This article can be cited before page numbers have been issued, to do this please use: M. Á. de Carvalho Servia, K. K. M. Hii, K. Hellgardt, D. Zhang and A. Del Rio Chanona, *Chem. Sci.*, 2025, DOI: 10.1039/D5SC01473E.

View Article Online
DOI: 10.1039/D5SC01473E

Simplest Mechanism Builder Algorithm (SiMBA): An Automated Microkinetic Model Discovery Tool

Miguel Ángel de Carvalho Servia
Department of Chemical Engineering
Imperial College London
South Kensington, London, SW7 2AZ, UK
m.de-carvalho-servia21@imperial.ac.uk

King Kuok (Mimi) Hii
Department of Chemistry
Imperial College London
White City, London, W12 0BZ, UK
mimi.hii@imperial.ac.uk

Klaus Hellgardt
Department of Chemical Engineering
Imperial College London
South Kensington, London, SW7 2AZ, UK
k.hellgardt@imperial.ac.uk

Dongda Zhang
Department of Chemical Engineering
The University of Manchester
Manchester, M13 9PL, UK
dongda.zhang@manchester.ac.uk

Ehecatl Antonio del Rio Chanona *
Department of Chemical Engineering
Imperial College London
South Kensington, London, SW7 2AZ, UK
a.del-rio-chanona@imperial.ac.uk

August 5, 2025

Abstract

Microkinetic models are key for evaluating industrial processes' efficiency and chemicals' environmental impact. Manual construction of these models is difficult and time-consuming, prompting a shift to automated methods. This study introduces SiMBA (Simplest Mechanism Builder Algorithm), a novel approach for generating microkinetic models from kinetic data. SiMBA operates through four phases: mechanism generation, mechanism translation, parameter estimation, and model comparison. Our approach systematically proposes reaction mechanisms, using matrix representations and a parallelized backtracking algorithm to manage complexity. These mechanisms are then translated into microkinetic models represented by ordinary differential equations, and optimized to fit available data. Models are compared using information criteria to balance accuracy and complexity, iterating until convergence to an optimal model is reached. Case studies on an aldol condensation reaction, and the dehydration of fructose demonstrate SiMBA's effectiveness in distilling complex kinetic behaviors into simple yet accurate models. While SiMBA predicts intermediates correctly for all case studies, it does not chemically identify intermediates, requiring expert input for complex systems. Despite this, SiMBA significantly enhances mechanistic exploration, offering a robust initial mechanism that accelerates the development and modeling of chemical processes. By automating microkinetic model generation from a data-first approach, SiMBA opens new avenues for future research in automated mechanism discovery.

Keywords: chemical reaction engineering, microkinetic model generation, automated knowledge discovery

View Article Online
DOI: 10.1039/D5SC01473E

1 Introduction

Microkinetic models are indispensable tools in both business and policymaking due to their ability to evaluate the efficiency of industrial processes and the environmental impact of chemicals. These models are particularly vital in sectors such as pharmaceuticals [1, 2, 3], petrochemicals [4, 5, 6], and environmental engineering [7, 8, 9], where they help optimize production processes, reduce costs, and improve sustainability. For instance, in the pharmaceutical industry, microkinetic models facilitate the understanding of drug interactions and optimize synthesis pathways, accelerating drug development [10, 11, 12]. Similarly, in environmental policy, these models provide insights into the behavior of chemical reactions, supporting the formulation of regulations and safety standards, for example the Montreal Protocol and Stockholm Convention [13, 14]. By simulating the steps of chemical reactions at the molecular level, microkinetic models offer a detailed understanding of reaction mechanisms, which is essential for making informed decisions in various sectors, thus balancing economic growth with environmental protection.

Despite their importance, the manual construction of microkinetic models is a complex, time-consuming, and error-prone process [15, 16]. Traditional methods require experts to manually identify possible reaction steps and intermediates, a task that can involve analyzing hundreds of thousands of potential interactions. This meticulous process is not only slow but also susceptible to human error and often results in models that are either overly simplified or unnecessarily complex. The increasing complexity of modern chemical systems further exacerbates these challenges, highlighting the need for more efficient and reliable approaches. Consequently, there has been a significant shift towards the development of automated methods for constructing these models [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27], exploiting advances in data-driven methodologies and computational resources to streamline and enhance the accuracy of the modeling process.

The general trend towards automation, or scientific machine learning, offers substantial benefits, including increased efficiency and reduced error rates, compared to traditional manual methods [28]. Various algorithms for generating mechanisms have been developed, typically falling into two categories: combinatorial algorithms and algorithms based on reaction classes [16]. In the former approach, the generation of the entire set of possible reactions is based solely on the congruence of the electronic configurations of reactants and products, utilizing graph theory and bond-electron matrix representations of molecules [29, 30, 18, 31]. These methods can produce highly detailed and comprehensive reaction networks. The latter approach involves algorithms that, after recognizing the compounds as belonging to a certain class, generate only those reactions known to be characteristic of that class. While this method produces more compact networks, it requires prior knowledge of existing reaction classes [32, 33, 34, 35]. Combinatorial algorithms often yield overly complex mechanisms that can hinder computational efficiency and interpretability whilst making experimental validation and parameter estimation challenging (and often impossible). Conversely, algorithms based on reaction classes are limited by the availability of pre-existing reaction knowledge which may be limited within the scope of mechanism discovery for novel reactions. For a more in-depth discussion of these methodologies, reviews by Ratkiewicz and Truong [34], and by Van de Vijver et. al. [36] provide valuable insights.

Recently, alternative automated approaches leveraging artificial intelligence have also emerged, such as the method introduced by Burés and Larrosa, which employs deep learning techniques to classify organic reaction mechanisms directly from kinetic data without requiring explicit derivation of rate laws or enumerating reaction structures beforehand [37]. While conceptually distinct from our proposed method, their method similarly addresses the complexity and interpretability challenges inherent to mechanistic discovery, highlighting the growing role and complementary potential of machine learning-based strategies in reaction engineering.

In this work, we propose a new approach, the Simplest Mechanism Builder Algorithm (SiMBA), which is designed to circumvent the necessity for substantial prior knowledge required by reaction class approaches, and to avoid the proposal of overly complex mechanisms yielded by combinatorial approaches. This is done by tackling the problem of automated generation of mechanisms from a data-first perspective, ensuring that whatever mechanism is proposed, is both physically reasonable and only as complex as the data allows. SiMBA generates microkinetic models that progressively increase in complexity based on the provided data. The algorithm begins with the simplest possible mechanism, yielding the most straightforward microkinetic model. The complexity of the mechanism is then incrementally increased, thus increasing the number of parameters of the corresponding microkinetic model. This process continues as long as there is informational gain from the added parameters, which is evaluated using the Akaike Information Criterion (AIC). By balancing model simplicity and accuracy, SiMBA ensures the generation of robust and sensible microkinetic models, effectively bridging the gap between theoretical exploration and practical applicability. While alternative model discrimination measures could be employed, we chose the AIC based on prior work demonstrating its superior performance in selecting data-generating kinetic models from a set of candidates [38]. This minimalist approach is structurally and fundamentally different than previous methods, in that the main objective is to discover the most accurate and parsimonious mechanism given the dataset available, with as little prior information as possible.

View Article Online
DOI: 10.1039/D5SC01473E

This research represents an advancement in the field of microkinetic modeling, offering a novel approach that overcomes many of the challenges associated with existing automated methods. By systematically generating, refining, and evaluating microkinetic models, SiMBA provides a robust framework for developing accurate and experimentally viable reaction mechanisms. The algorithm's ability to distill complex chemical processes into simple, yet precise models has the potential to accelerate the design and optimization of chemical processes across various industries. Ultimately, SiMBA can become a useful tool for chemists and engineers, facilitating the rapid discovery and refinement of microkinetic models, thereby advancing our understanding of chemical reactions in diverse contexts.

The rest of the paper is organized as follows: in Section 2 our proposed method is motivated and described in detail; in Section 3 we introduce the three case studies that are used to analyze the performance of SiMBA highlighting the data-generation procedure and the results of the study are presented and amply discussed along with the shortcomings of the proposed methodology; and in Section 4 the key findings are presented with a brief outlook on future research.

2 Methodology

SiMBA (Simplest Mechanism Builder Algorithm) has been tailored to develop microkinetic models using kinetic data, focusing on identifying the informationally smallest reaction mechanism that accurately describes the available data. By focusing on the balance between model accuracy and simplicity, SiMBA aims to make the process of mechanism discovery more accessible, efficient, and reliable.

SiMBA is comprised of four key phases:

1. **Mechanism generation phase**: utilizes a parallelized backtracking algorithm to generate all physically-sensible mechanisms for a given set of complexity parameters: the number of elementary steps and intermediates. This phase ensures that only feasible reaction pathways are considered, significantly reducing the computational burden;
2. **Mechanism translation phase**: the proposed mechanisms, represented by a matrix, are converted into executable microkinetic models, specifically systems of ordinary differential equations (ODEs). This translation is crucial for transforming reaction networks into practical models that can be analyzed and simulated;
3. **Parameter estimation phase**: the kinetic parameters of the proposed microkinetic models are estimated by minimizing the error between the model predictions and the observed kinetic data. This is achieved using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm;
4. **Model comparison phase**: involves evaluating the generated models using the AIC to determine the best microkinetic model for a given iteration. This phase also decides whether further iterations and additional complexity provide enough informational gain to justify continuing the algorithm.

By systematically progressing through these phases, SiMBA ensures the development of robust, accurate, and computationally efficient microkinetic models.

Our methodology also offers a closed-loop approach for refining models if the SiMBA's output is unsatisfactory, whether due to conflicts with prior knowledge (e.g., belief that the microkinetic model should involve more/less chemical species) or due to poor model fitting (e.g., the model failing to accurately capture the non-linearities in the kinetic data). In such cases, the modeler can opt to conduct an optimal experiment specifically designed to enhance model discovery – using model-based design of experiments (MBDoE), more specifically the Hunter-Reiner criterion [39] – and then integrate this new data with the initial dataset. With the additional experimental data, the methodology can be re-applied, allowing for iterative refinement and re-evaluation of the output. Practically, this discriminatory experiment could also serve to validate the models proposed in earlier iterations, rather than relying solely on the AIC. The process can be repeated as many times as necessary or until the experimental budget is exhausted. Figure 1 visually represents the SiMBA workflow, highlighting the key phases of the methodology. The following subsections provide a detailed account of each of these phases.

2.1 Mechanism Generation

The primary goal of the this initial phase is to systematically generate all physically plausible reaction mechanisms given a set of specified constraints. This phase sets the foundation for the microkinetic modeling process by exploring the potential reaction pathways that could feasibly describe the overall chemical reaction under investigation. By considering physical and chemical constraints, we ensure that only realistic and meaningful mechanisms are carried forward for further analysis.

