

*SUPPORTING INFORMATION FOR*

**Model-based evaluation and data requirements for parallel  
kinetic experimentation and data-driven reaction identification and  
optimization**

Nathan Jiscoat, Evgeny Uslamin<sup>\*,‡</sup>, Evgeny A. Pidko<sup>\*</sup>

*Inorganic Systems Engineering, Department of Chemical Engineering, Delft University of  
Technology, Van der Maasweg 9, 2629 HZ Delft, The Netherlands*

*\* Corresponding authors: [e.uslamin@cyberhydra.nl](mailto:e.uslamin@cyberhydra.nl) (E. U.); [e.a.pidko@tudelft.nl](mailto:e.a.pidko@tudelft.nl) (E.A.P.)*

## Index

S1.	Kinetic parameter derivation .....	3
S2.	Versatility.....	4
S2.1.	Simple system.....	4
S2.2.	Simple catalytic system .....	5
S2.3.	Hydrogen diffusion limited system .....	6
S2.4.	Competing reactions .....	7
S2.5.	Consecutive reactions .....	8
S2.6.	Heterogeneous catalysis .....	9
S2.7.	Missing data .....	11
S3.	Reliability .....	12
S4.	SMAPE information .....	13

## S1. Kinetic parameter derivation

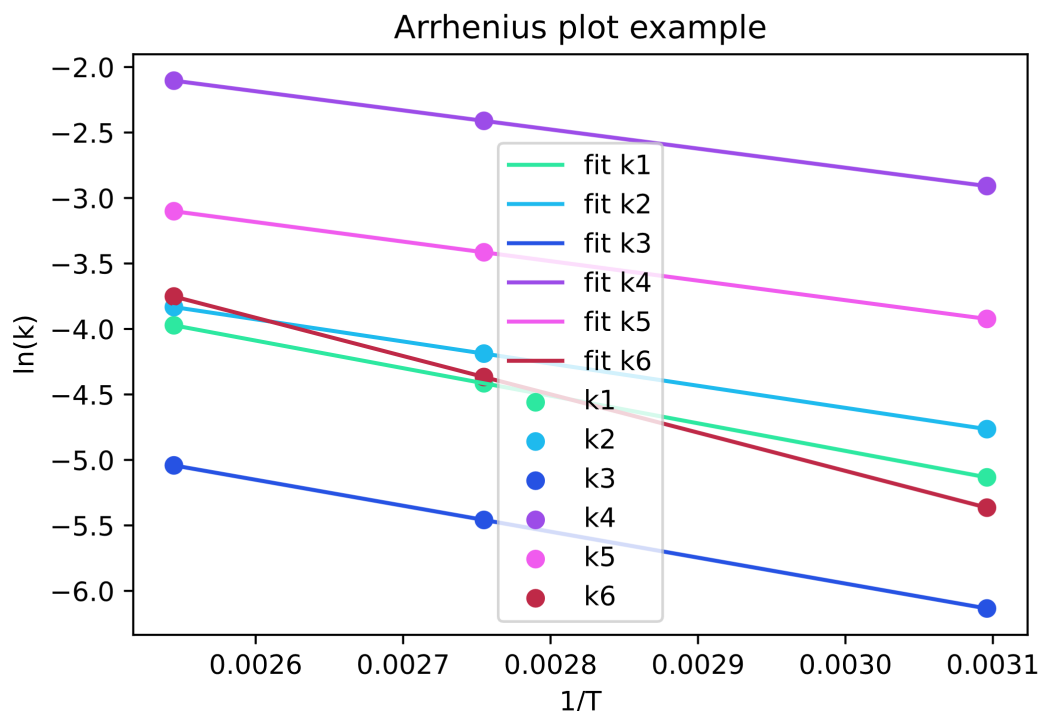


Figure S1. Example of how kinetic parameters are estimated from the rate constant estimates at different temperatures with an Arrhenius plot. Arrhenius pre-factors can be estimated from the intercept of the linear fit, and activation energies can be estimated from the slope

$$k_x = A_x * e^{\frac{-E_{act,x}}{R*T}} \quad Eq.S0$$

If reactions are done at different temperatures the algorithm can be used to guess activation energies and pre-constants. Equation S0 shows that when we have a rate it is separable into constants, a pre factor **A**, and an activation energy **Eact**. The fact that 2 unknown constants govern the final rate constant that is expressed, means that at least 2 temperature measurements are needed to separate them. Since estimating these kinetic parameters is much harder than directly guessing the rate constants, first guesses for the activation energies and pre-constants are made based on algorithm guesses of the rate constants for all temperature points. Guesses are made by linearly fitting the natural logarithm of a rate constant against the reciprocal of the temperatures they are measured at as shown in Figure S1.

## S2. Versatility

One of the big strengths of the algorithm is the versatility of the kinds of data and systems that can be put in. The following section is an exhibit of some different types of chemical datasets that can be fit to show the versatility of the method. Besides varying the models, dimensions of the dataset, data resolution, and noise parameters can be varied as well.

