Deep Reaction Network Exploration of Glucose Pyrolysis

Qiyuan Zhao and Brett M. Savoie*

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

E-mail: bsavoie@purdue.edu

Abstract

Resolving the reaction networks associated with biomass pyrolysis is central to understanding product selectivity and aiding catalyst design to produce more valuable products. However, even the pyrolysis network of relatively simple β-D-Glucose remains unresolved due to its significant complexity in terms of the depth of the network and the number of major products. Here, a transition-state guided reaction exploration has been performed that provides complete pathways to all significant experimental pyrolysis products of β-D-Glucose. The resulting reaction network involves over 31,000 reactions and transition states computed at the semi-empirical quantum chemistry level and approximately 7,000 kinetically relevant reactions and transition states characterized at the DFT level, comprising the largest reaction network reported for biomass pyrolysis. The exploration was conducted using graph-based rules to explore the reactivities of intermediates and an adaption of Dijkstra algorithm to identify kinetically relevant intermediates. This simple exploration policy surprisingly (re)discovered pathways to all major experimental pyrolysis products, many intermediates proposed by previous computational studies, and also identified new low-barrier

reaction mechanisms that resolve outstanding discrepancies between reaction pathways and yield in isotope labeling experiments. This network also provides explanatory pathways for the high yield of hydroxymethylfurfural (HMF) and the reaction pathway that contributes most to the formation of hydroxyacetaldehyde (HAA) during glucose pyrolysis. Due to the limited domain knowledge required to generate this network, this approach should also be transferable to other complex reaction network prediction problems in biomass pyrolysis.

1 Introduction

Fast biomass pyrolysis has been heavily investigated as a potential source of inexpensive chemicals and sustainable energy.[1–9] D-Glucose is a biomass feedstock that has gained wide interest for decades because it is readily available from cellulose and exhibits high selectivity in producing furan products.[10–17] However, resolving the reaction pathways for glucose pyrolysis is challenging because of the multi-step nature of plausible mechanisms and the likelihood of competing reactions. Due to this complexity, even computational studies of glucose pyrolysis pathways have focused on a handful of curated reaction classes or hypothesis-driven mechanisms. For example, many have focused on identifying the dominant pathways for forming hydroxymethylfurfural (HMF) from glucose using a series of β-elimination and isomerization reactions.[18–21] However, the activation energies of the proposed pathways to HMF formation are still relatively high ($\sim$ 60 kcal/mol),[15,20–22] and do not provide convincing explanations for higher yields of HMF compared to other products, such as hydroxyacetaldehyde (HAA).

Automated reaction prediction methods have recently made great advances in terms of cost, accuracy, and throughput that enable them to often find low-barrier reaction mechanisms with minimal or no mechanistic guidance.[23–32] However, many of these algorithms scale extremely poorly with reactant size, limiting their application to very small systems. Even algorithms that can be

applied to reactants the size of glucose still need to be combined with a network exploration policy that constrains the number of potential intermediates as the network grows. For example, a naïve exploration, in which all products are retained as potential reactants, grows exponentially with respect to network depth, even without considering bimolecular reactions or catalyzed reactions. To circumvent this, deep reaction networks are typically generated using sampling heuristics meant to discover plausible reaction sequences. The paradigmatic example is the *ab initio* nanoreactor (AINR) algorithm developed by Wang et al. that uses low-level quantum chemistry with high pressures and temperatures to accelerate reactions on the molecular dynamics (MD) timescale.[33] To reduce the high computational costs associated with *ab initio* MD in the AINR, Liu et al. developed the stochastic surface walking with neural network (SSW-NN) algorithm to speed up the reaction space exploration.[34] The SSW-NN was recently applied to glucose pyrolysis by combining and analyzing multiple SSW trajectories (this study is later compared with this work).[35] In the past year, several additional algorithms have been reported with applications for individual multi-step reactions,[36] and catalysis at surfaces.[37]

Despite the large number of reaction steps that can be discovered by some of these algorithms, sampling biases and limited transferability are common problems. For example, the AINR requires reaction conditions that are likely to over-sample high barrier reactions and endothermic products compared with algorithms that use minimum energy pathway searches. Conversely, methods that rely on system-specific heuristics or ML approximations are not transferable to new systems and require extensive customization to apply. It seems accurate to summarize that there are currently no general purpose network exploration policies that have been demonstrated to be accurate when compared against experimental observations of deep reaction networks (e.g., capable of finding plausible low-barrier pathways to all major experimental products, or predicting the relative mass flux of various products based on pathways). This gap motivated the current effort to revisit the problem of glucose pyrolysis using a simple and generic exploration policy based on network theory.

Here, the dilemma between high computational costs and comprehensive reaction exploration was addressed by combining an ultra-low cost reaction prediction program, the Yet Another Reaction Program (YARP), with an efficient modified Dijkstra algorithm for exploring kinetically relevant reactions.[38] Dijkstra's algorithm is formally the most efficient single-source exploration method for finding minimum cost pathways on directed graphs. The algorithm is based on the simple rule of always exploring off of the lowest cost node that has been discovered up to that point. The algorithm can also be run in a parallel multi-source fashion from both the starting and ending nodes (if they are known) until finding a point at which the minimum cost pathways overlap. Considering the unimolecular glucose pyrolysis network as a directed graph, applying the single-source Dijkstra's algorithm is the equivalent of recursively exploring the reactivity of the products most likely form from glucose in the network at each stage of exploration. The activation energy of the rate-limiting step for forming each product was used as the "cost" function for the algorithm and YARP was used to characterize the reactions the lowest-cost products could undergo. The algorithm was run in single-source fashion until spontaneously finding major experimental products, then run in multi-source fashion to find pathways to other experimental products.

Despite the simplicity of this exploration policy, the resulting pyrolysis reaction network includes well-known reaction mechanisms and intermediates, and also new low activation energy pathways to the formation of several major products, including hydroxymethylfurfura (HMF), hydroxyacetaldehyd (HAA), furfural (FF), 3-(2H)-furanone (3FO), and dihydroxyacetone (DHA). These pathways exhibit diverse mechanisms and challenge the long-standing understanding of the rate-limiting steps of various pyrolysis products. The accuracy of the reaction network is also benchmarked against an experimental isotope labeling study,[39] for which it provides the first self-consistent explanation for the nine distinct pathways for forming five distinct HAA isotopomers and their relative abundance.

2 Results and Discussion

The reaction network exploration for β-D-Glucose presented here is the largest that has been published to date in terms of depth, number of reactions, and finding pathways with transition states for all major experimental pyrolysis products. To manage the complexity of this data, the discussion of results has been organized around several physically relevant questions. The first section (2.1) discusses the pyrolysis network in its entirety and compares it to prior work. The second section (2.2) discusses reoccurring low-barrier reaction mechanisms that lead to the formation of experimentally observed products. The third (2.3) and fourth (2.4) sections discuss the detailed pathways predictions for the formation of HMF and HAA, two major pyrolysis products by mass, respectively. A detailed description of all methods can be found in the Computational Methods section.

2.1 The Reaction Network.

The eight-layer forward reaction network generated by the comprehensive graph-based exploration is shown in Figure 1. Initialized with the β-D-Glucose (beta isoform of D-Glucose, Node 1-1), two kinetically accessible (reactions with free energy of activation [$\Delta G^{\dagger}$] lower than 45 kcal/mol) products were formed in the first step, namely D-glucose (node 2-1, $\Delta G^{\dagger}$=23.66 kcal/mol. Energy units are kcal/mol unless otherwise stated) and 1,5-Anhydro-D-Fructose (node 2-2, $\Delta G^{\dagger}$=44.17). 1,5-Anhydro-D-Fructose can only convert into a six-membered ring structure with $\Delta G^{\dagger}$ of 41.28 kcal/mol. On the contrary, D-glucose is more reactive and can transform and decompose through nine kinetically accessible reaction pathways. The single-source modified Dijkstra algorithm was used to control the search space, with up to five nodes being explored in parallel at each step. As a result, five out of ten third-layer products (denoted in gold in Fig. 1) were selected to seed the next step reaction exploration based on the overall barrier ($\Delta \mathrm{G}^{\dagger}_{\mathrm{max}}$, defined as $\Delta G^{\dagger}$ of the rate-

