**Automated Reaction Mechanism Constructor**

Miguel Ángel de Carvalho Serviaa, King Kuok (Mimi) Hiib, Klaus Hellgardta, Dongda Zhangc\*, Antonio del Rio Chanonaa\*

aDepartment of Chemical Engineering, Imperial College London, South Kensington, London, SW7 2AZ, United Kingdom

bDepartment of Chemistry, Imperial College London, White City, London, W12 0BZ, United Kingdom

cDepartment of Chemical Engineering, The University of Manchester, Oxford Road, Manchester, M13 9P, United Kingdom

[dongda.zhange@manchester.ac.uk](mailto:dongda.zhange@manchester.ac.uk); [a.del-rio-chanona@imperial.ac.uk](mailto:a.del-rio-chanona@imperial.ac.uk)

Abstract

In catalytic reaction engineering, the discovery of reaction mechanisms is paramount but challenging due to the intricate involvement of intermediates and limited prior knowledge. This study introduces a novel approach to generate the smallest feasible reaction mechanism (SFRM) that can accurately represent kinetic data sets. We propose an iterative algorithm based on a rule-based formulation, aiming to uncover the SFRM with minimal prior information about a system. Starting with the simplest conceivable mechanism, the algorithm advances by adding layers of complexity, estimating kinetic parameters, and evaluating mechanisms through the Akaike information criterion (AIC). Once a simpler mechanism demonstrates a smaller AIC value, it is selected as the optimal solution. Applied to the fructose to 5-hydroxymethylfurfural (HMF) mechanism, the methodology successfully uncovered the structure utilized for kinetic rate data generation. This novel framework offers experts a robust foundation for suggesting reaction mechanisms, addressing current limitations in mechanism discovery, emphasizing its utility not as a replacement but as a useful tool for experts.

**Keywords**: Chemical reaction engineering, automated reaction mechanism constructor.

* 1. Introduction

Understanding reaction mechanisms is crucial for constructing microkinetic models, which play a pivotal role in various sectors, including business and public policy decision-making. In the business domain, microkinetic models are important tools for chemical engineers to design and theoretically assess the profitability and viability of chemical processing plants. On the public policy-making side, international treaties like the Stockholm Convention on Persistent Organic Pollutants (POPs), aimed at minimizing the production and use of POPs, relied on kinetic models to understand the environmental behavior and degradation of chemicals such as dichlorodiphenyltrichloroethane. These models need to strike a balance between accuracy and dimensionality to ensure interpretability and efficient computational evaluation.

Traditionally, the development of reaction mechanisms and corresponding microkinetic models has been the responsibility of domain experts. Their knowledge is used to propose possible reaction intermediates, and the reaction steps by which they are formed and consumed. However, given the vast number of possible interactions among reactants, intermediates, and products – potentially hundreds of thousands – the manual assembly of these models is not only tedious and error-prone but also incredibly time-consuming.

The uptake of data-driven methodologies and enhanced analytical capabilities, which have significantly increased the amount of available kinetic data, has catalyzed the creation of automated approaches for mechanism development. These have been previously divided into two categories: one relies on combinatorial algorithms to generate all possible reactions based on electronic configuration congruence, while the other generates characteristic reactions of a known reaction class based on the reactants and products involved.

The combinatorial approach, bounded only by electronic congruence, often results in enormous reaction networks that require significant reduction. These large reaction mechanisms compromise computational efficiency and model interpretability, as mentioned. The latter approach generates more tractable networks but is dependent upon pre-existing knowledge of reaction classes, which may not always be accessible. For in-depth discussions on these methodologies, reviews by van de Vijver et al. (2014) and Ratkiewicz and Truong (2005) are recommended.

