Electronic Supplementary Information

Semi-supervised Machine Learning Approach for Reaction Stoichiometry and Kinetic Model Identification using Spectral Data from Flow Reactors

Manokaran V[a], Sreeja Shanmuga Doss[a,c,‡], Sridharakumar Narasimhan[a,c,d*], and Nirav Bhatt[b,c,d*]

[a]Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai-600036, India
[b]Department of Biotechnology, Indian Institute of Technology Madras, Chennai-600036, India
[c]Robert Bosch Centre for Data Science and Artificial Intelligence, Indian Institute of Technology Madras, Chennai-600036, India
[d]Prospective Centre of Excellence for Network Systems Learning, Control, and Evolution, Indian Institute of Technology Madras, Chennai-600036, India
[‡]Present address: Department of Chemical Engineering, Indian Institute of Technology Bombay, Maharashtra - 400076, India.
E-mail id: *sridharkrn@iitm.ac.in, *niravbhatt@iitm.ac.in

Contents

S1 Modelling of Micro-reactors

In this section, a mathematical model to describe reaction kinetics in micro-reactor is presented. Although flow behaviour in micro-reactor is laminar, these reactors are often modelled as plug flow reactor (PFR) as it has minimal axial dispersion [1, 2]. The material balance for a homogeneous isothermal reaction system in a PFR of volume V with R reactions and S species is given by

$$\frac{d\mathbf{f}}{dV} = \mathbf{N}^T\mathbf{r}(\mathbf{c}, \boldsymbol{\theta}), \quad \mathbf{f}(0) = \mathbf{f}_0 \tag{S1}$$

In Equation (S1), $\mathbf{f}$ and $\mathbf{c}$ correspond to S-dimensional vectors of molar flow rates and concentrations, respectively, $\mathbf{N}$ is the stoichiometric matrix with dimension $R \times S$, $\mathbf{r}$ is an r-dimensional vector of reaction rates, $\boldsymbol{\theta}$ is a P-dimensional vector of parameters associated with the kinetic model and $\mathbf{f}_0$ is an S-dimensional vector of initial molar flow rates of all species at the entrance of the reactor. Under the assumption of constant density and volumetric flow rate ν_0, Equation (S1) is rewritten in terms of residence time τ as ($\mathbf{f} = \mathbf{c}\ \nu_0$ and $\tau = \frac{V}{\nu_0}$)

$$\frac{d\mathbf{c}}{d\tau} = \mathbf{N}^T\mathbf{r}(\mathbf{c}, \boldsymbol{\theta}), \quad \mathbf{c}(0) = \mathbf{c}_0 \tag{S2}$$

In Equation (S2), τ corresponds to the residence time or the amount of time spent by the reactants inside the reactor. Also, $\mathbf{c}_0$ is an S-dimensional vector of concentrations of all species at $\tau = 0$. The R reactions in this model reaction set are assumed to be independent reactions. The definition of independent reactions and an illustration for describing the reaction systems in terms of independent reactions can be found in Section S2.1.

S2 Preliminaries

S2.1 Independent reactions

Definition 1 (Independent reactions) *A set of R reactions are said to be independent if rank($\boldsymbol{N}$) = R and there exists a finite time interval in which reaction rates $\boldsymbol{r}(t)$ are independent i.e., $\boldsymbol{\beta}^T \boldsymbol{r}(t) = 0 \Leftrightarrow \boldsymbol{\beta} = \boldsymbol{0}_R$.*

One should note that the number of independent reactions in a system depends on both $\mathbf{N}$ and $\mathbf{r}$. The concept of independent reactions is illustrated using the following example. Consider a reaction system with the following set of reactions

$$
\begin{array}{lllllllll}
\text{R1:} & \text{A} & + & \text{B} & \rightarrow & \text{C}, & & r_1 \\
\text{R2:} & \text{A} & + & \text{B} & \rightarrow & \text{D}, & & r_2 \\
\text{R3:} & \text{C} & + & \text{B} & \rightarrow & \text{E}, & & r_3 \\
\text{R4:} & & & \text{E} & \rightarrow & \text{C} & + & \text{B}, & r_4
\end{array}
$$

With $\mathbf{c}$=$[c_A, c_B, c_C, c_D, c_E]^T$, the stoichiometric matrix (denoted as $\mathbf{N}_d$) and the reaction rate vector (denoted as $\mathbf{r}_d$) can be written as follows:

$$
\mathbf{N}_d = \begin{bmatrix} -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 0 & 1 & 1 & 0 & -1 \end{bmatrix}, \quad \mathbf{r}_d = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix}
$$

Equation (S2) for this example reaction set can be written using $\mathbf{c}$, $\mathbf{N} = \mathbf{N}_d$, and $\mathbf{r} = \mathbf{r}_d$. However, note that the rank($\mathbf{N}_d$) = 3. This is due to the fact that the reactions $R3$ and $R4$ are linearly dependent. Hence, $\mathbf{N}_d$ needs to be re-written in terms of the independent stoichiometric matrix $\mathbf{N}_i$:

$$
\mathbf{N}_d = \mathbf{P}\mathbf{N}_i, \quad \mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{bmatrix}, \quad \mathbf{N}_i = \begin{bmatrix} -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 & 1 \end{bmatrix}
$$

