Supporting Information: 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

S1 Thermodynamic Reversibility Evaluation

S1.1 Gas-Phase Species

To evaluate the thermodynamic reversibility of reaction i (rev_i), we use Eqs. (S1)-(S3).[1] If rev_i is approximately 1 at all species concentrations (c_j) throughout the reaction profile, then the reaction is considered irreversible. If rev_i is 0.99 or lower at any of these realistic species' concentrations, this indicates at some point in the reaction profile the reverse reaction is within two orders of magnitude of the forward reaction and the reaction is reversible.

$$rev_i = \left(1 - \frac{Q_a}{K_a}\right) \tag{S1}$$

$$Q_a = \prod_j \left(\frac{c_j}{c_0}\right)^{v_{ij}} \tag{S2}$$

$$K_a = exp\left(\frac{-\Delta G^\circ{}_{rxn,i}}{RT}\right) \tag{S3}$$

Q_a is the reaction quotient, K_a is the equilibrium constant, c_0 is the standard state concentration, v_{ij} is the stoichiometric coefficient of species j in reaction i, $\Delta G^\circ{}_{rxn,i}$ is the standard state Gibbs free energy of reaction i, R is the ideal gas constant, and T is the reaction temperature.

For each reaction, the challenge is estimating the $\Delta G^\circ{}_{rxn,i}$ value, as defined by the $\Delta G^\circ{}_j$ values (the standard state Gibbs free energy of every species j at the reaction temperature). This estimation is straightforward for the gas-phase; we discuss two approaches here.

For the ethane dehydrogenation example in Section 2 in the main text and the CO_2-assisted ethane dehydrogenation example in Section S4, all stable species are common chemicals. Therefore, we used the thermosolver software to evaluate all $\Delta G^\circ{}_j$ values.[2] To simplify the example in Section 2 in the main text, we do not consider the reverse of reaction 4; however, this possibility is considered in the more complex example in Section S4 as indicated by thermodynamics.

The second approach uses NASA polynomials to estimate the $\Delta G°_j$ values. These NASA polynomials can be looked up in databases, such as the Burcat thermochemical database,[3] or they can be estimated using group additivity with a tool like Green and coworkers created.[4]

S1.2 Liquid-Phase Species

Estimating the $\Delta G°_j$ values for a liquid-phase reaction is more complex, but necessary since the biomass-derived cross-ketonization reactions occur in the liquid-phase. The liquid-phase value, $\Delta G°_{liq,j}$, depends upon the gas-phase value, $\Delta G°_{gas,j}$, and the solvation free energy, $\Delta G°_{solv,j}$, as shown in Eq. (S4).

$$\Delta G°_{liq,j} = \Delta G°_{gas,j} + \Delta G°_{solv,j} \tag{S4}$$

The $\Delta G°_{gas,j}$ values can be estimated as described above, but the $\Delta G°_{solv,j}$ values require another set of tools developed by Green and coworkers.[5, 6] We can now calculate the $\Delta G°_{liq,j}$ value for each species, and subsequently the $\Delta G°_{rxn,i}$ and K_a values.

The final complexity sources from the liquid concentration at reaction conditions differing from the measured concentrations; the measurement occurs at room temperature of 25 °C, while the liquid concentration needed to evaluate Q_a occurs at 350 °C. We must calculate the liquid concentration at reaction conditions ($c_{liq,j,350°C}$) from the measured liquid concentration ($c_{liq,j,25°C}$).

We begin from the relationship between $\Delta G°_{solv,j}$ and the ratio between the concentrations in the liquid and gas phases, as defined in Eq. (S5).

$$exp\left(\frac{-\Delta G°_{solv,j,350°C}}{RT}\right) = \frac{c_{liq,j,350°C}}{c_{gas,j,350°C}} = \frac{c_{liq,j,350°C}}{\frac{n_{total,j,350°C} - n_{liq,j,350°C}}{V_{gas}}} = \frac{c_{liq,j,350°C}}{\frac{n_{total,j,350°C}}{V_{gas}} - \frac{n_{liq,j,350°}}{V_{gas}}} \tag{S5}$$

In the experimental setup, the volumes of the liquid (V_{liq}) and gas (V_{gas}) are identical, allowing Eq. (S5) to reduce to Eq. (S6).

$$exp\left(\frac{-\Delta G°_{solv,j,350°C}}{RT}\right) = = \frac{c_{liq,j,350°C}}{\frac{n_{total,j,350°C}}{V_{liq}} - \frac{n_{liq,j,350°C}}{V_{liq}}} = \frac{c_{liq,j,350°C}}{\frac{n_{total,j,350°C}}{V_{liq}} - c_{liq,j,350°C}} \tag{S6}$$

Evaluating the $\Delta G°_{solv,j}$ values at 25 °C, we discern Dod, Tri, and F are all approximately entirely in the liquid-phase; no moles of these species were lost to the gas-phase, so $c_{liq,j,25°C}$ enables the calculation of $n_{total,j,350°C}$ for these species. Using stoichiometry and volumes, we estimate the total moles ($n_{total,j,350°C}$) of FA, LA, H_2O, and CO_2 at each measurement. Eq. (S6) can now be rearranged to solve for $c_{liq,j,350°C}$, as shown in Eq. (S7). These calculations are well documented in the online data repository of this work.[7]

$$c_{liq,j,350°C} = \frac{exp\left(\frac{-\Delta G°_{solv,j,350°C}}{RT}\right) \times \frac{n_{total,j,350°C}}{V_{liq}}}{1 + exp\left(\frac{-\Delta G°_{solv,j,350°C}}{RT}\right)} \tag{S7}$$

These are now the concentrations of each species at reaction conditions, which allow for the accurate evaluation of Q_a and the thermodynamic reversibility of each proposed reaction.

S2 Reaction Network Graph Evolutions

For the example reaction network introduced in Figure 2 of the main text, we document all feeds of one or two species and their RNG evolutions in Figure S1-Figure S15.

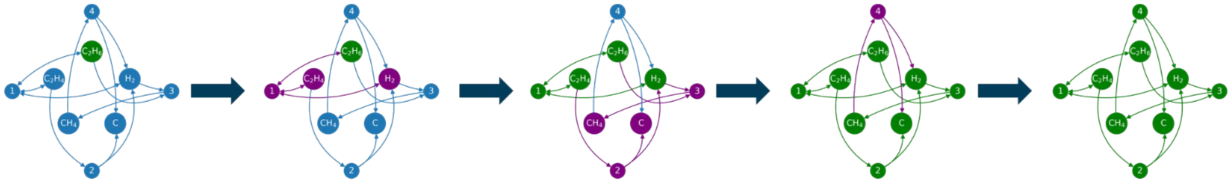

Figure S1: RNG evolution for the feed of C_2H_6.

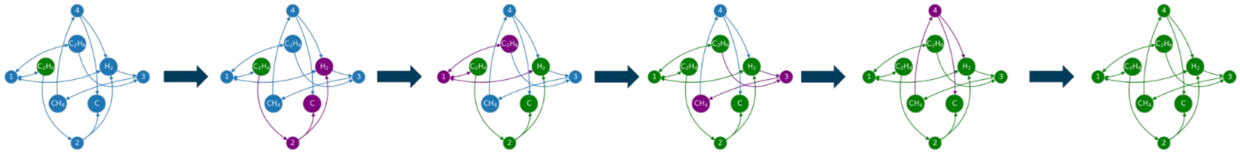

Figure S2: RNG evolution for the feed of C_2H_4. This is the same RNG evolution as shown in Figure 3 in the main text.

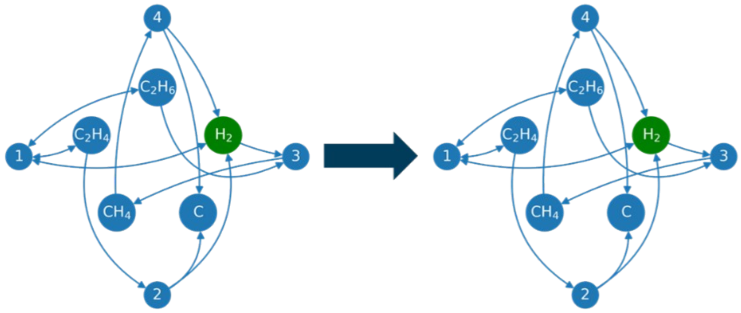

Figure S3: RNG evolution for the feed of H_2.

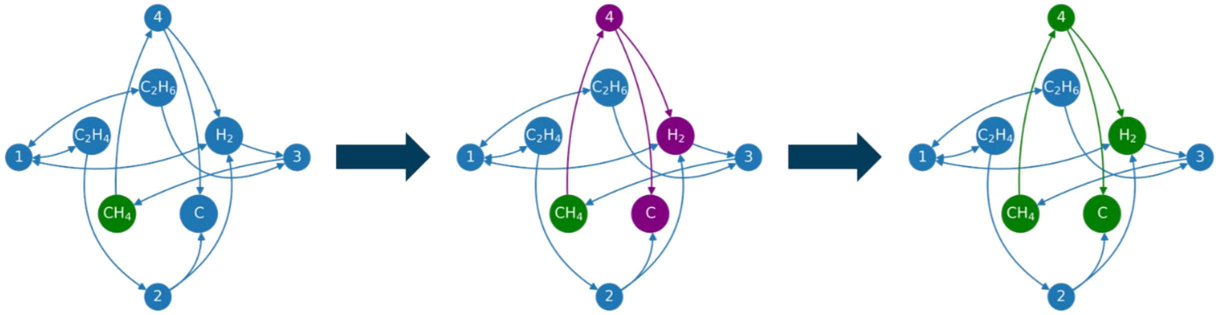

Figure S4: RNG evolution for the feed of CH_4.

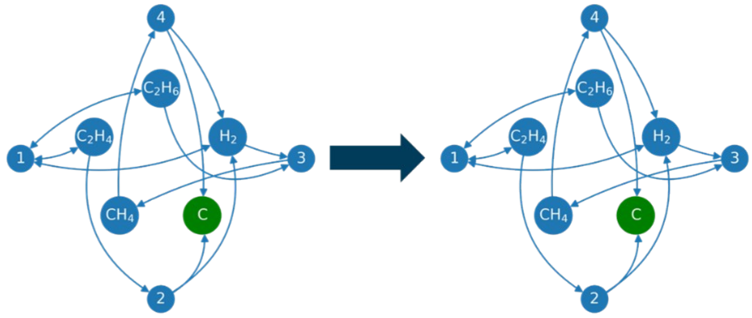

Figure S5: RNG evolution for the feed of C.

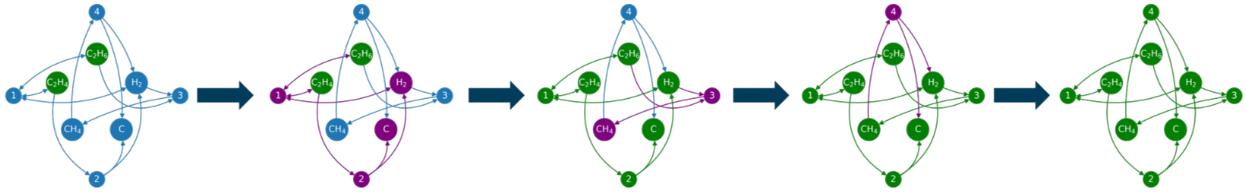

Figure S6: RNG evolution for the feed of C_2H_6 and C_2H_4.

