	Y1	870.9791	1.4481
MA-FQM	Y1	1004.1482	0.9810
NN-H5	Y1	579.3675	1.1598
NN-H7	Y1	657.6862	0.9808
NN-H9	Y1	659.0485	1.0139
NN-H13	Y1	772.1819	1.0469
FQM-FWS-AICc	Y2	2.8606	0.8432
SVEM-FQM-FWS	Y2	2.6188	0.8608
MA-FQM	Y2	3.6497	0.9099
NN-H5	Y2	3.6195	1.0018
NN-H8	Y2	3.0412	1.0100
NN-H11	Y2	2.8015	1.0003
FQM-FWS-AICc	Y3	0.6320	4.7682
SVEM-FQM-FWS	Y3	0.9051	-2.2742
MA-FQM	Y3	1.0594	-1.0264
NN-H2	Y3	0.8927	1.1222
NN-H5	Y3	0.7680	4.9556
NN-H11	Y3	0.8098	2.9753
FQM-FWS-AICc	Y4	0.5142	0.9574
SVEM-FQM-FWS	Y4	0.5885	0.9262
MA-FQM	Y4	0.5182	1.3124
NN-H1	Y4	0.5069	0.9666
NN-H23	Y4	0.5160	0.7200
NN-H28	Y4	0.5105	0.7464





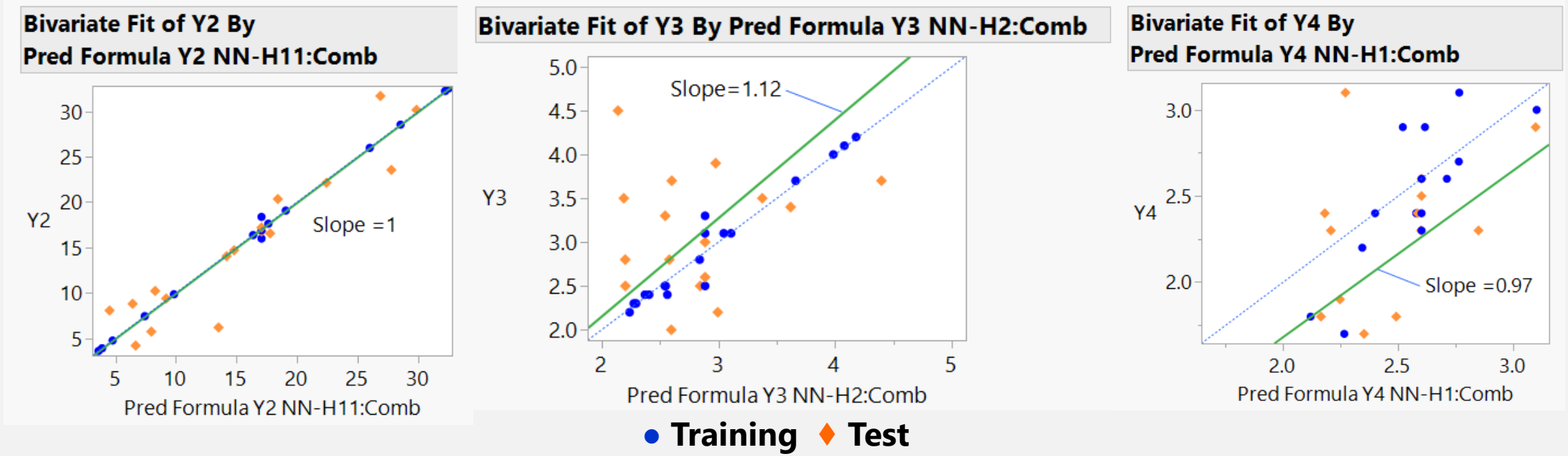




The Bio-process Experiment: Holdback Test Set

Here are shown actual by predicted plots for Y2, Y3, and Y4. For the low complexity Y2 response training set model is predictive of test behavior; the model interpolates.

Again, for the high complexity responses models fit to the training set are not predictive of the test set; the models do not interpolate. The result suggests one needs complete coverage of the design space to build predictive models Y3, Y4.



The Bio-process Experiment: Full Data Set

The final approach to building predictive models is to use the entire data set and forgo the test set.

Since we use the full data **K-fold cross validation** was added to the Forward Selection and to Neural Network modeling; the non-SVEM versions.