As a result, the reaction rate is also re-written so that Equation (S2) is consistent. The reaction rate corresponding to $\mathbf{N}_i$ (denoted as $\mathbf{r}_n$) can be written as:

$$\mathbf{r}_n = \mathbf{P}^T \mathbf{r}_d = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 - r_4 \end{bmatrix}$$

If the reaction rate $r_1 = \beta r_2$, then, the reaction rate vector $\mathbf{r}_n$ can be written as

$$\mathbf{r}_n = \begin{bmatrix} \beta r_2 \\ r_2 \\ r_3 - r_4 \end{bmatrix} = \mathbf{Q}\mathbf{r}_i, \quad \mathbf{Q} = \begin{bmatrix} \beta & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \ \mathbf{r}_i = \begin{bmatrix} r_2 \\ r_3 - r_4 \end{bmatrix}$$

Since the reaction rate is defined in terms of $\mathbf{r}_i$, the stoichiometric matrix $\mathbf{N}_i$ is redefined to make Equation (S2) consistent as follows

$$\mathbf{N}_{ir} = \mathbf{Q}^T \mathbf{N}_i = \begin{bmatrix} -(\beta+1) & -(\beta+1) & \beta & 1 & 0 \\ 0 & -1 & 1 & 0 & 1 \end{bmatrix}$$

Due to the dependency of the reaction rates, the number of independent reactions reduces from three to two. The resulting independent reactions as per definition are as follows:

$$\begin{array}{llllllllll} \text{R1:} & (\beta{+}1)\ \text{A} & + & (\beta{+}1)\,\text{B} & \rightarrow & \beta\,\text{C} & + & \text{D}, & r_2 \\ \text{R2:} & \text{C} & + & \text{B} & \rightarrow & \text{E}, & & & r_3 - r_4 \end{array}$$

Equation (S2) can be written in terms of independent reactions with $\mathbf{N} = \mathbf{N}_{ir}$, and $\mathbf{r} = \mathbf{r}_i$. Note that redefining the stoichiometric matrix and reaction rates in terms of independent reactions does not change the material balance equations. The concept of independent reactions allows us to represent the balance equations in a parsimonious manner [3]. Further, it also helps in identifying reaction kinetics from data as described in Section 2.4.

S2.2 Extents of reaction

Another quantity of interest is the extents of reaction. The extents of reaction for the constant density isothermal PFR is defined as follows:

Definition 2 *The extents of reaction is defined as the change in concentration of species in a reaction after τ residence time in the reactor.*

The extent of the ith reaction ($x_{r,i}$) at τ residence time is defined as

$$x_{r,i}(\tau) = \int_0^\tau r_i(\mathbf{c}, \boldsymbol{\theta}) \mathbf{d}\tau \tag{S3}$$

The relationship between concentrations at τ and the extents of reaction is given by

$$\mathbf{c}(\tau) = \mathbf{N}^T \mathbf{x}(\tau) + \mathbf{c}_0 \tag{S4}$$

where $\mathbf{x}(\tau)$ is an R-dimensional vector of the extents of reaction at τ residence time and $\mathbf{c}_0$ is an S-dimensional vector of initial concentrations. For L observations, the measured concentration data at different τ are concatenated as follows, where $\mathbf{C}$ is a $(L \times S)$ matrix :

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}^T(\tau_0) \\ \mathbf{c}^T(\tau_1) \\ \mathbf{c}^T(\tau_2) \\ \vdots \\ \mathbf{c}^T(\tau_{L-1}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^T(\tau_0)\mathbf{N} + \mathbf{c}_0^T \\ \mathbf{x}^T(\tau_1)\mathbf{N} + \mathbf{c}_0^T \\ \mathbf{x}^T(\tau_2)\mathbf{N} + \mathbf{c}_0^T \\ \vdots \\ \mathbf{x}^T(\tau_{L-1})\mathbf{N} + \mathbf{c}_0^T \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{x}^T(\tau_0) \\ \mathbf{x}^T(\tau_1) \\ \mathbf{x}^T(\tau_2) \\ \vdots \\ \mathbf{x}^T(\tau_{L-1}) \end{bmatrix}}_{\mathbf{X}} \mathbf{N} + \begin{bmatrix} \mathbf{c}_0^T \\ \mathbf{c}_0^T \\ \mathbf{c}_0^T \\ \vdots \\ \mathbf{c}_0^T \end{bmatrix} \tag{S5}$$

$$\mathbf{C} = \mathbf{X}\mathbf{N} + \mathbf{1}_L \mathbf{c}_0^T \tag{S6}$$

In the Equation (S6), $\mathbf{X}$ is the extents of reaction matrix[1] of size $L \times R$ and $\mathbf{1}_L$ is the L-dimensional vector of ones.

S3 Factorisation of Spectral Data

This section presents a generalised factorisation of spectral data from reaction systems in continuous plug-flow reactors. The factorisation of spectral data is derived under the following assumptions. (A1) Measured spectra have a linear relationship with concentration, i.e., Beer-Lambert's law is valid, (A2) experiments are conducted under isothermal conditions, (A3) the measured spectral data do not contain unwanted contributions such as drift or presence of absorbing interferents, (A4) reaction system is considered to be homogeneous with constant density, and (A5) all species involved in the system are absorbing with independent pure component spectra.

