Reaction Chemistry & Engineering

Active Learning of Chemical Reaction Networks via Probabilistic Graphical Models and Boolean Reaction Circuits

Journal:	*Reaction Chemistry & Engineering*
Manuscript ID	RE-ART-08-2022-000315.R1
Article Type:	Paper
Date Submitted by the Author:	28-Oct-2022
Complete List of Authors:	Cohen, Maximilian; Univ. of Delaware Goculdas, Tejas; Univ. of Delaware Vlachos, Dionisios; Univ. of Delaware,

Active Learning of Chemical Reaction Networks via Probabilistic Graphical Models and Boolean Reaction Circuits

Maximilian Cohen[#1], Tejas Goculdas[#1], and Dionisios G. Vlachos[1,2]*

[1]Department of Chemical and Biomolecular Engineering, 150 Academy St., University of Delaware, Newark, DE 19716, USA
[2]Catalysis Center for Energy Innovation, RAPID Manufacturing Institute, and Delaware Energy Institute, 221 Academy St., Newark, DE 19716

[#] Equal contribution
* Corresponding author: vlachos@udel.edu

Abstract

Discerning networks of many reactions among multiple interconverting species is challenging. Here, we present a reaction network identification methodology. Our methodology enumerates all stoichiometrically and chemically feasible reactions and requires statistical evidence from effluent concentrations for the inclusion or exclusion of each from the reaction network, contrasting with the commonly seen incremental approach and other work of relying heavily upon chemical intuition and assuming the reactions occurring. Using graph theory alongside an active learning design of experiments that propose maximally informative feeds, we identify the underlying reaction network with minimal laboratory runs. We introduce chemistry-probabilistic graphical modeling and Boolean reaction circuits to statistically quantify which reactions occur from effluent concentrations. Our methodology accurately discerns active reactions, as showcased upon a laboratory network of cross-ketonization of furoic and lauric acid and validated upon simulated networks of thermal and CO_2-assisted ethane dehydrogenation.

Keywords

Reaction networks, kinetic modeling, probabilistic modeling, design of experiments, cross-ketonization, ethane dehydrogenation

1. Introduction

An inherent complexity of many reaction systems is the occurrence of multiple chemical reactions among several species in a reaction network. To properly model these systems, each reaction within the network must be identified for subsequent rate parameterization.[1, 2]

This task of reaction network identification (RNI) is common in catalysis research. Weingarten et al. created a kinetic model to explore optimal reactor configurations for the biomass upgrading of glucose to 5-hydroxymethylfurfural.[3] Comparing their proposed reaction network to those conjectured previously, they obtained a more accurate network using product co-feeding experiments. Years later, Vlachos and coworkers improved the kinetic model with additional reaction identification using isotopic labeling experiments,[4] and subsequently improved the process yield.[5] While investigating bio-oil upgrading with an anisole hydrodeoxygenation reaction over a $CoMo/Al_2O_3$ catalyst, Marin and coworkers proposed a detailed reaction network,[6] which they elucidated using delplots.[7, 8]

While RNI is a common component of evaluating biomass reaction systems,[9, 10] its application extends to petrochemical and pharmaceutical processes.[11, 12] The research of Bhan and coworkers has especially highlighted the importance of identifying the reaction network to construct reliable kinetic models.[13-17] They effectively confirmed the occurring reactions by applying Wojciechowski's criteria to first-rank delplots[18, 19] and conducting experiments with product, isotopic labeling, and probe molecule co-feeds.[20]

Various experimental data collection and analysis techniques exist to enable kinetic model development; design of experiments (DOE) can propose optimal experiments to minimize the number of measurements. However, most DOE work has focused upon reaction rate equations rather than reactions. Such DOE, building from the foundational work of Box, Hunter, Froment, Buzzi–Ferraris, and their collaborators,[21-25] is well complemented by the incremental approach of Marquardt, Bonvin, and coworkers.[26-29] This approach effectively attributes measured rate data to individual reactions, enabling their independent determination; however, the authors acknowledge their incremental approach is vulnerable to inaccurate RNI.[30-34] DOE specifically for RNI is needed.

Building from the work of Tsu et al.,[35] Bourne and coworkers created an automated RNI approach.[36-38] Their approach assesses full reaction profiles against all stoichiometrically feasible kinetic models to identify which fits best. It assumes a rate equation, a common assumption in kinetic model discrimination.[23, 39, 40] When assuming a rate equation and evaluating its statistical evidence, it is unknown whether the lack of fit is due to an incorrect rate equation or the reaction not occurring. A rate agnostic analysis is preferable. This limitation, combined with the acknowledgement of Bourne et al. that DOE for individual experiments offers potential improvement (i.e., inferring reactions from a single experiment rather than their approach requiring many experiments at different conversions for full reaction profiles),[37] clearly identifies the need for improving the RNI methodology.

Herein, we propose a framework for RNI using model-based DOE.[41] We introduce chemistry-probabilistic graphical models[42, 43] (PGMs) to propose the causal structure of reactions inducing chemical conversion and Bayesian inference to identify the statistically non-zero reaction extents. We combine this causal structure with Boolean logic to construct Boolean reaction circuits (BRCs) to infer the reactions given effluent concentrations. We recommend implementing both techniques using our new DOE methodology based on graph theory; its objective is to identify the reaction network graph (RNG) with minimal experiments and maximal certainty. Overall, we perform

optimally informative experiments, analyze the effluent data with PGMs and BRCs, update the RNG structure, and repeat the process in an active learning cycle until the entire network is identified. This RNI methodology is generalizable and applicable to traditional experiments and industry's thrust towards automated, optimized laboratories.[44, 45] We demonstrate this RNI framework upon a simulated ethane dehydrogenation network in Section 2, new biomass-derived cross-ketonization laboratory experiments in Section 3, and a simulated CO_2-assisted ethane dehydrogenation network in Section S4 of the supporting information (SI).

2. Demonstration of Reaction Network Identification Methodology Upon an Ethane Dehydrogenation Network

2.1 Overview

We propose an RNI methodology that is rate equation agnostic and requires minimal experiments using an active learning cycle until the network is confirmed with statistical confidence. Figure 1 depicts the key steps.

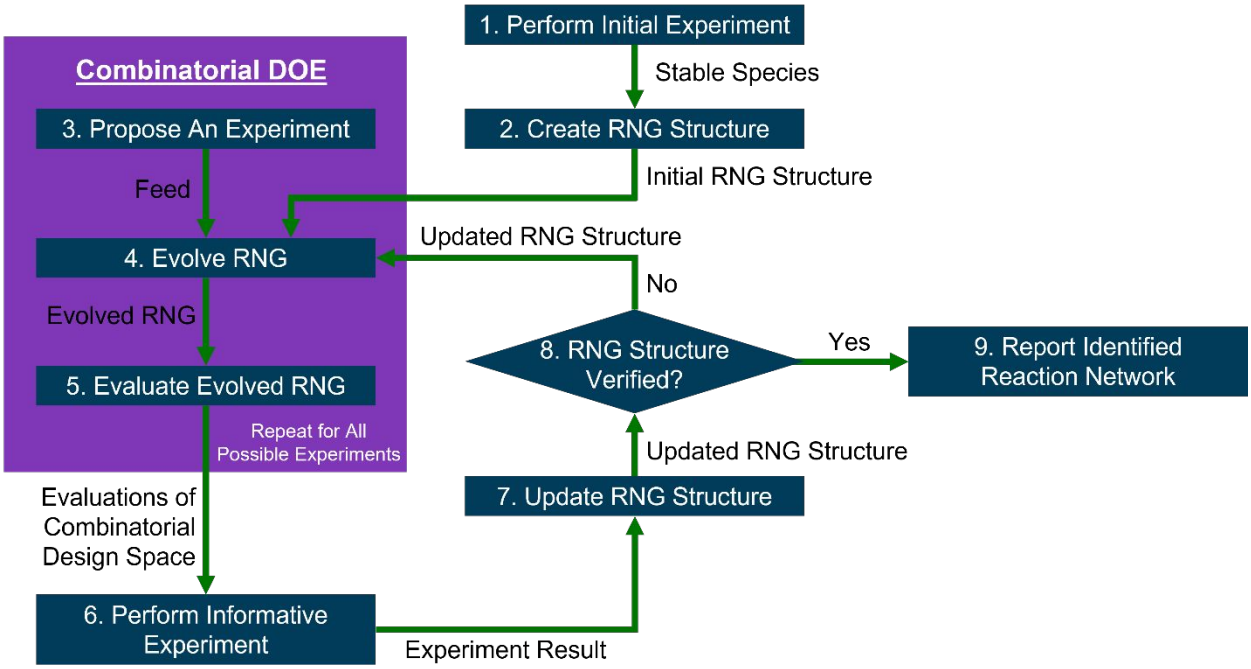

Figure 1: Workflow of RNI methodology. Information flow is depicted with green arrows; decisions (diamond) and actions (rectangle) are depicted with dark blue quadrilaterals. Boxed in purple, steps 3-5 are repeated for all possible experiments to implement a combinatorial DOE.

