Chemical Networks from Scratch with Reaction Prediction and Kinetics-Guided Exploration

Michael Woulfe and Brett M. Savoie*

Davidson School of Chemical Engineering, Purdue University, West Lafayette, IN, 47906

E-mail: bsavoie@purdue.edu

Abstract

Algorithmic reaction explorations based on transition state searches can now routinely predict relatively short reaction sequences involving small molecules. However, applying these algorithms to deeper chemical reaction network (CRN) exploration still requires the development of more efficient and accurate exploration policies. Here, an exploration algorithm, which we name Yet Another Kinetic Strategy (YAKS), is demonstrated that uses microkinetic simulations of the nascent network to achieve cost-effective and deep network exploration. Key features of the algorithm are the automatic incorporation of bimolecular reactions between network intermediates, compatibility with short-lived but kinetically important species, and the incorporation of rate uncertainty into the exploration policy. In validation case studies of glucose pyrolysis, the algorithm rediscovers reaction pathways previously discovered by heuristic exploration policies and also elucidates new reaction pathways to experimentally obtained products. The resulting CRN is the first to connect all major experimental pyrolysis products to glucose. Additional case studies are presented that investigate the role of reaction rules, rate uncertainty, and bimolecular reactions. These case studies show that naïve exponential growth estimates can vastly overestimate the actual number of kinetically relevant pathways in physical reaction networks. In light of this, further improvements in exploration policies and reaction prediction algorithms make it feasible that CRNs might soon be routinely predictable in many contexts.

1 Introduction

Reaction prediction methods with minimal heuristic guidance have recently achieved qualitative improvements in accuracy, cost, and throughput that make predicting relatively short reaction sequences involving small molecules routine in many scenarios.[1–11] Although emerging strategies vary in detail, they all ultimately rely on characterizing the transition states of prospective reactions to determine reaction outcomes. In this, the field as a whole has benefited from new low-cost potential energy surfaces,[12–14] double-ended algorithm refinement including string and band methods,[15–18] and ongoing developments in machine learning (ML).[19–24] Nevertheless, even as it has become possible to predict the few-step reactivity of smaller reactants, more sophisticated network exploration methods are still required to manage the exponential explosion of potential reactions with respect to network size. General solutions for bridging this gap between small-scale reaction prediction and the larger reaction network prediction problem have yet to emerge.

A chemical reaction network (CRN) is composed of the minimal set of molecular species (i.e., network nodes) and reactions (i.e., network edges) necessary to accurately model the concentration fluxes of a chemical process (Fig. 1).[25] In practice, the whole CRN doesn't emerge fully formed, and its elaboration is often a painstaking and haphazard process. The general problem of CRN exploration consists of discovering the full CRN starting from a set of initial conditions (Fig. 1A). With the development of *de novo* reaction exploration methods, more systematic explorations of CRNs have become possible with the aspiration of eventually being able to predict CRNs from scratch.

However, as the CRN grows, so does the potential range of reactions and intermediates. Every reaction exploration yields a new set of products that can serve as potential reactants for further exploration, or "terminal" nodes in the graph terminology owing to their position on the edge of the network with no outward reaction paths (shown as green in Fig. 1A). The number of potential unimolecular reactions to explore per terminal node scales factorially in the worst case with respect to molecular size, making it imperative to selectively sample terminal nodes for further exploration. Including bimolecular reactions amplifies the problem as the number of unique bimolecular pairs grows quadratically with network size, with each pair having (worst case) factorial scaling with respect to their combined size (Fig. 1B). Selecting terminal nodes for further exploration is further complicated by the fact that many important intermediates are short-lived and can be easily overlooked by naïve greedy algorithms. For example, this means that exploration algorithms cannot trivially filter single-step endergonic reactions because they may be consequential upon further exploration (Fig. 1C). Finally, computational reaction exploration carries unavoidable errors that must be propagated through exponential rate equations. As the CRN deepens, and depending on the network topology and relevant temperatures, it becomes increasingly unrealistic to model concentration fluxes without error estimates (Fig. 1D). These three problems–prioritizing reaction exploration amongst possible terminal nodes and bimolecular reactions, retention of short-lived but kinetically important intermediates, and uncertainty propagation–constitute a minimum set of challenges for any general CRN exploration algorithm.

In response, various network-level exploration algorithms have been developed that manage the trade-offs of deep CRN exploration in different ways.[26–38] Recent algorithms include the *ab initio* nanoreactor and its descendants that use reactive molecular dynamics simulations on approximate potential energy surfaces under conditions that accelerate reaction observations.[29,35–38] Instead of using low-level quantum chemistry, stochastic surface walking with neural network (SSW-NN), uses a system-specific neural-network potential energy surface (PES) and biased potential-climbing

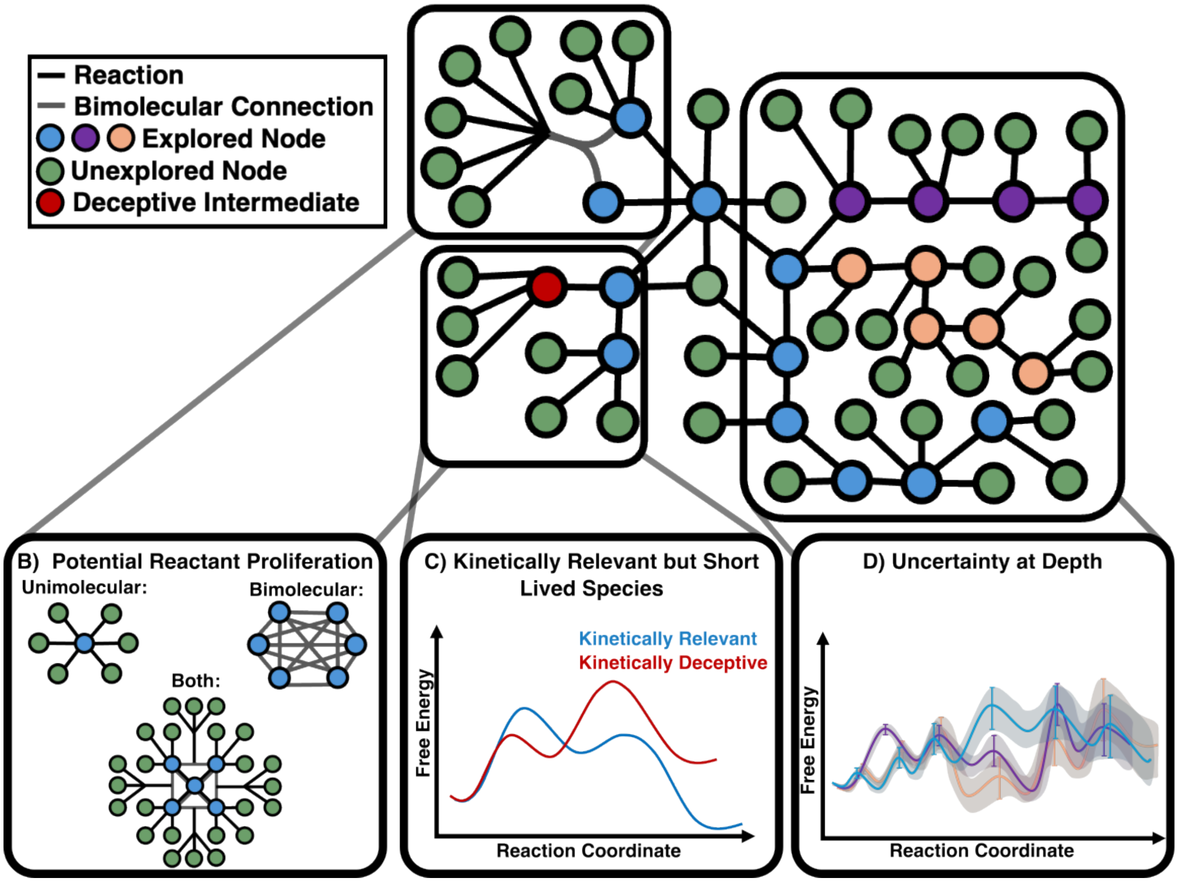

Figure 1: Core challenges for chemical reaction network exploration. A) A sample chemical reaction network with highlighted challenges of deep exploration. B) The number of potential bond rearrangements grows with the number of atoms in the system. For unimolecular reactions this scaling is with respect to the number of atoms in the molecule; for bimolecular reactions the space expands with all possible combinations of reactants and their combined numbers of atoms. In an exhaustive search, if each reactant generates five products, the fifth exploration step will contain 3906 total species, of which over 7.5 million bimolecular reactant pairs could be explored. C) Highlights the harm of premature reaction pruning. Kinetically accessible endergonic intermediates can prove critical to the kinetics of a system and should be retained. D) Failing to propagate uncertainties can obfuscate the most kinetically relevant reaction pathway. Concentration flux variation due to reaction rate uncertainty increases as CRNs deepen.

to explore plausible reaction sequences.[30] Relevant to the case studies presented here, SSW-NN has been used to discover several novel low-barrier pathways for glucose pyrolysis.[8] Broadbelt popularized kinetics-guided network exploration with an algorithm to explore combustion systems based on kinetic changes upon species addition[39] and later also applied a similar kinetics-guided exploration to glucose pyrolysis.[40–42] Kinetics-guided algorithms have undergone continuous development since their introduction owing to their relatively systematic approach for adding new species and reactions to the CRN.[43,44] For example, Reaction Mechanism Generator (RMG) uses a similar approach for defining an expanding "core" species when growing a reaction network.[44] Most recently, Reiher's group has developed two algorithms based on monitoring concentration fluxes within partially explored reaction networks to select intermediates for further exploration.[28] The most recent iteration is kinetics-interlaced exploration algorithm (KIEA) that iteratively conducts sensitivity analysis on the kinetics of the network, prunes inaccessible pathways, and refines important pathways at higher levels of theory.[33] Our group has also leveraged a kinetics-guided policy to automate unimolecular exploration with a modified Djistkra algorithm (MDA) that used the activation energy of the rate-limiting formation step as a cost function for node selection. Combined with comprehensive reaction exploration, this simple heuristic algorithm elucidated several lower barrier pathways to terminal products missed by earlier glucose pyrolysis studies.[26]

While all of these exploration algorithms have found use in specific contexts, none offer generic solutions to the CRN prediction problem (Fig. 1). Many of the CRN exploration algorithms face common difficulties: sampling bias, computational expense, and limited transferability to new systems. Even with approximate TS methods or relying on reaction templates, exhaustively searching through all possible reactions and intermediates in a CRN becomes intractable after only a few exploration steps, particularly with larger molecules.[45] Available CRN exploration algorithms that are meant to prioritize reactions when exploring deep reaction sequences still typically run into cost limitations. System-specific heuristic exploration algorithms and ML-based methods may reduce cost or expand the degree of exploration scope, but these are largely nontransferable and can show reaction biases or other uncontrolled errors.[43,44,46–48] In contrast, kinetics-based algorithms are at least in principle systematically improvable with perfect information, but in practice can be prone to prioritize greedy searches that follow the low barrier pathway to the exclusion of others. For example, the overall lowest barrier pathway may be hidden behind a slow reaction that microkinetic modeling may overlook while the network is still being explored.

Here, we develop a new network exploration algorithm that we call Yet Another Kinetic Strategy (YAKS), owing to its shared conceptual elements with prior work. YAKS is also thematic with the Yet Another Reaction Prediction (YARP) method that serves as the reaction prediction engine that we combine here with YAKS. Nevertheless, many aspects of YAKS are unique in implementation and meant to address the challenges associated with the CRN exploration problem as generally as possible. In particular, YAKS can automatically explore both unimolecular and bimolecular search spaces, discover pathways involving local kinetic bottlenecks, and uses concentration-flux uncertainty estimates during exploration. The key aspects of YAKS are simple kinetics-informed rules for selecting reactants for further reaction exploration. These rules are formulated to provide well-defined guarantees on the types of CRN topologies that can be discovered and to be systematically improvable. After describing its implementation details, several YAKS explorations with varying configurations are performed using β-D-Glucose pyrolysis as a model exploration problem. These case studies reveal the important role of uncertainty estimation on deep network explorations and demonstrate that bimolecular reactions can be automatically and tractably handled by YAKS for this system.

