JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
Cite This: J. Am. Chem. Soc. 2019, 141, 20525−20536
pubs.acs.org/JACS

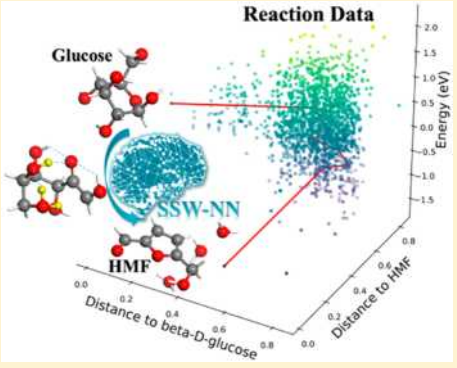
Glucose to 5-Hydroxymethylfurfural: Origin of Site-Selectivity Resolved by Machine Learning Based Reaction Sampling

Pei-Lin Kang, Cheng Shang,* and Zhi-Pan Liu*

Collaborative Innovation Center of Chemistry for Energy Material, Shanghai Key Laboratory of Molecular Catalysis and Innovative Materials, Key Laboratory of Computational Physical Science, Department of Chemistry, Fudan University, Shanghai 200433, P. R. China

Supporting Information

ABSTRACT: Glucose pyrolysis, a model system in biomass utilization, is renowned for its great complexity, deep in reaction network hierarchy and rich in reaction patterns. The selectivity in glucose pyrolysis, e.g., the high yield of 5-hydroxymethylfurfural (HMF), a value-added platform product, remains an intriguing puzzle even after 60 years of experimental study. Here we resolve the whole reaction network of glucose pyrolysis using a global-to-global technique for reaction pathway sampling. This is achieved by establishing the first organic chemistry reaction database via stochastic surface walking (SSW) global optimization, building the global neural network (G-NN) potential via machine learning and extensively exploring the reaction network of glucose pyrolysis. In total, 6407 elementary reactions, screened out from more than 150 000 reaction pairs in glucose pyrolysis, are collected in our reaction database. The established reaction network from SSW-NN, further validated by first-principles calculations, reveals that for glucose to HMF, the lowest energy reaction pathway involves fructose and 3-deoxyglucos-2-ene (3-DGE) as key intermediates and a site-selective reaction type, retro-Michael-addition, for three consecutive dehydration steps. The overall barrier is determined to be 1.91 eV, being at least 0.19 eV lower than all previously proposed mechanisms, which assumes direct β-H elimination dehydration. The lowest pathways to the other two major products, furfural (FF) and hydroxyacetaldehyde (HAA), are also discovered with a similar barrier 1.95 eV, which exhibit a competing nature by sharing the same key intermediate, 3-ketohexose. Since chemical reactions occurring in fast glucose pyrolysis are generally present in biomass chemistry, containing essentially all reaction patterns of C−H−O elements, the methodology designed and the results presented would help to advance reaction design and mechanistic modeling in renewable fuels from biomass.



1. INTRODUCTION

Biomass is a renewable feedstock for the production of chemicals and transportation fuels. D-Glucose, commonly known as sugar, can be obtained from cellulose, the most abundant component of biomass, by the ring-opening of the β-D-glucopyranose monomer of cellulose. Glucose conversion (e.g., pyrolysis) is therefore a straightforward route in biomass degradation and has raised considerable interest since the 1960s.[1−17] Due to the presence of many hydroxyl functional groups, glucose pyrolysis is notoriously complex with a huge number of possible reaction pathways, and how to improve the product selectity is a key question in the field. Mathematically, this would require a detailed knowledge of the potential energy surface (PES) of the glucose reaction system to identify all low energy reaction pathways and thus to determine the reaction kinetics. However, due to the lack of effcient PES sampling tools, the prediction of chemical reactivity for a given molecule without recourse to experiment has long been a dream in chemistry.

It is known that β-D-glucose pyrolysis (β-D-glucose can facilely convert to D-glucose) at 350−550 °C leads to several major products, including 5-hydroxymethylfurfural (HMF), hydroxyacetaldehyde (HAA), levoglucosan (LG), and furfural (FF) with the selectivity larger than 5% among more than 15 observed products with >0.1% selectivity (see Table 1). Since

Table 1. Typical Product Distribution of β-D-Glucose Pyrolysis in Experiments (FF: Furfural; LG: Levoglucosan; HMF: 5-Hydroxymethylfurfural; and HAA: Hydroxyacetaldehyde)

experiment[a]	1	2	3	4	5
HMF	20.03	7.7	11.93	3.28	12.91
HAA	13.47	6.62	12.8	16.45	16.45
FF	15	8.35	9.74	1.17	8.54
LG	1.04	7	0.15	8.34	1.26

[a]Experimental data **1** ref 26, (350 °C); **2**: ref 6, (500 °C); **3**: ref 26, (500 °C); **4**: ref 23, (500 °C); and **5**: ref 26, (550 °C). GC−MS were utilized for the product detection in all experiments.