We analyze the ethane dehydrogenation reaction network throughout Section 2, as a simple illustrative example upon a simulated system informed from our prior work.[46] An initial experiment is performed (Step 1) where ethane (C_2H_6) is fed. Ethylene (C_2H_4), methane (CH_4), coke (C), and hydrogen (H_2) are observed as products; our objective is to determine the overall reactions among these stable species. New species may be measured at new reaction conditions. If a new species is detected, then new overall reactions need to be considered and the entire methodology needs to be performed from the beginning. Therefore, we recommend performing

experiments over a range of conditions for reaction systems until the species present are known with confidence. With measured C_2H_6, C_2H_4, CH_4, C, and H_2, there are several potential reactions (see below). In addition to the main dehydrogenation reaction, hydrogenolysis involving C-C bond scission leading to methane, and coking from the various hydrocarbons, can happen. It is unknown at the outset of a new catalyst investigation which reactions occur.[46-48]

2.2 Traditional Stoichiometrically Feasible Reactions and Graph Theory-based Representation of Reaction Networks

First, we exhaustively list all stoichiometrically feasible, overall reactions (rather than elementary, microkinetic ones). The general formula is Eq. (1).

$$\forall_{i,k}: \quad \sum_j \nu_{ij}\varepsilon_{jk} = 0 \tag{1}$$

Here, ν_{ij} is the stoichiometric coefficient of species j in reaction i, and ε_{jk} is the number of atoms of element k in species j. Specifically, we consider all combinations of two or fewer reactants with two or fewer products. In this test case, 20 stoichiometrically feasible reactions are identified (see data repository[49]). For example, the reaction $a_1H_2 \leftrightarrow a_2CH_4$ is infeasible, whereas $a_1C_2H_6 \leftrightarrow a_2C_2H_4 + a_3H_2$ is stoichiometrically feasible. These 20 stoichiometrically feasible reactions are pruned to 4 physically realistic reactions using expert knowledge from the literature: ethane dehydrogenation, ethylene coking, ethane hydrogenolysis, and methane coking. While reliance upon expert knowledge might be completely removed from this RNI framework in the future to facilitate complete automation, for now we include it for two reasons. First, expert knowledge can eliminate physically unrealistic reactions and thereby reduce the number of required experiments for RNI. Second, the reduction of considered reactions enables showcasing our new analyses by reducing linear dependence between reactions. In the future, the role of expert knowledge may be filled by using chemistry rules constructed from automated literature searches or from software specifying elementary reaction rules such as RING, NETGEN, COMGEN, and KING or overall reactions.[50-53]

These research tools[50-53] approach reaction enumeration differently from our method. They specify elementary reaction rules that explain how species can interconvert between one another and then enumerate reactions consistent with these rules. This contrasts with our enumeration approach including all stoichiometrically feasible reactions. Our approach is more inclusive ensuring no possible chemistry is overlooked; the traditional approach is more focused, providing a smaller subset that likely contains relevant chemistries. Both approaches rely upon regressing to experimental data to discern reactions and rate parameters for quality model-experiment agreement, but while our approach focuses on overall reactions, the traditional approach can identify multiple microkinetic pathways of elementary reactions leading to the same overall reaction. Our work can be extended by exploring competing microkinetic pathways after identifying the overall reactions, and we recommend our more inclusive enumeration technique to mitigate the risk of neglecting key chemistries.

These feasible reactions form a reaction network, displayed in Figure 2a. Arrows indicate each reaction's reversibility, determined using thermodynamics and expert knowledge from the literature (discussed in Section S1 of the SI). These traditional reaction networks communicate reaction fluxes but lack mathematical uniqueness, i.e., there are multiple, valid depictions since species can appear multiple times such as H_2 in Figure 2a. In the traditional approach, one typically writes overall rate expressions, e.g., power law or Langmuir Hinshelwood, and attempts to regress

to experimental data. Experimental data may entail time dependent or space velocity series and/or reaction kinetic type (reaction orders, apparent activation energies). More often, one does not even invoke exhaustive stoichiometric relations; rather, one simply hypothesizes rate expressions and regresses for a single potential network.

Next, we construct a RNG using graph theory (Step 2 in Figure 1) from our list of stoichiometrically viable, pruned reactions. We represent each species and reaction as a node. The directed edges display the flux within the system, as shown in Figure 2b. These graph theory-based RNGs depict the connectivity of the network, preserving mass because the net flux into each reaction node is zero, i.e., atomic species in are equal to atomic species out, as prescribed by Eq. (1). They also provide mathematical uniqueness since each species, reaction, and flux is represented only once, in contrast to the more traditional form of Figure 2a.

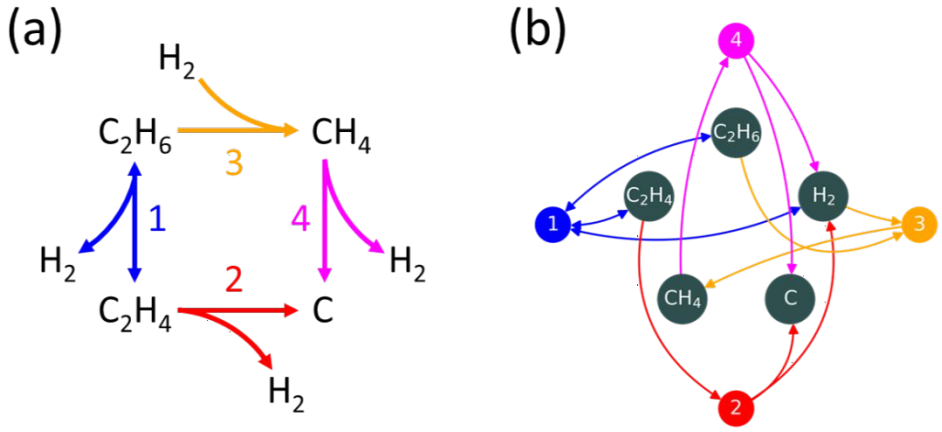

Figure 2: Alternative depictions of the ethane dehydrogenation reaction network. (a) Traditional overall reaction network. Reaction 1 (R1): ethane dehydrogenation in blue ($C_2H_6 \rightarrow C_2H_4 + H_2$). Reaction 2 (R2): coking from ethylene in red ($C_2H_4 \rightarrow 2C + 2H_2$). Reaction 3 (R3): ethane hydrogenolysis in orange ($C_2H_6 + H_2 \rightarrow 2CH_4$). Reaction 4 (R4): coking from methane in magenta ($CH_4 \rightarrow C + 2H_2$). (b) Graph theory representation of the reaction network. Each species and reaction are represented as a node. Relationships between nodes and colors correspond to (a) using directed edges to represent reaction flux and reversibility. Numbers indicate the reaction.

2.3 Design of Experiments and Active Learning

Next, we confirm the reactions taking place using maximally informative experiments, i.e., the experiments our DOE indicates are the most informative from the combinatoric set proposed. The following describes the mathematical operations of Steps 3-5 (Figure 1).

First, an experiment is proposed (Step 3) by suggesting a feed, for example feeding C_2H_4. The feed space is defined by the stable species of the RNG, but there is no restriction on specific concentrations. The RNG is then evolved to show which reactions occur (Step 4), as demonstrated in Figure 3, following a simple rule: if all reactants of a proposed reaction are present, we assume the reaction occurs and the products form. These products are then evolved; as reactions occur, products form, leading to more reactions. For example, feeding C_2H_4 allows reaction 2 (R2), $C_2H_4 \rightarrow 2C + 2H_2$ to occur. Then, H_2 reacts with C_2H_4 in R1 to form C_2H_6, $C_2H_4 + 2H_2 \rightarrow C_2H_6$. This evolution of the RNG continues: the C_2H_6 undergoes hydrogenolysis with H_2 to form CH_4 in R3, $C_2H_6 + H_2 \rightarrow 2CH_4$, and this CH_4 decomposes to C and H_2 via R4, $CH_4 \rightarrow C + 2H_2$. At this stage, the RNG evolution is complete since no more reactions occur, as shown in Figure 3. It describes the causality of chemical reactions, i.e., which reactions are caused to occur by the species present and which additional species are produced by the occurring reactions.

The RNG evolution is evaluated (Step 5) to determine which reactions would be revealed if the experiment were run; the number of revealed reactions is the information rating of a proposed experiment. For the example of the proposed experiment of Figure 3, the RNG evolution indicates that BRC analysis (discussed below) of this experiment's effluent concentrations would reveal three reactions (i.e., it has an information rating of three) while PGM analysis (discussed below) would reveal two reactions (i.e., it has an information rating of two). The procedure for obtaining these information ratings is discussed in Section 2.4.

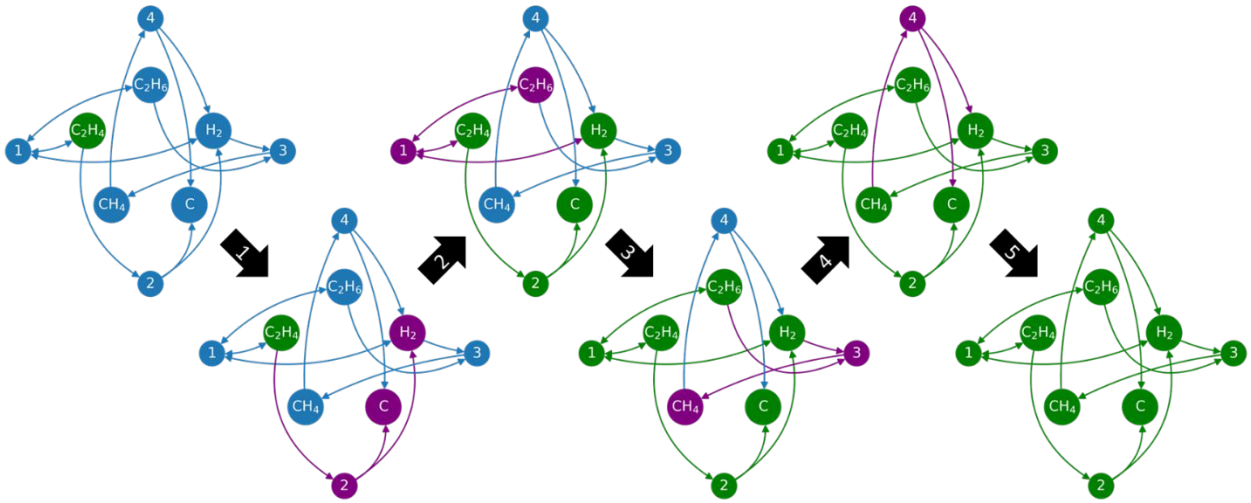

Figure 3: RNG representation of the evolution of active species and reactions within the ethane dehydrogenation reaction network. Each of the six RNGs represents a different state of evolution, and labeled arrows denote the sequence between states. Species and reaction nodes are either currently uninvolved (blue), newly introduced (purple), or previously incorporated (green). The initial state shows the initial feed; the final state denotes the active species and reactions given that feed.