View Article Online
DOI: 10.1039/D5SC01473E

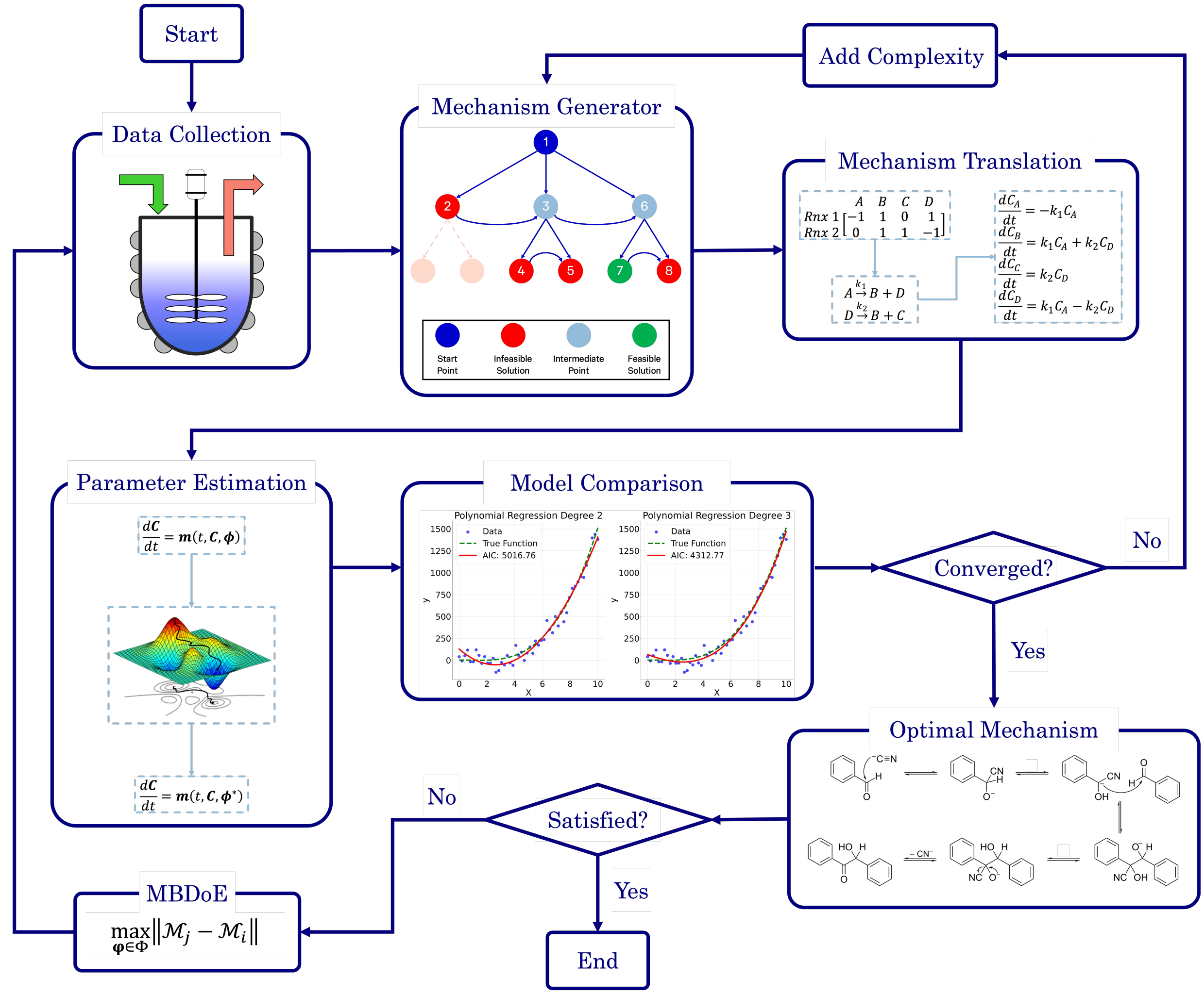

Start
Data Collection
Mechanism Generator
1
2
3
6
4
5
7
8
Start Point
Infeasible Solution
Intermediate Point
Feasible Solution
Add Complexity
Mechanism Translation
A B C D
Rnx 1 [−1 1 0 1]
Rnx 2 [0 1 1 −1]
$A \xrightarrow{k_1} B + D$
$D \xrightarrow{k_2} B + C$
$\frac{dC_A}{dt} = -k_1 C_A$
$\frac{dC_B}{dt} = k_1 C_A + k_2 C_D$
$\frac{dC_C}{dt} = k_2 C_D$
$\frac{dC_D}{dt} = k_1 C_A - k_2 C_D$
Parameter Estimation
$\frac{dC}{dt} = \boldsymbol{m}(t, \boldsymbol{C}, \boldsymbol{\phi})$
$\frac{dC}{dt} = \boldsymbol{m}(t, \boldsymbol{C}, \boldsymbol{\phi}^*)$
Model Comparison
Polynomial Regression Degree 2
Polynomial Regression Degree 3
Data
True Function
AIC: 5016.76
AIC: 4312.77
1500
1250
1000
750
500
250
0
y
0
2
4
6
8
10
x
Converged?
No
Yes
Optimal Mechanism
C≡N
CN
H
O
O⁻
OH
HO H
− CN⁻
NC O⁻
O⁻ H
NC OH
Satisfied?
No
Yes
MBDoE
$\max_{\varphi \in \Phi} \|\mathcal{M}_j - \mathcal{M}_i\|$
End

Figure 1: The workflow of the SiMBA methodology.

The algorithm utilizes matrix representations to model molecular transformations, where each matrix corresponds to a potential reaction mechanism (i.e., each row accounts for an elementary step and each column accounts for a chemical species). This formalism allows the algorithm to handle complex molecular interactions in a structured manner, making it easier to apply checks and balances on the proposed mechanisms.

SiMBA is designed to ensure that only chemically sensible and stoichiometrically balanced reactions are proposed. It does so by adhering to specific rules, which will be elaborated on later in this Section. These checks are crucial in maintaining the physical plausibility of the generated mechanisms. But before initiating the mechanism generation process, several key inputs are required:

- Number of elementary reactions: this input defines the smallest possible number of reactions, or elementary steps, that could lead to a feasible microkinetic model. These elementary steps are constrained by physical principles, such as the requirement that reactions typically involve at **most** two molecules (bimolecular interactions) and usually produce a **maximum** of two product molecules (i.e., four possible elementary reactions: (i) $A \rightarrow B$, (ii) $A + B \rightarrow C$, (iii) $A \rightarrow B + C$, and (iv) $A + B \rightarrow C + D$). This consideration significantly reduces the complexity of the potential mechanisms and aligns the generated reactions with known ones.
- Number of chemical species: this input specifies the minimum (i.e., lower limit) number of chemical species needed to form the smallest possible mechanism. The number of species is critical because it defines the scope of the mechanism generation, ensuring that all necessary reactants, products, and intermediates are considered. This parameter also helps in maintaining the balance between complexity and feasibility in the

View Article Online
DOI: 10.1039/D5SC01473E

proposed mechanisms. We emphasize that these reactant/product limits are defaults chosen for the present case studies; the SiMBA code allows these bounds to be increased trivially. If multi-molecular complexes or fast pre-equilibria are suspected, the user may simply raise the maximum number of reactants or products per step, and SiMBA will enumerate and analyze the resulting higher-order elementary reactions without further modification.

- Stoichiometry: the stoichiometry input dictates the overall chemical reaction being analyzed. It specifies the roles of different species in the reaction, with negative numbers representing reactants, positive numbers representing products, and zeros indicating intermediates (species that do not appear in the overall reaction). This ensures that the generated mechanisms adhere to the correct chemical balance and respect the conservation of mass.
- Time budget: the time budget defines the amount of computational time allocated to the mechanism generator algorithm for exploring the search space and identifying physically feasible reaction mechanisms at each iteration. This constraint helps in managing computational resources effectively and ensures that the generation process is both thorough and time-efficient.

In principle, SiMBA could operate with only the stoichiometry of the reaction provided by the user. From the stoichiometric information alone, the smallest feasible mechanism – defined by the minimum number of elementary reactions and chemical species – could technically be derived automatically. However, the current version of SiMBA deliberately retains the options for users to input these parameters explicitly. This design choice stems from recognizing that users may possess valuable prior knowledge or estimates about their reaction systems, including a realistic minimal number of elementary reactions or chemical species. Starting from an informed position can greatly enhance efficiency, allowing SiMBA to focus computational resources effectively. Thus, while reducing inputs is feasible, maintaining these user-defined inputs provides critical flexibility and practical advantages in guiding the mechanism discovery process.

Once the inputs are defined, we can allow the backtracking algorithm to explore the mechanistic possibilities. The backtracking algorithm – a branch-and-prune method within the field of constrained optimization – is used to systematically explore the vast search space of possible reaction mechanisms by incrementally building potential solutions and backtracking when a solution is found to be infeasible. In the context of mechanism generation, this algorithm starts with an empty matrix representation (mechanism) and progressively starts filling the matrix with possible numbers, ensuring at each step that the proposed mechanism adheres to physical and chemical constraints, such as stoichiometry and the limits on the number of reactants and products in each step. When the algorithm encounters a dead end, where a proposed mechanism violates any of the predefined constraints, it backtracks to the previous step and tries an alternative pathway. This process allows the algorithm to efficiently prune the search space, focusing only on chemically valid and feasible mechanisms, thereby avoiding the exhaustive and brute-force approach through enumeration of all possibilities. This is of particular importance because of the combinatorial nature of the problem. For example, for a small 4x5 matrix, there are 95,367,431,640,625 possible combinations assuming that $x_{i,j} \in \{-2, -1, 0, 1, 2\}$, where a brute-force approach would be intractable. Thus, employing smart methods for an efficient exploration of the space is paramount, even when dealing with small problems. For a more detailed discussion on backtracking, the interested reader should refer to Chapter 2 of Erikson's book [40]. Figure 2 gives an illustrative example of how the backtracking algorithm works.

To further improve the efficiency of the exploration of the possible reaction pathways, we employ a parallelized version of the backtracking algorithm. This allows multiple 'trees' or potential mechanisms to be explored simultaneously, significantly accelerating the search process. The degree of parallelization is primarily constrained by the number of available processors, making this approach highly scalable with increased computational power. Parallelizing the backtracking algorithm offers significant benefits in terms of computational efficiency and scalability. By exploring multiple potential pathways concurrently, the algorithm can cover a much larger portion of the search space within the same amount of time, making it feasible to generate comprehensive sets of candidate mechanisms even for complex reactions.

The algorithm includes several rules to ensure that the generated mechanisms are chemically plausible, to name a few:

- Stoichiometric consistency: the proposed mechanisms must adhere to the stoichiometry defined by the input, ensuring that the overall reaction remains balanced.
- Elementary step constraints: each elementary step is restricted to having at most two reactants and two products, with at least one of each, reflecting the typical nature of an elementary step and ensuring that there is no redundant elementary steps (i.e., a row full of zeros).
- Intermediate formation: intermediates must be generated in the reaction network before they are consumed, maintaining a logical and sequential flow of the reaction mechanism.

View Article Online
DOI: 10.1039/D5SC01473E

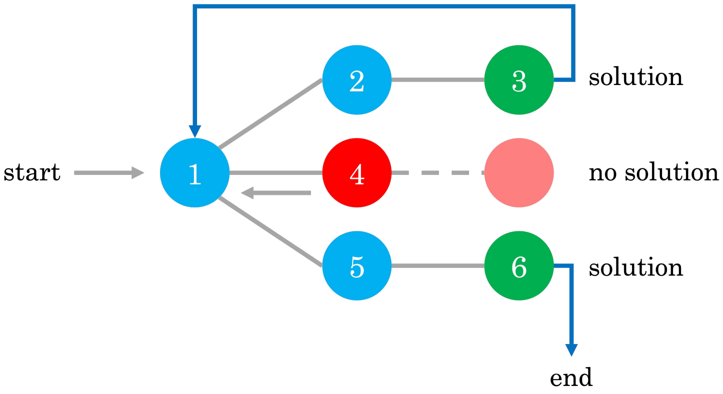


Figure 2: Example of a backtracking algorithm flowchart, where the algorithm explores potential pathways from node 1 by systematically advancing to connected nodes (2, 4, 5) while evaluating constraints. If a path fails to meet criteria (e.g., reaching a red node), the algorithm ”backtracks” to the previous node, exploring alternative paths until a viable solution path is found (ending at a green node).

These rules allow for SiMBA to filter out unfeasible or non-physical mechanisms, ensuring that the outputs are not only mathematically valid but also chemically meaningful. Thus, at the end of this phase, a comprehensive set of candidate reaction mechanisms is generated, each represented by a matrix and each having the same level of complexity (i.e., same number of elementary steps and chemical species). These matrices serve as the basis for further analysis in subsequent phases of the SiMBA methodology. This phase lays the groundwork for the rest of the SiMBA methodology, providing a robust and physically plausible set of reaction mechanisms that will be refined, validated and compared in the subsequent phases.

A noteworthy scenario arises when the overall chemical reaction is incompletely characterized due to unknown side products, such as when significant yield losses occur through mechanisms like coking or volatile byproduct formation. In such cases, SiMBA may encounter difficulty in proposing chemically meaningful reaction mechanisms if kinetic data from these unknown pathways are unavailable. However, two practical workarounds could partially mitigate this challenge. First, one could adjust the kinetic data for reactants by subtracting the fraction lost to side reactions, thereby defining an ”effective reactant” profile. This manipulation would focus SiMBA’s mechanism generation exclusively on the known reaction pathway of interest. Alternatively, one could introduce ”pseudo-side product” variables to represent all mass lost through unidentified side reactions, thus preserving mass balance. SiMBA would then propose mechanisms involving both the target reaction and a generalized pathway to the pseudo-side products. Although helpful, this second method implicitly assumes an arbitrary number of side pathways, potentially oversimplifying the actual chemical processes involved. Consequently, both approaches have inherent limitations and should be applied with careful consideration of the reaction system under study.

