

# Data-driven Identification of Stoichiometric and Kinetic Models for Complex Reaction Mixtures

An Honors Thesis for the Department of Chemical & Biological Engineering.

Jenna Fromer

Tufts University, 2021.

## Abstract

Target Factor Analysis (TFA) quantifies whether a hypothesized candidate stoichiometry is in agreement with time-resolved composition data from a reacting mixture. This has been recently aided by the Dynamic Response Surface Modeling (DRSM) methodology (Klebanov, 2016), where continuous concentration profiles are estimated from discrete compositional measurements. In the present thesis, we remove the requirement to postulate candidate stoichiometries. We define an algorithm for the identification of stoichiometric and kinetic models that accurately model the reaction mixture data. Initially, all possible reaction stoichiometries that satisfy the mass balance constraint and pass the TFA test are identified. The resulting stoichiometric candidates are combined to form full rank candidate reaction networks of multiple reactions each. Several of these networks are filtered out through additional constraints. Kinetic models of different forms are estimated separately for each reaction of the remaining reaction networks. The accuracies of competing reaction networks are compared among themselves using F-tests of the corresponding regression sums of squares. We subsequently test if all non-random data are represented by the most accurate network. Three case studies have been analyzed involving three, six, and eleven species, respectively. For the first two systems, involving two and three reactions, respectively, the true stoichiometric and kinetic models were identified. For the complicated case of eleven species and eight reactions, the obtained models from the first pass of the proposed algorithm fail to represent all non-random data. This necessitates the need of a second modeling cycle.

## **Acknowledgements**

First, I would like to acknowledge my thesis advisor Professor Christos Georgakis for sharing this project with me and helping me grow into an independent researcher. His academic and research insights have made progress in this project possible, and his mentorship has largely inspired my own career aspirations. I would also like to thank Professor Kyongbum Lee and Dr. Jason Mustakis, who served on the thesis committee and provided invaluable suggestions during the progression of the project. I especially appreciate Dr. Mustakis' insights, which ensured that the research problems addressed accurately represented realistic ones seen in the pharmaceutical industry. Finally, I would like to acknowledge Dr. Yachao Dong for beginning work on this project and developing the DRSM, which aided my progress significantly.

# Table of Contents

<b>Abstract</b> .....	<b>ii</b>
<b>Acknowledgements</b> .....	<b>iii</b>
<b>Table of Contents</b> .....	<b>iv</b>
<b>List of Tables</b> .....	<b>v</b>
<b>List of Figures</b> .....	<b>v</b>
<b>Chapter 1: Introduction</b> .....	<b>1</b>
<b>Chapter 2: Methodology</b> .....	<b>4</b>
2.1. DATA COLLECTION AND DRSM FITTING .....	4
2.2. STOICHIOMETRIC REACTION ENUMERATION .....	5
2.3. TARGET FACTOR ANALYSIS .....	7
2.4. ENUMERATION OF POSSIBLE NETWORKS .....	8
2.5. DYNAMIC FILTERS .....	9
2.6. KINETIC MODEL IDENTIFICATION .....	12
2.7. SELECTION OF MOST ACCURATE REACTION NETWORK .....	13
<b>Chapter 3: Example Reacting Systems</b> .....	<b>17</b>
3.1. CASE STUDY A: THREE SPECIES INVOLVED IN TWO REACTIONS.....	17
3.1.1. <i>Data Collection</i> .....	17
3.1.2. <i>Network Enumeration and Filtering</i> .....	19
3.1.3. <i>Kinetics and Statistical Analysis</i> .....	21
3.2. CASE STUDY B: SIX SPECIES INVOLVED IN THREE REACTIONS .....	24
3.2.1. <i>Data Collection</i> .....	24
3.2.2. <i>Network Enumeration and Filtering</i> .....	25
3.2.3. <i>Kinetics and Statistical Analysis</i> .....	26
3.3. CASE STUDY C: 11 SPECIES INVOLVED IN EIGHT REACTIONS .....	29
3.3.1. <i>Data Collection</i> .....	29
3.3.2. <i>Network Enumeration and Filtering</i> .....	30
3.3.3. <i>Kinetics and Statistical Analysis</i> .....	31
<b>Chapter 4: Discussion</b> .....	<b>34</b>
<b>Chapter 5: Conclusions</b> .....	<b>39</b>
<b>References</b> .....	<b>41</b>
<b>Appendix</b> .....	<b>42</b>

## List of Tables

Table 1: Design of Experiments for Case Study A.....	18
Table 2: Passing Stoichiometric Reactions in Case Study A.....	19
Table 3: Stoichiometric Reactions and Kinetic Models of Case Study A .....	21
Table 4: Statistical Comparison of Networks for Case A.....	22
Table 5: True and Estimated Kinetic Parameters .....	23
Table 6: Molecular Weights of Species in Case B .....	25
Table 7: Factors of Experimental Design for Case B .....	25
Table 8: Passing Stoichiometric Reactions in Case Study B.....	26
Table 9: Stoichiometric Reactions and Kinetic Models of Case B.....	27
Table 10: Comparison of Kinetic Parameters for Case B.....	27
Table 11: Molecular Weights of Species in Case C .....	30
Table 12: Best Performing 5-Reaction Networks for Case C.....	31
Table 13: Overall Best Performing Networks for Case C .....	33

## List of Figures

Figure 1: Reaction Rates of Failed Network.....	20
Figure 2: Estimated and True Concentration Profiles for Case A .....	23
Figure 3: Estimated and True Concentration Profiles for Case B .....	28

## Chapter 1: Introduction

The increasing popularity of high-throughput instrumentation has facilitated the availability of large quantities of reaction mixture robotic experiment data. With increasing ease of collecting time-resolved concentration measurements, scientists and engineers inevitably observe unexpected compounds in the reaction mixture, which might be by-products and intermediates. While the overall stoichiometry between the major reactants and products is usually well-understand, the stoichiometric reactions which generate intermediates and by-products are often unknown.

The desire to extract information and knowledge about the reacting systems of interest from the collected data has motivated research into data-driven modeling. Target factor analysis (TFA), for example, determines whether a candidate stoichiometry is in agreement with time-resolved composition data<sup>1</sup> and has been shown to aid in the identification of pharmaceutical reactions<sup>2</sup>. The dynamic response surface methodology (DRSM), a means to estimate input-output measurement from design of experiments (DoE) data sets, is shown to enhance TFA performance in stoichiometric reaction identification<sup>2</sup>. In short, the

DRSM models provide a smoothed concentration profile of the collected data, enabling TFA to yield more accurate projection scores. A projection score greater than 90%<sup>2</sup> indicates that the candidate reaction passes the TFA test and is in agreement with the mixture data.

A major limitation of TFA is the postulation of candidate stoichiometries. Prior publications have thus assumed that it is sufficient for the chemist to hypothesize all reactions that might be occurring in the reaction system under investigation. However, target factor analysis will not be sufficient in identifying all reaction stoichiometries if the set of candidate reactions postulated are not exhaustive. Additionally, any stoichiometry that is a linear combination of passing stoichiometries will also pass the TFA test. Even with sufficient stoichiometric candidates, TFA does not reveal which linear combinations of qualified reactions constitute the most desirable reaction network.

To alleviate the aforementioned issues, we propose an approach in which all potential reaction stoichiometries are considered, removing the necessity to intuitively define reaction candidates for the system. Static constraints, such as mass balance and target factor analysis, are used to limit the possibilities of qualified reactions. Next, linearly independent sets from the qualified reactions, defined as reaction networks, are examined with respect to the monotonicity of the corresponding back-calculated reaction rates. Additional filters, motivated by chemical reaction engineering knowledge to be described later, limit the network possibilities as well. Kinetic models are subsequently postulated and fitted to the data. To finally select the most appropriate stoichiometric and kinetic model,

statistical tools for the goodness of the fit are used. Kinetic model identification approaches have been studied in depth<sup>3,4</sup>, but these studies assume that the stoichiometric reactions occurring in the system are known. Our innovation here is in the use of kinetic modeling to aid in stoichiometric identification.

While the most general solution of the above challenging problem is beyond the scope of the present thesis, three example case studies are examined in detail to determine the effectiveness of our proposed approach in aiding exhaustive stoichiometric identification.

## **Chapter 2: Methodology**

Automatic stoichiometric and kinetic model identification necessitates evaluating all possible combinations of the measured species, both as reactants or products, assuming mass balance between those participating in a candidate reaction is satisfied. We design our algorithmic approach to filter out infeasible reaction choices before forming reaction networks. Next, we further limit the reaction network choices by ensuring that the rate of a reaction does not change sign during an experiment. Kinetic models for each reaction in a candidate networks are estimated, and statistical tests on the residual sums of squares are used to compare the performance of competing networks through their kinetic models.

### **2.1. Data Collection and DRSM Fitting**

We begin by designing a set of experiments for the studied reaction mixture using the Design of Experiments methodology. In order to estimate the activation energies of participating reactions, reaction temperature must be one experimental factor. Initial concentrations of participating species may also be considered factors.