We repeat the procedure for all possible feeds of a single species and two species (feeds of more species were found to be less informative). For our test case, this entails feeds of C_2H_6, C_2H_4, CH_4, H_2, beginning with coked catalyst, or combinations of any two of these. The RNG evolutions of all these feeds considered are documented in Section S2 of the SI. Combinatorically exploring the design space of feeds ensures evaluating the information rating of each potential experiment. This DOE approach is codified in Section S3 of the SI with algorithms. The maximally informative experiment (highest information rating) is identified and performed in the laboratory (Step 6). For this example, the experiment selected and conducted is feeding C_2H_4; the data collected is the effluent concentrations.

Analysis of this experiment's data confirms or rejects proposed reactions within the RNG, as detailed in Section 2.4. For this example, R1, R2, and R3 are confirmed from the effluent concentrations of the experiment of feeding C_2H_4. By updating the RNG with this information (Step 7), we then determine if any unconfirmed reactions remain (Step 8). If not, the reaction network has been fully identified for subsequent kinetic modeling (Step 9). However, if additional reactions require identification, the combinatorial DOE is repeated. This repeat of the DOE is an active learning cycle; an initial DOE identifies an informative experiment, the experiment is conducted, the RNG is updated, and this new information is used to inform another DOE. In this example, R1, R2, and R3 are confirmed, but R4 is still unconfirmed requiring an additional experiment. The DOE identifies the promising experiment of feeding CH_4, and once performed the effluent concentrations confirm R4 occurs. With only two experiments (one feeding C_2H_4,

another feeding CH_4), we confirm all four reactions occur to fully identify the ethane dehydrogenation network.

Occasionally, specific reactions cannot be identified with this approach, so alternate methods are employed. We demonstrate this in Section 3 by employing a modified delplot method.[7, 20] While these alternate methods may offer less statistical confidence or require more experiments, the structure of some RNI problems requires them.

2.4 RNG Trajectory Evaluation Methods

Herein, we present two methods for analyzing evolved RNGs to determine their information ratings: BRCs and PGMs. Depending on the RNG, either may provide higher information ratings.

Through a one-to-one mathematical transformation, BRCs convert the evolution of the RNG (Figure 3) into a Boolean logic circuit (Figure 4); certain species (inputs) must be present (on) to produce additional species (outputs). Even if one of the reactants is missing, the reaction does not occur. This behavior makes an "and" gate appropriate to describe each chemical reaction. Depending on which reactions occur (i.e., which "and" gates are activated), the effluent differs. For the ethane dehydrogenation network, if R2 does not occur, the effluent only contains C_2H_4, whereas if R2 occurs, H_2 and C are produced. Evaluating the effluent identifies the reaction(s) in a rate equation agnostic approach. The proposed experiment of feeding C_2H_4 is evaluated as having an information rating of 3 (i.e., this experiment identifies 3 reactions: R2, R1, and R3, as shown in Figure 4).

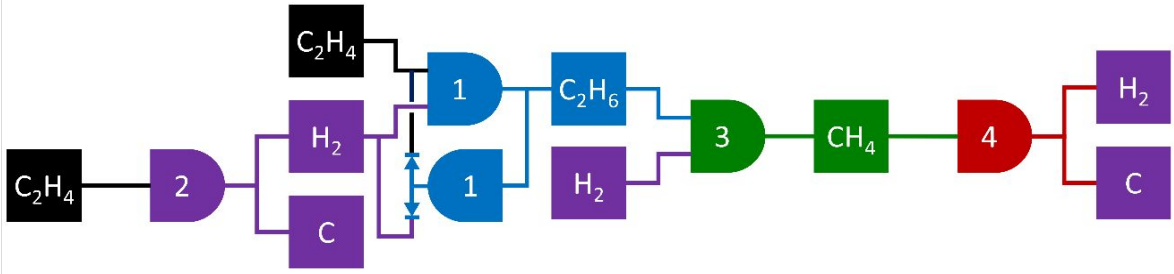

Figure 4: BRC corresponding to the evolved RNG of Figure 3. C_2H_4 is fed as the leftmost value node. Each "and" gate (reaction) has a unique color. The color of a species value node corresponds to the "and" gate that first generates it. While using single value nodes for each species is more structurally accurate, we allow repeat value nodes for visual clarity. R1 is the only reversible reaction as inferred from thermodynamics (Section S1 in the SI), which is implemented using diodes.

PGMs are more complex. They also employ a mathematical one-to-one transformation of the RNG's evolution but only require the final evolved state. A strength of PGMs is their focus on reaction extents, similar to the incremental approach for modeling reaction kinetics.[27, 28, 30] In contrast to other approaches,[23, 35-40] the extents do not assume a rate equation or information about the rates. It relates directly to experimental measurements of species concentration changes. The generalized formula relating the vector of reaction extents (ξ_i) to the vector of net changes in species j concentration ($\Delta[C]_j$) depends upon the stoichiometric matrix $A_{i\times j}$, as shown in Eq. (2). The number of rows and columns reflect the species j and reactions i, respectively. An example of this formulation is displayed in Figure 5a.

$$\underline{A}_{i\times j} \cdot \underline{\xi}_i = \underline{\Delta[C]}_j \quad (2)$$

This formulation is similar to the traditional method of rates balancing, i.e., relating species rates (w_j) to reaction rates (r_j) as shown in Eq (3).

$$\underline{A}_{i \times j} \cdot \underline{w}_i = \underline{r}_j \quad (3)$$

With noise-free experimental data and a full rank stoichiometric matrix, the solution of Eq. (2) would be exact, and linear algebra would identify the reactions: reactions with non-zero extents occur, and reactions with extents of zero do not. However, significant noise generally accompanies laboratory measurements, so we require a statistically rigorous evaluation using PGMs. This is the first challenge in analyzing experimental data. This noise of the measured concentration changes is generally evenly distributed around values that fall within the column space of the stoichiometric matrix; systematic noise (i.e., a significant offset from the column space) indicates there may be an issue measuring species and requires additional analysis, as demonstrated in Section 3.2.

Our PGM approach uses the same core formulation of Eq. (2) but implements a Bayesian framework to provide probability distributions for each reaction extent.[42, 43, 54] We constrain the prior distributions of each reaction extent with the reversibility of the reaction (a prior probability bias) inferred by thermodynamics. The PGM expresses Eq. (2) with statistical deviation from noisy experimental measurements (σ), as shown in Eq. (4), where P() denotes a probability.

$$P(\underline{\xi}_i, \sigma | \underline{\Delta[C]}_j) \propto P(\underline{\Delta[C]}_j | \underline{\xi}_i, \sigma) \times P(\underline{\xi}_i, \sigma) \quad (4)$$

The Bayesian formulation is expressed visually with a probabilistic graphical model (Figure 5a) displaying the causal relations of the reaction extents changing the species concentrations. The posterior distribution of the reaction extents accounts for the correlations between the estimated extents while probabilistically quantifying their estimated values. These probability distributions enable statistical conclusions regarding which reactions occur.

To evaluate the PGM information rating, we calculate the rank of the stoichiometric matrix. If the matrix is full rank, the PGM analysis promises a fully determined result with an information rating equal to the rank. However, if the rank of the stoichiometric matrix is less than the number of reactions, then the extents are linearly dependent with infinite linear combinations. This results in an underdetermined PGM with extents not probabilistically quantifiable. This is the second routine challenge in chemical kinetics.

With reaction stoichiometries often possessing linear dependence, obscured identifiability is common in kinetics, as demonstrated in Figure 5a. The incremental approach, i.e., decomposing measurements to individual rates, addresses this concern by assuming an identifiable set of reactions from the possible linear combinations.[30] For the example in Figure 5a, the stoichiometric matrix is underdetermined; the rank is three while there are four reactions. Therefore, the reaction extents cannot be solved for uniquely because the reactions' stoichiometries are linearly dependent. For example, suppose ΔC_2H_6, ΔC_2H_4, ΔCH_4, and ΔH_2 are measured as -6M, 1M, 4M, and 8M, respectively. These changes correspond exactly to values of 3M, 2M, 3M, and 2M for the extents of ξ_1, ξ_2, ξ_3, and ξ_4, respectively. However, the measured changes in species also correspond to extents of 4M, 3M, 2M, and 0M, wherein R4 does not occur. These scenarios are indistinguishable given the underdetermined structure. To ensure unique solutions for each reaction extent, the incremental approach might assume R1, R2, and R3 occur while R4 does not; this would eliminate the linear dependence, but at the cost of an inaccurate RNI since R4 does occur. To avoid this inaccuracy, we must create a different approach for analyzing underdetermined systems.