2 Methods

This section is organized to first provide a description of the Yet Another Kinetics Strategy (YAKS) algorithm (Subsection 2.1), followed by illustrative thought-experiments and examples for understanding the limitations of the algorithm (Subsection 2.2), an illustration of a YAKS cycle (Subsection 2.3), discussion of termination condition and relevant hyperparameters (Subsection 2.4 and the SI Section 2), the reactivity characterization engine Yet Another Reaction Program (YARP) (Subsection 2.6) and then the computational details associated with the microkinetic modeling and reaction characterizations that are specific to the current case studies.

2.1 Yet Another Kinetic Strategy (YAKS) Stages

YAKS uses a three-stage recurrent cycle to explore CRNs. In the first stage, the kinetics of the available CRN are simulated under application-specific conditions (Fig. 2, **Stage 1**, microkinetic simulations). In the second stage, a selection process is performed that uses the results from the microkinetic simulations to identify a subset of species within the CRN for additional reaction exploration (Fig. 2, **Stage 2**). In the third stage, the reactivities of the selected species are characterized (Fig. 2, **Stage 3**), which results in the addition of new species and reactions (nodes and edges) to the CRN. These stages are then repeated until reaching a user-specified termination condition.

2.1.1 YAKS: Stage 1

In the first stage of the exploration cycle, YAKS conducts a microkinetic simulation of the available CRN using application-specific initial conditions to obtain approximate steady-state concentrations for various species within the CRN. The minimal inputs for microkinetic simulations are the initial concentrations and rate equations for all of the reactions that are being modeled. The method

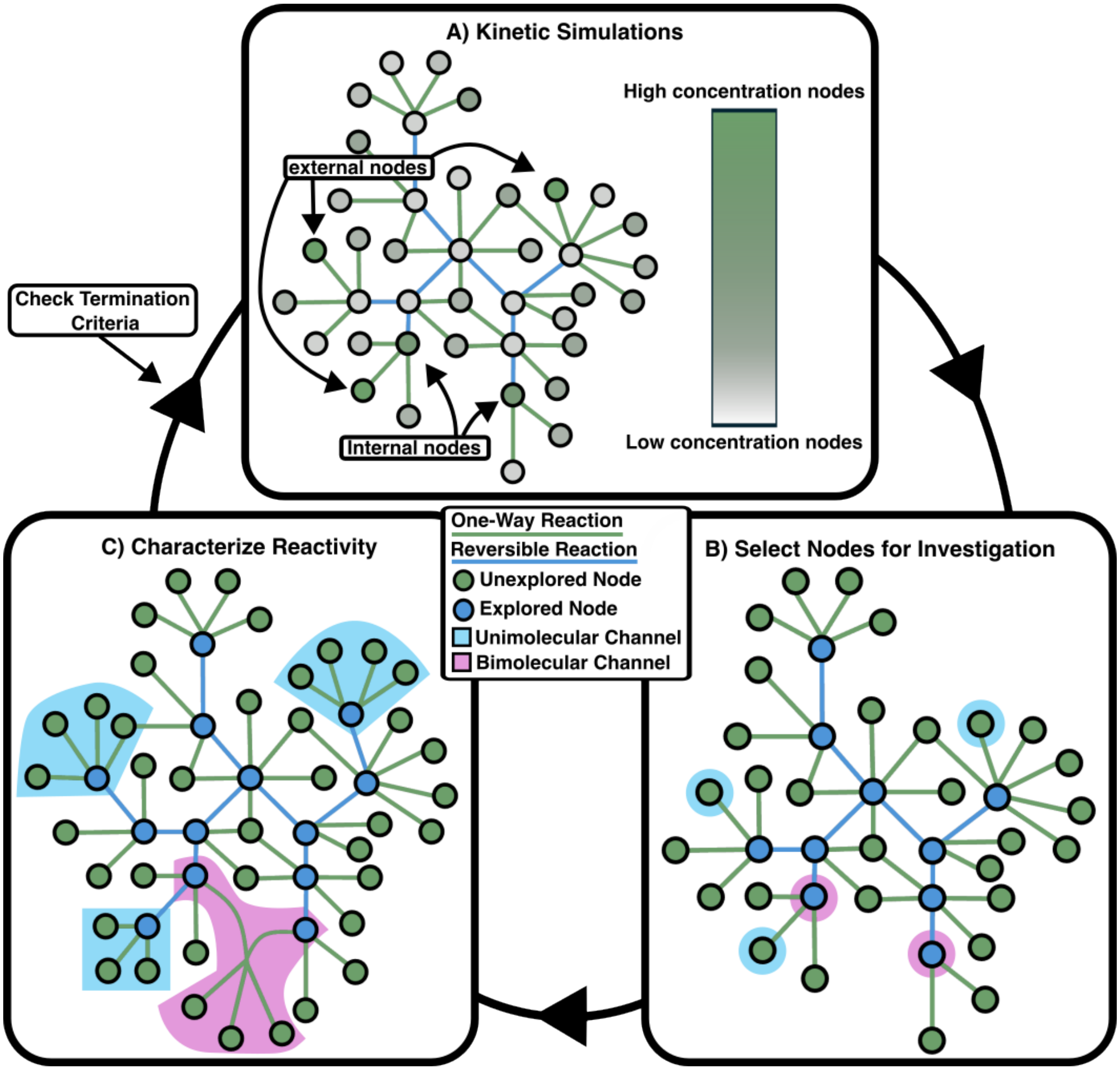

Figure 2: Overview of the Yet Another Kinetic Strategy (YAKS) algorithm. In **Stage 1** (A), microkinetic simulations are performed of the currently available CRN subject to some strategic topology manipulations. In **Stage 2** (B), species are selected for unimolecular (blue) and bimolecular (pink) reactivity exploration based on their steady-state concentration. In **Stage 3** (C), an exploration engine characterizes the reactivity of new reactants in unimolecular and bimolecular scenarios, which expands the CRN and creates the possibility of continuing the YAKS cycle via **Stage 1**.

for generating rate equations (e.g., estimating A-factors and activation energies) is separate from YAKS itself and the manner in which these are calculated in the current case studies will be described later (Subsection 2.6).

One of the key distinctions in YAKS is that the CRN topology is manipulated to bias the pseudo steady-state concentrations from the **Stage 1** microkinetic simulations towards potentially useful intermediates for further reactivity exploration. Topology manipulation occurs in two ways. First, YAKS keep track of which species have already been selected for unimolecular exploration and those that have not. Unless a species has already been selected for unimolecular exploration, its bimolecular reactions are not included in the **Stage 1** microkinetic simulations. The rationale for this is that bimolecular reactivity is more expensive to explore than unimolecular reactivity, so it is advantageous to first search for unimolecular reactions that might siphon off concentration and forestall bimolecular reactivity. Second, YAKS considers the graphical distance, d_0, of each species in the CRN from the nearest species that has yet to undergo a reactivity exploration. That is, $d_0 = 0$ for any species in the CRN that has yet to be selected in **Stage 2** as a reactant for exploration, and d_0 is defined for all other species as the minimum number of reactions required to reach a $d_0 = 0$ species. YAKS uses d_0 to manipulate the topology of the network for all species in the CRN for which $d_0 \leq n_\mathrm{d}$, where n_d is a hyperparameter for the exploration that is greater than or equal to zero. For species satisfying this condition, no reactions are included in **Stage 1** for which d_0 of the products is larger than d_0 of the reactants. In the simplest case of $n_\mathrm{d} = 0$, this has the effect of excluding the reverse (i.e., “consumption” reactions) for terminal species in the CRN. The rationale for this manipulation is that it allows kinetically relevant endergonic intermediates to be discovered that would be otherwise not collect concentration in a microkinetic simulation inclusive of reverse-reactions.

The manipulation of the CRN topology so that the $d_0 \leq n_\mathrm{d}$ nodes are irreversible concentration sinks means that sufficiently long microkinetic simulations will result in all steady-state concentration accumulating in these species. However in practice, a pseudo steady-state between the low overall barrier $d_0 \leq n_\mathrm{d}$ species and exergonic $d_0 > n_\mathrm{d}$ portions of the network is arrived at very quickly with a much longer time constant associated with the slower equilibration with high barrier $d_0 \leq n_\mathrm{d}$ species (See Fig. S8 for an illustration and additional discussion). That is, the topology manipulation is designed to equilibrate the $d_0 \leq n_\mathrm{d}$ species that are *kinetically* accessible with the $d_0 > n_\mathrm{d}$ species that are both *thermodynamically and kinetically* accessible. For the remainder of the work, we will drop the reference to “pseudo” when referring to steady-state concentration for simplicity.

In **Stage 1**, YAKS also incorporates uncertainty estimates for the steady-state concentrations obtained from microkinetic simulations. These estimates are obtained by resampling the CRN activation energies from independent normally sampled distributions. The default behavior is to use means centered on the values supplied by the reactivity characterization engine (in this case YARP, but they could be from other sources), and standard deviations set by the user. The kinetics of the CRNs are rerun with these resampled rate parameters until converging the rank-ordering of the highest concentration species. In practice, this can involve thousands of microkinetic simulations, but given the relatively low costs of simulations this is not a significant bottleneck for YAKS (SI Fig. S7).

2.1.2 YAKS: Stage 2

In the second stage of the exploration cycle, YAKS selects species from the CRN for further unimolecular and bimolecular reactivity exploration based on the results of the **Stage 1** microkinetic simulations. The default exploration rules are based on the steady-state concentration (c_{ss}) of the species in the CRN, which is a consequence of the **Stage 1** CRN topology manipulation. Other plausible selection criteria are maximum instantaneous flux or maximum concentration, which would perhaps capture transiently important species but are not further explored here.

All species in the CRN are rank-ordered in **Stage 2** by c_{ss}. Different criteria are used to select species for unimolecular exploration versus bimolecular exploration based on the c_{ss} ranking. For unimolecular exploration, the top-n_{uni} species with $d_0 = 0$ are selected for **Stage 3** characterization, where n_{uni} is a user-specified parameter. When $n_{\mathrm{uni}} = 1$, the exploration will only be performed on the species with the highest c_{ss}. Selecting $n_{\mathrm{uni}} > 1$ results in parallel unimolecular explorations of different species in **Stage 3**. This has the practical effect of better utilizing typical high-performance computing resources as well as promoting the discovery of important reaction sequences that proceed through relatively high-barrier intermediates.

Bimolecular reactions introduce additional complexity to CRN exploration but are critical to accurately describe many systems. Possible rules range from neglecting bimolecular reactions entirely, conducting all bimolecular combinations from a core of reactants, to conducting every possible reaction combination between all species in the CRN. As highlighted in Figure 1B, the last option is intractable for large networks, nor does it seem to be physically necessary. YAKS manages this trade-off by restricting bimolecular reactions to species in the network that are sufficiently high in concentration and that have already been explored for unimolecular reactivity (i.e., $d_0 \geq 1$). The rationale for first exploring unimolecular reactivity is that it avoids a premature and expensive bimolecular reactivity exploration if a rapid unimolecular reaction path exists. If two $d_0 \geq 1$ species appear within the top-n_{bi} by c_{ss} ranking, then they are selected for bimolecular reaction characterization in **Stage 3**. The rationale for this is that bimolecular reactivity will be favored between species that maintain high-concentration in spite of available unimolecular reaction channels. With up to n_{bi} new species per exploration step, bimolecular characterizations are limited to $\binom{n_{\mathrm{bi}}}{2}$, or 10 for the YAKS default of $n_{\mathrm{bi}} = 5$. In practice, far fewer bimolecular characterizations will occur if the concentration of intermediates does not sufficiently accumulate so as to satisfy the top-n_{bi} criteria. Apart from these direct bimolecular explorations, YAKS also discovers many bimolecular reactions as the reverse reactions of unimolecular decompositions.

2.1.3 YAKS: Stage 3

In the third stage the exploration cycle, the species that have been identified for unimolecular and bimolecular reactivity exploration are passed to an external reaction exploration engine that returns a set of new reactions involving these species. To be compatible with YAKS, the external engine must return sufficient information to evaluate the rate laws associated with the reaction, such that the microkinetic modeling in **Stage 1** can be performed.