It is important to note, however, that despite SiMBA’s systematic approach and rigorous filtering criteria, the mechanisms returned are fundamentally justified by the kinetic data rather than guaranteed to represent the actual underlying chemical pathways. The inherent limitation here is that concentration-time datasets, especially when incomplete, inherently underconstrain the reaction network. Consequently, SiMBA-generated mechanisms should be interpreted as being consistent with the available kinetic data and chemically plausible within the defined constraints (which can always be augmented), but not necessarily as uniquely true representations of the fundamental reaction mechanisms. This caveat is particularly critical when applying SiMBA to real-world experimental datasets, which may be incomplete or subject to measurement uncertainties. Users should therefore view the generated mechanisms as robust hypotheses warranting further experimental verification and refinement.

2.2 Mechanism Translation

The purpose of this step of the SiMBA algorithm is to convert matrix representations of mechanisms into executable models that can be used for parameter estimation, simulation, and analysis. The translation process consists of two main steps: first, converting the matrix representation into reaction strings, and second, converting these reaction strings into systems of ODEs which are executable in Python.

View Article Online
DOI: 10.1039/D5SC01473E

In the matrix representation, each row corresponds to an elementary reaction, while each column represents a chemical species. The elements within the matrix indicate stoichiometric coefficients: negative for reactants, positive for products, and zero for species not involved in the elementary step.

In the first step, each reaction string is generated by identifying the reactants, products and the stoichiometric coefficients for every row of a given matrix. For example, the below matrix would be converted into the following reaction strings, where A and B are reactants, C is a product and D is an intermediate:

$$\begin{bmatrix} -1 & -1 & 0 & 1 \\ 0 & -1 & 1 & -1 \end{bmatrix} \left.\begin{matrix} A + B \xrightarrow{k_1} D \\ B + D \xrightarrow{k_2} C \end{matrix}\right\} \tag{1}$$

Using mass-action kinetics, the reaction strings are then converted into ODEs. The rate of a reaction is proportional to the product of the concentrations of the reactants. For example, for the reaction string $A + B \xrightarrow{k_1} D$ the rate equation is expressed as $r = k_1 C_A C_B$, where k_1 is the rate constant. Our code automates the translation of these reaction strings into ODEs by systematically identifying unique species, constructing rate equations, and assembling the differential equations into a comprehensive kinetic model. For instance, the system of ODEs for the above matrix representation is:

$$\begin{matrix} A + B \xrightarrow{k_1} D \\ B + D \xrightarrow{k_2} C \end{matrix} \left\{\begin{aligned} \frac{dC_A}{dt} &= -k_1 C_A C_B \\ \frac{dC_B}{dt} &= -k_1 C_A C_B - k_2 C_B C_D \\ \frac{dC_C}{dt} &= k_2 C_B C_D \\ \frac{dC_D}{dt} &= k_1 C_A C_B - k_2 C_B C_D \end{aligned}\right. \tag{2}$$

The translation process is automated using a Python script, primarily leveraging the regex library [41]. Regex is employed to parse and extract key components from reaction strings, such as chemical species and reaction operators. Specifically, regex identifies species by matching patterns of letters and helps distinguish between different parts of the reaction strings, including reactants, products, and the reaction arrow (→). This parsing ensures accurate separation and interpretation of reactants and products, facilitating the automated construction of corresponding rate equations in the ODE system.

Given the potential to generate a vast number of models in any iteration of SiMBA, automating the translation from matrix notation to executable Python functions is crucial. Manual conversion would be impractical, if not impossible, due to the sheer volume of candidate mechanisms. Therefore, this automated approach not only improves efficiency but also ensures that the subsequent phases of the SiMBA algorithm can proceed smoothly. Much of the code for the translation of reaction strings to systems of ODEs has been adapted from the work of Jiscoot et. al. [42].

2.3 Parameter Estimation

The objective of this aspect of SiMBA is to determine the kinetic parameters that best fit the generated models to the available data. This step is crucial for ensuring that the proposed reaction mechanisms are optimized so that they reflect, as accurately as possible, the observed dynamics of the chemical system. This will enable the algorithm to compare different models fairly in the next phase. The parameter estimation step is a standard procedure in model building frameworks.

To solve the parameter estimation problem, we use simulated concentration-time profiles as the dataset. These profiles provide time-series data of species' concentrations, which are critical for fitting the kinetic models.

The parameter estimation problem is defined in Eq. (3), where $\hat{y}_m^{(i)}$ denote the prediction of a value coming from a proposed model m at a given time $t^{(i)}$ (i.e., $\hat{y}_m^{(i)} = m(t^{(i)} \mid \theta_m)$), and $y^{(i)}$ represents the target value at a given time $t^{(i)}$ (i.e., in-silico data, in this study). Furthermore, SSE represents the sum of squared errors and n_t is defined as the sampling times, which are set within the fixed time interval, $t^{(i)} \in \Delta t$ where $\Delta t = [t_0, t_f]$.

$$\theta_m^* = \arg\min_{\theta} \sum_{i=1}^{n_t} SSE\left(\hat{y}_m^{(i)}, y^{(i)}\right). \tag{3}$$

View Article Online
DOI: 10.1039/D5SC01473E

The Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm is employed for solving the parameter estimation problem [43]. L-BFGS is well-suited for handling this problem due to its performance in tasks pertaining to parameter estimation and optimization [44, 43].

To ensure a thorough exploration of the parameter space, we may use expert-informed or random initial guesses for the parameters, with bounds set within the range $[0, 10]$ to maintain physically meaningful values in the chosen case studies (this can be changed on as-needed basis). The stopping criteria for the optimization are left to the default options in the Scipy package [45], and a multi-start approach is employed, where multiple runs are initiated with different starting points, and the best solution is retained.

2.4 Model Comparison

SiMBA uses an information criterion approach for model selection rather than a data-splitting approach, enabling the entire dataset to be used for model construction while still providing a robust and reliable method for testing the proposed models. This is especially advantageous in low-data scenarios, as it ensures that we make full use of the available information when identifying suitable microkinetic models.

We use the Akaike Information Criterion (AIC) because in previous work we compared various information criteria to determine if any offered superior performance. We found that AIC consistently outperformed other criteria in the context of kinetic discovery; further details of these studies can be found in de Carvalho Servia et. al. [38]. However, SiMBA's architecture is entirely agnostic to the particular metric used for selecting the best mechanism. Should a user prefer a classical train/test error, cross-validation score, Bayes factor, or any other statistic instead of AIC, they can simply swap in their desired metric via the "metrics.py" module without altering the core enumeration, translation, or parameter-estimation workflows. Users can thus tailor model selection to their data and domain requirements.

Given a model m with parameters θ_m of dimension d_m, the AIC is defined as:

$$\text{AIC}_m = 2NLL(\theta_m \mid \mathcal{D}) + 2d_m, \tag{4}$$

where NLL represents specifically the negative log-likelihood [46]. Given two competing models, m_1 and m_2, the preferred model would be the one with the lowest AIC value calculated by Eq. (4).

If iteration $n + 1$ in the SiMBA algorithm results in an improvement of the AIC value compared to iteration n, SiMBA will continue running, further refining the model output by considering more complicated mechanisms. This approach ensures that the algorithm is consistently moving towards a model that better balances complexity with goodness of fit. However, if the best model in iteration $n + 1$ displays a worsening in the AIC value compared to the best model in iteration n, indicating that the model has become less optimal, SiMBA will terminate the process. In this scenario, the algorithm concludes that additional iterations are implausible to produce a superior model, and it will return the best solution found during iteration n, which is considered the most accurate and parsimonious model according to the AIC evaluation.

2.5 Model-Based Design of Experiments

If the dataset used for mechanism discovery is insufficient to yield an adequate model, and provided that the experimental budget has not been exhausted, we can use the insights from the optimized models to design a more informative experiment. Specifically, we can identify operating conditions that maximize the difference between the state predictions $\mathbf{X}$ of the two best proposed models, ν and μ based on the existing dataset. The rationale behind using the two best proposed models is discussed in de Carvalho Servia et. al. [47]. The MBDoE approach adopted in this work was developed by Hunter and Reiner [39]:

$$\mathbf{x}_{k+1} = \arg\max_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^{T}\sum_{j=1}^{d}(\mathbf{X}^{\nu}_{i,j} - \mathbf{X}^{\mu}_{i,j})^2 \tag{5}$$

$$\mathbf{X}^{\nu} = \int_{t_0}^{t_f} f^{\nu}(\mathbf{x}, t, \theta^{\nu})\,\mathrm{d}t \tag{6}$$

$$\mathbf{X}^{\mu} = \int_{t_0}^{t_f} f^{\mu}(\mathbf{x}, t, \theta^{\mu})\,\mathrm{d}t \tag{7}$$

View Article Online
DOI: 10.1039/D5SC01473E

In this equation, **x** represents the operating conditions within a set $\mathcal{X}$. Using the identified initial conditions, a new experiment can be conducted to generate additional data points, which are then added to the original dataset. With this updated dataset SiMBA can be executed again, thereby closing the loop between informative experimental design and optimal model discovery.

3 Catalytic Kinetic Case Studies

The purpose of the case studies presented in this work is to serve as proof-of-concept validations for the newly developed methodology, SiMBA. Before deploying this data-driven method in experimental environments or attempting to propose and discover new reaction mechanisms, it is important to ensure that the methodology is both sound and capable of delivering reliable results. To achieve this, we selected case studies where experimentalists have already proposed mechanisms or rate models, allowing us to generate in-silico data through computational simulations and subsequently test SiMBA's ability to accurately rediscover these mechanisms from the generated datasets.

The selection of the case studies was made to demonstrate SiMBA's effectiveness across a range of scenarios. The case studies include a hypothetical reaction, an aldol condensation between benzaldehyde and acetophenone [48], and the dehydration of fructose to 5-hydroxymethylfurfural (HMF) [49, 50]. These studies were chosen to showcase SiMBA's ability to distill complex kinetic behaviors into simple, accurate models.

The hypothetical reaction serves as an initial proof-of-concept, illustrating whether SiMBA can generate microkinetic models purely from fundamental principles without relying on prior knowledge. This first case study also demonstrates SiMBA's versatility in handling both first-order and second-order elementary steps. Next, the aldol condensation – a classic reaction with a well-understood mechanism – tests SiMBA's ability to reconstruct mechanistic pathways solely from dynamic data on main reactants and products. By successfully modeling this reaction, SiMBA shows that it can go beyond hypothetical examples to accurately capture established mechanistic pathways. Finally, the dehydration of fructose to HMF introduces a different challenge: rather than relying on a microkinetic simulation of stoichiometric reactants and products, the in-silico dataset comes from a rate model that has been experimentally validated. In this scenario, SiMBA is challenged to discover a plausible kinetic mechanism, aligning with established literature and demonstrating its capacity to derive robust reaction models from realistic data sources.

Although our case studies focus on homogeneous reaction networks, SiMBA's matrix-and-ODE framework naturally accommodates both homogeneous catalytic cycles and heterogeneous surface chemistry. In a homogeneous catalytic cycle, one treats the catalyst resting state and any activated forms (e.g., Cat–A, Cat–B) as additional species in the matrix. Turnover steps – substrate binding ($A \rightarrow \mathrm{Cat}$–$A$), intramolecular transformation (Cat–$A \rightarrow \mathrm{Cat}$–$B$), and product release ($\mathrm{Cat}$–$B \rightarrow \mathrm{Cat} + P$) – are handled just like any elementary reaction. Likewise, in heterogeneous catalysis, adsorption/desorption steps ($A \rightleftharpoons Z$) and surface–surface reactions ($X + Y \rightarrow Z$) map directly onto bimolecular or unimolecular matrix rows: free species adsorb ($A \rightarrow X$, $B \rightarrow Y$), surface intermediates react ($X + Y \rightarrow Z$), and products desorb ($Z \rightarrow P$). Because every elementary step – whether substrate isomerization, catalyst turnover, or surface adsorption – is represented in the same stoichiometric matrix and translated automatically into ODEs, SiMBA requires no algorithmic changes to discover, fit, and rank mechanisms in either homogeneous or heterogeneous catalytic systems.