Leveraging PGMs, we propose a new approach. If an evolved RNG shows an active species node with only one active flux edge directed into it, then a single reaction contributes to this species' generation. Similarly, if an evolved RNG shows a fed species node with only one active

flux edge directed away from it, then there is a single reaction responsible for this species' consumption. In these cases, the physics offers only one possible cause to explain the observed effect, and the evolved RNGs indicate that PGM analysis can identify these reactions, despite the PGM being underdetermined.

In our example, the stoichiometric matrix (Figure 5a) rank is three, indicating the PGM is underdetermined. However, the production of C_2H_6 and CH_4 (Figure 3) can only be causally explained by R1 and R3, respectively. Therefore, this underdetermined PGM possesses an information rating of 2 (i.e., revealing 2 reactions). We validate this information rating by performing a hypothetical underdetermined PGM analysis (Figure 5b-e) using simulated effluent concentrations from the proposed experiment of feeding C_2H_4; simulated data is available in the data repository accompanying this work.[49] The resulting distributions confirm that R1 and R3 have statistically non-zero extents, while the extents of R2 and R4 cannot be discerned from zero due to the identifiability issue.

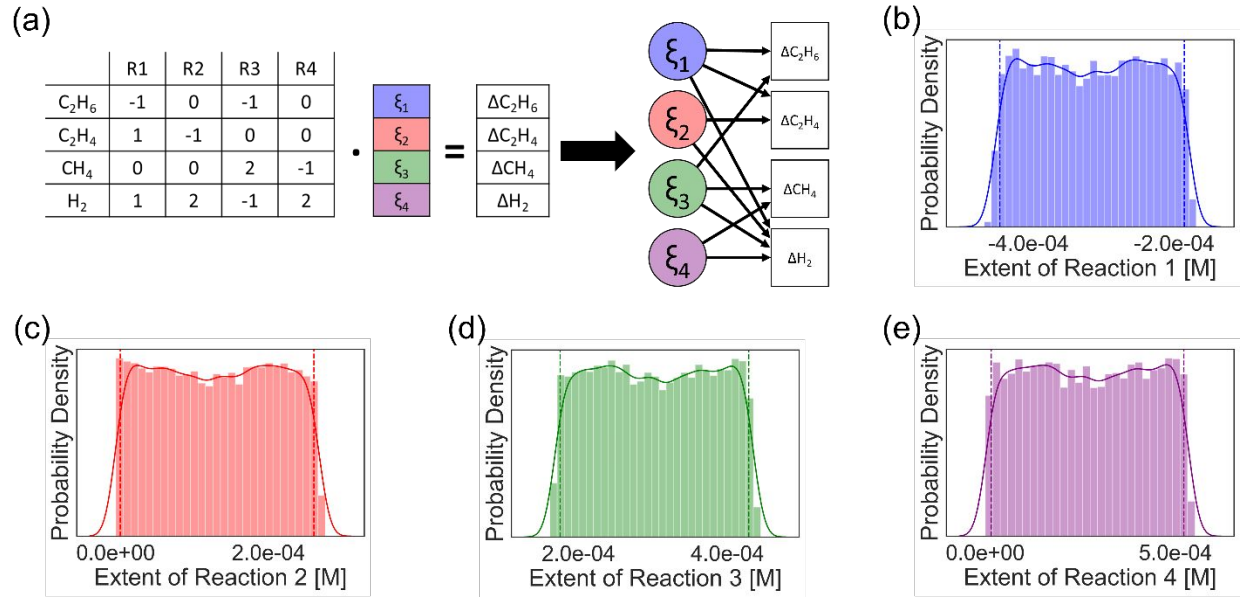

Figure 5: PGM analysis of the evolved RNG of Figure 3. (a) Matrix and PGM formulation of the reaction extents. The rows and columns of the matrix are labeled with their corresponding species and reactions. The same stoichiometric relationships described in the matrix formulation are embedded within the PGM edges connecting the corresponding nodes. The standard deviation of the experimental noise (σ) is omitted for clarity. (b-e) Probability distribution and histogram of extent of R1-R4, respectively. Histograms are included for clarity in b-e to show for R2 and R4 that no extents less than zero are included in the estimations following their irreversibility. Vertical dashed lines in b-e denote the 95% credible intervals, which do not overlap with zero for R1 and R3.

Predicting the information rating of hypothetical experiments from the BRC and PGM analyses screens proposed experiments to enable DOE for RNI. With these information ratings, the experiment is conducted and the resulting data is processed with BRC or PGM analysis to confirm the reactions. These hypothetical experiments also provide an initial guess of how many experiments will be required for RNI by cataloging which reactions can be identified using PGM or BRC analysis and which require other methods such as delplot analysis. Reaction networks with linear dependence and high connectivity are less compatible with PGM and BRC analysis. However, this initial guess should be treated with caution. As experiments are conducted and reactions are eliminated from consideration, the RNG structure is updated. DOE is performed upon the updated RNG, and new informative experiments are identified, as demonstrated in Section 3.3. Leveraging the strength of this active learning approach, the methodology consistently identifies

the full reaction network in fewer experiments than predicted by the initial guess in our experience; therefore, we recommend using such initial guesses as a qualitative metric prone to overestimation.

To summarize, this simulated ethane dehydrogenation network is identified with our RNI methodology using only two experiments. We feed C_2H_4 and use BRC to confirm the occurrence of R1, R2, and R3 (Figure 4); the underdetermined PGM analysis also statistically confirms R1 and R3 as active (Figure 5). The second experiment, feeding CH_4, confirms the occurrence of R4 by BRC analysis once H_2 is measured at the effluent. From only two sets of experimental measurements, all four reactions composing the underdetermined system are statistically and logically discerned as active.

2.5 RNI for CO_2-Assisted Ethane Dehydrogenation

An additional, more complex RNI example is provided in Section S4 of the SI upon a simulated system of CO_2-assisted ethane dehydrogenation network. With fifteen proposed reactions in the RNG, seven of which occur, the reaction network is challenging to identify. From noisy concentration data of our simulated reactor's outlet, we attempt to discern the active reactions using fully determined and underdetermined PGMs, detailed BRCs, and supplementary delplots. Specifically, we use simulated effluent concentrations from twenty experiments for PGM analysis (five feeds each measured four times), eight experiments for BRC analysis (two feeds each measured four times), and sixty-four experiments for delplot analysis (four feeds each measured four times at four conversions). The PGM and BRC analyses inform the RNG of nine reactions from their twenty-eight experiments, while the sixty-four delplot experiments inform the RNG of the remaining eight reactions. This highlights the significant reduction of number of experiments necessary for PGM and BRC analyses compared to delplots, which are only employed when all informative PGM and BRC experiments are exhausted. We validate our RNI methodology by correctly identifying the reactions present in the underlying model generating the effluent data (Figure 6).

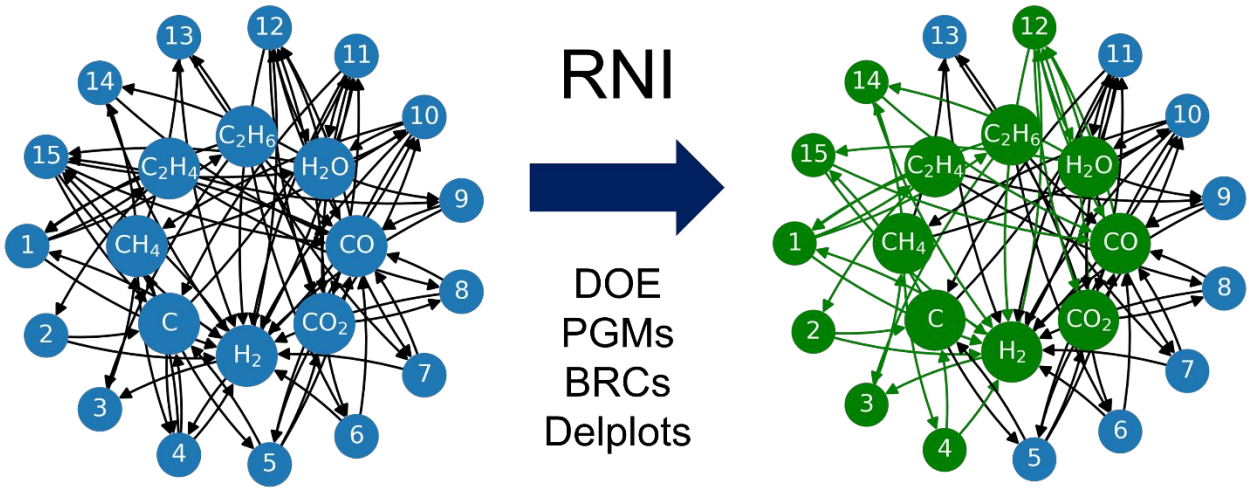

Figure 6: Overview of the RNI performed on the CO_2-assisted ethane dehydrogenation network. From a proposed RNG of fifteen reactions, the correct seven are identified (green) from the effluent measurements collected as directed by our DOE and analyzed by PGMs, BRCs, and delplots. Details are provided in Section S4 of the SI.

