Overview of the DoDE and DRSM Methodologies for Enhanced Process Understanding and Optimization

Christos Georgakis

Department of Chemical and Biological Engineering

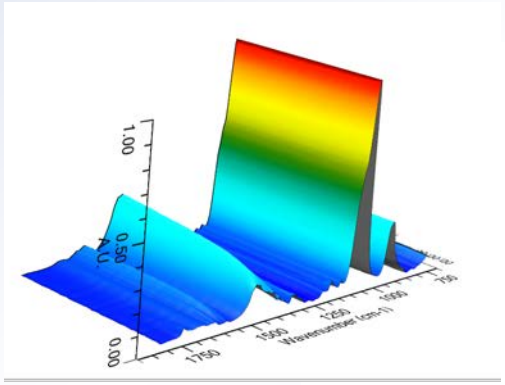
and Systems Research Institute

TUFTS University

Models: Types & Purposes

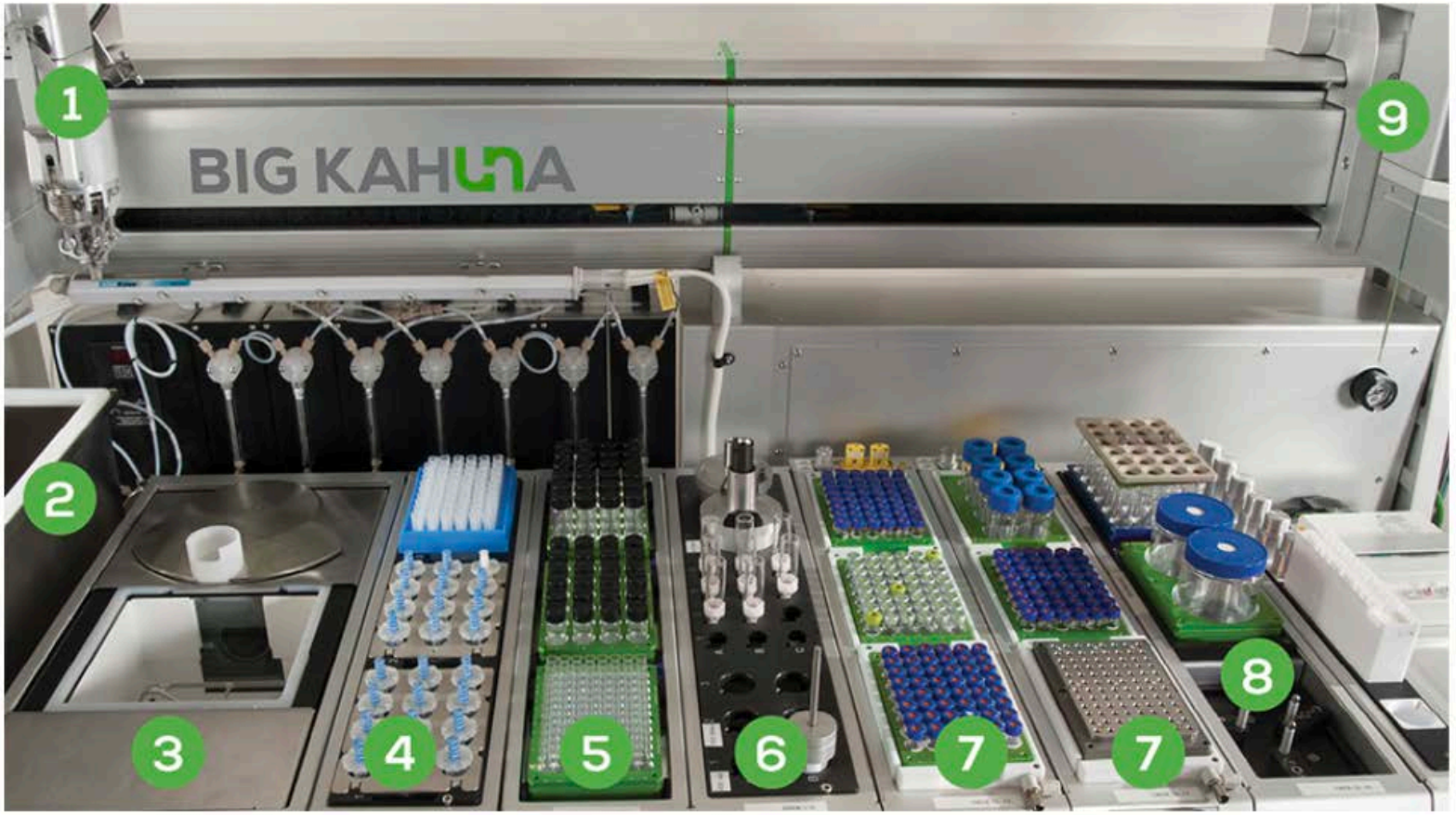
- Types of Models
  - Knowledge-Driven Models
  - Data-Driven Models *(Machine Learning)*
  - Hybrid Models
    - Partial Knowledge + Data
- ***Models Should Have a Purpose***
  - Change the Purpose ➔ Change the Model
    - Models for Design, Optimization, Control …
    - Conceptual, Physical (Pilot Plant), Mathematical, …

Plethora of Robotic Devices



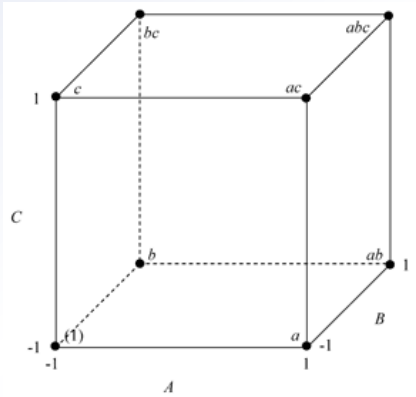
If You have One … Million $

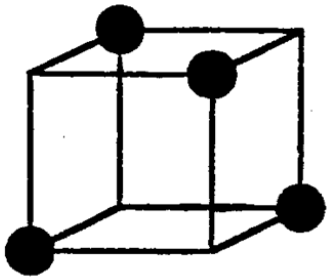
The Age of Big Data

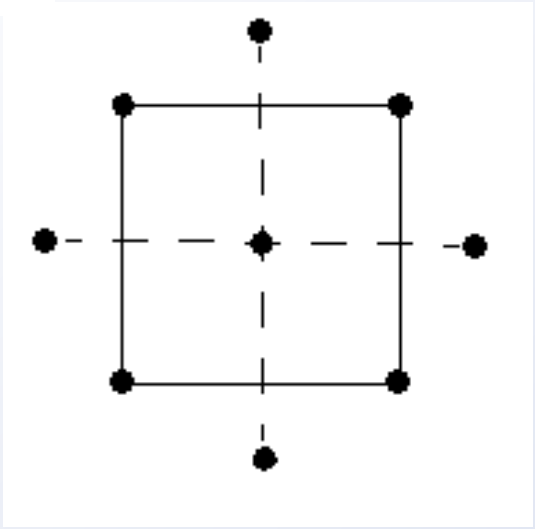


Data ➔ Model: *Which Way?*

- DoE: Powerful Methodology (50 Years Young!)
  - Full Factorial Designs,
  - Fractional Factorial Designs,
  - Center Composite Designs
  - …







Is DoE Sufficient? **NO**

DoE Can NOT Model Dynamics

How About Data Science?

Two Generalizations of DoE/RSM

❖ DoDE: Design of Dynamic Experiments

- ➢ Time-Varying Inputs (Factors)
  - ❑ Reactor Temperate vs. Time
  - ❑ Bioreactor Nutrients (t)
  - ❑ Bioreactor pH(t)

❖ DRSM: Modeling Time Resolved Data

- ➢ From: $y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij} X_i X_j + \sum_{i=1}^{n} \beta_{ii} X_i^2$
- ➢ TO: $y(t) = \beta_0(t) + \sum_{i=1}^{n} \beta_i(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij}(t) X_i X_j + \sum_{i=1}^{n} \beta_{ii}(t) X_i^2$
  - ❑ $\beta_q(t)$ = Polynomial of t

Georgakis, C., (2013) *"Design of Dynamic Experiments: A Data-Driven Methodology for the Optimization of Time-Varying Processes"* Ind. Eng. Chem. Res. **52** (35):12369-12382

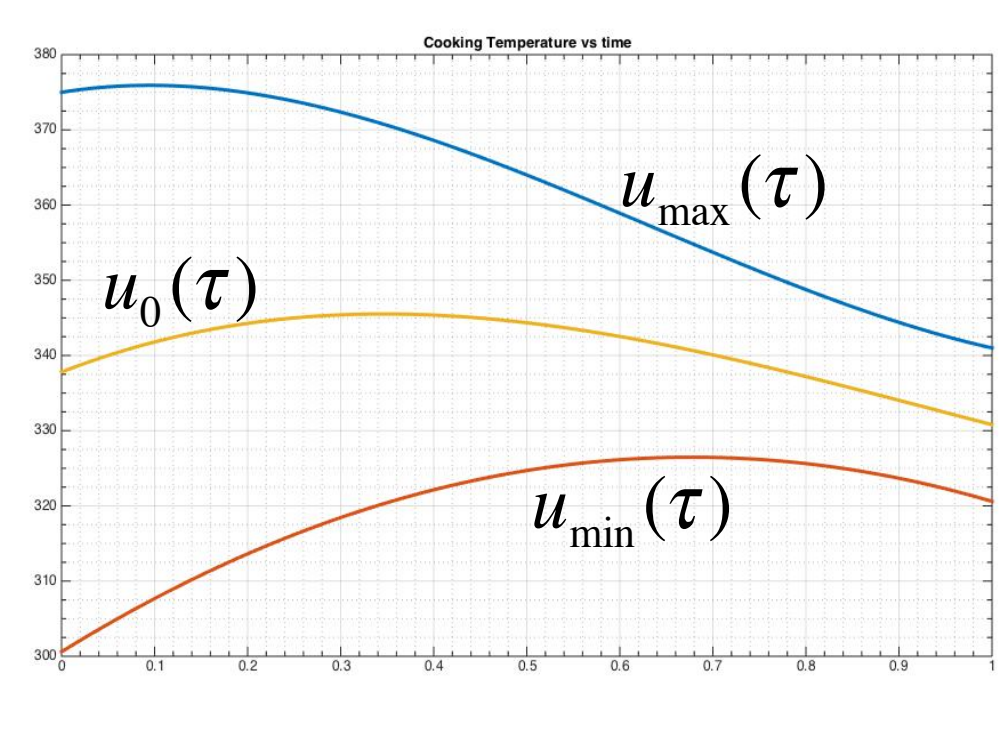
Klebanov, N., and C. Georgakis. 2016. "Dynamic Response Surface Models: A Data-Driven Approach for the Analysis of Time-Varying Process Outputs", *Ind. & Eng. Chem. Res.*, **55**: 4022-34.