### S2.1. Simple system

First is the simplest chemical system, a single 1 to 1 reaction.

$$\frac{dC_a}{dt} = -k_1 * C_a \quad \text{Eq.S1}$$

$$\frac{dC_b}{dt} = k_1 * C_a \quad \text{Eq.S2}$$

Figure S2 shows a dataset comprised of 12 runs at temperatures of 330,360 and 390 K, and with starting concentrations of 0.5,1,3, and 5. The noise was 10% proportional noise and 0.05 structural noise, and a resolution of 20. Though there is a lot of noise the final SMAPE is 0.0689. The accuracy is due to the dataset size, and the small error is due to the fact that the noise always results in the best fit being a small amount off.

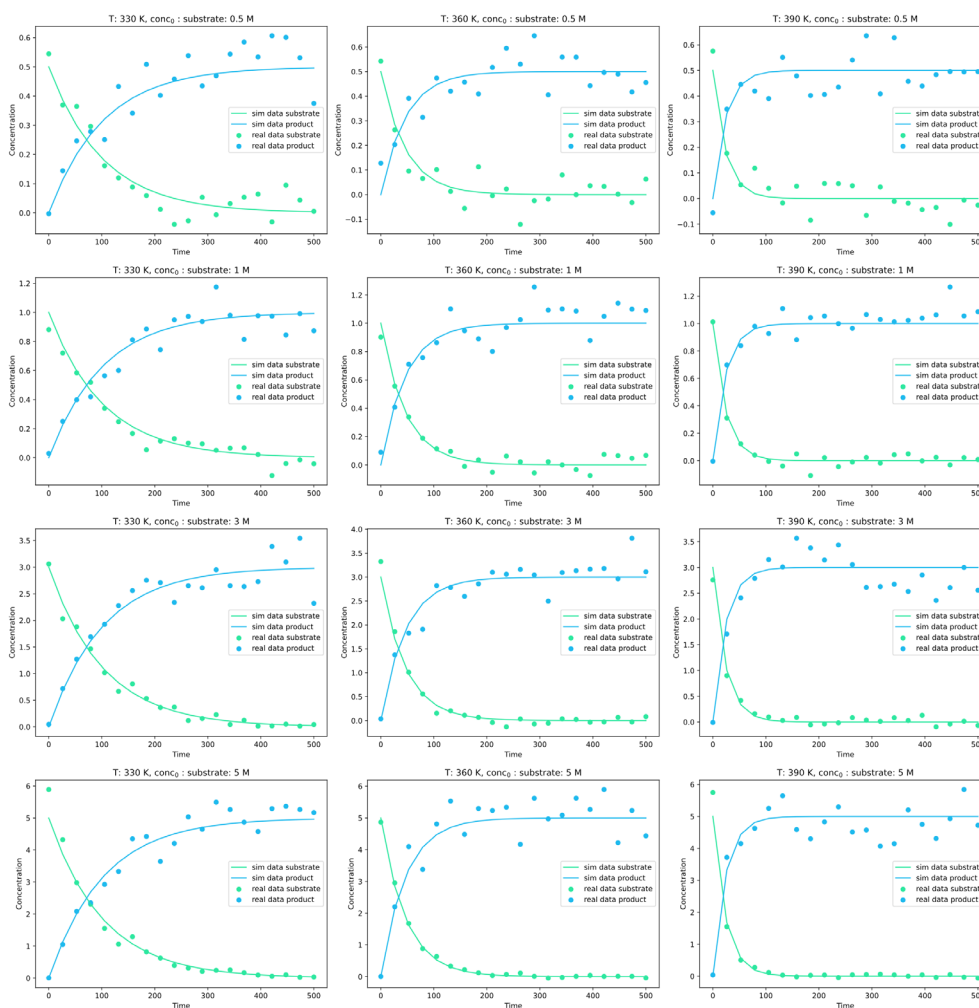


Figure S2. Results for fitting a dataset made with a simple differential system

## S2.2. Simple catalytic system

The simplest catalytic process can be described as simple  $A \rightarrow B$  reaction running in presence of catalyst that can deactivate.

$$\frac{dC_a}{dt} = -k_1 * C_a * C_{cat} \quad \text{Eq.S3}$$

$$\frac{dC_b}{dt} = k_1 * C_a * C_{cat} \quad \text{Eq.S4}$$

$$\frac{dC_{cat}}{dt} = -k_2 * C_{cat} \quad \text{Eq.S5}$$

Only the substrate and product concentrations were made available to the algorithm. For Figure S3 the dataset included 4 runs at 300, and 400 K, and with (substrate, catalyst) initial concentrations of (1,1) and (2,2) respectively. The datasets were modelled with the noise level of 2.5% and 0.01 for proportional and structural noise respectively and with the time resolution of 50 a.u. The final SMAPE is 0.156, showing that decent accuracy can be gained even for small datasets if they are simple and the noise levels are low.