3. RNI for the Cross-Ketonization of Furoic and Lauric Acids

3.1 Reaction Network Introduction

Herein, we demonstrate our RNI methodology upon a cross-ketonization network with laboratory data. This reaction produces a precursor toward the production of detergents, analogous to linear alkylbenzene sulfonates (LAS). LAS are the world's largest biodegradable surfactant commodity commonly used in cleaners, detergents, and laundry powders.[55] The global market for LAS was estimated at $6.74 billion in 2011 and was projected to be over $8 billion in 2020.[56] However, their continued use is undesirable due to their dependence upon non-renewable feedstocks and toxicity concerns.[57, 58]

Recent work has demonstrated biomass-derived oleo-furan sulfonate surfactants outperforming commercial LAS in detergency and showing greater stability in hard water conditions.[59] Targeting the bottleneck reaction in their production, Nguyen et al. developed a new synthesis of 2-alkoyl furan via cross-ketonization of 2-furoic acid (FA) and lauric acid (LA) using earth-abundant, commercially-available iron oxides.[60] A promising alternative to LAS, these oleo-furan sulfonate surfactants suffered from a low yield of 41% due to competitive side reactions. Chen et al. applied similar chemistry at milder reaction temperatures to synthesize short-chain alkyl furan ketones in higher yields of 80% with a nanoparticle magnesium oxide (MgO) catalyst.[61] While reaction networks have been proposed, rigorous RNI has not been performed.

3.2 The Initial Experiment and RNG Construction

An initial experiment over MgO is performed at 350 °C at 0.25 M FA and 0.15 M LA. Experimental methods are detailed in Section S5 of the SI and are the same as used in prior work.[62] Effluent analysis with the GCMS reveals in the liquid phase 2-dodecanoyl furan (Dod), 12-tricosanone (Tri), and furan (F). Prior work identified carbon dioxide (CO_2) and water (H_2O) as additional products.[60] With stables species identified, reaction enumeration commences.

Combinatorically listing all stoichiometrically feasible reactions,[63] we find six reactions with two or fewer reactants. The six reactions are diagrammed in Figure 7 with arrows indicating thermodynamic reversibility, as evaluated from reaction quotients and equilibrium constants estimated by the machine learning toolkits of Green and coworkers (details in Section S1 of the SI).[64-67] From the stoichiometry matrix (Eq. (2)) of the six potential reactions, the rank is evaluated to be three, indicating the system is underdetermined.

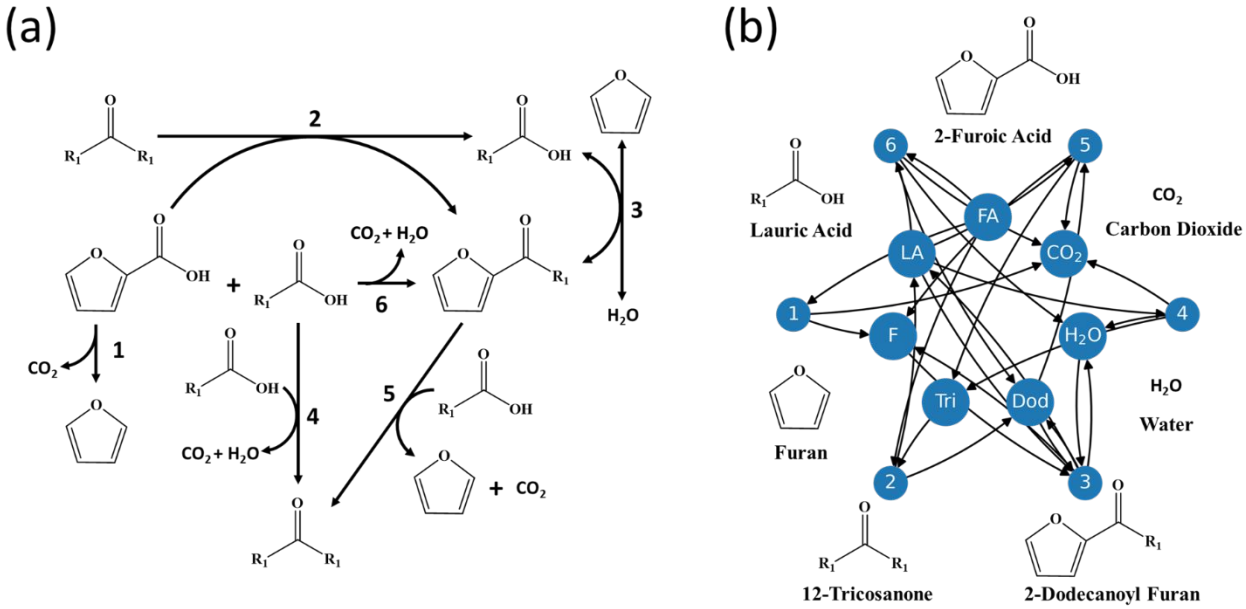

Figure 7: Reaction network depicting potential overall reactions between 2-furoic acid (FA), lauric acid (LA), 2-dodecanoyl furan (Dod), 12-tricosanone (Tri), and furan (F) over MgO. Thermodynamic reversibility is indicated by the directionality of the reactive flux arrows. (a) Reaction network diagramed with each reaction represented with a labeled arrow. (b) Reaction network graph with each reaction represented with a labeled node. Labeled chemical structures are on the periphery. The reversibility of R3 is shown using arrows in both directions rather than bidirectional arrows for convenience.

First, we analyze the initial experimental data to ensure the measurements fall within the stoichiometry matrix column space. Specifically, we examine the measured effluent concentrations of three repeat experiments with a feed of FA and LA, as described above. If the concentration changes deviate strongly from the stoichiometric subspace, this indicates a mass balance violation stemming from a calibration issue or an unaccounted reaction, e.g., coke or humins formation. Subspace evaluation is performed using a full rank PGM; with a stoichiometric matrix of rank three, we select the three linearly independent reactions R1, R4, and R6 to define the PGM structure (Figure 8a).

Evaluating the PGM, we identify a significant offset in the measured change of FA (Figure 8e); specifically, there is more FA consumed than can be explained by the generated products. FA and LA form complexes with the catalyst, preventing their quantification in the effluent. We classify these species' concentrations as "hidden data," denoted by rounded edges in the PGM (Figure 8a). This is yet another challenge in RNI common in catalysis. Reactive species and intermediates often preferentially adsorb and block active sites, leading to undetected products, mass imbalances, or catalytic deactivation. Deactivation will not introduce errors in our rate agnostic analyses, and errors from mass imbalances caused by undetected products or complexation with catalytic sites are mitigated by this check on hidden data. While not measured in the effluent, these undetected species can be determined through PGM analysis; using the estimated extents of reactions, the unmeasured species FA, LA, H_2O, and CO_2 are inferred. The reminder of our analysis depends only upon the measured concentrations of Dod, Tri, and F, showing the adaptability of the RNI methodology in the context of physical constraints.

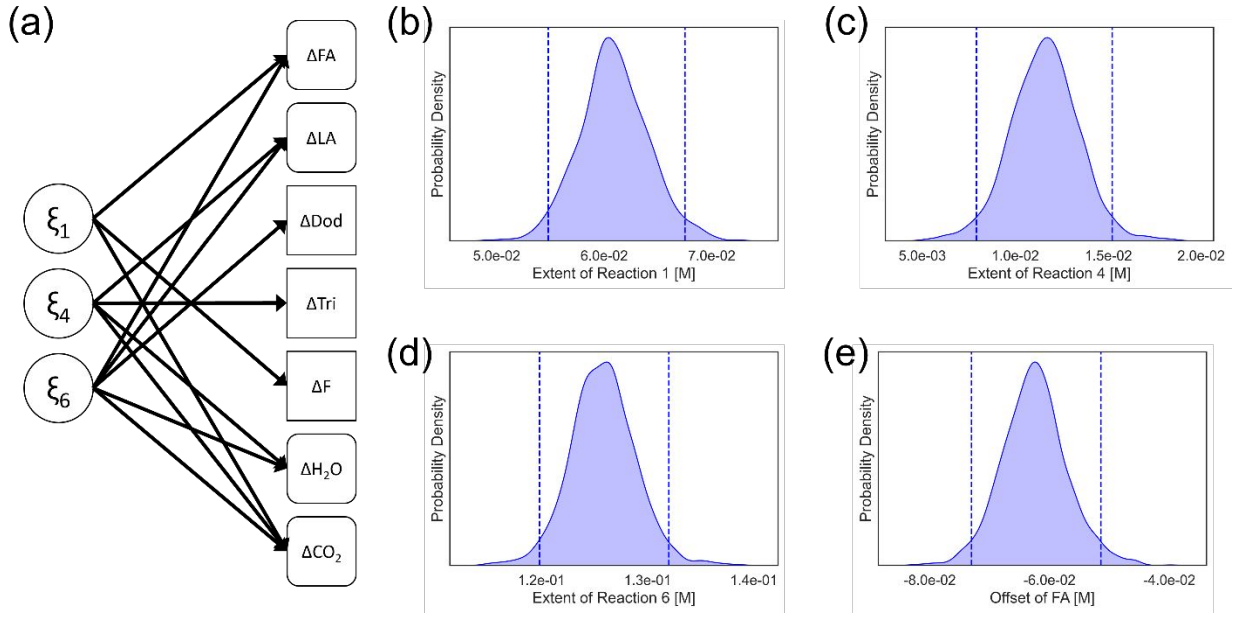

Figure 8: PGM analysis of an initial experiment to perform a stoichiometry check. (a) PGM structure. Reactions extents (ξ_i) are parent nodes (circles) and change in species concentrations are child nodes (squares). Unmeasured species nodes are rounded to denote them as hidden data. (b)-(d) Probability distribution of extent of R1, R4, and R6 with dashed vertical lines denoting the 95% credible intervals. (c) Probability distribution of the offset of FA with dashed vertical lines denoting the 95% credible interval. The credible interval does not overlap with zero indicating a statistically significant mass imbalance in FA consumed.