In this study, we use simulated time-resolved concentration data as the starting point. The time-resolved data, collected at  $n_k$  time instants throughout each experiment for each of the  $n_s$  species, are fit to a Dynamic Response Surface Model (DRSM)<sup>5</sup>. The DRSM smoothens the data, allows for interpolation between measured time points and at different values of the factors, and provides a means of estimating the time derivative of the species' concentrations at any point throughout any experiment. We define an  $n_k \times n_s$  matrix  $\mathbf{y}_e(\mathbf{t})$  for each experiment as:

$$\mathbf{y}_e(\mathbf{t}) = \begin{bmatrix} \frac{dC_{A,1}}{dt} & \frac{dC_{B,1}}{dt} & \dots & \frac{dC_{n_s,1}}{dt} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{dC_{A,n_k}}{dt} & \frac{dC_{B,n_k}}{dt} & \dots & \frac{dC_{n_s,n_k}}{dt} \end{bmatrix} \quad (1)$$

where  $\frac{dC_{s,k}}{dt}$  denotes the rate of appearance or disappearance of a species  $s$  at time point  $k$  for experiment  $e$ . We also define the data matrix  $\mathbf{D}$  which includes the time derivatives for all experiments stacked:

$$\mathbf{D} = \begin{bmatrix} \mathbf{y}_1(\mathbf{t}) \\ \mathbf{y}_2(\mathbf{t}) \\ \vdots \\ \mathbf{y}_{n_e}(\mathbf{t}) \end{bmatrix} \quad (2)$$

The  $(n_k n_e) \times n_s$  matrix  $\mathbf{D}$  includes the concentration time derivatives for all  $n_e$  experiments.

## 2.2. Stoichiometric Reaction Enumeration

We next enumerate all possible stoichiometric reactions between the measured species that adhere to specific constraints. First, the absolute value of any stoichiometric coefficient may not exceed two. Second, there cannot be more

than two reacting species in a stoichiometric reaction, and similarly, there may be no more than two products species. Third, the sum of the absolute values of all stoichiometric coefficients in a stoichiometric reaction may not exceed five. In other words, we do not include reactions which involve more than five molecules. Finally, if all stoichiometric coefficients can be evenly divided by two, the reaction is not considered.

These constraints are described by the following statements:

$$\begin{aligned}
 & \text{(i) } |v_i| \leq 2 \quad \text{for } i = 1, 2, \dots, n_{ps} \\
 & \text{(ii) } n_{reac} \leq 2 \\
 & \text{(iii) } n_{prod} \leq 2 \\
 & \text{(iv) } \sum_{i=1}^{n_{ps}} |v_i| \leq 5 \\
 & \text{(v) } \text{gcd}(v_1, v_2, \dots, v_{ps}) = 1
 \end{aligned} \tag{3}$$

for  $n_{reac}$  as the number of reacting species,  $n_{prod}$  as the number of produced species,  $n_{ps}$  as the total number of participating species, and  $v$  as a stoichiometric coefficient. We subsequently eliminate reactions which do not fulfill a mass balance between reactants and products, defined by the following equation:

$$\sum_{i=1}^{n_{ps}} v_i MW_i = 0 \tag{4}$$

where  $MW_i$  is the molecular weight of species  $i$ . Note that, here, the reactants' stoichiometric coefficients are assumed to be negative, while those of the products are positive. Reactions which do not fulfill this mass balance are filtered out. As a result of this mass balance constraint, a substantial number of candidate reactions

are eliminated. Here we assume that the molecular weights for all participating species are known accurately.

### 2.3. Target Factor Analysis

Each of the stoichiometric reactions that fulfill mass balance constraints is evaluated with Target Factor Analysis (TFA), as described in detail by Dong et al.<sup>2</sup>. In summary, TFA begins with the singular value decomposition (SVD) of matrix  $\mathbf{D}$  as follows<sup>1</sup>:

$$\mathbf{D} = \mathbf{U}_D \mathbf{S}_D \mathbf{V}_D^T \quad (5)$$

Here, we used the reduced SVD representation where  $\mathbf{S}_D$  is a square matrix which contains the non-zero singular values of  $\mathbf{D}$  on the diagonal, and  $\mathbf{U}_D$  and  $\mathbf{V}_D$  contain the corresponding left-hand and right-hand singular vectors, respectively. Because it is unlikely that  $\mathbf{S}_D$  is full rank, we must calculate the number of *significant* singular values which correspond to the number of linearly independent reactions occurring in the system, often less than  $n_s$ . We therefore calculate the number of significant singular values,  $n_{SV}$ , using the Malinowski test<sup>6</sup>, and data corresponding to the remaining insignificant singular values are not used in the remaining TFA calculations. To test if a candidate reaction is in agreement with the data, a projection matrix is calculated with the right-hand vectors in  $\mathbf{V}_D$  as follows:

$$\mathbf{P} = \mathbf{V}_{D1} \mathbf{V}_{D1}^T \quad (6)$$

in which  $\mathbf{V}_{D1}$  contains  $n_{SV}$  right-hand vectors which correspond to the significant singular values. The  $n_s$ -dimensional candidate stoichiometry vector  $\mathbf{c}$

$$\mathbf{c} = [v_1, v_2, \dots, v_{n_s}]^T \quad (7)$$

defines the candidate stoichiometry with  $v_i$  negative for reactants. The candidate  $\mathbf{c}$  and its response vector  $\mathbf{r}$  are used to calculate the projection score  $p_{sc}$  using:

$$p_{sc} = 100 \left( 1 - \frac{\|\mathbf{r} - \mathbf{c}\|}{\|\mathbf{t}\|} \right), \quad \mathbf{r} = \mathbf{P}\mathbf{c} \quad (8)$$

Candidates with a projection score above 90 are considered to be sufficiently in agreement with the time-resolved reaction mixture data as modeled by the DRSM. Reactions with scores lower than 90 will be filtered out at this stage.

We label the mass balance and TFA constraints as “static” constraints, as they do not relate to the time evolution of the reactions. They are applied to single candidate stoichiometric reactions before reaction networks have been enumerated.

#### 2.4. Enumeration of Possible Networks

The rank of all stoichiometric reactions which passed the TFA constraint is calculated. If the rank is equal to the number of significant singular values, the number of reactions per network is equal to the number of significant singular values. Alternatively, if the rank of passing stoichiometries is less than the number of significant singular values, the number of reactions per network,  $n_r$ , will be equal to the determined rank.

All stoichiometric reactions that fulfill the static constraints are grouped together into networks of  $n_r$  reactions. If  $n_{sc}$  stoichiometric reactions pass the static constraints, the number of potential networks is defined as:

$$n_{net} = \binom{n_{sc}}{n_r} = \frac{n_{sc}!}{n_r! (n_{sc} - n_r)!} \quad (9)$$

where  $n_{net}$  is the number of possible networks. For each candidate network, we define the  $n_r \times n_s$  stoichiometric matrix for each network as

$$\mathbf{N} = [\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{n_r}]^T \quad (10)$$

where each row of  $\mathbf{N}$  contains a stoichiometric reaction vector  $\mathbf{c}$  defined by Equation 7. If  $\mathbf{N}$  is not full rank, the candidate network is filtered out. Therefore, the total number of networks considered in the next stage is equal to or less than  $n_{net}$ , as some of those initially enumerated may not be full rank.

In complex cases, the overall stoichiometric reaction is known, and we then include an additional constraint in which the known overall reaction must be some linear combination of the included reaction in a network.

All possible combinations of stoichiometric reactions and corresponding  $\mathbf{N}$  matrices that fulfill the applied constraints are enumerated. This list of candidate networks is further filtered in the following steps.

## 2.5. Dynamic Filters

At this point, we employ a set of “dynamic” filters that may only be applied once a network has been defined. For each network, we first calculate the rates of involved reactions by multiplying the concentration derivative data modeled by the DRSM by the pseudo-inverse of the stoichiometric matrix, as follows:

$$\mathbf{r}(\mathbf{t}) = \mathbf{D}\mathbf{N}^+ \quad (11)$$

where  $\mathbf{D}$  and  $\mathbf{N}$  are as defined in equations 2 and 10, respectively. Also,  $\mathbf{r}(\mathbf{t})$  is an

$(n_e n_k) \times n_r$  matrix with the rates of each reaction in the network at each time point in all experiments. We next consider the confidence intervals of the calculated reaction rates. The half-widths,  $B_y$ , of the confidence interval of each concentration derivative at all observations are determined from the DRSM model<sup>5</sup> as

$$B_y = t_{0.975,n} \sigma_y \quad (12)$$

where  $t_{0.975,n}$  is the t-statistic with 95% confidence and  $n$  degrees of freedom, and  $\sigma_{y_d}$  is the standard deviation of the concentration time derivative at specified observation conditions.

We assume that the degrees of freedom for all observations are equal and that the degrees of freedom for the reaction rate uncertainties are equal to those of the concentration time derivative. Therefore, we can calculate the reaction rate half-width matrix  $\mathbf{B}_r^2$ , of size  $(n_e n_k) \times n_r$ , from the half-widths of the concentration time derivatives collected in matrix  $\mathbf{B}_y^2$ . Because all values in matrix  $\mathbf{N}$  have zero variance, we can estimate the half-width matrix for rates of reaction as:

$$\mathbf{B}_r^2 = \mathbf{B}_y^2 \mathbf{N}^+ \quad (13)$$

The half widths of the confidence interval for reaction rates at each observation are calculated as the square root of the individual entries in  $\mathbf{B}_r^2$ .