In general, the reaction exploration stage is the most expensive step in exploration and this motivates the choices in **Stages 1-2** to limit the number of species advanced for characterization. Reaction exploration engines can vary from programs that apply a fixed set of contextual reaction templates, to programs that perform searches based on activation energy characterizations. YAKS was developed to be fully compatible with the Yet Another Reaction Program (YARP), which is a reaction prediction engine developed by our group that uses generic graphical rules to enumerate potential products associated with inputted reactants and then uses accelerated activation energy characterizations to predict reactions. **Stage 3** concludes with a clean-up phase to ensure that there are no duplicated reactions, updates a list of reactions that have been attempted but are discovered to be infeasible, and the addition of new products and reactions to the CRN. Returning to **Stage 1**, YAKS incrementally explores the CRN, seeding products from one generation as reactants for a subsequent explorations.

2.2 Motivating Thought Experiments

The YAKS stages are meant to effectively coarse-grain the dynamics of the real CRN. To understand this, the following thought experiments might be useful. In these thought experiments we will assume that there is an oracle that can reveal the reactivity of any species in the CRN (e.g., all associated unimolecular and bimolecular reactions, such as occurs in **Stage 3** of YAKS), and the goal is to query this oracle as little as possible.

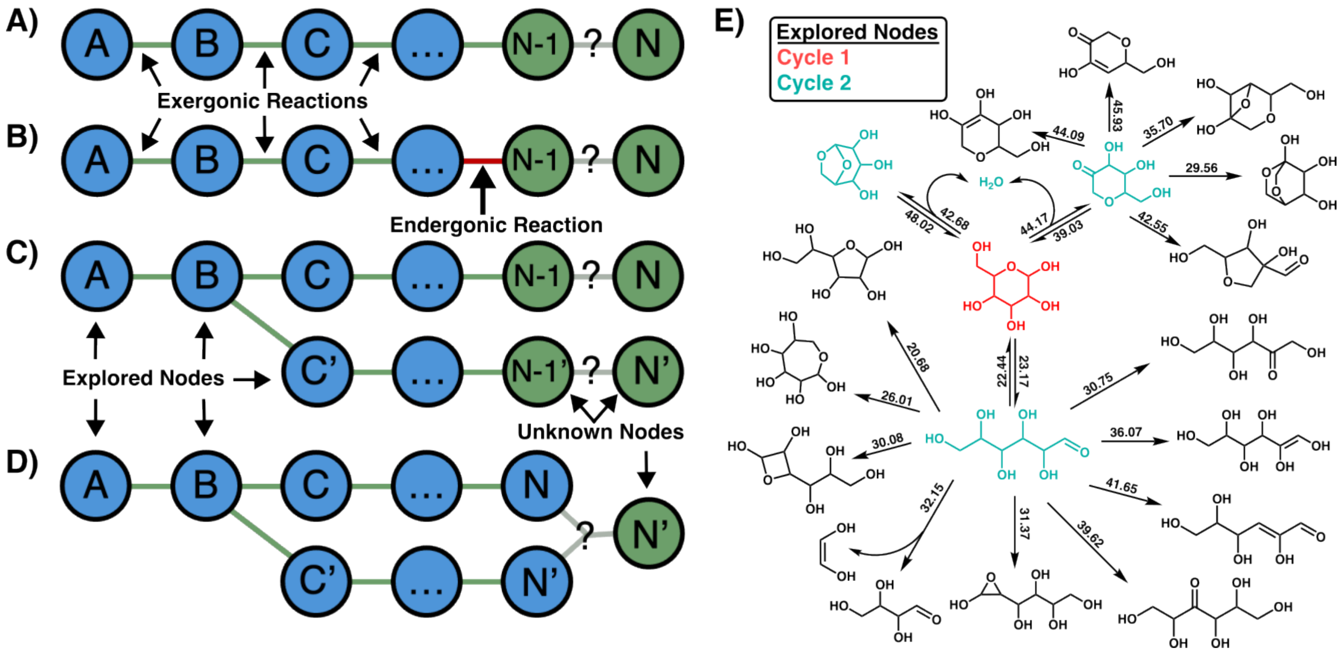

Figure 3: Motivating thought experiments. (A) An idealized linear CRN with unbroken exergonic reactions from source, A, to terminal species, N. A reaction is missing between species N and N-1. (B) The same linear CRN as in (A), except now N-1 has a higher free energy than N-2. This example motivates the topology manipulations performed by YAKS. (C) The same CRN as in (A), except that an isoenergetic branch has been added at species B. This example motivates the parallel beam search performed by YAKS. (D) The same CRN as in (C), except that the branch termini can participate in a favorable bimolecular reaction. This example motivates the the bimolecular reaction rule used by YAKS. (E) The intial 2 cycles of YAKS glucose exploration. After discovering only three accessible reactions, YAKS simulates the kinetics and then conducts reactivity calculations for the four Cycle 2 reactants.

To start, suppose we have a "complete" CRN consisting of a sequence of unimolecular reactions $A \rightarrow B \rightarrow C \ldots \rightarrow N-1 \rightarrow N$, where all net fluxes flow from A to N (i.e., all reactions are exergonic), and completeness means that all intermediates and reactions are present such that starting from relevant initial conditions the transient and steady-state concentration distributions match the ground truth for all appreciable species (Fig. 3A). Now, suppose we were to delete the species with the highest steady-state concentration, N, from the network along with all of its reactions. Where would be the best place to look for the missing node? If microkinetic simulations were performed, then the steady-state concentration that had been in N would have to be redistributed upstream with the greatest excess in N-1 if the CRN were as simple as described. Thus, the best place to look for the missing node (N) would be by exploring the reactivity of the node with the highest steady-state concentration. If N-1 had also been deleted, then the largest excess concentration would occur in N-2 and this would be the species whose reactivity was most promising to explore.

By induction, the same logic could be applied all the way back to a fully pruned CRN starting from A. Working in reverse, the exploration would proceed by characterizing the reactivity of A, which, by the definition of this CRN would reveal reactions to (potentially many) irrelevant endergonic species and B. Microkinetic simulations of this expanded network would have concentration pooling in B, whose reactivity would be explored, leading to the discovery of C, and so on until discovering N, after which no more exergonic species would be discovered.

Now let's consider a modified thought-experiment with a realistic complication. Suppose we again have a complete CRN consisting of a linear sequence of reactions $A \rightarrow B \rightarrow C \ldots \rightarrow N-1 \rightarrow N$, however N-1 is endergonic with respect to N-2 and N (Fig. 3B). In other words, N-1 is a short-lived but kinetically crucial intermediate. Now what would occur if we were to delete N from the network and perform microkinetic modeling? The largest concentration increase would occur in N-2, *not N-1.* So in this case, the concentration increase would occur in the next-nearest neighbor of the species that actually needs to be explored, rather than in the species with the concentration accumulation itself. Thus, if we adopted a policy of exploring the reactivity of the nodes with highest concentration and their nearest neighbors, we would still be able to rediscover N, while a purely greedy exploration would not. By induction, this would again work even if we had to start from the fully pruned CRN consisting only of A. In general, the number of endergonic intermediates determines how far away within a linear CRN that the concentration will pool from the species that are actually relevant for further reactivity exploration. In the worst case, concentration could accumulate in species arbitrarily far away from the deleted species. However, YAKS is guided by the heuristic that physical reaction networks tend to only include a relatively small number of consecutive endergonic reaction steps. In its default application, YAKS assumes that only one kinetically favorable yet endergonic reaction step occurs in the CRN, which means that it explores the reactivity of the highest c_{ss} nodes and their nearest neighbors. This is the rationale for the default $n_{\mathrm{d}} = 0$ hyperparameter used in **Stage 1** of YAKS to exclude consumption reactions for terminal nodes. If sequential endergonic reaction steps are sought, then the next-nearest neighbors would also be included in the reactivity exploration (i.e., $n_{\mathrm{d}} = 1$), but also with the associated increase in cost.

The thought experiment can be extended to include branches. Suppose the CRN were modified to include an isoenergetic branch such that $A \rightarrow B(\rightarrow C' \rightarrow \ldots \rightarrow N') \rightarrow C \ldots \rightarrow N$, where the parentheses indicate a branch off of B consisting of an exergonic sequence of reactions leading eventually to N' which is isoenergetic with N (Fig. 3C). If N and N' were both deleted, then there would be two sites of excess concentration along each branch. In the simplest case, they could be investigated in parallel by using an algorithm that searched the reactivity of the top-n species by concentration at each stage, rather than just the top species. The situation becomes more complicated where the branch occurs (and more generally wherever the branches interact with one another, such as through bimolecular reactions). Suppose we pruned the network back to the branch with $A \rightarrow B(\rightarrow C') \rightarrow C$. Unless the two branches were identical in energy, one of them would be explored at the expense of the other until fully exploring to N (and possibly other irrelevant side-reactions) then backtracking to C'. Alternatively, if multiple points of concentration accumulation are investigated for reactivity exploration at every step, then both branches could be explored. This is the rationale for selecting $n_{\mathrm{uni}} > 1$ in **Stage 2** of YAKS. The YAKS default of $n_{\mathrm{uni}} = 5$ used here means that YAKS could simultaneously explore up to five separate network branches at a time.

Finally, the thought experiment can be extended to include bimolecular reactions. Suppose the CRN were modified to include an important bimolecular reaction between the ends of each isoenergetic branch such that $A \to B(\to C' \to \ldots \to N') \to C \ldots \to N \cup N + N' \to P_{NN'}$, where $P_{NN'}$ is an exergonic product formed from reacting the two isoenergetic species N and N' (Fig. 3D). If the bimolecular reaction is deleted from the CRN, then there would be excess concentration in N and N' and it could be rediscovered by allowing reactions between the top-2 species with highest steady-state concentration. If N and N' were also deleted from the network, then N-1 and N'-1 would be the sites of accumulation. Here, it would be wasteful to bimolecularly react N-1 and N'-1 because there is a unimolecular for each that is relatively inexpensive to discover. This is the rationale in YAKS for limiting bimolecular explorations to species that have already been unimolecularly characterized.

These thought-experiments highlight the behaviors of partially explored CRNs under idealized scenarios. Under such conditions, YAKS provides discoverability guarantees of reaction sequences involving up to n_{d} exergonic intermediates, the discovery of unimolecular branch points with up to n_{uni} exergonic products, and the discovery of up to $\binom{n_{\mathrm{bi}}}{2}$ bimolecular channels per exploration step. No guarantees are possible when the CRNs deviate from these idealized topologies. One such complication is CRNs with physically relevant branches with large energy differences (e.g., suppose that N and N' had very different energies in the last thought experiment). However, the benchmarks of this work support the conclusion that the YAKS exploration heuristics are still useful for economically exploring more complex networks.

2.3 Illustrative Cycle

For the sake of illustration, the first two YAKS exploration cycles for glucose pyrolysis are briefly explained (Fig. 3E). Initially, the CRN consists only of the initial reactants and reactions provided by the user, which in this case would be D-glucose. **Stages 1-2** are trivial when starting with a lone reactant, because there is no CRN yet and the reactant is the only species with concentration. If the user had started with a subset of known reactions, then a non-trivial **Stage 1** would need to be performed. Thus the first YAKS cycle for D-glucose trivially advances to **Stage 3** and passes D-glucose itself to YARP for unimolecular reactivity characterization. The results from **Stage 3** expand the CRN about D-glucose.

In **Stage 1** of the second cycle, $d_0 = 1$ for D-glucose and $d_0 = 0$ for all of the newly discovered products. The CRN used for microkinetic simulations in this stage consists of all the reactions involving D-glucose, but none of the reverse reactions involving the $d_0 = 0$ species as reactants. Because of the great difference between the low-barrier reaction rate and all other reactions, the uncertainty in the rates plays no role in rank ordering the species by c_{ss}, but it potentially would for more complicated CRNs. In **Stage 2**, the n_{uni} products of D-glucose with the highest c_{ss} are selected for unimolecular characterization in **Stage 3**. Because of the d_0-rule, the rank ordering of c_{ss} is determined only by activation energy at this stage and not by free energies of reaction. No species are selected for bimolecular characterization, because D-glucose is the only available $d \neq 0$ species and it has no steady-state concentration due to the d_0-rule and the kinetically accessible intermediates. After **Stage 3** of the second cycle, a non-trivial CRN topology emerges with several branches and over 25 $d_0 = 0$ products upon entering the third cycle.