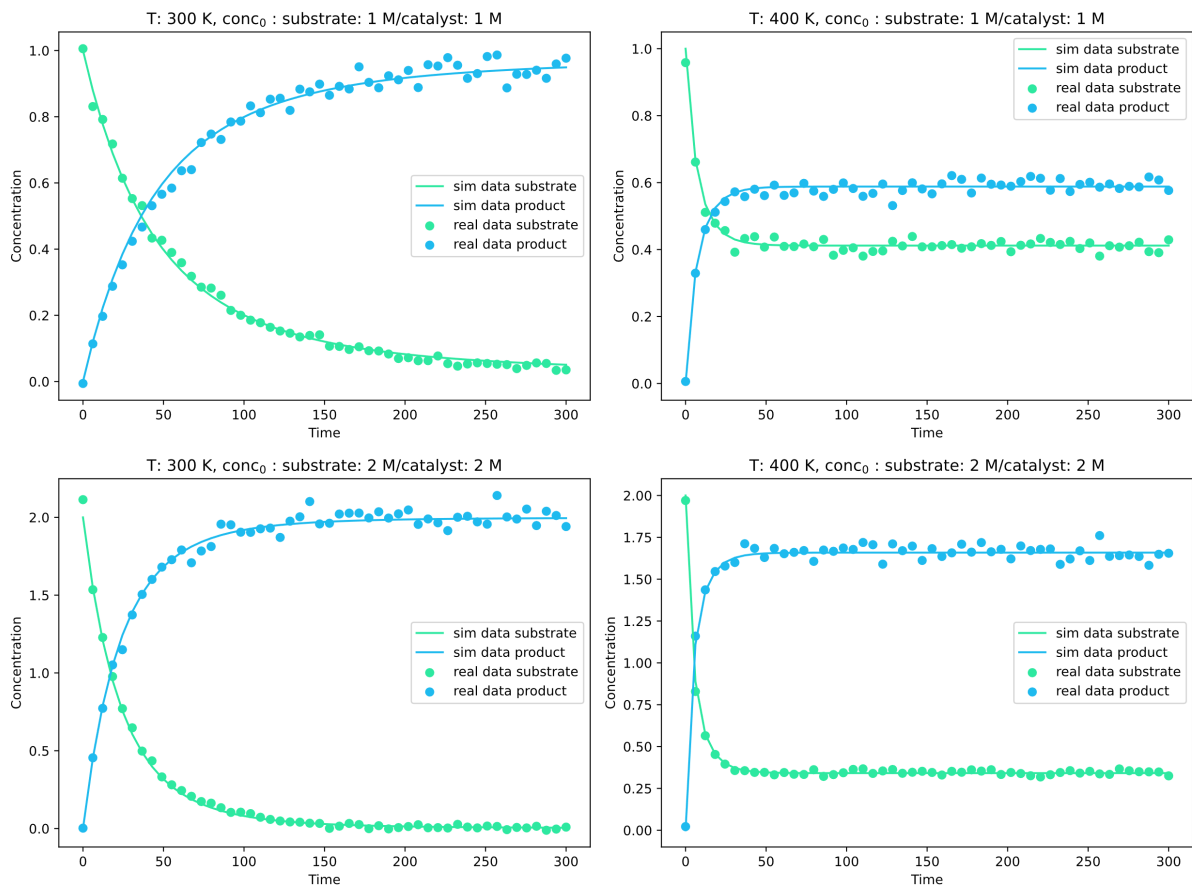


Figure S3. Results for fitting a dataset made with a simple catalytic differential system.

### S2.3. Hydrogen diffusion limited system

Another example exhibits how effects related to diffusion or limits can also be included into calculations. Here hydrogen is a reactant that can limit the reaction rate.

$$\frac{dH_2}{dt} = k_1 * (C_{H_2,max} - C_{H_2}) \quad \text{Eq.S6}$$

$$\frac{dC_a}{dt} = -k_2 * C_a * C_{H_2} \quad \text{Eq.S7}$$

$$\frac{dC_b}{dt} = k_2 * C_c * C_{H_2} \quad \text{Eq.S8}$$

Figure S4 shows a dataset comprised of 16 runs at temperatures of 320, 330, 340 and 350 K, and with starting concentrations of 0.5, 1, 2, and 4, the noise was 1% proportional noise and 0.005 structural noise, with a resolution of 15 points that are denser towards the beginning. The final SMAPE is 0.0612, which proves that including transport phenomena can be done while preserving accurate measurement.