Part A

DoDE: Design of Dynamic Experiments

DoDE: Time-Varying Domain

❖ Define Domain: $u_{max}(\tau)$, $u_{min}(\tau)$, $\tau = t/t_b$



❖ Coded Dynamic Factor

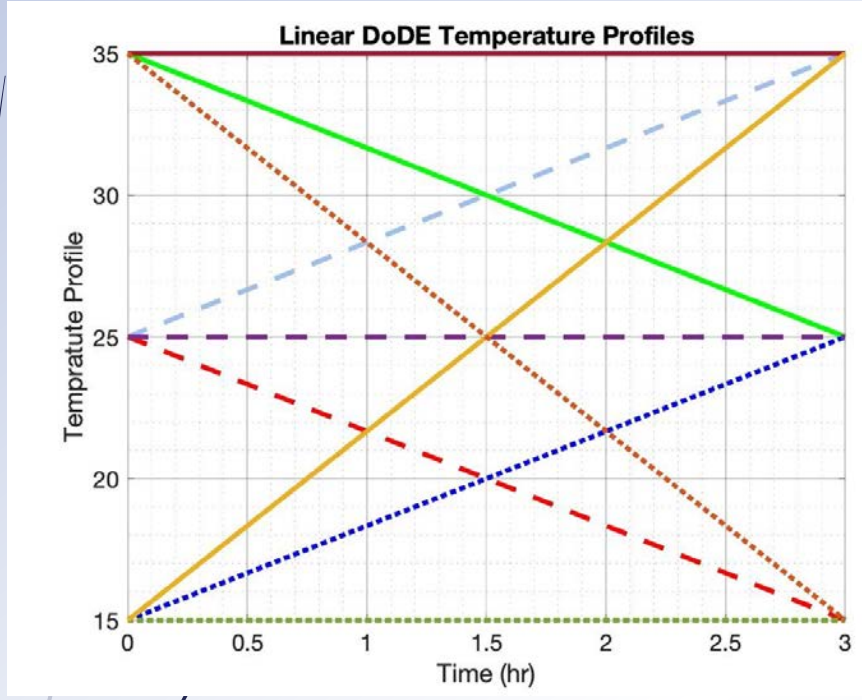
- $u(\tau) = u_0(\tau) + \Delta u(\tau) z(\tau)$
- $-1 \leq z(\tau) \leq +1$

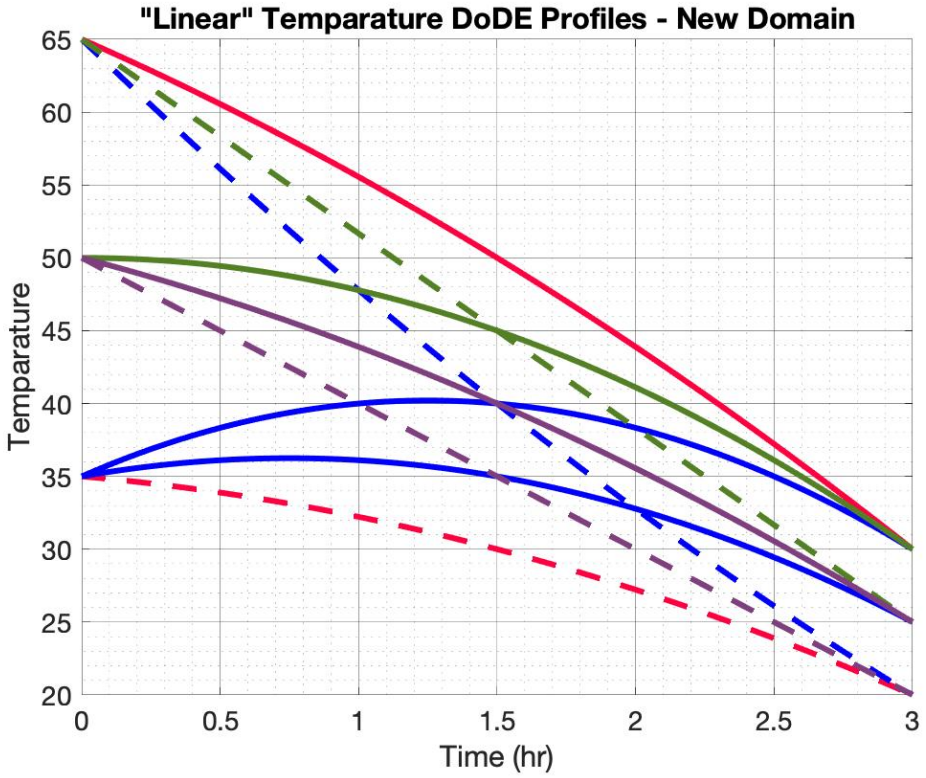
$$u_0(\tau) = 0.5\left(u_{max}(\tau) + u_{min}(\tau)\right)$$
$$\Delta u(\tau) = 0.5\left(u_{max}(\tau) - u_{min}(\tau)\right)$$

Main Idea: $z(\tau) = a_1 P_0(\tau) + a_2 P_1(\tau) + a_3 P_2(\tau) + \cdots$

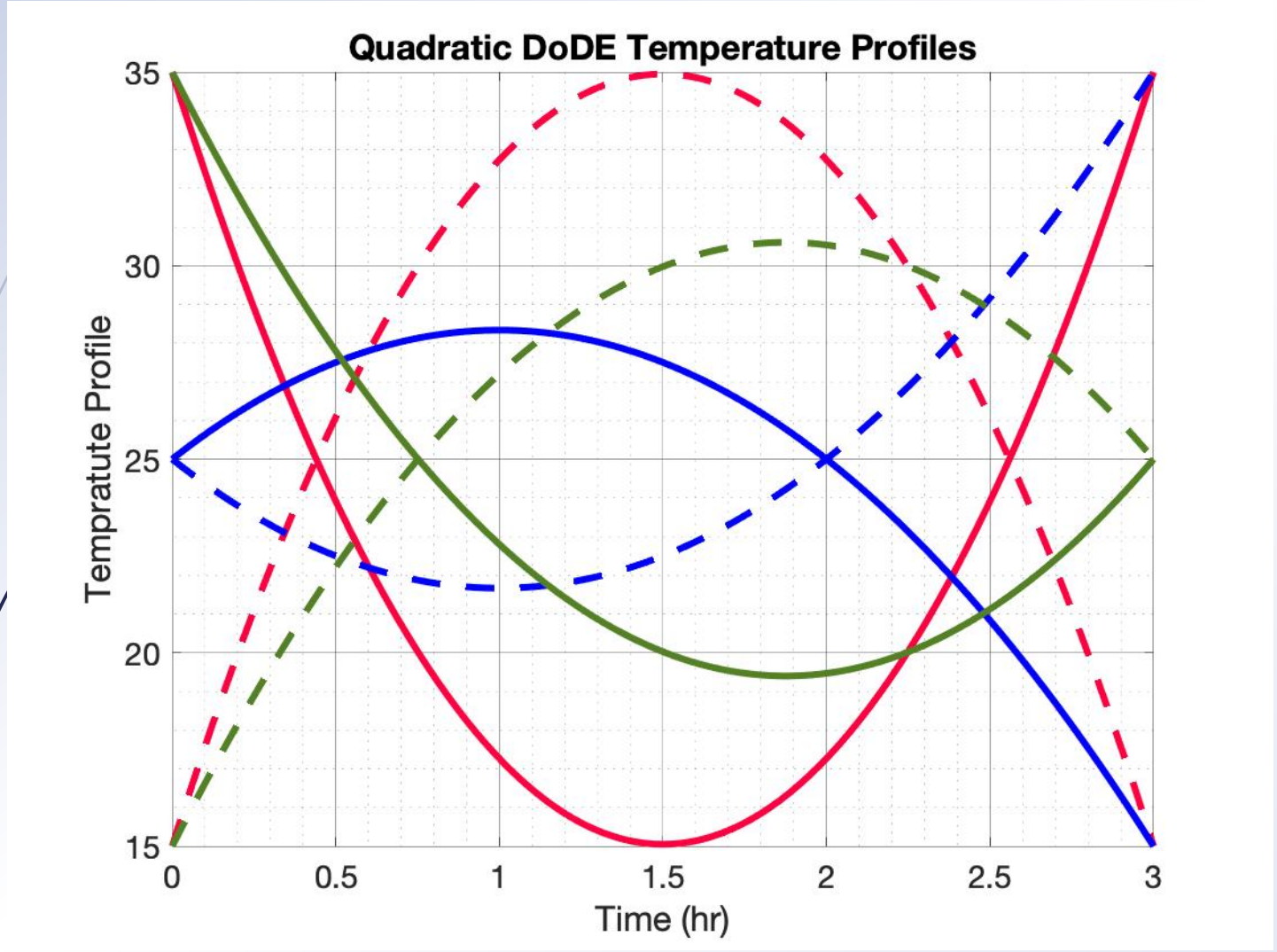
$P_i(\tau)$ $i-th$ Shifted Legendre Polynomial

Nine (9) Linear Time-Varying Inputs





Some Quadratic Profiles



DoDE Example: Batch Reactor

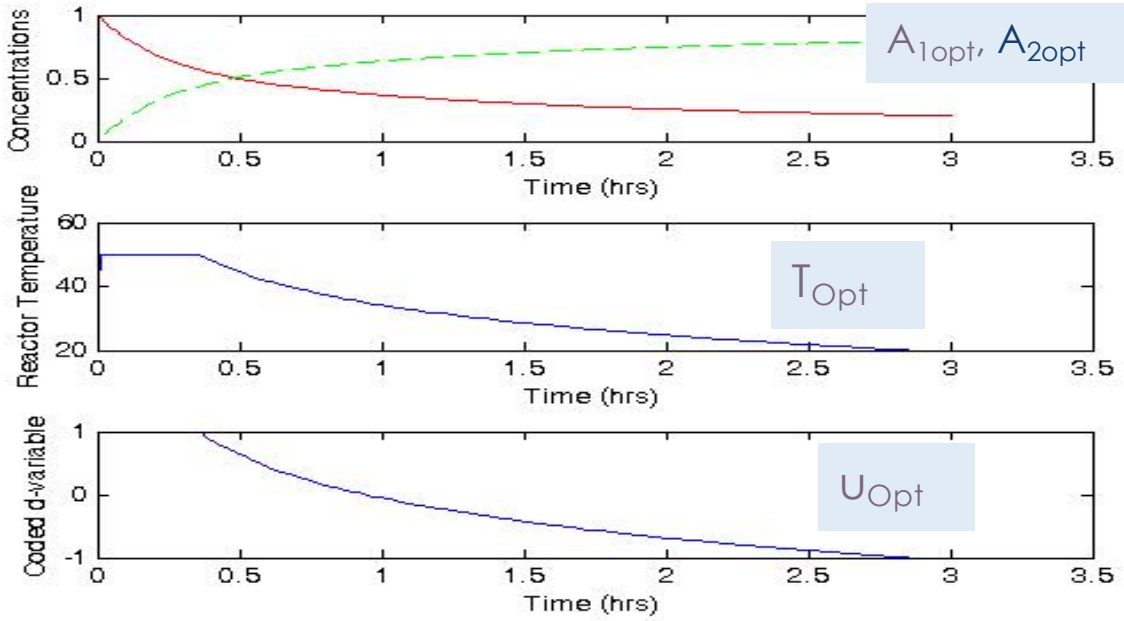
Reversible Reaction in Batch: $A_1 \rightleftharpoons A_2$ (15 < T < 50 °C)

$$r = k_1 A_1 - k_2 A_2 \qquad k_i = k_{i0} \exp\left(-\frac{E_i}{RT}\right) \; with \quad E_2 > E_1$$