With the rates of reaction and corresponding uncertainties calculated, we now employ the dynamic constraints. Three such constraints are defined below:

- a) The reaction rate of each reaction in the candidate network must always be positive or negative throughout each experiment.
- b) At least one reaction in the network must be able to proceed with the species present at  $t = 0$ , the beginning of the experiment. (14)
- c) If a species acts only as a reactant in an experiment, it must be present at the beginning of the experiment.

We apply these constraints to each network in parallel.

First, with the time resolved reaction rates for the studied network, we can confirm that reaction rates remain positive or negative for all times that their values are significant within each experiment. We test whether a change in the reaction rate sign is solely due to error by determining whether a reaction rate of zero is included in the confidence interval of the calculated rate with an opposite sign. If a zero rate is included in the confidence interval at that time point, the change in sign is due to error. If the zero rate is not included in the confidence interval, the change in sign is not due to error, and the network is eliminated.

Second, once the direction in which each reaction in a network proceeds is determined through the calculation of the corresponding reaction rates, we identify which reactants must be initially present for each reaction network to proceed. If the reactants of at least one reaction in the network are present in the initial mixture in each experiment, this network passes the second dynamic constraint. Networks in which no reaction can proceed with the given starting materials are eliminated.

Similarly, the directions of all reactions in a network dictate which species act only as products, only as reactants, or both throughout experiments. A species that acts only as a reactant throughout an experiment must have a non-zero initial concentration. Networks that do not fulfill this constraint are filtered out.

## 2.6. Kinetic Model Identification

Finally, several kinetic models are estimated for each reaction in a network that passed all prior constraints. We assume that each reaction can be accurately modeled by reversible or irreversible elementary rate laws shown below:

$$\begin{aligned}
 \text{a) } r &= k \prod_{i=1}^{n_{\text{react}}} C_i^{\nu_i} \\
 \text{b) } r &= k_f \prod_{i=1}^{n_{\text{react}}} C_i^{\nu_i} - k_r \prod_{j=1}^{n_{\text{prod}}} C_j^{\nu_j} \\
 k_i &= k_{i0}(T_0) \exp\left(-\frac{E_i}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right)
 \end{aligned} \tag{15}$$

We define the reactants and products of the system to be those dictated by reaction rate and determined during the dynamic filtering stage.

For all  $n_r$  reactions in each network, we fit the reaction rate and concentration data to the above two models. Note that the power laws are set as the stoichiometric coefficient of corresponding species for the explored kinetic models. We consider the models for both the forward and reverse reaction directions, in which case reactants would instead be products, and vice versa, totally four different kinetic models. The MATLAB function *fmincon* is utilized to estimate the optimal rate constants  $k$  and activation energies  $E$  which minimize the following objective function:

$$\Phi_{min} = \sum_{i=1}^{n_o} (r_{m,i} - r_i)^2 \quad (16)$$

where  $n_o$  is the total number of observations,  $r_{m,i}$  is the modeled rate at observation  $i$ , and  $r_i$  is the observed rate, as calculated in Equation 11, at observation  $i$ . Because the DRSM model allows for interpolation between experimental measurement time points, we increase the number of observations used in the parameter estimation by 3 times, calculating  $n_o$  as:

$$n_o = 3n_e n_k \quad (17)$$

Our parameter estimation task will therefore include an increased number of data points than the experimental observations.

We apply constraints on the kinetic constant and activation energy parameters to ensure they remain non-negative. The BIC values of all models are calculated, and the kinetic model with the lowest BIC is selected to best represent the data. This process is repeated for each reaction of all remaining networks.

We complete this phase with a collection of networks, each of which contains a set of stoichiometric reactions and corresponding kinetic models and parameters. The final step is to identify which reaction network, represented by its stoichiometric and kinetic model, best represents the data.

## 2.7. Selection of Most Accurate Reaction Network

Once the kinetic models for all reactions in a network are identified, the network model is complete. The MATLAB ODE solver *ode15s* enables prediction of concentration profiles based on the network's stoichiometric and kinetic models. We simulate each experiment in the original data collection by

setting the temperature and initial species concentrations. The difference between the original concentration profiles and the modeled profiles is quantified by calculating the model's regression sum of squares as follows:

$$SS_r = \sum_{s=1}^{n_s} \sum_{e=1}^{n_e} \sum_{k=1}^{n_k} (C_{s,e,mod}(T_k) - C_{s,e,obs}(T_k))^2 \quad (18)$$

where  $n_e$  is the number of experiments performed.  $C_{s,e,mod}(T_k)$  and  $C_{s,e,obs}(T_k)$  are, respectively, the modeled and observed concentration for species  $s$  at the  $k$ -th time point in experiment  $e$ . The degrees of freedom for the sum of squares is calculated as:

$$DoF_r = n_s n_k - n_p \quad (19)$$

in which  $n_p$  is the total number of parameters used in the network's kinetic models. With each network's  $SS_r$  and  $DoF_r$  at hand, we are able to compare the accuracy with which each reaction network models the concentration data. In order to compare the performance of two networks, we perform an F-test between the two regression sums of squares. We first rank the networks with from the smallest to the largest  $SS_r$  values. The series of F-tests we perform aim to discover which  $SS_r$  values are statistically different than that of the minimum value. Each test is formally stated as:

$$\text{null hypothesis: } H_0: SS_{r,0} = SS_{r,i} \quad (20)$$

$$\text{alternative hypothesis: } H_1: SS_{r,0} < SS_{r,i}$$

where  $SS_{r,0}$  is of the model with the smallest  $SS_r$  and  $SS_{r,i}$  is of any other network  $i$ . The appropriate statistic is the ratio of the two corresponding means sum of squares:

$$F_0 = \frac{SS_{r,i}/DoF_{r,i}}{SS_{r,0}/DoF_{r,0}} \quad (21)$$

The null hypothesis is rejected if

$$F_0 > F_{\alpha, DoF_{r,0}, DoF_{r,i}} \quad (22)$$

For 95% confidence level,  $\alpha$  is 0.05. Instead of performing this inequality,

though, we calculate the  $p$ -value corresponding to the  $F_0$  statistic as

$$p = 1 - \int_0^{F_0} F_{x, DoF_{r,0}, DoF_{r,i}} dx \quad (23)$$

which is equal to the upper tail of the  $f$  distribution at  $F_0$ . The resulting  $p$  values allows us to make the following decision:

- (1) If  $p \leq 0.05$ , the null hypothesis is rejected.
- (2) If  $p > 0.05$ , the null hypothesis fails to be rejected.

If  $p \leq 0.05$  for the comparison with the second-best performing network, we confirm that this network outperforms all others and is the most accurate model for the studied reacting system. We may also perform this test for all other remaining networks, comparing the best performing network with every other competing network. Comparisons with  $p$ -values greater than 0.05 reveal a network that performs equally well as the network with the lowest  $SS_r$ .

If we cannot confirm with certainty that one network outperforms all others, we are left with a less straightforward outcome and may continue in one of multiple directions. First, we may propose additional experiments to perform which may reveal data that distinguishes the performance among the two networks. Second, we may consider controlling the number of significant singular values used in Target Factor Analysis, instead of relying on the Malinowski test<sup>6</sup>,

which may alter the projection scores of stoichiometric reactions. We must consider that the true reacting system may be a set of linearly dependent reactions, in which case the *true* complete set of stoichiometric reactions will not be possibly discovered from our algorithm.

## Chapter 3: Example Reacting Systems

In order to evaluate the performance of our approach, we apply our methodology to three separate reacting systems and observe the stoichiometric and kinetic models identified.

### 3.1. Case Study A: Three Species Involved in Two Reactions

We first consider a set of simulated data for a system of three species involved in two reactions. In the kinetics and statistical analysis section (3.1.3), we show that the identified stoichiometric and kinetic model is the same as the true one. The estimated kinetic parameters are also almost identical to the true parameters.

#### 3.1.1. Data Collection

Here, we apply the described methodology to identify the best stoichiometric and kinetic model for a reacting mixture with three species ( $A$ ,  $B$ , and  $C$ ) with identical molecular weights, involved in two irreversible reactions with the following characteristics:



in which

$$\begin{aligned}
k_1 &= k_{10} \exp\left(-\frac{E_1}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right) \\
& k_{10} = 0.5 \text{ hr}^{-1}, E_1 = 1.987 \frac{\text{kcal}}{\text{mol}}, T_0 = 473.15 \text{ K} \\
k_2 &= k_{20} \exp\left(-\frac{E_2}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right) \\
& k_{20} = 1.3 \text{ hr}^{-1}, E_2 = 1.987 \frac{\text{kcal}}{\text{mol}}, T_0 = 473.15 \text{ K}
\end{aligned}
\tag{26}$$

To simulate actual reaction data, we added a constant measurement error to the simulation results for each collected concentration measurement, as follows:

$$C_e = C_s + N(0, \sigma) \tag{27}$$

Here,  $C_e$  is the observed concentration and  $N(0, \sigma)$  is a randomly distributed number with mean 0 and a standard deviation equal to  $\sigma$ . The simulated concentration ( $C_s$ ) is thus observed as  $C_e$ . In this case,  $\sigma$  is set to 0.02.

**Table 1: Design of Experiments for Case Study A**

#	1	2	3	4	5	6	7	8	9
Temperature(°C)	0	0	100	100	0	50	50	50	50
$[B]_0$ (mol/L)	0	0.5	0	0.5	0.25	0	0.25	0.25	0.25

The center composite design is used for this set of experiments which includes two factors: temperature and initial concentration of species  $B$ .