This physical example of concentration data with significant offsets highlights two concerns of relying upon more traditional fittings of different proposed rate models for RNI. First, if unsupervised fittings of proposed rate equations are implemented without evaluating how well the measured concentration changes fall within the stoichiometric subspace, offsets such as this

one caused by reactants complexing with the catalyst would not be detected and fitted rate models would be inaccurate. At best, these rate models would over-predict the rate of reactions consuming FA; at worst, these unsupervised fits might select a different and inaccurate set of reactions that better explain the non-physical concentration offset. Second, even if the offset is identified from data supervision before rate model fitting, it can be challenging to discern if the offset is statistically significant or can be discounted as noise (i.e., if the offset is a fundamental issue requiring further investigation, or if the concentration data can be used as is by relying on the fitting to provide an average within the noise). This approach we present in Figure 8 demonstrates how both concerns can be overcome: implementing a PGM to provide a Bayesian quantification of the statistical significance of potential offsets.

3.3 DOE and RNG Active Learning

To determine the first optimal experiment, we perform DOE. Every combination of one or two stable species are proposed as a feed; each of these feeds is used to simulate an RNG evolution trajectory, which in turn yields an information rating for the proposed experiment. All proposed experiments feeding CO_2 or Dod are eliminated from consideration to reflect the physical constraint that the experimental setup and supplies cannot facilitate feeding these species.

We determine that the highest information rating is one, i.e., even the most informative experiment can only guarantee the identification of a single reaction. From multiple proposed experiments with information ratings of one, we choose a feed composed of FA and Tri to confirm R2. We perform this experiment three times for statistical confidence. The results of the experiment are conclusive: no Dod is measured in the effluent. Thus, the BRC analysis confirms R2 is inactive; R2 is removed from the RNG.

With an updated RNG, we perform a new round of DOE in an active learning approach. While again no proposed experiment has an information rating higher than one, there is a surprising finding: the previous experiment can now identify R1. In the previous RNG including R2, the feed of FA and Tri led to an RNG evolution where multiple reactions produce F. In the updated RNG excluding R2, the feed of FA and Tri leads to an RNG evolution where only R1 produces F. This result emphasizes the importance of performing DOE after every RNG update, in an active learning framework.

We reanalyze the previously collected experimental data rather than obtain new data, as suggested by the DOE. BRC analysis of the previous results of feeding FA and Tri now confirms R1 occurs since significant F is produced. A PGM analysis confirms it as shown in Figure 9; R1 is active (non-zero extent) and R2 is inactive (zero extent).

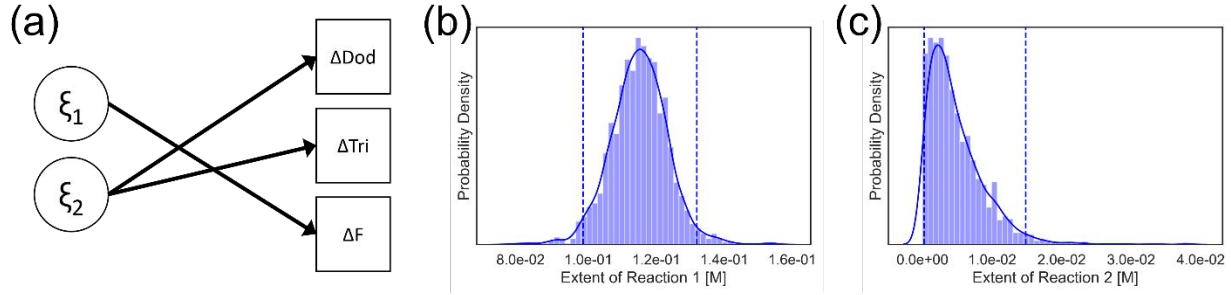

Figure 9: PGM analysis of feeding FA and Tri over MgO. (a) PGM structure of measurable species assuming only R1 and R2 may occur. (b) Probability density of extent of R1. (c) Probability density of extent of R2. The 95% credible intervals denoted by dashed vertical lines indicate that ξ_2 cannot be statistically distinguished from zero since the left bound is zero. While there is some probability density of a non-zero ξ_2 value, we attribute this to experimental noise in mass loss of Tri rather than activity of R2.

With no new reactions excluded, there is no need to run new DOE; the RNG has not been changed. From the previous round of DOE, there are multiple proposed experiments with information ratings of one. We select the feed of LA and F and perform the experiment in triplicate. BRC analysis confirms R3 does not occur since no Dod is produced.

With the elimination of R3, our next DOE active learning cycle is conducted: the RNG is updated and DOE is performed. Again, the DOE identifies that a previous experiment can be reevaluated in the context of the updated RNG; without R3 occurring, only R4 can possibly produce Tri from a feed of LA and F. With significant Tri measured in the effluent of the previous experiment, BRC analysis confirms R4 occurs. This conclusion is supported by additional PGM analysis in Figure 10 showing that the extent of R4 can be statistically distinguished from zero while the extent of R3 cannot.

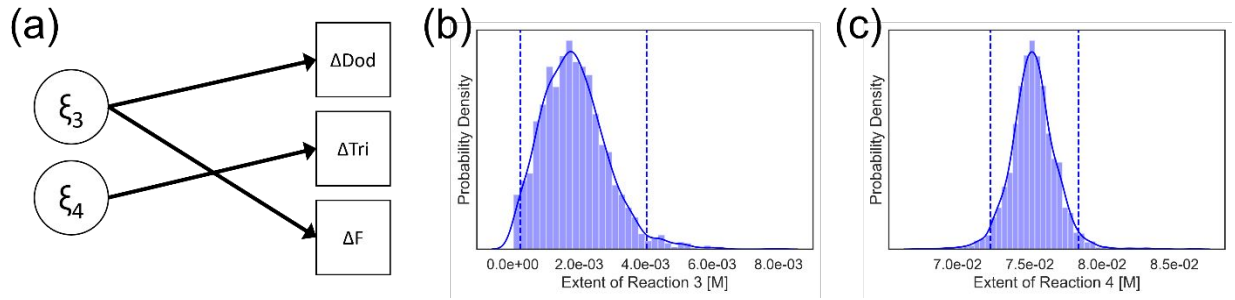

Figure 10: PGM analysis of feeding LA and F over MgO. (a) PGM structure of measurable species assuming only R3 and R4 may occur. (b) Probability density of extent of R3. (c) Probability density of extent of R4. The 95% credible intervals denoted by dashed vertical lines indicate that ξ_3 cannot be statistically distinguished from zero since the left bound is zero. While there is some probability density of a non-zero ξ_3 value, we attribute this to experimental noise in mass loss of F rather than activity of R3.

Revisiting the proposed experiments' information ratings from the DOE, R6 can be identified by feeding LA and FA; according to BRC analysis, if Dod is produced, R6 occurs but if Dod is absent, then R6 does not. Fortunately, we can again return to a previous experiment, this time the initial experiment, for the required data. Dod is measured in significant quantities, confirming the occurrence of R6 within the system.

Having confirmed the occurrence of R1, R4, and R6 along with the inactivity of R2 and R3, only R5 remains unconfirmed. However, our DOE indicates there are no experiments that can identify R5 using BRC or PGM analysis. Therefore, we must employ an alternate technique. We select the delplot method[7, 8] instead of other options such as isotopic labeling or probe molecule co-feed experiments[20] because delplots can be constructed without requiring additional species be procured.

A first rank delplot is constructed from reactor effluent concentration data collected at low conversions. To create this delplot, the product species' selectivity (Eq. (5)) is plotted against a reactant's conversion (Eq. (6)) relative to the reactant's initial concentration (R_0), where P and R are the product and reactant species' effluent concentrations respectively.

$$y = \text{selectivity} = \frac{P/R_0}{1 - R/R_0} \tag{5}$$

$$x = \text{conversion} = 1 - R/R_0 \tag{6}$$

Applying Wojciechowski's criteria[18, 19] to such a delplot can identify reactions: a negative slope indicates the product species is consumed in a later reaction while a slope of zero indicates there are no secondary consumption reactions.[20] Delplot analysis is less preferred to BRC or PGM analysis because it is less statistically reliable: its conclusion depends upon the evaluation of a line's slope, which is more susceptible to experimental noise. Additionally, delplot analysis requires more experiments to construct the plot at a range of conversions. However, when BRC and PGM analyses cannot identify specific reactions, delplot analysis often can.

Feeding FA and LA and evaluating the effluent's Dod concentration at six different reaction times, we construct the delplot shown in Figure 11a. Given the potential reaction network diagramed in Figure 7, if R5 occurs then Dod will be consumed in a secondary reaction and have a negative delplot slope. Performing Bayesian inference upon the delplot, we determine the distribution of best fit lines (visualized in Figure 11a) and the probability distribution of the slope (Figure 11b). With a slope indistinguishable from zero, there is no evidence of the consumption of Dod with increasing conversion. R5 is eliminated from consideration, and the RNG is fully identified. Our RNI methodology required 15 total experiments to be conducted (including the triplicate experiments): the initial experiment (3), feeding FA and Tri (3), feeding LA and F (3), and feeding FA and LA (6). This number could have been reduced further since BRCs do not require replicates and the delplot could have been constructed with only three measurements instead of the nine we chose. However, to demonstrate our PGM analysis and improve the statistical confidence of our conclusions, we elected to include more replicates.

