Engineering Conferences International
ECI Digital Archives

Cell Culture Engineering XVI | Proceedings

5-9-2018

More accurate process understanding from process characterization studies using Monte Carlo simulation, regularized regression, and classification models

Cary Opel
Gilead Sciences, USA, cary.opel@gilead.com

Cerintha J. Hui
Gilead Sciences, USA

Patrick Y. Yang
Gilead Sciences, USA

Daniel J. Tien
Gilead Sciences, USA

Gayle E. Derfus
Gilead Sciences, USA

See next page for additional authors

Authors
Cary Opel, Cerintha J. Hui, Patrick Y. Yang, Daniel J. Tien, Gayle E. Derfus, and Rajesh Krishnan

More Accurate Process Understanding from Process Characterization Studies Using Monte Carlo Simulation, Regularized Regression, and Classification Models

Cary Opel, Research Scientist II

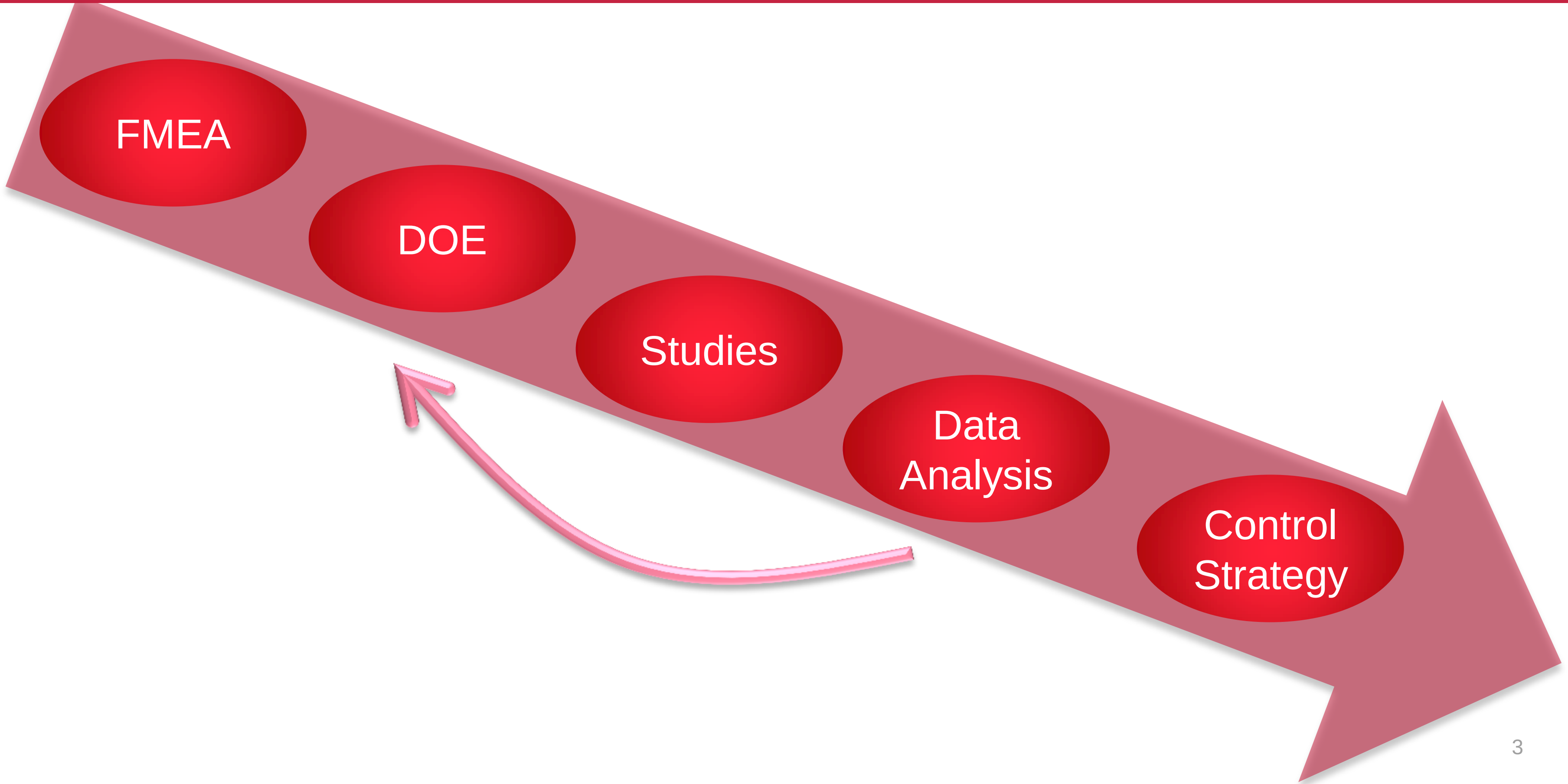
May 9th, 2018

GILEAD

Key Takeaways

- Cross Validation and Monte Carlo techniques can establish accurate CPPs and control strategies that enable a robust manufacturing process.
- Uncertainty affects model outcomes and should be taken into account when making risk-based predictions.
- The best models are created when researchers evaluate the models, not just rely on rules.
- More accurate model construction can make QbD programs more efficient, enable refinement of DOE studies, and inform future programs.

Process Characterization



Regression and Model Selection

- DOE generated data lends itself to linear regression models:

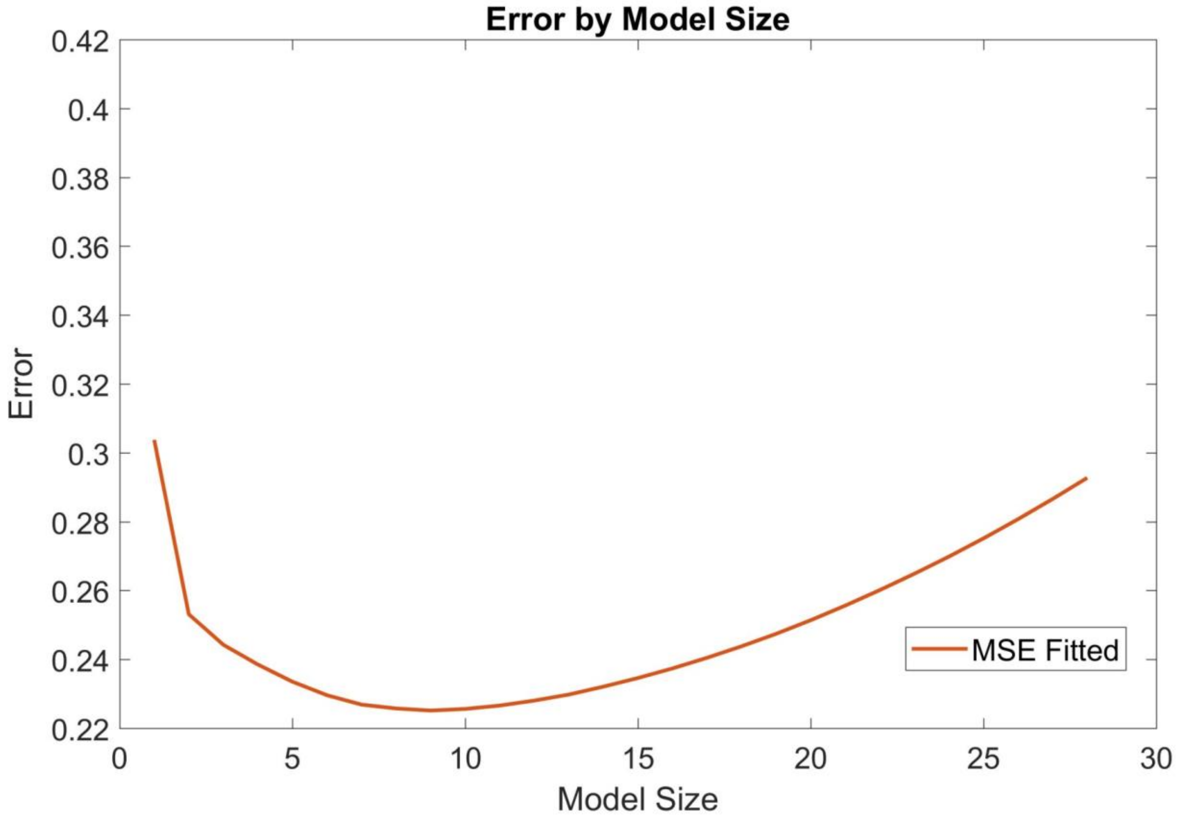
$$\mathrm{y} = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