[1]Extents of reaction for constant density homogeneous reaction systems is represented in concentration units (mol/m^3 or mol/L).

$$\mathbf{a}^T(t) = \mathbf{c}^T(t)\mathbf{E} \tag{S7}$$

where $\mathbf{c}(t)$ denotes the concentration of the t^{th} sample, and $\mathbf{E}$ is the $S \times M$-dimensional pure component spectra matrix. Here, each row of the pure component spectra matrix denotes the pure signals of individual species involved in the system. For L observations, Equation (S7) can be expressed as:

$$\mathbf{A} = \mathbf{CE} \tag{S8}$$

where $\mathbf{A}$ is the absorbance matrix of dimension $L \times M$, and $\mathbf{C}$ is the concentration matrix of dimension $L \times S$. The concentrations of all the S species at the i^{th} residence time τ_i, $i = 0, 1, \ldots, L-1$ is denoted as $\mathbf{c}(\tau_i)$. Using Equation (S4), the concentration can be written in terms of the extents of reaction $\mathbf{x}$ as:

$$\mathbf{c}(\tau_i) = \mathbf{N}^T\mathbf{x}(\tau_i) + \mathbf{c}_0 \tag{S9}$$

For L observations, the relationship for concentrations obtained from Equation (S6) is substituted in Equation (S8) which leads to the following equation

$$\mathbf{A} = (\mathbf{XN} + \mathbf{1}_L\mathbf{c}_0^T)\mathbf{E} = \mathbf{XNE} + \mathbf{1}_L\mathbf{a}_0^T \tag{S10}$$

where $\mathbf{a}_0$ is the initial absorbance vector of the reactants at residence time $\tau = 0$ with dimension $M \times 1$. In Equation (S10), the contributions to spectral data from reaction systems are spanned by two subspaces: (1) R-dimensional reaction space spanning the rows of $\mathbf{NE}$, and (2) one-dimensional space spanning the contribution to initial absorbance. The *reaction variant form* (RV-form) of spectral data ($\mathbf{H}$)[3] can be derived by subtracting the contributions from initial absorbance as shown below:

$$\mathbf{H} = \mathbf{A} - \mathbf{1}_L\mathbf{a}_0^T = \mathbf{XNE} \tag{S11}$$

It should be noted from Equation (S11) that the RV-form of spectral data contains contribution only from the reactions. This representation will be useful to isolate contributions from various reactions into the extents of reaction.

S4 Identification of stoichiometric matrix from SSML approach

S4.1 Lipase-catalysed hydrolysis reaction

The number of independent reactions for the lipase-catalysed hydrolysis reaction from the rank analysis of spectral data is found to be equal to *one* (refer to Section 4.1). The maximum number of independent reactions possible for this system can be computed from Equation (2) using the information of the species involved and their corresponding molecular formula from Table S1.

Table S1: Molecular formulas and weights for the species involved in the lipase-catalysed hydrolysis reaction system.

Species	**Molecular formula**	**Molecular weight**
pNPA	$C_8H_7NO_4$	181.15
Water	H_2O	18.01
pNP	$C_6H_5NO_3$	139.11
Acetic acid (AA)	$C_2H_4O_2$	60.05

The atomic matrix for this reaction system involving elements $e = \{\text{C, H, N, O}\}$ and species $S = \{\text{pNPA, } H_2O\text{, pNP, AA}\}$ is given below.

$$\mathbf{A}_m = \begin{bmatrix} 8 & 0 & 6 & 2 \\ 7 & 2 & 5 & 4 \\ 1 & 0 & 1 & 0 \\ 4 & 1 & 3 & 2 \end{bmatrix} \tag{S12}$$

From the rank analysis of $\mathbf{A}_m$, it is found that $R_{max} \leq 1$ for this particular system. The candidate stoichiometric matrices for this reaction are generated by solving the optimisation problem in Equation (5) subject to the following constraints.

- $\text{rank}(\mathbf{N}) = 1$
- $\|\mathbf{N}^T(:, i)\|_0 \leq 4$, i.e., maximum number of species involved in a reaction should be ≤ 4
- Stoichiometric coefficient of all species should belong to the set $\{-1, 0, 1\}$

The only possible stoichiometric matrix for this system is $\mathbf{N} = [-1, -1, 1, 1]$. However, as H_2O is used in excess, the change in concentration of H_2O is negligible compared to other reactants. Hence a reduced stoichiometric matrix $\mathbf{N} = [-1, 1, 1]$ involving species {pNPA, pNP, AA} will be used throughout this work for this reaction system.

S4.2 Wittig reaction

The number of independent reactions for the Wittig reaction system obtained from the rank analysis of spectral data is found to be *three* (refer to Section 4.2). As before, the maximum number of independent reactions (R_{max}) in a reaction system can be found from the analysis of atomic matrix ($\mathbf{A}_m$) formulated using the information from Table S2. It should be noted that species which are

hypothesised to be involved in the Wittig reaction are based on the literature studies for similar starting materials [4].