2.4 YAKS Termination Conditions

YAKS explorations can terminate based on a number of criteria, such as reaching a fixed depth, encountering no external nodes in the top-n species, the discovery of a particular product, reaching a minimum confidence threshold, a computational time limit, or even a combination of several methods. Apart from two exceptions, the case-studies reported here were terminated once the top-5 unexplored species consisted of less than 30% of the overall concentration of the system.

One noiseless unimolecular case-study terminated because the top-5 highest concentration species had all previously been explored. The uncertainty-guided case-study terminated at a fixed depth of 20 cycles.

2.5 Comparison of YAKS with Other Methods

The distinguishing features of YAKS are the topology manipulation of the partially explored network, the maintenance of n_{uni} parallel search beams across the network, and the even-handed incorporation of bimolecular and unimolecular reactions based on intermediate steady-state concentrations. However, the use of microkinetic simulations is shared by many other algorithms. The most modern example is the Kinetics-interlaced exploration algorithm (KIEA), which explores CRNs based on microkinetic modeling and quantum chemistry based reactivity characterization in a gradual fashion.[28] However, KIEA approaches exploration protocols much differently. YAKS considers all reactions within the CRN at every microkinetic simulation step, which allows for backtracking, while KIEA permanently prunes any species with negligible concentration flux in future microkinetic modeling steps. KIEA also relies on manually set thresholds based on mean and maximum concentration fluxes to seed species for future reactivity characterization. YAKS uses a relative rule to characterize important species while also limiting computational costs. Additional detailed comparisons can be found in the SI (Section 5).

2.6 Yet Another Reaction Program (YARP)

The YAKS algorithm identifies intermediates and reactants in the system whose reactivity needs to be characterized, but it still relies on an external engine to actually do this characterization. We have designed YAKS with modularity in mind, such that users could for example query their own library of reaction templates for this step. Here, all bimolecular and unimolecular reaction explorations were performed with the YARP 2.0 package.[11,49,50] YARP is a method developed by our group for TS-based and template-free reaction exploration. The reader is directed to the dedicated methods publications for a detailed review of YARP, here we briefly summarize its general features and the settings specific to this study.

To characterize reaction pathways, YARP enumerates all possible products using generic graph-based elementary reaction steps (ERS). These ERSs are defined in terms of a fixed number of bond-breaks and bond-formations, such as break 2 bonds and form 2 bond (b2f2). For neutral closed-shell organic systems such as glucose, the simplest ERS that yields closed-shell products is b2f2. In our earlier glucose study we explored conditional b3f3 (Cb3f3), both b2f2 reactions and b3f3 reactions that involved at least one π-bond breaking.[26] The latter was empirically motivated by earlier studies showing that b3f3 reactions exclusively involving σ-bonds yielded very few competitive reactions.[11,50] These ERSs were retained here. From the reactant and ERS-generated product graphs, YARP applies standardized routines to generate reactant and product conformations and localize transition states. As glucose pyrolysis liberates water through many channels, it is important to also consider water-catalyzed proton transfers in the exploration. Here, all reactions involving at least one proton transfer were separately tested in water-catalyzed and non-catalyzed scenarios. The protocol for water-catalyzed convergence has been previously described and involves re-performing the TS localization as a b3f3 (or b4f4) water-mediated reaction rather than a b2f2 (or b3f3) uncatalyzed proton transfer.[45] After TS convergence, intrinsic reaction coordinate (IRC) calculations were performed on all TS to confirm that they corresponded to the intended reactant-product pair. Final activation energies were calculated as the free energy difference between the lowest energy TS and the lowest energy conformation(s) of the isolated reactant(s).

Several YARP settings were adjusted to be more permissive than in the earlier glucose study. These changes make the reactions explored here a superset of those explored in the earlier study. In addition to the Cb3f3 ERS described in the last paragraph, all σ-bond b3f3 reactions were also characterized to investigate whether concerted reaction mechanisms missed by the earlier Cb3f3 exploration are potentially consequential. The earlier study also pre-filtered reactions with an enthalpy of reaction (ΔH_r) $>$ 20 kcal/mol to limit kinetically irrelevant explorations.[26,51] Here the ΔH_r filter was dispensed with leading to some notable differences, including the discovery of a D-Glucose dehydration reaction to form Levoglucosan with a barrier of 42.67 kcal/mol that was previously missed. To avoid more expensive DFT-level TS optimizations, YARP can also optionally pre-filter reactions based on low-level estimates of the activation energy. Here, any TS with a barrier $>$ 65 kcal/mol at the GFN2-xTB level was excluded from DFT-level exploration, which is 15 kcal/mol higher than the previous study.

2.7 Computational Details

Reaction characterization was performed by YARP v2.0.[50] The Conformer-Rotamer Ensemble Sampling Tool (CREST)[52] was used to generate reactant and product conformers with the GFN2-xTB potential,[12] then joint-optimization and conformer selection routines were used to align and select up to five conformers per attempted reaction.[53] Double-ended growing string searches were used to generate approximate TSs using nine images per string.[17,54,55] The approximate TSs were then optimized to saddle points using Berny optimization as implemented in Gaussian 16.10.[56] GFN2-xTB was used as a low-level method for GSM and Berny optimization prior to a final DFT-level Berny optimization. All GFN2-xTB calculations were performed with the xTB program (version 6.4.0). DFT calculations were carried out using Gaussian 16.10. Unless stated otherwise, all results are reported using optimized geometries, energies, and frequencies calculated at the B3LYP-D3/TZVP level of theory, all energy units are kcal/mol, and thermally dependent properties use 298.15 K as a reference temperature. This is the same level of theory used in earlier studies and so has been adopted here. Energies are generally reported to two decimal places for reproducibility, but our previous benchmarks on the accuracy of DFT and conformational uncertainty for similar classes of reactions suggest that these values are only accurate to within ∼3 kcal/mol on average, and so the discussion focuses on differences on that scale or larger.[53,57] These errors are uncorrelated and together imply a possible 4.25 kcal/mol error. This study used these two uncertainty regimes, corresponding to DFT only uncertainties and DFT and conformational sampling uncertainties combined.

The version of Cantera used in this study is 2.6.0.[58] Guides on how to use Cantera are available at [https://cantera.org] with documentation at [https://zenodo.org/record/6387882]. Under default conditions, our microkinetic modeling simulates an ideal gas mixture in an isothermal reactor. For this study, the system was modeled at 623 K and 101.3 kPa. Microkinetic simulations ran for 1200 0.1 second time steps, sufficient time for the system to resolve towards a pseudo-steady state between the kinetically accessible terminal nodes in the CRN and the kinetically and thermodynamically accessible internal nodes. Cantera supports more complicated reactors, but this setup is inexpensive and proved sufficient to supersede previous glucose pyrolysis explorations. At every time integration step, Cantera updates the system density, mean molecular weight, internal energy, entropy, and enthalpy as well as all mole fractions and chemical potentials. Cantera tracks species production/destruction rates defined as $\frac{dc_i}{dt} = R_i$ where c_i is the molar volume in units of $\frac{mol}{m^3}$ and R_i is the production rate of volume-specific species in units of $\frac{mol}{m^3 s}$. Cantera further tracks individual reaction rates, defined as $R_i = \sum_{j=1}^{N_{rxns}} v_{ij} r_j$ where v_{ij} is the stoichiometric coefficient of species i in reaction j and r_j is the volume-specific stoichiometric reaction rate for reaction j. Individual reaction fluxes are used to map the highest flux pathways through the network during later uncertainty analysis and pruning (Fig. 7). The primary Cantera reaction type used by YAKS is the elementary reaction, which relies on Transition State Theory and the Arrhenius equation to calculate rate constants. The Arrhenius equation is of the form $k = AT^b e^{\frac{-E_a}{RT}}$, where k is the rate constant, A is the pre-exponential factor, T is the simulation temperature, b is the temperature exponent, E_a is the activation energy, and R is the universal gas constant. No additional temperature dependency was assumed, so b was set to 0 in all simulations. A was approximated as $\frac{k_B T}{h}$ where k_B is the Boltzmann constant and h is Planck's constant. The free energy of activation calculated by YARP was assigned as E_a for each reaction.

3 Results and Discussion

To directly compare with previous studies, YAKS was applied to explore the reaction networks associated with D-Glucose pyrolysis.[26,46,47,59] The ultimate goal of this case-study is to elaborate a network consisting of low-barrier pathways to the major experimental products of glucose pyrolysis. By mass percent these are hydroxymethylfurfural (HMF), hydroxyacetaldehyde (HAA), furfural (FF) with high yields, and 3-(2H)-furanone (3FO), dihydroxyacetone (DHA), and 3-hydroxy-γ-butyrolactone (HBL) with lower yields.[60] The discovery of pathways to all of these products in a single unified network is still an unresolved problem. Additionally, recent studies have revealed new low barrier pathways to individual products that suggest the individual reaction mechanisms have yet to be established.

This section is organized to first discuss the full D-Glucose pyrolysis CRN discovered by YAKS (i.e., inclusive of all elementary reaction steps, bimolecular reactions, and with flux uncertainty estimates) followed by subsections discussing comparative case studies to investigate the importance of each YAKS component.

3.1 The Overall CRN

The uncertainty-guided Unimolecular Cb3f3 YAKS CRN is shown in Figure 4. This network has been condensed for clarity to show only the three lowest barrier reactions from any node under 45 kcal/mol. After 20 YAKS cycles, the uncertainty-guided CRN included 931 species and 983 unique reactions with activation energies less than 65 kcal/mol. Of these, 756/931 species were not further explored as YAKS did not consider them kinetically relevant (i.e., they never met **Stage 2** selection criteria), 95/931 were intermediates that YAKS selected for only unimolecular reactivity characterization, 3/931 were species that YAKS selected for both unimolecular and bimolecular reactivity characterization, and 80/931 were terminal species that were newly discovered in the last cycle and thus not considered for further characterization. That the overwhelming number of species in the network are unexplored is an illustration of the work being performed by YAKS in down-selecting important reactants for reaction characterization. Within 9 exploration steps (the same number of steps explored in the earlier MDA study), YAKS identified pathways to 5/6 major experimental products, HMF, FF, HAA, DHA, and HBL. Backward reaction searches—manually conducted explorations from the experimental products back to the explored CRN—starting from FF, 3FO, and HMF were able to connect along the low barrier pathway to the forward-explored network within one, two, and two reaction steps, respectively. The full CRN, composed of all forward and backward searches, recovers the low barrier pathways and multiple routes to all six of the major experimental products. The full CRN comprises 4733 species with 5395 unique reactions under 65 kcal/mol and is the first unified reaction network connecting all major experimental products.

The YAKS exploration is more efficient and accurate than the MDA exploration. The simpler MDA exploration was limited to unimolecular chemistry and was only able to discover pathways to 2/6 of the major experimental products as part of the forward search.[26] The MDA network was sufficiently broad that backwards searches were able to connect an additional three products to the network, with equal or lower barriers as discovered in the SSW study.[8] In contrast, YAKS rediscovered the low barrier pathways to DHA and HAA one and four steps earlier than the MDA exploration, respectively. YAKS also found new pathways to levoglucosan, 1-Hydroxy-2-propanone (HA), and HBL that were missed by the simpler forward MDA approach, owing to its reliance on simple ERS.

YAKS also overcomes the choice paralysis that faces later stage MDA explorations. During D-Glucose pyrolysis, all significant reaction channels diverged from a single rate-limiting reaction step, making downstream intermediates equally desirable and severely hindering the ability to select species to explore.[26] The 8th MDA step suggested 33 different intermediates to characterize, more than all species explored up to that point, effectively halting exploration. In contrast, YAKS can distinguish between species that share a common rate-limiting step by accounting for secondary bottlenecks through microkinetic modeling. The YAKS exploration performed here ran nearly 3x deeper than the MDA exploration without any selection issues.