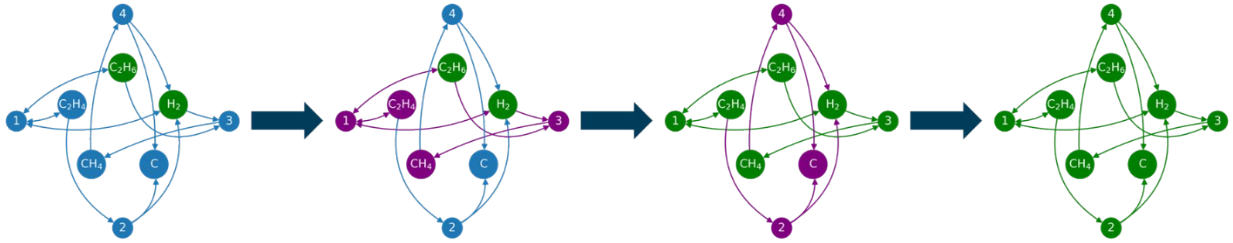

Figure S7: RNG evolution for the feed of C_2H_6 and H_2.

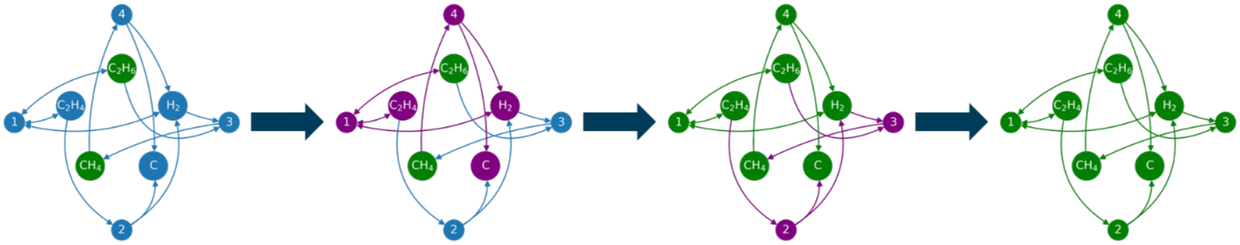

Figure S8: RNG evolution for the feed of C_2H_6 and CH_4.

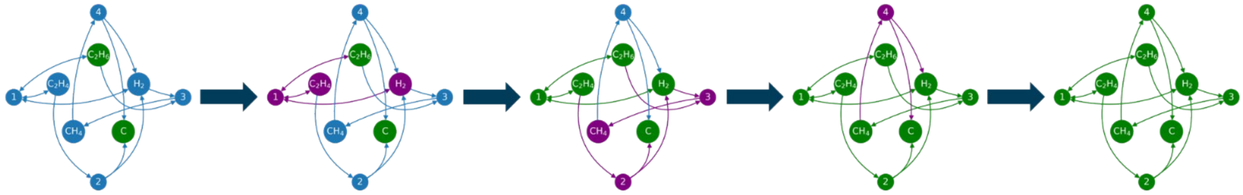

Figure S9: RNG evolution for the feed of C_2H_6 and C.

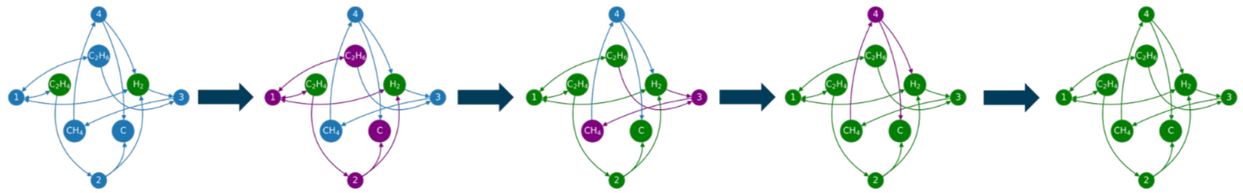

Figure S10: RNG evolution for the feed of C_2H_4 and H_2.

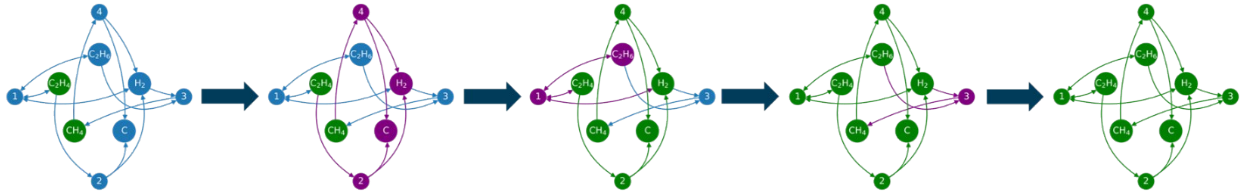

Figure S11: RNG evolution for the feed of C_2H_4 and CH_4.

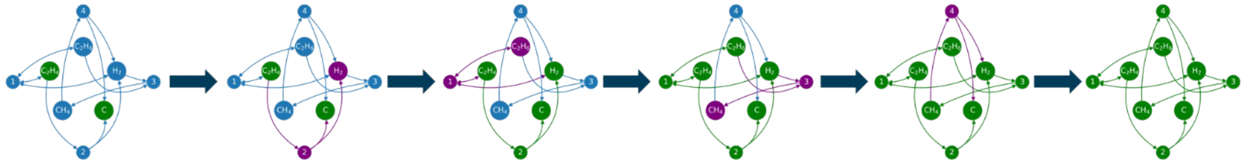

Figure S12: RNG evolution for the feed of C_2H_4 and C.

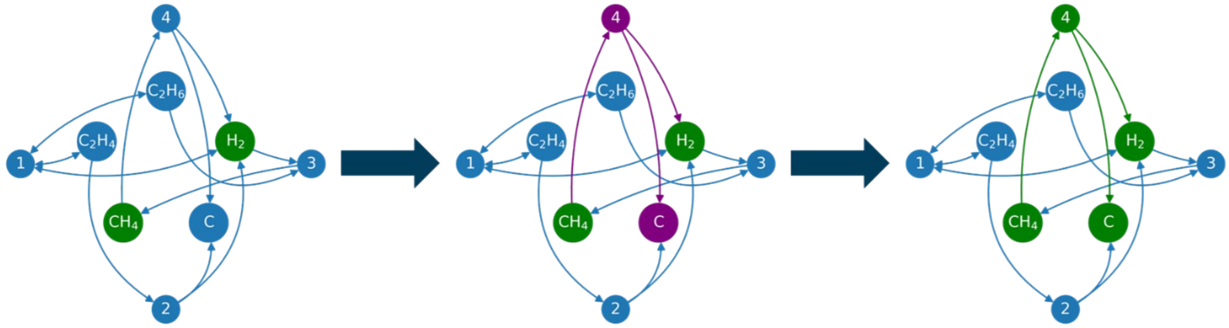

Figure S13: RNG evolution for the feed of H_2 and CH_4.

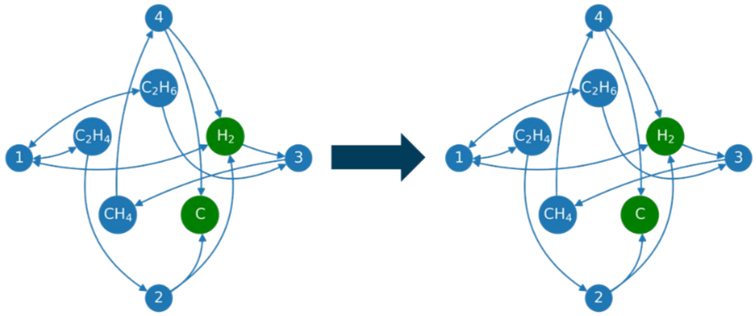

Figure S14: RNG evolution for the feed of H_2 and C.

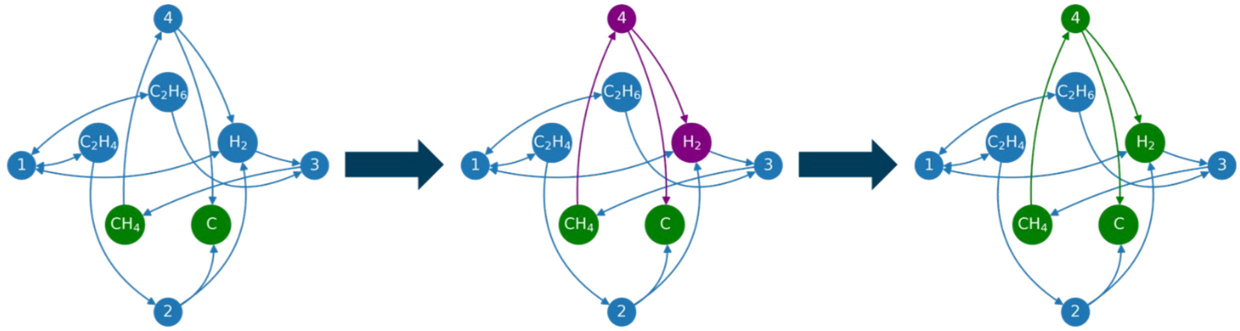

Figure S15: RNG evolution for the feed of CH_4 and C.