Received: October 26, 2019
Published: December 4, 2019

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

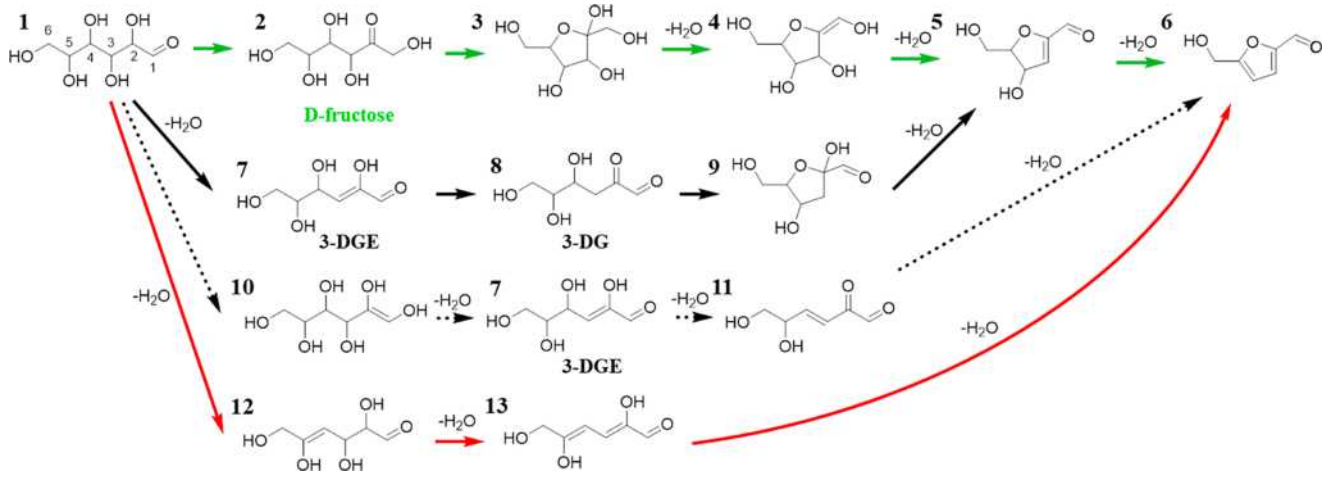


Figure 1. Proposed pathways in literature from glucose to HMF, namely the fructose path (green), 3-DG paths (black and black dotted), and direct path (red). The molecules are indicated by numbers and some key molecules are named as follows: **1**. D-glucose; **2**. D-fructose; **3**. D-fructofuranose; **6**. 5-hydroxymethylfurfural (5-HMF); **7**. 3-deoxyglucos-2-ene (3-DGE); **8**. 3-deoxyglucosone (3-DG); and **10**. hex-1-ene-1,2,3,4,5,6-hexaol (enol form of glucose).

the furan products, especially HMF, are highly valuable platform chemicals, how to control the selectivity in glucose conversion with mild conditions is the key task in research. This is however very challenging without a deep understanding of the reaction mechanism, considering that virtually all CHO fragments can be generated from glucose's $C_6H_{12}O_6$ composition. The past six decades have seen many different reaction pathways proposed, mainly based on basic knowledge of the organic chemistry, but they still cannot rationalize the observed selectivity in glucose pyrolysis, e.g., why HMF is the major product instead of smaller CHO fragments. It becomes even more surprising that the HMF selectivity can be extremely high under catalytic conditions.[2,3,9] Except for levoglucosan that can be formed in one step via dehydration of β-D-glucose with certainty, more than three pathways to HMF from D-glucose were cited popularly in the literature, not mentioning other pathways to HAA and FF, and low yield furan byproducts.[18−27] To be more specific, the representative pathways to HMF are outlined below and also shown in Figure 1.

Fructose path[23] (Figure 1, green line): The pathway proceeds via an initial isomerization of D-glucose to D-fructose. After one step of ring closure and three sequential dehydration steps, fructose can finally transform to HMF.