Model-based Optimum Conversion:

Decreasing Temperature Profile

74.6%

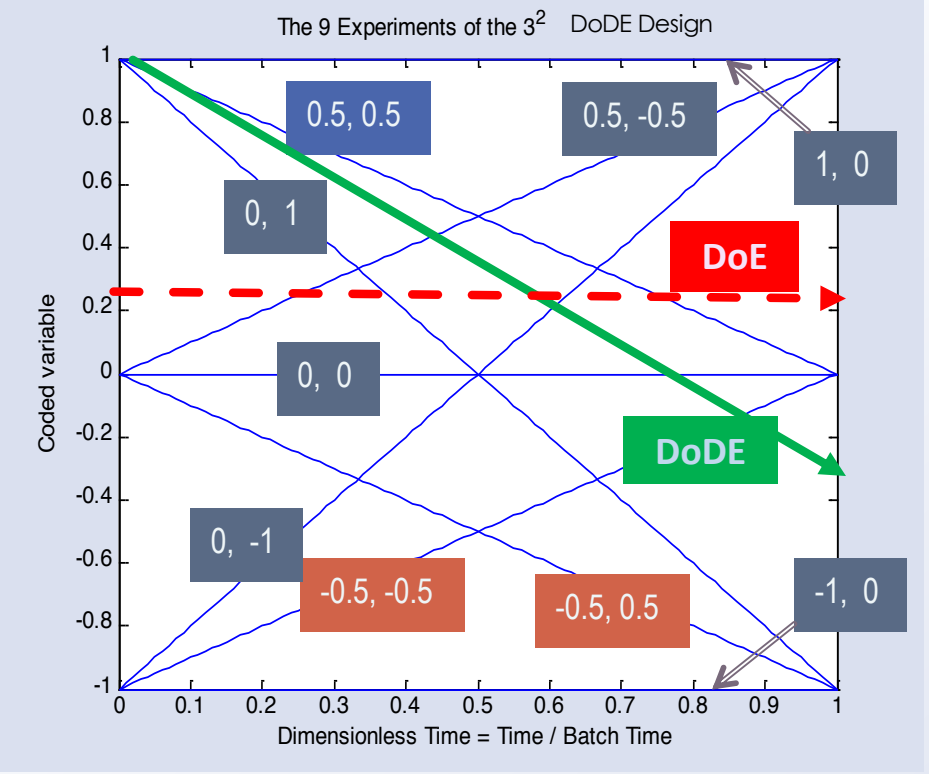


Optimization via DoDE

- Two Factors: T Level & Linear Slope
  - Nine DoDE Experiments
    - Linear in Time
    - between 15^{0}C to 50^{0}C

74.6%

- Optimization:
  - Max DoE Conversion=71.4%
  - **Max DoDE Conversion 74.3%**
    - T*(t): 50^{0}C → 28^{0}C



	DoE	DoDE	MBO
Conversion	71.4%	74.3	74.6%
Difference from MBO	3.2	0.3	

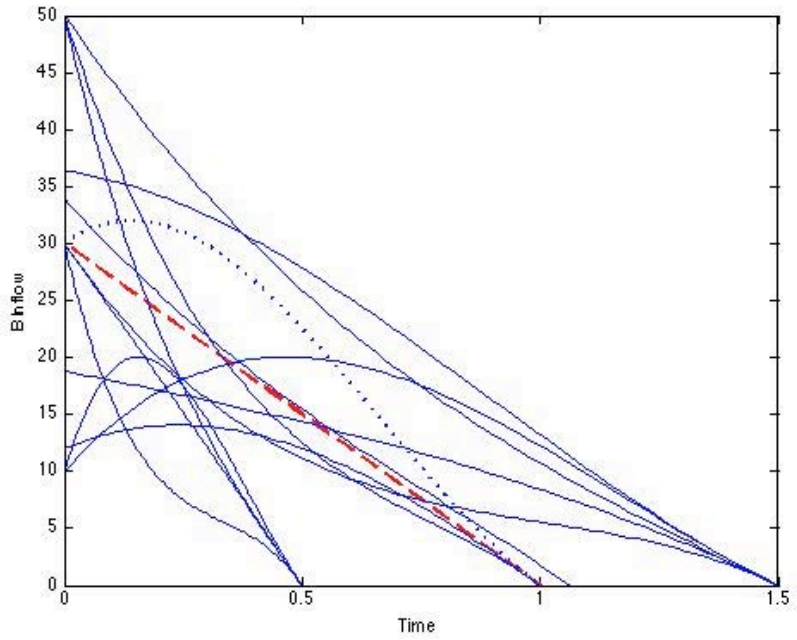
VERY Small Difference

DoDE for Semi-Batch Reactor

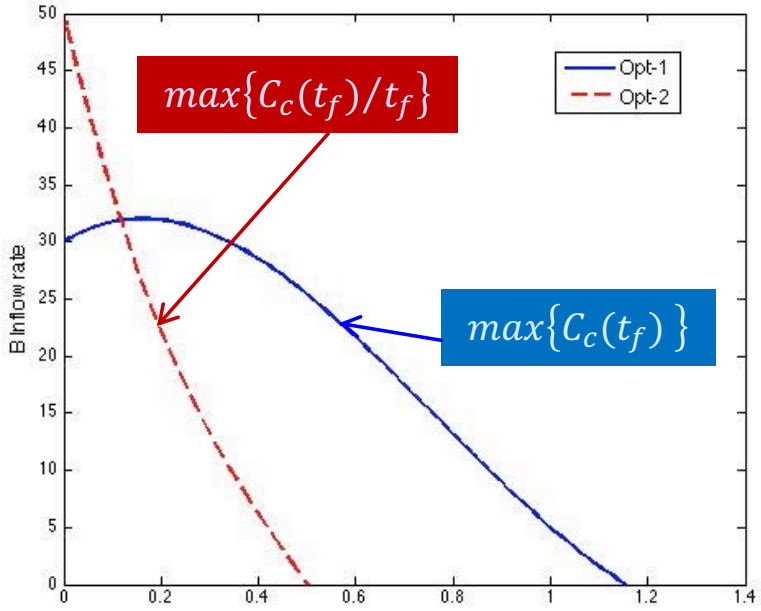
$Rxn1: A + B \longrightarrow C \quad r_1 = k_1 C_A C_B, k_1 = 2\ l\ mol^{-1} h^{-1}$
$Rxn2: 2B \longrightarrow D, \quad r_2 = k_2 C_B^2, \ k_2 = 1\ l\ mol^{-1} h^{-1}$
$Rxn3: C \longrightarrow E, \quad r_1 = k_3 C_C, \ k_3 = 1\ h^{-1}$

B in Semi-batch Mode
$q_B(t) = ?$

DoDE Runs: Feeding B



Optimal Runs



Part B

DRSM: Dynamic Response Surface Methodology

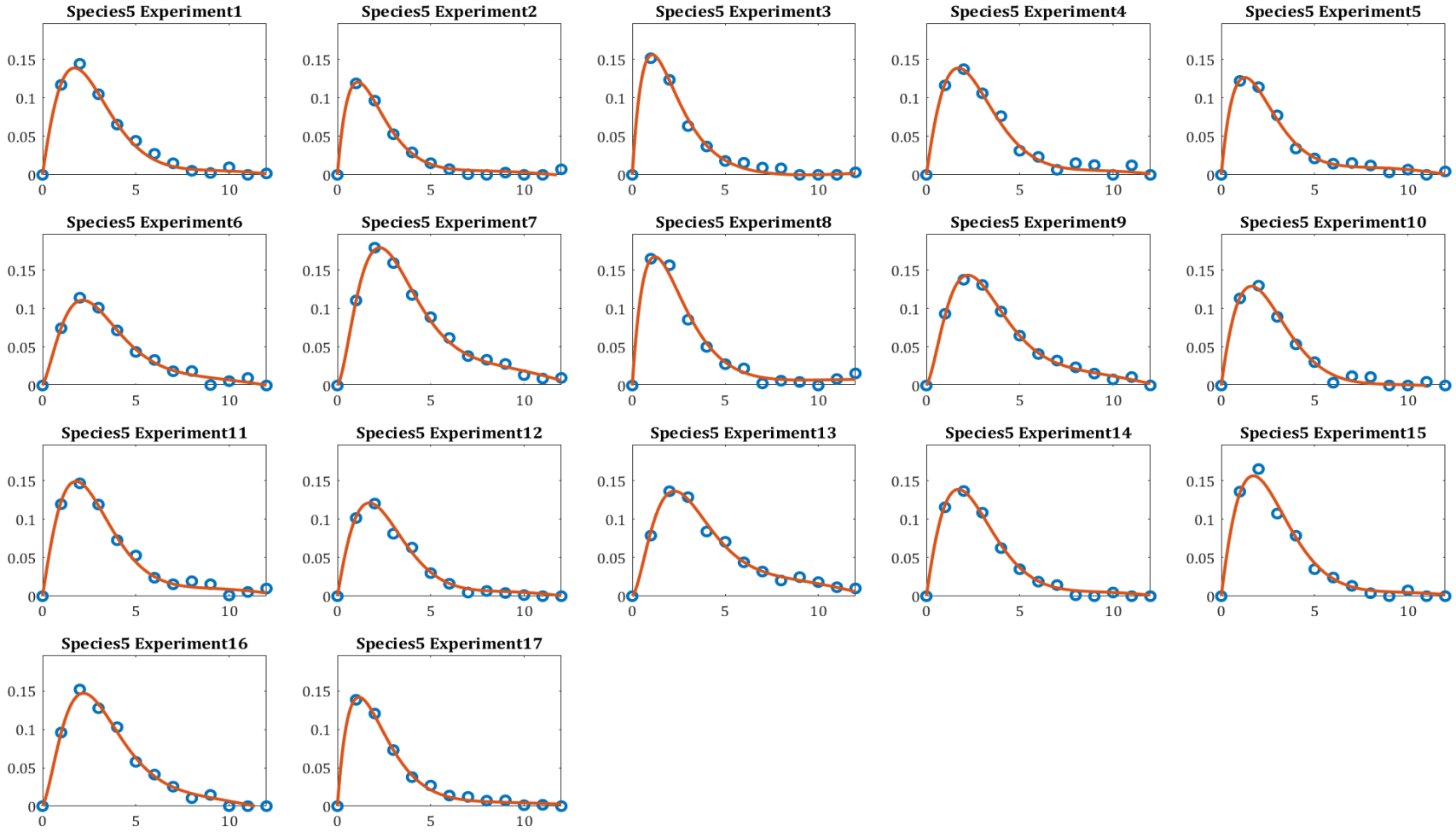
The DRSM Idea

- From RSM:
  - $y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij} X_i X_j + \sum_{i=1}^{n} \beta_{ii} X_i^2$
- To DRSM:
  - $y(t) = \beta_0(t) + \sum_{i=1}^{n} \beta_i(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij}(t) X_i X_j + \sum_{i=1}^{n} \beta_{ii}(t) X_i^2$
- Parameterization:
  - $\beta_q(t) = \gamma_{q,1} P_0(t) + \gamma_{q,2} P_1(t) + \cdots \gamma_{q,R} P_{R-1}(t)$
    - $q = i, ij, or\ ii$ with $i, j = 1, 2, \ldots, n; j < i$
  - R(parameters) < K(Data per Experiment)
- **DRSM-1: Parametrization with t** ➡ Has Oscillations
- **DRSM-2: Parametrization with $\theta = \left\{1 - \exp(-\frac{t}{t_0})\right\}$**
  - $0 \le t < \infty \iff 0 \le \theta < 1$