Both methods present inherent limitations, prompting the need for a ‘middle-ground’ approach that can construct a minimalistic reaction network without pre-existing system knowledge. This gap motivated our formulation of SIMBA (SImplest Mechanism Builder Algorithm), which aspires to create the smallest feasible reaction mechanism (SFRM) that can accurately explain available kinetic data without prior knowledge. The following sections of this paper are laid out as follows: Section 2 delves into the inner-workings of SIMBA, detailing the iterative construction, optimization, and selection of the SFRM; Section 3 presents the case study employed to validate the approach, including the generation of in-silico data; Section 4 showcases the results where SIMBA successfully uncovers the underlying microkinetic model of the case study; and Section 5 concludes with the primary contributions of the study and proposes enhancements and future research directions.

* 1. Methodology

SIMBA (SImplest Mechanism Builder Algorithm) is tailored to develop microkinetic models using kinetic data, focusing on identifying the informationally smallest reaction mechanism that accurately describes the available data. It comprises four key phases: (I) reaction chain generation phase, where it proposes increasingly complex reaction mechanisms; (II) ODE system builder phase, translating these mechanisms into microkinetic models represented by ordinary differential equations (ODE); (III) system integration phase, estimating kinetic parameters using an optimization algorithm; and (IV) comparison phase, assessing each model's performance with the Akaike Information Criterion (AIC) to decide if further iterations are necessary. Each phase contributes to SIMBA’s goal of efficiently deriving an accurate and minimalistic representation of the kinetic system. The following subsections delve into each of these phases in greater detail.

* + 1. Reaction Chain Generation

The initial phase of SIMBA involves the proposition of a series of reaction mechanisms, starting from the most basic possibility and gradually escalating in complexity with each iteration. This reaction chain generation phase begins by assuming reaction stoichiometry is known and discounting termolecular or higher-order interactions due to their relative rarity. For instance, consider a straightforward reaction stoichiometry such as ‘A → Y + Z’, with A, Y, and Z representing arbitrary reactants and products.

With the reaction stoichiometry known and constraining reactions to be either bimolecular or unimolecular, SIMBA identifies the simplest possible mechanism for this example as a single-step reaction: ‘A → Y + Z’. Upon completing the first iteration, the algorithm proceeds to a second iteration, introducing an intermediate, labeled ‘B’. This addition broadens the scope of potential mechanisms, now encompassing two-step mechanisms. In this second iteration, SIMBA generates and considers all three feasible mechanisms: (I) ‘A → B’, followed by ‘B → Y + Z’; (II) ‘A → B + Y’, followed by ‘B → Z’; and (III) ‘A → B + Z’, followed by ‘B → Y’.

Should a third iteration be needed, as determined by SIMBA’s fourth phase, another intermediate, ‘C’ in this case, would be introduced. The algorithm then constructs the corresponding possible mechanisms incorporating this new intermediate. At each stage, SIMBA systematically generates all possible reaction mechanisms under the set of constraints and feeds these into the next phase of the algorithm for further evaluation and optimization. This structured approach ensures a thorough exploration of potential reaction pathways, progressively building in complexity, to accurately model the kinetic behavior of the system under study in the simplest way possible.

* + 1. ODE System Builder

Following the initial phase of generating reaction mechanisms, SIMBA advances to the next stage: the transformation of these theoretical mechanisms into microkinetic models, each represented by a system of ODEs. This automatic formulation is critical for translating theoretical reaction mechanisms into quantifiable models that we can then optimize and evaluate.

To illustrate this process, Figure 1 depicts the expected inputs and outputs of the ODE system builder, referred to as ‘make\_system’, for two sample reaction mechanisms discussed in the previous subsection. From Figure 1, it is evident that the output of the ODE system builder is a function characterized by unknown kinetic parameters. In the third phase of SIMBA, these unknown parameters are estimated, enabling the proposed microkinetic model to be compared against available kinetic data.

A screen shot of a computer program

Description automatically generated

Figure 1: Representation of the inputs required by the ODE builder and its expected output. Two examples of increasing complexity are shown.