We designed the experiments using the DoE methodology with two factors: temperature and initial concentration of  $B$ . A center composite design was used with three center point replicates. The initial concentration of  $A$  was 1 mol/L for all experiments, and species  $C$  was not present at the beginning of any experiment. Factors for each of the nine experiments are listed in Table 1 for each of the nine experiments. Experiments were run for a total batch time of 12 hours,

and concentration measurements were collected at seven points throughout the batch time. After collecting the data, a DRSM model was calculated.

### 3.1.2. Network Enumeration and Filtering

Following the described methodology above, we begin this stage by enumerating stoichiometries that fulfill the mass balance constraint between reactants and products. In this case, 6 reaction stoichiometries passed the mass balance constraint. All of them passed the TFA test as well, and the specific reactions are shown in Table 2. The Malinowski test<sup>6</sup> revealed two significant singular values, and the rank of the passing stoichiometric reactions was two. Therefore, from these stoichiometric reactions, 15 full rank networks of two reactions each were enumerated.

**Table 2: Passing Stoichiometric Reactions in Case Study A**

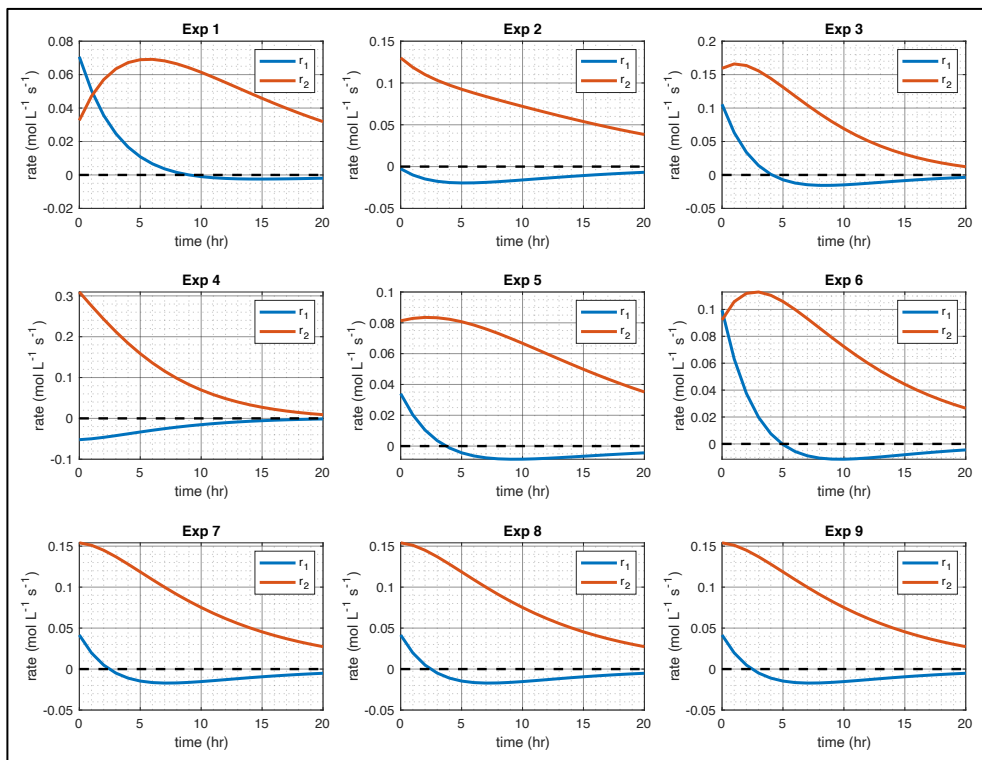
$r_1:$	$A + C \rightleftharpoons 2B$
$r_2:$	$A \rightleftharpoons B$
$r_3:$	$B \rightleftharpoons C$
$r_4:$	$A \rightleftharpoons C$
$r_5:$	$2A \rightleftharpoons B + C$
$r_6:$	$A + B \rightleftharpoons 2C$

The six stoichiometric reactions which fulfilled the mass balance and TFA constraints are listed above.

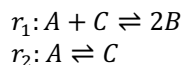
As described in Section 2.5, the reaction rates were calculated for each network using Equation 11. Networks that included at least one reaction rate whose significant values changed from positive to negative, or vice versa, in an experiment were filtered out. The calculated reaction rates of the two reactions shown in Figure 1 correspond to a reaction network that does not pass the dynamic filter. In multiple experiments, the rate of reaction 1 ( $r_1$ , blue) changes from positive to negative. We confirm that this change is significant.

Of the 15 networks possible, consisting of two of the six reactions in Table 2, six pass this first dynamic constraint related to the sign change of a reaction rate.

**Figure 1: Reaction Rates of Failed Network**



The calculated sets of reaction rates for each experiment are plotted. The first reaction transitions from a positive to negative rate in multiple experiments, and thus the corresponding network is filtered out by this constraint. A zero rate is denoted by a black dotted line. The two reactions of this network are:



The second dynamic constraint was implemented to ensure that at least one reaction in the network under consideration can proceed given the starting species, and 3 of the 6 remaining networks were filtered out. One such network that did not pass this constraint is defined by the two reactions below:



Note that the direction and reversibility of these reactions are unknown at this stage before the rate calculation. The rates of this reaction for each experiment are shown in Appendix (Figure A1). We observe that the reactions proceed in the forward direction for all experiments. However, multiple experiments only begin with species A, with the initial concentration of other species being zero. Neither reaction can proceed forward in these experiments, yet the reaction rates as calculated through Equation 11 result in forward rates. Thus, this network and others that failed an equivalent test were removed by the second dynamic constraint.

The remaining three networks which passed to the kinetic evaluation are listed in Table 3.

**Table 3: Stoichiometric Reactions and Kinetic Models of Case Study A**

Network	Stoichiometric Reactions	Kinetic Models	BIC Values for Kinetic Models	Kinetic Model SS
1	$r_1: A + C \rightleftharpoons 2B$ $r_2: B \rightarrow A$	$r_1 = k_1 C_A C_C - k_{-1} C_B^2$ $r_2 = k_2 C_B$	$r_1: -1.2 \times 10^3$ $r_2: -1.2 \times 10^3$	$r_1: 0.29$ $r_2: 0.28$
2	$r_1: A \rightarrow B$ $r_2: B \rightarrow C$	$r_1 = k_1 C_A$ $r_2 = k_2 C_B$	$r_1: -2.1 \times 10^3$ $r_2: -1.8 \times 10^3$	$r_1: 0.002$ $r_2: 0.01$
3	$r_1: A \rightleftharpoons B$ $r_2: A + B \rightarrow 2C$	$r_1 = k_1 C_A - k_{-1} C_B$ $r_2 = k_2 C_A C_B$	$r_1: -2.0 \times 10^3$ $r_2: -1.5 \times 10^3$	$r_1: 0.005$ $r_2: 0.07$

### 3.1.3. Kinetics and Statistical Analysis

The values of the two reaction rates are calculated for each reaction network under consideration for each experiment. They are regressed against the corresponding values for four different kinetic models (Equation 15). We consider both the forward and reverse models as possibilities. For each reaction, the kinetic model with the lowest BIC was chosen as the most accurate model, and the

resulting models are shown in Table 2. In the same table, we list the corresponding BIC and regression sum of squares values.

To compare the accuracy of the estimated stoichiometric-kinetic models, we calculated their predicted time-resolved concentration profiles and compared them to the measured concentration at each of the experiments. The regression sum of square errors of the model’s prediction was calculated according to Equation 18. To determine if the reaction network with the smallest sum of squares performs significantly better than both other networks, we performed an  $F$  test that is formally stated as follows:

$$\begin{aligned} \text{null hypothesis, } H_0: & \quad SS_{r,0} = SS_{r,i} \\ \text{alternative hypothesis, } H_1: & \quad SS_{r,0} < SS_{r,i} \end{aligned} \quad (29)$$

Here,  $SS_{r,0}$  represents the minimum sum of squares, in this case corresponding to Network 2, and  $SS_{r,i}$  is the sum of squares of the stoichio-kinetic models of another reaction network. The  $F_0$  statistic and corresponding  $p$ -value were calculated according to Equations 21 and 23, respectively. We fail to reject the null hypothesis if  $p$  is greater than 0.05. See Table 4 for the statistical results. Because the  $p$  values of both network comparisons against Network 2 are less than 0.05, we reject the null hypothesis and conclude that the stoichio-kinetic model of Network 2 is statistically better than the other two alternatives.