  - y's are outcomes (e.g. product quality) and x's are parameters (e.g. temperature)
- How to pick the "best" variables to fit the data?
  - Minimize error
  - Avoid over-fitting
- Move from "**descriptive**" analysis to "**predictive**" analysis
  - Mean Squared Error Fitted (MSE Fitted) to Mean Squared Error Predicted (MSE Predicted)

Standard Stepwise Analysis

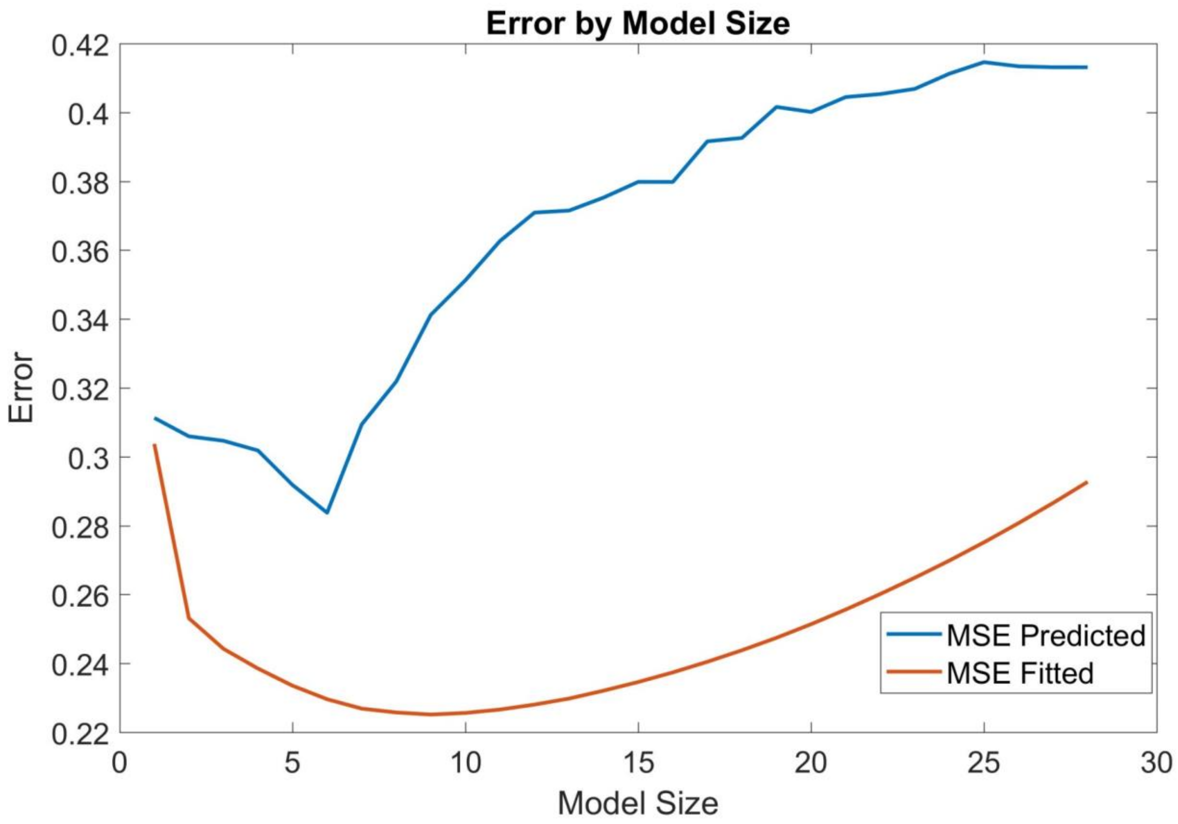
- Emphasis on Rules-Based Model Selection
- Backwards Stepwise
  - Start with all main, interaction, and/or quadratic effects included
  - Eliminate one by one based on single p-Value or AIC/BIC criteria
  - When no more parameters meet the elimination criteria, the model is final
- Impact Assessment
  - A final round of variable elimination is performed based on the magnitude of the effect
  - This is often accomplished by some kind of Impact Ratio
  - For example, aggregates can be significantly impacted by Temperature, but if the change in HMW is ~0.5% over the range studied, should it be considered a CPP?

The Problem with Fitting By Error



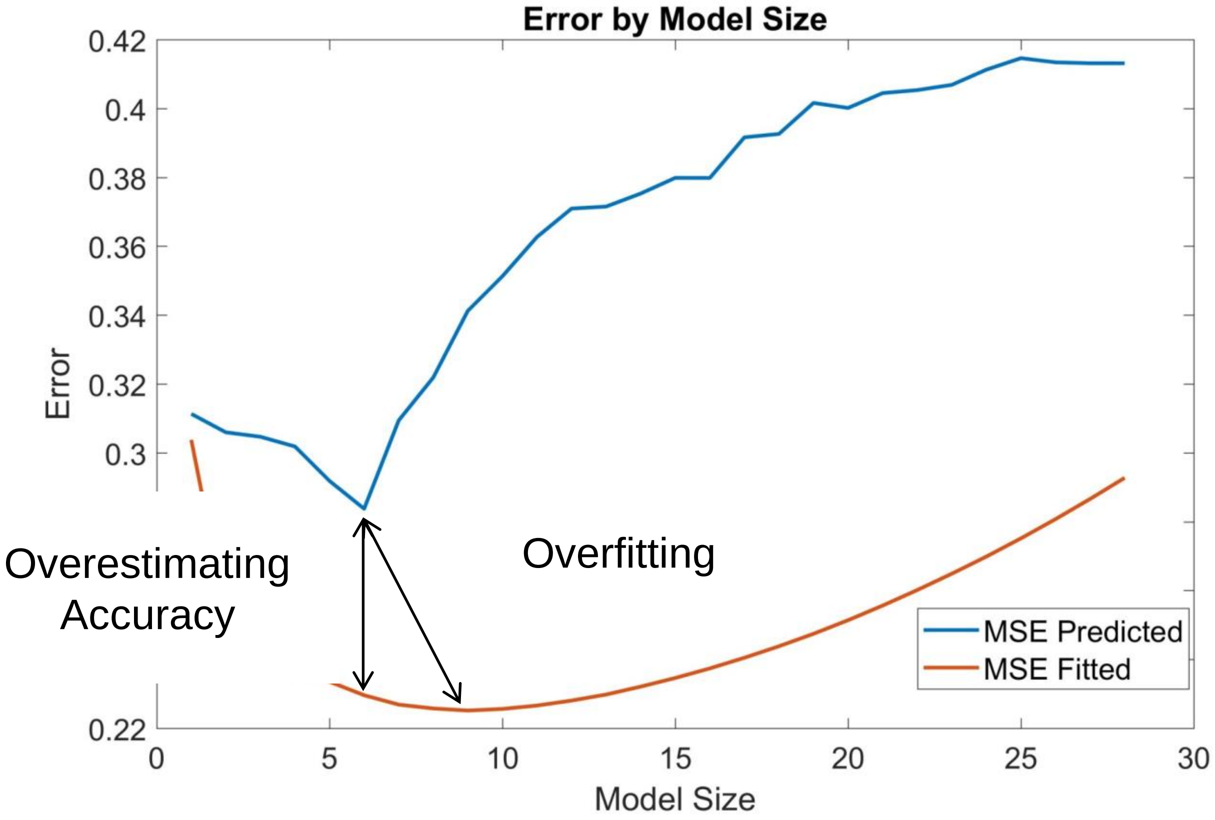
MSE Fitted is the error of the model when used on the data that was used to generate the model itself