Table S2: Species which are assumed to be involved in the Wittig reaction

Species	**Notation**	**Molecular formula**	**Molecular weight**
(4-Nitrobenzyl)triphenylphosphonium bromide	A	$C_{25}H_{21}O_2NBrP$	478.32
Benzaldehyde	B	C_7H_6O	106.12
KOH	C	KOH	56.10
(4-nitrobenzylidene)triphenylphosphorane (Ylide)	D	$C_{25}H_{20}O_2NP$	397.41
KBr	E	KBr	119
H_2O	F	H_2O	18.01
Triphenylphosphine oxide	G	$C_{18}H_{15}OP$	278.29
trans-4-nitrostilbene	H	$C_{14}H_{11}O_2N$	225.24
cis-4-nitrostilbene	I	$C_{14}H_{11}O_2N$	225.24
4-nitrotoluene	J	$C_7H_7O_2N$	137.14
HBr	K	HBr	80.91

The atomic matrix for this reaction system involving elements $e = \{C, H, O, N, Br, P, K\}$ is given below.

$$\mathbf{A}_m = \begin{bmatrix} 25 & 7 & 0 & 25 & 0 & 0 & 18 & 14 & 14 & 7 & 0 \\ 21 & 6 & 1 & 20 & 0 & 2 & 15 & 11 & 11 & 7 & 1 \\ 2 & 1 & 1 & 2 & 0 & 1 & 1 & 2 & 2 & 2 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{S13}$$

The rank of $\mathbf{A}_m$ for the Wittig reaction is equal to 6, denoting the maximum possible number of independent reactions $R_{max} \leq 5$. However, from the literature studies, we find that only four reactions are possible, which are the formation of ylide reaction (R1), ylide reaction with benzaldehyde to give geometric isomers (from the reactions R2 and R3), and hydrolysis of ylide reaction (R4). The following constraints are used to generate different stoichiometric matrices for the Wittig reaction system based on discussions from Section 2.2.

- $\text{rank}(\mathbf{N}) = 4$
- $\| \mathbf{N}^T(:, i) \|_0 \leq 5$
- Stoichiometric coefficient of all species should belong to the set $\{-1, 0, 1\}$

A total of 70 stoichiometric candidates are generated by solving the optimisation problem in Equation (5). Additional constraints, as shown below, are used to reduce the number of candidate stoichiometric matrices tested from the proposed SSML approach.

- All reactions in the system are bimolecular
- Each species listed in Table S2 participate at least in one reaction, i.e., there is no non-zero column in the generated stoichiometric matrix
- Reactions R2 and R3 should have the same species as the reactant, as it leads to the formation of geometric isomers H and I
- For ease of comparison, all the stoichiometric matrices generated are arranged in the order of {R1, R2, R3, and R4}

From these conditions, only five stoichiometric candidates are possible, which are shown below in Table S3. These stoichiometric matrices are used to compute the extents of reaction as discussed in Section 2.3.

Table S3: Candidate stoichiometric matrices generated for representing the Wittig reaction system

$\mathbf{N}_1$	$\mathbf{N}_2$
$\begin{bmatrix} -1 & 0 & -1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} -1 & 0 & -1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$
$\mathbf{N}_3$	$\mathbf{N}_4$
$\begin{bmatrix} -1 & 0 & -1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$	$\begin{bmatrix} -1 & 0 & -1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$
$\mathbf{N}_5$	
$\begin{bmatrix} -1 & 0 & -1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$	

It is found that stoichiometric candidates $\mathbf{N}_1$ and $\mathbf{N}_2$ results in negative values for the extents of reaction (computed from rotation matrix using offline measurements), which is not physically meaningful. Hence, these two candidates are removed from the candidate set. The candidate with minimum error for computing the extents of reaction from Equation (14) is used as the best stoichiometric matrix for representing the reactions in the Wittig reaction system, which is $\mathbf{N}_3$ as shown in Table S4.

Table S4: Stoichiometric candidates, nature of the solution and the objective function values for the computation of rotation matrix from the SSML approach

Stoichiometric Candidate	**Nature of Solution**	**Objection function value**
$\mathbf{N}_1$	Physically infeasible	$--$
$\mathbf{N}_2$	Physically infeasible	$--$
$\mathbf{N}_3$	Feasible	0.3493
$\mathbf{N}_4$	Feasible	0.3637
$\mathbf{N}_5$	Feasible	0.3921

S5 Analysis of independent reactions

The number of independent reactions in a reaction system can be found from rank analysis of (i) Material balance equation and (ii) Singular value decomposition (SVD) of spectral data. The following sections describe the rank analysis from the material balance equation for the reaction systems considered in Sections 3.2 and 3.3. The analysis based on SVD is discussed in Section 2.1.

S5.1 Lipase-catalysed hydrolysis reaction

For the chosen enzymatic hydrolysis reaction, the stoichiometric matrix $\mathbf{N}$ is equal to $[-1\,1\,1]$. From the definition of independent reactions, we find that $\text{rank}(\mathbf{N}) = 1$ and $\mathbf{r}(\text{t})$ is independent. Hence, the number of independent reactions is equal to 1. This is also verified by the rank analysis of spectral data (refer Section 4.1) and atomic matrix (Section S4.1).