**Table 4: Statistical Comparison of Networks for Case A**

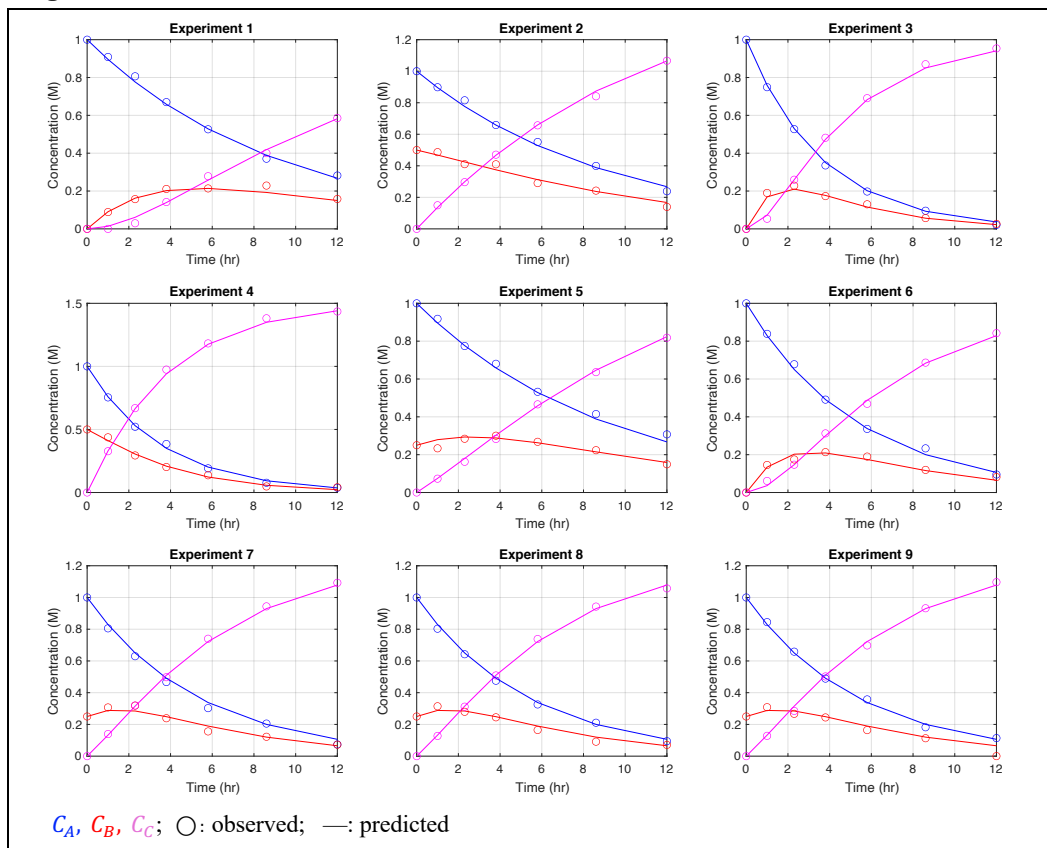
	DoF	MSE	$p$ value against Network 2
Network 1:	15	3.75	$1.05 \times 10^{-22}$
Network 2:	17	0.003	N/A
Network 3:	15	0.073	$1.01 \times 10^{-7}$

**Table 5: True and Estimated Kinetic Parameters**

Parameters	$k_{10}$ [hr <sup>-1</sup> ]	$k_{20}$ [hr <sup>-1</sup> ]	$E_1$ [kcal/mol]	$E_2$ [kcal/mol]
True	0.50	1.30	1.99	1.99
Estimated	0.47	1.25	1.86	1.92

We note that our algorithm correctly identified the true stoichiometric and kinetic model. The true and estimated model parameters are compared in Table 5.

Finally, we evaluate the performance of the identified stoichio-kinetic model visually in Figure 2. By comparing the model's predicted concentration profiles with the observed ones, we confirm that the model does accurately represent the reacting system.

**Figure 2: Estimated and True Concentration Profiles for Case A**

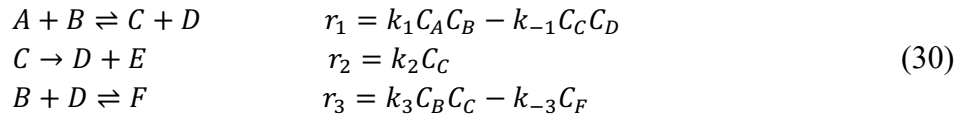
The true concentration profiles and estimated profiles using the identified stoichio-kinetic model are shown in circles and lines, respectively. Concentrations of species  $A$  (blue),  $B$  (blue), and  $C$  (magenta) are shown for all experiments performed.

### 3.2. Case Study B: Six Species Involved in Three Reactions

We next apply our methodology to identify the stoichiometric and kinetic models for a larger system consisting of six species and three reactions. As in the first case, we identify the correct stoichio-kinetic model, and the estimated parameters are quite close to the true ones.

#### 3.2.1. Data Collection

This system includes six species ( $A$ ,  $B$ ,  $C$ ,  $D$ ,  $E$ , and  $F$ ) involved in three reactions with the following characteristics:



in which

$$k_i = k_{i,0} \exp\left(-\frac{E_i}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right) \quad (31)$$

$$\begin{array}{ll}
 k_{1,0} = 0.5 \text{ L mol}^{-1} \text{ hr}^{-1} & E_1 = 3.974 \frac{\text{kcal}}{\text{mol}} \\
 k_{-1,0} = 0.05 \text{ L mol}^{-1} \text{ hr}^{-1} & E_{-1} = 3.775 \frac{\text{kcal}}{\text{mol}} \\
 k_{2,0} = 1 \text{ hr}^{-1} & E_2 = 1.987 \frac{\text{kcal}}{\text{mol}} \\
 k_{3,0} = 0.1 \text{ L mol}^{-1} \text{ hr}^{-1} & E_3 = 0.994 \frac{\text{kcal}}{\text{mol}} \\
 k_{-3,0} = 0.005 \text{ L mol}^{-1} \text{ hr}^{-1} & E_{-3} = 1.987 \frac{\text{kcal}}{\text{mol}} \\
 T_0 = 50^\circ\text{C} &
 \end{array}$$

Molecular weights are shown in Table 6. A constant measurement error was added to the simulation by using a  $\sigma$  of 0.004 (Equation 27). The factors of the designed experiments are listed in Table 7. The initial concentration of  $A$  was 1 mol/L for all experiments, and neither  $E$  nor  $F$  were present at the start of any experiment. A DRSM was evaluated with the simulated data.

**Table 6: Molecular Weights of Species in Case B**

Species	A	B	C	D	E	F
Molecular Weight (g/mol)	320	210	274	256	18	466

**Table 7: Factors of Experimental Design for Case B**

Experiment #	Temperature (°C)	$C_{B0}$ ( $\frac{\text{mol}}{\text{L}}$ )	$C_{D0}$ ( $\frac{\text{mol}}{\text{L}}$ )
1	70	1.0	1
2	90	0.8	2
3	90	1.2	2
4	70	1.0	1
5	90	0.8	0
6	50	0.8	2
7	50	1.2	0
8	90	1.2	0
9	50	1.0	1
10	70	1.0	2
11	70	1.0	0
12	70	0.8	1
13	50	0.8	0
14	70	1.0	1
15	70	1.2	1
16	50	1.2	2
17	90	1.0	1

### 3.2.2. Network Enumeration and Filtering

For this case, seven stoichiometric reactions, shown in Table 8, fulfilled the mass balance constraint between reactants and products. The Malinowski Test<sup>6</sup> revealed three significant singular values, and thus  $n_r$  was defined to be three for this system. All seven of the aforementioned stoichiometric reactions passed the TFA test, and the reactions were combined to form 28 full rank networks of three reactions each.

**Table 8: Passing Stoichiometric Reactions in Case Study B**

$r_1:$	$A + F \rightleftharpoons C + 2D$
$r_2:$	$A + B \rightleftharpoons C + D$
$r_3:$	$C \rightleftharpoons D + E$
$r_4:$	$A + B \rightleftharpoons 2D + E$
$r_5:$	$B + D \rightleftharpoons F$
$r_6:$	$A + 2B \rightleftharpoons C + F$
$r_7:$	$C + D \rightleftharpoons E + F$

Three of the 28 networks passed the first dynamic constraint, requiring no sign change of each reaction rate during an experiment. One network, which was defined by the following reactions,



was filtered out during the second dynamic constraint. Thus, we were left with two networks, defined in Table 9, on which to perform kinetic analysis.

### 3.2.3. Kinetics and Statistical Analysis

Kinetic models with the lowest BIC for the two reaction in each network were identified and are shown in Table 9.

After simulating the reaction data from the estimated kinetic models and parameters for each network, we calculate the sum of square errors according to Equation 18. Because there were only two remaining networks, we performed an F test to determine if the error of the second network was significantly greater than that of the first network, as defined in Equations 21, 22, and 23, which resulted in a  $p$  value of  $8 \times 10^{-113}$ . We reject the null hypothesis that the two

network errors are equivalent and conclude that Network 2 is the most accurate model for this case.

Network	Stoichiometric Reactions	Kinetic Models	BIC Values for Kinetic Models	Kinetic Model SS
1	$r_1: A + B \rightleftharpoons C + D$ $r_2: C \rightarrow D + E$ $r_3: B + D \rightleftharpoons F$	$r_1 = k_1 C_A C_B - k_{-1} C_C C_D$ $r_2 = k_2 C_C$ $r_3 = k_3 C_B C_C - k_{-3} C_F$	$r_1: -6.6 \times 10^3$ $r_2: -6.0 \times 10^3$ $r_3: -7.5 \times 10^3$	$r_1: 0.03$ $r_2: 0.07$ $r_3: 0.01$
2	$r_1: F \rightleftharpoons B + D$ $r_2: A + 2B \rightarrow C + F$ $r_3: B + C \rightarrow E + F$	$r_1 = k_1 C_F - k_{-1} C_B C_D$ $r_2 = k_2 C_A C_B^2$ $r_3 = k_3 C_B C_C$	$r_1: -2.8 \times 10^3$ $r_2: -4.1 \times 10^3$ $r_3: -4.3 \times 10^3$	$r_1: 9.3$ $r_2: 1.3$ $r_3: 0.96$

**Table 9: Stoichiometric Reactions and Kinetic Models of Case B**

Remark: For each of the two networks that passed the dynamic constraints for Case B, the stoichiometric reactions are shown. The rate laws to the right of the reactions denote the kinetic models with the lowest BIC, and the respective BIC and sum of squares for the rate calculations are shown.