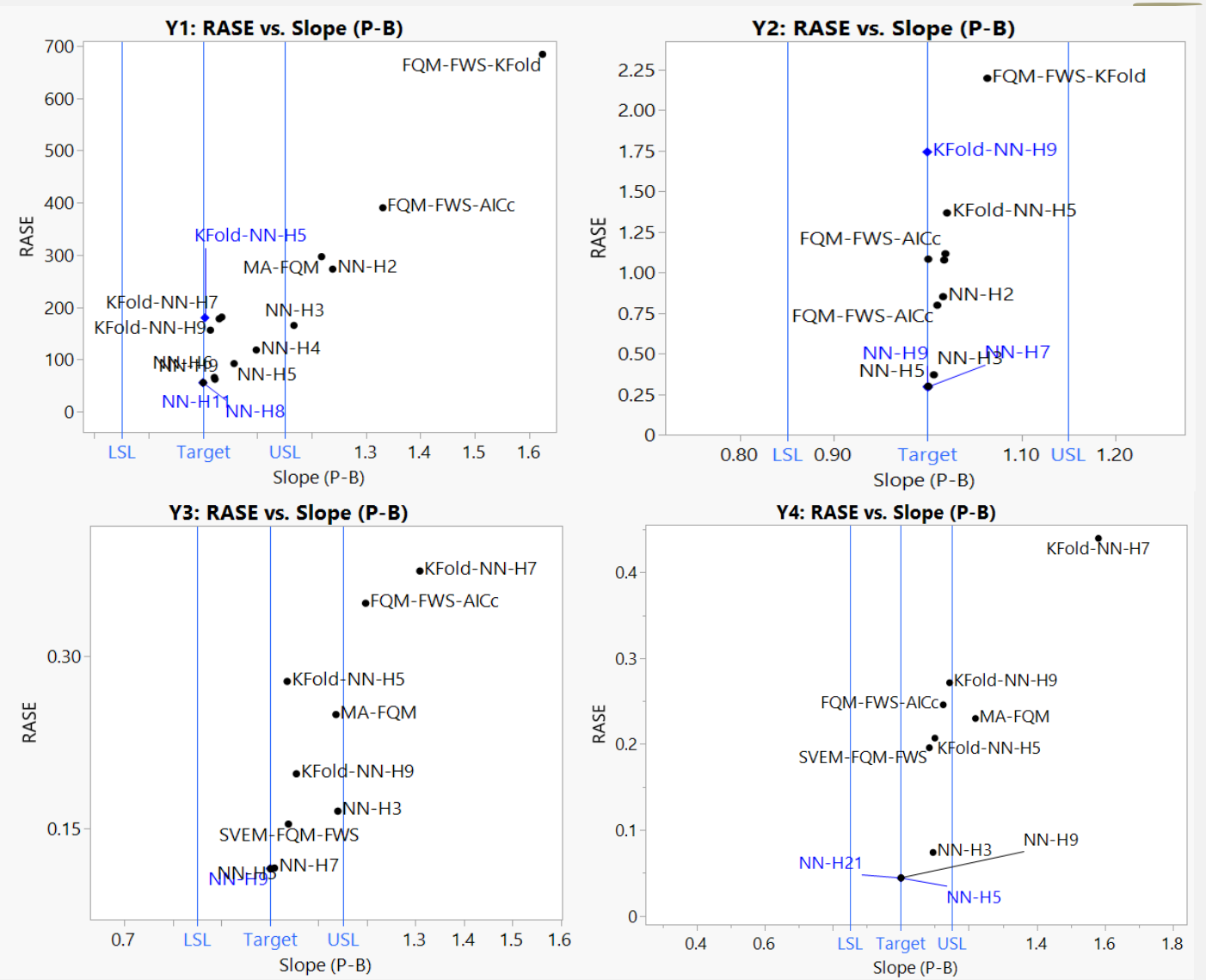
Although the approach is quite common in DoE it is not without risk as one has no direct measure of how well the models may interpolate over the design space.

However, given responses Y3 and Y4 exhibit considerable complex kinetic behavior over the design space, it is desirable to use all the data covering as much of the design space as possible with the experiment as performed.

The fact that 16 of the observations form a space filling design we have some evidence of how well the models might interpolate within the interior of the region.