S3 Design of Experiments Algorithms

This section contains pseudocode detailing the algorithms used in this work. All algorithms are applied upon the reaction systems investigated herein; the corresponding code for these applications is available in the online data repository.[7]

S3.1 Reaction Network Graph Evolution

```
input = RNG, inlet_composition
reactor_composition = inlet_composition
reactions_involved = 0
evolving = True
while evolving = True:
    for each reaction in RNG:
        if reactants in reactor_composition:
            add reaction to reactions_involved
            add products of reaction to reactor_composition
    if reactions_involved is unchanged:
        evolving = False
output = RNG_evolutionary_path
```

S3.2 Design of Experiments for Fully Determined Probabilistic Graphical Models

```
input = RNG, potential_feeds
proposed_experiments = 0
for each species_fed in potential_feeds:
    RNG_evolutionary_path = Algorithm_1(RNG, species_fed)
    evolved_RNG = final_state(RNG_evolutionary_path)
    PGM = PGM_transformation(evolved_RNG)
    rank = stoichiometric_matrix_rank(PGM)
    if rank = full:
        add species_fed and evolved_RNG to proposed_experiments
for each combination of 2_species_fed in potential_feeds:
    RNG_evolutionary_path = Algorithm_1(RNG, 2_species_fed)
    evolved_RNG = final_state(RNG_evolutionary_path)
```

```
        PGM = PGM_transformation(evolved_RNG)
        rank = stoichiometric_matrix_rank(PGM)
        if rank = full:
            add 2_species_fed and evolved_RNG to proposed_experiments
outputs = proposed_experiments
```

S3.3 Design of Experiments for Underdetermined Probabilistic Graphical Models

```
input = RNG, potential_feeds
proposed_experiments = 0
for each species_fed in potential_feeds:
    RNG_evolutionary_path = Algorithm_1(RNG, species_fed)
    evolved_RNG = final_state(RNG_evolutionary_path)
    for each species_active in evolved_RNG:
        if directed_edge_condition = True and feed_condition = True:
            add species_fed and evolved_RNG to proposed_experiments
for each combination of 2_species_fed in potential_feeds:
    RNG_evolutionary_path = Algorithm_1(RNG, 2_species_fed)
    evolved_RNG = final_state(RNG_evolutionary_path)
    for each species_active in evolved_RNG:
        if directed_edge_condition = True and feed_condition = True:
            add 2_species_fed and evolved_RNG to proposed_experiments
outputs = proposed_experiments
```

S3.4 Design of Experiments for Boolean Reaction Circuits

```
input = RNG, potential_feeds
proposed_experiments = 0
for each species_fed in potential_feeds:
    RNG_evolutionary_path = Algorithm_1(RNG, species_fed)
    BRC = BRC_transformation(RNG_evolutionary_path)
    default_active_species = run_logic(BRC)
    for each reaction "and" gate in BRC:
        altered_BRC = deactivate_reaction(BRC, reaction)
        altered_active_species = run_logic(altered_BRC)
        if default_active_species not equal altered_active_species:
            add species_fed, default_active_species, and altered_active_species
            to proposed_experiments
for each combination of 2_species_fed in potential_feeds:
    RNG_evolutionary_path = Algorithm_1(RNG, species_fed)
    BRC = BRC_transformation(RNG_evolutionary_path)
    default_active_species = run_logic(BRC)
    for each reaction "and" gate in BRC:
        altered_BRC = deactivate_reaction(BRC, reaction)
        altered_active_species = run_logic(altered_BRC)
        if default_active_species not equal altered_active_species:
```

```
                    add 2_species_fed, default_active_species, and
                    altered_active_species to proposed_experiments
      outputs = proposed_experiments
```

S4 RNI Upon a CO_2-Assisted Ethane Dehydrogenation Network

S4.1 Reaction System Introduction

We demonstrate our reaction network identification (RNI) methodology for a physical system we recently investigated[8] on the conversion of ethane (C_2H_6) and carbon dioxide (CO_2) over a Ga/Al_2O_3 catalyst to generate ethylene (C_2H_4), methane (CH_4), hydrogen (H_2), coke (C), water (H_2O), and carbon monoxide (CO).

The kinetic model simulating effluent data is constructed from seven reactions (Figure S16a): ethane dehydrogenation and hydrogenolysis, coking from ethylene and methane, the reverse water-gas shift, methane steam reforming, and coke gasification. Our RNI methodology aims to identify all seven reactions as occurring while not including any inactive pathways. Model details are in Table S1.

An initial experiment feeding C_2H_6 and CO_2 is simulated, generating the measurements represented as data points in Figure S16b. Based on the real system, we simulate plug flow reactor effluent measurements (4 repeats) as concentration data characterized by gas chromatography with Gaussian noise added. In line with the measurement constraints of the physical system, i.e., that the H_2O and coke concentrations cannot be measured, we perform our analysis without using these values.

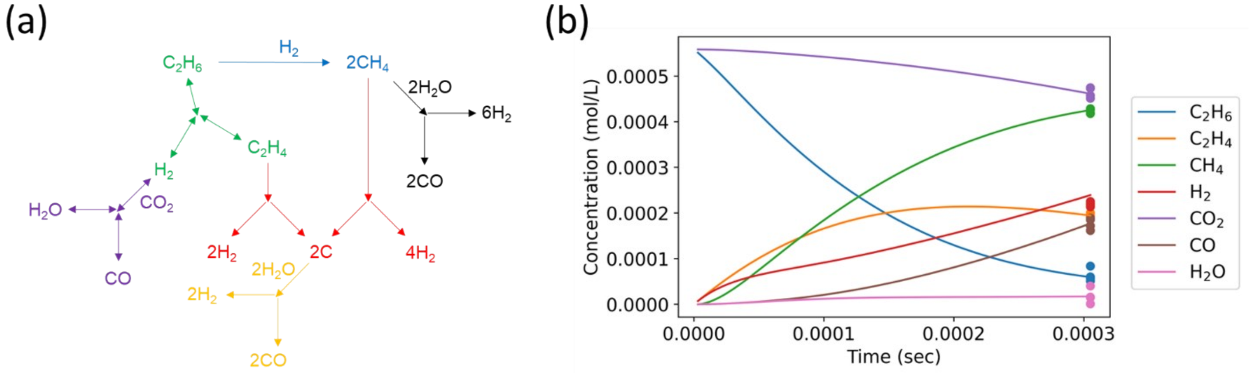

Figure S16: Visualization of the kinetic model. (a) Reaction network of ethane dehydrogenation (green), coking from ethylene (red), ethane hydrogenolysis (blue), coking from methane (red), reverse water-gas shift (purple), methane steam reforming (black), and coke gasification (orange). (b) Concentration profiles over time within the plug flow reactor at a pressure 1 atm and temperature of 873 K. We model a tubular reactor with an inner diameter of 10 cm and a length of 20 cm, within which are packed 25 mg of catalyst and 1.3 g of quartz inert. The inlet feed is introduced at a flowrate of 20 mL/min at room temperature with a composition of 4 mole percent C_2H_6, 4 mole percent CO_2, and 92 mole percent He inert gas. Throughout the rest of this work, the balance of gas composition is always supplied by He inert gas but is not explicitly stated. Gaussian noise with a standard deviation of 0.1 mole percent is introduced to the outlet measurements, which are displayed as datapoints at the final residence time.

Table S1: Reactions occurring within the simulated system and their associated rate equations. The reversible reactions are indicated as such with two-way arrows in their chemical formulas, and they incorporate the equilibrium correction term into their rate equations of the reaction quotient (Q_P) divided by the equilibrium constant (K_P). A limitation of these reversible reactions is that they will not occur in reverse unless there are trace amounts of the reactants present due to their equation forms. Therefore, when these reverse reactions should be occurring in later simulations, 0.1 mole percent of each reversible equation's reactant is cofed to enable these reverse reactions to occur while influencing little else. The concentrations are all in units of mol/L, aside from the coke, which is in units of mol/L^(2/3) since its concentration is relative to area instead of volume. We note that the values of the rate constants were not sourced from our previous work[8] since not all rates were parameterized, but are instead set to values yielding reaction rates on similar orders of magnitude under our considered conditions to reflect realistic experimental results.

Reaction	Chemical Formula	Net Rate Equation [mol/L/s]	k value
Ethane dehydrogenation	$C_2H_6 \leftrightarrow C_2H_4 + H_2$	$k\times[C_2H_6]\times(1-(Q_P/K_P))$	4.2×10^3
Coking from ethylene	$C_2H_4 \rightarrow 2C + 2H_2$	$k\times[C_2H_4]$	5.8×10^2
Ethane hydrogenolysis	$C_2H_6 + H_2 \rightarrow 2CH_4$	$k\times[C_2H_6]\times[H_2]$	4.3×10^7
Coking from methane	$CH_4 \rightarrow C + 2H_2$	$k\times[CH_4]$	6.1×10^2
Reverse water-gas shift	$CO_2 + H_2 \leftrightarrow CO + H_2O$	$k\times[CO_2]\times[H_2]\times(1-(Q_P/K_P))$	5.2×10^6
Methane steam reforming	$CH_4 + H_2O \rightarrow CO + 3H_2$	$k\times[CH_4]\times[H_2O]$	6.9×10^7
Coke gasification	$C + H_2O \rightarrow CO + H_2$	$k\times[C]\times[H_2O]$	9.3×10^6

S4.2 Reaction Network Graph Creation

S4.2.1 Enumeration and Pruning of Feasible Overall Reactions

From the stable species identified from the initial experiment, we exhaustively list all stoichiometrically feasible reactions using Eq. (1) in the main text. Note that these are overall reactions, and not elementary ones describing microkinetic phenomena. For example, we investigate a combinatorically proposed reaction of C_2H_6 as the reactant and C_2H_4 and H_2 as products. This is a feasible stoichiometric reaction with coefficients of 1 for each species yielding balances for the elements of carbon and hydrogen. In contrast, the combinatorically proposed reaction of $aC_2H_6 \leftrightarrow bC + cC_2H_4$ is not stoichiometrically feasible for any values of a, b, and c, so it is discarded as infeasible.

Upon completing this combinatorial exploration, we identify 51 stoichiometrically feasible reactions for this system (listed in text file in the online data repository[7]). We limit our investigation to a maximum of two species on each side of the chemical equation to avoid infeasibly complicated reactions.

This list comprised of 51 stoichiometrically feasible reactions must now be scrutinized with expert knowledge. Many reactions are physically infeasible and can be eliminated. For example, C_2H_6 will not decompose directly into C_2H_4 and CH_4 ($2C_2H_6 \leftrightarrow C_2H_4 + 2CH_4$). This chemistry may happen indirectly as a combination of C_2H_6 dehydrogenation into C_2H_4 ($C_2H_6 \leftrightarrow C_2H_4 + H_2$) and then an additional C_2H_6 hydrogenolysis into CH_4 ($C_2H_6 + H_2 \leftrightarrow 2CH_4$). However, it will not occur directly, so that reaction is removed from consideration. We are optimistic that in the future, this application of expert knowledge need not be supplied by a user but instead from an automated literature search or a reliable software leveraging elementary or overall reaction rules given the chemistry involved such as RING, NETGEN, COMGEN, and KING.[9-12]

By applying similar chemical knowledge, we reduce the list into 15 chemically viable reactions, which are subjected to further analysis.

S4.2.2 Determining Thermodynamic Reversibility

An important consideration is the reversibility of each reaction. As discussed in Section S1, examination of the reactions' equilibrium constants at the reaction temperature provides indication of which reactions are reversible and which are not. For example, the equilibrium constant of ethane hydrogenolysis is 18,400 at 873 K. This corresponds to an equilibrium conversion of approximately 1.5% if feeding pure CH_4. Such a low conversion implies this reaction is practically irreversible. Similar analysis reveals that ethane dehydrogenation, the reverse Sabatier reaction, and the reverse water-gas shift are reversible, while all other reactions that do not involve coke are irreversible.

We do not apply this analysis to reactions involving coke due to experimental evidence in our original study[8] indicating theoretically reversible coking reactions were actually irreversible in the physical system. Specifically, we found that the equilibrium constant of the reaction of coking from CH_4 indicates that CH_4 could be formed from coke and H_2, but our experiments flowing H_2 over coke did not detect any CH_4. Similarly, reaction 15, the gasification of coke into CO, is thermodynamically feasible in reverse, but our experiments only show evidence of the forward reaction. Therefore, we apply our reversibility analysis to the reactions involving only gas-phase species and rely upon experimental measurements to evaluate the reversibility of the other reactions involving coke. To simplify the example of Figure 2 in the main text, we specify methane coking as irreversible.