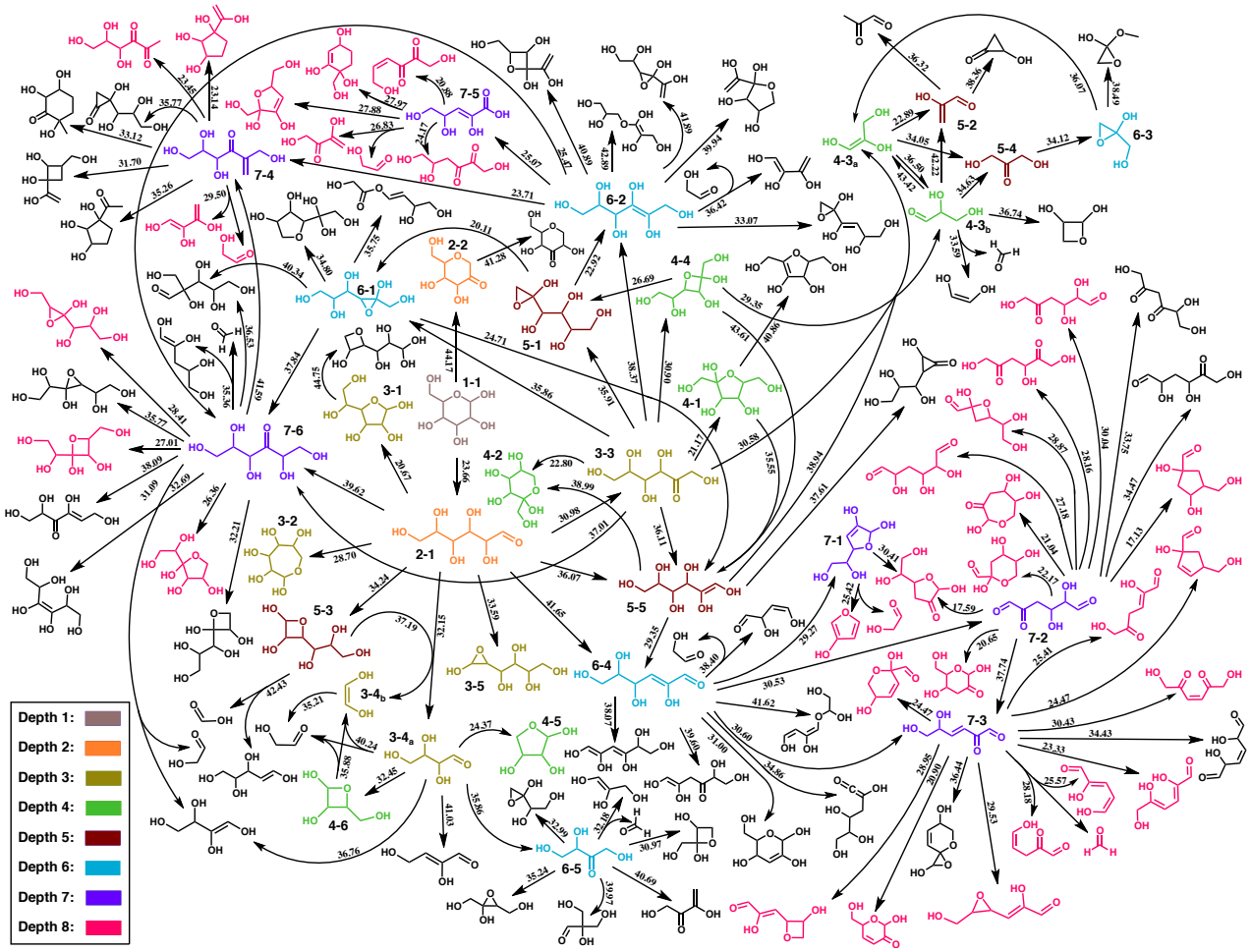

Figure 1: A subset of the reaction network of products and intermediates that can be formed after at most eight reactions starting from β-D-Glucose (denoted as grey, node 1-1). Reactions with activation energies less than 45 kcal/mol are shown in the network (all reactions and activation energies can be found in the supplementary information). The number above each arrow refers to the free energy of activation ($\Delta G^{\dagger}$) in kcal/mol. The arrows follow the direction of the network exploration. In some cases, $\Delta G^{\dagger}$ is lower for the reverse reaction to that shown. The intermediates highlighted by different colors served as reactants for exploration at the corresponding reaction depth. Species shown in black were considered terminal products by the exploration algorithm (i.e., no further exploration was performed using these species). If multiple arrows point at a species, this indicates that the exploration algorithm found multiple pathways to that species that satisfy the kinetic and thermodynamic thresholds used to make this figure.

limiting step). The reaction exploration continued for eight steps, resulting in the identification of 32 intermediates (denoted in pink) that have the same overall barrier. The common rate-limiting step of these intermediates occurs at the very beginning of glucose pyrolysis, which is the conversion of β-D-Glucose to D-Psicose with a $\Delta G^{\dagger}$ of 30.98 kcal/mol.

One important feature of this exploration policy is that the objective function (i.e., overall barrier) of each node are updated after each stage of exploration. This creates the possibility of backtracking to a node, (e.g., nodes previously considered as inaccessible are selected for exploration when low-barrier pathways connecting them are discovered or when downstream nodes prove to have higher overall barriers). A typical example is 3-hexulose (7-6 in Fig. 1), which was discovered during the second step of the exploration process, but was not considered a relevant intermediate for exploration until the seventh step of exploration. The overall barrier for 3-hexulose was updated from 39.62 (at depth 2) to 37.01 (at depth 3) and finally to 30.98 (at depth 6) due to the discovery of a low-barrier conversion from node 6-2, catalyzed by multiple hydroxyl groups (an important reaction mechanism discussed later in section 2.2). Terminated nodes are also common in the network. A terminated node is connected to the network by a low barrier reaction and thus is selected for further reaction exploration by the algorithm. However, no other kinetically accessible reactions are discovered on this node, which means most of the reaction flux arriving at terminated nodes eventually flows back upstream. Several intermediates produced through Korcek cyclization reactions, such as β-D-Allofuranose (node 3-1) and β-D-Galactoseptanose (node 3-2), are categorized as terminated nodes.

Two sub-networks related to the formation of two experimentally observed products are observed from the whole network. The first one is centered around D-Erythrose (node 3-4$_a$), which is produced by a retro-aldol reaction of D-glucose. After the elimination of 1,2-ethenediol (node 3-4$_b$), this sub-network mainly describes the conversions among $C_4H_8O_4$ isomers and the formation of hydroxyacetaldehyde (HAA), which is one of the major products of the glucose pyrolysis sys-

tem.[40] In addition to the water-catalyzed keto-enol tautomerization of 1,2-ethenediol ($\Delta G^{\dagger}$=35.21) that represents the shortest pathway of HAA formation, D-Erythrose can further decomposes into 1,2-ethenediol and HAA through another retro-aldol reaction. This series of reactions effects the conversion of glucose to three HAA molecules with an overall barrier of 40.24 kcal/mol. If considering an additional cyclization reaction from D-Erythrose to 4-(Hydroxymethyl)oxetane-2,3-diol (node 4-6), the overall barrier of the entire HAA conversion is reduced to 35.88 kcal/mol, which is the $\Delta G^{\dagger}$ of the retro-cycloaddition reaction of 4-(Hydroxymethyl)oxetane-2,3-diol. The other sub-network is centered on dehydroglycerol (node 4-3$_a$) and glyceraldehyde (node 4-3$_b$), which are two $C_3H_6O_3$ isomers formed by the retro-aldol reaction of D-Psicose (node 3-3). Both of these species can convert into a more stable isomer, dihydroxyacetone (DHA, node 5-4), through water-catalyzed keto-enol tautomerization reactions. Besides, glyceraldehyde can decompose into HAA through a retro-aldol reaction and a keto-enol tautomerization. HAA and DHA, being more stable compared with other isomers, act as two thermochemical sinks in this sub-network, providing an explanation for the experimental observation of these two products even apart from the detailed analysis considered in the subsequent sections.