The Bio-process Experiment: Full Data Set

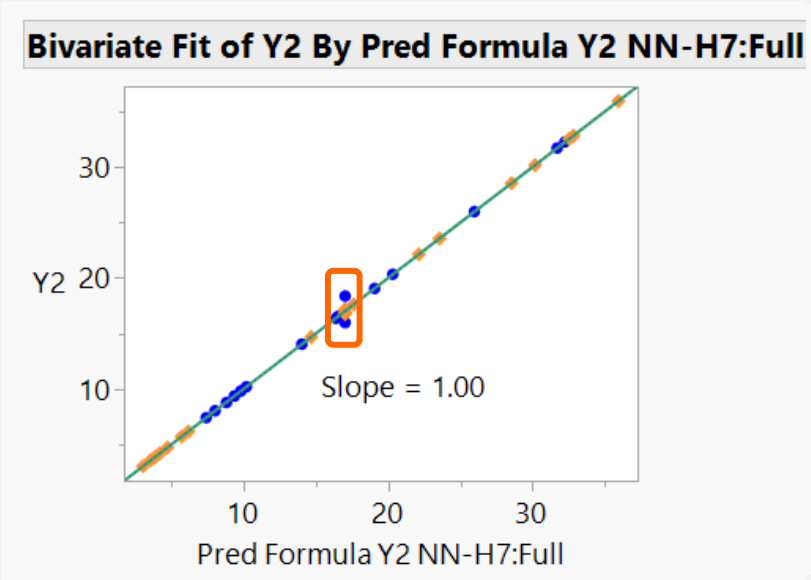
Model Algorithm	Response	RASE	Slope (P-B)	Intercept (P-B)
FQM-FWS-AICc	Y1	390.3875	1.3313	-785.7426
FQM-FWS-KFold	Y1	683.8600	1.6251	-1510.2000
SVEM-FQM-FWS	Y1	177.8926	1.0302	-87.3208
MA-FQM	Y1	296.8619	1.2184	-489.3009
KFold-NN-H5	Y1	179.9468	1.0039	8.1246
KFold-NN-H7	Y1	181.1458	1.0348	-78.5400
KFold-NN-H9	Y1	156.2522	1.0139	-19.8650
NN-H2	Y1	272.9519	1.2389	-507.1969
NN-H3	Y1	165.2965	1.1676	-373.4326
NN-H4	Y1	118.3173	1.0983	-244.8764
NN-H5	Y1	92.3041	1.0575	-159.7577
NN-H6	Y1	65.7075	1.0212	-53.7683
NN-H7	Y1	62.3799	1.0223	-59.4397
NN-H8	Y1	55.7469	1.0002	-0.6321
NN-H9	Y1	55.8185	1.0007	-1.8057
NN-H10	Y1	57.1456	1.0009	-3.4673
NN-H11	Y1	55.6238	1.0003	-0.9045
NN-H13	Y1	55.7820	1.0014	-2.9586
FQM-FWS-AICc	Y2	0.7977	1.0103	-0.3376
FQM-FWS-KFold	Y2	2.1964	1.0636	-0.6148
FQM-FWS-AICc	Y2	1.1152	1.0189	-0.2374
MA-FQM	Y2	1.0764	1.0176	-0.0769
KFold-NN-H5	Y2	1.3668	1.0206	-0.6442
KFold-NN-H7	Y2	1.0815	1.0006	0.0102
KFold-NN-H9	Y2	1.7423	0.9996	0.0107
NN-H2	Y2	0.8505	1.0164	-0.5584
NN-H3	Y2	0.3690	1.0066	-0.1703
NN-H4	Y2	0.2969	1.0003	-0.0239
NN-H5	Y2	0.2986	1.0006	-0.0142
NN-H6	Y2	0.2944	1.0002	-0.0047
NN-H7	Y2	0.2936	1.0002	-0.0026
NN-H9	Y2	0.2944	1.0001	-0.0037