Next, we evaluate the performance of our algorithm by comparing the most accurate stoichiometric and kinetic model with the true one used in data collection. First, we observe that the stoichiometric reactions and kinetic models identified by the algorithm match the true ones (Table 9). We compare the kinetic parameters to the true ones in Table 10.

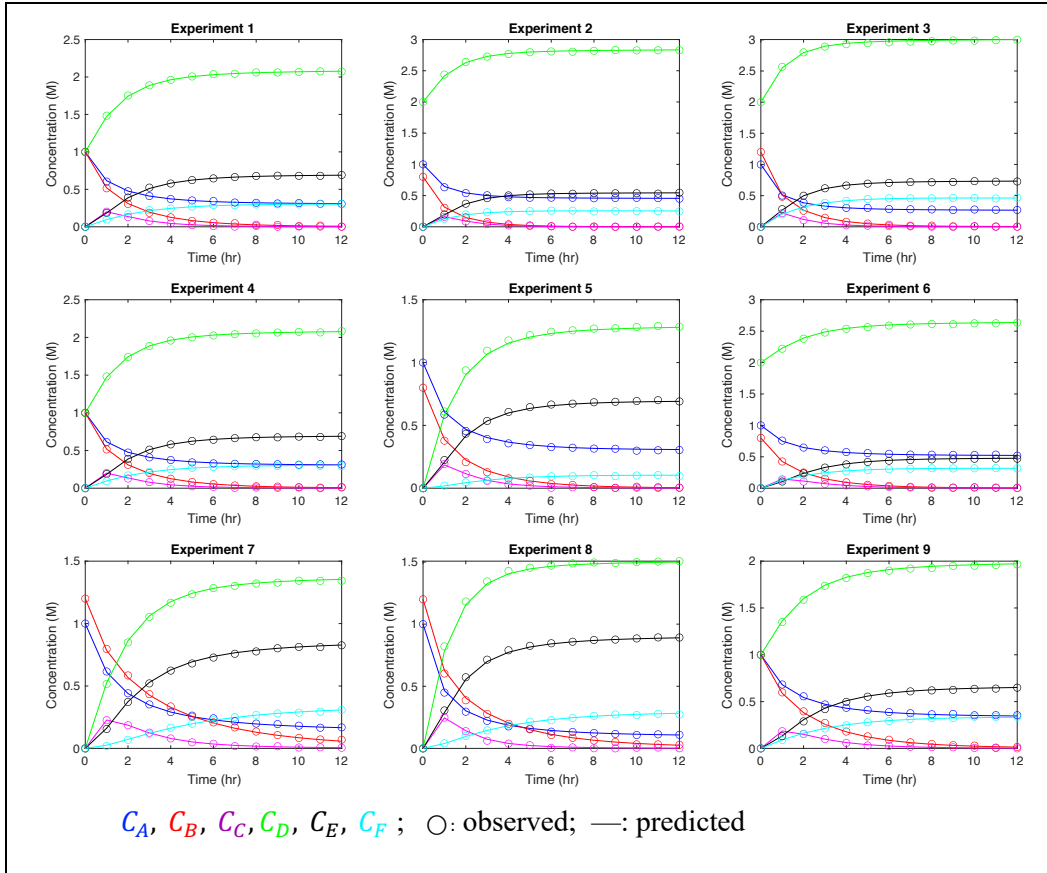
**Table 10: Comparison of Kinetic Parameters for Case B**

Parameters	$k_{1,0}$	$k_{-1,0}$	$k_{2,0}$	$k_{3,0}$	$k_{-3,0}$	$E_1$	$E_{-1}$	$E_2$	$E_3$	$E_{-3}$
True	0.50	0.050	1.0	0.10	0.005	3.97	3.78	1.20	0.994	1.99
Estimated	0.51	0.048	1.0	0.10	0.005	3.29	$4.5 \times 10^{-7}$	1.86	0.798	0.077

Here, we list the true kinetic parameters used to simulate the original data (Equation 31) and the estimated parameters using our algorithm. Units for kinetic constants and activation energies are as listed in Equation 31.

Finally, we visualize the performance of the discovered stoichio-kinetic model visually by comparing the model's predicted concentration profiles with the observed ones concentration profiles. The first nine of 17 experiments are

shown in Figure 3, and the remaining eight profiles are shown in Appendix Figure A1.



**Figure 3: Estimated and True Concentration Profiles for Case B**

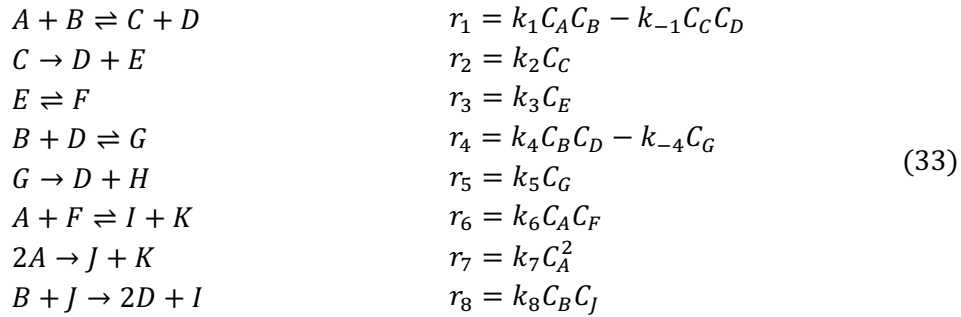
The true and estimated concentration profiles using the identified stoichio-kinetic model are shown in circles and lines, respectively. Concentrations of species  $A$  (blue),  $B$  (blue),  $C$  (magenta),  $D$  (green),  $E$  (black), and  $F$  (teal) are shown for the first nine experiments. The profiles for the remaining experiments are shown in Appendix Figure A1.

### 3.3. Case Study C: 11 Species Involved in Eight Reactions

The final case studied includes 11 species and eight reactions. The kinetic model used to generate the simulation was provided by Pfizer and is representative of the most challenging reaction networks encountered in pharmaceutical practice. Because of the added complexity of this case in comparison to the prior two, we are met with greater challenges when evaluating this system.

#### 3.3.1. Data Collection

The eight reactions and rate laws for this case are shown below:



in which

$$k_i = k_{i,0} \exp\left(-\frac{E_i}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right) \text{ with } T_0 = 50^\circ\text{C} \tag{34}$$

$$\begin{array}{ll}
 k_{1,0} = 0.5 \text{ L mol}^{-1} \text{ hr}^{-1} & E_1 = 3.974 \\
 k_{-1,0} = 0.05 \text{ L mol}^{-1} \text{ hr}^{-1} & E_{-1} = 3.775 \\
 k_{2,0} = 1 \text{ hr}^{-1} & E_2 = 1.987 \\
 k_{3,0} = 1 \text{ hr}^{-1} & E_3 = 2.9805 \\
 k_{4,0} = 0.1 \text{ L mol}^{-1} \text{ hr}^{-1} & E_4 = 0.9935 \\
 k_{-4,0} = 0.005 \text{ L mol}^{-1} \text{ hr}^{-1} & E_{-4} = 1.987 \\
 k_{5,0} = 0.01 \text{ hr}^{-1} & E_5 = 7.948 \\
 k_{6,0} = 0.02 \text{ L mol}^{-1} \text{ hr}^{-1} & E_6 = 3.974 \\
 k_{7,0} = 0.01 \text{ L mol}^{-1} \text{ hr}^{-1} & E_7 = 0.1987 \\
 k_{8,0} = 0.5 \text{ L mol}^{-1} \text{ hr}^{-1} & E_8 = 0.3974
 \end{array} \text{ (All in [kcal/mol])}$$

Molecular weights of the species are shown in Table 11. We added a constant measurement error to the simulated concentration measurements using a  $\sigma$  of 0.004 via Equation 27. The factors of the designed experiments are listed in the Appendix (Table A1), and a DRSM model was estimated with the data. The factors are temperature and initial concentrations of *B* and *D*. A center composite design was used with three replicates at the center point. Note that the eight reactions listed in Equation 33 are linearly dependent and have a rank of 7, implying that one is a linear combination of the others.

**Table 11: Molecular Weights of Species in Case C**

Species	A	B	C	D	E	F	G	H	I	J	K
Molecular Weight (g/mol)	187.3	170.2	339.5	18.0	321.5	321.5	188.2	170.2	448.7	314.5	60.1

### 3.3.2. Network Enumeration and Filtering

For this case, 32 stoichiometric reactions pass the mass balance constraint. The Malinowski test reveals six significant singular values, and TFA eliminates 15 reactions, leaving 17 candidate stoichiometric reactions that have an acceptable TFA projection score. We have previously stated that the number of significant singular values is equal to the number of linearly independent reactions occurring in a system. However, here, the rank of the resulting 17 reactions is only five. This implies that we can only identify reaction networks of five reactions, and we thus construct networks of five reactions each. In this case, we apply an additional constraint, requiring that the known overall reaction is a linear combination of the reactions within the network. We employ this additional constraint because, typically, the overall reaction is known. Over 3,000 full rank

networks of 5 reactions were enumerated that fulfill this constraint. Only 54 of the networks passed the first dynamic constraint, and 30 passed the second and third constraints.