2.2 Important Reaction Mechanisms and Pathways

A distinctive aspect of the current exploration is that generic graph-based elementary reactions were used to initiate all transition state searches. In contrast to template-based network exploration, this approach allows both conventional reactions and unexpected reactions to be discovered by the algorithm. A second distinction is that all discovered reactions were also tested for alternative water-catalyzed mechanisms. For example, the exploration rediscovered five reaction types that have frequently been invoked in glucose pyrolysis studies, including Korcek cyclizations (e.g., 2-1 $\rightarrow$ 3-1 in Fig. 1), hydrogen migration rearrangements (e.g., 2-1 $\rightarrow$ 3-3), keto-enol tautomerization (e.g., 2-1 $\rightarrow$ 5-5), retro-aldol reactions (e.g., 2-1 $\rightarrow$ 3-4), and β-elimination reactions (e.g.,

2-1 → 6-4). Detailed reaction mechanisms are provided in Supporting Information section 2. Water was observed to catalyze Korcek cyclization, keto-enol tautomerization, and β-elimination reactions, but had little effect on hydrogen migration rearrangement and retro-aldol reaction, and sometimes even increased the activation energy (see Fig. S5 for a comparison of water-catalyzed and non-catalyzed transition state geometries and activation energies).

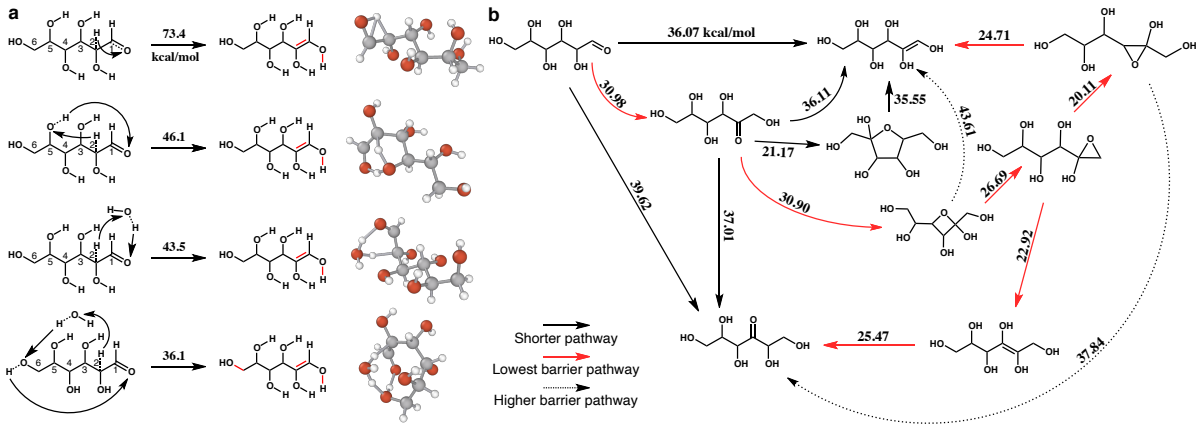

Figure 2: Unexpected reaction mechanisms and reaction pathways that reduce the activation energy. (a) Four different reaction mechanisms of a keto-enol tautomerization. From up to down: direct transfer of an hydrogen atom (b2f2); intramolecular hydroxyl group catalyzed hydrogen transfer (b3f3); single water-catalyzed hydrogen transfer (b3f3); inter- and intramolecular catalyzed hydrogen transfer (b4f4). (b) A sub-network of the formation pathways to two important intermediates: 3-hexulose and hexene-1,2,3,4,5,6-hexanol. The arrows denoted as red represent the reaction pathways with the lowest overall barriers while the arrows with dotted lines refer to alternative reaction pathways with higher barriers.

Catalyzed proton transfers feature prominently in the final network. Striking examples are provided by several low-barrier routes to keto-enol tautomerization of D-Glucose (Fig. 2). Tautomerization of D-Glucose to hexene-1,2,3,4,5,6-hexanol (HEH) involves the direct transfer of a hydrogen atom from an α carbon (index 2 in Fig. 2a) to the carbonyl oxygen through the breaking a σ and π bond. This uncatalyzed breaking of two bonds and forming two bonds (i.e., a b2f2 reaction mechanism) has a relatively high $\Delta G^{\dagger}$ of 73.4 kcal/mol, which is unlikely to occur at low temperatures (e.g. below 500 °C). However, the exploration revealed three other catalyzed

mechanisms that reduced the barrier by up to half. An intramolecularly catalyzed pathway was discovered utilizing the hydroxyl group at the 5-position (Fig. 2a) with a $\Delta G^{\dagger}$ of 46.1 kcal/mol. A similar catalyzed mechanism with water acting in place of the hydroxyl reduces $\Delta G^{\dagger}$ to 43.5 kcal/mol. Both mechanisms have been previously discussed but were rediscovered here without explicit guidance to the exploration algorithm.[18,35] YARP also identified an unreported mechanism catalyzed by a proton shuttle network formed by a water molecule and a hydroxyl group at the 6- position (Fig. 2a), which reduced the $\Delta G^{\dagger}$ to 36.1 kcal/mol. This new pathway strongly favors HEH formation and has hitherto been missed by studies relying on manual TS characterization. The use of conformational sampling and automated TS characterization revealed many examples of catalyzed reactions throughout the network.

The network also includes many examples of multi-step reaction pathways with three- and four-membered rings as intermediates with significantly reduced overall barriers compared with the analogous single-step conversion. For example, 3-hexulose (7-6 in Fig. 1) and HEH (5-5) are important intermediates of many major products from glucose pyrolysis. The commonly acknowledged lowest overall barrier reaction pathways of the formation of 3-hexulose and HEH are either from D-Gluclose or from D-Psicose with overall barriers between 40-45 kcal/mol, respectively.[35] These pathways were also discovered here, along with several unreported lower barrier reaction pathways (Fig. 2). When only considering D-Glucose and D-Psicose, the overall barriers of 3-hexulose and HEH formation are 37.01 and 36.07 kcal/mol, respectively (Fig. 2b). However, more relevant reactions were explored that further reduced the overall barrier. The formation pathway of HEH with β-D-Psicofuranos (4-1 in Fig. 1) as an intermediate reduced the overall barrier to 35.55 kcal/mol. More interestingly, a series of low barrier ring transformations connected both 3-hexulose and HEH with an overall barrier of 30.98 kcal/mol. The discovery of these reaction pathways challenged previous computational studies in which the barrier to glucose conversion was supposed to be more than 45 kcal/mol or even 60 kcal/mol and provides new insights into the

possible existence of intermediates and reaction mechanisms.