3-DG paths[21] (Figure 1, black line): The pathway starts from the loss of C3's hydroxyl group to form 3-deoxyglucos-2-ene (3-DGE, **7**). The 3-DGE then undergoes the enol-keto isomerization to 3-deoxy-D-glucosone (3-DG, **8**), the ring closure to **9**, and the dehydration to **5**. The final step from **5** to HMF employs the same step as that in the fructose path.

The pathway via 3-DG intermediate may follow an alternative mechanism proposed by Anet and Moreau (Figure 1, black dotted line).[18,19] The pathway proceeds via an initial tautomerization to intermediate **10** and then three sequential dehydration steps. However, their proposed final dehydration reaction cannot be proved by later theoretical studies.

Direct path[20] (Figure 1, red line): The pathway first undergoes two sequential dehydration steps to the enol intermediate (**12**) and then dienol (**13**), which continues to undergo simultaneous ring closure and dehydration to produce HMF.

To better resolve the pyrolysis reaction network, theoretical simulation has been considered as a new tool to complement experimental findings and even predict the reaction selectivity. It should ideally be able to fulfill the following key requirements for investigating glucose chemistry:

(i) Describe all possible bonding patterns of the three elements C, H, and O;
(ii) Distinguish sensitively different configurations with flexible hydrogen bonding (H-bonding) network; and
(iii) Explore the enormous reaction space to identify the lowest energy pathways.

While the state-of-the-art quantum mechanics (QM) methods can achieve a high accuracy in computing the electronic structure and thus satisfy the above (i) and (ii) requirements, they suffer from too high computational cost and in general fail in the fast reactivity prediction required in (iii).[28] The reaction sampling based on QM was typically limited to shallow levels in the reaction network hierarchy, despite many elegant methods that were developed to expansively interrogate complicated reaction paths.[29−33] Recently, Grambow et al. have achieved impressive progress on the reaction network of γ-ketohydroperoxide ($C_3O_3H_6$, i.e., six heavy-element atoms) by combining several automated reaction discovery methods based on density functional theory (DFT) calculations.[31] Alternatively, empirical force field calculations, although with a low computational cost, are often not reliable in reactivity prediction due to the intrinsic difficulty of simple analytic functions in describing different bond formation and dissociation curves.[34−36] The current dilemma thus calls for new methods to efficiently locate both minima (initial and final states, IS/FS) and transition states (TS) of chemical reactions.

The recent advance in machine learning techniques points to a promising direction for reactivity prediction, i.e., by constructing accurate and low-cost neural network (NN) PES.[37−44] By replacing the self-consistent Schrodinger equation solvation in QM, NN potential can be utilized as a powerful numerical solver to correlate the geometry of a structure to its total energy by learning the existing accurate PES data set, typically from first-principles DFT calculations. However, to predict organic reactions via fast and robust NN