### 3.3.3. Kinetics and Statistical Analysis

We performed kinetic analysis on the resulting 30 networks. The five best performing networks, defined as those with the smallest sum of square error, are shown in Table 12.

Network	$SS_r$	Stoichiometric Reactions	Kinetic Models
1	141	$r_1: A + H \rightleftharpoons C + D$ $r_2: A + B \rightarrow 2D + E$ $r_3: B + D \rightleftharpoons G$ $r_4: E + G \rightleftharpoons B + C$ $r_5: E + H \rightleftharpoons B + F$	$r_1 = k_1 C_A C_H - k_{-1} C_C C_D$ $r_2 = k_2 C_A C_B$ $r_3 = k_3 C_B C_D - k_{-3} C_G$ $r_4 = k_4 C_E C_G - k_{-4} C_B C_C$ $r_5 = k_5 C_E C_H - k_{-5} C_B C_F$
2	146	$r_1: A + B \rightleftharpoons 2D + E$ $r_2: C + E \rightarrow D + 2F$ $r_3: B + D \rightleftharpoons G$ $r_4: F + G \rightleftharpoons B + C$ $r_5: E + H \rightarrow B + F$	$r_1 = k_1 C_A C_B - k_{-1} C_D^2 C_E$ $r_2 = k_2 C_C C_E$ $r_3 = k_3 C_B C_D - k_{-3} C_G$ $r_4 = k_4 C_F C_G - k_{-4} C_B C_C$ $r_5 = k_5 C_E C_H$
3	146	$r_1: A + H \rightleftharpoons C + D$ $r_2: 2D + F \rightleftharpoons A + H$ $r_3: A + B \rightarrow 2D + E$ $r_4: B + D \rightleftharpoons G$ $r_5: E + G \rightleftharpoons B + C$	$r_1 = k_1 C_A C_H - k_{-1} C_C C_D$ $r_2 = k_2 C_D^2 C_F - k_{-2} C_A C_H$ $r_3 = k_3 C_A C_B$ $r_4 = k_4 C_B C_D - k_{-4} C_G$ $r_5 = k_5 C_E C_G - k_{-5} C_E C_G$
4	158	$r_1: A + B \rightleftharpoons 2D + E$ $r_2: C \rightarrow D + F$ $r_3: B + D \rightleftharpoons G$ $r_4: E + G \rightleftharpoons B + C$ $r_5: E + H \rightarrow B + F$	$r_1 = k_1 C_A C_B - k_{-1} C_D^2 C_E$ $r_2 = k_2 C_C$ $r_3 = k_3 C_B C_D - k_{-3} C_G$ $r_4 = k_4 C_E C_G - k_{-4} C_E C_G$ $r_5 = k_5 C_E C_H$
5	257	$r_1: C + E \rightarrow D + 2F$ $r_2: G \rightleftharpoons B + D$ $r_3: A + 2B \rightarrow C + G$ $r_4: B + C \rightarrow E + G$ $r_5: G + H \rightarrow 2B + D$	$r_1 = k_1 C_C C_E$ $r_2 = k_2 C_G - k_{-2} C_B C_D$ $r_3 = k_3 C_A C_B^2$ $r_4 = k_4 C_B C_C$ $r_5 = k_5 C_G C_H$

**Table 12: Best Performing 5-Reaction Networks for Case C**

Reactions and identified corresponding kinetic expressions are shown above for the five best performing networks, as well as the regression sum of squares for each network, as defined by Equation 18.

An F-test on each of the above five network sum of squares against Network 1, which has the lowest sum of squares, reveals that only Network 5 has a significantly larger error than Network 1. Networks 1-4 all have regression sums of squares that are indistinguishable from each other. Therefore, for this case, our algorithm fails to discover one single reaction network and instead identifies four.

At this stage, we have multiple potential routes to continue examining this system. We first observe that reaction involving species  $I, J$ , and  $K$  are not present in any of the four best performing networks. Considering both that we have multiple species that remain unmodeled and that we have multiple networks of indistinguishable performance, clearly some non-random data from the measured compositions is not modeled at this stage. We discuss these considerations and potential future explorations further in the Discussion section.

One option we explore is to increase the number of significant singular values used to calculate the target factor analysis scores. While the Malinowski test claims that there are six significant singular values, this test has not been proven to work for every system, in any case. We begin by setting the number of significant singular values to seven, which increases the number of reactions passing TFA to 28. These 28 stoichiometries have a rank of seven, so we form groups of seven reactions each. Over 65,000 full rank networks are formed, and over 1,500 pass the dynamic constraints. Due to the substantially increased computational cost for this many networks, we choose to not explore all of these networks at this point.

A second alternative is to consider smaller networks of less than five reactions using the original number of significant singular values. We now consider reaction networks of size 2, 3, or 4 reactions. While we expect that networks with more reactions can model the data with greater accuracy, we have no proof that this hypothesis is indeed true.

After applying the dynamic constraints to networks with two, three, and four reactions, we are left with 75 networks and proceed with kinetic evaluation. Interestingly, the smallest sum of square errors found in the networks is 10.5, which is much smaller than the minimum sums found for networks of five reactions. Comparing all other networks, including those with five reactions, to the best performing network with the previously defined F-test, we observe that the regression sum of squares of only two networks are indistinguishable. These networks are shown in Table 13.

Network	SS	Stoichiometric Reactions	Kinetic Models
1	10.5	$r_1: A + B \rightarrow 2D + E$ $r_2: E \rightarrow F$ $r_3: 2B + D \rightarrow G + H$	$r_1 = k_1 C_A C_B$ $r_2 = k_2 C_E$ $r_3 = k_3 C_B^2 C_D$
2	12.9	$r_1: A + B \rightarrow 2D + F$ $r_2: 2B + D \rightarrow G + H$	$r_1 = k_1 C_A C_B$ $r_2 = k_2 C_B^2 C_D$

**Table 13: Overall Best Performing Networks for Case C**

Stoichiometric reactions and kinetic expressions are shown for the two networks that performed best of a set of networks with 2 – 5 reactions. The regression sum of squares for each network is shown as well.

Surprisingly, the top networks with only two and three reactions outperform those with four and five reactions. We further examine the implications of these results in the Discussion section.

## Chapter 4: Discussion

In Case A, we explored a relatively simple simulated system with only two reactions and three species. We successfully identified the true stoichiometric reactions and corresponding kinetic models. Our optimized kinetic constants and activation energies (Table 5) are close to those used in the data simulation. However, while our determined parameters look promising for this case, their estimation is incomplete without defined confidence intervals. A next step in this work is to identify the variance of the parameter estimates.

After the methodology was applied to a simple case, we simulated a more complex scenario with six species and three linearly independent reactions in Case B. As in Case A, our algorithm successfully identified the correct stoichiometric reactions and the corresponding kinetics, and our estimated kinetic constants (Table 10) accurately represent those of the true network. However, the estimated activation energies for the true reverse reactions vary from the true ones significantly (Table 10). The calculation of the confidence intervals for estimated parameters would provide greater insight into this discrepancy. We acknowledge, however, that the activation energies of Cases A and B were generally of the same

order of magnitude. Furthermore, in Case B, the experiments ranged from 50°C to 90°C, while those in Case A ranged from 0°C to 100°C. We expect that the narrower temperature range in experiments resulted in the inaccuracy of activation energy estimations. Because the performed experiments had a narrower temperature range, we hypothesize that the estimation of the parameters was less sensitive to activation energy changes. Considering the close fit demonstrated in Figure 3, the observed concentration profiles were not sensitive to the reverse reaction activation energies. Perhaps performing the experiments for a longer batch time or beginning with higher concentrations of species will increase the sensitivity of the data to such activation energies.

Case C provided a much more realistic example with added complexity. The increase in the number of species and reactions introduced additional computational cost, which limited some of the calculations. Further, the eight reactions occurring in the system were linearly dependent. Because our current algorithm cannot test linearly dependent networks of reactions, any network might have been unable to model the data completely. Finally, some of the reactions within the network had quite small reaction rates compared to the others. If the magnitudes of concentration derivatives related to slower reaction rates are close to the magnitude of the measurement error, identification of those reactions would be especially difficult.

While we were able to identify networks that fulfilled dynamic constraints for Case C and fit kinetic models to the resulting reaction rates, we were

ultimately unable to identify a set of reactions that closely matched the true network.

We focus now on questioning our assumptions and modifying them to better fit reality. First, our method assumes that the true network is a set of linearly independent reactions. This assumption clearly does not apply to this network and limits the possibilities of network explored by the algorithm. If we were to remove one reaction from the set, leaving a set of seven linearly independent reactions, and subsequently evaluate the system, we may analyze how the algorithm identifies a linearly independent set of reactions of similar complexity. Additionally, if we were able to consider linearly dependent sets of reactions, we may discover stoichiometric models that model the system with greater accuracy.