The microkinetic modeling used by YAKS constitutes a negligible computational cost relative to the reactivity characterization. For example, YAKS Stage 3 activities for the exploration associated with Figure 4 exceeded 1,000 node-hours on our local cluster, while the microkinetic modeling involved minutes ($> 0.3\%$ of the total exploration time). By construction, this YAKS workload distribution generalizes to other systems. The automation associated with YAKS also saves untold hours of human toil associated with manual job initiation that are difficult to quantify.

Parallel noiseless CRN explorations with different ERS types and with or without seeded bimolecular reactions were performed to investigate how sensitive the CRN discovered in 4 was to the underlying YAKS settings. Specific differences between these CRNs and the full CRN are discussed in the following sections, but overall characteristics are as follows. The unimolecular Cb3f3 network, shown in SI Fig. S1, ran for 17 exploration steps, stopping after the top-5 highest concentration species had all previously been explored, and comprises 1145 unique reactions under 65 kcal/mol and 983 species. The bimolecular Cb3f3 CRN fell below the minimum confidence threshold in 14 YAKS cycles and comprised 797 species with 947 unique reactions. The unique bimolecular portion of the network is shown in Fig. 5 and a larger subnetwork is shown in SI Fig. S2. The unimolecular b3f3 CRN fell below the minimum confidence threshold after 17 cycles, comprised 1362 species and 1384 reactions and is shown in the SI Fig. S3. Lastly, the b3f3 bimolecular

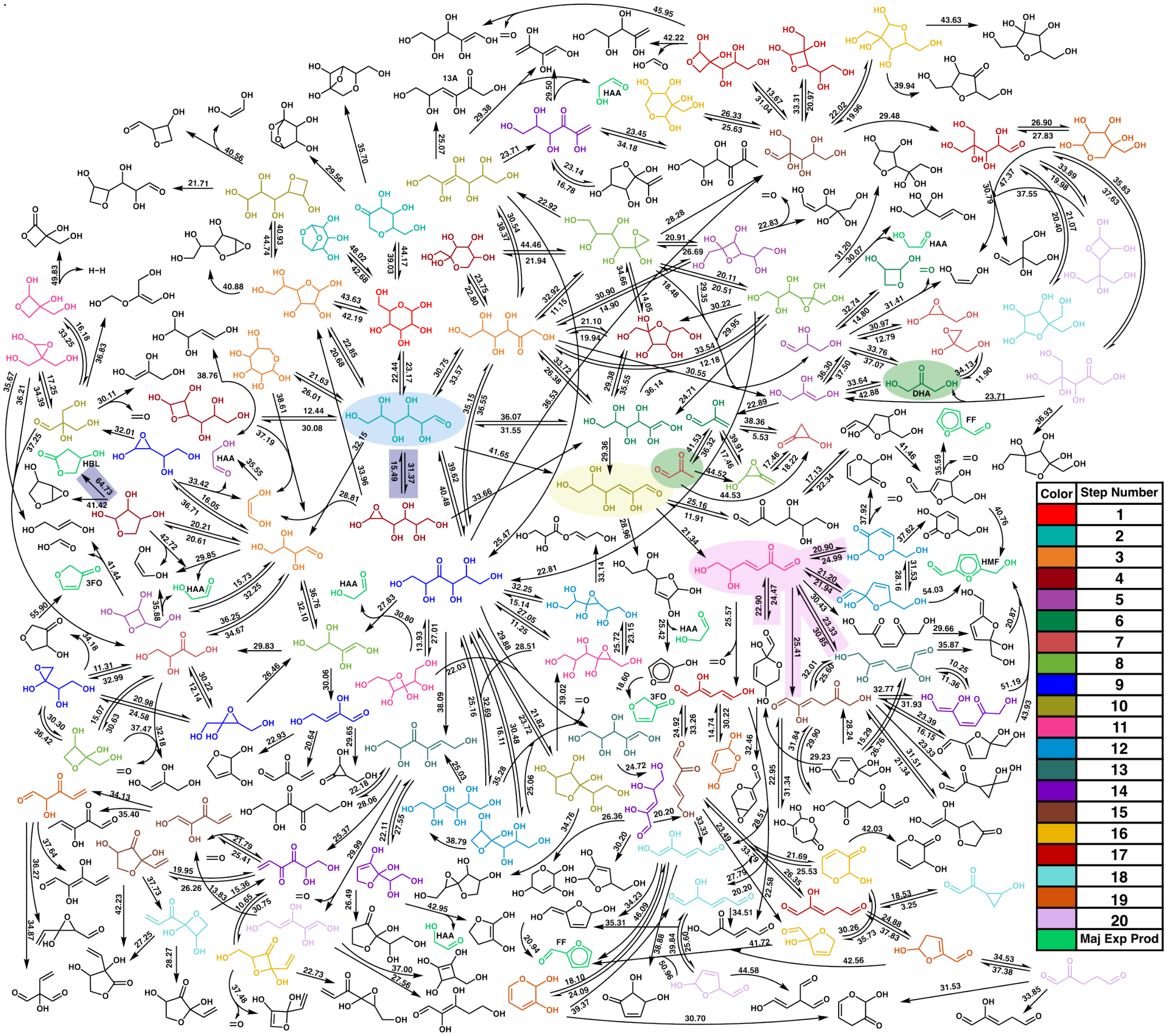

Figure 4: The D-glucose pyrolysis network explored by YAKS. Only unimolecular reactions are shown for clarity; bimolecular reactions are discussed in Section 3.2. The exploration was performed with rate uncertainty estimation, bimolecular reactions, and up to b3f3 ERS enabled. Starting from D-glucose (shown in red), molecules are colored according to their step number, but only experimental products are labeled. The number adjacent to each arrow refers to the free energy of activation ($\Delta G^{\dagger}$) in kcal/mol. The arrows follow the direction of the network exploration, but many reverse reactions $\Delta G^{\dagger}$ are shown above and below double arrows. From each explored intermediate, the graphic highlights the three reactions with lowest activation barriers under 45 kcal/mol and select pathways to experimental products. Species shown in black are unexplored intermediates. Species in shamrock green are experimental products (except the initial exploration of a major product which retains its original color). For graphic clarity, water is not shown during dehydration reactions. Highlighted sections of the network are discussed in the text.

network dropped beneath the confidence threshold after 15 cycles and comprised over 1521 species and 1424, a subnetwork is shown in SI Fig. S4. All configurations explored similarly in the shallow network, but the direction generally diverged and resulted in crucial pathway omissions as the CRN became deeper.

3.2 Consequences of Expanded Reaction Rules

The reaction rules that govern CRN exploration pose a compromise between breadth and computational cost. For D-Glucose, with 24 total atoms and 24 bonds, a b2f2 reaction enumeration has $\binom{24}{2} = 276$ possible rearrangements (without discounting symmetrically equivalent reactions) whereas b3f3 enumeration has $\binom{24}{3} = 2024$ possible rearrangements. All b3f3 reactions can be decomposed into sequential b2f2 reactions, and previous testing has confirmed this is usually kinetically preferable unless the reaction involves one or more π-bond rearrangements. For this reason, both the earlier MDA exploration and the uncertainty-guided exploration only explored Cb3f3 reactions involving one or more π-bond rearrangements.

B3f3 ERS. To investigate whether the Cb3f3 rule led to the exclusion of any important b3f3 (all σ-bond) reactions, the YAKS glucose exploration was reperformed with water-catalyzed reaction rules involving all b3f3 (Figs. S3-S4). The inclusion of the b3f3 reactions involving only σ-bonds reduced several barriers, introduced new intermediates inaccessible by Cb3f3, and introduced the leftmost blue highlighted reaction in Fig. 4, which represents a new pathway to HBL. However, virtually all relevant reactions discovered by including the unconditional b3f3 explorations are actually b2f2 reactions that were discovered as unintended reactions (i.e., the transition state connects a different reactant and product than the one that was used to initiate the search). The inclusion of unconditional b3f3 reactions thus mainly provided a form of conformational sampling for b2f2 reactions. Only two true b3f3 reactions were discovered that altered YAKS explorations (both highlighted in dark blue in Fig. 4), and both of these proved unproductive after further exploration. Unconditional b3f3 reactions thus contributed minimally to CRN knowledge and at considerable expense.

Bimolecular Reactions. Analysis of the bimolecular reaction pathways revealed by YAKS suggests that these play a minimal role in D-Glucose pyrolysis at the simulated conditions (Fig. 5). Across the D-glucose case studies, over 30 bimolecular reactivity calculations were performed. None of these contributed pathways to high-yield experimental products. For example, the same three bimolecular reactions between the combinations of DHA, methylglyoxal [SMILES: O=CC(=O)C], and water (highlighted in green in Fig. 4 and Fig. 5) were seeded during the eighth exploration step of both ERS case-studies. This new reaction channel occupied numerous costly YARP calculations and resulted in newly formed endergonic species that largely decomposed back to their original reactants or similar small stable compounds upon further exploration and microkinetic modeling. Bimolecular reactions did identify a pathway to form HA, a minor product of high temperature pyrolysis, highlighted in yellow in Fig. 5. Although HA is one of the few thermodynamically stable products of the bimolecular network, it does not experimentally form at the lower temperature at which this study was conducted and so even this does not constitute a clear accomplishment.[60] The relatively small number of bimolecular reactions seeded in the exploration ultimately reflects the tendency of the species in this CRN to unimolecularly react under pyrolysis conditions before accumulating sufficient concentration to bimolecularly react.

Notably, unimolecular exploration with YAKS still discovers many bimolecular reaction channels through the reverse reactions of unimolecular fragmentations and unintended reaction channels. Examples of the latter include hydration reactions that are discovered while attempting water-catalyzed reactions. All CRNs, regardless of unimolecular or bimolecular formulation include over 25 bimolecular reactions. The backward search did identify two routes to form HBL, both using hydration reactions, but both are kinetically and thermodynamically unfavorable.