The Problem with Fitting By Error



MSE Predicted is the error of the model when used on new data

The Problem with Fitting By Error

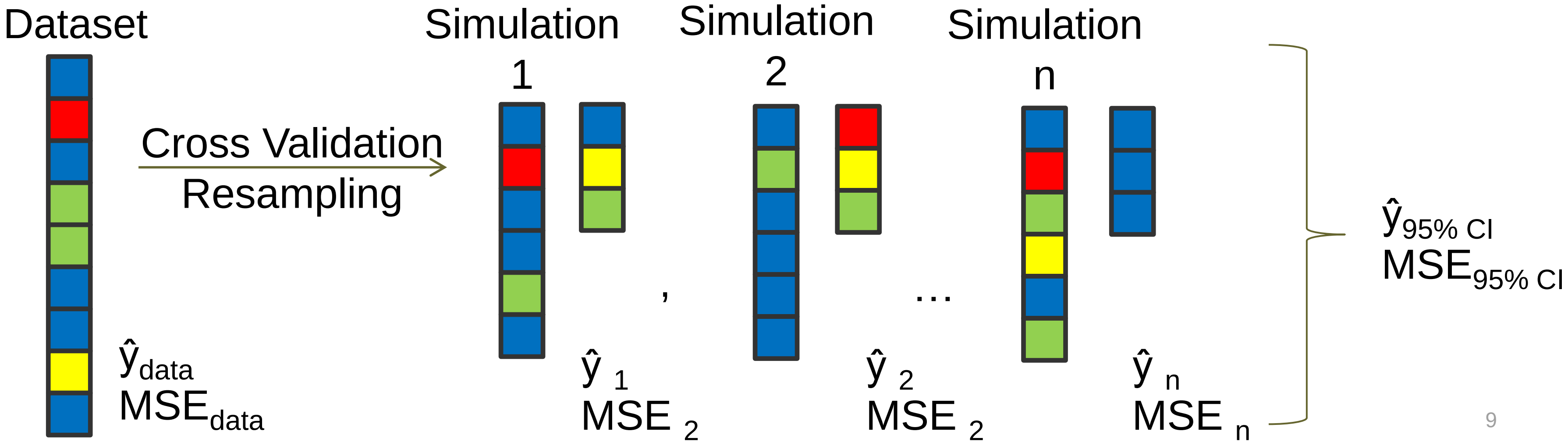


MSE Fitted error both overestimates the accuracy of the model and overfits the data by including too many terms

Monte Carlo / Cross Validation

Algorithm

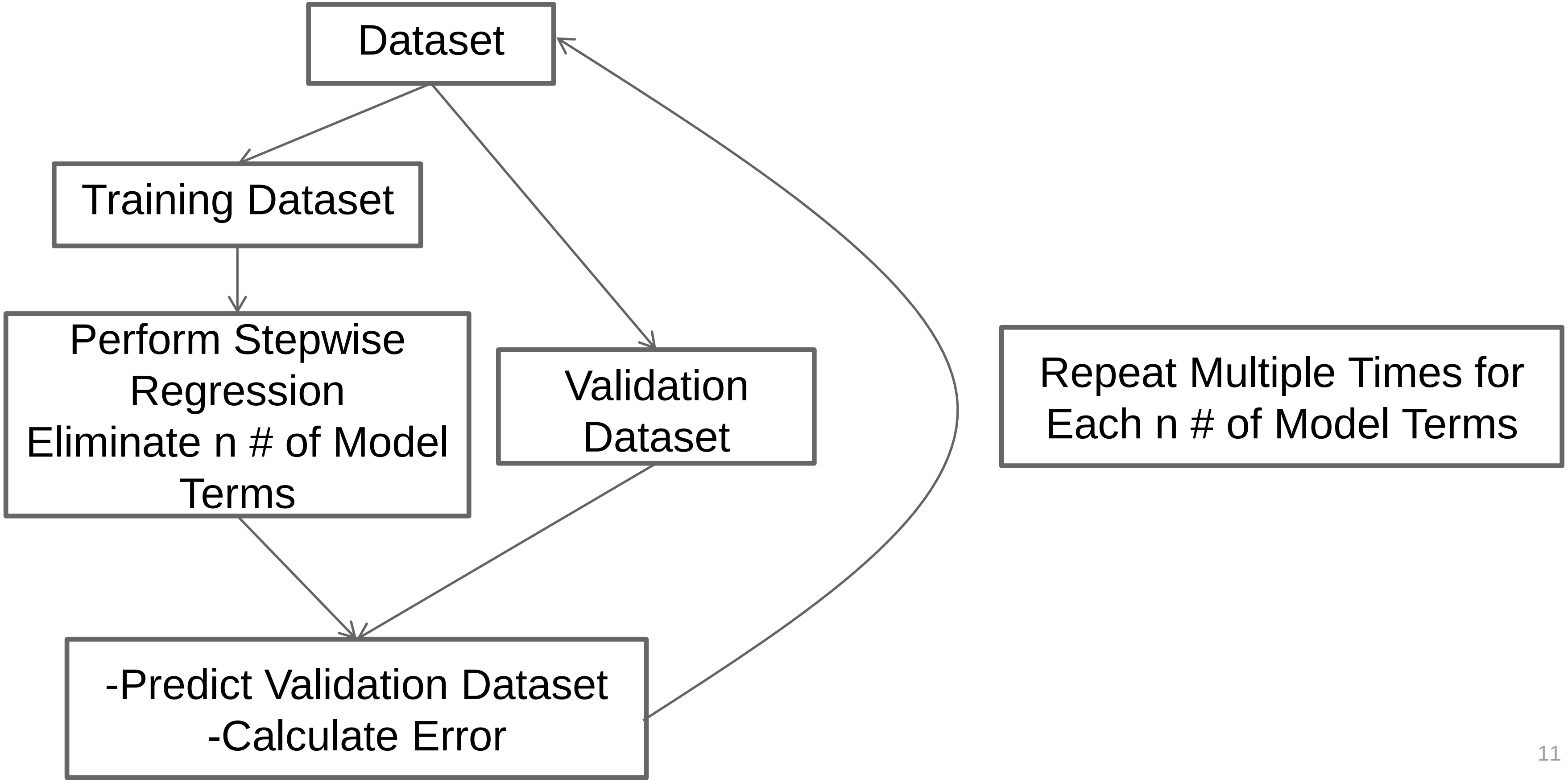
- Generate two data sets
  - Sample subset of data without replacement (Training Set)
  - Set aside the remaining data (Validation Set)
- Build model with Training Set
- Measure model performance on Validation Set