S4.2.3 Reaction Network Graph

Knowing the 15 reactions being considered and their reversibility, we can now construct the reaction network graph (RNG) to be used for our RNI methodology. The traditional reaction network and the RNG are displayed in Figure S17. This RNG is significant more complex than the one investigated in the main text, but our RNI methodology is still applicable.

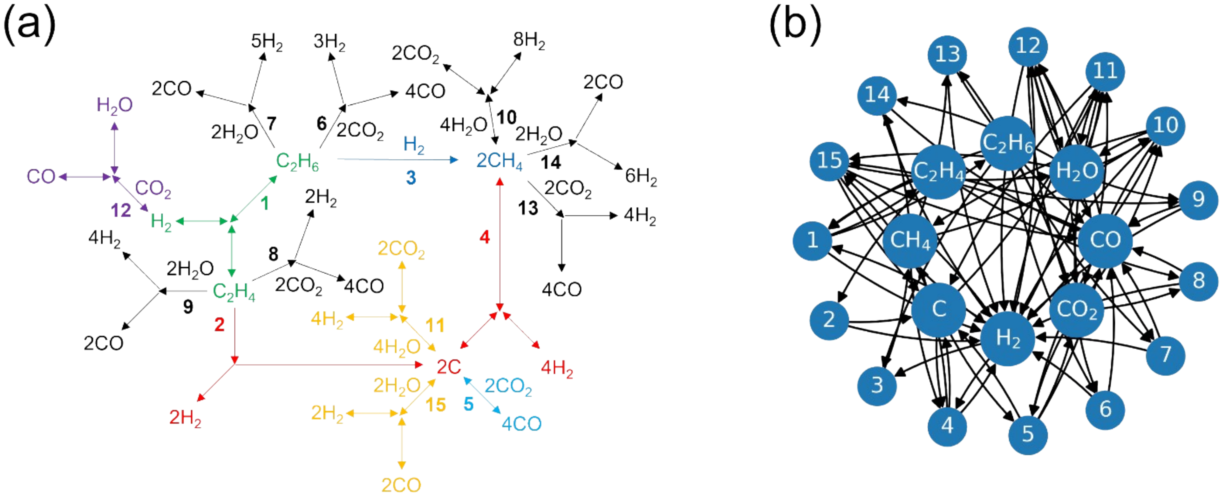

Figure S17: Alternative depictions of the CO_2-assisted ethane dehydrogenation reaction network. (a) Considered overall reaction network for CO_2-assisted ethane dehydrogenation over Ga/Al_2O_3 catalyst. Reaction 1 (R1): ethane dehydrogenation (green). R2: coking from ethylene (red). R3: ethane hydrogenolysis (blue). R4: coking from methane (red). R5: reverse Boudouard reaction (light blue). R6: ethane dry reforming (black). R7: ethane steam reforming (black). R8: ethylene dry reforming (black). R9: ethylene steam reforming (black). R10: reverse Sabatier reaction (black). R11: coke gasification to CO_2 (orange). R12: reverse water-gas shift (purple). R13: methane dry reforming (black). R14: methane steam reforming (black). R15: coke gasification to CO (orange). (b) Reaction network graph with each species and reaction represented as a node. Relationships between nodes correspond to those displayed in (a) using directed edges to represent reaction flux and reversibility or irreversibility.

S4.3 Active Learning of Reaction Network

S4.3.1 Overview

The following sections will demonstrate learning the reaction network from probabilistic graphical models (PGMs), both fully determined and underdetermined, and Boolean reaction circuits (BRCs). If we use our optimal design of experiments (DOE) with both analysis techniques in tandem, then there will not be opportunities to use underdetermined PGMs. To better demonstrate and discuss each technique rather than omit one given this specific problem structure, we instead use DOE for each analysis technique sequentially. Once all reactions identifiable with these techniques have been exhausted, we rely upon the delplot method[13, 14] to demonstrate how additional analysis techniques can be incorporated into this RNI framework.

S4.3.2 Probabilistic Graphical Models

Using the PGM formulation of Eq. (4) in the main text, we can infer the probability distributions of the reaction extents from our effluent concentrations. The likelihood term is described in Eq. (S8) as Gaussian distributions with means defined by the summations of reaction extents multiplied by their corresponding stoichiometric coefficients (v_{ij}), which is the equivalent of each row of the matrix formulation in Eq. (2) in the main text.

$$P\left(\underline{\Delta[C]}_j | \xi_i, \sigma\right) = \prod_j N\left(\sum_i v_{ij}\xi_i, \sigma^2\right) \tag{S8}$$

An example can be derived from the PGM in Figure 5a in the main text. Eq. (2) in the main text with all variables substituted in reads $P(\xi_1, \xi_2, \xi_3, \xi_4, \sigma | \Delta[C_2H_6], \Delta[C_2H_4], \Delta[CH_4], \Delta[H_2]) \propto P(\Delta[C_2H_6], \Delta[C_2H_4], \Delta[CH_4], \Delta[H_2] | \xi_1, \xi_2, \xi_3, \xi_4, \sigma) \times P(\xi_1, \xi_2, \xi_3, \xi_4, \sigma)$. With the likelihood expanded with Eq. (S8), the full PGM is $P(\xi_1, \xi_2, \xi_3, \xi_4, \sigma | \Delta[C_2H_6], \Delta[C_2H_4], \Delta[CH_4], \Delta[H_2]) \propto P(\Delta[C_2H_6] | \xi_1, \xi_3, \sigma) \times P(\Delta[C_2H_4] | \xi_1, \xi_2, \sigma) \times P(\Delta[CH_4] | \xi_3, \xi_4, \sigma) \times P(\Delta[H_2] | \xi_1, \xi_2, \xi_3, \xi_4, \sigma) \times P(\xi_1, \xi_2, \xi_3, \xi_4, \sigma)$ where $P(\Delta[C_2H_6] | \xi_1, \xi_3, \sigma) = N(-\xi_1 - \xi_3, \sigma^2)$, $P(\Delta[C_2H_4] | \xi_1, \xi_2, \sigma) = N(\xi_1 - \xi_2, \sigma^2)$, $P(\Delta[CH_4] | \xi_3, \xi_4, \sigma) = N(2\xi_3 - \xi_4, \sigma^2)$, and $P(\Delta[H_2] | \xi_1, \xi_2, \xi_3, \xi_4, \sigma) = N(\xi_1 + 2\xi_2 - \xi_3 + 2\xi_4, \sigma^2)$.

To identify promising experiments for fully determined PGM analysis, all feeds involving one or two species are combinatorically explored with an RNG evolution, i.e., simulated as shown in Figure 3 in the main text without requiring any effluent data from an experiment or simulation. Specifically, for each feed, the RNG is evolved to its final state following the algorithm in Section S3.1. From this evolved RNG, the PGM's stoichiometric matrix is abstracted, and the rank is evaluated. Feeds of full rank are reported as promising experiments for fully determined PGM analysis, while the others are discarded. This effectively explores the feed space of the RNG up to two fed species, though combinations involving more can also be investigated. This DOE algorithm is outlined in Section S3.2.

Using this algorithm to explore all possible feed conditions, the following promising experiments are revealed for our RNG:

Feeding CH_4 to investigate the forward direction of R_4 (methane coking).
Feeding H_2 over C to explore the reverse direction of R_4 (methane production from coke).
Feeding CO to investigate the forward direction of R_5 (Boudouard reaction).
Feeding CO_2 over C to investigate the reverse direction of R_5 (reverse Boudouard reaction).

Since all experiments had an information rating of one, each was simulated to generate synthetic data for PGM analysis. R_4 is identified as active forward with a statistically non-zero extent, and all other reactions are inactive with extents statistically indiscernible from zero. The data of PGM analyses and all following RNI analyses can be found in the online data repository.[7] This information leads to an updated RNG, which may yield new DOE suggestions. In this case, rerunning Algorithm 2 with the updated RNG does not propose new experiments.

Having exhausted the experiments resulting in fully determined PGMs, we focus on identifying experiments that yield underdetermined yet informative PGMs. Feeds are combinatorically explored, with the RNG evolved to its final state in accordance with the algorithm in Section S3.1. This final RNG state is analyzed to discern its information rating for an underdetermined PGM analysis. This combinatorial DOE approach explores the feed space up to two fed species. While combinations involving more species can be investigated, we found exploration of this subsection of the design space sufficient. This DOE algorithm is outlined in Section S3.3.

Evaluating potential feed with this DOE, three informative experiments are proposed that identify new reactions:

Feeding C_2H_6 to investigate R_1 (ethane dehydrogenation) and R_3 (hydrogenolysis).
Feeding C_2H_4 to investigate R_1 (ethane dehydrogenation) and R_3 (hydrogenolysis).
Feeding CO and H_2O to investigate R_{10} (Sabatier reaction).

The first two reactions have information ratings of two and identify the same reactions. We elect to simulate feeding C_2H_6; the results are displayed in Figure S18. We note that the probability estimates for each extent of reaction bear similarity to bounded uniform distributions, confirming this as an underdetermined PGM analysis. However, the extents of R_1 and R_3 are statistically guaranteed to be non-zero; our DOE procedure correctly predicted this experiment as informative, even though the PGM is underdetermined. Confirmation of these reactions' occurrence cements their inclusion in the RNG.

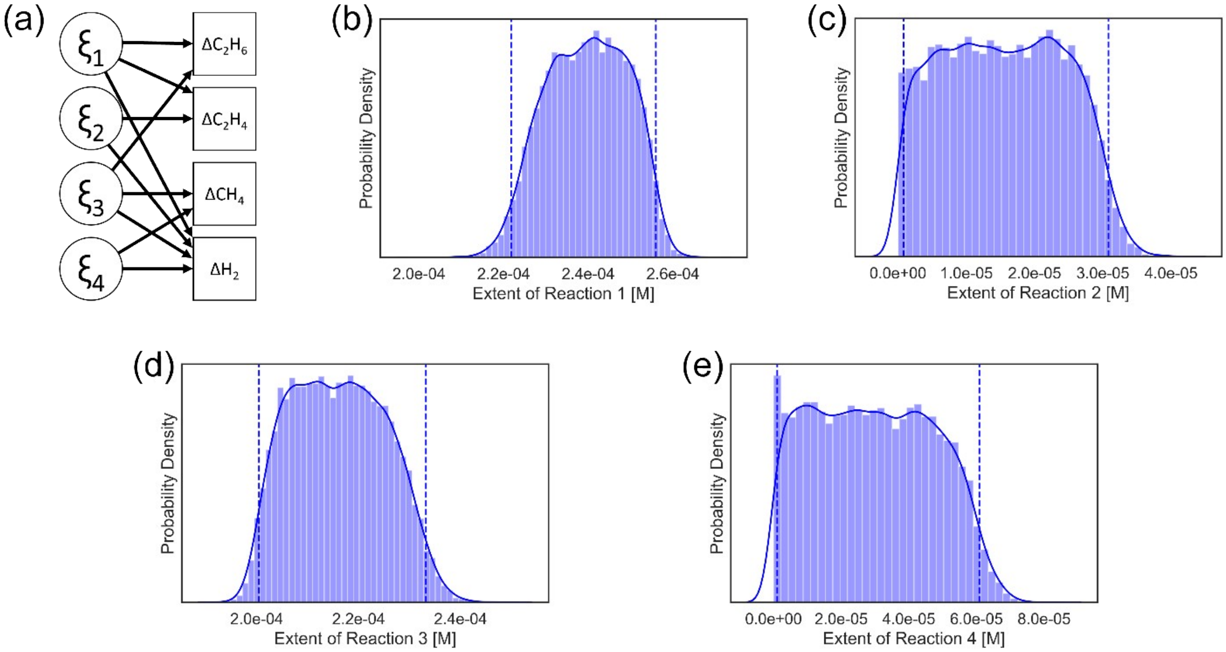

Figure S18: Underdetermined PGM analysis of reactions 1-4. (a) PGM of extents of reactions 1-4. The standard deviation of the experimental noise (σ) would be included as a parent node but is omitted for visual clarity. (b) Probability distribution and histogram of extent of R_1. (c) Probability distribution and histogram of extent of R_2. (d) Probability distribution and histogram of extent of R_3. (e) Probability distribution and histogram of extent of R_4. Histograms are included for clarity in b-e to show for R_2 and R_4 that no extents less than zero are included in the estimations in accordance with their irreversibility. Vertical dashed lines in b-e denote the 95% credible intervals.