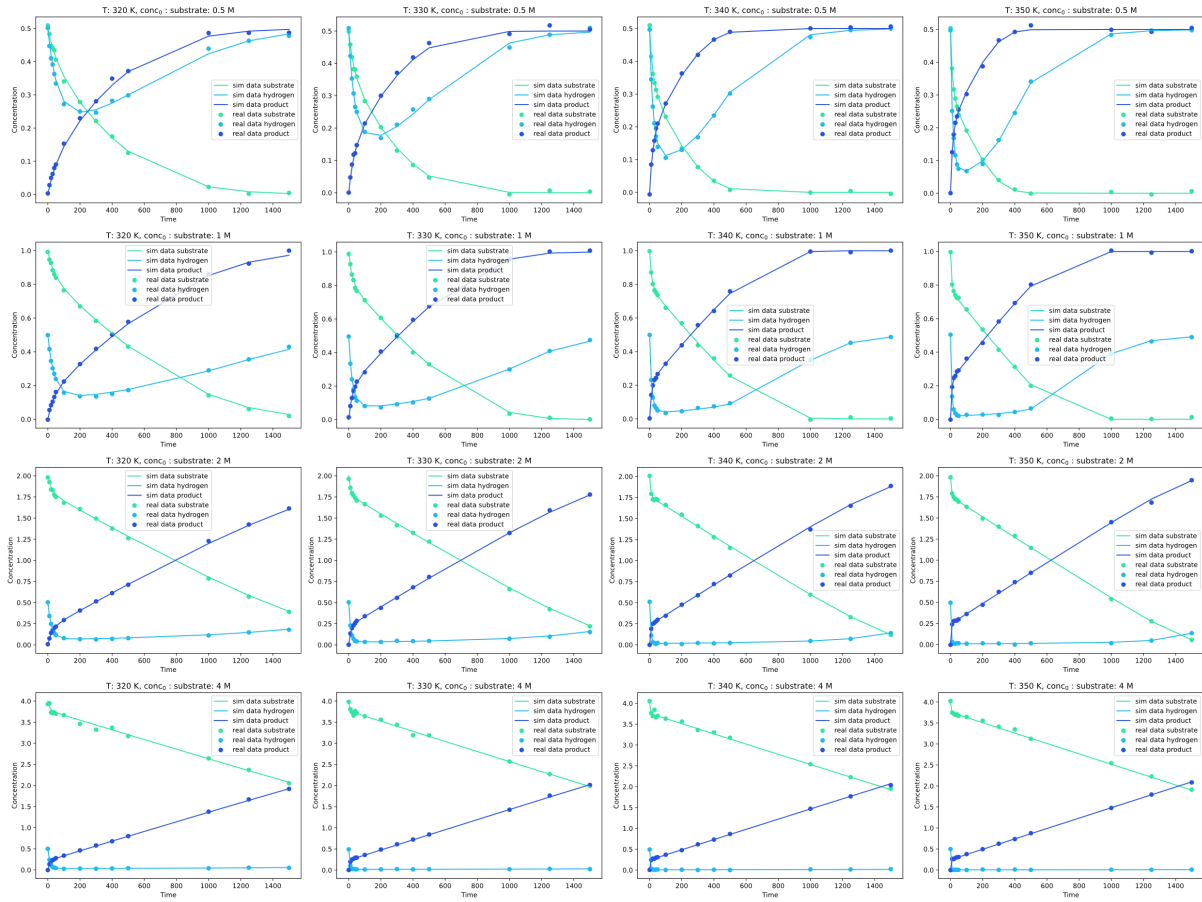


Figure S4. Results for fitting a dataset made with a set of differentials mimicking reactions with limiting transport phenomena.

## S2.4. Competing reactions case

Another common mechanism involves competing reactions. A system of 4 competing reactions was made to show that the constants of such systems are easily retrievable.

$$\frac{dC_a}{dt} = -k_1 * C_a - k_2 * C_a - k_3 * C_a - k_4 * C_a \quad \text{Eq.S9}$$

$$\frac{dC_b}{dt} = k_1 * C_a \quad \text{Eq.S10}$$

$$\frac{dC_c}{dt} = k_2 * C_a \quad \text{Eq.S11}$$

$$\frac{dC_d}{dt} = k_3 * C_a \quad \text{Eq.S12}$$

$$\frac{dC_e}{dt} = k_4 * C_a \quad \text{Eq.S13}$$

Figure S5 shows a dataset comprised of 6 runs at temperatures of 400, 500 and 600 K, and with starting concentrations of 0.1 and 0.4, the noise was 10% proportional noise and 0.005 structural noise, with a resolution of 50 points. The final SMAPE is 0.106. Even though some graphs would be unintelligible to the naked eye, the fits show very clearly how the reactions relate to each other.