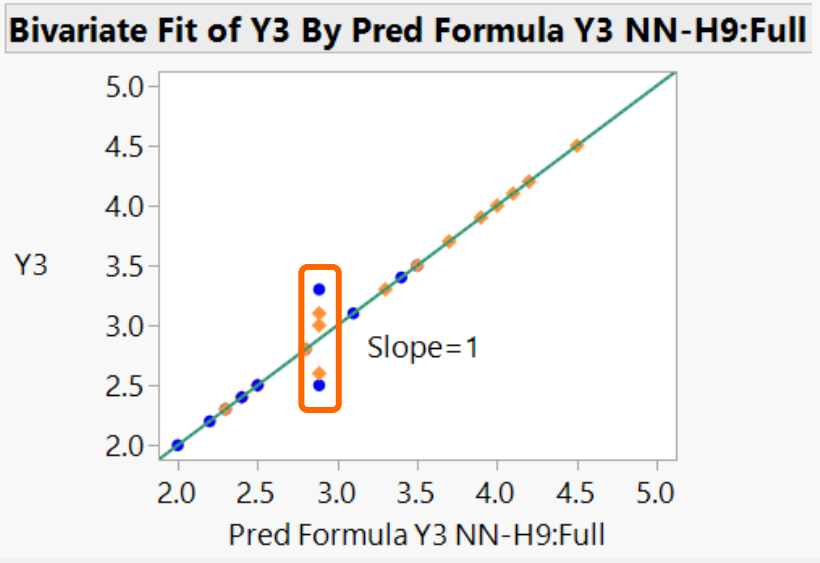


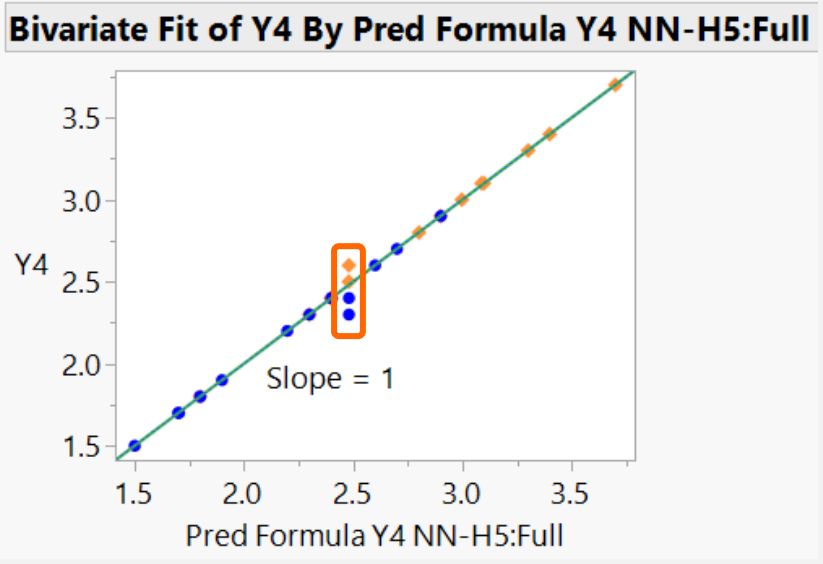
The Bio-process Experiment: Full Data Set

Here are shown actual by predicted plots for the best models for each response. In all three cases a SVEM NN model fits best.

In this case the DSD and SFD points are aligned for all three models and one can see that the observed center point values for the three responses are close in value between the DSD and SFD