* + 1. System Integration

In this phase, SIMBA employs an optimization algorithm to estimate the kinetic parameters for each proposed model. The accuracy of these models in representing the kinetic behavior of the system heavily depends on this phase, making it essential for refining the models based on the available experimental data. For solving this optimization problem, commonly known as parameter estimation, in this work we employed the limited-memory bounded Broyden-Fletcher-Goldfarb-Shanno (L-BFGS-B) algorithm. The optimization problem tackled by SIMBA is formally defined as:

**(1)**

In this formulation, represents the set of unknown kinetic parameters that need to be estimated; denotes the number of sampling points at which experimental data is available, are the predicted concentrations of the observed species at time , based on the model with parameters ; are the actual measured concentrations of the observed species at the corresponding time ; and is designed to compute the sum of squared errors between and . The L-BFGS-B algorithm’s role in this phase is to iteratively adjust the parameters to minimize the error computed by , effectively tuning the model to align with the observed kinetic behavior as close as possible. Once this phase is concluded and the kinetic parameters are estimated, the models can be evaluated and compared against each other.

* + 1. Model Comparison

Having developed various microkinetic models with their respective optimized kinetic parameters to best fit the observed data, the final stage in SIMBA involves a quantitative comparison of these models. This comparative analysis is essential to decide whether SIMBA should continue generating new reaction mechanisms or conclude its operation. The decision-making process in this phase employs the Akaike information criterion (AIC) to evaluate and compare the models.

For a given model *m* with parameters of dimension *dm*,the AIC is computed as follows:

, **(2)**

where represent the negative log-likelihood given a data set . When comparing two models, *m1* and *m2*, the model with the lower AIC value is considered the better one. The choice of the AIC as the model selection criteria is not arbitrary. In fact, in literature, it has been shown that AIC is a great information criterion that tends to perform better than other criteria within the kinetic modeling paradigm (de Carvalho Servia et al., 2023). The usage of AIC is also important because it helps us balance model complexity and overfitting through the penalty term that it includes.

The routine by which we determine if SIMBA terminates or continues is straightforward: if the current iteration produces a mechanism with a lower AIC than the previous iteration , then SIMBA proceeds to the next iteration ; otherwise, SIMBA terminates. This iterative process ensures that SIMBA not only explores a wide range of potential mechanisms but also converges towards the smallest and most representative model in describing the kinetic behavior of the system under study.

* 1. Case Study

The case study used for the performance analysis of SIMBA is the dehydration of fructose to 5-hydroxymethylfurfural (HMF), catalyzed by [BmimHSO3][HSO4]. The overall reaction is show below:

. **(3)**

A microkinetic model can be developed using the reaction mechanism proposed in Hu et al. (2023), where HMF is formed through three dehydration steps and three different intermediates are produced and consumed. The proposed mechanism is shown below:

**(4)**

where A, B and C represent fructose, water and HMF respectively; and Int1, Int2 and Int3 denote the three intermediates that are formed during the conversion of fructose to HMF.

To generate the in-silico data set, three computational studies were carried out with the following initial conditions (in molar units): (CA(t=0), CB(t=0), CC(t=0) ∈ {(4, 0, 0), (6, 2, 1), (4, 2, 0)}. For each experiment, the concentration of the reactant and products are recorded 30 times, at evenly spaced intervals between time t0=0 h and tf=2 h. The kinetic parameters were defined as: k1=1.514 h-1, k2=5.259 h-1, k3=9.352 h-1 and k4=2.359 h-1.

Gaussian noise is added to the in-silico measurements to simulate a realistic chemical system. The added noise had zero mean and a standard deviation of 0.2 for all observed species. This noise addition allows the approximation of the response of a real system.

* 1. Results and Discussion

The implementation of SIMBA on the in-silico data indicated that only three iterations were necessary to identify the smallest feasible reaction mechanism (SFRM) for our study. In other words, the optimal SFRM was discovered during the second iteration, as indicated by the AIC values. For clarity and conciseness, Table 1 presents the microkinetic models, lists the estimated kinetic parameter values, along with the AIC values for the best reaction mechanism of each iteration executed by SIMBA. For reference, the sum of squared errors (SSE) between the in-silico data and the data-generating model is 10.780 M2, whereas the SSE between the in-silico data and the model discovered by SIMBA is 12.664 M2.