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

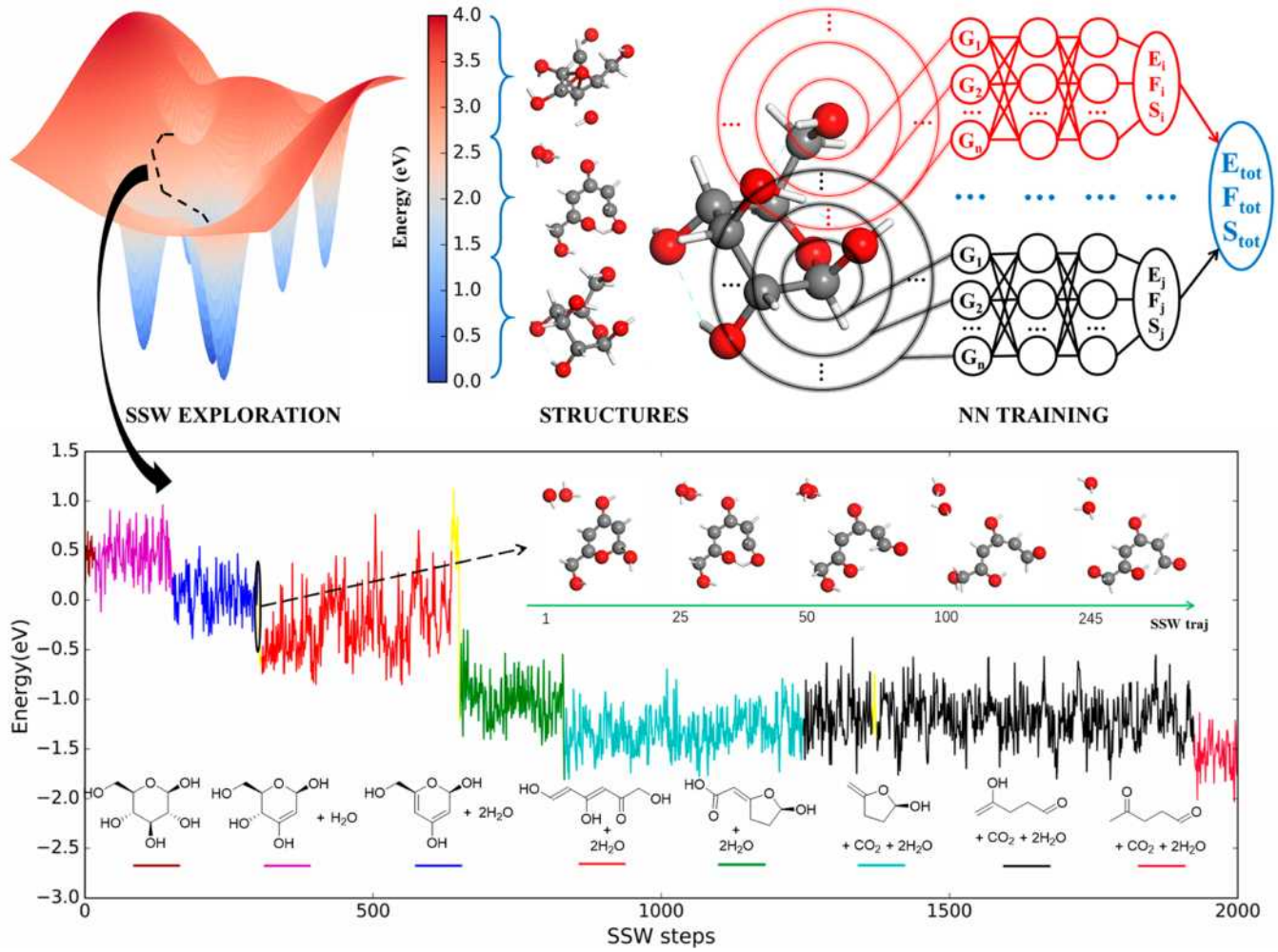


Figure 2. Scheme of SSW-NN method to build global NN potential for organic reactions. The top pannel illustrates the SSW for PES exploration to generate representative global PES data that are used by the atom-centered NN architecture for PES learning (also see SI Section S1 for details); The bottom pannel shows a typical 2000-step SSW trajectory where a β-D-glucose molecule evolves into different products. In the bottom pannel, the color of lines represents major intermediate molecules evolved in the trajectory (also see the plotted molecular structure), except that yellow lines represent various products that appear only occasionally. The 3D structural changes in one step SSW step between two specific minima are also highlighted in the inset as indicated by the dashed arrow. Energy zero is defined by the lowest energy conformation of β-D-glucopyranose. Gray balls: C; red balls: O; and white stick: H.

potential, there are at least two outstanding difficulties related to the PES of the organic chemistry. First, the configurational and reactive space of organic molecules is astronomical huge, involving many elements (at least C, H, O, N) and many different bonds with different bond orders (C−C, C═C, C−H, C−O, ···). Second, the transition region of organic reactions often has a high energy due to the bond making/breaking (e.g., >0.7 eV). As a result, the traditional PES sampling methods, such as normal mode sampling, molecular dynamics (MD), or even the enhanced MD techniques (e.g., metadynamics) meet with great difficulty in generating a proper data set to represent the vast reaction space.[42−44] The NN potential reported so far is either specific for a single reaction, e.g., the one generated by Gastegger et al. for Claisen reaction using metadynamics,[42] or valid largely for minimum structures, e.g., the ANN generated by Roitberg group[43] using the normal mode sampling based on a subset of GDB-11 data set[45,46] (GDB-X is a small organic molecule database with up to X atoms and up to C, N, O, and F four heavy elements per molecule).

With the advent of stochastic surface walking (SSW) global optimization in combination with neural network (NN) potential method, i.e., SSW-NN,[39,40,47] developed by us recently, here we manage to generate the first global NN (G-NN) potential for organic reactions and utilize it to resolve the reaction network in glucose pyrolysis. The lowest energy pathways to key products, 5-HMF, FF, and AA are discovered, which lead to resolution of the long-standing selectivity puzzles in the field.