The DOE proposes an additional experiment: to identify R_{10} by co-feeding CO and H_2O. However, this analysis is more illuminating when using BRCs rather than PGMs, so for the demonstrative purposes of this example we move on to BRC analysis.

S4.3.3 Boolean Reaction Circuits

Updating the RNG with our conclusive results of the 4 investigated reactions of R_1, R_3, R_4, and R_5, we next employ BRC analysis to discern which of the remaining uncertain reactions occur.

To explore the utility of a BRC, its “and” gates are individually deactivated, and the resulting species in the effluent are predicted. Suppose the effluent species change from the default case with all “and” gates activated. In that case, this indicates a definitive result from the proposed BRC experiment: either the reaction occurs to yield the default effluent species, or the reaction does not occur to yield the altered effluent species. Using this principle, valuable BRC experiments can be identified through a combinatorial search of the feed design space, as outlined in the algorithm in Section S3.4.

Exploring possible feeds with this DOE, two informative experiments are proposed that identify new reactions:

Feeding C_2H_4 to investigate R_2 (ethylene coking).

Feeding CO and H_2O to investigate R_{12} (water-gas shift reaction) and R_{10} (Sabatier reaction)

Simulating the first proposed experiment confirms the inclusion of R_2 into the RNG once H_2 is measured in the effluent; we describe this analysis in Section 2.4 and Figure 4 in the main text.

With an information rating of two, the second proposed experiment provides an even more illuminating result from BRC analysis. We only need the first two layers of the BRC for the analysis, as diagramed in Figure S19. The measurement of CO_2 and H_2 in the effluent alongside the absence of CH_4 confirms that R_{12} occurs while R_{10} does not. The more interesting result comes from the coking. While the physical constraint described in Section S4.1 dictates that the C concentration cannot be measured, the catalyst would be checked for coking after the run. Our simulation indicates no C would be observed; however, the BRC predicted C would form regardless of the removal of any single reaction from the RNG. We must look deeper into the BRC analysis to understand why its prediction of C formation was unreliable.

Examining the BRC, it becomes clear that C will not form if both R_{11} and R_{15} are inactive in the direction of coke formation. We had limited our BRC search to the omission of a single "and" gate at a time to avoid combinatorial explosions of possible omissions. Therefore, there may be species missing from the effluent that the BRC claims must be present when multiple reactions are inactive, as exemplified here. These can be regarded as opportunities to eliminate multiple reactions, which we do for the coke formation (R_{11} and R_{15}). This experiment actually corresponded to identifying four reactions, even though it only had an initial information rating of two. Updating the RNG and rerunning the DOE confirms all promising experiments using the BRC methodology have been conducted.

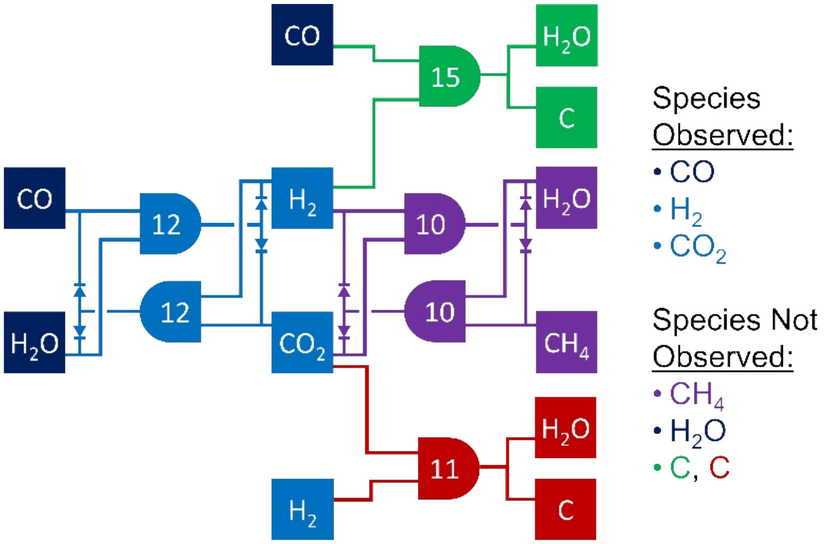

Figure S19: BRC analysis of R_{10}, R_{11}, R_{12}, and R_{15}. There are more reactions that would follow from these two layers, but we only diagram these first two to focus the analysis on these reactions. Each "and" gate (reaction) has a unique color. The color of a species value node corresponds 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. R_{12} and R_{10} are reversible reactions, as implemented using diodes. The observed species from the simulated effluent are listed on the right. H_2O is not observed as per the physical constraint described in Section S4.1. C is listed twice, once in each color of R_{11} and R_{15} to indicate formation from either reaction is possible.

S4.3.4 Delplots

A common approach in determining a reaction product's rank (i.e., primary, secondary, etc.) is the use of delplots[13-15]. A delplot is constructed from concentration data collected at multiple residence times at low conversions. One calculates the selectivity of product (P), as shown in Eq. (S9) as the y-axis value, and conversion of a reactant (R), as shown in Eq. (S10) as the x-axis value, and the first rank delplot of selectivity vs. conversion, is graphed. If the y-intercept is non-zero, the product forms in a primary reaction. If the intercept is zero, the product forms by a secondary or later reaction. Identifying the products' ranks offers valuable utility for identifying the RNG.

$$y = selectivity = \frac{P/R_0}{1 - R/R_0} \tag{S9}$$

$$x = conversion = 1 - R/R_0 \tag{S10}$$

Obtaining data for a delplot is seemingly simple: feed the reactant(s) of the reaction of interest, measure the reaction's reactant and product concentrations at various residence times at low conversion, and construct the selectivity versus conversion delplot. However, physical practicalities can prevent this approach's application; sometimes, key species are not measurable.

To overcome this concern, we extend delplots with a more general formulation, as specified in Eqs. (S11)-(S12). The x-value is defined as the change in a species that certainly changes in a primary reaction (S_c) normalized by the initial amount of a reactant species (R_0). The y-value is equal to the change in a species that is questionably changing in a primary reaction (S_Q) normalized by R_0 divided by the x-value. The parities of the species changes are defined to ensure the terms are positive, regardless of whether the species are being consumed or produced. This delplot formulation produces the revealing y-intercept of the ratio of rates between the two species of S_Q and S_C, while providing greater flexibility by allowing each species to be a reactant or a product.

$$y = \frac{\pm \Delta S_Q / R_0}{\pm \Delta S_C / R_0} \tag{S11}$$

$$x = \pm \Delta S_C / R_0 \tag{S12}$$

With this delplot approach, we focus on creating and applying DOE to optimally identify the remaining reactions. Delplots identify non-zero rates of primary reactions, so our associated DOE will analyze the first evolution state of the RNG, identifying the possible primary reactions. To reiterate, an informative delplot analysis to identify a reaction in question requires 2 species to be present: a species whose consumption or generation can be uniquely attributed to the primary reaction (S_Q), and a species whose consumption or generation is certain given the possible primary reactions (S_C). We term this the 2 species criterion.

We offer an example in Figure S20a, where the first RNG evolution of an example experiment has been redrawn and simplified. In this example of feeding H_2O over C, CO_2 will be primarily produced only by $R_{11,}$ CO will be primarily produced only by R_{15}, and H_2 is produced by both reactions. Note that this system is not analyzable by standard delplots because neither reactant H_2O nor C is measurable; therefore, one cannot quantify their conversion. However, our extension to delplots is more flexible.

The first of the 2 species criterion is met by CO_2 for R_{11} and CO for R_{15}. The second is met by H_2, which will be a primary product regardless of which of R_{11} or R_{15} occur. Therefore, this is an informative DRP experiment that can identify two reactions simultaneously, as demonstrated in Figure S20b-d. One final nuance of the 2 species criterion is of note: if a reaction lacks an identifiable species, but all other primary reactions have unique identifiers, the lacking reaction can be evaluated using a shared species and removing contributions from the other identifiable primary reactions.

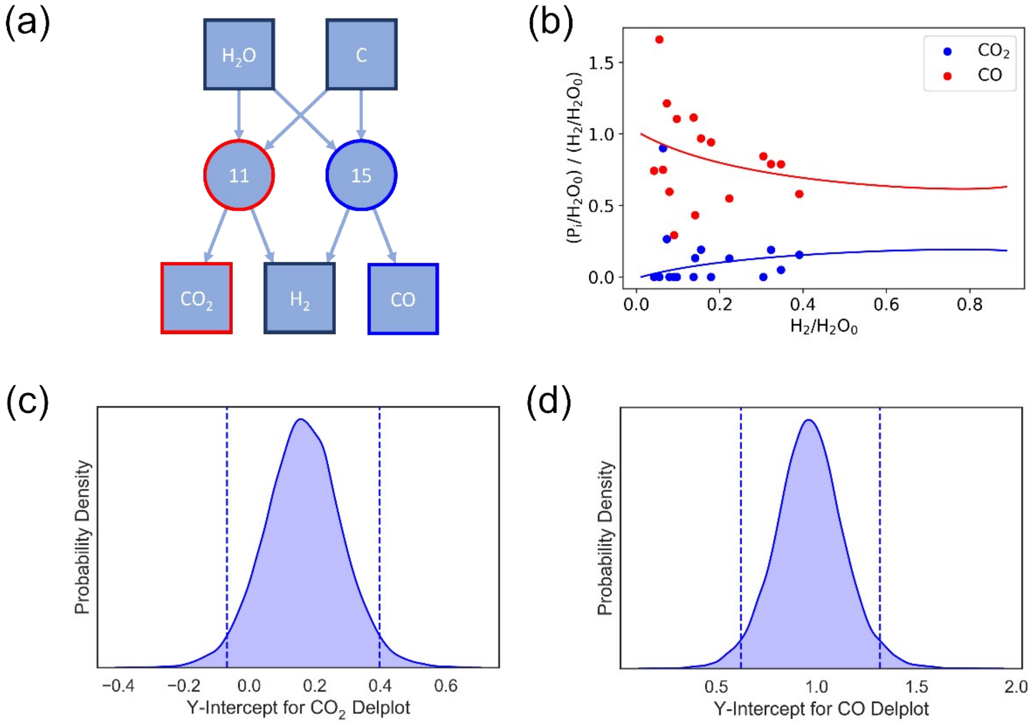

Figure S20: Delplots and associated analysis for determination of R_{11} and R_{15}. (a) Redrawn subsection of the RNG when H_2O is fed over C leading to CO_2 produced by R_{11}, CO produced by R_{15}, and H_2 produced by both reactions. (b) Delplot for CO_2 and CO versus the production of H_2. Solid lines display ground truth, and datapoints display measured values, displaying the significant noise present. (c) Probability distribution of the CO_2 delplot y-intercept estimated by the Bayesian inference. (d) Probability distribution of the CO delplot y-intercept estimated by the Bayesian inference.