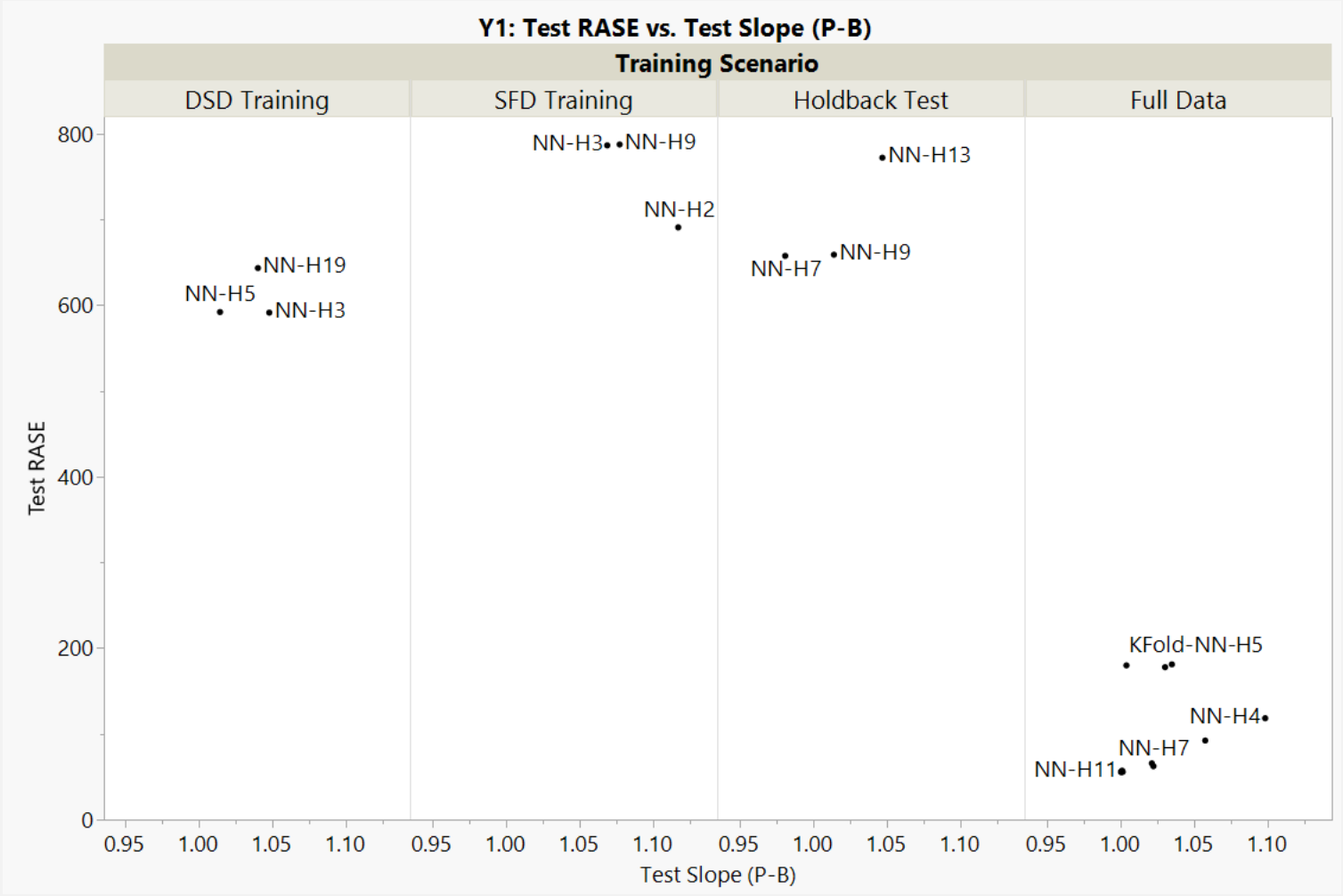
● SFD ♦ DSD

The Bio-process Experiment: Final Results Y1

Here is shown a display of the final best model results for all training scenarios and model algorithms for Y1; moderate complexity

The SVEM NN algorithms performed the best across all scenarios.

As expected, the best RASE scores are for the Full Data as there is no test set and the RASE scores are calculated for the training data.