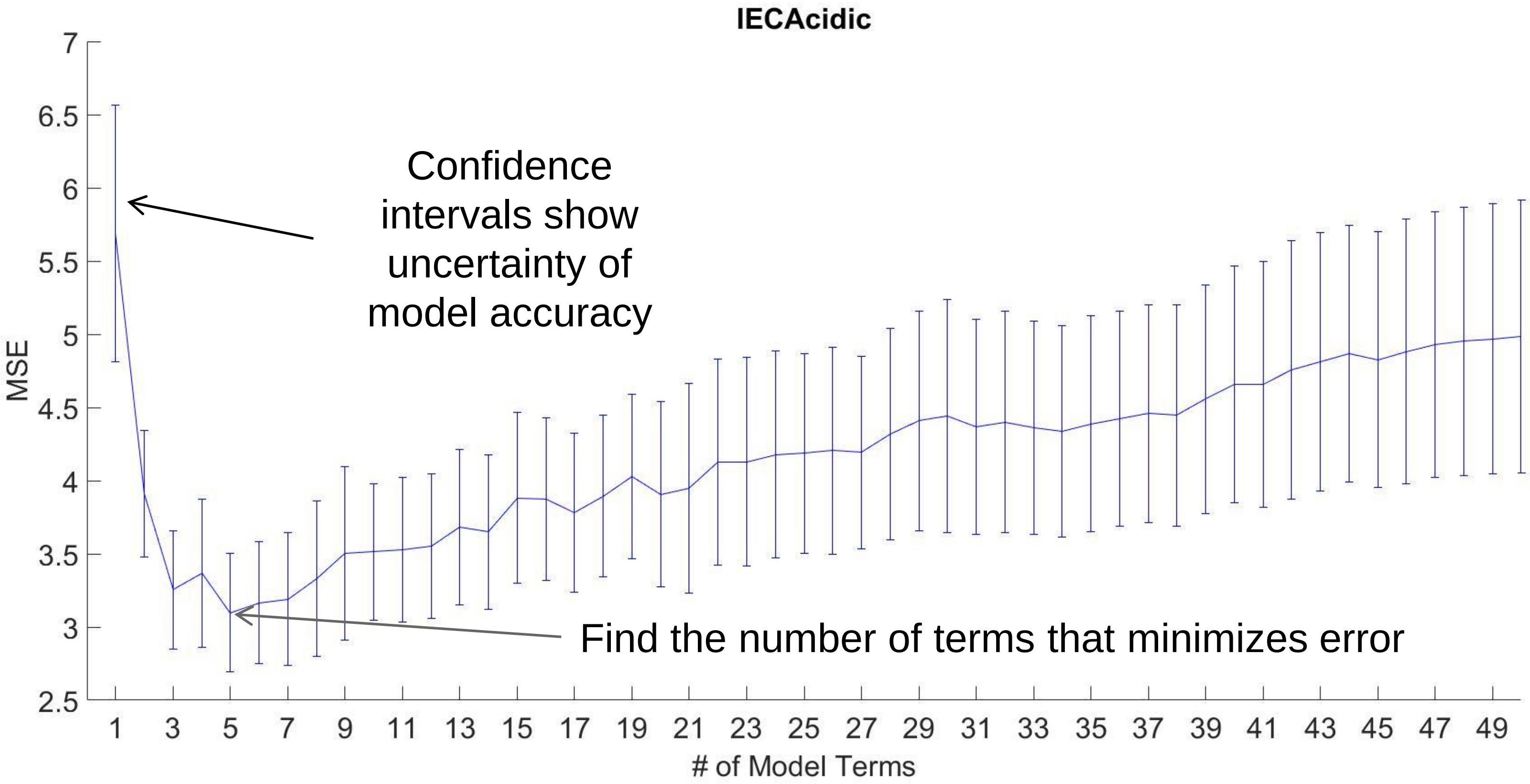
Workflow

- Define Model Size
- Select Process Parameters
- Simulate Product Quality
- Compare Different Models

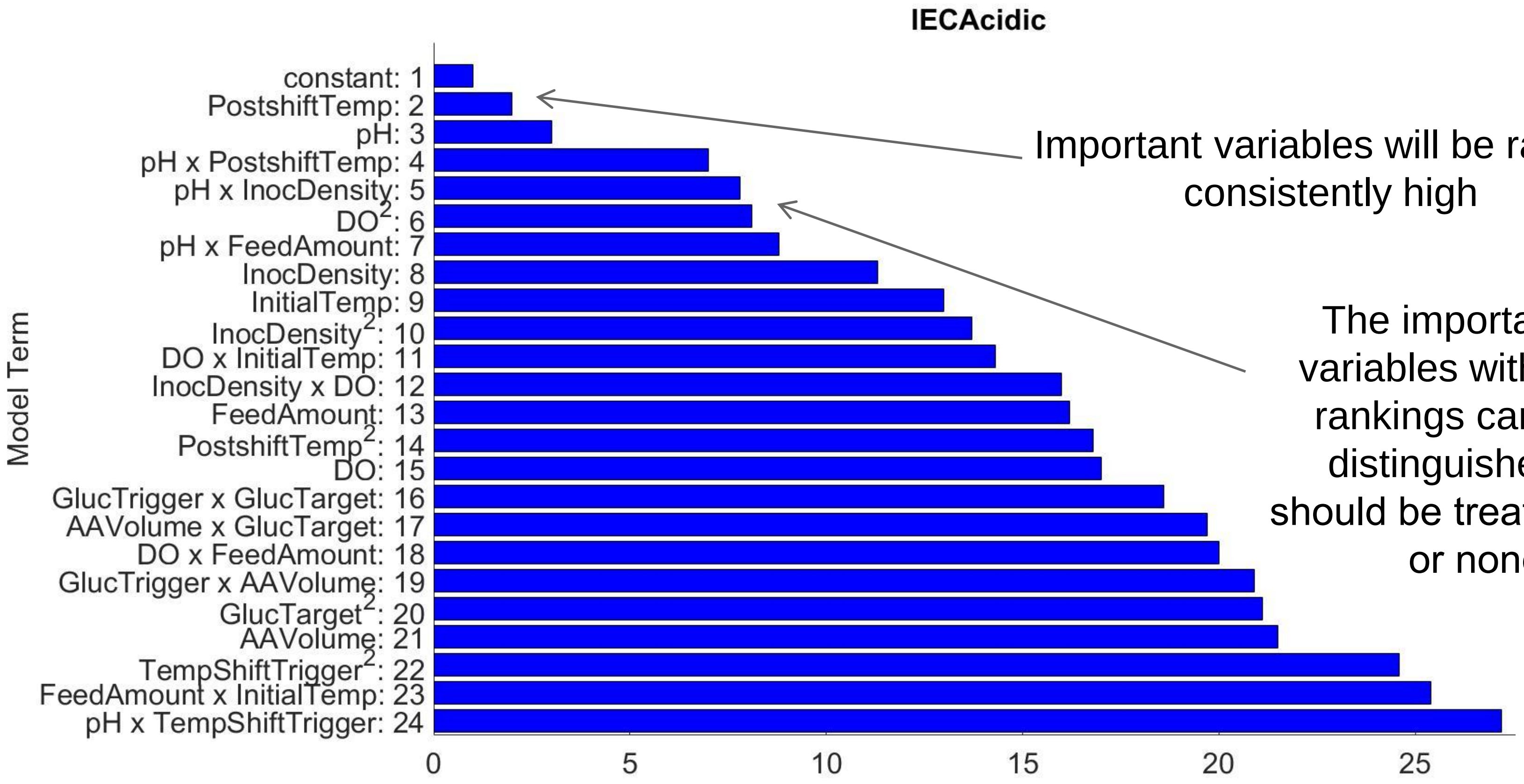
Define Model Size



Define Model Size



Select Variables



Important variables will be ranked consistently high

The importance of variables with similar rankings cannot be distinguished and should be treated as “all or none”