NO Oscillations - Excellent Model

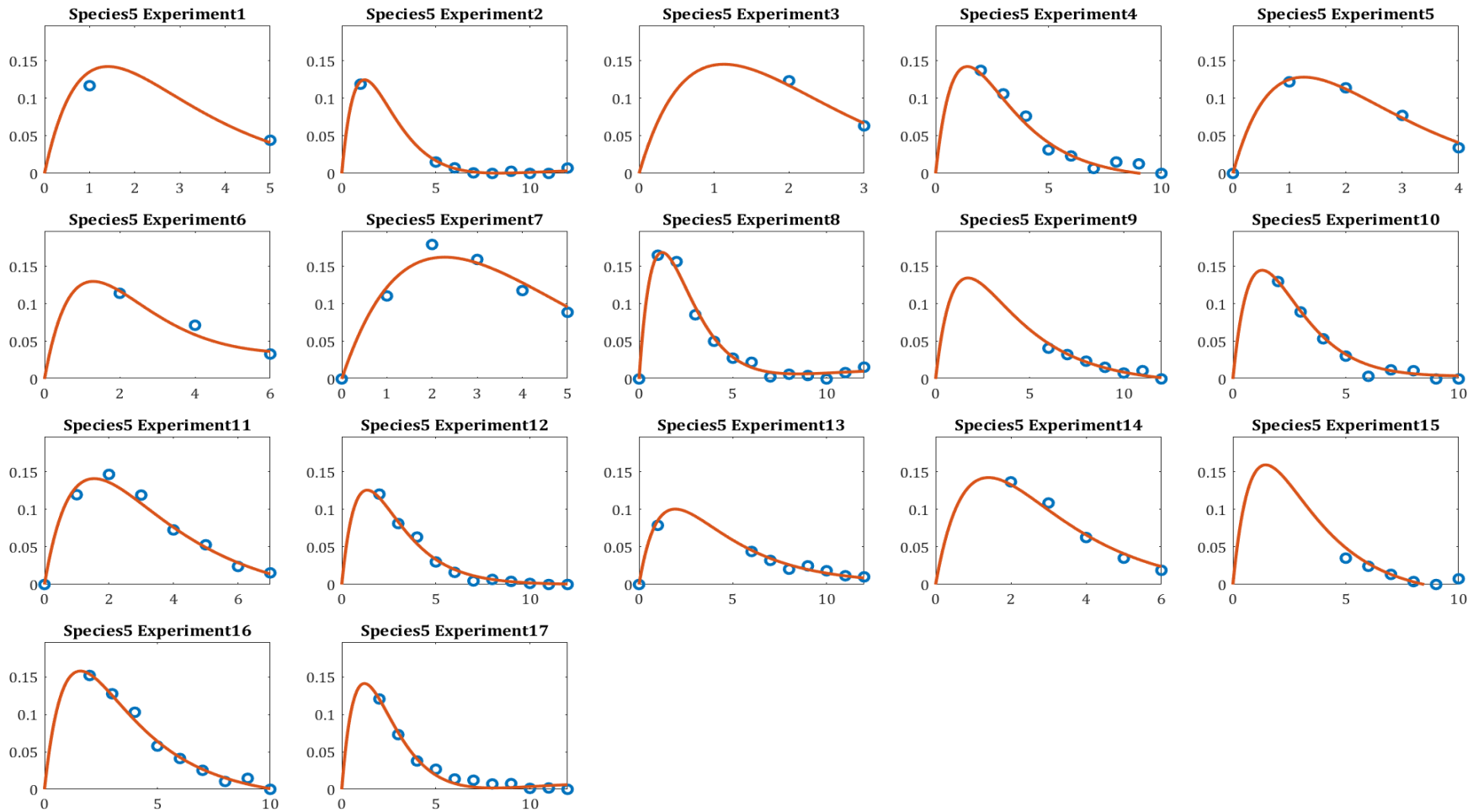
DRSM-2c for **ALL** Pfizer Data: $C_5(t)$

❖ Species 5: $R = 5, t_c = 5.4$

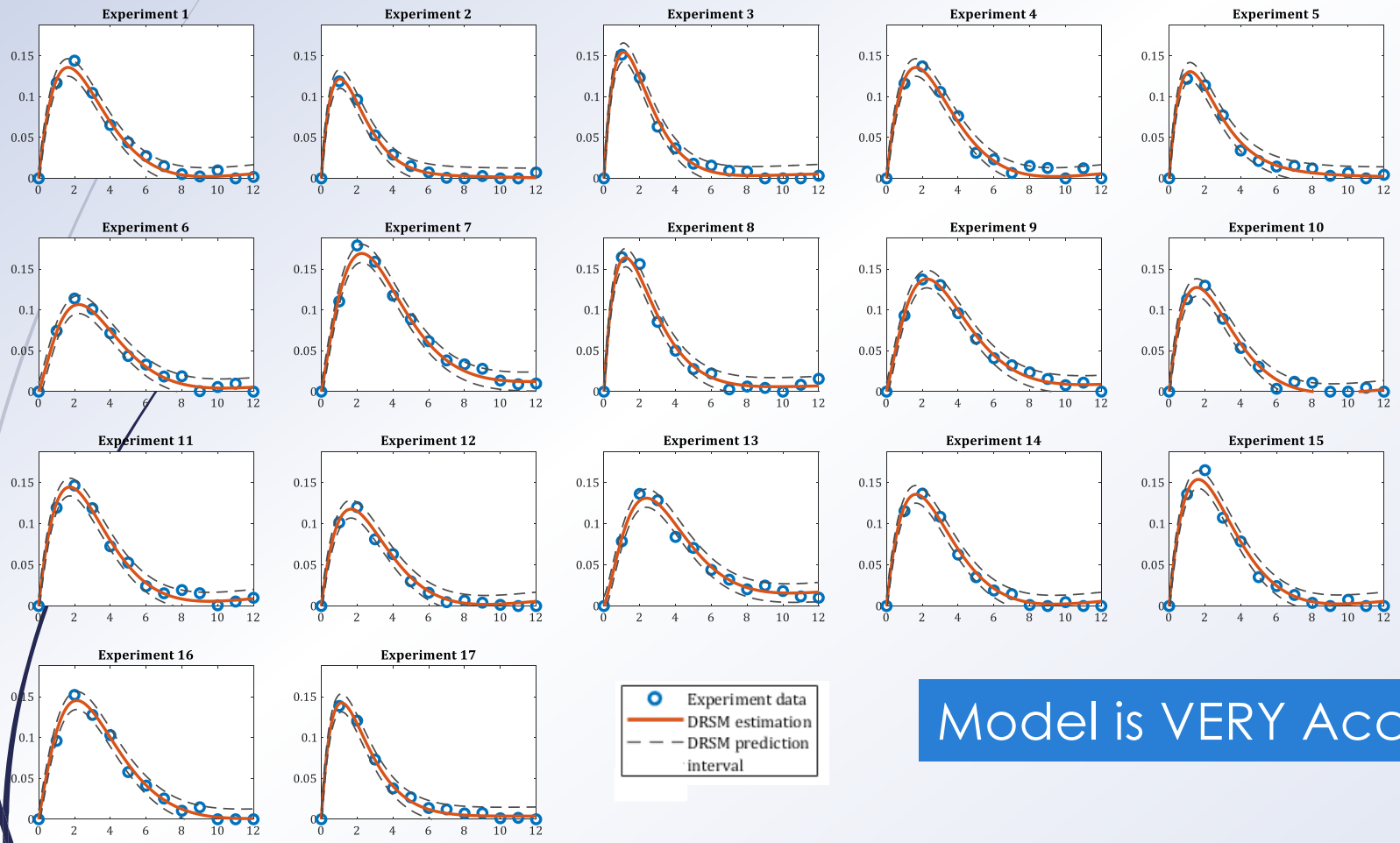


DRSM-2c: **Missing** Pfizer Data: $C_5(t)$

- Species 5: R=3 Tc=3.3



Species E with Prediction Interval



Model is VERY Accurate

Fractional Factorial Design

- 3 Species and 5 Factors: A, B, C, D, and E
- 2 Blocks: Robotic & Manual
- 6 Samples per Batch at ***Unequal*** Intervals
  0, 20,40, 60, 120, 240 mins
- LC area converted to concentration

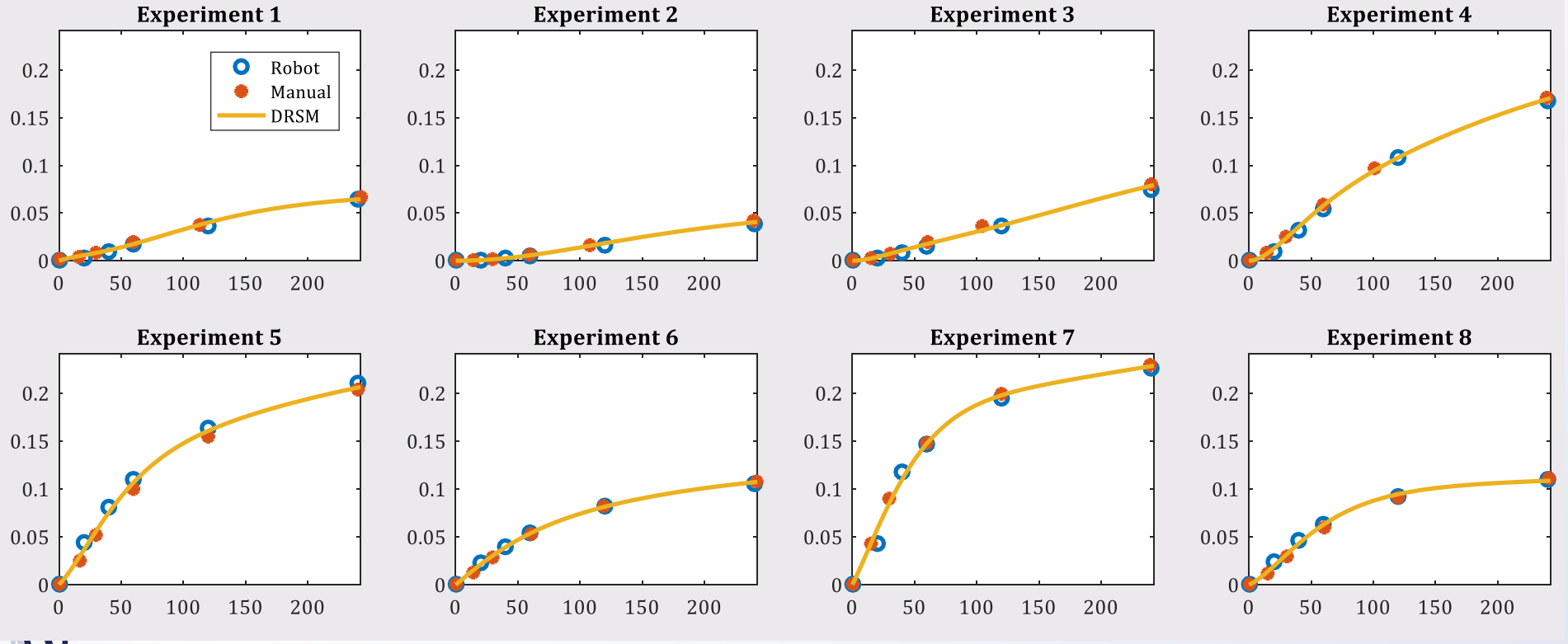
The 5 FACTORS

A: Methanol ratio, (% wt/wt solvent)
B: Starting material, wt% **C**: Base, wt%
D: Water wt% **E**: Temperature