Second, we calculate the uncertainty in the reaction rates according to Equation 13, requiring the assumption that the concentration derivative variances may be manipulated in this way to derive the reaction rate variances. We must consider that concentration error may propagate to the rates in a different manner. Specifically, we assume that the degrees of freedom of the reaction rates and the concentration time derivatives are the same, or close enough that any discrepancies will not affect calculations. We may consider altering our error propagation calculations to include these differences in degrees of freedom. Our first dynamic filter, which requires no sign change of reaction rates, eliminates over 90% of 5-reaction networks in Case C. If, in fact, the rate uncertainties are calculated incorrectly, and the calculations are adjusted, this dynamic filter may

allow a greater number of networks to pass. Perhaps some networks eliminated from this dynamic filter would outperform the current “best” networks.

In evaluating the best performing networks for Case C, we first observe that the first reaction of the Network 1 shown in Table 13 is the sum of the first two true reactions. The second reaction in Network 1 is the third true reaction, and the third reaction in Network 1 is a linear combination of the fourth and fifth true reactions. We similarly observe for Network 2 in Table 13 that the first and second reactions are both linear combinations of the first five true reactions. Our algorithm, therefore, did identify reaction networks which involve the primary reactions in the system. We might ask: why does the network of the first five true reactions fail our dynamic filter if a smaller network of linear combinations of those networks do pass? In this case, we observe that our approach favors networks of fewer reactions that fail to distinguish intermediate reactions. Future work must therefore focus on modifying the algorithm to possibly identify intermediate reactions which might have slower reaction kinetics.

A possible route to address this obstacle is to subtract the data explained by our best network from the original data and subsequently perform a second modeling cycle with the resultant differences. For example, in Case Study C, we may subtract data explained by models associated with Network 1 or 2 in Table 13 from the original data set. Next, we would evaluate the resulting data in a second cycle of our approach. Such analysis will focus more closely on reactions with smaller reaction rates or expose intermediates that were not relatively dominant during the previous modeling cycle. Then, we must develop an

appropriate methodology to fuse the two reaction stoichiometric identified in each cycle. While each network alone will have linearly independent reactions, the combination of the two networks may be a set of reactions that are not linearly independent. This approach would allow for an overall linearly dependent set of reactions, from a combination of two or more linearly independent sets of reactions.

## **Chapter 5: Conclusions**

We have developed a methodology that employs not only TFA, a static filter, but also dynamic filters and kinetic modeling, to aid in our search for the appropriate stoichiometric models. We have examined its effectiveness in identifying reactions and corresponding kinetic models in three different systems. The methodology involves data fitting with the DRSM methodology, target factor analysis, and kinetic modeling. We utilize mass balance and reaction engineering constraints to limit the numerous possibilities of reaction stoichiometries and networks, and statistical analysis of final kinetic models is used to identify the best performing reaction network.

The first two cases explored demonstrate that our novel approach is effective in identifying stoichiometric and kinetic models for simpler systems. However, we acknowledge that a complete model requires confidence intervals on kinetic parameters and intend to address this in future work. The third case studied revealed the need for a reevaluation of our approach and modification for more complex systems.

Future work will explore the possibility of subtracting modeled data from the original collected data, and subsequently identifying reactions from the resulting set of modified data. We anticipate that this approach may reveal reactions with slower kinetic rates and potentially intermediate reactions as well. We also acknowledge limitations that arise by assuming that reaction networks are linearly independent, and future work that explores linearly dependent networks would advance our methodology.

Despite the remaining obstacles with our stoichiometric and kinetic model identification method, we have made significant progress toward the issue of automatic and exhaustive model identification. Our work indicates that stoichiometric and kinetic models for at least some systems can be identified without the requirement that a list of candidate stoichiometries first be proposed. Nevertheless, future work will ensure that our approach is robust enough to handle complex reacting systems.

## References

1. Bonvin, D.; Rippin, D. W. T. Target factor analysis for the identification of stoichiometric models. *Chemical Engineering Science* **1990**, *45* (12), 3417-3426.
2. Dong, Y.; Georgakis, C.; Mustakis, J.; Hawkins, J. M.; Han, L.; Wang, K.; McMullen, J. P.; Grosser, S. T.; Stone, K. Stoichiometry identification of pharmaceutical reactions using the constrained dynamic response surface methodology. *AIChE Journal* **2019**, *65* (11).
3. Brendel, M.; Bonvin, D.; Marquardt, W. Incremental identification of kinetic models for homogeneous reaction systems. *Chemical Engineering Science* **2006**, *61* (16), 5404-5420.
4. Yeow, Y. L.; Wickramasinghe, S. R.; Han, B.; Leong, Y.-K. A new method of processing the time-concentration data of reaction kinetics. *Chemical Engineering Science* **2003**, *58* (16), 3601-3610.
5. Klebanov, N.; Georgakis, C. Dynamic Response Surface Models: A Data-Driven Approach for the Analysis of Time-Varying Process Outputs. *Industrial & Engineering Chemistry Research* **2016**, *55* (14), 4022-4034.
6. Malinowski, E. R. Statistical F-tests for abstract factor analysis and target testing. *Journal of Chemometrics* **1989**, *3* (1), 49-60.

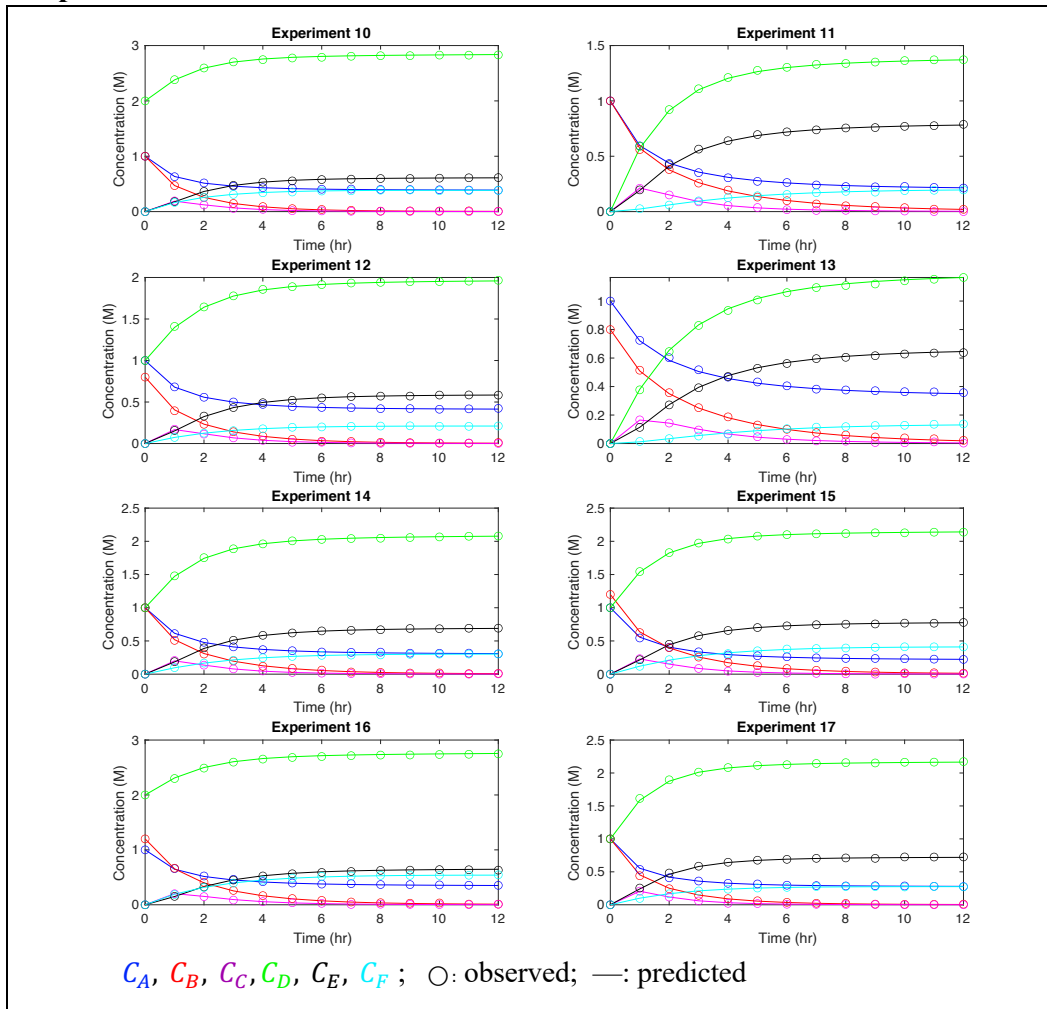
## Appendix

**Table A1: Factors of Experimental Design for Case C**

Experiment #	Temperature (°C)	$C_{B0}$ ( $\frac{\text{mol}}{\text{L}}$ )	$C_{D0}$ ( $\frac{\text{mol}}{\text{L}}$ )
1	70	1.0	1
2	90	0.8	2
3	90	1.2	2
4	70	1.0	1
5	90	0.8	0
6	50	0.8	2
7	50	1.2	0
8	90	1.2	0
9	50	1.0	1
10	70	1.0	2
11	70	1.0	0
12	70	0.8	1
13	50	0.8	0
14	70	1.0	1
15	70	1.2	1
16	50	1.2	2
17	90	1.0	1

The factors for each experiment in Case C including temperature, initial concentration of B, and initial concentration of D, are listed above.

**Figure A1. Estimated and True Concentration Profiles for Case B, Experiments 10–17.**



The true and estimated concentration profiles using the identified stoichiokinetic model are shown in circles and lines, respectively. Concentrations of species *A* (blue), *B* (blue), *C* (magenta), *D* (green), *E* (black), and *F* (teal) are shown for experiments 10–17.