Simulate Product Quality

Randomize Dataset
and Build Model
$$y = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n$$

Generate Random Run
with Conditions Inside
Operating Range

$1<x_1<2$

$3<x_2<5$

$\ldots$

$3<x_n<8$

Calculate
Predicted PQ
and Repeat

$y_1=$

$y_2=$

$\ldots$

$y_n=$

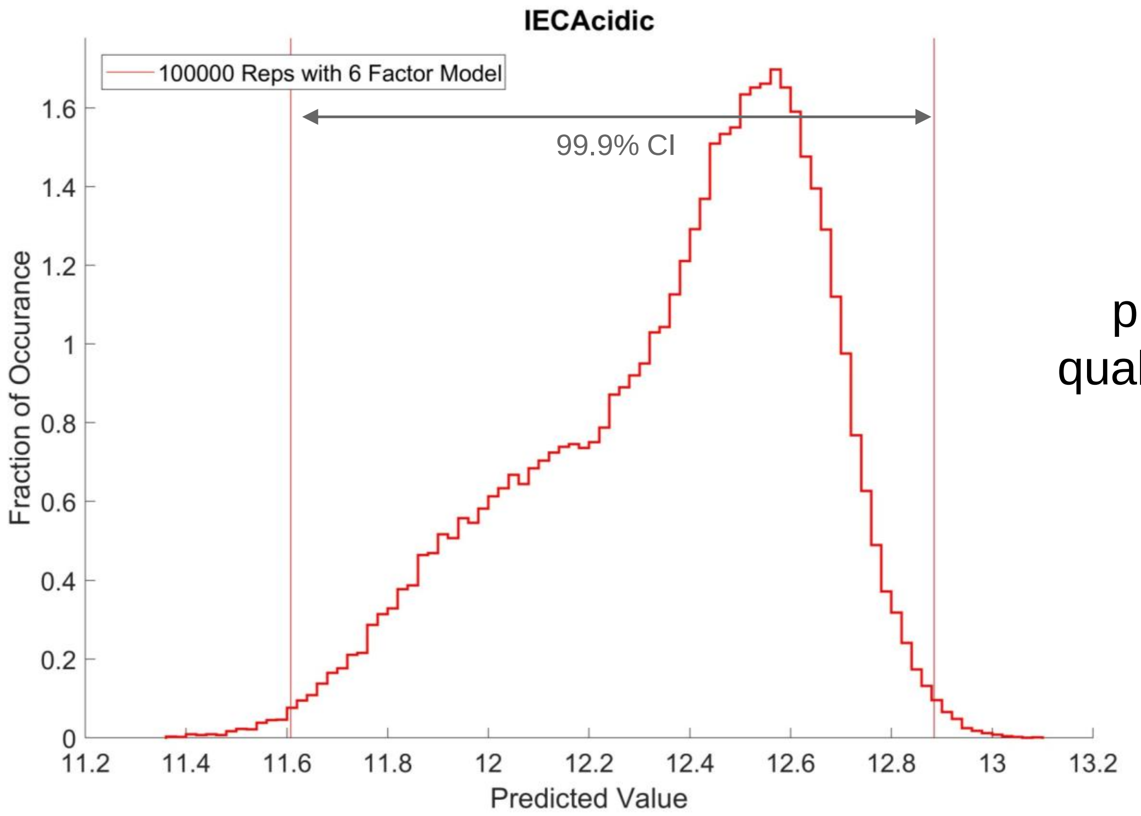
Calculate
Summary
Statistics

Mean

Median

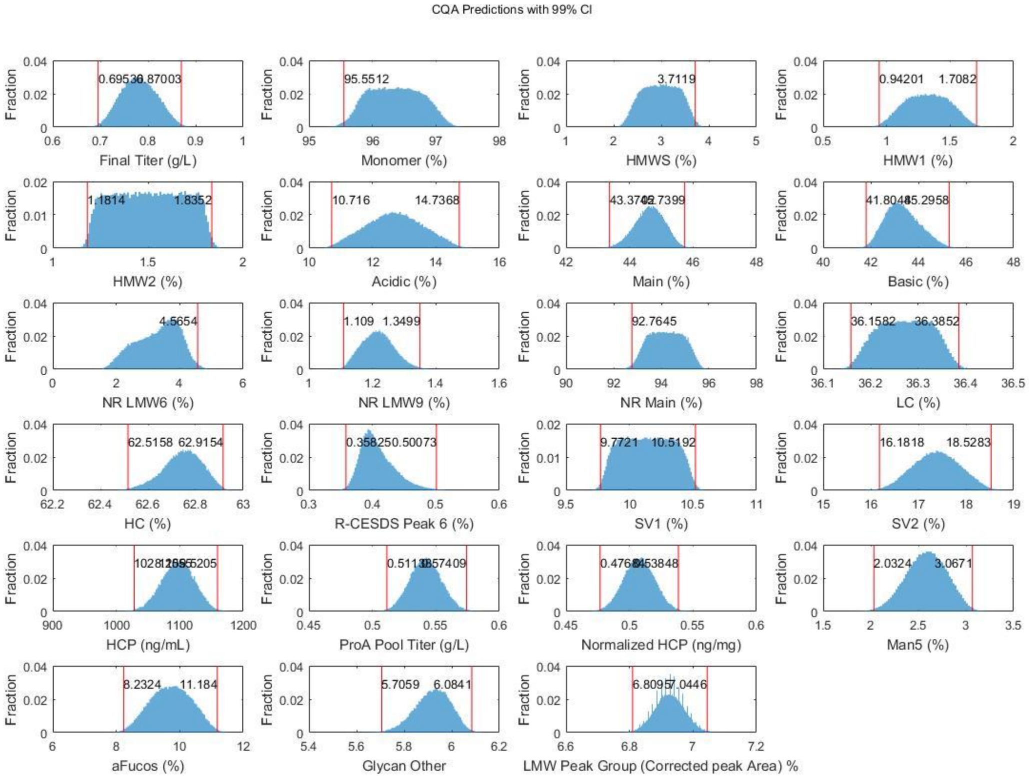
95% CI

Simulate Product Quality



A set of Operating Ranges produces a simulated product quality outcome, with measureable confidence intervals