- **¼ Fractional factorial design**: 2^{5-2} design
  - 8 experiments
  - Aliasing Structure: D = AB, and E = AC

2FI Model: Species B

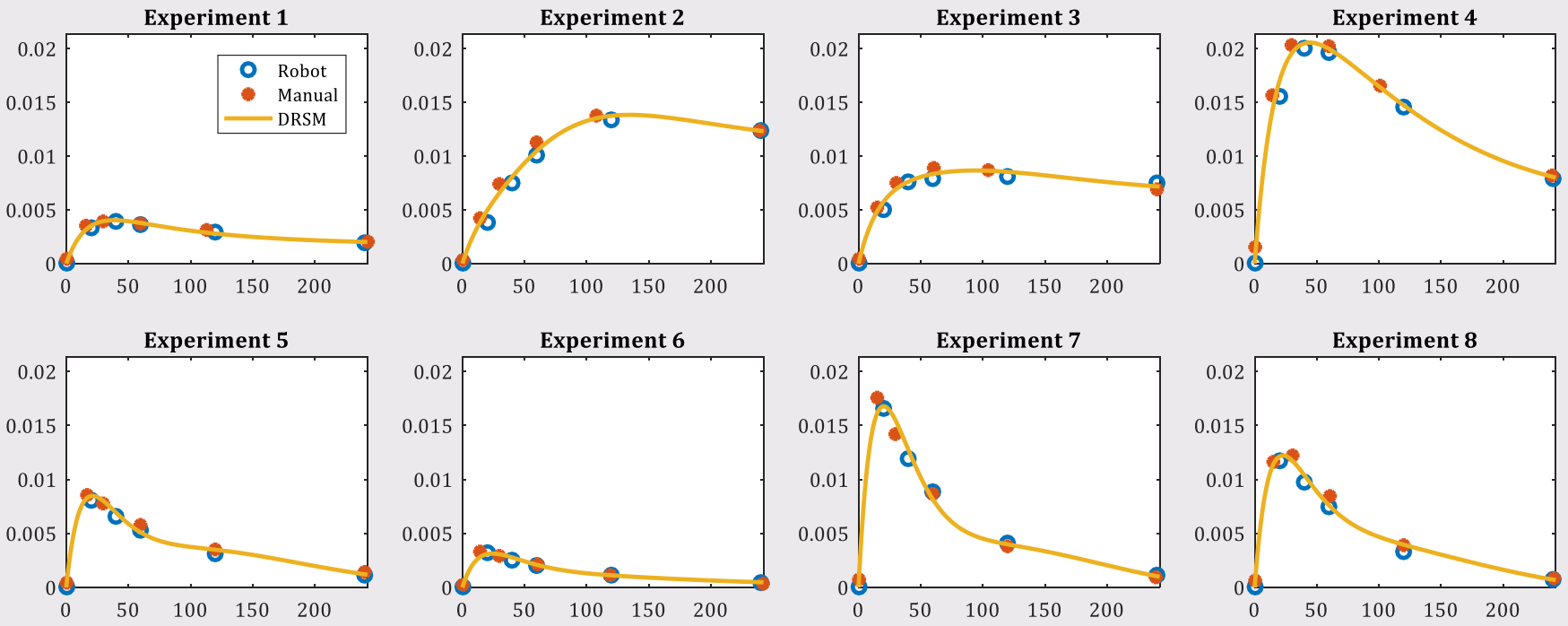
LoF *p*-value = 0.99 ➔ Perfect model



$$\tilde{y}(t) = \beta_0(t) + \beta_A(t)A + \beta_B(t)B + \beta_C(t)C + \beta_D(t)D + \beta_E(t)E +$$
$$+\boldsymbol{\beta_{BC}(t)BC + \beta_{CD}(t)CD}$$

2FI Model: Species C

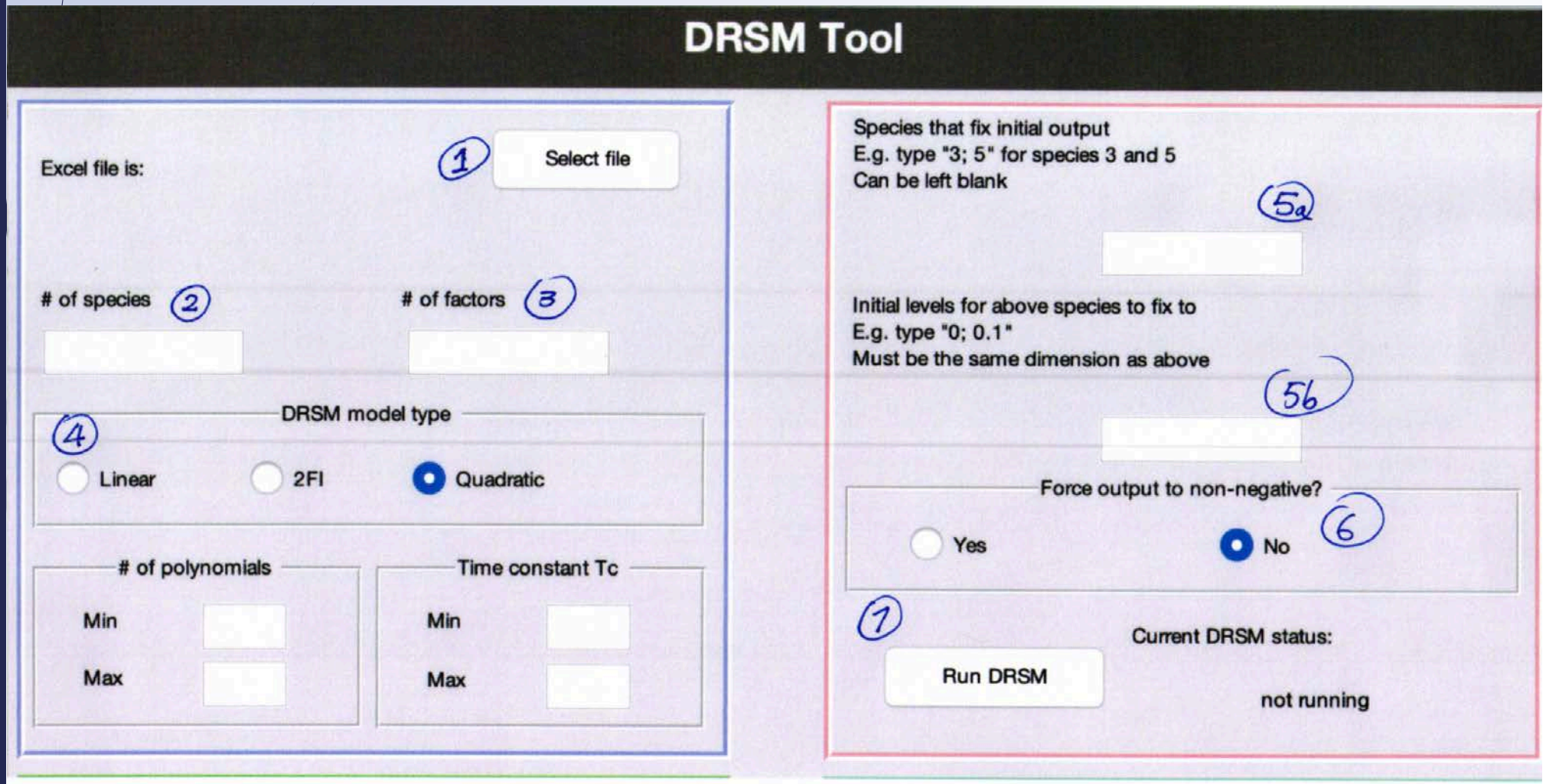
LoF p-value = 0.06 ➔ Good Model



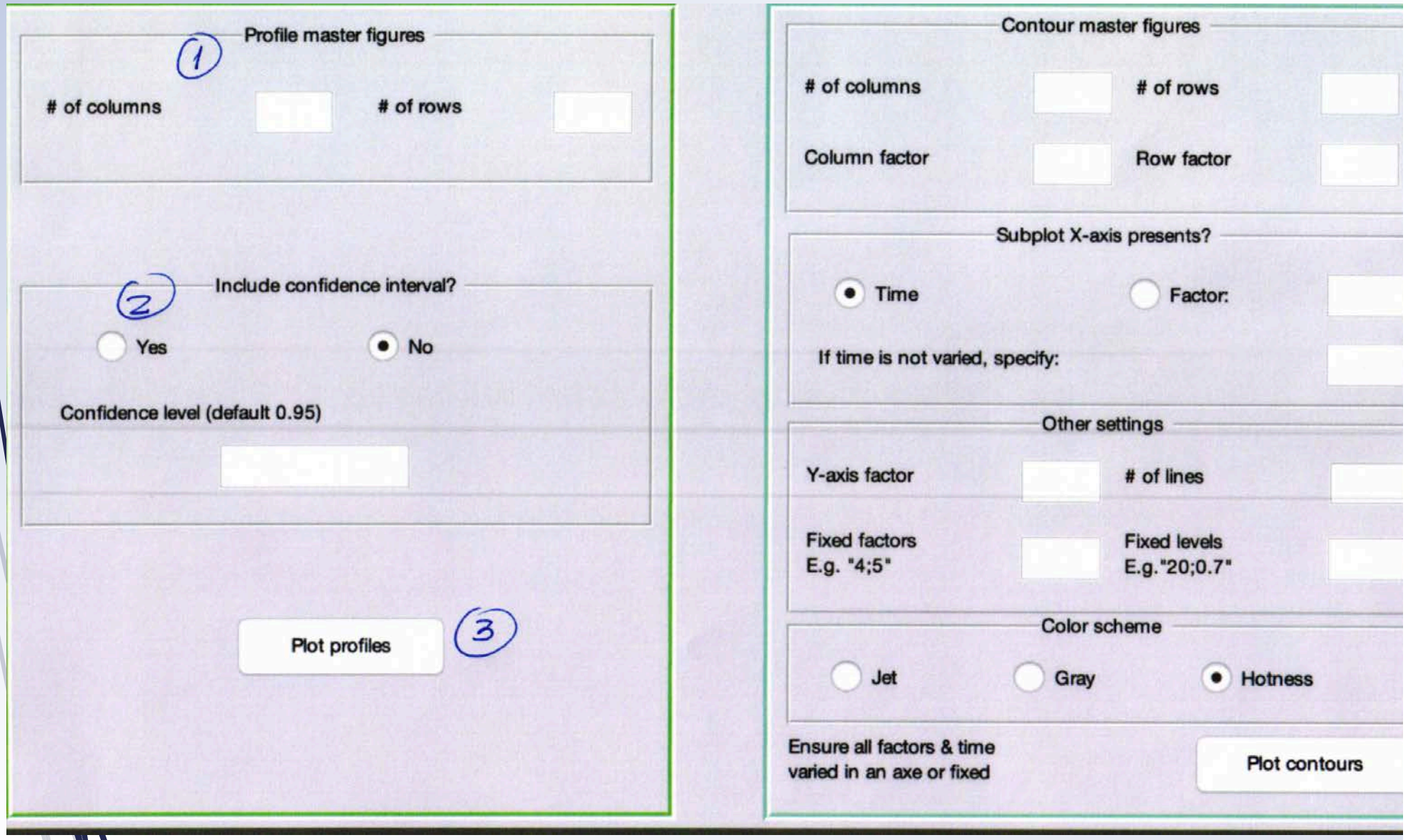
$$\tilde{y}(t) = \beta_0(t) + \beta_A(t)A + \beta_B(t)B + \beta_C(t)C + \beta_D(t)D + \beta_E(t)E +$$
$$+\boldsymbol{\beta_{BC}(t)BC + \beta_{CD}(t)CD}$$