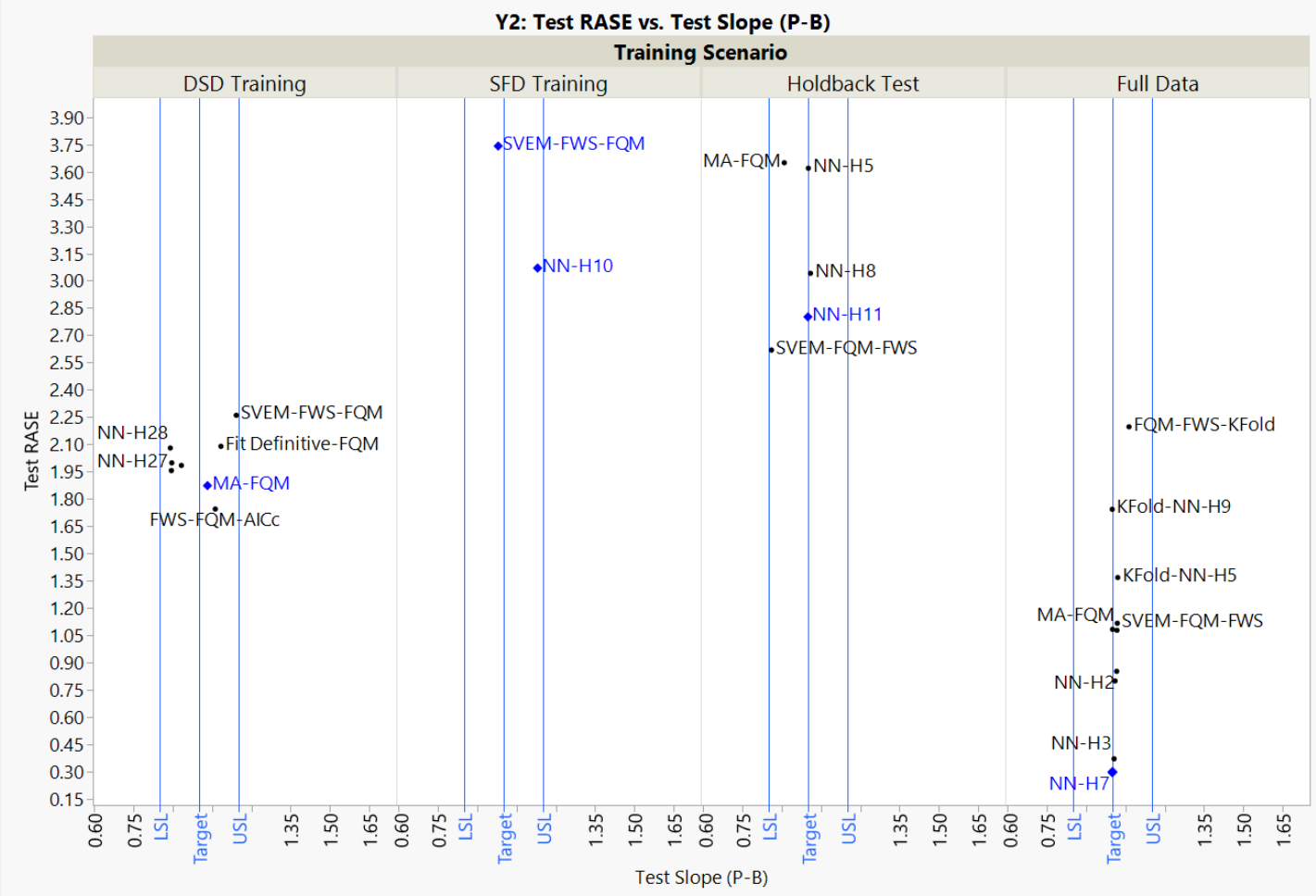


The Bio-process Experiment: Final Results Y2

Here is shown a display of the final best model results for all training scenarios and model algorithms for Y2; the lowest complexity.

The SVEM NN algorithms performed the well across all scenarios, however other algorithms did as well in many cases.

Again, the RASE scores for the full data are training set values so they should be lower.

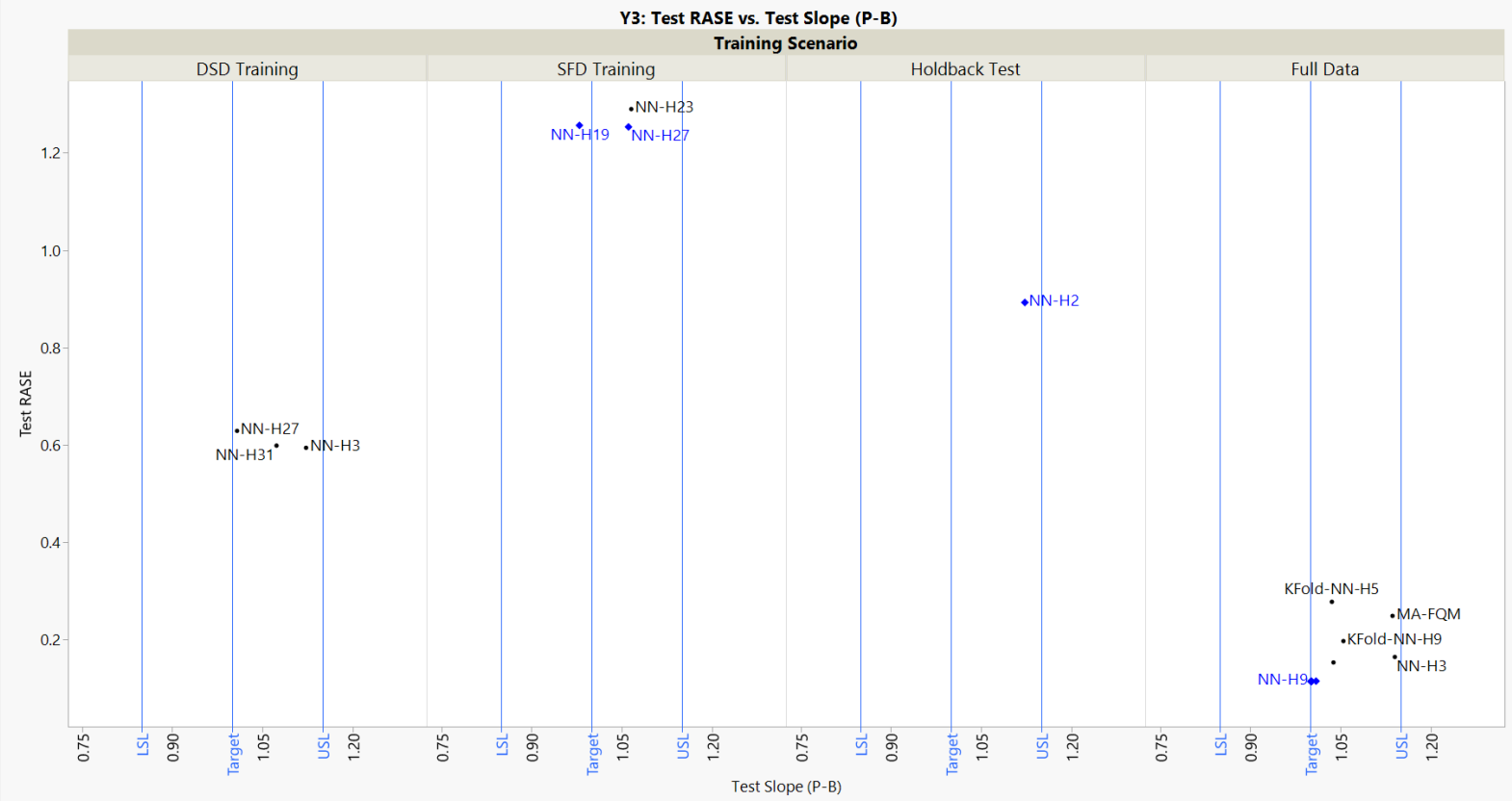


The Bio-process Experiment: Final Results Y3

Here is shown a display of the final best model results for all training scenarios and model algorithms for Y3; high complexity.