Simulate Product Quality



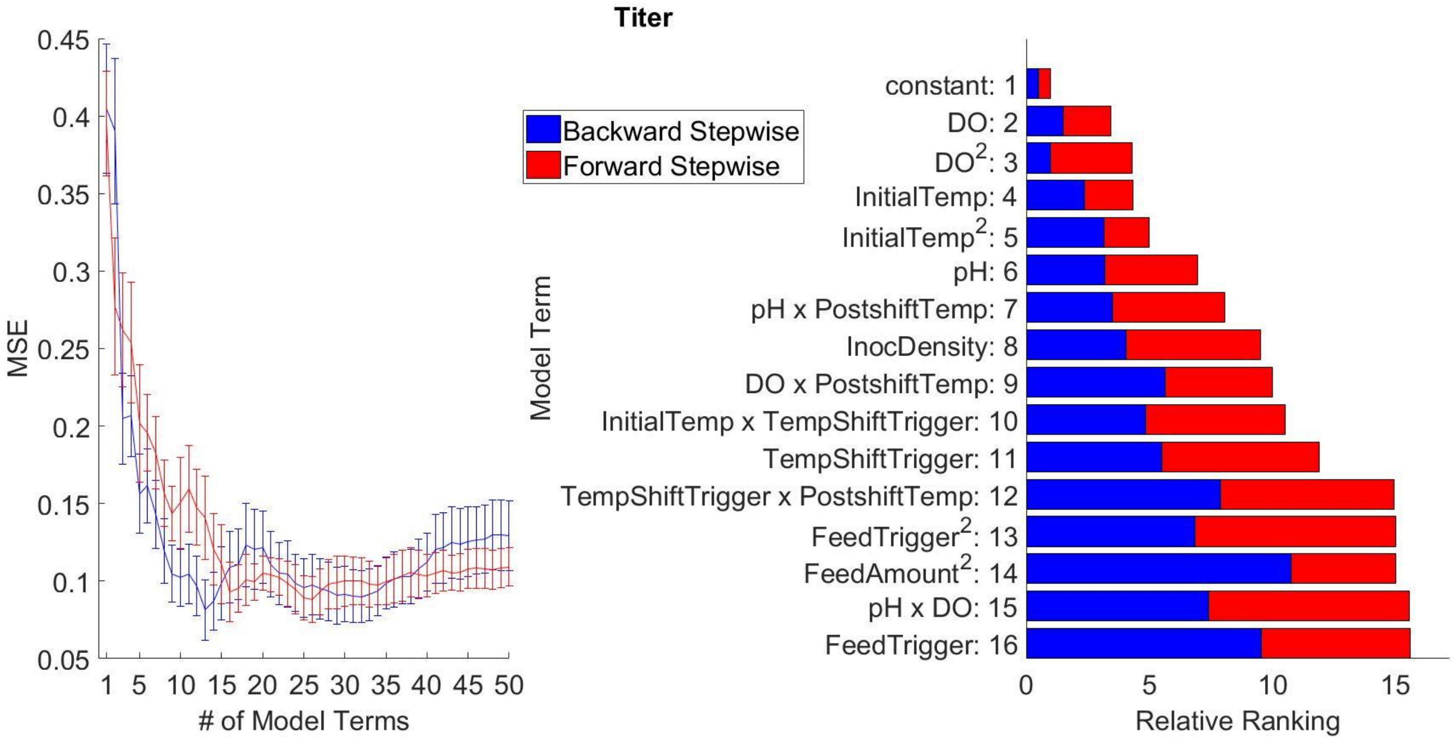
Candidate control strategies can generate simulated quality profiles to allow Operating Ranges to be set

Compare Different Models

- Goals
  - Accurate predictions
  - Clear parameter selection

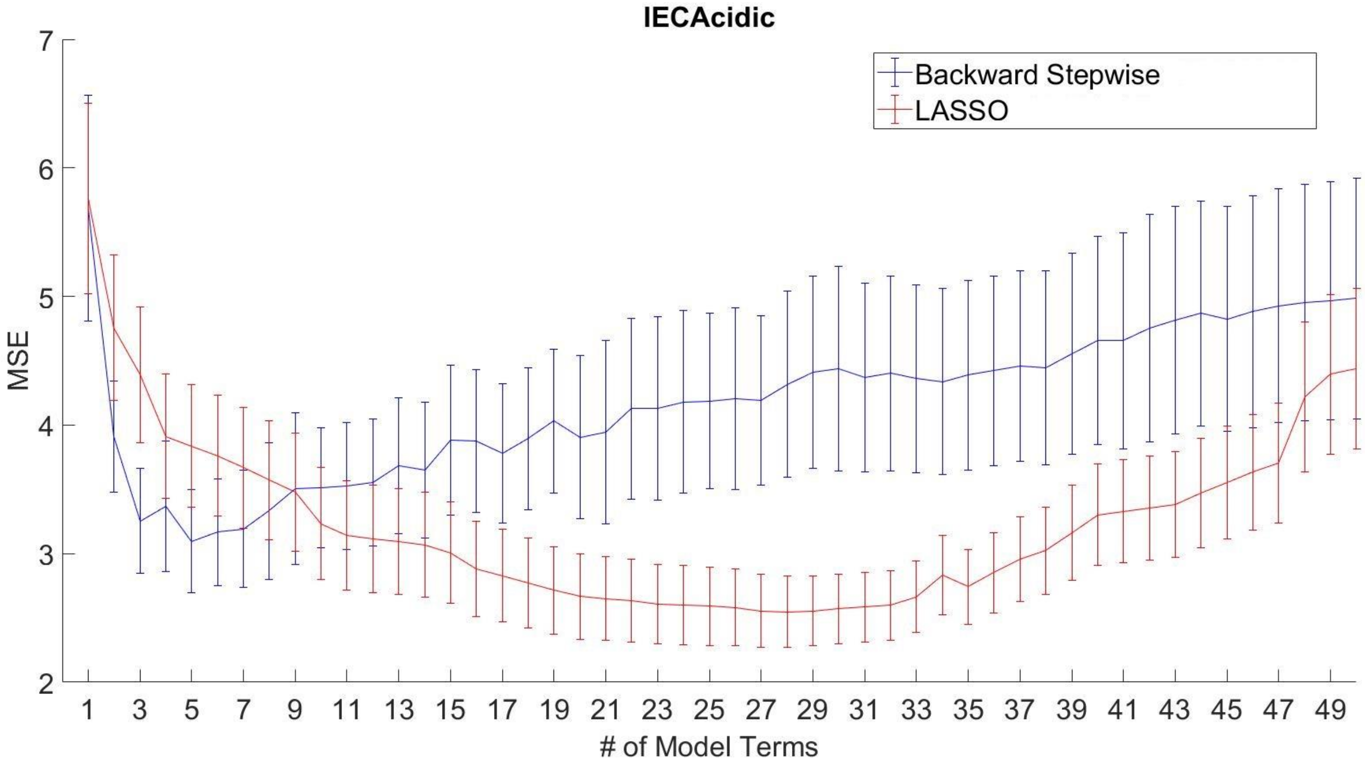
- Models
  - Stepwise Regression
    - Backwards
    - Forwards
  - Regularization
    - LASSO
  - Classification Models
    - Decision Trees

Compare Different Models: Stepwise



Comparing different elimination rules like Forward Stepwise regression can help discriminate borderline significant parameters.