Block Effect Insignificant: Robotic vs. Manual Operation

DRSM Menu --- Top part



DRSM Menu – Bottom Part



Part C

From DRSM To Stoichiometry & Kinetics

DRSM ⟹ from **Data** to … **Knowledge**

DISCOVER Stoichiometry and Kinetics

- Simple Semi-Batch Reactor Example
  - Five DRSMs for $C_A(t), \ldots, C_E(t)$,

➔ $\begin{cases} R1: \ A + B \rightarrow C \\ R2: \ 2B \rightarrow D \\ R3: \ C \rightarrow E \end{cases}$

$$y_A(t) = \beta_{A0}(t) + \sum_{i=1}^{n} \beta_{Ai}(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{Aij}(t) X_i X_j + \sum_{i=1}^{n} \beta_{Aii}(t) X_i^2$$

$$\vdots$$

$$y_E(t) = \beta_{E0}(t) + \sum_{i=1}^{n} \beta_{Ei}(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{Eij}(t) X_i X_j + \sum_{i=1}^{n} \beta_{Eii}(t) X_i^2$$

- Calculate Derivatives with Time for ALL Models

$$y'_A(t) = \beta'_{A0}(t) + \sum_{i=1}^{n} \beta'_{Ai}(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta'_{Aij}(t) X_i X_j + \sum_{i=1}^{n} \beta'_{Aii}(t) X_i^2$$

Rate Data ⟹ Stoichiometry

$$\rightarrow \begin{cases} Rxn1:\ A + B \rightarrow C \\ Rxn2:\ 2B \rightarrow D \\ Rxn3:\ C \rightarrow E \end{cases}$$

- Rates of appearance for Each Species
  - $\boldsymbol{D}_k = \begin{pmatrix} r_{Ak}(t_1) & r_{Bk}(t_1) & r_{Ck}(t_1) & r_{Dk}(t_1) & r_{Ek}(t_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{Ak}(t_i) & r_{Bk}(t_i) & r_{Ck}(t_i) & r_{Dk}(t_i) & r_{Ek}(t_i) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{Ak}(t_{n_K}) & r_{Bk}(t_{n_K}) & r_{Ck}(t_{n_K}) & r_{Dk}(t_{n_K}) & r_{Ek}(t_{n_K}) \end{pmatrix}$

  $t_i = i\Delta t$
  $i = 1, \dots, n_K$
  $\Delta t = 1/n_K$

  - For $n_K = 100$ matrix D_k is a 100×5
- Data Matrix for Rates of ALL Species and ALL Experiments: R_c
  - $\boldsymbol{R}_c = \begin{pmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \\ \vdots \\ \boldsymbol{D}_{n_e} \end{pmatrix}$ D_k = Data from k−th experiment
    $k = 1, 2, \dots, n_e$
  - For $n_e = 9$ experiments R_c is a 900×5 matrix
- SVD=Singular Value Decomposition of $\boldsymbol{R}_c$

Number of Significant SVs = ?

  - $\boldsymbol{R}_c = \boldsymbol{U\Sigma V}^{\mathrm{T}},\ \Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_5 \end{pmatrix}$, $\boldsymbol{U\Sigma V}^{\mathrm{T}}$ $sizes$: $900X5, 5X5$ & $5X5$

SVD: $\boldsymbol{R_c = U\Sigma V^{\mathrm{T}}}$ & Projections

- # of Reactions ⇔ Significant σ_i Values= 3
  - $\boldsymbol{R_c = U_3\Sigma_3 V_3^T}$ $\quad \boldsymbol{V_3^T} = \begin{pmatrix} \boldsymbol{v_1^T} \\ \boldsymbol{v_2^T} \\ \boldsymbol{v_3^T} \end{pmatrix} = \begin{pmatrix} 0.41 & 0.84 & -0.26 & -0.21 & -0.15 \\ -0.26 & 0.21 & 0.79 & -0.23 & 0.50 \\ 0.60 & -0.28 & 0.01 & 0.44 & -0.61 \end{pmatrix}$
- **IS (-1,-1,1,0,0) a Linear Combination of the $\boldsymbol{V_3^T}$ rows ?**
- Projection Matrix: $\boldsymbol{P = V_3^T V_3}$
- Projection of Candidate Stoichiometry: $\boldsymbol{n_{ir} = n_i V_3^T V_3}$
- Is it TRUE that: $\boldsymbol{n_{ir} \cong n_i}$?
- Projection Score: $PS = 100\{1 - \|\boldsymbol{n_{ir} - n_i}\|/\|\boldsymbol{n_i}\|\}$
  - **PS ≥ 90 is GOOD**

Bonvin & Rippin (1998)Target Factor Analysis (TFA)

Identifying Pfizer Stoichiometries

Additive error = 0.005 on Concentrations $0.005 < C_i(t_k) < 0.9$

Scores of True reactions		
1	A + B⇆ C + D	96.5
2	C → D + E	90.8
3	E → F	92.3
4	B + D ⇆ G	99.1
5	G → D + H	96.3
6	A + F → I	82.4
7	2A → J	77.4
8	B + J → 2E + I	24.8

Scores of Untrue reactions		
1	A → J	57.6
2	C → J	38.8
3	2A + B → J	72.5
4	J → 2D + I	65.0
5	B + J → E + I	21.2
6	B + J → D + I	51.2

Blind Test: Excellent Result

$\text{Score}_i = 100(1 - \|\boldsymbol{n}_{ir} - n_i\| / \|\boldsymbol{n}_i\|)$
$\boldsymbol{n}_i$ = Candidate Stoichiometry
$\boldsymbol{n}_{ir}$ = Response Vector

Seven (7) Significant SVs via an *F*-test
σ_i = **81, 9.7, 6.3, 1.5, 1.0, 0.92, 0.22**, 0.18, 0.15, 0.09

Identifying Pfizer Stoichiometries

NO Measurement error

Scores of True reactions (Without Measurements Error)		
1	A + B ⇆ C + D	99.5
2	C → D + E	99.0
3	E → F	99.5
4	B + D ⇆ G	99.9
5	G → D + H	99.5
6	A + F → I	99.9
7	2A → J	97.9
8	B + J → 2E + I	92.7

Scores of Untrue reactions (Without Measurement error)		
1	A → J	74.6
2	C → J	57.3
3	2A + B → J	81.9
4	J → 2D + I	82.0
5	B + J → E + I	85.8
6	B + J → D + I	88.6

Confirmation of Method

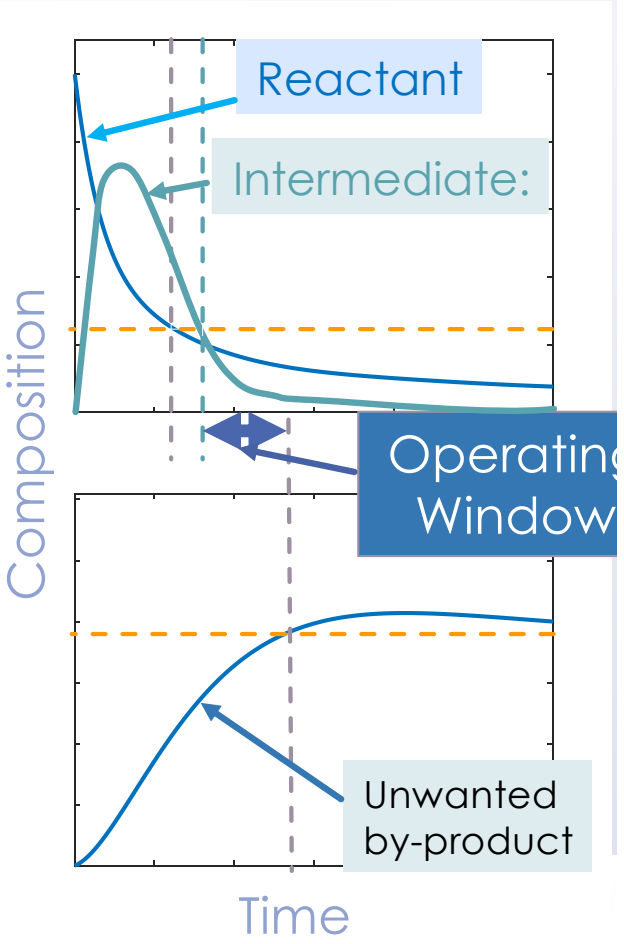
Eight (8) Significant SVs

$\sigma_i =$ **81, 9.7, 6.3, 1.5, 1.0, 0.91, 0.19, 0.10**, 0.04, 0.02

Dw

Process Optimization via DRSM

- Calculate Operating Window (OW):
  - Concentrations of Impurities Below Specs
    - Reactants, Intermediates, by-products
- Maximize OW
  - Select Operating Conditions
  - Account for Uncertainties
- Can also Use
  - Peak Area of HPLC Data