This 2 species criterion is the foundation of our DOE approach. From the list of reactions whose inclusion in the RNG is uncertain, one is selected. This reaction's reactants are set as the RNG initial state, and the RNG evolves to reveal the possible primary reactions (i.e., no effluent data is necessary). Then, the 2 species criterion can be confirmed to hold for the uncertain reaction in question. Once the criterion is confirmed, other reactions that might be investigated by the same experiment are explored. First, all other primary reactions that are also uncertain are evaluated with the 2 species criterion, and valid candidate reactions for delplot analysis are added to a list of revealed reactions. Second, an additional species is added to the feed, and the evaluations of the 2 species criterion for all uncertain reactions are repeated. In some cases, new reactions will be added to the list of revealed reactions, while in other cases, reactions will be removed. All remaining possible species are iteratively added to the feed to explore all combinations. When the addition of a species does not cause the list of revealed reactions to increase, this search ends, and the proposed reaction with the largest revealed reactions list is selected. Note that the search terminates after one addition fails to increase the list of revealed reactions, but multiple additions can be explored if appropriate. The identified reactions are removed from the uncertain reactions list, and the DOE process is repeated with a new uncertain reaction as the starting condition. The specific uncertain reaction selected does not matter since all uncertain reactions will eventually need to be investigated. This process continues until no more uncertain reactions remain. The final result is a list generated of feed conditions and associated reactions identified, from which an information rating can be determined for optimal DOE. This approach starting from each unidentified reaction is summarized in this algorithm:

```
input = RNG, measurable_species
revealed_reactions = 0
proposed_experiments = 0
select 1 uncertain_reaction in RNG
species_fed = reactants in uncertain_reaction
RNG_evolutionary_path = Algorithm_1(RNG, species_fed)
primary_RNG = 1st_evolution(RNG_evolutionary_path)
if 2_species_criteria(primary_RNG, uncertain_reaction) = True:
    add uncertain_reaction to revealed_reactions
    for each other_uncertain_reaction in RNG:
        if 2_species_criteria(primary_RNG, other_uncertain_reaction) = True:
            add other_uncertain_reaction to revealed_reactions
    proposed_experiment = species_fed, revealed_reactions
    increased_reveal = True
    count_reactions = count(revealed_reactions)
    while increased_reveal = True:
        increased_reveal = False
        for each other_species in measurable_species:
            add other_species to species_fed
            RNG_evolutionary_path = Algorithm_1(RNG, species_fed)
            primary_RNG = 1st_evolution(RNG_evolutionary_path)
            for each reaction in revealed_reactions:
                if 2_species_criteria(primary_RNG, reaction) = True:
                    keep reaction in revealed_reactions
                else:
                    remove reaction from revealed_reactions
            for each other_uncertain_reaction in RNG:
                if 2_species_criteria(primary_RNG,
                other_uncertain_reaction) = True:
                    add other_uncertain_reaction to revealed_reactions
            count_reactions_updated = max(count(revealed_reactions),
            count_reactions)
            if count_reactions_updated > count_reactions:
                increased_reveal = True
                count_reactions = count_reactions_updated
                proposed_experiment = species_fed, revealed_reactions
            remove other_species from species_fed
else:
    proposed_experiment = Warning: DRP cannot identify uncertain_reaction
outputs = proposed experiment
repeat for all unidentified reactions
```

determine which proposed experiment has the highest information rating

The above algorithm demonstrates that additional analysis techniques each have their own DOE algorithms which can be incorporated into our RNI methodology. Using this DOE to explore possible experiments, the following feeds are suggested:

Feeding H_2O over C to investigate R_{11} (coke gasification to CO_2) and R_{15} (coke gasification to CO).
Feeding CH_4, CO_2, and H_2O to investigate R_{13} (methane dry reforming) and R_{14} (methane steam reforming).
Feeding C_2H_4, CO_2, and H_2O to investigate R_8 (ethylene dry reforming) and R_9 (ethylene steam reforming).
Feeding C_2H_6, CO_2, and H_2O to investigate R_6 (ethane dry reforming) and R_7 (ethane steam reforming).

Our DOE for delplot analysis effectively halves the number of necessary experiments; investigating each reaction with an individual delplot would require eight experiments, whereas the DOE accomplished the same with only four. While delplots can identify reactions that our PGM or BRC methods cannot, there are tradeoffs. Delplots intercepts are difficult to evaluate if there is experimental noise, which can be especially significant at low conversion (Figure S20b). Additionally, delplots require more measurements to be taken at different conversions; therefore, what our DOE proposes above as individual experiments are really multiple experiments with the same feed compositions at different residence times. These practicalities inform our decision to prefer PGM and BRC analysis over delplots for RNI when possible.

Following the DOE, the four experiments are conducted and the final 8 reactions are identified. R_{14} and R_{15} are determined to occur and the others are confirmed to be inactive. Specifically, delplot analysis indicates that feeding H_2O over C produces CO as a primary product but not CO_2; this confirms R_{15} occurs and R_{11} does not. When CH_4, CO_2, and H_2O are co-fed, delplot analysis shows that CO is a primary product while CO_2 is not a primary reactant. This result indicates that R_{14} is active while R_{13} is not. Investigating the reforming reactions of C_2H_6 and C_2H_4 with delplot analysis revealed that CO is not a primary product in these experiments; therefore, R_6, R_7, R_8, and R_9 are identified as inactive. The case for R_{11} and R_{15} is demonstrated in Figure S20; the rest are documented in the online data repository.[7]

The RNG is updated accordingly to be fully identified (Figure S21). Assessing this identified RNG against the ground truth of the kinetic model generating all data, we confirm all identifications are correct and validate our RNI methodology upon this simulated reaction network.

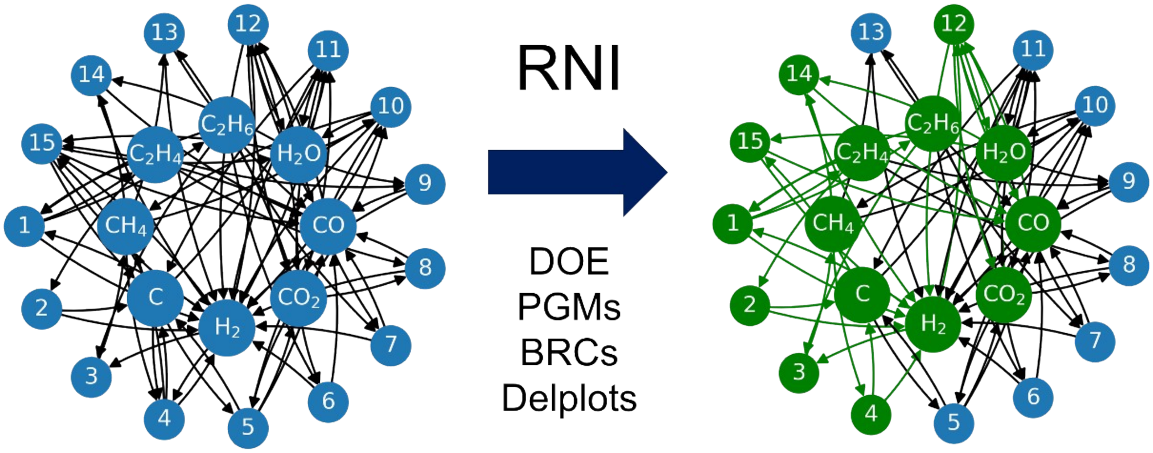

Figure S21: Identification of the reaction network involving R_1, R_2, R_3, R_4, R_{12}, R_{14}, and R_{15} from a possible set of 15 reactions. RNG on left shows potential reactions being considered. RNG on right shows identified reactions in green with R_4 and R_{15} having their reverse reaction edges removed.

S5 Experimental Information

S5.1 Materials Used

Furoic acid, lauric acid, n-dodecane, furan, 12-tricosanone and chloroform were purchased from Sigma Aldrich. Dimethyl sulfoxide was acquired from Fisher Scientifics. Magnesium oxide was obtained from Fisher Scientifics.

S5.2 The Catalytic Reaction and Effluent Characterization

Reactions were conducted in a 100 mL batch Parr reactor. The reactants 2-furoic acid and lauric acid, the solvent n-dodecane, and the catalyst were placed in a glass liner with a magnetic stir bar. The ketone products and furan were quantified using a gas chromatogram (GC) and gas chromatogram-mass spectrometer (GCMS) system. Since 2-dodecanoyl furan was not commercially available, it was quantified using the effective carbon number method using hexane and 2-acetyl furan standards. The other reactants and products were quantified using standard calibration curves (Section S5.5).

S5.3 Mass Transfer Limitations Investigation

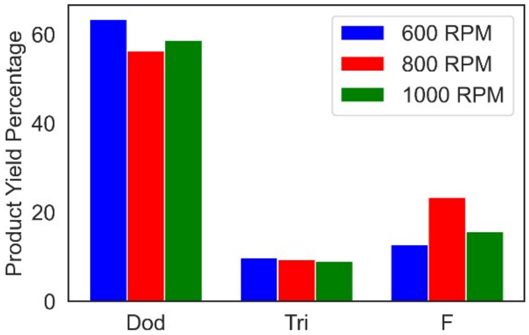

Figure S22: Yields at different stir rates of products in the cross-ketonization reaction system. No trends are evident between yields and stir rates, indicating the reaction system does not experience mass transport limitations in this stir rate regime.