This paper is organized as follows. In Section 2, we will briefly introduce the methodology utilized, including SSW-NN architecture, SSW reaction sampling (SSW-RS), and DFT calculations. In Section 3, we first validate the G-NN potential by analyzing the data set and then benchmark the reaction network of the glucose pyrolysis. In Section 4, we apply the G-NN PES to search for the lowest energy pathways of glucose pyrolysis. The key pathways are confirmed by high accuracy DFT calculations, and the results are discussed thoroughly in the context of experimental findings.

2. METHODS

2.1. SSW-NN Method. The SSW-NN method developed in the group as implemented in the LASP code[48] (accessible from www.

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

lasphub.com) is utilized to generate the PES data for fitting the G-NN potential. The method is based on a "global-to-global" idea,[34] utilizing the global optimization method to generate global PES that is utilized to iteratively train a global neural network potential. While SSW-NN has demonstrated success in a number of solid materials (e.g., TiO_2, B and ZnCrO),[39,40,49] its application in generating the reactive G-NN for organic molecules is the first attempt and involves much greater effort. We have designed a self-learning procedure for generating the reactive NN potential, which will be detailed in Section 3.1 in the context of data set analyses, and here we briefly overview the SSW method and G-NN technique, as schematically shown in Figure 2.

The SSW algorithm is a global optimization method that can explore both minima and saddle points on PES[50,51] without any predetermined guess. SSW implements an automated climbing mechanism to manipulate a structural configuration moving smoothly from a local minimum to a high-energy configuration along one random mode direction. The method was initially developed for aperiodic systems, such as molecules and clusters, and has been extended to periodic crystals.

Our G-NN potential[39,40,49] follows the atom-centered high dimensional neural network (HDNN) architecture,[37,38] as also schematically shown in Figure 2 (upper-right corner). The total energy E_{tot} is written as a linear combination of individual atomic energy E_i, which is the output of NN. The atomic force (F_i) and the static stress tensor (S_i) can then be analytically derived by relating the total energy to the coordinate. We utilize our recently proposed power type structural descriptor (PTSD)[40] to construct the input layer for NN, which was demonstrated to be sensitive for distinguishing the local chemical environment of the atom. The PTSDs include two-body functions, S1 and S2; three-body functions, S3, S4, and S5; and a four-body function, S6. In PTSD, the radial part of the functions are power functions, and the angular part contains either trigonal or spherical functions.[40]

2.2. SSW-RS Method for Pathway Search. With the G-NN potential, it becomes feasible to explore the whole reaction network even for a complex reaction process. The SSW reaction sampling (SSW-RS) developed previously[29] is extensively used in this work to identify reaction pathways in glucose pyrolysis. The SSW-RS method is an automated method derived upon SSW global optimization targeted to identify reaction pathways without the need of a priori information on the reaction. It can simultaneously sample the reactant space of conformations, often a flat PES with very low barriers, and the reaction pathways that could have high barriers leading to unexpected products. The method has been applied to many different reactions, from molecules[29] to solids.[52] More details on the SSW-RS simulation can be found in the Supporting Information (SI) and our previous work.[29] For completeness, we briefly introduce the procedure of the SSW-RS simulation below.

The SSW-RS simulation can explore the likely reaction pathways near a predefined reactant and identify the lowest energy pathway leaving it. The two stages in simulation are as follows: (i) Reaction sampling via extensive SSW global search, which collects the reaction pairs from the reactant; and (ii) Pathway building and TS determination via the double-ended surface walking (DESW).[53,54] By progressively changing the reactant of SSW-RS as guided by the determined low energy pathways, we can therefore explore the whole reaction network in an automated way (see SI Figure S2).

2.3. DFT Calculations. The data set for fitting NN potential is generated using first-principles periodic DFT calculations with plane wave basis set as implemented in VASP package.[55] The ionic core electrons are described using the projector augmented wave (PAW) pseudopotential.[56,57] The electron exchange and correlation effects are described by the GGA-PBE functional.[58] The kinetic energy cutoff for plane wave basis is 450 eV and the fully automatic Monkhorst–Pack k-mesh grid is generated with 25 times the reciprocal lattice vectors. This setup is a standard in building the LASP G-NN potential library across the Periodic Table.[48]

Because the GGA-PBE functional tends to underestimate the reaction barrier for molecular reactions, we have also utilized the B3LYP functional together with the 6-311++G(2d,p) basis set as implemented in the Gaussian-09 program[59] to refine the reaction kinetics for all lowest energy reaction channels obtained from SSW-NN simulation. These results have been compared with the previously published data and showed good consistency for the same reaction. We note that the current DFT calculations generally have intrinsic errors in predicting the reaction barrier and thus the kinetics.[60] Here by focusing on the relative barrier difference between different pathways and extensively comparing our data with known experimental facts, we aimed to provide better insights into glucose chemistry with DFT-based reaction network.