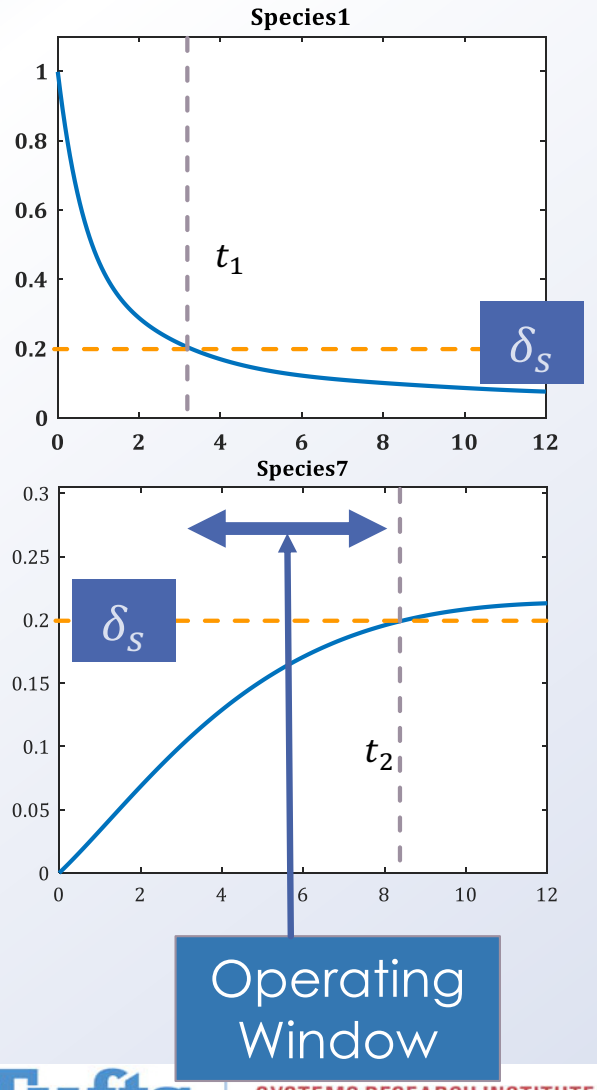


Maximize Operating Window

Optimization Results

- When $\delta_s = 0.1$ for all species,
  - window does not exist.
- Study different specifications

δ_s	Factor1	Factor2	Factor3	Optimal window (hr)
0.14	Infeasible			---
0.15	90	1.02	0	3.07
0.16	90	1.03	0.06	4.35
0.17	90	1.03	0.17	4.59
0.18	90	1.03	0.28	4.81
0.19	90	1.03	0.39	5.01
0.2	90	1.03	0.50	5.20