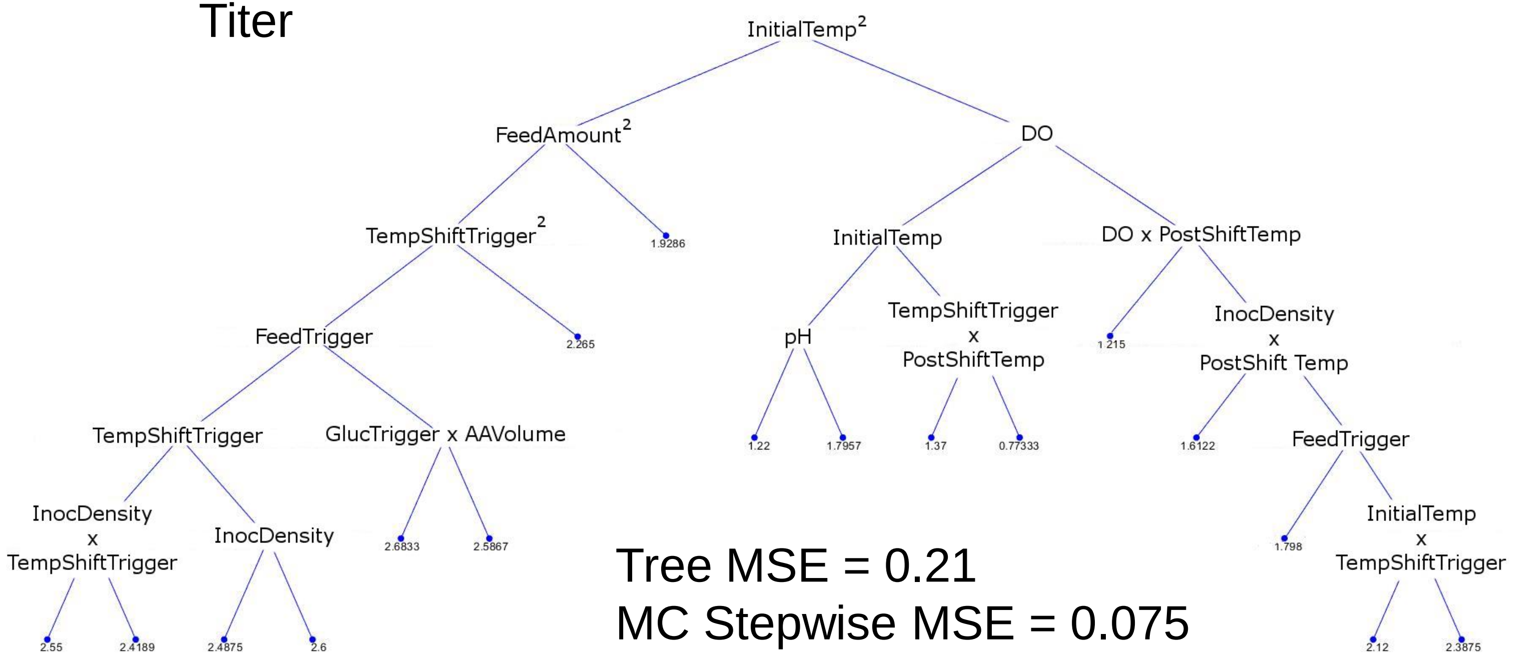
Compare Different Models: LASSO



Regularization methods like LASSO can do a good job minimizing error, but fail to clearly designate critical parameters.

Compare Different Models: Decision Trees

Titer



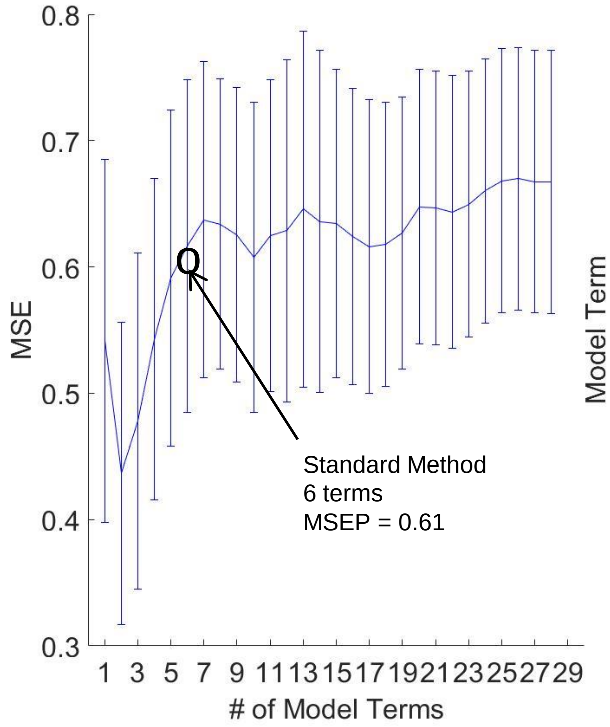
Tree MSE = 0.21
MC Stepwise MSE = 0.075

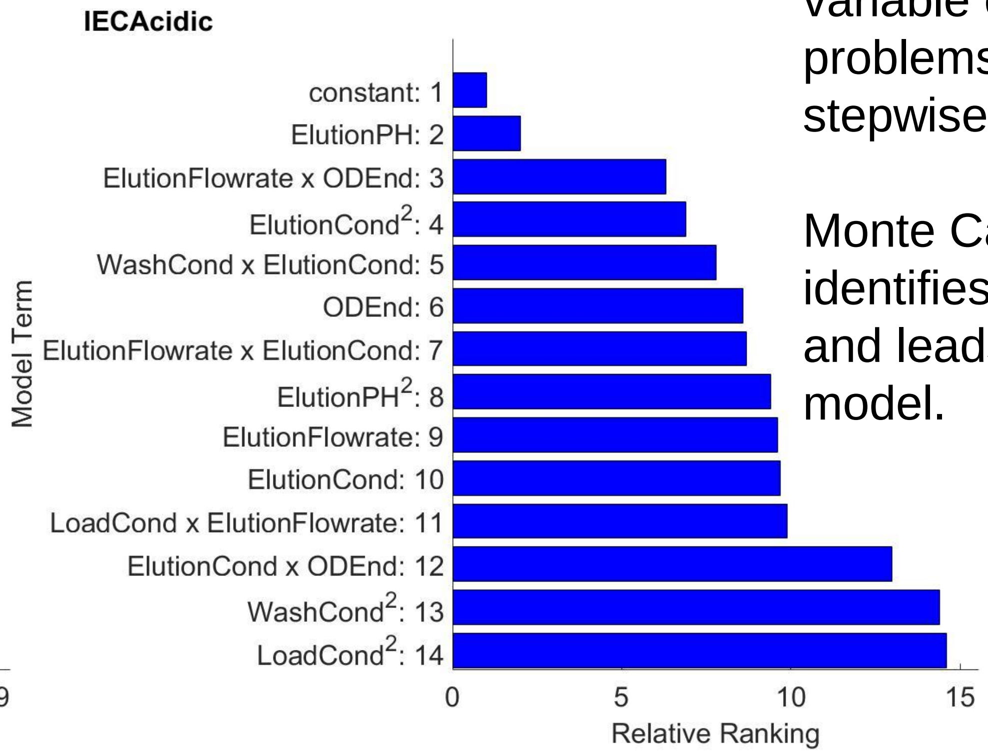
Classification and Regression Trees can provide clear parameter selection, but often fail to achieve the accuracy of linear regression techniques.

Example Process Characterization Program

- mAb Process Characterization Program
- D-optimal DOE Designs
  - Upstream
    - 102 runs / 11 factors
  - Protein A
    - 52 runs / 6 factors
  - Anion
    - 83 runs / 6 factors
  - Cation
    - 64 runs / 7 factors