S5.2 Wittig reaction system

For the Wittig reaction system, the best stoichiometric matrix $\mathbf{N}$ identified from the SSML approach is given by

$$\mathbf{N} = \begin{bmatrix} -1 & 0 & -1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \tag{S14}$$

Even though the rank of $\mathbf{N}$ is equal to 4, it is found that the number of independent reactions is equal to 3 based on the analysis from Section 4.2. This is true when the reactions R2 and R3 are assumed to be dependent as it leads to the formation of *trans* and *cis* isomers (species H and I) based on Figure 12. The material balance equations for the Wittig reaction system are written as

$$\frac{d\mathbf{c}}{d\tau} = \begin{bmatrix} -1 & -1 & -1 & -1 \\ 0 & -1 & -1 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} \tag{S15}$$

where $\mathbf{c} = [c_A, c_B, c_C, c_D, c_E, c_F, c_G, c_H, c_I, c_J, c_K]^T$. In Equation (S15), the reaction rates r_2 and r_3 are dependent as r_3 can be written as βr_2 where $\beta = k_3/k_2$. Thus, Equation (S15) is rewritten in terms of independent reactions as

$$\frac{d\mathbf{c}}{d\tau} = \begin{bmatrix} -1 & -1 & -1 & -1 \\ 0 & -1 & -1 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_4 \end{bmatrix} \tag{S16}$$

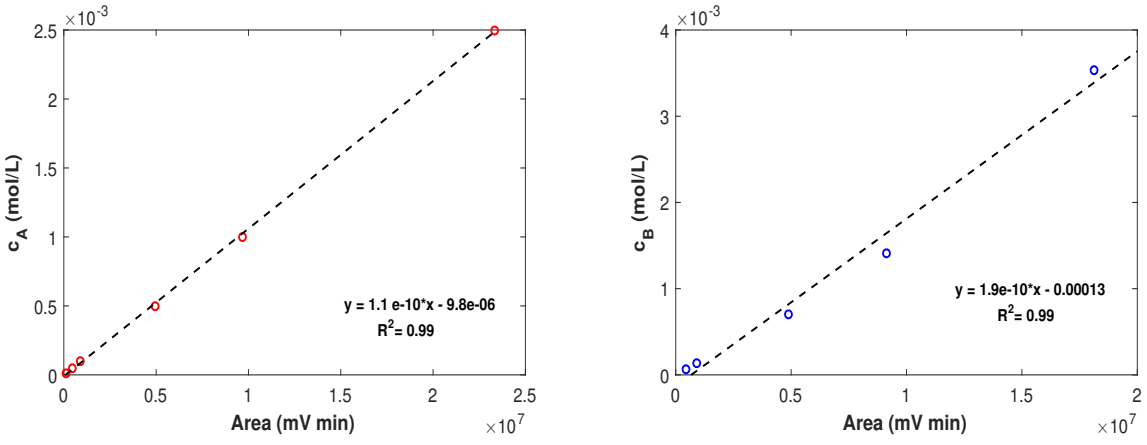
$$\frac{d\mathbf{c}}{d\tau} = \begin{bmatrix} -1 & -(\beta+1) & -1 \\ 0 & -(\beta+1) & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & -1 \\ 0 & \beta+1 & 1 \\ 0 & 1 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \\ 0 & \beta+1 & 1 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_4 \end{bmatrix} \tag{S17}$$

From Equation (S17), rank($\mathbf{N}$) is equal to 3 and hence we proved that the rank of a reaction system is dependent on both $\mathbf{N}$ and $\mathbf{r}$.

S6 HPLC analysis - Wittig reaction

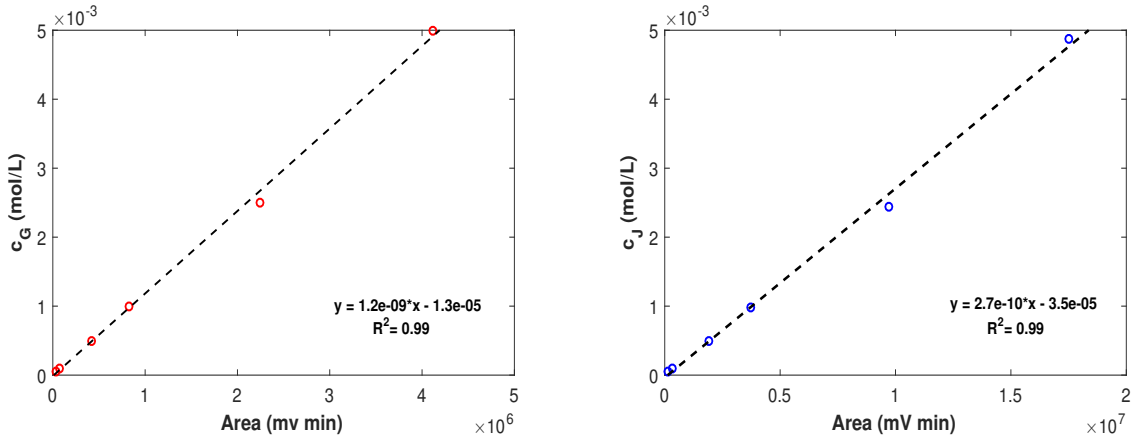
In order to illustrate the SSML approach for the Wittig reaction, offline concentration measurements are obtained from the analysis of the reaction samples using HPLC. It is assumed that only species A, B, G, and J are available for calibration. Pure standards of these species are used to prepare solutions of different concentrations to create the calibration set. Samples from the calibration set are analysed using an HPLC system (JASCO) integrated with the multi-wavelength detector. The column used is a reverse-phase C18 column (SunQsil 4.6 mm I.D, 250 mm length, 5 μ) with the mobile phase being methanol/water in a 3:1 ratio. The eluent is pumped at an isocratic flow rate of

0.5 ml/min, and compounds are analysed at the wavelength of 254 nm. Samples are introduced into the column using a stainless steel sample injector with 20 μl sample loop (Rheodyne™7725i, Thermo Scientific™). It is found from HPLC analysis that compounds A, B, G, and J elute at 25.61, 8.67, 12.51, 13.53 min. Univariate calibration between the concentration of calibration sample and area under the curve for each measured species is used to predict the concentration of these species in the reaction mixture. The calibration plots for all the species available for calibration from HPLC analysis are shown in Figures S1 and S2.