2.3 Reaction Pathways to Major Experimental Pyrolysis Products

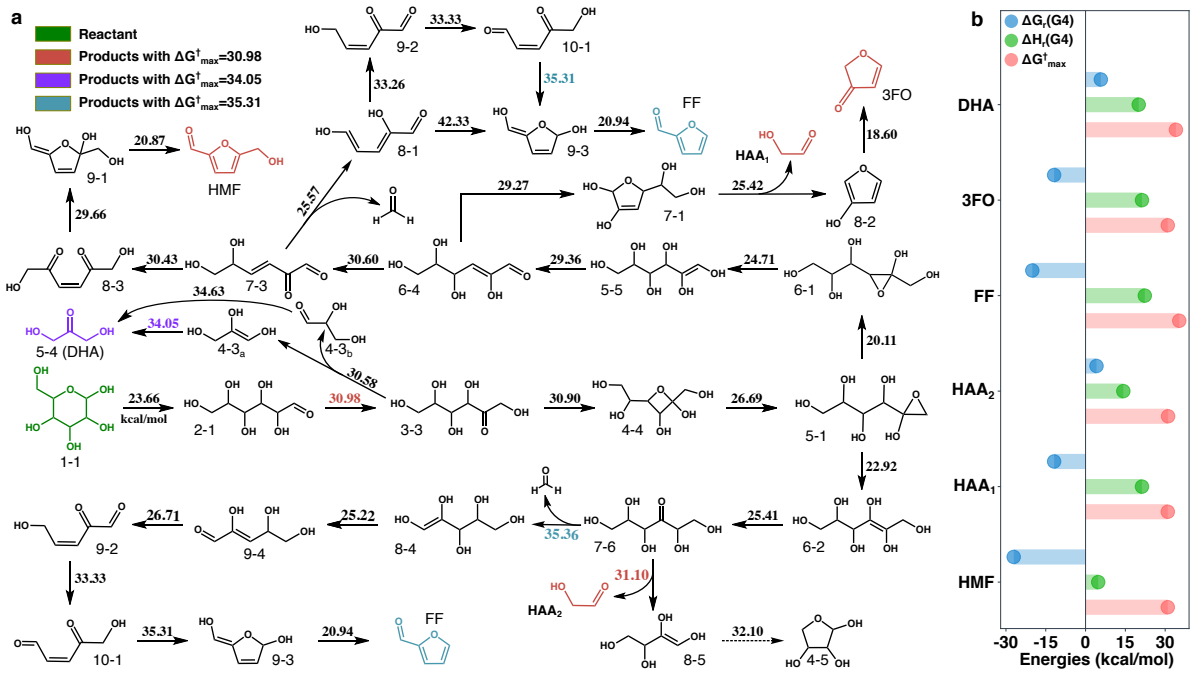

Figure 3: Reaction pathways to major products. (a) Summary of the low activation energy pathways identified by combining the forward and backward reaction explorations. Five major products which are observed in the experimental work are included, namely HMF, HAA, FF, DHA and 3FO. The index of each intermediate is identical to Figure 1. (b) Comparisons of the Gibbs free energies of reaction (ΔG_r), enthalpies of reaction (ΔH_r) and the overall barriers ($\Delta G^{\dagger}_{max}$) of each reaction pathway. ΔG_r and ΔH_r are computed at Gaussian-4 level while $\Delta G^{\dagger}_{max}$ is computed at DFT level (B3LYP-D3/TZVP)

The reaction network discussed in the previous sections comprises an exploration starting from β-D-Glucose that followed lowest overall barrier pathways out to a depth of eight sequential reactions. At that stage, there is an explosion of 32 intermediates with the same overall barrier of 30.98 kcal/mol due to a shared rate limiting step in the network. Although other metrics, like the ΔG of the second rate-limiting step or direct microkinetic modeling, might have been used to facilitate further exploration, forward exploration was terminated here since this already constitutes

the most comprehensive exploration of the β-D-Glucose pyrolysis network with TS calculations to date. This network also includes complete reaction pathways to two of the major experimental pyrolysis products (in terms of mass percent), hydroxyacetaldehyde (HAA) and dihydroxyacetone (DHA). Nevertheless, we still wished to provide pathways to remaining products, even if they were not spontaneously discovered by the forward exploration.

To achieve this, we used the structures of the missing products to perform a series of "backward" reaction explorations starting at these products for a fixed number of steps or until they connected with the forward network. Based on recent experimental work by Fang et al., the major products obtained from glucose pyrolysis at 350 °C are hydroxymethylfurfural (HMF, 20.0%), furfural (FF, 15.0%), hydroxyacetaldehyde (HAA, 13.5%). In addition, 3-(2H)-furanone (3FO, 5.0%), dihydroxyacetone (DHA, 3.9%) and 3-hydroxy-γ-butyrolactone (HBL, 3.4%) are also observed with lower yields ($\leq$ 5%). Since the formation pathways of HAA and DHA were already established in the forward reaction exploration, backward explorations were performed for the remaining four products (Supporting information section 1). Backward reaction exploration successfully connected HMF, FF, and 3FO with the forward reaction network (Fig. 1, a specific discussion of the reaction pathways to the minor product HBL is provided in Supporting information Section 1).

The low activation energy pathways to the formation of the five major experimental products are summarized in Figure. 3. All products share the first two steps (1-1→2-1→3-3), then DHA is formed in a single step (shown in purple, with $\Delta G^{\dagger}_{max}$ of 34.05 kcal/mol), while the other four products share two additional steps (3-3→4-4→5-1) before branching through the intermediates 6-1 and 6-2 (i.e., precursors to HEH and 3-hexulose, respectively). These latter branches are shown in the upper and lower part of Figure 3a, respectively. In the discussion that follows several products can be formed through either branch with similar barrier and these branches will be referred to as the HEH and hexulose pathways.

HMF Formation Pathway. HEH is a key intermediate involved in the formation of HMF.

Starting from HEH (species 5-5), sequential 1,4-conjugated elimination reactions remove two water molecules to form species 7-3. Hydrogen migration within species 7-3 results in a symmetric diketone diol intermediate (8-3). Two double-water catalyzed reactions facilitate cyclization and a further water elimination to produce HMF. The transformation of 8-3 to 9-1 is a ring closure with an extremely long distance proton transfer (5 bonds separate the proton donor and acceptor, Fig. S6a). The dehydration step similarly involves a proton and hydroxyl separated by 5 bonds (Fig. S6c). These long distance proton transfers use two external water molecules as a bridge for proton-shuttling, resulting in low catalyzed activation energies of 29.66 and 20.87 kcal/mol, respectively. Based on this new pathway, we predict that the rate limiting step of HMF formation is the isomerization from D-glucose to D-Psicose with an overall barrier of 30.98 kcal/mol.

FF Formation Pathway. Two reaction pathways to form FF from either HEH or 3-hexulose, respectively, were identified from the reaction exploration with similar overall barriers. The HEH pathway to FF diverges from the analogous pathway to HMF at species 7-3, where a formaldehyde elimination reaction occurs (retro-aldol reaction) to yield species 8-1 instead of hydrogen migration in the case of the HMF pathway. Intermediate 8-1 can go through either a three-step reaction mechanism involving a proton transfer and [1,5] bond shift (Fig. S6d), hydrogen migration rearrangement and double water-catalyzed ring closure, or a one-step single water-catalyzed cyclization to form intermediate 9-3 (similar to Fig. S6b). Similar to HMF formation, double water-catalyzed dehydration reactions are the last step in all pathways to form FF (9-3→FF in Fig. 3a). For the alternate hexulose pathway, after the elimination of formaldehyde from 3-hexulose, the remaining steps have the same reaction mechanism as the pathway forming HMF. The rate-limiting step of these two pathways are cyclization and formaldehyde elimination with overall barriers of 35.31 and 35.36 kcal/mol, respectively.

HAA and 3FO Formation Pathways. The major pathways forming HAA also involve either 3-hexulose or HEH as an intermediate. Starting from 3-hexulose, HAA is formed through a

one-step retro-aldol reaction, with ΔG of 31.10 kcal/mol. Starting from HEH, the pathway to HAA diverges from the pathways to HMF and FF at species 6-4 (denoted as 3-DGE). Competing with the 1,4-conjugated elimination, a cyclization reaction with similar activation energy can occur on 3-DGE, resulting in the formation of a 5-membered ring (7-1). This species can undergo a surprising single-step reaction that produces HAA and 3-Furanol via a 1,4-conjugated elimination coupled with a β-elimination (Fig. S6e). From here, 3-Furanol can be converted into the more stable 3-(2H)-furanone (3FO) through a double-water catalyzed keto-enol tautomerization reaction with $\Delta G = 18.60$ kcal/mol (Fig. S6f). It is thus observed that the overall barriers to the formation of HAA and 3FO via the HEH-pathway are both 30.98 kcal/mol.

From the detailed analysis of the lowest overall barrier pathways for each product, the formation of HMF, HAA and 3FO have the lowest overall barrier of 30.98 kcal/mol, while the overall barriers for the formation of DHA and FF are slightly higher at 34.05 and 35.31 kcal/mol, respectively. Since these differences are within the accuracy of DFT, the Gibbs free energies of reaction (ΔG_r) and enthalpies of reaction (ΔH_r) were calculated at the Gaussian-4 level[41] to assist the analysis of the yields of each product (Fig. 3b). The ΔG_r, ΔH_r and $\Delta G^{\dagger}_{max}$ of HMF formation are the lowest among all the products, which indicates that HMF is both the kinetically and thermodynamically most favorable product. This conclusion is in agreement with the experimental observation that HMF is the predominant pyrolysis product. Although the overall barrier of HAA is the same as HMF, its ΔG_r and ΔH_r explain why its yield is lower, especially at low temperatures. However, considering that multiple reaction pathways can form HAA (see Fig.5 for more information), the yield of HAA is still predicted to be much higher than 3FO. Furthermore, the thermodynamic preference of FF formation contributes to the relatively high yield of FF, despite the sightly higher overall barrier than the other major products. To our knowledge, this represents the first reaction network that self-consistently describes the relative yields of these experimental products with complete pathways and transition states.