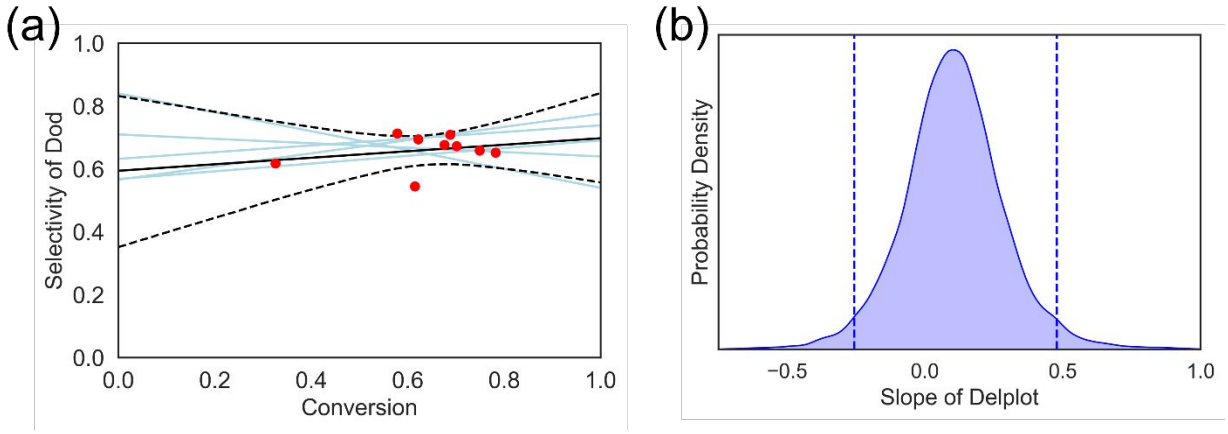

Figure 11: Delplot analysis of Dod for a feed of FA and LA over MgO. (a) Delplot showing the selectivity of Dod versus the conversion of FA (red data points). Consistent with delplot construction, selectivity is calculated as the yield of Dod divided by the conversion of FA. Black dashed lines indicate the 95% credible intervals of the best fit line for the data; from the sampled best fit lines, five were randomly selected and plotted in light blue for visualization. (b) Probability distribution of the slope of the delplot. With the 95% credible intervals (vertical dashed lines) encompassing zero, the slope cannot be statistically distinguished from zero and there is no evidence of the occurrence of R5.

3.4 Kinetic Modeling

With the reaction network identified (Figure 12a) and confirmation of no mass transfer limitations obtained (see Section S5 in the SI), a kinetic model can be constructed. Assuming power law kinetics and fitting to the delplot concentration data with Bayesian inference,[54] we construct the kinetic model detailed in Eq. (7). With an R^2 value of 0.94 and mean absolute error of 0.005 M, the model agrees well with the limited data (Figure 12b).

$$\begin{aligned} \text{rate}_1 &= k_1[\text{FA}] \quad \text{where } k_1 = 3.03 \times 10^{-3}\ \text{s}^{-1} \\ \text{rate}_4 &= k_4[\text{LA}]^2 \quad \text{where } k_4 = 6.66 \times 10^{-2}\ \text{M}^{-1}\text{s}^{-1} \\ \text{rate}_6 &= k_6[\text{FA}][\text{LA}] \quad \text{where } k_6 = 2.18 \times 10^{-1}\text{M}^{-1}\text{s}^{-1} \end{aligned} \tag{7}$$

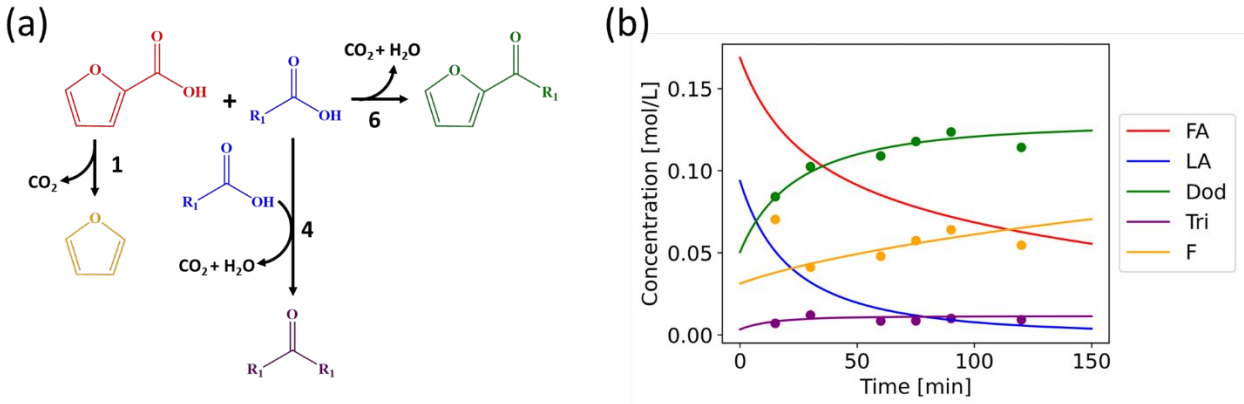

Figure 12: Reaction network model. (a) Identified reaction network with R1, R4, and R6. (b) Concentration profiles fit to data with colors corresponding to panel a. The hidden data species of FA and LA are inferred and plotted, but no data can be measured for comparison since these species complex with the catalyst.

Subsequent model refinement DOE and kinetic parameter estimation must be performed to obtain a reliable kinetic model for optimization, for which there are multiple developed techniques.[68-71] As more data is collected and the model is refined, the rate equation form may change, the reaction orders will be reassessed, the rate constants will update, and temperature dependence will be incorporated. However, the reactions being modeled will not change unless significantly new regimes are explored (i.e., temperatures or concentrations change enough to alter which reactions occur). The RNI methodology provides statistical confidence in the identified network establishing the foundation for future process optimization of this cross-ketonization network.

4. Conclusions

We constructed and demonstrated a methodology for reaction network identification. Using stoichiometry, chemical knowledge, and thermodynamics, we combinatorically enumerated possible reactions. Leveraging graph theory with rate equation agnosticism, we created a DOE approach of optimal experiments for identifying which of these reactions occur, simulating hypothetical reactions and ranking them on their information potential. Once run, these experiments' effluent concentrations were analyzed with BRCs and PGMs, methods we introduced for discerning reactions in complex networks. When feasible, these methods are preferred alternatives to delplots as our examples showed that they provide comparable reaction identification with at least a factor of two fewer experiments. The analysis identified a reaction, the RNG was updated, and a new iteration of DOE selected the next maximally informative experiment in an active learning framework that continued until the entire RNG is identified. We introduced our RNI methodology upon a simulated ethane dehydrogenation network, and we further validated it upon a simulated CO_2-assisted ethane dehydrogenation network and a biomass-derived cross-ketonization network in the laboratory. For the cross-ketonization network with six potential reactions, we showcased how all can be identified using minimal experiments. Upon confirming this RNG, a simple rate model was constructed showing good agreement with

measured concentrations. These results were obtained within the physical constraints of the laboratory system including unquantifiable species, linearly dependent reactions, and experimental noise. This reaction network methodology is general, providing a path forward for accurate reactor simulations and efficient automated laboratories in future catalysis research.

5. Associated Content

A supporting information document includes additional data, examples, and discussions to support the work presented in this article.

The data repository[49] contains all code and supporting analysis presented in this article and the accompanying supporting information document. Example scripts can easily be modified for the users' own work. This repository will be made publicly available online at the time of publication.

6. Author Contributions

Maximilian Cohen: Data curation, formal analysis, methodology, software, validation, visualization, writing – original draft. Tejas Goculdas: Data curation, investigation, validation, visualization, writing – original draft. Dionisios Vlachos: Conceptualization, funding acquisition, project administration, resources, supervision, writing – review & editing.

7. Conflicts of Interest

There are no conflicts to declare.

8. Acknowledgments

We acknowledge support from the RAPID manufacturing institute, supported by the Department of Energy (DOE) Advanced Manufacturing Office (AMO), award number DE-EE0007888-9.5. RAPID projects at the University of Delaware are also made possible in part by funding provided by the State of Delaware. The Delaware Energy Institute gratefully acknowledges the support and partnership of the State of Delaware in furthering the essential scientific research being conducted through the RAPID projects.

The experimental work was supported as part of the Catalysis Center for Energy Innovation, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under award number DE-SC0001004.

The authors acknowledge valuable discussions with Professor Babatunde Ogunnaike and Professor Markos Katsoulakis.