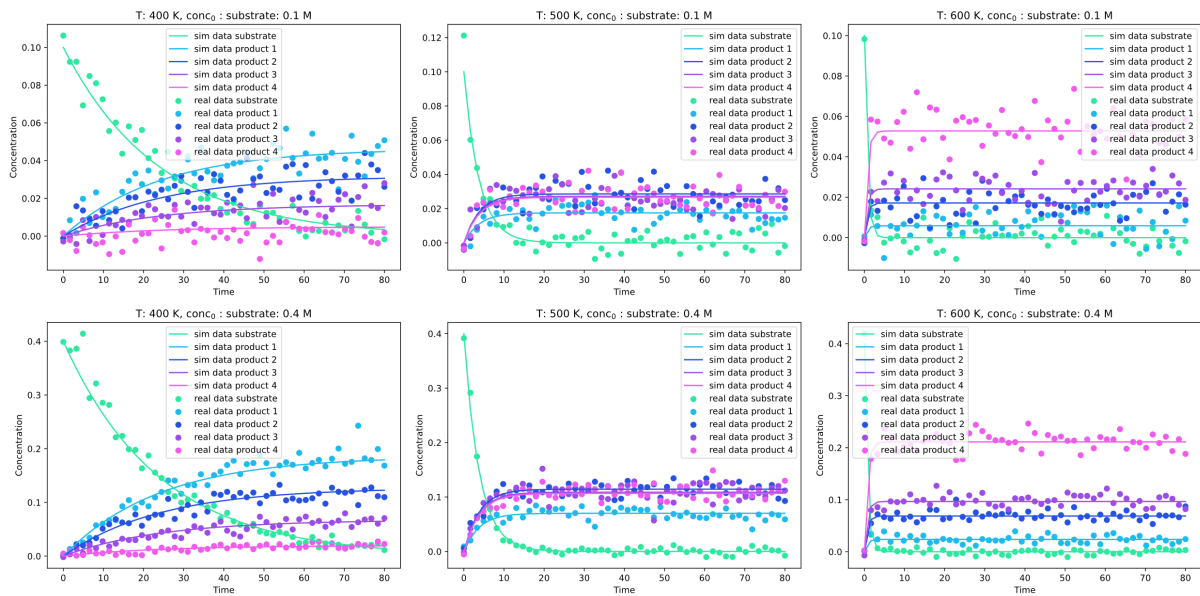


Figure S5. Results for fitting a dataset made with a set of differentials of many competing reactions.

## Consecutive reactions

Consecutive reaction is another common example. This system has 4 consecutive reactions.

$$\frac{dC_a}{dt} = -k_1 * C_a \quad \text{Eq.S14}$$

$$\frac{dC_b}{dt} = k_1 * C_a - k_2 * C_b \quad \text{Eq.S15}$$

$$\frac{dC_c}{dt} = k_2 * C_b - k_3 * C_c \quad \text{Eq.S16}$$

$$\frac{dC_d}{dt} = k_3 * C_c - k_4 * C_d \quad \text{Eq.S17}$$

$$\frac{dC_e}{dt} = k_4 * C_d - k_5 * C_e \quad \text{Eq.S18}$$

Figure S6 shows a dataset comprised of 6 runs at temperatures of 250, 260, 270, 280, 290 and 300 K, and with starting concentrations of 0.5, the noise was 0% proportional noise and 0.05 structural noise, with a resolution of 40 points. The final SMAPE is 0.144. The datasets are even more hard to interpret than the competing reaction dataset, especially at lower temperature, yet the fits are good, and the constants are quite accurate.

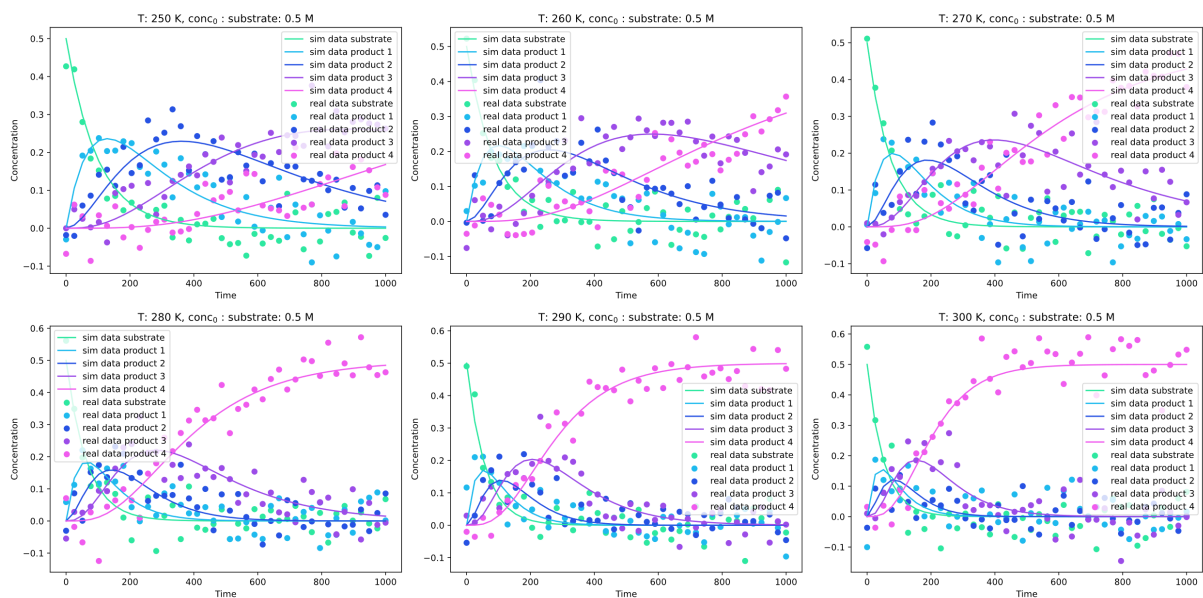


Figure S6. Results for fitting a dataset made with a set of differentials of many consecutive reactions.



## S2.5. Heterogeneous catalysis case

The last dataset shows that the method can be further extended to other complex processes such as surface reactions in heterogeneous catalysis. This last example demonstrates how even near-perfect data can be insufficient to reliably estimate some of the complexities of reactions.

Heterogeneous catalysis is vastly different in mechanics from homogenous catalysis. Heterogeneous processes involve, desorption, and often a presence of catalytic sites with different activities. The full model is illustrated in *Eqs. S17-21*. The model includes adsorption constants, degradation of the number of sites, and degradation of the ratio of certain sites.