3.1 The Hypothetical Reaction

The hypothetical reaction is one that involves five different chemical species – only three of which are observed – interacting in four different elementary steps. The overall stoichiometry of the hypothetical reaction can be represented by Eq. (8) while Eq. (9) provides a description of the mechanism of the reaction as well as the system of ordinary differential equations (ODEs) underpinning the dynamics of the reaction (and directly derived from the proposed mechanism). The kinetic parameters (rate constants) were defined as: $k_1 = 0.1$ M^{-1} h^{-1}, $k_2 = 0.2$ h^{-1}, $k_3 = 0.13$ h^{-1} and $k_4 = 0.25$ M^{-1} h^{-1}. Since this is a purely in silico proof-of-concept, the four rate constants were drawn randomly within a physically reasonable range.

$$4A \rightarrow B + C \tag{8}$$

View Article Online
DOI: 10.1039/D5SC01473E

$$\begin{array}{r} 2A \xrightarrow{k_1} B \\ A \xrightarrow{k_2} D \\ D \xrightarrow{k_3} E \\ A + E \xrightarrow{k_4} C \end{array} \left\{ \begin{aligned} \frac{dC_A}{dt} &= -k_1 C_A^2 - k_2 C_A - k_4 C_A C_E \\ \frac{dC_B}{dt} &= k_1 C_A^2 \\ \frac{dC_C}{dt} &= k_4 C_A C_E \\ \frac{dC_D}{dt} &= k_2 C_A + k_3 C_D \\ \frac{dC_E}{dt} &= k_3 C_D - k_4 C_A C_E \end{aligned} \right. \tag{9}$$

Starting from the ODE system in Eq. 9, an in-silico dataset is generated wherein $\Delta t = [0, 10]$ h and $n_t = 30$. This dataset is composed of five different experiments, each ran at different initial conditions (in molar units: $(C_A(t = 0), C_B(t = 0), C_C(t = 0), C_D(t = 0), C_E(t = 0)) \in \{(10, 0, 2, 0, 0), (10, 2, 0, 0, 0), (10, 2, 2, 0, 0), (5, 0, 0, 0, 0), (10, 0, 0, 0, 0)\}$); these experiments were randomly picked from a 2^k factorial design [51].

For all experiments, the system is assumed to be both isochoric and isothermal, and Gaussian noise is added to the in-silico measurements to simulate a chemical experiment. The added noise had zero mean and a standard deviation of 0.15 for A, B and C. To further approximate a realistic system, we assume that we cannot measure the intermediates D and E. The generated dataset for the one of the experiments are presented in Fig. 4 (a). The dataset, providing 150 datapoints, has a realistic size for kinetic studies [52, 53, 54], especially considering recent advancements in high-throughput setups.

3.1.1 Results and Discussions

The application of SiMBA to the hypothetical reaction case study successfully demonstrated its ability to recover the underlying microkinetic model with limited data and only access to the species present in the overall reaction (i.e., without direct data on intermediates). This provides initial validation of the algorithm's capacity to propose accurate and physically sensible models under constrained conditions, highlighting its potential for broader application in more complex chemical systems, which will be shown in subsequent subsections.

In this case, given the stoichiometry of the hypothetical reaction, shown in Eq. (8), the simplest possible mechanism involves two elementary steps. This is due to the fact that termolecular, and higher order interactions, are relatively rare – for the purpose of SiMBA, we consider them as impossible – as the simultaneous collision of three or more molecules in the correct orientation is a very unlikely occurrence. Thus, based on that constraint, we would need at least two elementary steps to react four moles of A. As such, in the first step, two moles of species A react to produce one mole of B, while in the second step, two moles of A react to produce one mole of C. Notably, the order in which B and C are produced is interchangeable, without affecting the model's performance. These two configurations represent the only physically feasible mechanisms that could be formed in the first iteration of SiMBA, based on a 2x3 matrix (representing two elementary steps and three species).

Upon identifying all possible permutations of mechanisms represented by this 2x3 matrix (in this case, two permutations), SiMBA translated them into ordinary differential equation (ODE) models that could be optimized. Through parameter estimation, we optimize the kinetic parameters of the model, which enable us to calculate the AIC values for each model and selected the best-performing one. Table 1 shows the optimal mechanism discovered in the first iteration, including the corresponding microkinetic model, and the AIC value which amounted to 1110.34. Figure 6 presents the model's fit against an arbitrary training experiment, visually illustrating its accuracy.

Following this initial step, SiMBA automatically proceeded to iteration 2, which leads to an increase of complexity by allowing an extra elementary step and an extra intermediate to be present in the modeling task. Consequently, in this iteration, the algorithm began with an empty 3x4 matrix representation of the reaction mechanism. Using the same process as in iteration 1, we identified and optimized the potential mechanisms for this iteration. To echo the point made in Section 2 regarding the importance of smart explorative methods, iteration 2 could generate 244,140,625 different matrix configurations; with backtracking, we only check the 31 configurations that make physical sense (0.00001% of all possibilities). The AIC value was again used to select the best model for iteration 2, which now amounted to an improvement to 106.28. A decrease in AIC from 1110.34 to 61.67 indicates a substantial improvement in model accuracy while maintaining parsimony.

View Article Online
DOI: 10.1039/D5SC01473E

SiMBA converged in four iterations, at which point the termination criterion was met, meaning that the informationally optimal mechanism was discovered in iteration 3. At iteration 4, no further reduction in AIC was observed, which met the termination criterion for model refinement, indicating that the added complexity did not yield better predictive accuracy. Table 1 summarizes the best mechanism identified in each iteration, along with the corresponding microkinetic models and AIC values. Figure 6 illustrates the model fit for each selected mechanism against an arbitrary training experiment, further demonstrating the progressive refinement of the models across iterations (with exception of the last one).

In iteration 1, the 2×3 matrix yields $5^6 = 15,625$ possible combinations, of which SiMBA identifies and evaluates all feasible candidates in just 3.40 s. In iteration 2, the search space jumps to $5^{12} \approx 2.44 \times 10^8$ matrices, yet SiMBA finds and evaluates all feasible mechanism arrangements in 14.91 s by pruning infeasible branches early. By iteration 3 ($5^{20} \approx 9.54 \times 10^{13}$, 1,110.34 s) and iteration 4 ($5^{30} \approx 9.31 \times 10^{20}$, 5,911.28 s), the exponential growth in possibilities becomes evident – even though the number of feasible matrices remains a tiny fraction and the execution of SiMBA stays tractable.

A comparison of the final selected model against the data-generating model indicates that SiMBA successfully uncovered the underlying mechanism driving the hypothetical reaction. This case study serves as a proof-of-concept, showcasing SiMBA's ability to generate accurate microkinetic models even in the absence of direct data on intermediates and with limited in-silico data. The results demonstrate the robustness of SiMBA in handling systems that feature both first- and second-order elementary steps, confirming its potential for more complex chemical systems and broader industrial applications, as demonstrated in the next subsections.

View Article Online
DOI: 10.1039/D5SC01473E

Table 1: Reaction mechanisms, corresponding microkinetic models, and AIC values for iterations 1 through 4. The table presents the reaction mechanisms discovered at each iteration, alongside their respective microkinetic models and AIC values. Each iteration reflects an increase in the complexity of the reaction mechanism. In iteration 1, the simplest model consists of two elementary steps involving species A, B, and C, with an AIC value of 1139.86. By iteration 2, an additional intermediate (D) is introduced, lowering the AIC to 106.28. The optimal mechanism (in this case, identical to the data-generating one), discovered in iteration 3, involves the introduction of an additional intermediate (E), achieving the lowest AIC of -317.99. Iteration 4 introduces yet another intermediate (F), but results in a higher AIC value of 390.92, indicating that iteration 3 provides the best balance between accuracy and complexity.

Iteration	Reaction Mechanism	Microkinetic Model	AIC Value
1	$2A \xrightarrow{k_1} B$ $2A \xrightarrow{k_2} C$	$\frac{dC_A}{dt} = -k_1 C_A^2 - k_2 C_A^2$ $\frac{dC_B}{dt} = k_1 C_A^2$ $\frac{dC_C}{dt} = k_2 C_A^2$	1139.86
2	$2A \xrightarrow{k_1} D$ $2A \xrightarrow{k_2} B$ $D \xrightarrow{k_3} C$	$\frac{dC_A}{dt} = -k_1 C_A^2 - k_2 C_A^2$ $\frac{dC_B}{dt} = k_2 C_A^2$ $\frac{dC_C}{dt} = k_3 C_D$ $\frac{dC_D}{dt} = -k_3 C_D$	106.28
3	$2A \xrightarrow{k_1} B$ $A \xrightarrow{k_2} D$ $D \xrightarrow{k_3} E$ $A + E \xrightarrow{k_4} C$	$\frac{dC_A}{dt} = -k_1 C_A^2 - k_2 C_A - k_4 C_A C_E$ $\frac{dC_B}{dt} = k_1 C_A^2$ $\frac{dC_C}{dt} = k_4 C_A C_E$ $\frac{dC_D}{dt} = k_2 C_A + k_3 C_D$ $\frac{dC_E}{dt} = k_3 C_D - k_4 C_A C_E$	-317.99
4	$2A \xrightarrow{k_1} B$ $2A \xrightarrow{k_2} D$ $D \xrightarrow{k_3} 2E$ $E \xrightarrow{k_4} F$ $E + F \xrightarrow{k_5} C$	$\frac{dC_A}{dt} = -k_1 C_A^2 - k_2 C_A^2$ $\frac{dC_B}{dt} = k_1 C_A^2$ $\frac{dC_C}{dt} = k_5 C_E C_F$ $\frac{dC_D}{dt} = k_2 C_A^2 - k_3 C_D$ $\frac{dC_E}{dt} = k_3 C_D - k_4 C_E - k_5 C_E C_F$ $\frac{dC_F}{dt} = k_4 C_E - k_5 C_E C_F$	390.92

View Article Online
DOI: 10.1039/D5SC01473E

3.2 The Aldol Condensation Reaction

An aldol condensation is a type of condensation reaction in organic chemistry between a ketone and an aldehyde to form a carbon-carbon double bond in the enone product, eliminating a molecule of water. The mechanism of the reaction is initiated with the formation of an enol or enolate intermediate from a ketone. This nucleophilic intermediate attacks the carbonyl group of the aldehyde, to form a β-hydroxyaldehyde or β-hydroxyketone, which in turn dehydrates to produce a conjugated enone. The aldol condensation therefore involves six different chemical species – only four of which are observed – interacting in three different elementary steps. Fig. 3 represents the overall reaction as well as the detailed mechanism of the aldol condensation reaction between acetophenone and benzaldehyde.

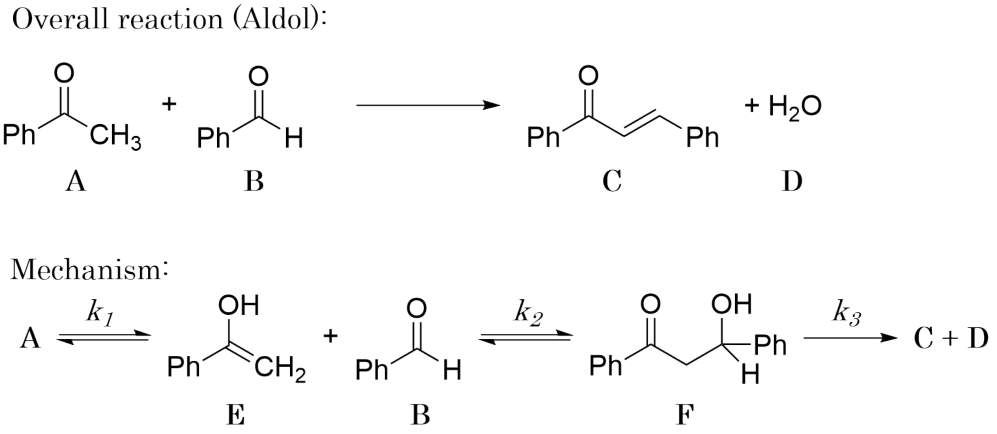


Figure 3: Schematic representation of the aldol condensation reaction between acetophenone (A) and benzaldehyde (B) to form the chalcone product (C) and water (D). The mechanism proceeds in three main steps: (i) enolization of A to give the enolate/enol intermediate (E), (ii) nucleophilic addition of E to B to form the β-hydroxy adduct (F), and (iii) dehydration to yield the final conjugated enone (C). Rate constants k_1, k_2, and k_3 are associated with each step. Phenyl groups are represented by "Ph."