9. References

1. G. N. Simm and M. Reiher, *Journal of Chemical Theory and Computation*, 2017, **13**, 6108-6119.

2. R. Vinu and L. J. Broadbelt, *Annual Review of Chemical and Biomolecular Engineering*, 2012, **3**, 29-54.
3. R. Weingarten, J. Cho, R. Xing, W. C. Conner Jr and G. W. Huber, *ChemSusChem*, 2012, **5**, 1280-1290.
4. T. D. Swift, C. Bagia, V. Choudhary, G. Peklaris, V. Nikolakis and D. G. Vlachos, *ACS Catalysis*, 2014, **4**, 259-267.
5. P. Desir, B. Saha and D. G. Vlachos, *Energy & Environmental Science*, 2019, **12**, 2463-2475.
6. D. Otyuskaya, J. W. Thybaut, R. Lødeng and G. B. Marin, *Energy & Fuels*, 2017, **31**, 7082-7092.
7. N. A. Bhore, M. T. Klein and K. B. Bischoff, *Industrial & Engineering Chemistry Research*, 1990, **29**, 313-316.
8. M. T. Klein, Z. Hou and C. Bennett, *Energy & Fuels*, 2012, **26**, 52-54.
9. J. Jae, W. Zheng, R. F. Lobo and D. G. Vlachos, *ChemSusChem*, 2013, **6**, 1158-1162.
10. J. Fu, E. S. Vasiliadou, K. A. Goulas, B. Saha and D. G. Vlachos, *Catalysis Science & Technology*, 2017, **7**, 4944-4954.
11. B. Antwi Peprah, O. Brown, J. M. Stryker and W. C. McCaffrey, *Energy & Fuels*, 2020, **34**, 16532-16541.
12. Y. Dong, C. Georgakis, J. Mustakis, J. M. Hawkins, L. Han, K. Wang, J. P. McMullen, S. T. Grosser and K. Stone, *AIChE Journal*, 2019, **65**, e16726.
13. L. Bui, R. Chakrabarti and A. Bhan, *ACS Catalysis*, 2016, **6**, 6567-6580.
14. J. H. Miller and A. Bhan, *ChemCatChem*, 2018, **10**, 5242-5255.
15. J. H. Miller and A. Bhan, *ChemCatChem*, 2018, **10**, 5511-5522.
16. L. Bui and A. Bhan, *Applied Catalysis A: General*, 2017, **546**, 87-95.
17. L. Bui and A. Bhan, *Applied Catalysis A: General*, 2018, **564**, 1-12.
18. T. M. John and B. W. Wojciechowski, *J. Catal.*, 1975, **37**, 240-250.
19. D. Best and B. W. Wojciechowski, *J. Catal.*, 1977, **47**, 11-27.
20. J. H. Miller, L. Bui and A. Bhan, *Reaction Chemistry & Engineering*, 2019, **4**, 784-805.
21. G. E. P. Box and W. J. Hill, *Technometrics*, 1967, **9**, 57-71.
22. W. G. Hunter and A. M. Reiner, *Technometrics*, 1965, **7**, 307-323.
23. G. F. Froment, *AIChE Journal*, 1975, **21**, 1041-1057.
24. G. Buzzi Ferraris, P. Forzatti, G. Emig and H. Hofmann, *Chemical Engineering Science*, 1984, **39**, 81-85.
25. G. Buzzi-Ferraris and P. Forzatti, *Chemical Engineering Science*, 1983, **38**, 225-232.
26. S. Srinivasan, J. Billeter and D. Bonvin, *AIChE Journal*, 2019, **65**, 1211-1221.
27. A. Bardow and W. Marquardt, *Chemical Engineering Science*, 2004, **59**, 2673-2684.
28. N. Bhatt, N. Kerimoglu, M. Amrhein, W. Marquardt and D. Bonvin, *Chemical Engineering Science*, 2012, **83**, 24-38.
29. N. Bhatt, M. Amrhein and D. Bonvin, *Industrial & Engineering Chemistry Research*, 2011, **50**, 12960-12974.
30. M. Brendel, D. Bonvin and W. Marquardt, *Chemical Engineering Science*, 2006, **61**, 5404-5420.
31. S. Srinivasan, J. Billeter and D. Bonvin, *Industrial & Engineering Chemistry Research*, 2016, **55**, 8034-8045.
32. J. Billeter, S. Srinivasan and D. Bonvin, *Analytica Chimica Acta*, 2013, **767**, 21-34.
33. K. Villez, J. Billeter and D. Bonvin, *Processes*, 2019, **7**, 75.

34. A. Mašić, J. Billeter, D. Bonvin and K. Villez, *IFAC-PapersOnLine*, 2017, **50**, 3929-3934.
35. J. Tsu, V. H. G. Díaz and M. J. Willis, *Computers & Chemical Engineering*, 2019, **121**, 618-632.
36. C. J. Taylor, H. Seki, F. M. Dannheim, M. J. Willis, G. Clemens, B. A. Taylor, T. W. Chamberlain and R. A. Bourne, *Reaction Chemistry & Engineering*, 2021, **6**, 1404-1411.
37. C. J. Taylor, M. Booth, J. A. Manson, M. J. Willis, G. Clemens, B. A. Taylor, T. W. Chamberlain and R. A. Bourne, *Chemical Engineering Journal*, 2021, **413**, 127017.
38. C. J. Taylor, J. A. Manson, G. Clemens, B. A. Taylor, T. W. Chamberlain and R. A. Bourne, *Reaction Chemistry & Engineering*, 2022, **7**, 1037-1046.
39. W. G. Hunter and R. Mezaki, *AIChE Journal*, 1964, **10**, 315-322.
40. Z. T. Wilson and N. V. Sahinidis, *Computers & Chemical Engineering*, 2019, **127**, 88-98.
41. G. Franceschini and S. Macchietto, *Chemical Engineering Science*, 2008, **63**, 4846-4872.
42. J. Feng, J. L. Lansford, M. A. Katsoulakis and D. G. Vlachos, *Science Advances*, 2020, **6**, eabc3204.
43. D. Koller, *Probabilistic Graphical Models : principles and techniques*, Massachusetts Institute of Technology, USA, 2009.
44. C. Nunn, A. DiPietro, N. Hodnett, P. Sun and K. M. Wells, *Organic Process Research & Development*, 2018, **22**, 54-61.
45. T.-C. Kuo, N. A. Malvadkar, R. Drumright, R. Cesaretti and M. T. Bishop, *ACS Combinatorial Science*, 2016, **18**, 507-526.
46. W. Chen, M. Cohen, K. Yu, H.-L. Wang, W. Zheng and D. G. Vlachos, *Chemical Engineering Science*, 2021, **237**, 116534.
47. Z. Yang, H. Li, H. Zhou, L. Wang, L. Wang, Q. Zhu, J. Xiao, X. Meng, J. Chen and F.-S. Xiao, *J. Am. Chem. Soc.*, 2020, **142**, 16429-16436.
48. H. Saito and Y. Sekine, *RSC Advances*, 2020, **10**, 21427-21453.
49. M. Cohen, T. Goculdas and D. G. Vlachos, *Journal*, 2022, Mendeley Data, V1, doi: 10.17632/86vkrpvbr4.1.
50. S. Rangarajan, T. Kaminski, E. Van Wyk, A. Bhan and P. Daoutidis, *Computers & Chemical Engineering*, 2014, **64**, 124-137.
51. L. J. Broadbelt, S. M. Stark and M. T. Klein, *Industrial & Engineering Chemistry Research*, 1994, **33**, 790-799.
52. A. Ratkiewicz and T. N. Truong, *Journal of Chemical Information and Computer Sciences*, 2003, **43**, 36-44.
53. F. P. Di Maio and P. G. Lignola, *Chemical Engineering Science*, 1992, **47**, 2713-2718.
54. B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li and A. Riddell, *2017*, 2017, **76**, 32.
55. J. J. Wang, Y. Y. Chuang, H. Y. Hsu and T. C. Tsai, *Catalysis Today*, 2017, **298**, 109-116.
56. personal communication.
57. P. p. b. ChemSystems, 2009, 1-7.
58. *US Pat.,* US 10906857 B2, 2021.
59. D. S. Park, K. E. Joseph, M. Koehle, C. Krumm, L. Ren, J. N. Damen, M. H. Shete, H. S. Lee, X. Zuo, B. Lee, W. Fan, D. G. Vlachos, R. F. Lobo, M. Tsapatsis and P. J. Dauenhauer, 2016, DOI: 10.1021/acscentsci.6b00208.
60. H. Nguyen, Y. Wang, D. Moglia, J. Fu, W. Zheng, M. Orazov and D. G. Vlachos, *Catalysis Science & Technology*, 2021, DOI: 10.1039/d0cy02349c, 0-7.
61. S. Chen and C. Zhao, 2021, DOI: 10.1021/acssuschemeng.1c02875.

62. T. Goculdas, S. Deshpande, S. Sadula, W. Zheng and D. G. Vlachos, 2022.
63. B. Dahlgren, *Journal of Open Source Software*, 2018, **3**, 565.
64. Y. Chung, F. H. Vermeire, H. Wu, P. J. Walker, M. H. Abraham and W. H. Green, *Journal of Chemical Information and Modeling*, 2022, **62**, 433-446.
65. Y. Chung, R. J. Gillis and W. H. Green, *AIChE Journal*, 2020, **66**, e16976.
66. C. W. Gao, J. W. Allen, W. H. Green and R. H. West, *Computer Physics Communications*, 2016, **203**, 212-225.
67. S. I. Sandler, *Chemical, biochemical, and engineering thermodynamics*, John Wiley & Sons, New York, 4. ed. edn., 2006.
68. C. Waldron, A. Pankajakshan, M. Quaglio, E. Cao, F. Galvanin and A. Gavriilidis, *Industrial & Engineering Chemistry Research*, 2019, **58**, 22165-22177.
69. S. Olofsson, L. Hebing, S. Niedenführ, M. P. Deisenroth and R. Misener, *Computers & Chemical Engineering*, 2019, **125**, 54-70.
70. S. Masoumi, T. A. Duever, A. Penlidis, R. Azimi, P. López-Domínguez and E. Vivaldo-Lima, *Macromolecular Theory and Simulations*, 2018, **27**, 1800016.
71. S.-H. Hsu, S. D. Stamatis, J. M. Caruthers, W. N. Delgass, V. Venkatasubramanian, G. E. Blau, M. Lasinski and S. Orcun, *Industrial & Engineering Chemistry Research*, 2009, **48**, 4768-4790.