Figure S7 shows a dataset comprised of 12 runs at temperatures of 320, 345, and 370, with starting concentrations of (reactant 1, reactant,2) of (1,1.1), and (1,1.5), and 2 catalysts, one with 10000 reactive sites and a corner site per edge site ratio of 0.1, and another with 5000 sites and a corner site per edge site ratio of 0.3, and noise was 1% proportional noise and 0 structural noise, with a resolution of 30 points.

$$\frac{dC_a}{dt} = -N_{sites} * (k_1 * \%corners + k_2 * \%edges) * \frac{k_a * C_a + k_b * C_b}{(1 + k_a * C_a + k_b * C_b)^2} \quad Eq.S19$$

$$\frac{dC_b}{dt} = -N_{sites} * (k_1 * \%corners + k_2 * \%edges) * \frac{k_a * C_a + k_b * C_b}{(1 + k_a * C_a + k_b * C_b)^2} \quad Eq.S20$$

$$\frac{dC_c}{dt} = N_{sites} * (k_1 * \%corners + k_2 * \%edges) * \frac{k_a * C_a + k_b * C_b}{(1 + k_a * C_a + k_b * C_b)^2} \quad Eq.S21$$

$$\frac{dN_{sites}}{dt} = N_{sites} * (k_1 * \%corners + k_2 * \%edges) * \frac{k_a * C_a + k_b * C_b}{(1 + k_a * C_a + k_b * C_b)^2} * 0.25 \quad Eq.S22$$

$$\frac{d\%corners}{dt} = k_3 * \%corners \quad Eq.S23$$

The model can still find good fits despite all the convolution. Yet the guessed constants led to a SMAPE of 0.512. This result shows that when metrics are a result of many effects at once, and the different metrics do not give a lot of different information, the estimation capacity is limited, and intricate tuning of initial conditions must likely be done to extract the maximum amount of information. Despite the high SMAPE, the errors are mostly attributable to the Arrhenius constants. All rate constants are still good approximations with an average SMAPE of 0.178, meaning that all orders of reaction are accurate and even close to the real values.

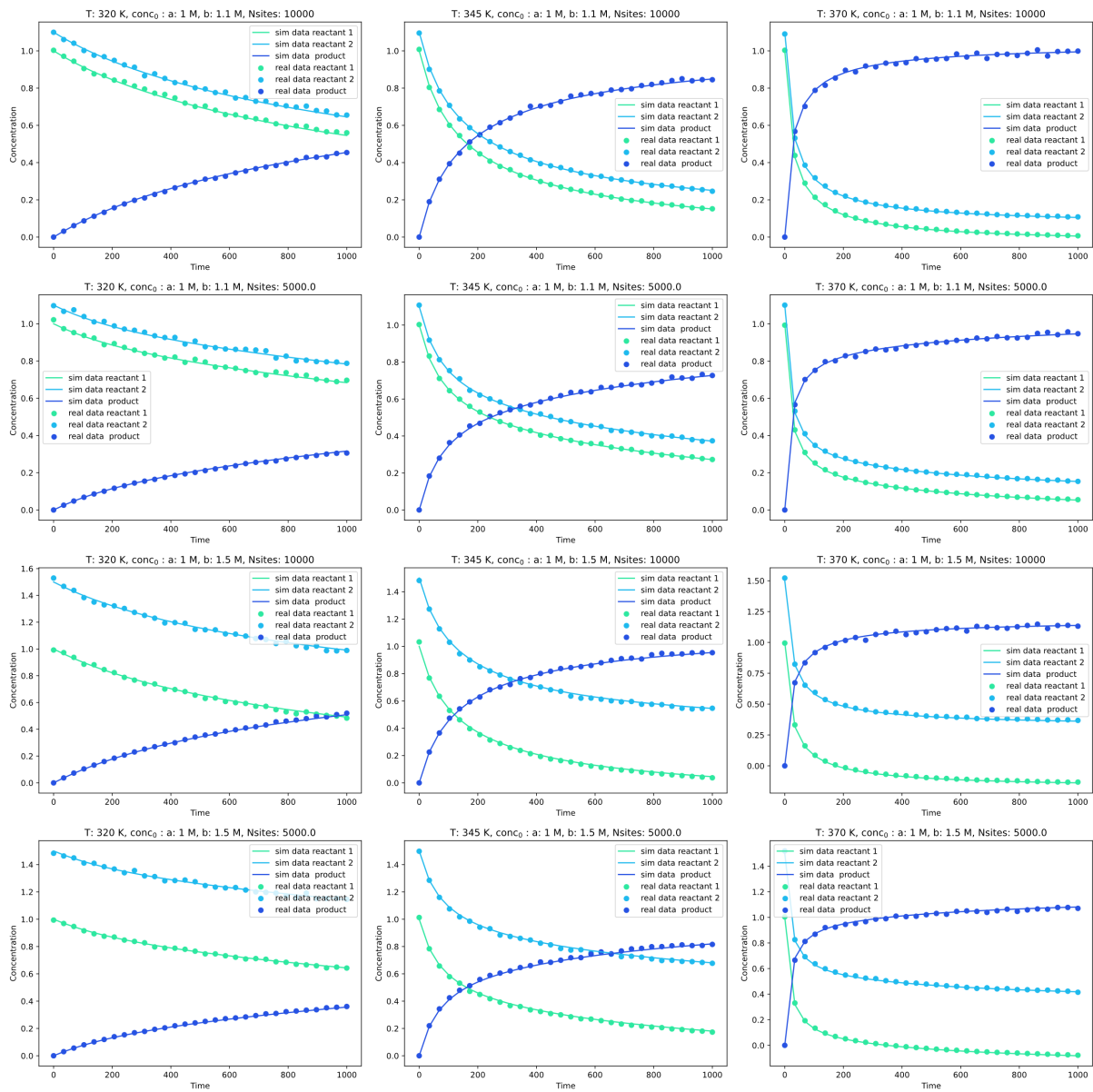


Figure S7. Results for fitting a dataset made with a set of differentials mimicking heterogeneous catalytic reaction behaviour.