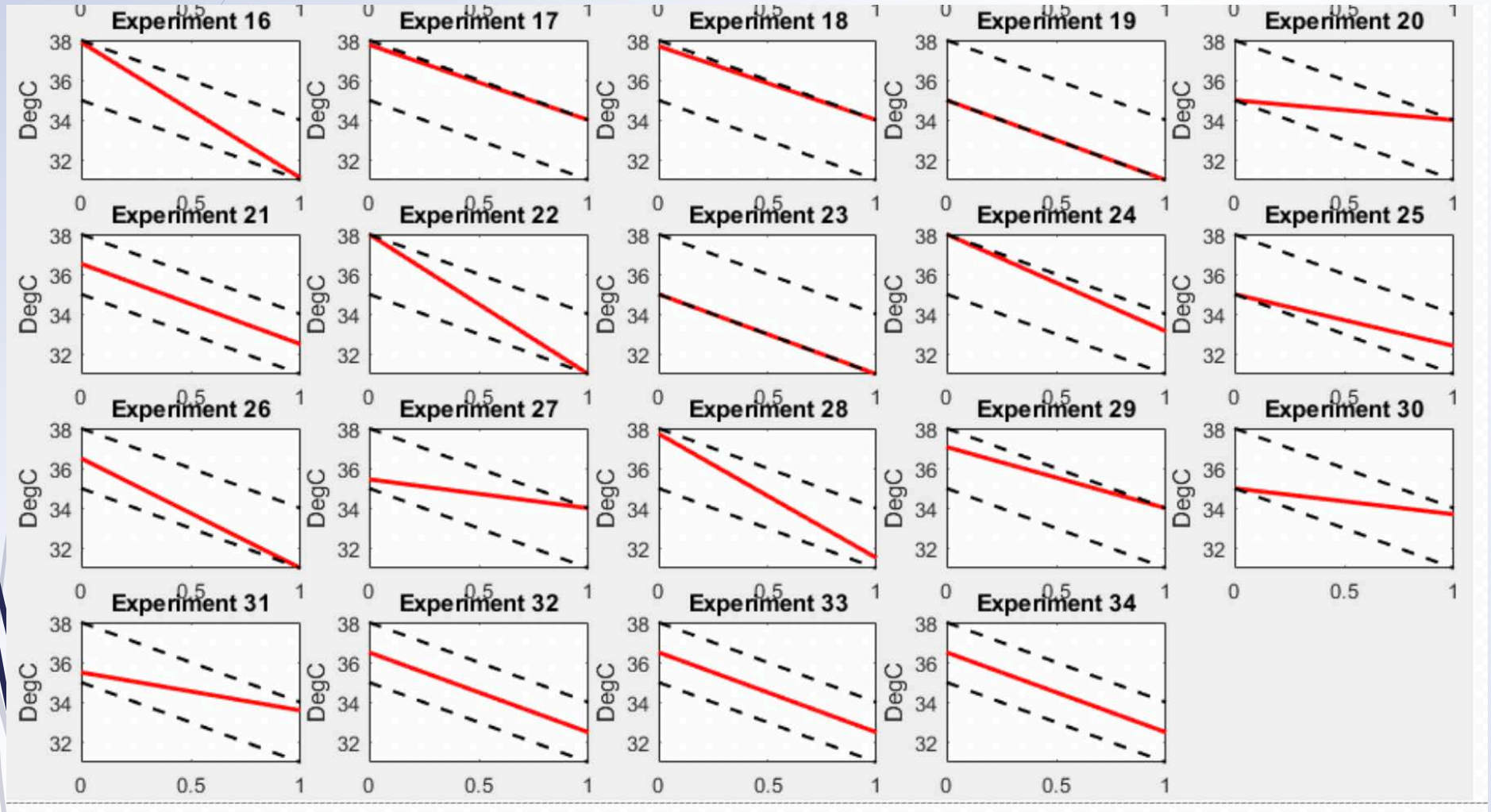


Part D

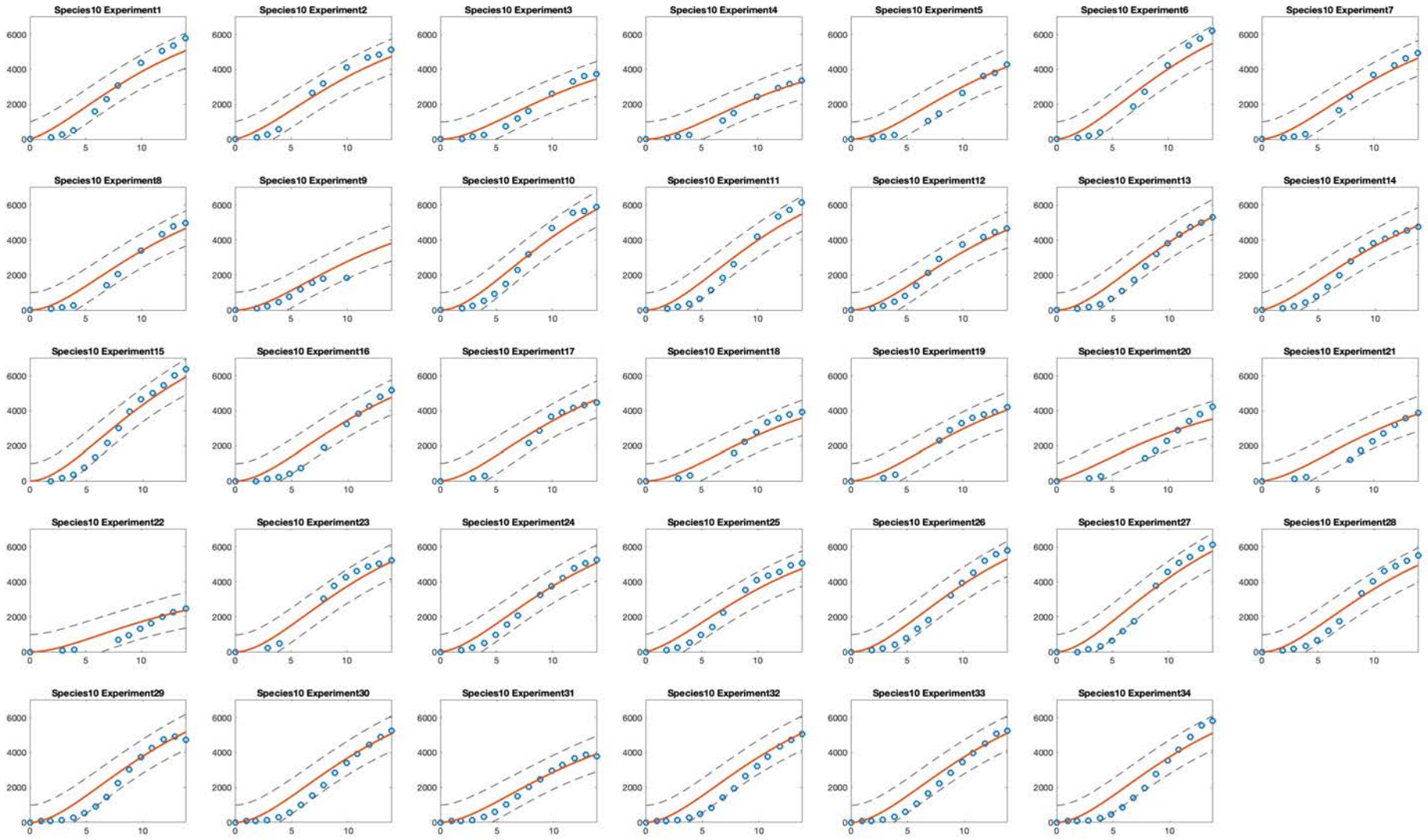
DoDE & DRSM for CELL Culture Processes

Some Temperature Profiles

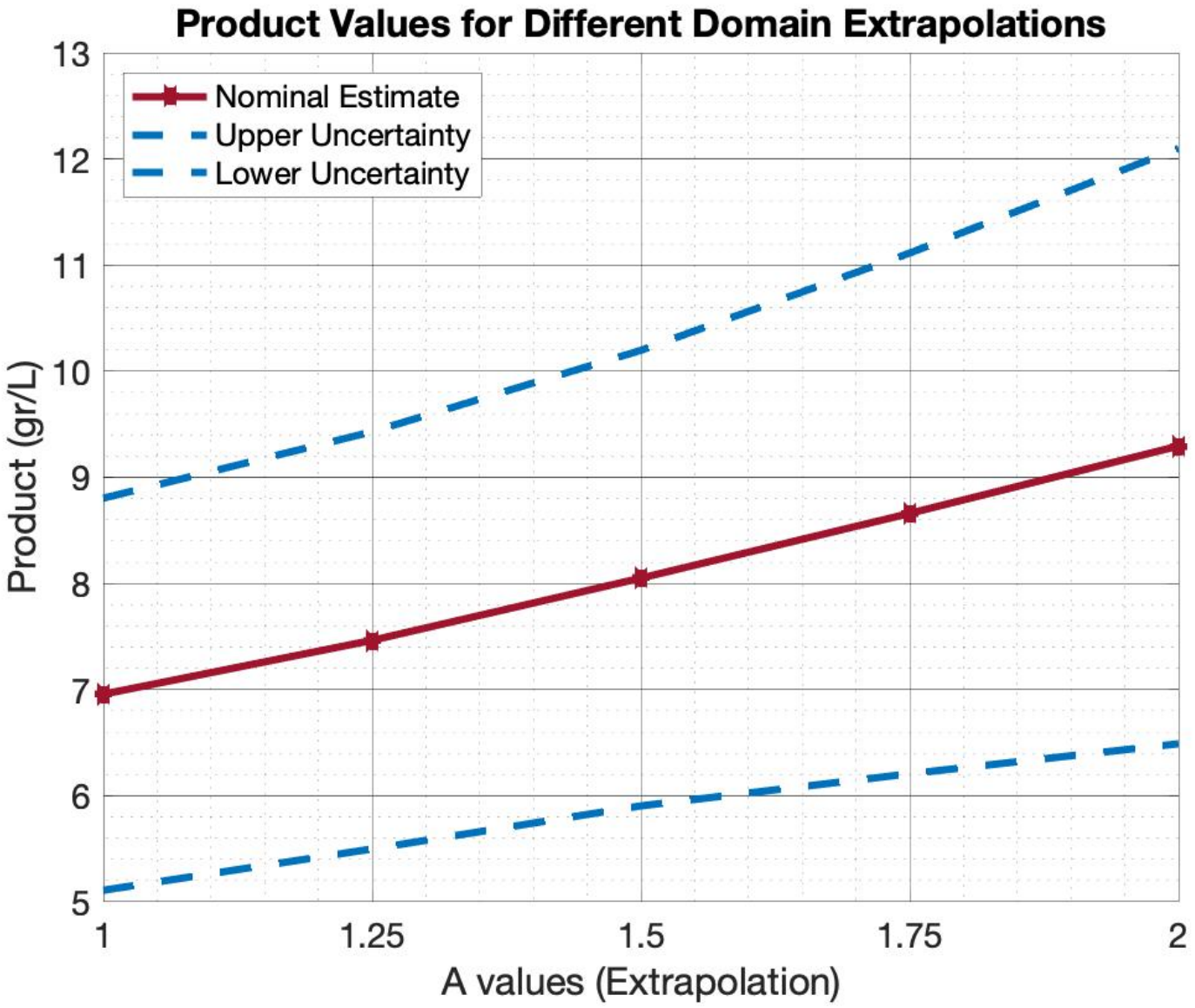
6 Factors (T, pH, nutrients) + Quadratic DRSM = 34 Experiments



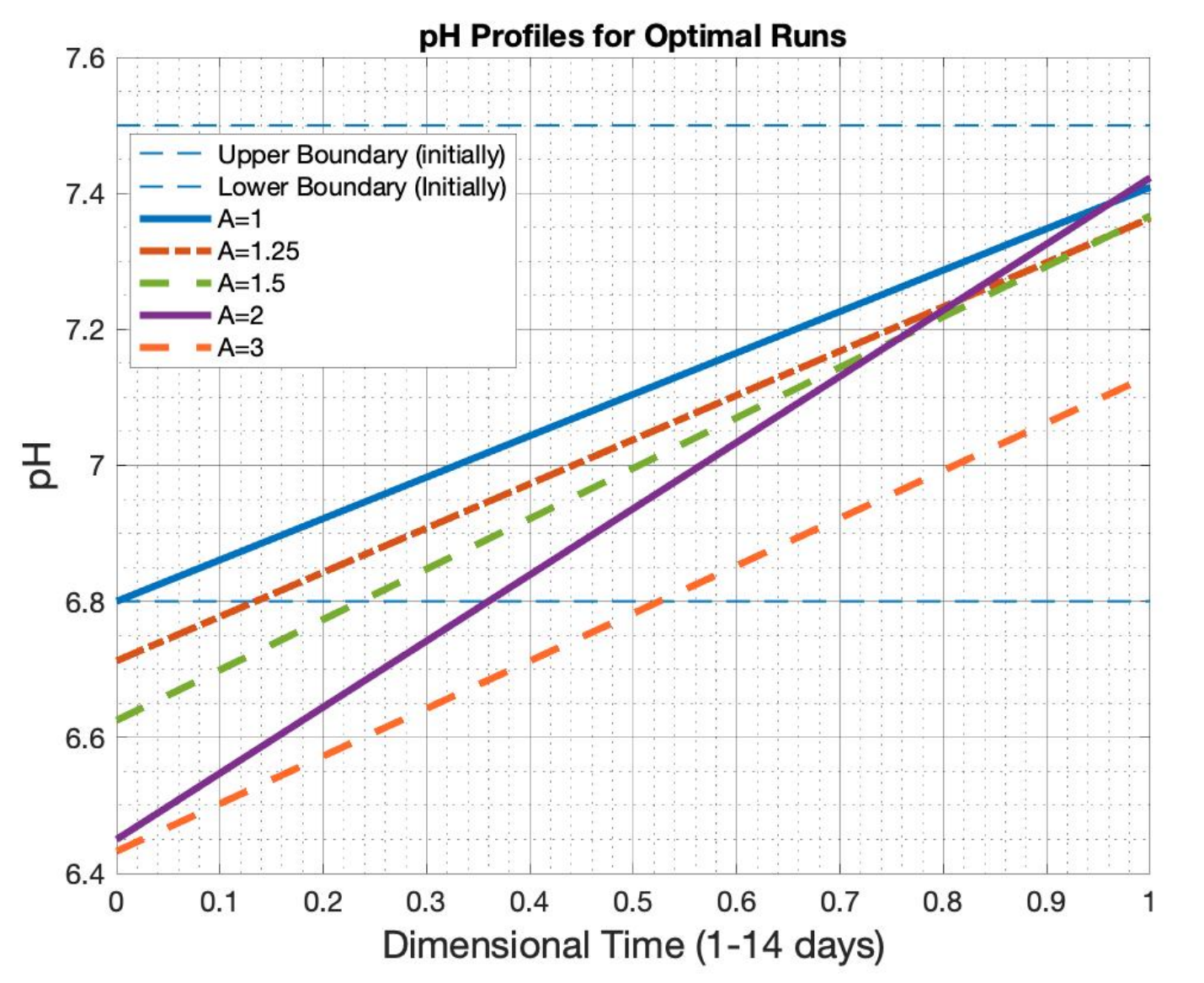
DRSM Model for One (of 10) Output



Optimal Values of Product (extrapolation)



Optimal pH Profiles



Part E

The FUTURE Challenges

Challenges: Stoichiometry & Kinetics

- Identify ALL Candidate Stoichiometries
  - Postulate All Reactions with
    - 1 or 2 Reactants or Products
    - Stoichiometric Coefficients: -2, -1, 1 , 2
  - Impose Mass Balance
  - Project to DRSM Models (TFA)
- Constraints For Rxn Stoichiometries
  - Monotonic Extents of Reaction
  - Estimate Kinetic Models
  - Which Fits Data Best?
- Additional Complications
  - Reversible Reactions
  - Form of Kinetic Models

Challenges (Cont'd)

- Mass Balances of Candidate Stoichiometries
  - Know MW Accurately OR ± 1 ?
  - Unmeasured Species
    - How Many and Which?
    - When Methodology Breaks Down?
- Atom Balances?
- Challenges with Reaction Orders
    - Elementary Step?
    - General Power Law – to be Estimated
    - Denominator Terms – Which Ones

Useful Toy Example

- Simulation of $A \rightarrow B \rightarrow C$, 1st Order Kinetics
  - DoE, Collect $A(t_k), B(t_k)$, $C(t_k)$ Data
- Estimate $DRSM_A$, $DRSM_B$, $DRSM_C$
  - Rank of Data = 2 ➔ 2 Reactions
- Reactions with Mass Balance & Projection
  - $A \leftrightharpoons B$, $B \leftrightharpoons C$, $A \leftrightharpoons C$,
  - $A + C \leftrightharpoons 2B$, $A + B \leftrightharpoons 2C$ $2A \leftrightharpoons B + C$
- Possible Stoichiometries with 2 Rxns: $\binom{6}{2}=15$
  - Group of 2 that are Linearly Independent

Screening Stoichiometries

❖ Monotonic Extent of Reactions: 6 of 15

Group	G1	G2	G3	G4	G5	G6
Rxn-1	$A \rightarrow B$	$A \rightarrow B$	$A \rightarrow B$	$A + C \rightarrow 2B$	$A + C \rightarrow 2B$	$A + B \rightarrow 2C$
Rxn-2	$B \rightarrow C$	$2B \rightarrow A + C$	$A + B \rightarrow 2C$	$B \rightarrow C$	$A + B \rightarrow 2C$	$C \rightarrow B$

❖ Starting Material Restriction (Only A)

Group	G1	G2	G3
Rxn-1	$A \rightarrow B$	$A \rightarrow B$	$A \rightarrow B$
Rxn-2	$B \rightarrow C$	$2B \rightarrow A + C$	$A + B \rightarrow 2C$

❖ Estimate Kinetics

➢ Fit 3 Models with 2 Reactions Each

- ❑ Assume Elementary Reactions
  - Stoichiometry ➔ Kinetic Order
- ❑ Fit Each Reaction Rate **Separately**

Fitting of Reaction Kinetics

	Group 1 (True)		Group 2		Group 3	
Each reaction						
Reactions	$A \rightarrow B$	$B \rightarrow C$	$A \rightarrow B$	$2B \rightarrow A + C$	$A \rightarrow B$	$A + B \rightarrow 2C$
SSE	0.28	2.6	3.8	46	18	44
RMSE	0.07	0.23	0.24	0.83	0.52	0.81
R^2	0.98	0.77	0.78	0.53	0.13	0.58
Overall group*						
SSE	1.44		24.9	(+1629%)	31	(+2053%)
RMSE	0.15		0.535	(+257%)	0.665	(+343%)
R^2	0.875		0.655	(-25%)	0.355	(-59%)

*Values of SSE, RMSE and R^2 are averaged over those reactions in the group

Correct Model Fits the Data the Best

How Can we Do this for Complex Mixtures?

Should Remember Tomorrow

- Novelty of **DoDE** & **DRSM**
  - **DoDE**: Time Varying Inputs
  - **DRSM**: Modeling Time-Varying Outputs
- Stoichiometric Identification
  - DoDE/DRSM ➔ Process Knowledge
- Optimization Using DoDE/DRSM
  - Results Close to Model-Based Optimization
- DoDE & DRSM Applicable to:
  - Batch & Continuous Processes &
  - Chemical & Biological Processes
- Future Challenges
  - Implications of Unmeasured Species
  - Automatic List of Stoichiometries
  - Comprehensive Kinetic Modeling

Thank You for Your Attention

Will You JOIN Us
in this Exiting Trip?