(a) (b)

Figure S1: HPLC calibration plots for (a) species A; (b) species B.



(a) (b)

Figure S2: HPLC calibration plots for (a) species G; (b) species J.

Samples from the reaction mixture are prepared for HPLC analysis by initially quenching the reaction mixture with dilute HCl (0.05 M) followed by dilution in absolute ethanol. The prepared samples from the reaction mixture are analysed under the same conditions as used in developing the calibration model. The measured HPLC chromatograms for reaction samples collected at different τ is shown in Figure 11a. The offline concentration of all measured species (A, B, G, and J) is predicted from the calibration model and is shown in Figure 11b. It is also found that *trans* and *cis* isomers formed from the reactions R2 and R3 elute at 42.1 and 43.6 min. The ratio of the area of *trans* and *cis* isomers (species H and I) from HPLC analysis for all the offline samples are used for fixing the β value in Equation (S17).

S7 Model discrimination criteria

S7.1 Akaike Information Criteria (AIC)

AIC criterion is one of the commonly used model discrimination criteria, where the goodness of fit and model complexity are addressed together to choose a simpler model from a set of model candidates [5]. For an independent and identically distributed Gaussian data ($\mu = 0$, σ^2), the residual sum of squares (RSS) from the maximum-likelihood estimate of variance (σ^2) will be used to calculate the AIC criteria as follows

$$AIC = 2k + n\ ln(RSS) \tag{S18}$$

In Equation (S18), k is the number of parameters in a particular model, and n is the number of data points. Using Equation (S18) AIC value is computed for each model candidate, and the model with minimum AIC value is selected as the best model. It should be noted that if all model candidates have an equal number of parameters, the model chosen from AIC criteria will be identical to the model selected from the maximum-likelihood method. In case of smaller datasets $\left(\frac{n}{k} < 40\right)$, modified AIC criterion (AIC_c) will be used as the discrimination criteria which is given by

$$AIC_c = 2k + n\ ln(RSS) + \frac{2k(k+1)}{n-k-1} \tag{S19}$$

S8 Calibration based approaches

S8.1 Classical calibration

In classical calibration, the errors are assumed to be associated with absorbance measurements ($\mathbf{A}_{cal}$) and a calibration model is developed using the calibration set ($\mathbf{C}_{cal}$, $\mathbf{A}_{cal}$) as follows

Calibration model: $$\mathbf{A}_{cal} = \mathbf{C}_{cal}\ \mathbf{E} + \epsilon$$

The unknown pure component spectra ($\mathbf{E}$) is estimated by minimising the square of Frobenius norm of error ϵ ($\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$)[2] and the solution is given by

$$\hat{\mathbf{E}} = \mathbf{C}_{cal}^{\dagger}\ \mathbf{A}_{cal}$$

The superscript † denotes the Moore-Penrose pseudo inverse of the corresponding matrix. Using the predicted pure component spectra, the unknown concentrations $\mathbf{c}_{unk}$ in the unknown spectral data $\mathbf{a}_{unk}$ is predicted as

Prediction: $$\hat{\mathbf{c}}_{unk} = \hat{\mathbf{E}}\ \mathbf{a}_{unk}$$

S8.1.1 Lipase-catalysed hydrolysis reaction

For the reaction system considered, pure standards of pNPA and pNP are used to prepare solutions of different concentrations that comprise the calibration set. Spectra is recorded for the each solution in the calibration set ($\mathbf{C}_{cal}$, $\mathbf{A}_{cal}$) and classical calibration is used for estimating pure component spectra of pNPA and pNP. The estimated pure component spectra is shown in Figure S3 and molar absorptivity values of pNPA and pNP at their λ_{max} is provided in Table S5.

[2] $\boldsymbol{\epsilon}$ is assumed to follow Normal or Gaussian distribution

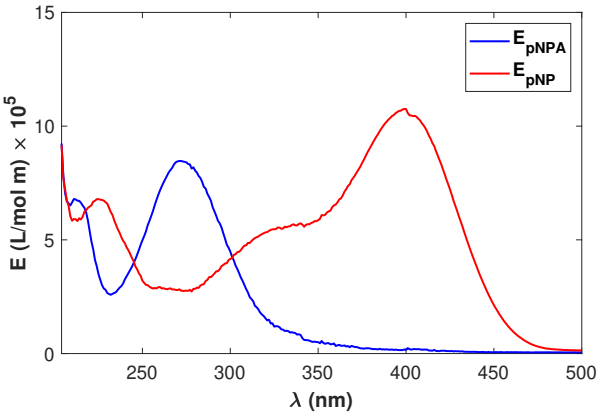


Figure S3: Pure component spectra of pNPA and pNP

Table S5: Molar absorptivity (E) of absorbing species at λ_{max} for lipase-catalysed hydrolysis reaction system measured using 0.1 M phosphate buffer as the reference solution[3]

Species	λ_{max} (nm)	E at λ_{max} (dm^3/mol cm)
pNPA	211, 271	6797, 8460
pNP	225, 399	6790, 10760

S8.1.2 Wittig reaction

In case of Wittig reaction, the absorbing species are A, B, D, G, H, I, and J with D being an unstable intermediate. However, only pure standards of A, B, G, and J are available for calibration. From pure standards of A (TCI 98%), B (Merck 99%), G (Avra 98%), and J (Spectrochem 98%) solutions at different concentrations are prepared to generate a calibration set. From the calibration set, pure component spectra are computed using classical calibration approach and is shown in Figure S4. The species H and I have characteristic peaks in the calibration range (200 to 400 nm) [6], but do not show up in Figure 9 due to very low concentration of these species after dilution. The formation of these species are verified from HPLC analysis of reaction mixture as shown in Figure 11a.