## S2.6. Incomplete datasets

In Figure S8 it is shown that the algorithm can even fit data if it does not have data for all combinations of temperatures and initial conditions. This is the base system used in the research, so the SMAPE can be compared to the ones found in the research. Besides, noise and other parameters can also be varied in many ways, as long as it stems from one reaction system.

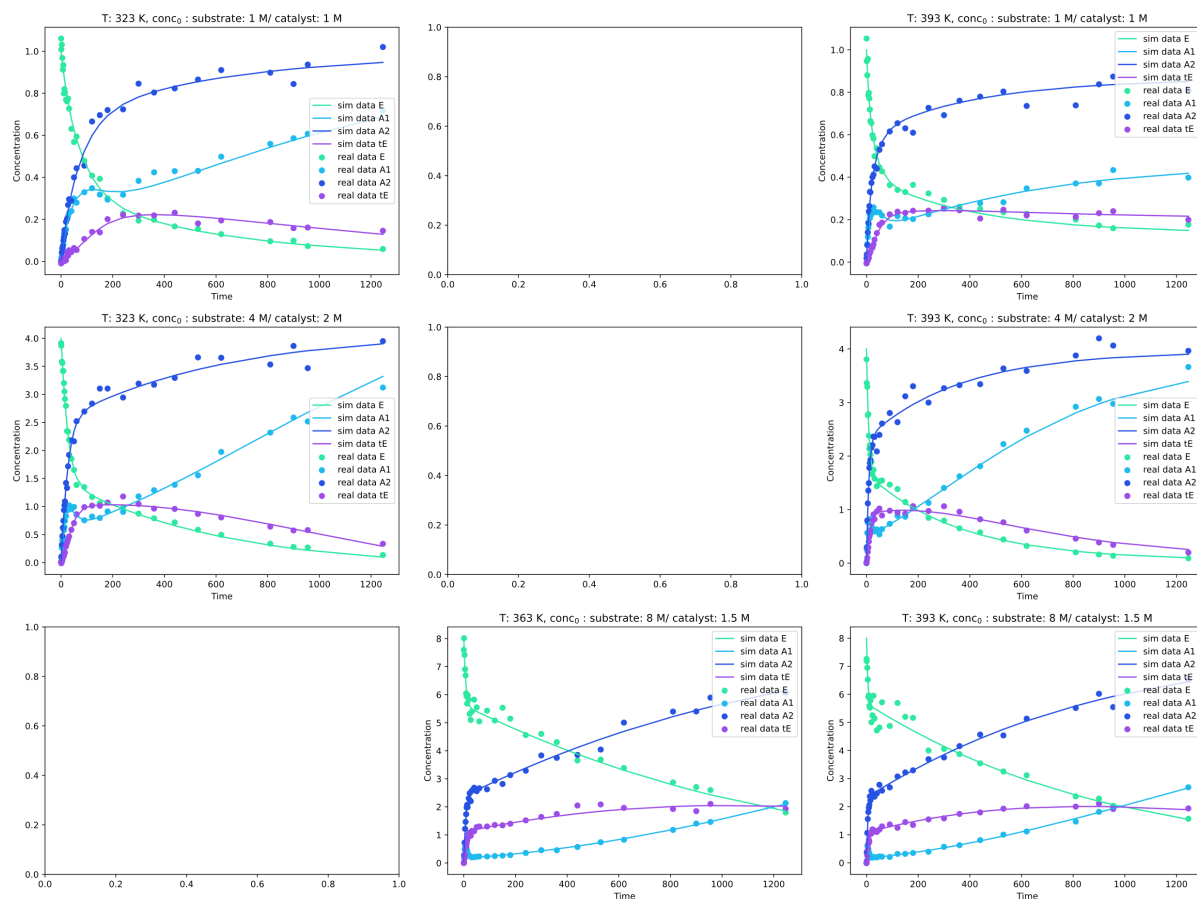


Figure S8. Results for fitting a dataset of the base homogeneous catalytic system that does not have a run for all combinations of included temperatures and initial conditions.

The retrieved SMAPE value for this run was 0.264, so the effect of removing random runs can be less severe than completely removing runs for one initial condition or temperature.