3. RESULTS

3.1. Global NN Potential for Organic Reactions. While the total numbers of organic molecules and their reactions are in principle infinite, the reaction pattern as a local property can be numerated. For example, there are already 12 million single-step reactions recorded in the Reaxys chemistry database, but the number of common reaction patterns (appearing more than 50 times) is limited to be only 17 134.[61] The key for building a reactive organic NN potential is therefore to collect the reaction patterns via efficient reaction space exploration.

In this work, we design an iterative self-learning procedure of SSW-NN for generating the organic reaction data set (see SI Section S1 for details). In brief, there are three key steps in expanding the reaction database. (i) The short-time SSW sampling of molecular crystal systems (in variable periodic cells) are performed using DFT calculations, which provide the most common bonding pattern of organic molecules with the C−H−O−N four elements. The starting molecular structures for these SSW-DFT simulations are randomly selected from the QM9 database (QM9 is a standard organic molecule database with no more than 9 heavy atoms generated from the GDB-17 data set by DFT calculations[62,63]); (ii) The SSW-NN global sampling of organic molecules and molecular crystals are performed iteratively to expand the reactive PES data set. The starting structures are also randomly selected from the QM-9 database; (iii) The SSW-NN global sampling for glucose pyrolysis reaction networks are used to further improve the transferability of the G-NN potential for glucose chemistry. The starting structure are updated and randomly selected from glucose global optimization SSW trajectories. It should be mentioned that both (ii) and (iii) steps are in an iterative self-learning manner, where NN potential is iteratively trained by incorporating a new DFT data set screened from the SSW-NN global optimization. Both step (ii) and step (iii) are essential for robustness and accuracy in describing the glucose reaction PES using NN.

To illustrate how the reaction space is sampled by SSW-NN and how the NN potential is improved iteratively, e.g., in step (iii) above, we show a typical SSW trajectory starting from β-D-glucopyranose in Figure 2. Within a 2000 step SSW sampling, 31 different molecules are encountered, which can be divided into 8 main minimum domains (marked by different color lines) and many other minority minimum structures (marked by yellow lines). In each minimum domain the diversity of conformations is also evident from the large oscillation in the energy scale. The entire trajectory gradually changes from a six-membered ring to dehydration products and finally to short-chain molecules. In the meantime, different functional groups emerge, e.g., the common alcohol, ether, alkenyl, and aldehyde groups, together with some exotic structures (e.g., with uncommon coordination). The appearance of exotic structures in SSW-NN trajectory reflects the wrong prediction of the NN potential because these structures are rare and not included in

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

the existing training set. Consequently, these structures should be added to the training set for further learning (we list our guidelines on how to select these structures in SI Section S1).

As a representative, the inset indicated by the dashed arrow in Figure 2 illustrates the structure snapshots in one step SSW, which transform 6-(hydroxymethyl)-2H-pyran-2,4-diol (blue line) to 3,5,6-trihydroxyhexa-2,4-dienal (minority minimum, yellow line), which is the enol-keto tautomeric precursor of 1,4,6-trihydroxyhexa-3,5-dien-2-one (red line). With such a process, the reaction space with the simultaneous O−H bond formation and C−O bond rupture is captured by SSW sampling and subsequently learned by the G-NN potential.

By using the SSW-NN method, we finally obtained the organic reaction data set with 94 854 structures in total. Our data set covers almost all likely local patterns (78) that obey the octet rules of C atom (in total 79 for C with the chemical environment of the C−H−O−N four elements), which is even larger than that in QM9 data set (63) (these local patterns are detailed in SI Section S2). The larger bonding pattern diversity in our data set demonstrates the power of SSW global search in exploring the chemical space. In addition, we also identify 410 bonding patterns that do not obey the octet rule. Not limited to low energy saddle point structures, they also include both too highly coordinated carbon atoms (e.g., up to six first neighbors and up to nine bond orders) and too low coordinated atoms (e.g., radials and single atoms). These minority structural patterns are in fact critical for G-NN to define the boundary of global PES, which eventually allows the reactivity prediction from first-principles (no predetermined guess).