The performance of SIMBA on the case study of the production of HMF from fructose is promising. Given the stoichiometry and acknowledging the rarity of ter- and higher-order molecular interactions, SIMBA successfully identified the underlying mechanism and microkinetic model that governs the system’s dynamics. This achievement underscores SIMBA’s potential as an effective tool in kinetic discovery. However, it is crucial to recognize SIMBA’s limitations: while it proposes a mechanism with intermediates, it does not chemically identify these intermediates. In our case, deducing the nature of these intermediates is feasible through basic atom balances and fundamental chemical knowledge. Nevertheless, in more complex chemical systems, identifying these intermediates might not be as straightforward, necessitating the insight of an experienced experts. This limitation highlights a notable challenge in the algorithm’s application.

Despite this, the results reinforce the utility of SIMBA as a valuable tool for chemists and reaction engineers. It is not a replacement for expert knowledge but rather a tool that can significantly expedite the process of mechanistic discovery. By providing a robust initial ‘guess’ of the reaction mechanism, SIMBA enables experts to efficiently put together the puzzle pieces of complex chemical processes, potentially shortening the timeline for understanding and optimizing such systems.

* 1. Conclusions

This study introduces SIMBA, an algorithm encompassing four phases: reaction chain generation, ODE builder, system integration, and model comparison. Tested on the catalyzed synthesis of HMF from fructose, SIMBA successfully identified the smallest feasible reaction mechanism from in-silico data, matching the literature-sourced microkinetic model. This highlights SIMBA’s potential in accelerating mechanistic discovery, despite its limitation in not chemically identifying reaction intermediates.

Future research should focus on integrating chemical knowledge into SIMBA, crucial for aiding the discovery of complex systems, and on incorporating uncertainty quantification in model predictions. Enhancing SIMBA in these aspects will significantly improve its effectiveness and reliability in discovering and optimizing microkinetic models, making it a more comprehensive tool for chemists and reaction engineers.

Table 1: The best reaction mechanism at each iteration of SIMBA, along with the corresponding microkinetic model, the estimated kinetic parameters, and the AIC value. It demonstrates that the SFRM was discovered in iteration 2, where the chosen mechanism perfectly matches with the data-generating one.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration | Reaction Mechanism | Microkinetic Model | Estimated Kinetic Parameters | AIC Value |
| 1 |  |  | h-1  h-1  h-1 | 287.030 |
| 2 |  |  | h-1  h-1  h-1  h-1 | 281.528 |
| 3 |  |  | h-1  h-1  h-1  h-1  h-1 | 287.931 |

References

R. van de Vijver, N. M. Vandewiele, P. L. Bhoorasingh, B. L. Slakman, F. S. Khanshan, H.-H. Carstensen, M.-F. Reyniers, G. B. Marin, R. H. West, K. M. van Geem, 2014, Automatic Mechanism and Kinetic Model Generation for Gas- and Solution-Phase Processes: A Perspective on Best Practices, Recent Advances, and Future Challenges, International Journal of Chemical Kinetics, 47, 4, 199-231

A. Ratkiewicz, T. N. Truong, 2005, Automated Mechanism Generation: From Symbolic Calculation to Complex Chemistry, International Journal of Quantum Chemistry, 106, 1, 244-255

M. Á. de Carvalho Servia, I. O. Sandoval, K. K. Hii, K. Hellgardt, D. Zhang, E. A. del Rio Chanona, 2023, The Automtaed Discovery of Kinetic Rate Models – Methodological Frameworks, arXiv

J. Hu, M. Yu, Y. Li, X. Shen, S. Cheng, T. Xu, C. Ge, Y. Yu, Z. Ju, 2023, Dehydration mechanim of fructose to 5-hydroxymethylfurfural catalyzed by functionalized ionic liquids: a density functional theory study, New Journal of Chemistry, 47, 11525-11532.