[3]Molar absorptivity computed from calibration set is represented in ($\times 10^5$ L/mol m) in the Figure S3. Here it is represented in (dm^3/mol cm). Also molar absorptivity is denoted as E instead of ϵ

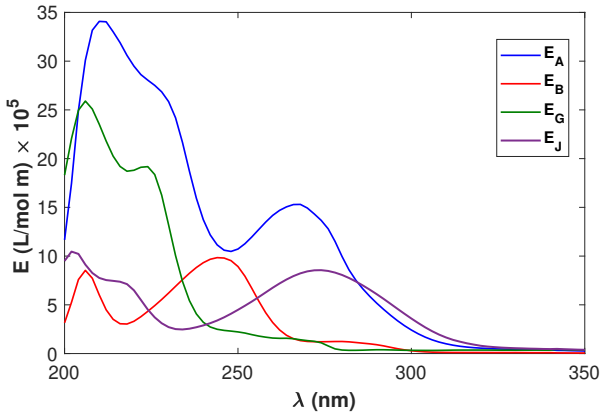


Figure S4: Pure component spectra of species A, B, G, and J for Wittig reaction

Table S6: Molar absorptivity (E) of absorbing species at λ_{max} for Wittig reaction system measured using ethanol as the reference solution[4]

Species	λ_{max} (nm)	E at λ_{max} (dm^3/mol cm)
A	210, 228, 268	34080, 26850, 15300
B	206, 244, 280	8535, 9841, 1238
G	206, 224	25900, 19180
J	218, 274	7126, 8556

S8.2 Inverse calibration

In inverse calibration, the errors are assumed to be in concentration ($\mathbf{C}_{cal}$) and using the calibration set ($\mathbf{C}_{cal}$, $\mathbf{A}_{cal}$) an inverse calibration model is developed.

Calibration model: $$\mathbf{C}_{cal} = \mathbf{A}_{cal}\ \boldsymbol{\beta} + \gamma$$

The unknown $\boldsymbol{\beta}$ is estimated by minimising the square of Frobenius norm of error γ ($\boldsymbol{\gamma} \sim \mathcal{N}(0, \sigma^2)$)[5] and the solution is given by

$$\hat{\boldsymbol{\beta}} = \mathbf{A}_{cal}^{\dagger}\ \mathbf{C}_{cal}$$

[4]Molar absorptivity computed from calibration set is represented in ($\times 10^5$ L/mol m) in the Figure S4. Here it is represented in (dm^3/mol cm). Also molar absorptivity is denoted as E instead of ϵ

[5]$\boldsymbol{\gamma}$ is assumed to follow Normal or Gaussian distribution

$\hat{\boldsymbol{\beta}}$ is used to predict the concentrations in unknown (reaction) spectral data ($\mathbf{a}_{unk}$) as follows.

Prediction: $$\hat{\mathbf{c}}_{unk}^T = \mathbf{a}_{unk}^T \, \hat{\boldsymbol{\beta}}$$

S9 Linear regression for estimating parameters $\boldsymbol{\theta}$ of Michaelis-Menten (MM) model

Using the inverse calibration model, the concentration of pNPA and pNP in the reaction set is predicted. Integrated form of MM model as in Equation (S20) is used for linear regression. The predicted concentration of substrate (pNPA) is used for estimation of parameters (V_{max}, K_m) using linear regression between τ and [log(c_{pNPA}) c_{pNPA}]. Here c_{pNPA_0} is the initial concentration of pNPA at $\tau = 0$ inside the reactor.

$$\frac{K_m}{V_{max}} \, log\Big(\frac{c_{pNPA}}{c_{pNPA_0}}\Big) + \frac{1}{V_{max}} \, (c_{pNPA} - c_{pNPA_0}) = -\tau \tag{S20}$$

S10 Rank of a matrix–Important properties

P1: Let $\mathbf{Q}$ be a matrix of dimension $m \times n$. The $rank(\mathbf{Q}) \leq min(m, n)$.
P2: Let $\mathbf{Q}$ be a matrix of dimension $m \times n$. If $\mathbf{Q} = \mathbf{BF}$ where $\mathbf{B}$ is an $m \times r$ matrix and $\mathbf{F}$ is an $r \times n$ matrix, then $rank(\mathbf{Q}) \leq min(rank(\mathbf{B}), rank(\mathbf{F}))$.
P3: Let $\mathbf{Q}$ be a matrix of dimension $m \times n$. If $\mathbf{Q}$ is decomposed into two matrices $\mathbf{B}_{m\times r}$ and $\mathbf{F}_{r\times n}$, and another matrix $\mathbf{G}$ is constructed such that $\mathbf{G} = \mathbf{BF} + \mathbf{1}_m \, \mathbf{g}_{ref}^T$, the rank of $\mathbf{G} = min(rank(\mathbf{Q}) + 1, m, n)$, when $\mathbf{1}_m$ does not lie in $\mathcal{C}(\mathbf{B})$ (column space of $\mathbf{B}$) and $\mathbf{g}_{ref}^T$ does not lie in $\mathcal{C}(\mathbf{F}^T)$ (row space of $\mathbf{F}$).