Eq. (10) provides a simplified description of the mechanism of the reaction as well as the ODE system underpinning the dynamics of the reaction (which was directly derived from the proposed mechanism) [48]. It is worth noting that we assume that the elementary steps are all irreversible. The kinetic parameters (rate constants) were defined as: $k_1 = 0.759\ \mathrm{h}^{-1}$, $k_2 = 0.293\ \mathrm{M}^{-1}\ \mathrm{h}^{-1}$ and $k_3 = 0.681\ \mathrm{h}^{-1}$. Although the sequence of elementary steps is taken from literature, we likewise randomized the three rate constants within [0.1, 1], imposing only the constraint that the C–C bond-forming step (i.e., second step) is the slowest, consistent with mechanistic studies showing carbon–carbon coupling as rate-determining [48].

$$\begin{array}{c} A \xrightarrow{k_1} E \\ E + B \xrightarrow{k_2} F \\ F \xrightarrow{k_3} C + D \end{array} \left\{ \begin{aligned} \frac{dC_A}{dt} &= -k_1 C_A \\ \frac{dC_B}{dt} &= -k_2 C_E C_B \\ \frac{dC_C}{dt} &= k_3 C_F \\ \frac{dC_D}{dt} &= k_3 C_F \\ \frac{dC_E}{dt} &= k_1 C_A - k_2 C_E C_B \\ \frac{dC_F}{dt} &= k_2 C_E C_B - k_3 C_F \end{aligned} \right. \tag{10}$$

In Eq. (10), A, B, C, D, E and F correspond to acetophenone, benzaldehyde, chalcone, water, α-phenylvinyl enolate and 4-hydroxy-1,3-diphenylbutan-1-one, respectively.

Starting from the ODE system in Eq. 10, an in-silico dataset is generated wherein $\Delta t = [0, 10]$ h and $n_t = 30$. This dataset is composed of five different experiments, each ran at different initial condi-

View Article Online
DOI: 10.1039/D5SC01473E

tions (in molar units: $(C_A(t=0), C_B(t=0), C_C(t=0), C_D(t=0), C_E(t=0), C_F(t=0)) \in \{(5,10,0,0,0,0),(5,5,2,0,0,0),(5,10,0,2,0,0),(10,10,0,2,0,0),(10,10,2,2,0,0)\}$); these experiments were randomly picked from a 2^k factorial design [51].

For all experiments, the system is assumed to be both isochoric and isothermal, and Gaussian noise is added to the in-silico measurements to simulate a chemical experiment. The added noise had zero mean and a standard deviation of 0.15 for A, B, C and D. To further approximate a realistic system, we assume that we cannot measure the intermediates E and F. The generated dataset for one of the experiments are presented in Fig. 4 (b).

3.2.1 Results and Discussions

The aldol condensation reaction provided a different challenge for SiMBA due to the complexity of the overall system, given that there is a higher number of chemical species involved. Despite the absence of direct data on intermediates, SiMBA demonstrated its capability to infer a reliable microkinetic model, even in this constrained setting, which matches perfectly to the data-generating model.

In the first iteration, based on the stoichiometric relationship between the reactants and products, SiMBA identified a single elementary step where one mole of acetophenone (A) reacted with one mole of benzaldehyde (B) to produce chalcone (C) and water (D). The simplicity of this mechanism, represented by a 1x4 matrix (one step involving four species), aligned with the overall stoichiometry of the aldol condensation reaction and was deemed the only feasible configuration for this initial phase.

Following this, SiMBA translated the 1x4 matrix into a set of ODEs that could be optimized computationally. Parameter estimation was performed, and the AIC value was calculated and used to gauge the model's fit. Table 2 does not present the mechanism identified during this initial iteration for spacing reasons, but Figure 6 shows how well the model predictions aligned with experimental data from a selected training experiment.

Upon completion of iteration 1, SiMBA advanced to iteration 2, where an additional elementary step and species were incorporated. This expanded the search space to a 2x5 matrix, increasing the complexity of the possible mechanisms. As in the previous step, the new sets of ODEs were optimized, and AIC values were calculated. The results from iteration 2 showed a notable improvement, as the added complexity contributed to a better overall fit, without overfitting the system. Table 2 and Figure 6 detail the refined mechanism and its improved accuracy.

The iterative process continued, and SiMBA reached its optimal solution at iteration 3, where no further reduction in the AIC value was observed in subsequent iterations. The complexity added in iteration 4 did not yield a better AIC value, signaling that the best mechanism had already been identified, since further complexity was improving the fit negligibly. The termination criterion was therefore met after iteration 4, confirming that iteration 3 provided the most accurate and parsimonious model. The results from the second, third and fourth iterations are summarized in Table 2, while Figure 6 illustrates the fit for every iteration.

A similar pattern to that of the first case study emerges for the aldol case regarding the execution of SiMBA: iteration 1 ($5^4 = 625$ total possible matrices) executes in 3.06 s; iteration 2 ($5^{10} \approx 9.77 \times 10^6$) in 4.41 s; iteration 3 ($5^{18} \approx 3.81 \times 10^{12}$) in 75.18 s; and iteration 4 ($5^{28} \approx 3.73 \times 10^{19}$) in 7,989.69 s. These results confirm that backtracking keeps the search tractable even as the combinatorial possibilities skyrocket.

The comparison between the selected model and the original data-generating mechanism demonstrates SiMBA's ability to successfully uncover the fundamental dynamics of the aldol condensation reaction, even when working with limited data. This case study serves as further evidence of SiMBA's strength in identifying complex reaction mechanisms in realistic systems. The results not only validate SiMBA's accuracy but also highlight its potential for broader application in mechanistic discovery across diverse chemical reactions.

View Article Online
DOI: 10.1039/D5SC01473E

Table 2: Reaction mechanisms, microkinetic models, and AIC values for iterations 2 to 4 identified by SiMBA in the aldol condensation case study. Due to space constraints, iteration 1 is omitted from the table but its performance is shown in Figure 6. The table presents the reaction mechanisms for iterations 2 through 4, the corresponding microkinetic models in the form of ordinary differential equations, and the AIC values, which assess model quality. Iteration 2 begins with the introduction of intermediate E, resulting in a substantial improvement in the AIC value to 866.58. By iteration 3, the inclusion of intermediate F yields the optimal model with the lowest AIC value of -351.17, indicating the best balance between fit and parsimony. In iteration 4, an additional intermediate (G) is introduced, but the AIC value of -349.21 indicates that the added complexity is unnecessary, suggesting that the optimal model was discovered in iteration 3.

Iteration	Reaction Mechanism	Microkinetic Model	AIC Value
2	$A \xrightarrow{k_1} E$ $B + E \xrightarrow{k_2} C + D$	$\frac{dC_A}{dt} = -k_1 C_A$ $\frac{dC_B}{dt} = -k_2 C_B C_E$ $\frac{dC_C}{dt} = k_2 C_B C_E$ $\frac{dC_D}{dt} = k_2 C_B C_E$ $\frac{dC_E}{dt} = k_1 C_A - k_2 C_B C_E$	866.58
3	$A \xrightarrow{k_1} E$ $B + E \xrightarrow{k_2} F$ $F \xrightarrow{k_3} C + D$	$\frac{dC_A}{dt} = -k_1 C_A$ $\frac{dC_B}{dt} = -k_2 C_B C_E$ $\frac{dC_C}{dt} = k_3 C_F$ $\frac{dC_D}{dt} = k_3 C_F$ $\frac{dC_E}{dt} = k_1 C_A - k_2 C_B C_E$ $\frac{dC_F}{dt} = k_2 C_B C_E - k_3 C_F$	-351.17
4	$A \xrightarrow{k_1} E$ $E + B \xrightarrow{k_2} F + G$ $F \xrightarrow{k_3} C$ $G \xrightarrow{k_4} D$	$\frac{dC_A}{dt} = -k_1 C_A$ $\frac{dC_B}{dt} = -k_2 C_B C_E$ $\frac{dC_C}{dt} = k_3 C_F$ $\frac{dC_D}{dt} = k_4 C_G$ $\frac{dC_E}{dt} = k_1 C_A - k_2 C_B C_E$ $\frac{dC_F}{dt} = k_2 C_B C_E - k_3 C_F$ $\frac{dC_G}{dt} = k_2 C_B C_E - k_4 C_G$	-349.21

View Article Online
DOI: 10.1039/D5SC01473E

3.3 The Dehydration of Fructose

The dehydration of fructose refers to the process of removing water molecules from fructose to produce 5-hydroxymethylfurfural (HMF), a valuable platform chemical. This reaction is important because HMF can be further converted into various high-value chemicals and biofuels, making it a crucial step in the conversion of biomass into renewable energy and materials. The overall stoichiometry of the dehydration reaction can be represented by Eq. (11) whilst Eq. (12) shows the rate model extracted from the literature [49], which is derived from experimental data and governs the reaction dynamics. A brief note on assumptions: the energy balance is not included in this study, instead treating the reaction as if it proceeds isothermally at 137 °C. This choice reflects the experimental setup – heating a 0.5 mL reaction mixture in sealed glass ampoules – where the small volume and thin walls likely minimize heat-up time and heat transfer limitations. Under these conditions, the parameters used were: $C_{acid} = 3.3 \times 10^{-2}$ M of sulfuric acid, $k_{ref} = 0.9$ M^{-1} min^{-1}, $E_a = 124$ J mol^{-1}, $R = 8.314$ J K^{-1} mol^{-1} and $T = 410.15$ K (directly taken from van Putten et. al. [49], which alike the model, were derived from experimental data).

$$A \rightarrow 3B + C \tag{11}$$

$$r = kC_A C_{acid} \tag{12a}$$

$$k = k_{ref} \exp\left(-\frac{E_a}{RT}\right) \tag{12b}$$

In Eq. (11), A, B and C correspond to fructose, water and HMF respectively. Starting from the rate model in Eq. 12, we can derive an ODE system ($r = \frac{1}{\nu_i}\frac{dC_i}{dt}$) and generate an in-silico dataset wherein $\Delta t = [0, 90]$ min and $n_t = 30$. This dataset is composed of five different experiments, each ran at different initial conditions (in molar units: $(C_A(t=0), C_B(t=0), C_C(t=0)) \in \{(4,0,0), (6,2,1), (4,2,0), (4,0,1), (6,2,0)\}$); these experiments were randomly picked from a 2^k factorial design [51].

For all experiments, the system is assumed to be both isochoric and isothermal, and Gaussian noise is added to the in-silico measurements to simulate a chemical experiment. The added noise had zero mean and a standard deviation of 0.2 for A, B and C. In this example, resembling a real system, we do not have any measurement on possible intermediates. The generated dataset for the one of the experiments are presented in Fig. 4 (c).

3.3.1 Results and Discussions

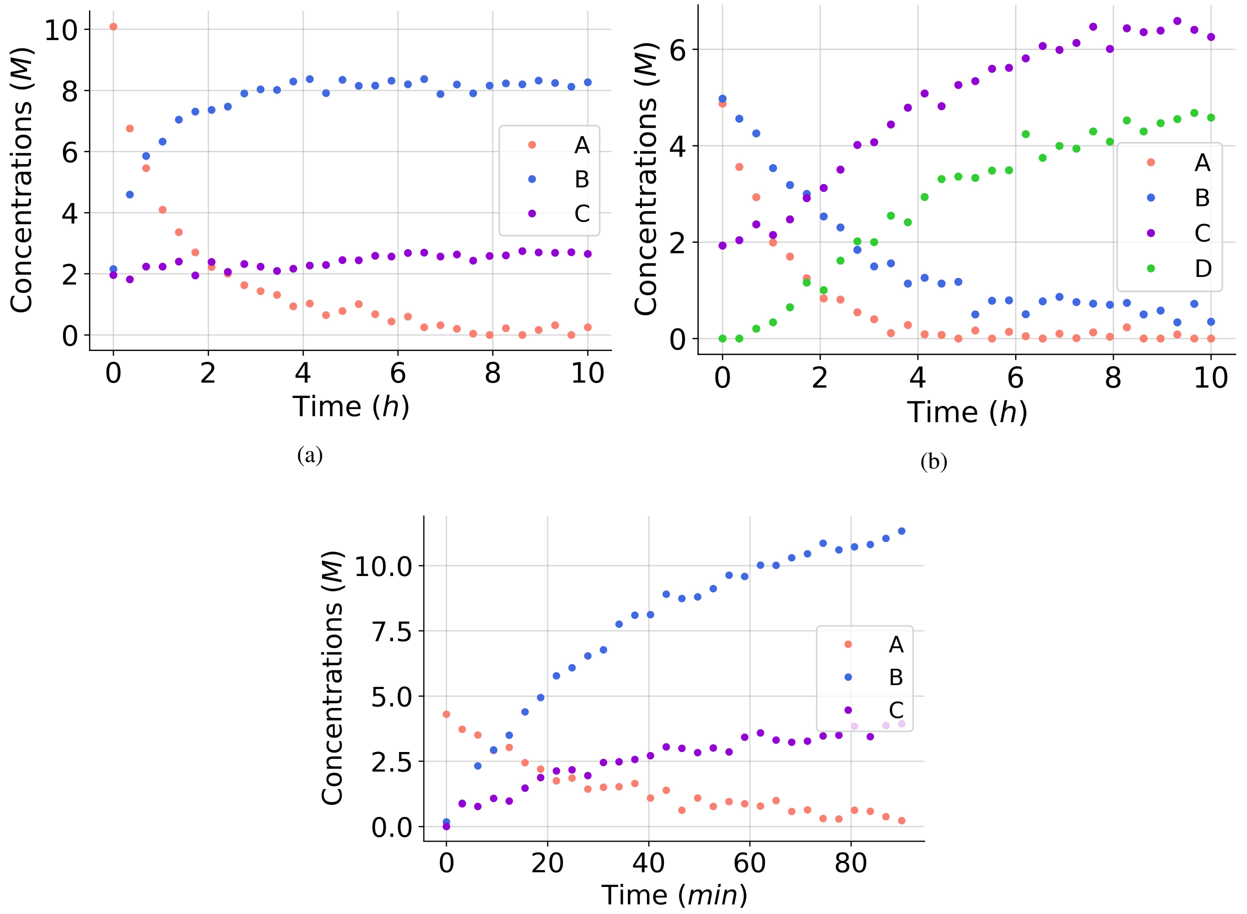
The application of SiMBA to the dehydration of fructose case study demonstrated its ability to uncover a mechanistic pathway that aligns with literature-accepted models [50], even though the data originated from a rate law validated by experimental findings [49] rather than from a constructed microkinetic model with hidden intermediates (alike the other two presented case studies). By working with a system where only the concentrations of fructose (A), water (B), and hydroxymethylfurfural (C) were available, SiMBA inferred the presence and behavior of unobserved species in a manner that remained consistent with a widely accepted reaction mechanism in literature.