In order to train a high-quality G-NN potential using the vast amount of global PES data, we have adopted a large set of PTSDs, i.e., 407 structural descriptors in total for each element, including 148 two-body, 229 three-body, and 30 four-body descriptors. In addition, the network utilized is also large with 120−80−80 three-hidden layers, equivalent to 65 201 network parameters per element and 260 804 in total (the guidelines to choose a suitable NN architecture can be found in our previous work).[40] Both the large number of structural descriptors and the large NN parameters are found to be essential to achieve high accuracy for predicting organic reactions in general: the final NN PES can achieve the RMS values for energy and force with 10.05 meV/atom and 0.242 eV/Å, respectively. The overall accuracy is satisfactory considering that the energy of the structures in our global data set spans in a large window from 1.5 to 6 eV/atom. We also demonstrate the G-NN performance on the geometry and reactions for common organic molecules and reactions that are compared with the DFT results in SI Section S3. We will further our benchmark results for glucose pyrolysis later in Section 3.2.

It should be mentioned that the PTSDs utilized in constructing the G-NN potential are selected to maximally distinguish structures. To visualize the data set and illustrate the performance of PTSD, we have performed principle component analysis (PCA) on the data set. We randomly selected 6000 carbon atoms from data set, which belong to six typical functional groups in the first neighbor, i.e., s-alcohol, alkene, s-alkane, amine, *p*-alkane, and *p*-alcohol. PCA was then utilized to identify the principal component projections of the 407 PTSDs of our G-NN potential that describe the geometrical environment of the carbon atoms. Figure 3 plots the 6000 data points projected on the first three components obtained from PCA, which rank as the most valued

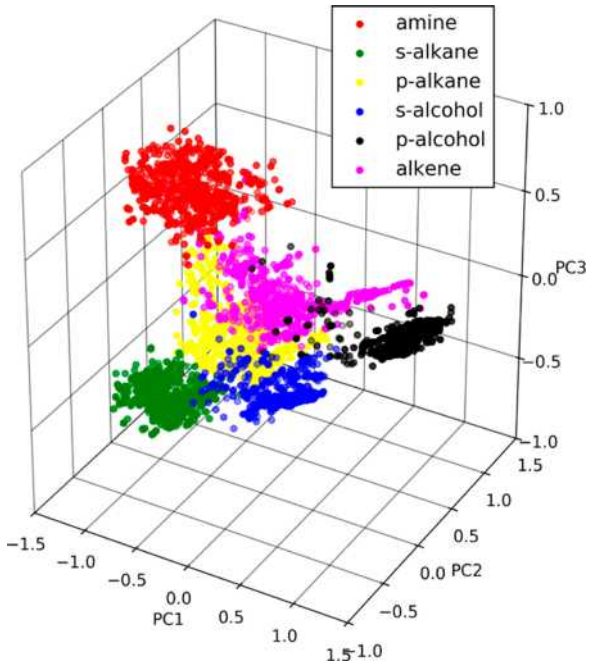


Figure 3. Principle component analysis plot of 6000 carbon atoms with six different functional groups that are randomly selected from the global training data set. The data are projected onto the first three PCs (PC1−PC3) of PTSDs.

components to distinguish these carbon atoms. It shows that there is an obvious clustering of data points, exhibiting different domains of functional group as colored by different zones, which implies that the first neighboring environment for the six functional groups can be distinguished nicely by our PTSDs. In the meantime, the data points belonging to the same functional group remain to be scattered inside each domain, indicating the long-range minority structural differences can also be described well by PTSDs.

While our SSW-NN simulation has covered most local molecular patterns for C−H−O−N elements, there is of course no guarantee that the reactions associated with all these patterns can be predicted correctly using the current G-NN potential due to the enormous chemical space of organic reactions. We emphasize that the G-NN potential can always be retrained to incorporate more reaction data that are relevant for any particular purpose, and thus can grow its predictivity with time and usage.

3.2. Glucose Pyrolysis Pathway Sampling via SSW-NN and Benchmark with DFT. With the G-NN potential, it is now feasible to extensively explore the reaction space of glucose. By using the SSW-RS method to automatically sample the likely pathways, we have iteratively searched the reaction tree starting from glucose, where the simulation records the elementary reactions, i.e., IS, TS, and FS, to establish a reaction database. Specifically, in each cycle of SSW-RS, one intermediate is selected as the starting molecule, and the simulation then samples the possible reaction route leaving from this molecule and collects all reactant/product (R/P) pairs. After the pathway search, all recorded R/P pairs are connected and verified to determine the reaction barrier. After removing the duplicate elementary reactions, these new reactions are added to the reaction database. From the updated reaction network of glucose pyrolysis, a new molecule is selected as the next starting molecule for the SSW-RS in the

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

next cycle. A threshold overall barrier, i.e. 3.0 eV, is utilized to select the molecule that is not sampled previously by SSW-RS, i.e., the overall barrier to reach this molecule from glucose should be less than 3.0 eV (obviously, a too high barrier is not necessary since pyrolysis occurs below 550 °C).