### S3. Reliability and performance

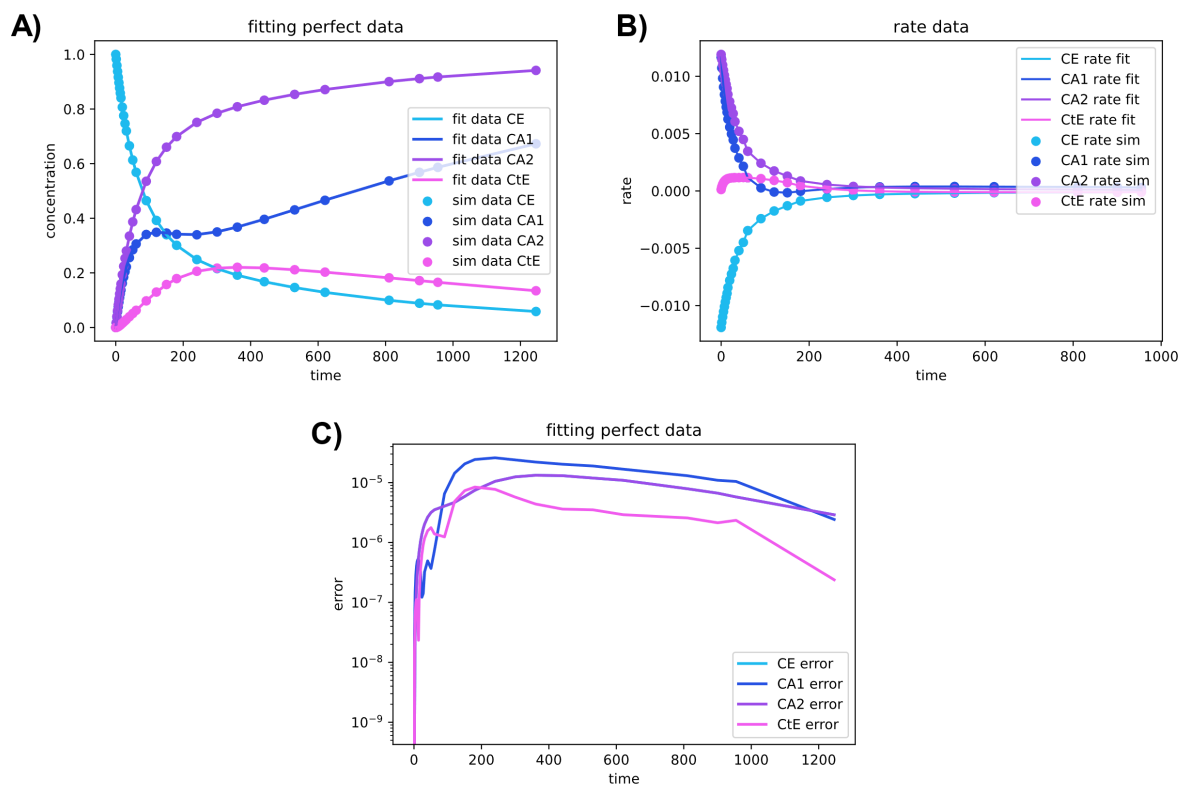


Figure S9. Graphs to illustrate how a fit of the data used for ODE fitting looks. A) concentration profiles, B) rate profiles, C) algorithm guess residual.

The data was first fit per temperature. An example of one result of this for the reaction profiles, rate profiles, and the estimation error of one of those fits is Figure S9, it is based on runs at 3 different initial conditions and 3 different temperatures. The residuals of the R2\_scores for the per temperature fitting are 2.41e-10, 2.65e-9, and 2.775e-5 for each temperature. For the total fitting the residual is 5.5e-6. Python has another simultaneous fitting package called symfit. The performance is similar to the algorithm in the current research. The input menu of the current research is more geared towards chemistry, given that it has a more compact input structure and that it allows for smooth workflows from fitting data per temperature to getting the total fit across different temperatures. Given the low error it is valid to assume any errors are attributable to the error in the data, not the algorithm.

Table S1. Kinetic parameter estimation time for different datasets.

Dataset\Estimation time	Per temperature fitting time	Simultaneous fitting time
Perfect data, 3 temperatures, 3 initial concentrations	147 seconds	2
Noisy data, 3 temperatures, 3 initial concentrations	158 seconds	2
Maximum used noise, 3 temperatures, 3 initial concentrations	65 seconds	3
Perfect data, 5 temperatures, 5 initial concentrations	446 seconds	8

For all data the algorithm estimation time is negligible compared to the amount of time it would take to run reactions that would generate such a dataset. Table S1 shows that estimation time for the per temperature fitting of the data in *Figure S9* is about 2 and a half minutes, and the estimation time for the subsequent full simultaneous fitting is about 2 seconds. Noise does not significantly influence the algorithm estimation time. The amount of data does influence the estimation time. For a dataset of 5 temperatures and 5 initial conditions the estimation times increase to about 7 and a half minutes, and 8 second respectively. Other alterations that influence dataset size will also affect the estimation time, but given that the reaction estimated is very complex, and that the data evaluated is very rich It is unlikely that the estimation time will be too long to be integrated into kinetic chemical measurements.

## S4. SMAPE information

Table S2. Reference for errors in guesses of parameters for certain SMAPE scores

SMAPE	Percent deviance
0.05	-5%/+5%
0.10	-10%/+11%
0.20	-19%/+22%
0.40	-33%/+50%
1	-66%/+200%
2	-inf%/inf%