Proof:

$$\begin{aligned}
\mathbf{G} &= \mathbf{BF} + \mathbf{1}_m \mathbf{g}_{ref}^T \\
&= \begin{bmatrix} \mathbf{B} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{F} \\ \mathbf{g}_{ref}^T \end{bmatrix} \\
&= \mathbf{B}^*_{m\times r+1} \mathbf{F}^*_{r+1\times n}
\end{aligned}$$

If $\mathbf{1}_m$ does not lie in $\mathcal{C}(\mathbf{B})$ and $\mathbf{g}_{ref}^T$ does not lie in $\mathcal{C}(\mathbf{F}^T)$, then both $\mathbf{B}^*$ and $\mathbf{F}^*$ are of full rank. Thus, rank of $\mathbf{G} = \min(rank(\mathbf{Q}) + 1, m, n)$ based on the P2.
P4: Let $\mathbf{B}$ and $\mathbf{F}$ be $m \times r$ and $r \times n$ dimensional matrices. If $\mathbf{BF} = 0$, then $rank(\mathbf{B}) + rank(\mathbf{F}) \leq r$ according to Sylvester nullity theorem

Proof for Property 1 (Rank of A and H): Here, we derive the rank of $\mathbf{H}$ using the factorisation in Equation (S11). Using Equation (S11),

$$rank(\mathbf{H}) = rank(\mathbf{XNE}) \tag{S21}$$

Using the P1 and P2, Equation (S21) can be written as:

$$rank(\mathbf{H}) = min(rank(\mathbf{X}), rank(\mathbf{N}), rank(\mathbf{E}))$$

Since $L, M >> R$ and $S > R$ (independent reactions), the rank of $\mathbf{H} = R$. Similarly, we can derive the rank of $\mathbf{A}$ using Equation (S10) and the P3 as follows

$$rank(\mathbf{A}) = rank(\mathbf{H}) + \mathbf{1}_L \mathbf{a}_0^T = min(rank(\mathbf{H}) + 1, L) = R + 1$$

.

Proof for Property 2 (Maximum number of independent reactions): For a reaction system involving R reactions and S species, the stoichiometric matrix $\mathbf{N}$ should lie in the null space of the atomic matrix $\mathbf{A}_m$ so that all the elements involved in the reaction are conserved. In mathematical notation, this can be written as

$$\mathbf{A}_m \mathbf{N}^T = \mathbf{0} \tag{S22}$$

Using P4 in Equation (S22), we obtain the relationship

$$rank(\mathbf{A}_m) + rank(\mathbf{N}) \leq S \tag{S23}$$

The maximum number of linearly independent vectors in $\mathbf{N}$ is equal to the rank of the matrix. Then, the maximum number of linearly independent reactions R_{max} is given by

$$R_{max} \leq S - rank(\mathbf{A}_m) \tag{S24}$$

Proof for Property 3 (Minimal offline concentration measurements): Here, we will derive the rank conditions for minimal offline concentration measurements required for unique resolution of extents of reaction as discussed in Section 2.3. For the optimisation problem solved in Equation (14) (for estimating rotation matrix) to have a unique solution, the rank of $\mathbf{X}_{off}$ should be equal to R. The rank of $\mathbf{X}_{off}$ according to P2 from Equation (13) can be written as follows:

$$rank(\mathbf{X}_{off}) = min(rank(\mathbf{D}_{off}), rank(\mathbf{N}_{off})) \tag{S25}$$

Using P1, the minimum number of species to be measured is equal to R (independent reactions), hence the dimension of $\mathbf{N}_{off}$ should be at least $R \times R$. As at least R species concentrations have to be measured, the dimension of $\mathbf{D}_{off}$ should be $L_{off} \times R$. Using P1 for $\mathbf{D}_{off}$, we find that $L_{off} \geq R$ so that $rank(\mathbf{X}_{off})$ is equal to R. Thus, for a unique resolution of extents of reaction using offline concentration measurements, it is required to have at least R offline concentration measurements of R species provided the $rank(\mathbf{D}_{off}) = R$.

References

(1) J. P. McMullen, M. T. Stone, S. L. Buchwald and K. F. Jensen, *Angew. Chem. Int. Ed.*, 2010, **49**, 7076–7080.

(2) K. D. Nagy, B. Shen, T. F. Jamison and K. F. Jensen, *Org. Process Res. Dev.*, 2012, **16**, 976–981.

(3) M. Amrhein, B. Srinivasan, D. Bonvin and M. Schumacher, *Chemometr. Intell. Lab Syst.*, 1996, **33**, 17–33.

(4) M.-L. Wang, C.-J. Lin and J.-J. Jwo, *Chem. Eng. Commun.*, 1989, **79**, 189–205.

(5) H. Akaike, in *Selected papers of hirotugu akaike*, Springer, New York, 1998, pp. 199–213.

(6) P. Schmid, Masters Thesis, University of California, California, 1959.