In the first iteration, SiMBA identified all permutations of the simplest possible reaction configuration consistent with the overall stoichiometry. For this case study, these permutations resulted in 10 candidate reaction matrices, each describing a minimal three-step mechanism involving five total species. For an in-depth discussion of these initial candidates, please refer to the Supplementary Information. The best of these initial models, shown in the first row of Table 3, achieved an AIC value of -166.18, indicating a reasonable fit to the in-silico data. As seen in the top right plot of Figure 6, this initial mechanism satisfactorily captures the concentration profiles of A, B, and C, yet leaves open the possibility that additional chemical complexity could yield a still better match to the observed dynamics.

It is interesting to see that in iteration 1, SiMBA proposes a mechanism that is analogous to a widely accepted one in literature in which the 5-membered ring of fructose remains intact [49]. The first dehydration step yields intermediate D, which can exist either as the enol or keto tautomers. The second elimination introduces a unit of unsaturation in the ring (E). Finally, HMF can be obtained from the last dehydration from the ring. Given that the tautomerism is a very fast process, this mechanism can be described kinetically by three consecutive elementary dehydration steps. The detailed mechanism can be found in Fig. 5 (i).

Iteration 2 introduced a more elaborate mechanism by appending an additional elementary step and including an extra intermediate species (F). The resulting 4x6 matrix improved the fit of the model, offering a more nuanced description of how fructose converts into HMF through an additional intermediate stage. As reported in Table 3, this enhanced

View Article Online
DOI: 10.1039/D5SC01473E



(a)

(b)

(c)

Figure 4: (a) The in-silico data of one of the computational experiments for the hypothetical reaction. (b) The generated data of one of the computational experiments for the aldol condensation reaction. (c) The generated data of one of the computational experiments for the the dehydration of fructose to HMF.

mechanism yielded an AIC of -169.40 – a slight improvement that underscores a better balance between complexity and predictive accuracy. The middle-right plot of Figure 6 shows a slightly improved alignment between the simulated trajectories and the measured concentrations, particularly for water (B), which now follows the experimentally validated rate model's curvature slightly more precisely.

Remarkably, in this iteration, SiMBA also recovers an underlying sequence of elementary steps that is analogous to another widely accepted mechanism in the literature for the dehydration of fructose [49, 55]. This mechanism starts with the acyclic form of fructose (A), which is thought to be more abundant. In order for dehydration steps to occur, the formation of an enediol intermediate (D) is thought to be critical. Following two sequential dehydration steps, the dideoxyhexosulose intermediate (F) can cyclise very readily to form the 5-membered ring, prior to the elimination of the final water molecule to yield HMF (C). The detailed mechanism can be found in Fig. 5 (ii).

Iteration 3 further expanded the proposed mechanism by introducing yet another species (G). Although this more complex mechanism continued to reproduce the trajectory of the reaction reasonably well, its increased complexity did not translate into further gains in predictive power; the AIC rose to -161.52, signifying that the added steps merely inflated model complexity without meaningful improvement in data fitting. This is corroborated by the bottom-right plot of Figure 6, where the concentration profiles remain comparable to those from iteration 1, highlighting diminishing returns on model complexity. Consequently, the algorithm converged in iteration 3, as no additional refinement offered a better trade-off between model simplicity and accuracy. Among all iterations, the mechanism discovered in iteration 2 proved optimal.

View Article Online
DOI: 10.1039/D5SC01473E

The fructose example further illustrates the exponential scaling: iteration 1 ($5^{15} \approx 3.05 \times 10^{10}$ possible matrices) takes 12.80 s; iteration 2 ($5^{24} \approx 5.96 \times 10^{16}$) takes 1,978.81 s; and iteration 3 ($5^{35} \approx 2.91 \times 10^{24}$) takes 10,217.54 s. Two key conclusions follow. First, the total enumeration size grows exponentially with matrix dimensions, but the backtracking time – and thus the feasible subset – is strongly problem-dependent (i.e., dependent on SiMBA's starting point and the stoichiometry of the reaction) but remains small. Second, as network complexity grows, most of the runtime shifts from backtracking to the parameter-estimation phase, which becomes the dominant cost in evaluating every feasible candidate.

In this example, unlike our other case studies – where we deliberately hid intermediates in a microkinetic simulation – this example underscores SiMBA's value in a realistic setting. The results establish that SiMBA can parse complex, experimentally grounded datasets and distill them into verifiable mechanistic pathways, further solidifying its robustness and practical relevance for catalytic reaction discovery.

View Article Online
DOI: 10.1039/D5SC01473E

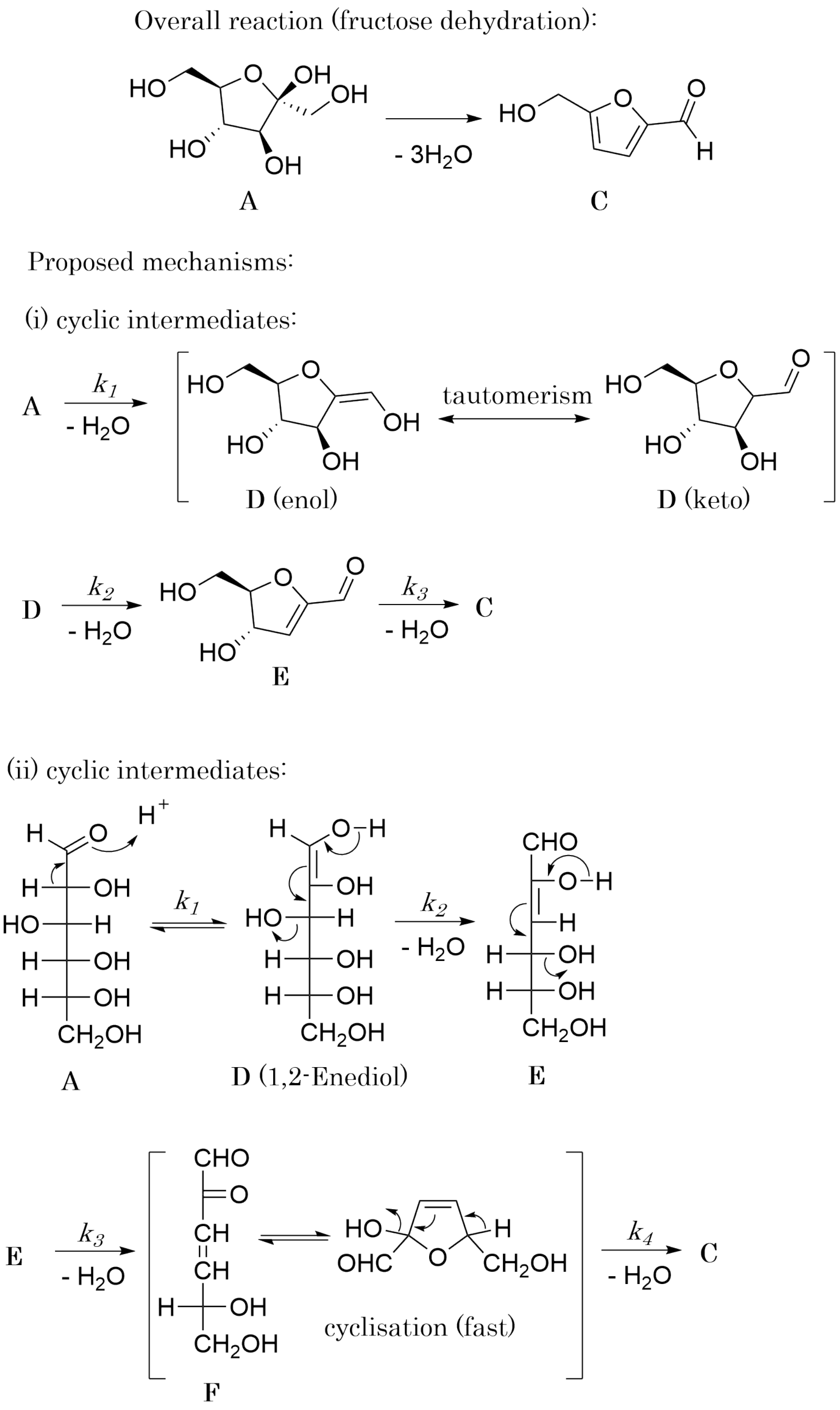


Figure 5: The transformation of fructose (A) to HMF (C) is known to be facile and involves three dehydration steps, eliminating 3 molecules of water. There are two general mechanistic pathways that are commonly proposed in literature. In the cyclic pathway (found in iteration 1), the five-membered ring remains intact and undergoes three consecutive dehydration steps: the first step yields intermediate D (enol or keto tautomer), followed by a second dehydration to produce intermediate E, and a final dehydration to form HMF. In the acyclic pathway (found in iteration 2 and chosen by SiMBA), fructose is proposed to adopt an open-chain form, which tautomerizes through an enediol intermediate (also labeled D). After two sequential dehydration steps, the resulting intermediate F cyclises readily, and the last dehydration step produces HMF. Both routes eliminate a total of three water molecules.

View Article Online
DOI: 10.1039/D5SC01473E

Table 3: Evolution of reaction mechanisms, microkinetic models, and AIC values across three iterations of the SiMBA process for the dehydration of fructose case study. The table shows the progression of SiMBA through iterations 1, 2, and 3 for the given reaction system. For each iteration, the reaction mechanism, corresponding microkinetic model, and the AIC values are presented. In iteration 1, the simplest mechanism is identified with an AIC of -166.18. As complexity increases in iterations 2 and 3, intermediates such as F and G are introduced, and the model structure becomes more intricate. The best-fit mechanism is achieved in iteration 2 with an AIC of -169.40, while iteration 3, despite introducing additional complexity, yields a higher AIC value of -161.52, indicating that further refinement may not improve model accuracy considerably.