The SVEM NN algorithms performed the best across all scenarios

Again, the RASE scores for the full data are training set values so they should be lower.

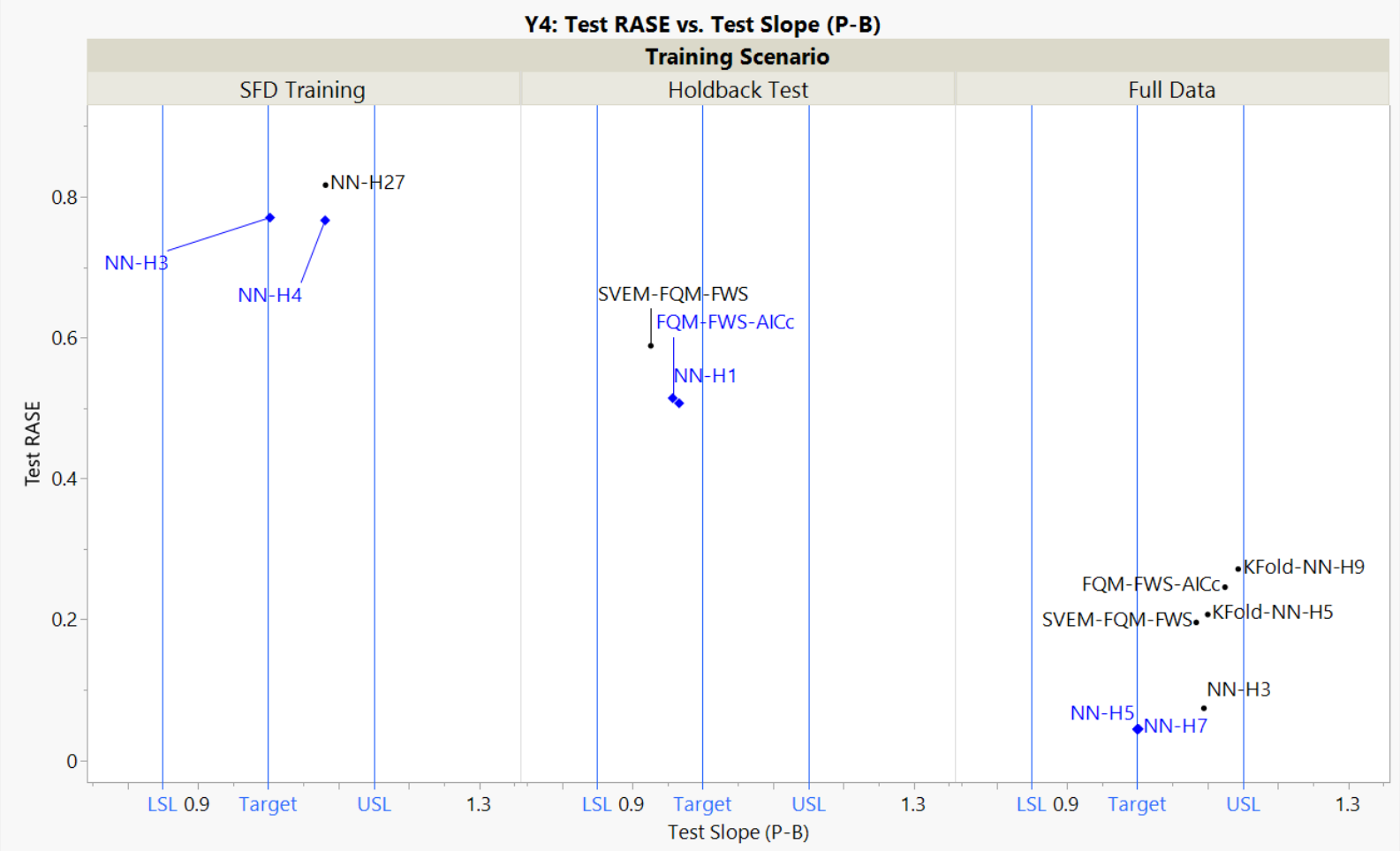


The Bio-process Experiment: Final Results Y4

Here is shown a display of the final best model results for all training scenarios and model algorithms for Y4; very high complexity.

No models for the DSD Training scenario resulted in acceptable slopes. The response is difficult to model. The SVEM NN models tended to do the best.