2.4 Comparing Competing Reaction Pathways for HMF Formation.

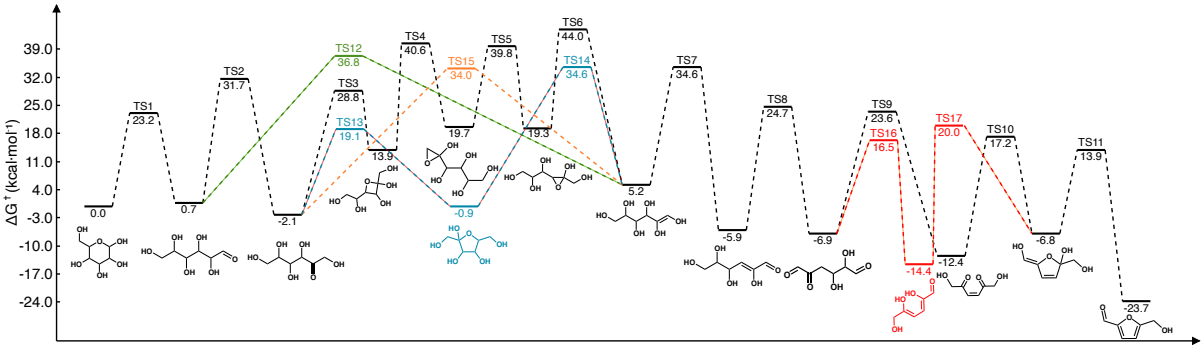

Figure 4: A summary of all competitive reaction pathways of forming HMF. The black lines refer to the reaction pathway with the lowest overall barrier. The reaction pathways denoted as green, blue orange and red are other competing pathways with lower absolute barrier or more kinetically accessible intermediates. In addition, the reaction steps denoted by orange lines plus red lines represent a similar mechanism recently reported by Kang et al., but contains consistent consideration of the catalytic effect of water

Many reaction mechanism studies have been performed over the past decade to resolve the mystery of why HMF is the major product of glucose pyrolysis. Although the lowest overall barrier pathway to HMF formation was discussed in detail in the previous section, the reaction network reveals other reaction pathways with similar barriers that would also contribute to the HMF formation (Fig. 4). Although the reaction pathway denoted as black has the lowest overall barrier of 30.98 kcal/mol, the corresponding absolute barrier ($\Delta G^{\dagger}_{abs}$), which is defined as the energy of the highest-energy TSs with respect to the initial reactant (β-D-Glucose), is not the lowest over all pathways. In addition, the competing reactions of each intermediate also affect the analysis of kinetic preference. Taking these factors into account, the reaction pathways indicated in blue, green and orange are predicted to play an important role in the formation of HMF. For example, instead of sequential multi-step ring transformations, the reactions marked in green and orange are direct keto-enol tautomerization catalyzed by both water and intramolecular hydroxyl groups (Fig. 2a). Starting from D-Glucose and D-Psicose, the overall and absolute barriers of these two

reaction pathways are 36.07 and 36.8 (TS12), and 36.11 and 34.6 (TS7) kcal/mol, respectively. The reaction pathway denoted as the blue line goes through β-D-Psicofuranose, which is a commonly hypothesized intermediate in HMF formation studies, with an overall and absolute barriers of 35.55 and 34.6 (TS14 and TS7) kcal/mol, respectively. In contrast to previous studies that proposed several β-elimination steps occurring on β-D-Psicofuranose to form HMF, YARP discovered a conversion pathway from β-D-Psicofuranose to HEH through a water-catalyzed ring opening reaction, thereby reducing the activation energy (ΔG = 35.55 kcal/mol, $\sim$ 15 kcal/mol lower than the water-catalyzed β-elimination), and making the β-D-Psicofuranose pathway kinetically accessible. More importantly, β-D-Psicofuranose is the most kinetically favorable reactant at a depth of four in the network, which makes the β-D-Psicofuranose pathway more competitive. The other two competing reaction pathways were identified from the backward search (Fig. S1) and are distinguished by the black line (TS9 and TS10) and the red line (TS16 and TS17). The reaction pathway denoted as black contains a hydrogen migration rearrangement (ΔG = 30.43 kcal/mol) and a double water-catalyzed ring closure with a long-distance hydrogen shift (ΔG = 29.66, Fig. S6a), while the reaction pathway denoted as red contains a [1,5] shift (ΔG = 23.33, similar to Fig. S6d) and a single-water-catalyzed ring closure with a shorter-distance hydrogen shift (ΔG = 34.37, Fig. S6b). Both pathways have comparable overall and absolute barriers and predicted to be of similar kinetic relevance. Notably, the reaction mechanism represented by the combination of orange and red lines is similar to the HMF formation pathway recently proposed by Kang et al.[35] However, the corresponding overall barrier of the reaction mechanism proposed by Kang et al. is 12 kcal/mol higher than that discovered by YARP, which discovered lower-barrier water catalyzed TSs in several steps.

2.5 Comparing Competing Reaction Pathways and Experimental Data for HAA Formation

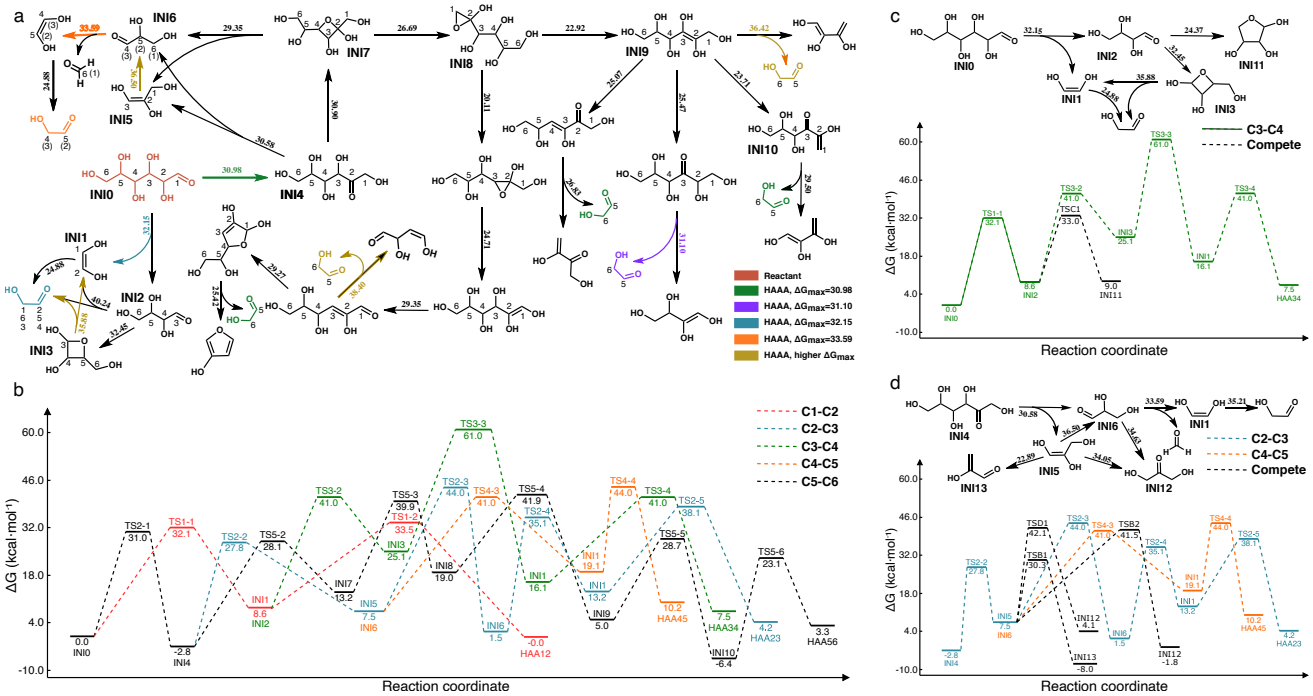

Figure 5: Competing reaction pathways of HAA formation. (a) A summary of kinetic favorable reaction pathways to HAA formation. The labeled carbon atoms reveal how 5 types of HAA molecules with carbon index of (1,2), (2,3), (3,4), (4,5) and (5,6) are formed. (b) A potential energy diagrams with most favorable reaction pathways to each type of HAA. The lines with different color represent different types of HAA. (c) Illustration of a competing reaction pathway against the formation of (3,4) HAA. (d) Illustration of two competing reaction pathways against the formation of (2,3) and (4,5) HAA.