Iteration	Reaction Mechanism	Microkinetic Model	AIC Value
1	$A \xrightarrow{k_1} B + D$ $D \xrightarrow{k_2} B + E$ $E \xrightarrow{k_3} B + C$	$\frac{dC_A}{dt} = -k_1C_A$ $\frac{dC_B}{dt} = k_1C_A + k_2C_D + k_3C_E$ $\frac{dC_C}{dt} = k_3C_E$ $\frac{dC_D}{dt} = k_1C_A - k_2C_D$ $\frac{dC_E}{dt} = k_2C_D - k_3C_E$	-166.18
2	$A \xrightarrow{k_1} D$ $D \xrightarrow{k_2} B + E$ $E \xrightarrow{k_3} B + F$ $F \xrightarrow{k_4} B + C$	$\frac{dC_A}{dt} = -k_1C_A$ $\frac{dC_B}{dt} = k_2C_D + k_3C_E + k_4C_F$ $\frac{dC_C}{dt} = k_4C_F$ $\frac{dC_D}{dt} = k_1C_A - k_2C_D$ $\frac{dC_E}{dt} = k_2C_D - k_3C_E$ $\frac{dC_F}{dt} = k_3C_E - k_4C_F$	-169.40
3	$A \xrightarrow{k_1} B + D$ $D \xrightarrow{k_2} E + F$ $E + F \xrightarrow{k_3} D$ $D \xrightarrow{k_4} B + G$ $G \xrightarrow{k_5} B + C$	$\frac{dC_A}{dt} = -k_1C_A$ $\frac{dC_B}{dt} = k_1C_A + k_4C_D + k_5C_G$ $\frac{dC_C}{dt} = k_5C_G$ $\frac{dC_D}{dt} = k_1C_A - k_2C_D + k_3C_EC_F - k_4C_D$ $\frac{dC_E}{dt} = k_2C_D - k_3C_EC_F$ $\frac{dC_F}{dt} = k_2C_D - k_3C_EC_F$ $\frac{dC_G}{dt} = k_4C_D - k_5C_G$	-161.52

View Article Online
DOI: 10.1039/D5SC01473E

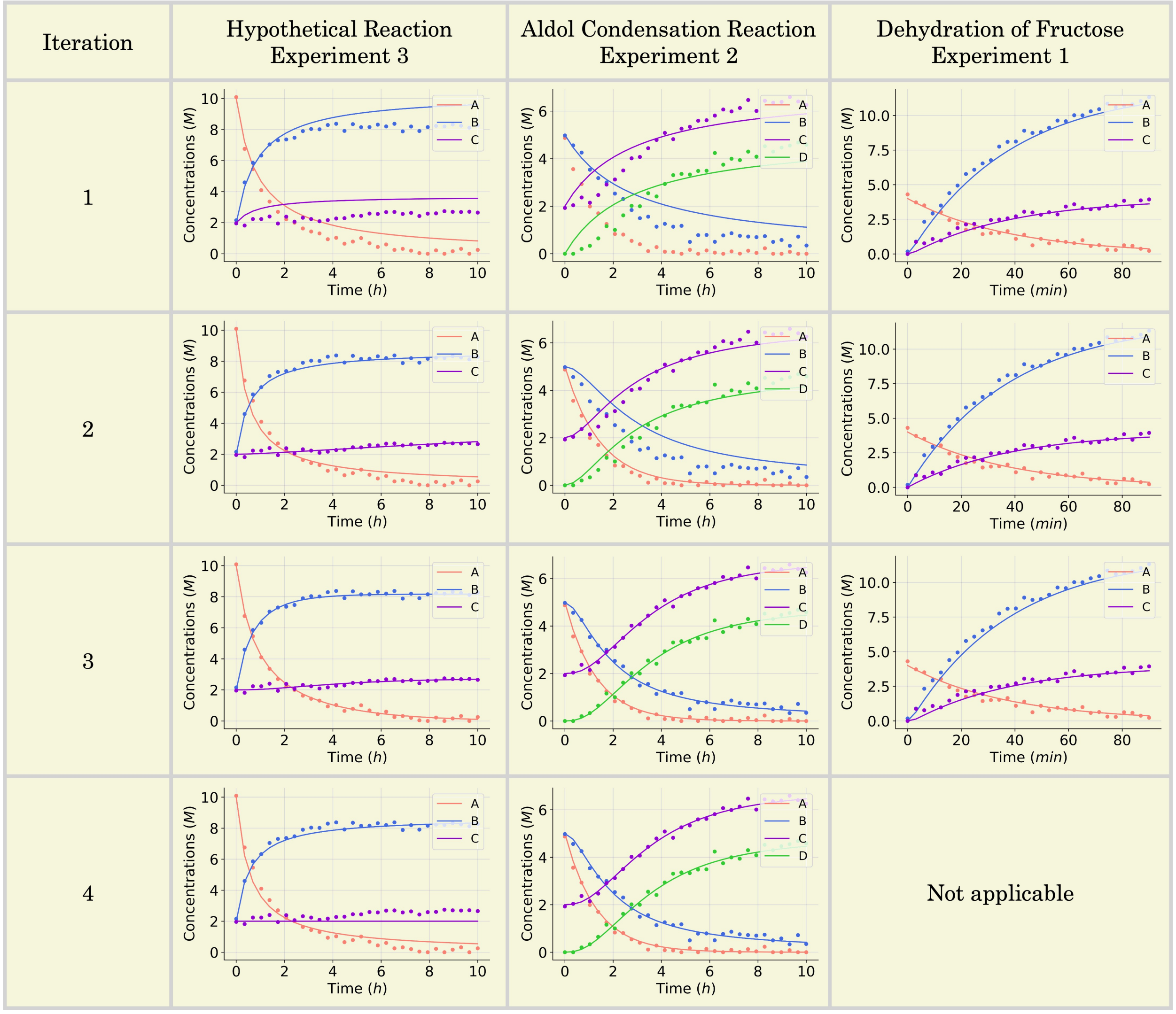


Figure 6: Model fit of the selected mechanisms across iterations for three case studies: the Hypothetical Reaction (Experiment 3), Aldol Condensation Reaction (Experiment 2), and Dehydration of Fructose Reaction (Experiment 1). Each plot shows the concentration profiles of species A (blue), B (red), C (yellow), and D (where applicable) over time, with solid lines representing model predictions and dotted points corresponding to in-silico data. For the hypothetical reaction (left column), SiMBA progressively refines the mechanism through iterations, achieving the best model fit by iteration 3. The aldol condensation reaction (middle column) shows notable improvement in fit by iteration 3, where the model captures the observed data well for all species. Iteration 4 introduces additional complexity but does not improve the fit, as demonstrated by the increased AIC value. For the dehydration of fructose reaction (right column), the model performs well from iteration 1, with iteration 2 yielding the optimal mechanism, which aligns with the data-generating model. The AIC value increases in iteration 3, signaling that additional complexity does not enhance the fit, and the process is terminated.

3.4 Methodological Limitations

While SiMBA represents an advancement in the automated construction of microkinetic models, it is not without its limitations. One of the primary challenges is the lack of inherent chemical identification for intermediates, which requires expert input when dealing with complex systems. This limitation can constrain the algorithm's utility in scenarios where the identification of novel intermediates is crucial for understanding the reaction mechanism. Additionally, SiMBA's approach to exploring extensive mechanism spaces is computationally demanding, particularly as the complexity of the potential mechanisms increases. Furthermore, the optimization process can be sensitive to initial parameter guesses, which might lead to suboptimal solutions if not managed carefully.

View Article Online
DOI: 10.1039/D5SC01473E

It is also important to note that despite SiMBA's systematic approach and rigorous filtering criteria, the mechanisms returned are fundamentally justified by the kinetic data rather than guaranteed to represent the actual underlying chemical pathways. The inherent limitation here is that concentration-time datasets, especially when incomplete, inherently underconstrain the reaction network. Consequently, SiMBA-generated mechanisms should be interpreted as being consistent with the available kinetic data and chemically plausible within the defined constraints (which can always be augmented), but not necessarily as uniquely true representations of the fundamental reaction mechanisms. This caveat is particularly critical when applying SiMBA to real-world experimental datasets, which may be incomplete or subject to measurement uncertainties. Users should therefore view the generated mechanisms as robust hypotheses warranting further experimental verification and refinement.

To mitigate the other limitations, several strategies have been implemented within the current study. For instance, to address the computational demands associated with exploring large mechanism spaces, we have employed a backtracking technique, as detailed in Section 2.1. This method significantly reduces the search space by eliminating unfeasible pathways early in the process, and the exploration has been parallelized to further improve computational efficiency. To counter the potential sensitivity to initial parameter guesses during optimization, as discussed in Section 2.3, we have utilized a well-established optimization algorithm, specifically the BFGS algorithm, with a multi-start option. This approach increases the likelihood of finding the global optimum by starting the optimization from multiple initial guesses.

Looking ahead, future work will focus on overcoming the lack of inherent chemical knowledge as well as continuing to reduce the computational cost to further enhance SiMBA's capabilities. For the issue of chemical identification of intermediates, we plan to explore the use of quantum chemistry methodologies. While quantum chemistry workflows (e.g., as reviewed in Simm et. al. [56]) provide a systematic route to enumerate and validate reaction pathways, their high computational cost often limits the breadth of mechanism exploration – particularly for complex networks with many intermediates. We therefore envision SiMBA serving as a low-cost "exploration" engine that rapidly identifies the simplest skeleton mechanisms consistent with kinetic data available. These skeletal networks can then be subjected to more expensive DFT or other quantum-chemical calculations ("exploitation") to assign chemical identities, compute activation barriers, and confirm intermediate stabilities. In cases where quantum-chemical results diverge from SiMBA's proposal (e.g., predicting additional intermediates), targeted MBDoE can be employed to generate discriminating data and refine the mechanism (as explained in Section 2.5). This hybrid workflow would thus combine the speed and parsimony of data-driven exploration with the physical rigor of first-principles validation.

Additionally, we are considering to add canonicalization rules into SiMBA, which will decrease the number of duplicates that, at the moment, the methodology inevitably computes and explores (a more in-depth discussion can be found in the Supplementary Information). We are also considering the integration of uncertainty quantification methods, which will increase the robustness of the models proposed by SiMBA. These enhancements aim to make SiMBA a powerful tool for chemists and engineers, capable of addressing the diverse challenges encountered in kinetic discovery and reaction mechanism elucidation.

4 Conclusions

In this paper, we have presented SiMBA (Simplest Mechanism Builder Algorithm), an efficient approach to microkinetic model discovery that aims to address key limitations in both manual and automated methods. Microkinetic models play a crucial role in various industries, including pharmaceuticals, petrochemicals, and environmental engineering, by helping to optimize chemical processes and understanding their environmental impact. However, traditional methods for constructing these models are often time-consuming, complex, and prone to human error, as they require extensive expertise to manually identify reaction mechanisms and intermediates. While automated approaches have emerged to overcome these challenges, they tend to generate overly complex models or rely heavily on prior knowledge, limiting their practical application.

SiMBA was developed to fill this gap by introducing a minimalistic, data-driven approach that incrementally builds model complexity based on available information. Unlike other methods, SiMBA begins with the simplest possible mechanism and systematically adds complexity only if the additional parameters provide informational gain. This balance between simplicity and accuracy is achieved through four key phases: mechanism generation, mechanism translation, parameter estimation, and model comparison. The algorithm starts by proposing feasible reaction mechanisms using a parallelized backtracking algorithm, translates these mechanisms into systems of ODEs, optimizes their kinetic parameters, and selects the best model using the AIC to ensure the right trade-off between model complexity and fit.

The effectiveness of SiMBA was demonstrated through three case studies: a hypothetical reaction, an aldol condensation, and the dehydration of fructose. In each case, SiMBA successfully distilled complex reaction behaviors into accurate

View Article Online
DOI: 10.1039/D5SC01473E

models, even in situations where intermediates were not directly observable. These case studies highlight the algorithm's versatility and robustness in generating models.

While SiMBA has proven to be a powerful tool for microkinetic model discovery, it is not without limitations. The current version does not provide chemical identification of intermediates, necessitating expert input for it. Additionally, while the algorithm excels at balancing simplicity with accuracy, incorporating uncertainty quantification could further enhance the robustness of its predictions. Future work will focus on integrating more chemical knowledge and techniques for identifying intermediates, as well as expanding the algorithm's capabilities to address uncertainty in model predictions.

In conclusion, SiMBA offers a novel approach to overcoming many of the challenges associated with existing automated methods for microkinetic model discovery. By systematically generating, refining, and evaluating microkinetic models, SiMBA provides a new framework for mechanistic discovery. As SiMBA continues to evolve with future enhancements like uncertainty quantification and intermediate identification, we hope that it will become a useful tool for chemists and engineers, helping bridge the gap between theoretical exploration and industrial applications.

Author Contributions

Miguel Ángel de Carvalho Servia: Conceptualization, formal analysis, investigation, methodology, project administration, software development, validation, visualization, writing (original draft), and writing (review and editing).

King Kuok (Mimi) Hii: Conceptualization, formal analysis, funding acquisition, supervision, writing (original draft), and writing (review and editing).

Klaus Hellgardt: Conceptualization, formal analysis, funding acquisition, supervision, and writing (review and editing).

Dongda Zhang: Conceptualization, funding acquisition and supervision.

Ehecatl Antonio del Rio Chanona: Conceptualization, formal analysis, funding acquisition, methodology, project administration, supervision, and writing (review and editing).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments and Funding

This work was supported by the EPSRC Centre of Doctoral Training for Next Generation Synthesis & Reaction Technology (rEaCt) funding grant EP/S023232/1.

Appendix A. Supplementary Information

The accompanying supplementary information offers an in-depth discussion of the candidates proposed by SiMBA in the first iteration of the dehydration of fructose case study.