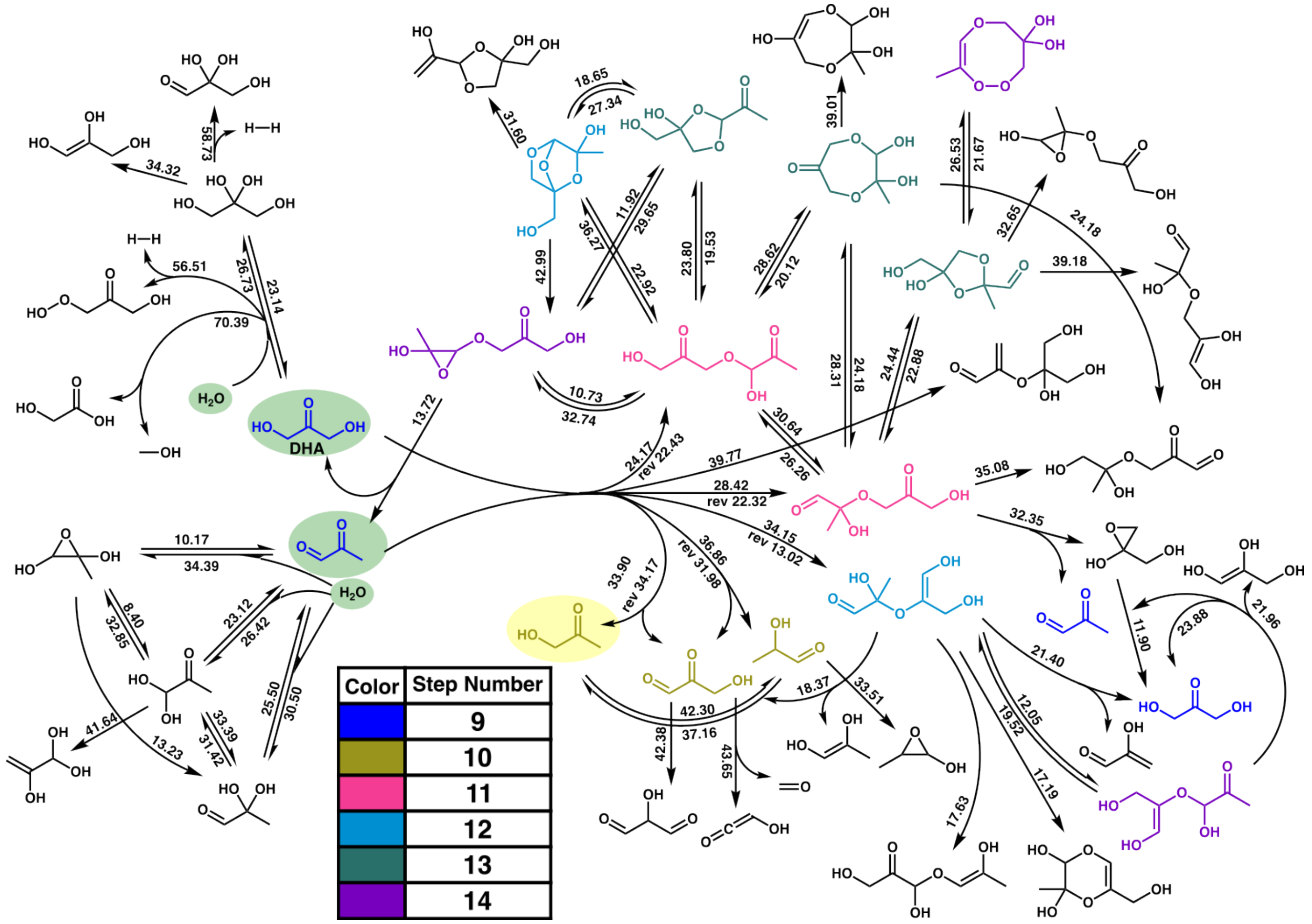

Figure 5: Bimolecular subnetworks resulting from bimolecular reactions between combinations of DHA, water, and methylglyoxal. Species are colored according to their bimolecular exploration step. Most bimolecular reactions available to the noiseless D-Glucose explorations decompose back toward original reactants or similar small thermodynamically stable compounds.

3.3 Uncertainty in Deep Reaction Networks

To quantify the impact of reaction rate uncertainty in the resulting CRN, we resimulated the kinetics of the four noiseless explorations Cb3f3 and b3f3 networks (SI Figs. S1-S4) with resampled reaction rates at each stage of the YAKS exploration (see methods). The number of unique primary terminal species (UPTS) (left-axis in Fig. 6) and the mean cumulative mass percent (right-axis) of the top-5 species was tracked at each exploration stage. A primary terminal species is any species with a plurality of the steady-state concentration. As exploration deepens, there is a dramatic increase in the number of top-1 species for each exploration step. The b3f3 bimolecular CRN has over 70 species that were the top-1 during any of the 1,000 simulations. Expanding the lefthand axis to include any species that appeared within the top-5 during any exploration step, the number of species grows 3x-11x depending on the step. Without accounting for such uncertainty, a naïve noiseless exploration is more likely to fall into shallow local minima and explore more deeply in kinetically misguided directions.

The observation that small deviations in activation barriers can lead to dramatically different CRN explorations motivated the use of uncertainty-guided exploration in YAKS. A particularly diabolical example from the noiseless D-glucose network involves the species highlighted in pink, shown on the bottom right of Fig. 4. The five highlighted reactions were explored during the noiseless unimolecular exploration, but the two unhighlighted reactions follow the low barrier pathways to form major experimental products, HMF and FF. The difference between the highest barriers of the noiseless explored species and the lowest barrier of the unexplored species is 0.16 kcal/mol, well within DFT and conformational sampling errors. By averaging the results of many noisy simulations, the CRN converges toward a more convincing solution.

The use of a beam-search is the second YAKS feature that is implemented to mitigate rate uncertainty. A wider beam search makes shallow YAKS explorations more robust by exploring all possible kinetically relevant branches. For example, YAKS explored the reactivity of every intermediate connected to the blue highlighted species in Fig. 4 with a barrier less than ~40 kcal/mol. Nevertheless, this benefit is diluted at later stages. For example, the fifth Cb3f3 exploration step characterized the reactivity of $\frac{20}{184}$ species in the CRN whereas the 17th step explored $\frac{80}{983}$ of the CRN. As the ERS expands, the problem magnifies. The 16th b3f3 exploration step contained 1362 distinct species, without an increase in the number of nodes explored per step, which causes each exploration beam to become increasingly isolated and greedy.

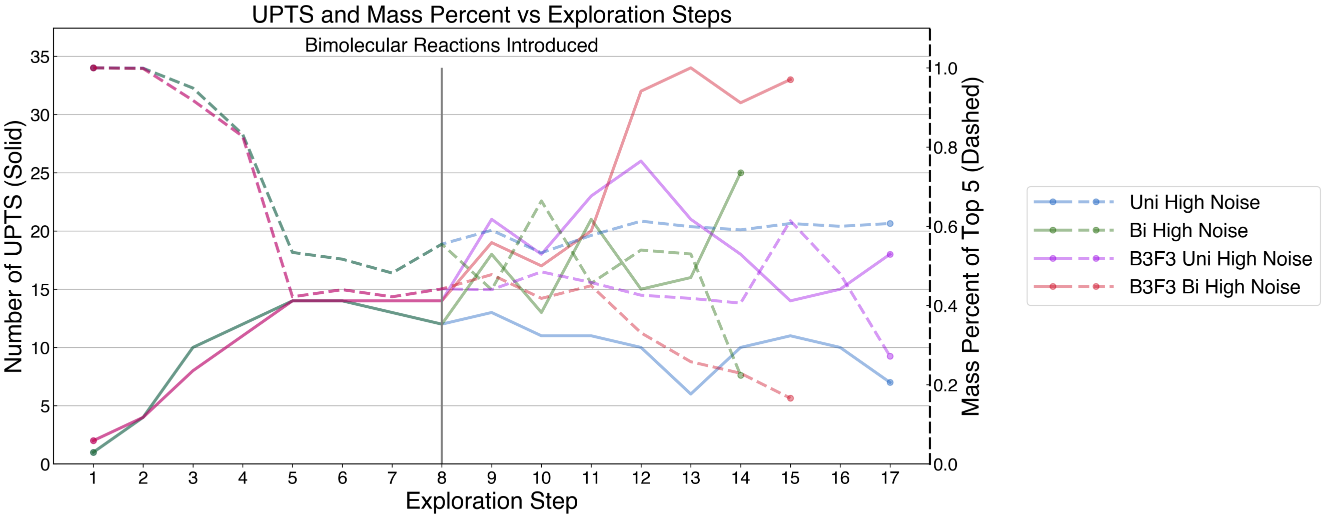

Figure 6: Unique primary terminal species (UPTS) and cumulative mass percent (CMP) of top five species across all exploration steps. UPTS hold a concentration pluarality at the conclusion of any of the 1,000 microkinetic simulations. Dashed lines correspond to the CMP of the top 5 species at each exploration step. The increase in concentration uncertainty with respect to network depth is reflected by the coincident increases in UPTS and decreases in CMP. Bimolecular and B3F3 reaction rules expand chemical space, aggravating the prioritization of the most pertinent intermediates.

The general effect of including uncertainty-guided exploration is to broaden the exploration of the network at the expense of depth. All else being equal, the noiseless exploration will prioritize the lowest barrier reaction sequences regardless of their number, whereas long sequences with marginally lower barriers are disfavored by error propagation. For example, the lowest barrier pathways to form HMF and FF are 11 and 12 reactions long, allowing a dozen opportunities for uncertain barriers to divert flux away. Although we have explored physically motivated ranges of activation energy uncertainties, the user could also use this phenomenology to tune the exploration.

With the benefit of the global network view afforded by Fig. 4 we also highlight two other strategies that could be used in conjunction with YAKS to assist exploration despite uncertainty. The first method involves starting YAKS anew from an important intermediate partway into the network, such as from the yellow highlighted species in Fig. 4 that is seven reactions deep along the lowest barrier pathway. YAKS explored only the lowest barrier reaction from the yellow species, but a wide beam search from this juncture would certainly identify additional pathways to terminal products, especially the low barrier path to 3FO. The second method involves performing a sequence of reactivity explorations between microkinetic simulations (i.e., running **Stage 3** multiple times in each cycle). This strategy would allow YAKS to explore sequential endergonic reactions in succession as an alternative to using a fixed $n_\mathrm{d} > 0$. Although deep exploration is inherently uncertain, these strategies can help in practical explorations.

3.4 The Critical Pathways: Low Barrier or High Flux?

Do the low-barrier reaction sequences always dominate the flux in large CRNs? To investigate this, the reaction fluxes from 1,000 microkinetic simulations with uncertainty sampling were compared between the lowest barrier pathways, the highest flux pathways of the CRN from the forward exploration, and the the highest flux pathways of the complete CRN with backward searches that connected to experimental products (Fig. 7). Blue pathways show the highest flux routes to form 5/6 experimental products (no pathway to 3FO was found during the forward search). Yellow shows the high flux pathways in the complete CRN. Red pathways are the low barrier pathways. When a species/reaction is both high flux during the forward and complete CRN, it is green. If in the full network, it is high flux and low barrier, it is shown as orange. Lastly, those reaction that are high flux in the forward CRN, full CRN, and low barrier paths are shown in purple.

Lower barrier reaction pathways often receive the most flux through a network. For example, the pathway that forms DHA involves a reaction sequence that exhibits the lowest overall barrier (LOB) to DHA and is also the highest flux. Longer discovered pathways to form DHA exhibit negligible flux and are functionally irrelevant. The majority of flux through Fig. 7 flows through the same low barrier reactions, but there are notable exceptions.

Shorter reaction pathways with higher overall barriers are often more kinetically relevant than longer reaction sequences with lower rate-limiting steps. The shortest route to form HAA, despite being nearly 1.5 kcal/mol larger in overall barrier, is 4x more favored over the lowest barrier route, which involves 2x more reactions. Similarly, the high-flux pathways to form HMF, FF, 3FO, and HBL all traverse a 33.72 kcal/mol reaction (OCC(C(C(C(=O)CO)O)O)O => OCC(C(C(C(=CO)O)O)O)O, also shown in green between two purple species), whereas the lowest barrier pathway has a nearly 3 kcal/mol lower rate-limiting step but involves four reactions.

One reason for this behavior is that longer reaction sequences siphon flux to more off-target channels, even if the overall barrier to a particular species is lower. A second reason is that rate uncertainty propagates with respect to sequence length. With randomly injected noise, each reaction has an opportunity to become unfavorable, but a single reaction is more likely to remain favorable. Thus, kinetically simulated terminal products are more likely to form if the pathways to their formation is shorter. The trend is reinforced by HMF and FF, whose LOB pathways are 11 and 12 reactions long, but highest flux pathways are only 8 and 9 reactions. Uncertain kinetics prefer direct reaction pathways.