A major challenge in reaction network characterizations is that, absent laborious experimental mechanistic studies, the predicted pathways can only be indirectly validated through comparisons with terminal product yields. Interesting experimental work by Lu et al that used ^{13}C-labeling to quantify the carbon fluxes from β-D-Glucose to specific positions in HAA provides a unique opportunity to directly test several of the competing mechanistic pathways predicted here (Fig. 5).[39] In total, five distinct ^{13}C-label patterns can be produced from the total of eleven kinetically

relevant pathways for HAA formation predicted here (i.e., HAA, which has two carbons, can inherit those carbons from the [1,2], [2,3], [3,4], [4,5], or [5,6] carbons in β-D-Glucose based on the labeling in Fig. 5a). These five distinct labelled products are consistent with those found experimentally. Seven distinct pathways were discovered to form (5,6)-HAA, while for all other labeled products only one kinetically relevant pathway was found (Fig. 5a). To compare the kinetic favorability of each HAA-labeling, the overall barriers, absolute barriers and the number of reaction steps are summarized in Table.1 and presented graphically in Figure 5b. To perform a quantitative comparison with the experimental fluxes, we also considered the possibility of double-water catalyzed reactions (i.e., the network exploration was done considering only single-water catalyzed reactions). A more stable double water-catalyzed TS was discovered for the keto-enol tautomerization of 1,2-Ethenediol ($\Delta G^{\dagger}$ = 24.88 kcal/mol; the single water-catalyzed barrier is $\Delta G^{\dagger}$ = 35.21 kcal/mol). An energy diagram with only single-water catalyzed results is presented in Fig. S7 for comparison.

The potential energy diagram shows that the reaction pathways with the two minimum $\Delta G^{\dagger}_{max}$ and $\Delta G^{\dagger}_{abs}$ produce (1,2)-HAA and (5,6)-HAA. Experimentally, these species are responsible for 86.7% of total HAA yield at 300°C.[39] Although similar overall barriers are associated with the formation of (1,2)-HAA and (5,6)-HAA, the absolute barrier of (1,2)-HAA formation is 10 kcal/mol lower than (5,6)-HAA, suggesting that (1,2)-HAA is kinetically more favorable. In addition, the reduced number of reaction steps for (1,2)-HAA formation limits the influence of competing reactions (not shown in typical energy level diagrams) compared with (5,6)-HAA. Furthermore, (1,2)-HAA exhibits the lowest free energy of reaction due to side-products associated with its formation pathway, which demonstrates that the formation of (1,2)-HAA is both kinetically and thermodynamically more favorable (Fig. 5b). The contribution of other reaction pathways to HAA yield is predicted to increase with temperature, and the difference between (1,2)- and (5,6)-HAA becomes smaller, which is in consistent with the experimental trend (Table.1).

Table 1: Summary of reaction properties and experimentally observed proportions of five types of HAA

HAA index	C1-C2	C2-C3	C3-C4	C4-C5	C5-C6[a]
$\Delta G^{\dagger}_{max}$ (kcal/mol)	32.1	36.5	35.9	33.6	31.0
$\Delta G^{\dagger}_{abs}$ (kcal/mol)	33.5	44.0	61.0	44.0	41.9
Number of reaction steps	2	5	4	4	6
Proportions at 300° C	58.5	0.0	13.3	0.0	28.2
Proportions at 500° C	39.5	2.3	18.8	5.3	32.9

a. Among the seven reaction pathways that lead to the formation of (5,6)-HAA, the pathway through the intermediate INI10 is presented in the table selected because it has the lowest overall and absolute barrier and the shortest pathway.

The potential energy diagram (Fig. 5b) also provides an explanation for the relatively low yields of the other three labelings, especially (2,3)-HAA and (4,5)-HAA, for which no previous studies have provided corresponding reaction mechanisms. Although these isotopic pathways have similar overall barriers as (1,2)-HAA, the absolute barriers are much higher. Comparing the formation of (2,3)-HAA and (4,5)-HAA, the formation pathway of (2,3)-HAA requires an additional transformation from INI5 to INI6 (glyceraldehyde), which leads to both a higher overall barrier and more reaction steps. This comparison also illustrates that the absolute barrier can sometimes be misleading, since one exothermic reaction can reduce the absolute barriers of all following steps. The occurrence of competing reactions also affects the yield of each pathway. Based on the entire reaction network (Fig. 1), one, two, and two relatively low barrier (with 2-14 kcal/mol reductions in $\Delta G^{\dagger}_{max}$) reactions compete with the formation of (3,4)-HAA, (2,3)-HAA, and (4,5)-HAA, respectively (Fig. 5c-d). On the one hand, as for (3,4) HAA, the formation of a 5-membered ring (node 4-5 in Fig. 1) has lower barrier than the formation of the intermediate INI3. However, this competing intermediate is less stable than the two HAA molecules produced throughout the degradation process (i.e. [3,4]-HAA and [5,6]-HAA), and has only one kinetic accessible reaction pathway, which is converting back to INI2. On the other hand, two competing reactions correspond

to the formation of (2,3)-HAA and (4,5)-HAA are both kinetically and thermodynamically more favorable. Specifically, the formation of DHA has similar barrier compared to the formation of (4,5)-HAA, but with 6 kcal/mol reduction in ΔG_r. Similarly, the formation of INI13 (which can be further converted into a more stable isomer, species 5-2 in Fig. 1) exhibits 13.6 and 12.2 kcal/mol reductions in $\Delta G^\dagger$ and ΔG_r, respectively, compared to the formation of (2,3)-HAA. The presence of these competing reaction pathways explains the experimental observation that (2,3)-HAA and (4,5)-HAA are only observed at higher temperatures (e.g. 500°C, Table. 1).

3 Conclusions

As it becomes routine to computationally characterize the reactivity of a given set of reactants, the next challenge is developing efficient algorithms for selecting reactants to characterize. Combining these capabilities (reactivity prediction and reactant exploration) is the crux of resolving deep reaction networks. In this study, we have shown one way in which this can be done, by combining YARP (a reactivity prediction tool) and Dijkstra's algorithm (a network exploration algorithm) to elucidate the water-catalyzed pyrolysis network of D-glucose. Despite the lack of domain expertise for this problem amongst the present authors, this algorithm (re)discovered most state-of-the-art pyrolysis pathways, provided low overall barrier pathways ($<$ 40 kcal/mol) connecting all major experimental products to D-Glucose, provided the first self-consistent explanation of experimental isotopomer yields, and predicted several new lowest overall barrier pathway to major products. The analysis of HMF formation pathways, in particular, revealed many newly reported pathways with overall and absolute barriers as low as $\sim$ 30 kcal/mol that challenge existing mechanistic proposals. The most surprising aspect of these results is the extreme simplicity of using the overall barrier as a cost function, contrasted with the subtlety of the new pathways proposed by the algorithm. The success of this remarkably simple exploration policy suggests that general purpose

exploration policies are both possible and likely simpler than anticipated.

There are still several aspects for improving the current approach. First, although it it is impressive how well using the overall barrier worked as a cost function, many more informative alternatives are readily at hand that will likely prove advantageous. A trivial extension would be to use the overall barrier, then next largest barrier, and so on, to break ties for nodes that share the same rate limiting step. A more quantitative approach would be to use the results of microkinetic modeling on the partially explored network to determine the most relevant nodes to explore. Second, the current network only considered catalyzed unimolecular reactions. This is a fine assumption for the vast majority of intermediates with negligible concentration, but bimolecular reactions amongst moderate to high concentration intermediates are certainly expected in general. Here too, a more sophisticated cost function or microkinetic modeling will be necessary for determining when an intermediate should be considered as potentially available for bimolecular reactions. Third, although the current study considered the catalytic effects of water, some reactions (e.g., keto-enol tautomerization) prefer double-water catalyzed mechanisms. Thus, a systematic investigation of the catalytic effects of multiple water molecules can provide even more accurate reaction network prediction. Finally, although the reaction network exploration was done automatically, reaction mechanisms summarization and classification remain manual tasks. Automatically interpreting reaction network data and utilizing this information to speed exploration will be essential as the scale of networks being explored increase, as we expect them to.