S5.4 Supporting Characterization Results

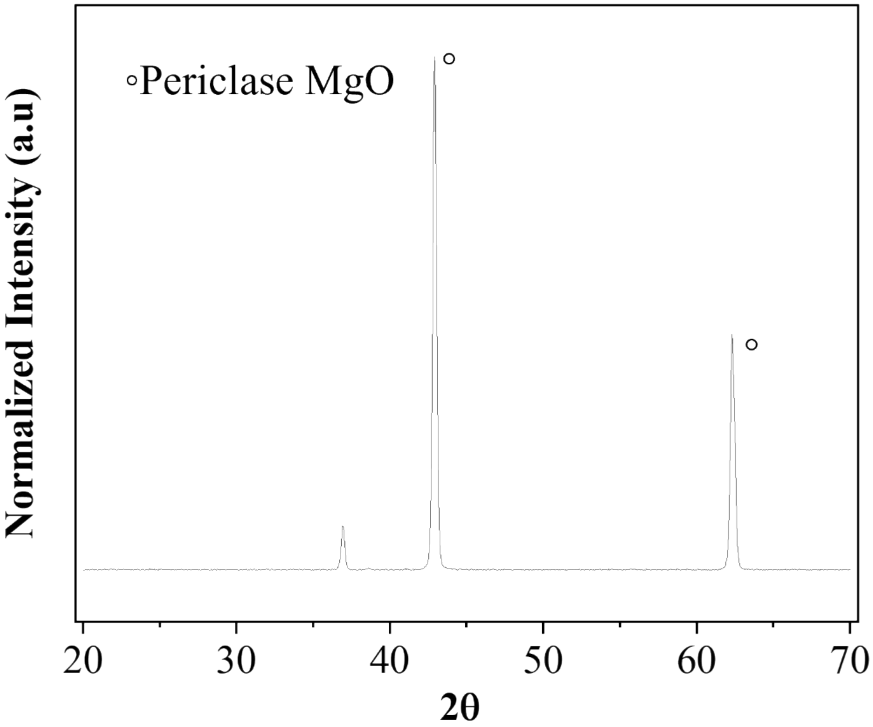

Figure S23: XRD Spectra of fresh MgO, identified as periclase MgO and displaying no impurities.

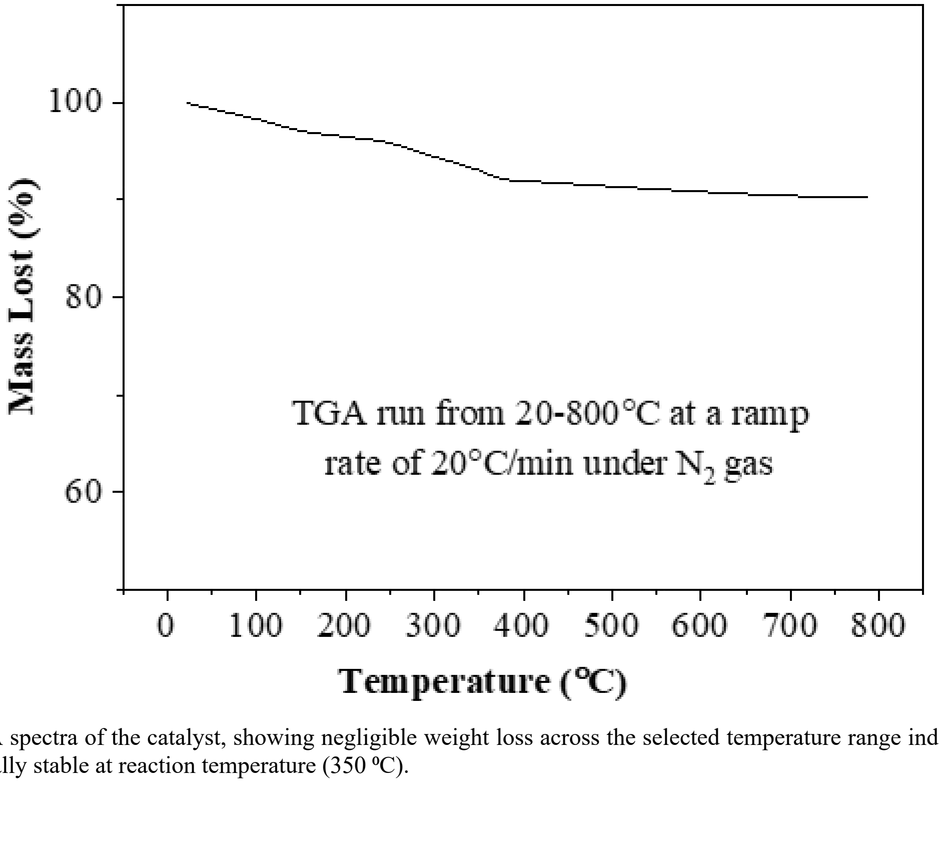

Figure S24: TGA spectra of the catalyst, showing negligible weight loss across the selected temperature range indicating that the catalyst is thermally stable at reaction temperature (350 ºC).

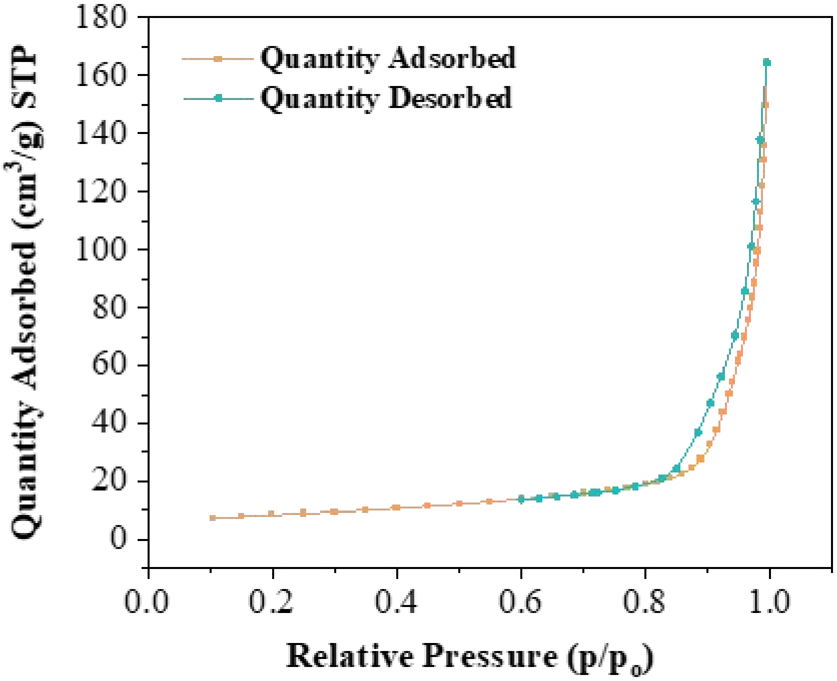

Figure S25: BET adsorption and desorption curves for MgO; the catalyst has a surface area of 29m^2/g.

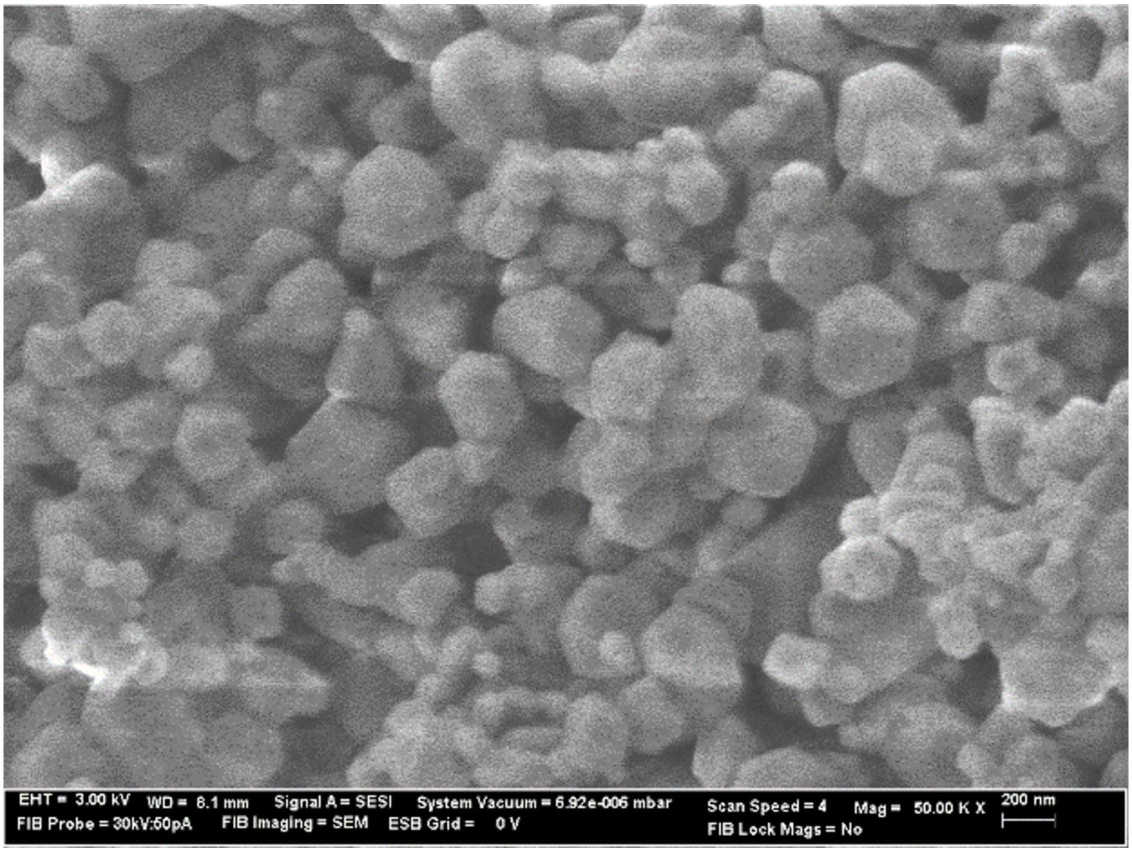

Figure S26: SEM image of MgO taken at Mag of 50kx and EHT of 3.00 kV.

S5.5 Calibration Curves for Quantification of Reactants and Products

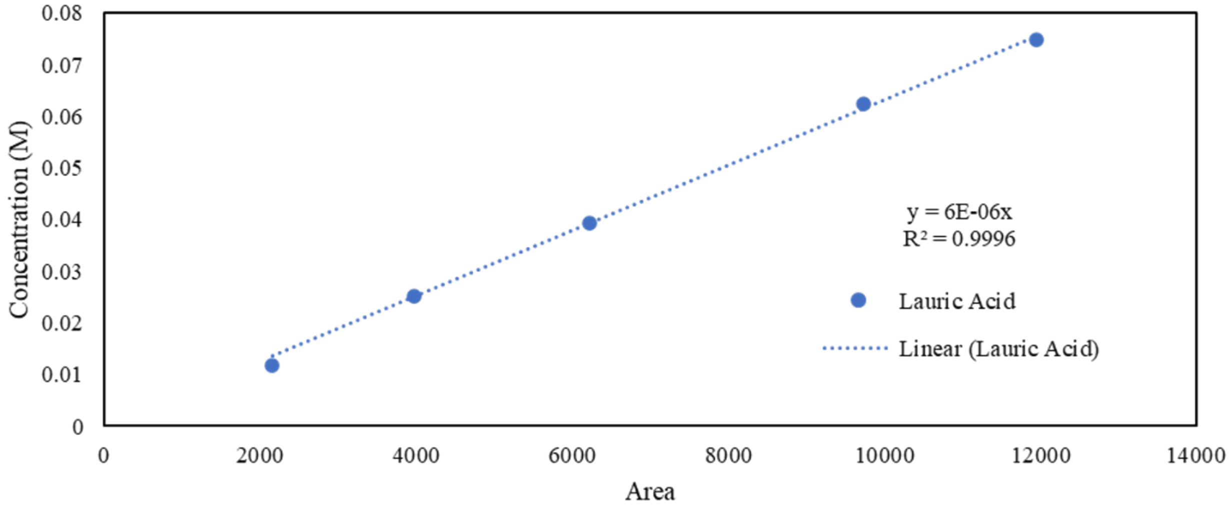

Figure S27: GC calibration curve used for quantification of the reactant lauric acid.