Example: Difficult to Analyze Data Set

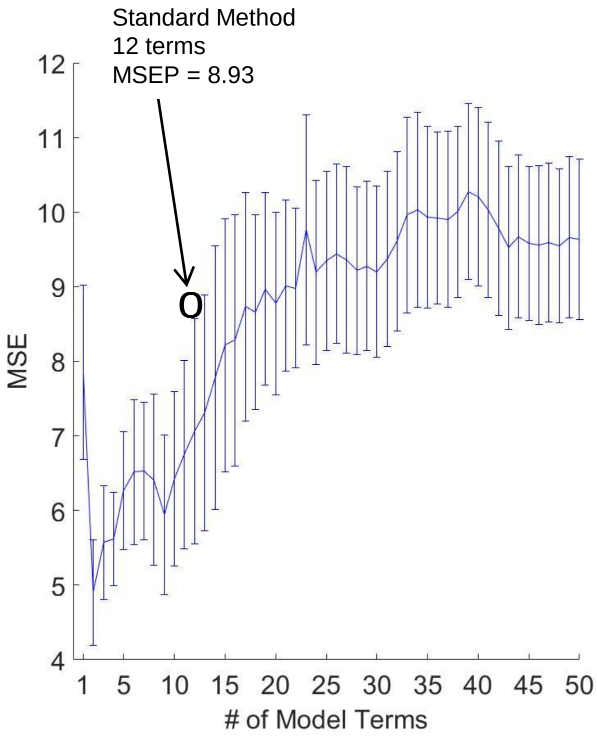


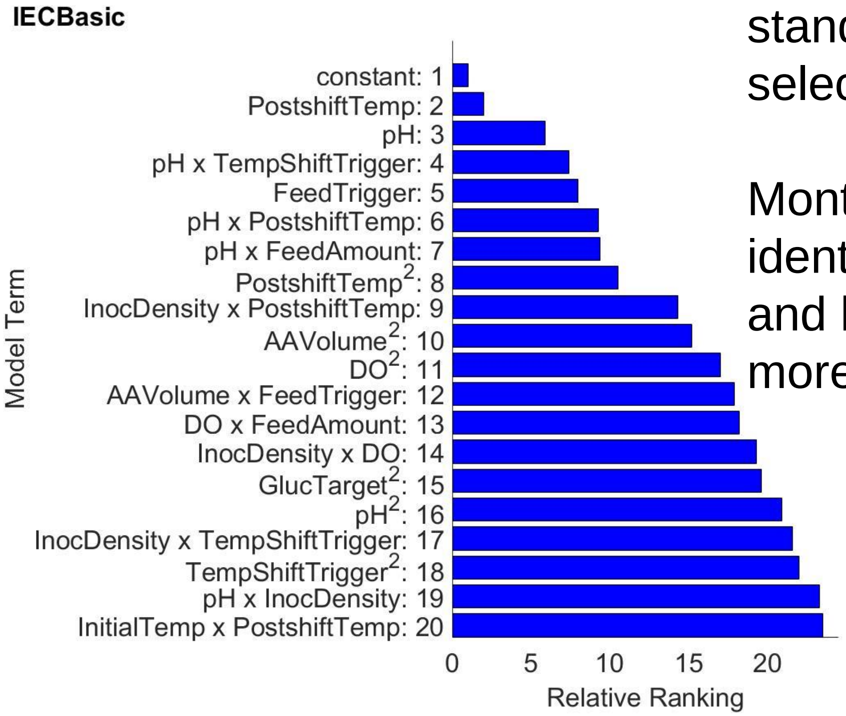


Nine-way tie for third variable caused problems for standard stepwise regression.

Monte Carlo method identifies this issue and leads to a simpler model.

Example: Many Terms Caused by Local Minima

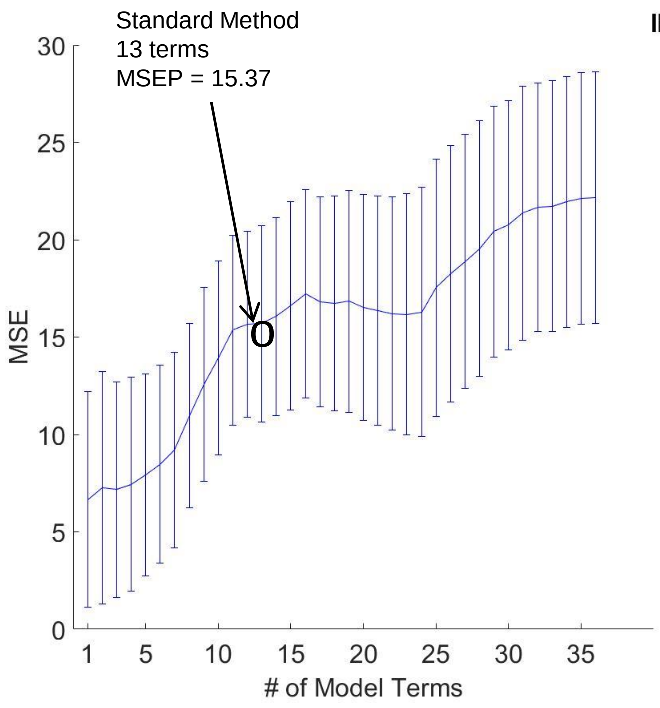


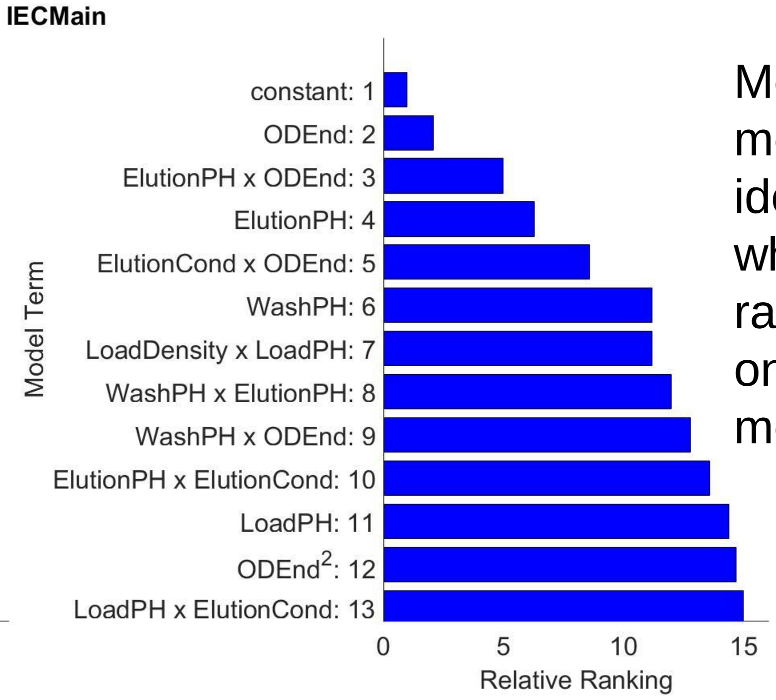


Local minima causes standard method to select larger model.

Monte Carlo method identifies this issue and leads to a simpler, more accurate model.

Example: Confidence in No Model





Monte Carlo method clearly identifies cases where the studied range has no effect on the outcome measured.

Improvements from Standard Stepwise

# Parameters per Upstream model using standard versus MC for PC						
	SEC Main	SEC HMWs	IEC Main	IEC Acidic	IEC Basic	Titer
Standard Backwards	13	14	13	17	16	22
Monte Carlo	10	10	1	5	2	13
Accuracy Difference	+10%	+5%	-2%	+1%	-5%	+20%

Conclusions

- Monte Carlo Methods, along with other advanced regression tools can improve researchers' ability to analyze their data.
- Reduction of overfitting in model selection can lead to simpler, more accurate process control, eliminating waste and improving efficiencies.
- Using advanced methods can help implement QbD, refine DOE studies and inform future programs.
- Data analysis should not be left to automated routines. There's no substitute for thoughtful scrutiny of models with the right tools.

Acknowledgements

- Co-Authors
  - Cerintha J. Hui
  - Patrick Y. Yang
  - Daniel J. Tien
  - Gayle E. Derfus
  - Rajesh Krishnan
- Executive Sponsorship
  - Yas Saotome
  - Reza Oliyai
- Upstream
  - Andy Snowden
  - Yunling Bai
  - Daniel Tong
  - Peter Zhang
  - Jennifer Autsen
  - Anne Thiel
  - Jeremy Wang
  - Garett Kasler
  - Brian Zedalis
- Downstream
  - Nooshafarin Sanaie
  - Andrew Quezada
  - Robert Vonder Reith
  - Deblina De Ghosh
  - James Woo
- Formulation
  - Charissa Towne
  - Juhi Firdos
  - Maria Fischer
  - Mohita Nimiya
- Analytical
  - Aabha Chordia
  - Darren Brown
  - Jen Kyauk
  - Josh Haleen
  - Luie Jaworski
  - Molly Roudabush
  - Noah Kiedrowski
  - Ron Seng
  - Steve Kauffman

Questions?