The code used to produce all results and graphs shown in this work can be accessed at `https://github.com/OptiMaL-PSE-Lab/auto_react_mech_construct`.

References

[1] Zhou X, Fu L, Wang P, Yang L, Zhu X, Li CG. Drug-herb interactions between Scutellaria baicalensis and pharmaceutical drugs: Insights from experimental studies, mechanistic actions to clinical applications. Biomed Pharmacother. 2021 Jun;138:111445.

[2] Subelzu N, Schöneich C. Near UV and Visible Light Induce Iron-Dependent Photodegradation Reactions in Pharmaceutical Buffers: Mechanistic and Product Studies. Mol Pharmaceutics. 2020 Sep;17(11):4163–4179.

[3] Uzondu B, Leung LY, Mao C, Yang CY. A mechanistic study on tablet ejection force and its sensitivity to lubrication for pharmaceutical powders. Int J Pharm. 2018 May;543(1–2):234–244.

View Article Online
DOI: 10.1039/D5SC01473E

[4] Zhang F, Zhang Q, Chen Y, Xu L, Li Z, Wang Q, et al. Advances and mechanistic insights in the catalytic semi-hydrogenation of acetylene over non-metallic catalysts. Appl Catal A-Gen. 2023 Oct;667:119447.

[5] Wang Y, Zhang L, Lin R, Han X, Xie K, Huang C, et al. Experimental and mechanistic study on the effect of active components and calcination temperatures on biochar-based catalysts for catalyzing heavy oil viscosity reduction. Geoenergy Sci Eng. 2024 Sep;240:213078.

[6] Qajar J, Razavifar M, Riazi M. A mechanistic study of the synergistic and counter effects of ultrasonic and solvent treatment on the rheology and asphaltene structure of heavy crude oil. Chem Eng Process. 2024 Jan;195:109619.

[7] Nelson PN. A DFT mechanistic study of two possible hydrolytic evolution pathways of thiamethoxam; implications in food and environmental safety. Comput Theor Chem. 2021 Aug;1202:113333.

[8] Kumari V, Sharma A, Kumar N, Sillanpää M, Makgwane PR, Ahmaruzzaman M, et al. TiO2-CeO2 assisted heterostructures for photocatalytic mitigation of environmental pollutants: A comprehensive study on band gap engineering and mechanistic aspects. Inorg Chem Commun. 2023 May;151:110564.

[9] Rajamohan N, Said Zahir Said Al Shibli F, Rajasimman M. Environmentally benign Prosopis juliflora extract for corrosion protection by sorption - Gravimetric, mechanistic and thermodynamic studies. Environ Res. 2022 Jan;203:111816.

[10] Gaud N, Gogola D, Kowal-Chwast A, Gabor-Worwa E, Littlewood P, Brzózka K, et al. Physiologically based pharmacokinetic modeling of CYP2C8 substrate rosiglitazone and its metabolite to predict metabolic drug-drug interaction. Drug Metab Pharmacokinet. 2024 Aug;57:101023.

[11] Juhász Luty-Błocho M, Wojnicki M, Tóth GK, Csapó E. General method for kinetic and thermodynamic evaluation of a receptor model peptide-drug molecule interaction studied by surface plasmon resonance. Microchem J. 2019 Jun;147:311–318.

[12] Maqbool T, Yousuf RI, Ahmed FR, Shoaib MH, Irshad A, Saleem MT, et al. Cellulose ether and carbopol 971 based gastroretentive controlled release formulation design, optimization and physiologically based pharmacokinetic modeling of ondansetron hydrochloride minitablets. Int J Biol Macromol. 2024 Sep;276:133841.

[13] Nations U. Montreal Protocol on Substances that Deplete the Ozone Layer Final Act 1987. J Environ Law. 1989 03;1(1):128-36.

[14] Lallas PL. The Stockholm Convention on Persistent Organic Pollutants. Am J Int Law. 2001;95(3):692–708.

[15] Puliyanda A, Srinivasan K, Sivaramakrishnan K, Prasad V. A review of automated and data-driven approaches for pathway determination and reaction monitoring in complex chemical systems. Digit Chem Eng. 2022 Mar;2:100009.

[16] Ratkiewicz A, Truong TN. Automated mechanism generation: From symbolic calculation to complex chemistry. Int J Quantum Chem. 2005 Aug;106(1):244–255.

[17] Liu M, Grinberg Dana A, Johnson MS, Goldman MJ, Jocher A, Payne AM, et al. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. J Chem Inf Model. 2021 May;61(6):2686–2696.

[18] Broadbelt LJ, Stark SM, Klein MT. Computer Generated Pyrolysis Modeling: On-the-Fly Generation of Species, Reactions, and Rates. Ind Eng Chem Res. 1994 Apr;33(4):790–799.

[19] Warth V, Battin-Leclerc F, Fournet R, Glaude PA, Côme GM, Scacchi G. Computer based generation of reaction mechanisms for gas-phase oxidation. Comput Chem. 2000 Jul;24(5):541–560.

[20] Vandewiele, Nick and Van Geem, Kevin and Reyniers, Marie-Françoise and Marin, Guy. Genesys: kinetic model construction using chemo-informatics. Chem Eng J. 2012;207:526-38.

[21] Rangarajan S, Bhan A, Daoutidis P. Language-oriented rule-based reaction network generation and analysis: Description of RING. Comput Chem Eng. 2012 Oct;45:114–123.

[22] Ranzi E, Dente M, Goldaniga A, Bozzano G, Faravelli T. Lumping procedures in detailed kinetic modeling of gasification, pyrolysis, partial oxidation and combustion of hydrocarbon mixtures. Prog Energy Combust Sci. 2001 Jan;27(1):99–139.

[23] Blurock ES. Reaction: System for Modeling Chemical Reactions. J Chem Inf Comput. 1995 May;35(3):607–616.

[24] Fontain E, Bauer J, Ugi I. Computer Assisted Bilateral Generation of Reaction Networks from Educts and Products. Chem Lett. 1987 Jan;16(1):37–40.

[25] Porollo AA, Lushnikov DE, Pivina TS, Ivshin VP. Computer representation and generation of possible pathways for thermal decomposition reactions of organic compounds. J Mol Struct: THEOCHEM. 1997 Feb;391(1–2):117–124.

[26] Karaba A, Zamostny P, Lederer J, Belohlav Z. Generalized Model of Hydrocarbons Pyrolysis Using Automated Reactions Network Generation. Ind Eng Chem Res. 2013 Jul;52(44):15407–15416.

View Article Online
DOI: 10.1039/D5SC01473E

[27] Chinnick SJ, Baulch DL, Ayscough PB. An expert system for hydrocarbon pyrolysis reactions. Chemom Intell Lab Syst. 1988 Nov;5(1):39–52.

[28] Peterson L, Gosea IV, Benner P, Sundmacher K. Digital twins in process engineering: An overview on computational and numerical methods. Comput Chem Eng. 2025;193:108917.

[29] Di Maio FP, Lignola PG. KING, a KInetic Network Generator. Chem Eng Sci. 1992 Jun;47(9–11):2713–2718.

[30] Clymans PJ, Froment GF. Computer-generation of reaction paths and rate equations in the thermal cracking of normal and branched paraffins. Comput Chem Eng. 1984 Jan;8(2):137–142.

[31] Susnow RG, Dean AM, Green WH, Peczak P, Broadbelt LJ. Rate-Based Construction of Kinetic Models for Complex Systems. J Phys Chem A. 1997 May;101(20):3731–3740.

[32] Warth V, Stef N, Glaude PA, Battin-Leclerc F, Scacchi G, Côme GM. Computer-Aided Derivation of Gas-Phase Oxidation Mechanisms: Application to the Modeling of the Oxidation of n-Butane. Combust Flame. 1998 Jul;114(1–2):81–102.

[33] Glaude P. Construction and simplification of a model for the oxidation of alkanes. Combust Flame. 2000 Sep;122(4):451–462.

[34] Ratkiewicz A, Truong TN. Application of Chemical Graph Theory for Automated Mechanism Generation. J Chem Inf Comput. 2002 Dec;43(1):36–44.

[35] Ranzi E, Faravelli T, Gaffuri P, Sogaro A. Low-temperature combustion: Automatic generation of primary oxidation reactions and lumping procedures. Combust Flame. 1995 Jul;102(1–2):179–192.

[36] Van de Vijver R, Vandewiele NM, Bhoorasingh PL, Slakman BL, Seyedzadeh Khanshan F, Carstensen H, et al. Automatic Mechanism and Kinetic Model Generation for Gas- and Solution-Phase Processes: A Perspective on Best Practices, Recent Advances, and Future Challenges. Int J Chem Kinet. 2015 Jan;47(4):199–231.

[37] Burés J, Larrosa I. Organic reaction mechanism classification using machine learning. Nature. 2023 Jan;613(7945):689–695.

[38] de Carvalho Servia M del Rio Chanona EA. In: Model Structure Identification. Royal Society of Chemistry; 2023. p. 85–108.

[39] Hunter WG, Reiner AM. Designs for Discriminating Between Two Rival Models. Technometrics. 1965 Aug;7(3):307-23.

[40] Erickson J. Algorithms. Independently published; 2019.

[41] Van Rossum G. The Python Library Reference, release 3.8.2. Python Software Foundation; 2020.

[42] Jiscoot N, Uslamin EA, Pidko EA. Model-based evaluation and data requirements for parallel kinetic experimentation and data-driven reaction identification and optimization. Digit Discov. 2023;2(4):994–1005.

[43] Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Program. 1989 Aug;45(1-3):503-28.

[44] Malouf R. A Comparison of Algorithms for Maximum Entropy Parameter Estimation. In: Proceedings of the 6th Conference on Natural Language Learning - Volume 20. COLING-02. USA: Association for Computational Linguistics; 2002. p. 1–7.

[45] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020;17:261-72.

[46] Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974;19(6):716-23.

[47] de Carvalho Servia M Sandoval IO, Hii KKM, Hellgardt K, Zhang D, Antonio del Rio Chanona E. The automated discovery of kinetic rate models – methodological frameworks. Digit Discov. 2024;3(5):954–968.

[48] Nielsen AT, Houlihan WJ. The Aldol Condensation. Wiley; 2011.

[49] van Putten R, Soetedjo JNM, Pidko EA, van der Waal JC, Hensen EJM, de Jong E, et al. Dehydration of Different Ketoses and Aldoses to 5-Hydroxymethylfurfural. ChemSusChem. 2013 Aug;6(9):1681–1687.

[50] Chen C, Lv M, Hu H, Huai L, Zhu B, Fan S, et al. 5-Hydroxymethylfurfural and its Downstream Chemicals: A Review of Catalytic Routes. Adv Mater. 2024 Jun;36(37).

[51] Mee R. A comprehensive guide to factorial two-level experimentation. Springer Science & Business Media; 2009.

[52] Schrecker L, Dickhaut J, Holtze C, Staehle P, Vranceanu M, Hellgardt K, et al. Discovery of unexpectedly complex reaction pathways for the Knorr pyrazole synthesis via transient flow. React Chem Eng. 2023;8(1):41-6.

View Article Online
DOI: 10.1039/D5SC01473E

[53] Waldron C, Pankajakshan A, Quaglio M, Cao E, Galvanin F, Gavriilidis A. Model-based design of transient flow experiments for the identification of kinetic parameters. React Chem Eng. 2020;5:112-23.

[54] Taylor CJ, Booth M, Manson JA, Willis MJ, Clemens G, Taylor BA, et al. Rapid, automated determination of reaction models and kinetic parameters. J Chem Eng. 2021 Jun;413:127017.

[55] van Putten RJ, van der Waal JC, de Jong E, Rasrendra CB, Heeres HJ, de Vries JG. Hydroxymethylfurfural, A Versatile Platform Chemical Made from Renewable Resources. Chem Rev. 2013 Feb;113(3):1499–1597.

[56] Simm GN, Vaucher AC, Reiher M. Exploration of Reaction Pathways and Chemical Transformation Networks. J Phys Chem A. 2018 Nov;123(2):385–399.

View Article Online
DOI: 10.1039/D5SC01473E

The code for SiMBA, an open-source Python automated microkinetic model discovery tool, can

be found at https://doi.org/10.5281/zenodo.14913720 with DOI -- 10.5281/zenodo.5510203.

The version of the code employed for this study is version v0.1.0.

Data for this article, including in-silico datasets are available at OptiMaL-PSELab/auto_react_mech_construct at https://doi.org/10.5281/zenodo.14913720.