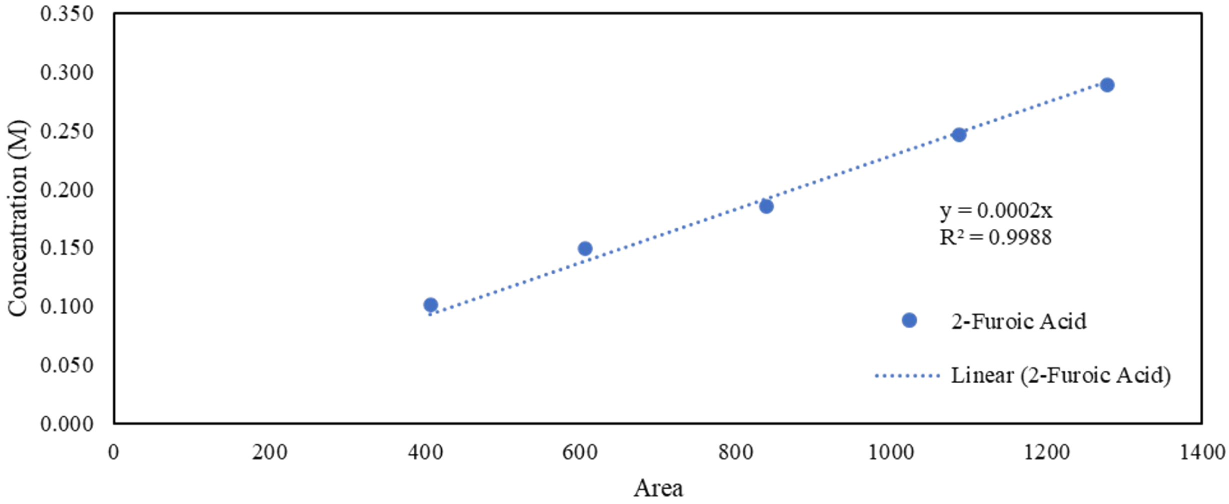

Figure S28: GC calibration curve used for quantification of the reactant 2-furoic acid.

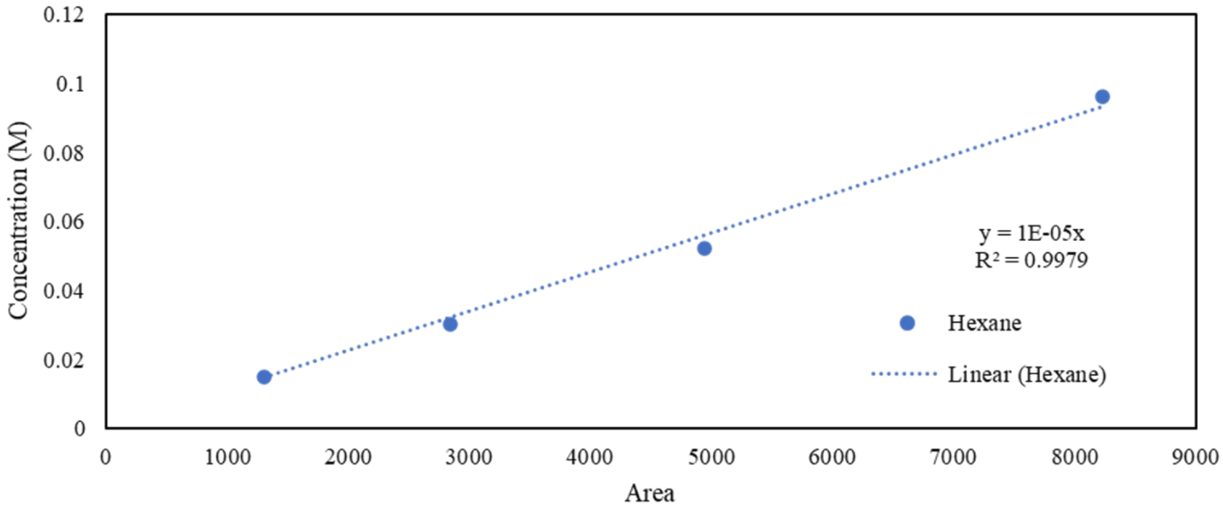

Figure S29: GC calibration curve used for quantification of the product 2-dodecanoyl furan derived from a calibration of hexane using the effective carbon number method.

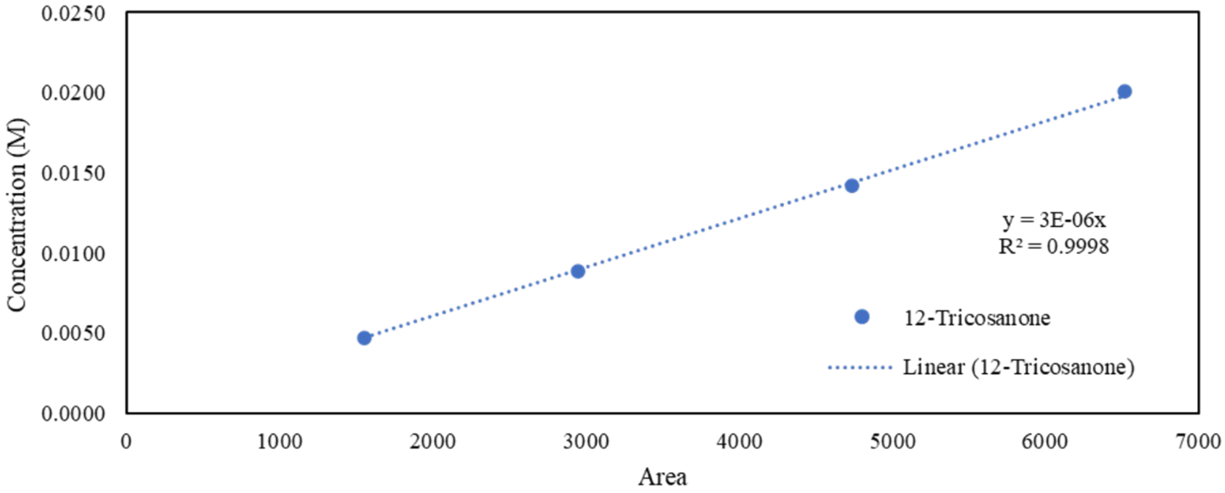

Figure S30: GC calibration curve used for quantification of the product 12-tricosanone.

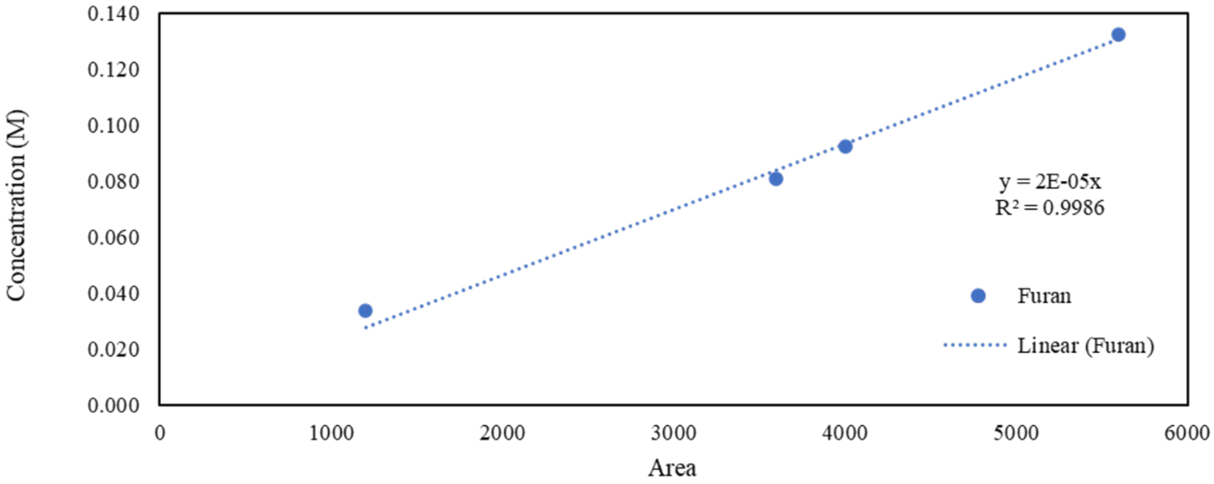

Figure S31: GC calibration curve used for quantification of the product furan.

S6 References

1. M. Cohen and D. G. Vlachos, *Industrial & Engineering Chemistry Research*, 2022, **61**, 5117-5128.
2. C. S. Barnes, DOI: http://localhost/files/tq57ns942Honors College Thesis, Oregon State University.
3. A. Burcat, in *Combustion Chemistry*, ed. W. C. Gardiner, Springer New York, New York, NY, 1984, DOI: 10.1007/978-1-4684-0186-8_8, pp. 455-473.
4. C. W. Gao, J. W. Allen, W. H. Green and R. H. West, *Computer Physics Communications*, 2016, **203**, 212-225.
5. Y. Chung, R. J. Gillis and W. H. Green, *AIChE Journal*, 2020, **66**, e16976.
6. 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.
7. M. Cohen, T. Goculdas and D. G. Vlachos, *Journal*, 2022, Mendeley Data, V1, doi: 10.17632/86vkrpvbr4.1.
8. W. Chen, M. Cohen, K. Yu, H.-L. Wang, W. Zheng and D. G. Vlachos, *Chemical Engineering Science*, 2021, **237**, 116534.
9. S. Rangarajan, T. Kaminski, E. Van Wyk, A. Bhan and P. Daoutidis, *Computers & Chemical Engineering*, 2014, **64**, 124-137.
10. L. J. Broadbelt, S. M. Stark and M. T. Klein, *Industrial & Engineering Chemistry Research*, 1994, **33**, 790-799.
11. A. Ratkiewicz and T. N. Truong, *Journal of Chemical Information and Computer Sciences*, 2003, **43**, 36-44.
12. F. P. Di Maio and P. G. Lignola, *Chemical Engineering Science*, 1992, **47**, 2713-2718.
13. N. A. Bhore, M. T. Klein and K. B. Bischoff, *Industrial & Engineering Chemistry Research*, 1990, **29**, 313-316.
14. M. T. Klein, Z. Hou and C. Bennett, *Energy & Fuels*, 2012, **26**, 52-54.
15. J. H. Miller, L. Bui and A. Bhan, *Reaction Chemistry & Engineering*, 2019, **4**, 784-805.