Similarly, even with an equal number of reactions, the higher overall barrier pathway doesn't imply lower flux. In the bottom left of Fig. 7, the high flux forward CRN pathway favors a seemingly unfavorable reaction (yellow) at 35.87 kcal/mol where the parallel (red) pathway just above has an overall barrier 5 kcal/mol lower. Both pathways involve only 2 reactions, with the

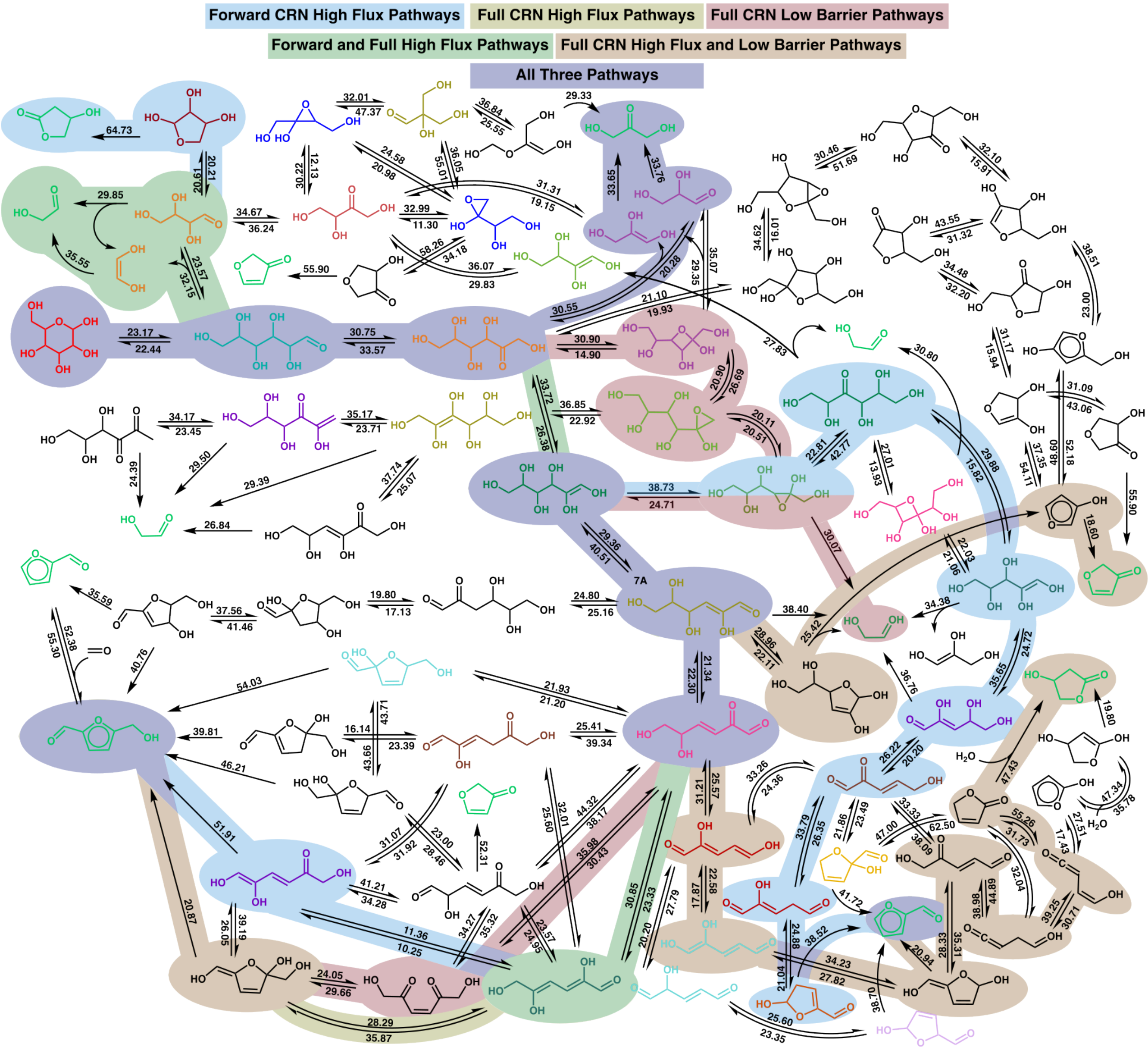
Forward CRN High Flux Pathways
Full CRN High Flux Pathways
Full CRN Low Barrier Pathways
Forward and Full High Flux Pathways
Full CRN High Flux and Low Barrier Pathways
All Three Pathways
7A

Figure 7: Characterization of reaction pathways to experimental products. The high flux pathways of the forward CRN (no backward searches) are shown in blue, the full CRN (with backward searches) are shown in yellow, and the low barrier pathways of the full CRN are shown in red. Coincident high flux pathways in the forward and full network are shown in green (yellow + blue = green). Coincident high flux and low barrier pathways in the full CRN are shown in orange (yellow + red = orange). If all three primary colors overlap, the pathway is purple. YAKS spontaneously identified pathways to 5 of 6 experimental products during the forward search, and low barrier pathways to all 6 during the backward search. Species are colored according to their exploration step in Figure 4.

same purple-shaded reactant and orange-shaded product. The major distinction is that the high flux pathway is initially 7 kcal/mol less than the low overall barrier (red) pathway which siphons flux away from the purple-shaded reactant instead of the seven other more energetically favorable reactions that would compete with the red pathway (reactions highlighted in pink in Fig. 4). The low barrier pathway is not as important as the topography of the network when determining the most kinetically relevant reactions.

3.5 Experimental Accuracy

To compare the calculated CRN with experimental results, it was selectively refined by retaining the three LOB reaction pathways that terminated in at least one experimental product and high flux pathways from the uncertainty-guided CRN (Fig. 7). In addition to the three lowest pathways to HAA, we also kept reactions to HAA from species already included in the CRN. Just as in a YAKS exploration, all internal reactions were considered reversible, but reverse reactions were not included for terminal edges.

Table 1: Experimental peak area %[60] vs. average peak concentration % results from 1,000 microkinetic simulations of the refined CRN.

Products	Fang et al Results	Critical Paths (CP)	CP Low Uncertainty	CP High Uncertainty
HMF	20%	24.5%	15.3%	9.7%
FF	15%	19.5%	8.2%	4.6%
HAA	13.5%	26.5%	40.4%	47.8%
DHA	3.9%	25.1%	29.9%	30.8%
3FO	3.5%	2.4%	3.0%	2.6%
HBL	3.4%	0%	0%	0%

A comparison of experimental product yields with the simulations of the pruned CRNs reveals some qualitative successes but also the limitations of current methods (Table 1).[60] Simulated yields are normalized to exclude the concentration percent of water so that they can be compared with the experimental values. In the noiseless CRN, HMF, FF, and 3FO are reasonably represented while

DHA and HAA are severely over represented and HBL is entirely absent. The major experimental products, except for HBL, are virtually omnipresent amongst the simulated major products even when considering rate uncertainty (Fig. S6).

Experimental products with longer reaction pathways and higher overall barriers tend to diminish rapidly under greater uncertainty (Table 2). In order, HBL, FF, and HMF lose the largest proportion of their noiseless population. Longer reaction pathways encounter more opportunities for flux to divert towards other products. As a useful comparison, HMF and 3FO have the same overall barriers for their LOB and shortest formation (SF) pathways, but both LOB and SF 3FO pathways are one reaction shorter than the HMF pathways. As a result, the 3FO population remains stable while HMF depletes by 60% in the high noise simulation scenario. DHA population grows moderately under uncertainty, because the DHA SF and LOB pathways are only slightly longer and higher barrier than the HAA SF.

Table 2: Lowest overall barrier (LOB) and shortest formation (SF) pathways for reaction sequences that produce major experimental products of D-glucose pyrolysis.

Products	LOB Reactions	LOB Barrier (kcal/mol)	SF Reactions	SF Barrier (kcal/mol)
HMF	11	30.90	8	33.72
FF	12	34.51	9	34.51
HAA	6	30.90	3	32.15
DHA	4	33.65	4	33.65
3FO	10	30.90	7	33.72
HBL	14	47.34	10	47.43

HAA concentration nearly doubles when simulated with rate uncertainty, because the highest yield pathway to form HAA is only three reactions long and the shortest of all critical pathways. A downward adjustment in either of the last two reactions will likely increase HAA yield, whereas a downward adjustment in both reaction barriers (25% chance) will dramatically increase its final concentration. If we were to prune the high flux 32.15 kcal/mol reaction shown in the upper left of Fig. 7, the concentration share of HAA would plummet to 13%. Without that pathway, the new

HAA SF pathway grows to five reactions, longer than the DHA SF. Considering rate uncertainty and additional off-target channels, the combination of LOB and SF is required to rationalize flux.

4 Conclusion

The improvements in cost, accuracy, chemical range, and throughput of automated reaction prediction methods create opportunities to elucidate comprehensive deep reaction networks. Despite these advances, work is still required to couple these methods with network exploration algorithms that prioritize physically relevant CRN explorations. This work elaborated the YAKS network exploration algorithm that uses microkinetic modeling on sequential subnetworks to prioritize relevant intermediates for further investigation. Salient features of YAKS are the use of rate uncertainty estimation, the manipulation of the network topology to prioritize kinetically accessible intermediates, the use of a parallel branch exploration, and the automatic treatment of bimolecular reactions involving intermediates. Application of YAKS to the problem of glucose pyrolysis yielded the first global reaction network that connects all major experimental products and glucose. This network supercedes the prior network generated using the simpler MDA exploration policy with the YARP reaction prediction engine. This new network is not substantially larger in terms of number of reactions and intermediates than the preceding MDA network; rather, it mainly reflects the alternative explorations selected by YAKS compared with the simpler algorithm.

Although large reaction networks are impressive, exploration efficiency is more important than the sheer number of reactions and intermediates that were characterized. Indeed, characterizing large numbers of reactions only to discover a few short reaction sequences should be regarded as a failure, and the field needs to standardize better metrics of exploration efficiency. Here, several case studies were performed where different aspects of the YAKS algorithm were removed. None of these changes affected the number of reactions and intermediates that could be characterized, but the omissions did lead to less physically relevant reactions being explored and some being missed entirely.

There are several avenues to further improve YAKS. Exploration algorithms need to address the potential for catalytically active intermediates. For example, water was utilized as a catalyst for proton transfers here as a hard-coded option, not because YAKS recognized the potential of liberated water to act as a catalyst. This differs from exploring bimolecular reactivity, but a similar framework could be applied to the two problems. Additionally, non-physical reactions returned by the reaction prediction engine can have large effects on the microkinetic simulations and ultimately mislead the exploration. Exploration algorithms like YAKS could more generally build in physical priors for certain reaction classes in order to make them more robust to artifacts from purely computational reaction prediction. These and other ongoing improvements will be necessary to expand the classes of CRNs that can be effectively explored from scratch.

5 Data Availability and Code Availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files.

Further raw data sources generated by this work are available at (XXX, figshare link will be populated upon publication XXX), including raw output files and molecular geometries. The YAKS software package can be accessed on GitHub (https://github.com/Savoie-Research-Group/yaks).

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

The work was made possible by the Office of Naval Research (ONR) through support provided by the Energetic Materials Program (MURI grant number:N00014-21-1-2476, Program Manager: Dr. Chad Stoltz). B.M.S also acknowledges partial support for this work from the Purdue Process Safety & Assurance Center (P2SAC).

References

(1) Suleimanov, Y. V.; Green, W. H. Automated discovery of elementary chemical reaction steps using freezing string and Berny optimization methods. *J. Chem. Theory Comput.* **2015**, *11*, 4248–4259, Publisher: ACS Publications.

(2) Habershon, S. Sampling reactive pathways with random walks in chemical space: Applications to molecular dissociation and catalysis. *J. Chem. Phys.* **2015**, *143*, 094106, Publisher: AIP Publishing LLC.

(3) Ismail, I.; Robertson, C.; Habershon, S. Successes and challenges in using machine-learned activation energies in kinetic simulations. *The Journal of Chemical Physics* **2022**, *157*, 014109.

(4) Habershon, S. Automated prediction of catalytic mechanism and rate law using graph-based reaction path sampling. *J. Chem. Theory Comput.* **2016**, *12*, 1786–1798, Publisher: ACS Publications.

(5) Grambow, C. A.; Jamal, A.; Li, Y.-P.; Green, W. H.; Zádor, J.; Suleimanov, Y. V. Unimolecular Reaction Pathways of a -Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *Journal of the American Chemical Society* **2018**, *140*, 1035–1048.

(6) Lee, C. W.; Taylor, B. L. H.; Petrova, G. P.; Patel, A.; Morokuma, K.; Houk, K. N.; Stoltz, B. M. An Unexpected Ireland–Claisen Rearrangement Cascade During the Synthesis of the Tricyclic Core of Curcusone C: Mechanistic Elucidation by Trial-and-Error and Automatic Artificial Force-Induced Reaction (AFIR) Computations. *Journal of the American Chemical Society* **2019**, *141*, 6995–7004.

(7) Ismail, I.; Stuttaford-Fowler, H. B.; Ochan Ashok, C.; Robertson, C.; Habershon, S. Auto-

matic proposal of multistep reaction mechanisms using a graph-driven search. *J. Phys. Chem. A* **2019**, *123*, 3407–3417, Publisher: ACS Publications.