4 Computational Methods

The detailed computational methodologies and settings are provided in this section, including basic YARP settings to explore the reaction space, water-catalyzed reaction mechanism generation, construction of forward and backward reaction networks, and computational details.

YARP calculations. The reaction characterization and analysis were performed by the YARP v2.0 package,[38] using GFN2-xTB[42,43] to pre-explore the potential energy surface and B3LYP-D3/TZVP as a more accurate DFT level of theory to quantitatively describe the reaction energies and activation energies. The YARP methodology has been described in detail in several places at this point.[38] Here, we only highlight the settings that are specific to this study. An enthalpy of reaction threshold of 20 kcal/mol was set to avoid exploring highly endothermic reactions. An activation energy threshold was also used to avoid exploring high barrier reactions. Reactions with $\Delta G^{\dagger}$ over 50 kcal/mol at the GFN2-xTB level were excluded from DFT characterizations. For reactions that passed the GFN2-xTB level filter, a DFT-level TS was characterized. Reactions with DFT-level $\Delta G^{\dagger}$ over 45 kcal/mol were excluded from IRC calculations since these are typically the most expensive step in the workflow. A conditional break three bonds and form three bonds (Cb3f3) elementary reaction step (ERS) was selected to enumerate possible uni-molecular reactions for each intermediate. The Cb3f3 ERS includes all b2f2 reactions and the subset of b3f3 reactions with at least one π-bond breaking and forming during the reaction (i.e., the π-bond changes positions but the underlying σ-bond does not break). We have previously shown that b3f3 reactions exclusively involving the breaking and forming of σ-bonds are rarely kinetically relevant. Once the reactions were enumerated, up to four reaction conformations were generated by a reaction conformational sampling algorithm.[44] For small or highly constrained systems, less than four conformations are sometimes generated. These conformers were used as initial geometries for double-ended TS searches.

Water catalyzed reaction generation. In previous work we have described an algorithm for including water in the TS search to assess whether proton shuttling can be catalyzed.[45] Here, water-catalyzed reaction mechanism(s) were tested for every reaction with at least one hydrogen atom as the reactive atom. The basic principle is to introduce an additional water molecule as a proton shuttle that may relax the transition state geometry and reduce the activation energy. The

procedure for characterizing the water-catalyzed reaction mechanisms is the same as in our previous study,[45] including updating the bond breaking and forming information via the bond-electron matrix, converting 2D reaction representations into 3D reactant-product geometries through a joint-optimization procedure,[32] and applying the reaction conformational sampling algorithm to generate up to five reaction conformations,[44] The Cb3f3 elementary reaction type may enumerate reactions that involve up to two reactive hydrogen atoms. For reactions with a single reactive hydrogen atom, there is only one possible catalyzed reaction mechanism.[45] For reactions with two reactive hydrogens, two water-catalyzed mechanisms are possible.[45] In these cases, all water-catalyzed mechanisms were characterized that were consistent with a given reaction.

Forward reaction exploration with a modified Dijkstra algorithm. Naive complete exploration of the reaction network by applying the ERS to all products at each step (i.e., characterizing all ERS products of the ERS products ... of the ERS products of L-glucose) grows exponentially and makes deep exploration impossible. For example, complete b2f2 exploration of the unimolecular decomposition network of γ-ketohydroperoxide of depth of two already contains hundreds of intermediates,[32] which is relatively simple compared with L-glucose. However, the number of kinetically relevant reactions in established reaction networks is linearly proportional to the number of intermediates, which suggests that the kinetically relevant sub-network only grows linearly rather than exponentially.[46,47] One means of avoiding exploration of kinetically irrelevant pathways is to use an activation energy threshold and avoid further exploration downstream of a high-barrier reaction.[45,48] However, this simple approach only reduces the base of the exponential scaling with network depth and is still impractical for deep exploration. Using thresholds that do not consider the whole network also neglects the possibility that a pathway that seems kinetically irrelevant might become relevant again if more severe bottlenecks are encountered downstream of the branching point.

Thus, here a modified Dijkstra algorithm was developed to guide the exploration of reactive

pathways that retains linear scaling with depth while permitting the possibility of backtracking. In the traditional Dijkstra algorithm, the number of search beams (i.e., the number of unvisited nodes that are explored at each step) is set to one and the objective function is the sum of costs along the minimum cost pathway. Here, the number of search beams was set to five to achieve higher parallelization efficiency while the minimum overall barrier for each intermediate was used as the objective function. Here, "overall barrier" refers to the maximum $\Delta G^{\dagger}$ along a potentially multi-step pathway. Because a given product might be reached by several different pathways, the "minimum overall barrier" refers to the pathway yielding the product with lowest overall barrier. This is in very close analogy to the minimum cost pathway of Dijkstra's original formulation.

To also take the uncertainty associating with the level of theory selection and conformational sampling into account, a soft margin of 1 kcal/mol was used for selecting candidates for further reaction exploration. For example, the difference between the overall barriers of intermediates 4-5 and 4-6 is only 0.3 kcal/mol (32.15 and 32.45 kcal/mol, respectively), thus both intermediates were considered reactants for the fourth-layer reaction exploration. Accounting for the soft margin sometimes led to the effective beam count being greater, but never less than five.

Another point to emphasize is the real-time update of overall barriers as the exploration progresses. For example, the overall barrier of the key intermediate 7-6 changed from 39.62 (depth2) to 37.01 (depth3) and was finally updated to 30.98 kcal/mol at depth6 due to the discovery of alternative lower barrier pathways to this product during the exploration. Although simply using the overall barrier as the objective function effectively controlled the searching scope in the first seven exploration steps, 32 intermediates with the same lowest overall barrier of 30.98 kcal/mol emerged after the eighth exploration step due to a shared upstream reaction bottleneck. A more sophisticated cost function could be developed to deal with such situations, but here explorations to this depth already discovered lowest barrier pathways to most experimentally observed products and the remaining (minor) products were connected to the network with backward reaction

exploration.

Backward reaction exploration. The backward reaction exploration is similar to the forward reaction exploration, but with the overall barriers of "backward" reaction as the objective function (i.e., for a reaction A→B, the barrier of B→A was used to determine if B should be a candidate for further exploration). The rationale for this is that the backward exploration is meant to connect with the forward network and so from that perspective the kinetic relevance is determined by the backward reactions. In addition, since dehydration is a common type of reaction that must be involved in the formation of certain products (e.g., HMF), unimolecular transformations as well as bimolecular reactions with a water molecule, and the corresponding water-catalyzed reaction mechanisms were generated and characterised for each product (or intermediate) in the backward reaction exploration. The exploration terminated once an intermediate from the forward exploration was identified in the backward exploration. Detailed descriptions of the backward exploration starting from HMF, 3FO, and HBL are provided in the supporting information section 1.

Computational details. DFT calculations were carried out using Gaussian 16.[49] All GFN2-xTB calculations were performed with the xTB program (version 6.4.0).[43] All simulations were run on a 448 node commodity cluster composed of two Rome CPUs (2.0GHz), 128 effective cores, and 256 GB of memory per node. DFT calculations were performed with 16-core parallelization, while all other calculations were performed as bundled single-core jobs. The enthalpy of reaction predictions were performed using the TCIT package.[50,51]

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.

The version of YARP used in this study and a guide to reproducing the results is available through GitHub under the GNU GPL-3.0 License [https://github.com/zhaoqy1996/YARP].

Further raw data sources generated by this work are available at (XXX, figshare link), including raw output files and molecular geometries.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

The work performed by Q.Z. and B.M.S 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 Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering and the Purdue Process Safety and Assurance Center.

References

(1) Bridgwater, A. V.; Meier, D.; Radlein, D. An overview of fast pyrolysis of biomass. *Org. Geochem.* **1999**, *30*, 1479–1493.

(2) Bridgwater, A. V.; Peacocke, G. V. C. Fast pyrolysis processes for biomass. *Renewable Sustainable Energy Rev.* **2000**, *4*, 1–73.

(3) Czernik, S.; Bridgwater, A. V. Overview of applications of biomass fast pyrolysis oil. *Energy Fuels* **2004**, *18*, 590–598.

(4) Mohan, D.; Pittman Jr, C. U.; Steele, P. H. Pyrolysis of wood/biomass for bio-oil: a critical review. *Energy Fuels* **2006**, *20*, 848–889.