Again, the RASE scores for are lower.



Summary

We evaluated multiple model building algorithms and four test set strategies to build predictive models for four responses from a bio-process experiment.

The fourth strategy was to use the complete data set for modeling and forgo a test set; not generally recommended for predictive model building.

In this case the presence of 16 space filling design points in the experiment supplied evidence of whether a model might interpolate well over the design space.

The responses varied from low complexity (Y2) to very high complexity (Y4).

A machine learning method SVEM combined with neural network (NN) modeling was evaluated over the four strategies and compared to other algorithms.

Overall, the combination of SVEM and NN provided the best predictive models, especially for the more complex responses.

Summary

Over the various predictive modeling techniques, full quadratic models (FQM) performed poorly compared to NN models on the complex responses (Y1, Y3, Y4).

Prediction is about interpolation; test sets are necessary to assess performance.

From the predictive modeling strategies with a test set it is observed that the DSD and SFD points exhibit different kinetic behavior for the more complex responses.

For the low complexity response Y2, the DSD and SFD points appear to aligned.

The implication is that kinetic behavior in complex systems may change substantially from the design center to boundaries; near boundaries often little data exists.

Operating regions (e.g., NORs) should avoid boundary areas; too much uncertainty.

The traditional boundary designs used in statistics for response surface analysis are inadequate for complex systems as no data is available on internal behavior.

Final Thoughts on Complexity, DoE, and Modeling

We are in the age of digital science; digital chemistry and digital biology are two prominent examples.

Systems theory and thinking about of complex systems is ascendent in science; 19th Century reductionism is in decline; its all about the interactivity.

In addition is the increase in both process and laboratory automation and new, affordable sensor technologies enabling evermore complex experiments.

Machine learning (e.g., SVEM) and deep learning algorithms are necessary to model complex system behavior.

Yet, DoE for response surface methods are mostly unchanged from Box and Wilson (1951) and certainly the1960s; new thinking is needed.

Boundary designs analyzed by FQMs dominate statistical approaches; both are inadequate for complex systems.

Final Thoughts on Complexity, DoE, and Modeling

Complex system behavior is driven by the interactive behavior of the inputs, therefore experimentation on complex systems requires designs incorporating large numbers of factors while maintaining reasonable run sizes.

New design criteria for complex systems in the digital era are at a minimum:

- Accommodate large number of factors with a reasonable number of runs.
- Cover the interior of the design space to capture the dynamic, kinetic behavior.
- Do not require the specification of a model to generate the design.
- Allow user input on the distribution of design points over the design space.
- Easily combined with existing data (e.g., process data) from the system in study.

At present space filling designs is the best option, but much more work is needed.

Final Thoughts on Complexity, DoE, and Modeling

Besides new designs and strategies, machine learning methods need to be adopted .

Traditional practices in DoE need be reconsidered, these include:

- **Screening** to remove factors from future consideration, it is a high-risk strategy; the active and inactive factor paradigm is not true for complex systems.
- With SFDs and machine learning traditional **screening is unnecessary**.
- **Screening** techniques are better used to refine design spaces for future work.
- Stop reflexively **reducing linear models**, it degrades prediction performance.
- The **Hierarchy principle** is false for complex systems. The system behavior is driven by interactivity not main effects.
- **Stop the ubiquitous use of FQMs**. The model is too stiff for complex systems.

References

1. D.B. White, H.M. Faessel, H.K. Slocum, L. Khinkis, and W.R. Greco, "Nonlinear Response Surface and Mixture Experiment Methodologies Applied to the Study of Synergism." *Biometrical Journal* 46, no. 1 (February 2004): 56–71.
2. P. Ramsey and C. Gotwalt, "Model Validation Strategies for Designed Experiments Using Bootstrapping Techniques with Applications to Biopharmaceuticals" (PowerPoint presentation, JMP Discovery Summit Americas, October 25, 2018).
3. C. Gotwalt, L. Xu, Y. Hong, and W.Q. Meeker, "Applications of the Fractional-Random-Weight Bootstrap" (research paper, ResearchGate GmbH, 2018).

References

4. T. Lemkus, P. Ramsey, C. Gotwalt, and M. Weese, “Self-Validated Ensemble Models for Design of Experiments,” (research paper, *Chemometrics and Intelligent Laboratory Systems* (2021): Version of Record; early-release article.

5. “Self-Validating Ensemble Modeling,” Predictum Inc., accessed August 12, 2021.

6. L. Breiman, “Heuristics of Instability and Stabilization in Model Selection” research paper, *Annals of Statistics* 24, no. 6 (December 1996): 2350–2383.