(8) Kang, P.-L.; Shang, C.; Liu, Z.-P. Glucose to 5-Hydroxymethylfurfural: Origin of Site-Selectivity Resolved by Machine Learning Based Reaction Sampling. *Journal of the American Chemical Society* **2019**, *141*, 20525–20536.

(9) Naz, E. G.; Paranjothy, M. Unimolecular Dissociation of -Ketohydroperoxide via Direct Chemical Dynamics Simulations. *J. Phys. Chem. A* **2020**, *124*, 8120–8127, Publisher: ACS Publications.

(10) Ramasesha, K.; Savee, J. D.; Zádor, J.; Osborn, D. L. A New Pathway for Intersystem Crossing: Unexpected Products in the O (3P)+ Cyclopentene Reaction. *J. Phys. Chem. A* **2021**, *125*, 9785–9801, Publisher: ACS Publications.

(11) Zhao, Q.; Savoie, B. M. Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks. *Nat. Comput. Sci.* **2021**, *1*, 479–490.

(12) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671.

(13) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications* **2019**, *10*, 2903.

(14) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97-3c: A revised low-cost

variant of the B97-D density functional method. *The Journal of Chemical Physics* **2018**, *148*, 064104.

(15) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics* **2000**, *113*, 9901–9904.

(16) Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. Efficient exploration of reaction paths via a freezing string method. *J. Chem. Phys.* **2011**, *135*, 224108, Publisher: American Institute of Physics.

(17) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **2004**, *120*, 7877–7886, Publisher: American Institute of Physics.

(18) Zimmerman, P. M. Growing string method with interpolation and optimization in internal coordinates: Method and examples. *J. Chem. Phys.* **2013**, *138*, 184102, Publisher: American Institute of Physics.

(19) Kim, S.; Woo, J.; Kim, W. Y. Diffusion-based Generative AI for Exploring Transition States from 2D Molecular Graphs. 2023; `http://arxiv.org/abs/2304.12233`, arXiv:2304.12233 [physics].

(20) Pattanaik, L.; Ingraham, J. B.; Grambow, C. A.; Green, W. H. Generating transition states of isomerization reactions with deep learning. *Physical Chemistry Chemical Physics* **2020**, *22*, 23618–23626.

(21) Makoś, M. Z.; Verma, N.; Larson, E. C.; Freindorf, M.; Kraka, E. Generative adversarial

networks for transition state geometry prediction. *The Journal of Chemical Physics* **2021**, *155*, 024116.

(22) Jackson, R.; Zhang, W.; Pearson, J. TSNet: predicting transition state structures with tensor field networks and transfer learning. *Chemical Science* **2021**, *12*, 10022–10040.

(23) Choi, S. Prediction of transition state structures of gas-phase chemical reactions via machine learning. *Nature Communications* **2023**, *14*, 1168, Number: 1 Publisher: Nature Publishing Group.

(24) Zhao, Q.; Anstine, D. M.; Isayev, O.; Savoie, B. M. 2 machine learning for reaction property prediction. *Chemical Science* **2023**, *14*, 13392–13401, Publisher: The Royal Society of Chemistry.

(25) Wen, M.; Spotte-Smith, E. W. C.; Blau, S. M.; McDermott, M. J.; Krishnapriyan, A. S.; Persson, K. A. Chemical reaction networks and opportunities for machine learning. *Nature Computational Science* **2023**, *3*, 12–24.

(26) Zhao, Q.; Savoie, B. Deep Reaction Network Exploration of Glucose Pyrolysis. 2023; `https://chemrxiv.org/engage/chemrxiv/article-details/64073643cc600523a3cd4782`.

(27) Unsleber, J. P.; Liu, H.; Talirz, L.; Weymuth, T.; Mörchen, M.; Grofe, A.; Wecker, D.; Stein, C. J.; Panyala, A.; Peng, B.; Kowalski, K.; Troyer, M.; Reiher, M. High-throughput ab initio reaction mechanism exploration in the cloud with automated multi-reference validation. *The Journal of Chemical Physics* **2023**, *158*, 084803.

(28) Bensberg, M.; Reiher, M. Concentration-Flux-Steered Mechanism Exploration with an Organocatalysis Application. *Israel Journal of Chemistry* **2023**, *63*, e202200123, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijch.202200123.

(29) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering chemistry with an ab initio nanoreactor. *Nat. chem.* **2014**, *6*, 1044–1048, Publisher: Nature Publishing Group.

(30) Huang, S.-D.; Shang, C.; Zhang, X.-J.; Liu, Z.-P. Material discovery by combining stochastic surface walking global optimization with a neural network. *Chem. Sci.* **2017**, *8*, 6327–6337, Publisher: Royal Society of Chemistry.

(31) Nakao, A.; Harabuchi, Y.; Maeda, S.; Tsuda, K. Leveraging algorithmic search in quantum chemical reaction path finding. *Phys. Chem. Chem. Phys.* **2022**, *24*, 10305–10310, Publisher: Royal Society of Chemistry.

(32) Kang, P.-L.; Shi, Y.-F.; Shang, C.; Liu, Z.-P. Artificial intelligence pathway search to resolve catalytic glycerol hydrogenolysis selectivity. *Chemical Science* **2022**, *13*, 8148–8160.

(33) Bensberg, M.; Reiher, M. Uncertainty-aware First-principles Exploration of Chemical Reaction Networks. 2023; `http://arxiv.org/abs/2312.15477`, arXiv:2312.15477 [physics].

(34) Zhang, S.; Makoś, M. Z.; Jadrich, R. B.; Kraka, E.; Barros, K.; Nebgen, B. T.; Tretiak, S.; Isayev, O.; Lubbers, N.; Messerly, R. A.; Smith, J. S. Exploring the frontiers of condensed-phase chemistry with a general reactive machine learning potential. *Nature Chemistry* **2024**, *16*, 727–734, Publisher: Nature Publishing Group.

(35) Chang, A. M.; Meisner, J.; Xu, R.; Martínez, T. J. Efficient Acceleration of Reaction Discovery in the Ab Initio Nanoreactor: Phenyl Radical Oxidation Chemistry. *The Journal of Physical Chemistry A* **2023**, *127*, 9580–9589, Publisher: American Chemical Society.

(36) Nishimura, Y.; Nakai, H. Species-selective nanoreactor molecular dynamics simulations based

on linear-scaling tight-binding quantum chemical calculations. *The Journal of Chemical Physics* **2023**, *158*, 054106.

(37) Stan-Bernhardt, A.; Glinkina, L.; Hulm, A.; Ochsenfeld, C. Exploring Chemical Space Using Ab Initio Hyperreactor Dynamics. *ACS Central Science* **2024**, *10*, 302–314.

(38) Wang, L.-P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J. Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *Journal of Chemical Theory and Computation* **2016**, *12*, 638–649, Publisher: American Chemical Society.

(39) Susnow, R. G.; Dean, A. M.; Green, W. H.; Peczak, P.; Broadbelt, L. J. Rate-Based Construction of Kinetic Models for Complex Systems. *The Journal of Physical Chemistry A* **1997**, *101*, 3731–3740, Publisher: American Chemical Society.

(40) Vinu, R.; Broadbelt, L. J. A mechanistic model of fast pyrolysis of glucose-based carbohydrates to predict bio-oil composition. *Energy & Environmental Science* **2012**, *5*, 9808–9826, Publisher: The Royal Society of Chemistry.

(41) Mayes, H. B.; Nolte, M. W.; Beckham, G. T.; Shanks, B. H.; Broadbelt, L. J. The alpha–bet(a) of glucose pyrolysis: computational and experimental investigations of 5-hydroxymethylfurfural and levoglucosan formation reveal implications for cellulose pyrolysis. *ACS Sustainable Chem. Eng.* **2014**, *2*, 1461–1473, Publisher: ACS Publications.

(42) Kostetskyy, P.; Coile, M. W.; Terrian, J. M.; Collins, J. W.; Martin, K. J.; Brazdil, J. F.; Broadbelt, L. J. Selective production of glycolaldehyde via hydrothermal pyrolysis of glucose: Experiments and microkinetic modeling. *Journal of Analytical and Applied Pyrolysis* **2020**, *149*, 104846.

(43) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Au-

tomatic construction of chemical kinetic mechanisms. *Computer Physics Communications* **2016**, *203*, 212–225.

(44) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J.; Blondal, K.; West, R. H.; Goldsmith, C. F.; Green, W. H. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *Journal of Chemical Information and Modeling* **2021**, *61*, 2686–2696, Publisher: American Chemical Society.

(45) Zhao, Q.; Garimella, S. S.; Savoie, B. M. Thermally Accessible Prebiotic Pathways for Forming Ribonucleic Acid and Protein Precursors from Aqueous Hydrogen Cyanide. *Journal of the American Chemical Society* **2023**, *145*, 6135–6143.

(46) Vadaddi, S. M.; Zhao, Q.; Savoie, B. M. Graph to Activation Energy Models Easily Reach Irreducible Errors but Show Limited Transferability. 2023; `https://chemrxiv.org/engage/chemrxiv/article-details/65410dc248dad23120c6e954`.

(47) Stulajter, M.; Rappoport, D. Reaction Networks Resemble Low-Dimensional Regular Lattices. 2024; `https://chemrxiv.org/engage/chemrxiv/article-details/6658fe89418a5379b0b45273`.

(48) Green, W. H. In *Computer Aided Chemical Engineering*; Faravelli, T., Manenti, F., Ranzi, E., Eds.; Mathematical Modelling of Gas-Phase Complex Reaction Systems: Pyrolysis and Combustion; Elsevier, 2019; Vol. 45; pp 259–294.

(49) Zhao, Q.; Savoie, B. More and Faster: Simultaneously Improving Reaction Coverage and Computational Cost in Automated Reaction Prediction Tasks. 2020; `https://chemrxiv.org/engage/chemrxiv/article-details/60c750b8567dfe44aeec58f9`.

(50) Zhao, Q.; Savoie, B. M. Algorithmic Explorations of Unimolecular and Bimolecular Reaction Spaces. *Angew. Chem., Int. Ed.* **2022**, *61*, e202210693.

(51) Zhao, Q.; Savoie, B. M. Self-Consistent Component Increment Theory for Predicting Enthalpy of Formation. *Journal of Chemical Information and Modeling* **2020**, *60*, 2199–2207.

(52) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **2020**, *22*, 7169–7192, Publisher: The Royal Society of Chemistry.

(53) Zhao, Q.; Hsu, H.-H.; Savoie, B. M. Conformational Sampling for Transition State Searches on a Computational Budget. *Journal of Chemical Theory and Computation* **2022**, *18*, 3006–3016.

(54) Zimmerman, P. M. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem.* **2013**, *34*, 1385–1392, Publisher: Wiley Online Library.

(55) Zimmerman, P. Reliable Transition State Searches Integrated with the Growing String Method. *Journal of Chemical Theory and Computation* **2013**, *9*, 3043–3050.

(56) Frisch, M. J. et al. Gaussian 16 Revision C.01. 2016.

(57) Zhao, Q.; Vaddadi, S. M.; Woulfe, M.; Ogunfowora, L. A.; Garimella, S. S.; Isayev, O.; Savoie, B. M. Comprehensive exploration of graphically defined reaction spaces. *Scientific Data* **2023**, *10*, 1–10, Number: 1 Publisher: Nature Publishing Group.

(58) David G. Goodwin,; Raymond L. Speth,; Harry K. Moffat,; Bryan W. Weber, ”Cantera: An Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics, and Transport Processes”. 2021; `https://www.cantera.org`.

(59) López, R.; Suárez, D. Pyrolytic Conversion of Glucose into Hydroxymethylfurfural and Furfural: A Survey of Mechanisms and Benchmark Quantum-Chemical Calculations. 2023; https://chemrxiv.org/engage/chemrxiv/article-details/654d5b9cdbd7c8b54bf9885e.

(60) Fang, Y.; Li, J.; Chen, Y.; Lu, Q.; Yang, H.; Wang, X.; Chen, H. Experiment and modeling study of glucose pyrolysis: Formation of 3-hydroxy--butyrolactone and 3-(2 H)-furanone. *Energy Fuels* **2018**, *32*, 9519–9529, Publisher: ACS Publications.