(5) Corma, A.; Iborra, S.; Velty, A. Chemical routes for the transformation of biomass into chemicals. *Chemical reviews* **2007**, *107*, 2411–2502.

(6) Yang, H.; Yan, R.; Chen, H.; Lee, D. H.; Zheng, C. Characteristics of hemicellulose, cellulose and lignin pyrolysis. *Fuel* **2007**, *86*, 1781–1788.

(7) Regalbuto, J. R. Cellulosic biofuels—got gasoline? *Science* **2009**, *325*, 822–824.

(8) Bridgwater, A. V. Review of fast pyrolysis of biomass and product upgrading. *Biomass Bioenergy* **2012**, *38*, 68–94.

(9) Pecha, M. B.; Arbelaez, J. I. M.; Garcia-Perez, M.; Chejne, F.; Ciesielski, P. N. Progress in understanding the four dominant intra-particle phenomena of lignocellulose pyrolysis: chemical reactions, heat transfer, mass transfer, and phase change. *Green Chem.* **2019**, *21*, 2868–2898.

(10) Román-Leshkov, Y.; Chheda, J. N.; Dumesic, J. A. Phase modifiers promote efficient production of hydroxymethylfurfural from fructose. *Science* **2006**, *312*, 1933–1937.

(11) Zhao, H.; Holladay, J. E.; Brown, H.; Zhang, Z. C. Metal chlorides in ionic liquid solvents convert sugars to 5-hydroxymethylfurfural. *Science* **2007**, *316*, 1597–1600.

(12) Patwardhan, P. R.; Satrio, J. A.; Brown, R. C.; Shanks, B. H. Product distribution from fast pyrolysis of glucose-based carbohydrates. *J. Anal. Appl. Pyrolysis* **2009**, *86*, 323–330.

(13) Vinu, R.; Broadbelt, L. J. A mechanistic model of fast pyrolysis of glucose-based carbohydrates to predict bio-oil composition. *Energy Environ. Sci.* **2012**, *5*, 9808–9826.

(14) Pagan-Torres, Y. J.; Wang, T.; Gallo, J. M. R.; Shanks, B. H.; Dumesic, J. A. Production of 5-hydroxymethylfurfural from glucose using a combination of Lewis and Brønsted acid catalysts in water in a biphasic reactor with an alkylphenol solvent. *ACS Catal.* **2012**, *2*, 930–934.

(15) 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.

(16) Chen, S.; Wojcieszak, R.; Dumeignil, F.; Marceau, E.; Royer, S. How catalysts and experimental conditions determine the selective hydroconversion of furfural and 5-hydroxymethylfurfural. *Chem. Rev.* **2018**, *118*, 11023–11117.

(17) 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. *J. Anal. Appl. Pyrolysis* **2020**, *149*, 104846.

(18) Assary, R. S.; Redfern, P. C.; Greeley, J.; Curtiss, L. A. Mechanistic insights into the decomposition of fructose to hydroxy methyl furfural in neutral and acidic environments using high-level quantum chemical methods. *J. Phys. Chem. B* **2011**, *115*, 4341–4349.

(19) Jadhav, H.; Pedersen, C. M.; Sølling, T.; Bols, M. 3-Deoxy-glucosone is an Intermediate in the Formation of Furfurals from D-Glucose. *ChemSusChem* **2011**, *4*, 1049–1051.

(20) Hu, B.; Lu, Q.; Jiang, X.; Dong, X.; Cui, M.; Dong, C.; Yang, Y. Pyrolysis mechanism of glucose and mannose: The formation of 5-hydroxymethyl furfural and furfural. *J. Energy Chem.* **2018**, *27*, 486–501.

(21) Easton, M. W.; Nash, J. J.; Kenttämaa, H. I. Dehydration pathways for glucose and cellobiose during fast pyrolysis. *J. Phys. Chem. A* **2018**, *122*, 8071–8085.

(22) Zhang, Y.; Liu, C.; Xie, H. Mechanism studies on β-d-glucopyranose pyrolysis by density functional theory methods. *J. Anal. Appl. Pyrolysis* **2014**, *105*, 23–34.

(23) Shang, C.; Liu, Z. P. Stochastic surface walking method for structure prediction and pathway searching. *J. Chem. Theory Comput.* **2013**, *9*, 1838–1845.

(24) Zimmerman, P. M. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem.* **2013**, *34*, 1385–1392.

(25) Maeda, S.; Taketsugu, T.; Morokuma, K. Exploring transition state structures for intramolecular pathways by the artificial force induced reaction method. *J. Comput. Chem.* **2014**, *35*, 166–173.

(26) Zimmerman, P. M. Navigating molecular space for reaction mechanisms: an efficient, automated procedure. *Mol. Simul.* **2015**, *41*, 43–54.

(27) 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.

(28) Simm, G. N.; Reiher, M. Context-driven exploration of complex chemical reaction networks. *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119.

(29) Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Efficient prediction of reaction paths through molecular graph and reaction network analysis. *Chem. Sci.* **2018**, *9*, 825–835.

(30) Van de Vijver, R.; Zádor, J. KinBot: Automated stationary point search on potential energy surfaces. *Comput. Phys. Commun.* **2020**, *248*, 106947.

(31) Martínez-Núñez, E.; Barnes, G. L.; Glowacki, D. R.; Kopec, S.; Peláez, D.; Rodríguez, A.; Rodríguez-Fernández, R.; Shannon, R. J.; Stewart, J. J. P.; Tahoces, P. G.; Vazquez, S. A. AutoMeKin2021: An open-source program for automated reaction discovery. *J. Comput. Chem.* **2021**, *42*, 2036–2048.

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

(33) 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.

(34) 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.

(35) Kang, P.-L.; Shang, C.; Liu, Z.-P. Glucose to 5-hydroxymethylfurfural: origin of site-selectivity resolved by machine learning based reaction sampling. *J. Am. Chem. Soc.* **2019**, *141*, 20525–20536.

(36) 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.

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

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

(39) Lu, Q.; Tian, H.-y.; Hu, B.; Jiang, X.-y.; Dong, C.-q.; Yang, Y.-p. Pyrolysis mechanism of holocellulose-based monosaccharides: The formation of hydroxyacetaldehyde. *J. Anal. Appl. Pyrolysis* **2016**, *120*, 15–26.

(40) 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-γ-butyrolactone and 3-(2 H)-furanone. *Energy Fuels* **2018**, *32*, 9519–9529.

(41) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108.

(42) Grimme, S.; Bannwarth, C.; Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z=1-86). *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

(43) 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. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(44) Zhao, Q.; Hsu, H.-H.; Savoie, B. M. Conformational Sampling for Transition State Searches on a Computational Budget. *J. Chem. Theory Comput.* **2022**, *18*, 3006–3016.

(45) Zhao, G., Qiyuan Sanjay; Savoie, B. Thermally Accessible Prebiotic Pathways for Forming RNA and Protein Precursors. *ChemRxiv* **2022**,

(46) Lu, T.; Law, C. K. Toward accommodating realistic fuel chemistry in large-scale computations. *Prog. Energy Combust. Sci.* **2009**, *35*, 192–215.

(47) Van de Vijver, R.; Vandewiele, N. M.; Bhoorasingh, P. L.; Slakman, B. L.; Seyedzadeh Khanshan, F.; Carstensen, H. H.; Reyniers, M. F.; Marin, G. B.; West, R. H.; Van Geem, K. M. Automatic mechanism and kinetic model generation for gas-and solution-phase processes: a

perspective on best practices, recent advances, and future challenges. *Int. J. Chem. Kinet.* **2015**, *47*, 199–231.

(48) Zhao, Q.; Xu, Y.; Greeley, J.; Savoie, B. M. Deep reaction network exploration at a heterogeneous catalytic interface. *Nat. Commun.* **2022**, *13*, 4860.

(49) Frisch, M. J. et al. Gaussian 16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(50) Zhao, Q.; Savoie, B. M. Self-Consistent Component Increment Theory for Predicting Enthalpy of Formation. *J. Chem. Inf. Model.* **2020**, *60*, 2199–2207.

(51) Zhao, Q.; Iovanac, N. C.; Savoie, B. M. Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds. *J. Chem. Inf. Model.* **2021**, *61*, 2798—2805.