Benefiting from the low cost of G-NN PES, we can now achieve a deep exploration of the reaction tree starting from D-glucose. In this work, we have managed to sample 1 200 000 minimum and collected more than 150 000 reaction pairs. With this reaction data, we are able to establish a reaction database where each elementary step is indexed uniquely by using the structural fingerprint of reactant and product and the reaction barriers separating them. After removing duplicate reactions and recording only the lowest barrier connection between pairs, the final reaction database contains 4455 unique molecules, and 6407 different reactions with 3488 reaction patterns.

Before we searched the lowest energy pathways using the reaction database, we have examined the accuracy of G-NN PES. We randomly selected 2000 points from a SSW trajectory starting from D-glucose and benchmarked them with DFT results. These data are not only minimum structures, but also mostly nonequilibrium and high energy structures from the SSW trajectory. As shown in Figure 4, we found that the

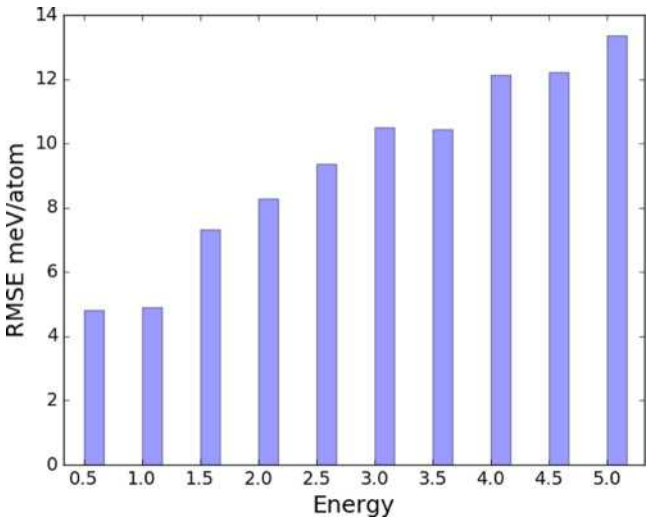


Figure 4. Energy-resolved RMSEs of energy for the G-NN PES performance on 2000 structures randomly selected from the SSW global search trajectory starting from glucose. The *x*-axis is the energy window of structures relative to the lowest energy minimum in the trajectory and the *y*-axis represent the RMSE of energy between G-NN and DFT calculations.

RMSE are 9.90 meV/atom for all data points and 7.20 meV/atom for those below 3 eV. It suggests that the error bar for the computed barrier for low energy reactions from G-NN PES should be less than 0.3 eV (24 atom of glucose), which provides a quantitative guideline to select the low energy pathways from the reaction database.

3.3. Mechanisms and Selectivity of Glucose Pyrolysis.

Now we are in a position to present our results on the lowest energy pathways of glucose pyrolysis. By estimating the reaction rate based on transition state theory (assuming the typical pre-exponential factor 10^{13} and the reaction temperature 500 °C) for each elementary reaction of the reaction database, we have identified the low energy pathways from glucose to different reaction products on the G-NN PES. To be accurate, we then selected 2060 reaction pairs (IS, TS, and FS) from 500 lowest energy pathways that are associated with the glucose to the three key products, i.e., HMF, FF, and AA, and refined their energetics using DFT where all minimum structures are fully relaxed and the TSs are researched. (A comparison between NN and DFT is detailed in SI Figure S2). All pathway results reported below are from DFT energetics (B3LYP with 6-311++G(d,p) basis sets, see also Section 2), which can be compared directly with previous calculations by other groups.[22]

To help us elaborate the key pathways, we first provide an overview on the basic reaction patterns in Figure 5. There are six reaction patterns frequently encountered for reactions in glucose pyrolysis, namely, cyclization, isomerization, retro-aldol, tautomerization, retro-Michael-addition, and β-H elimination as ordered by their reaction barrier from low to high. They are explained below.

(i) Cyclization (CR): The cyclization in glucose pyrolysis is mainly associated with the transformation of monosaccharide from aldehyde/ketone to inner hemiacetal[29] via ring formation by a hydroxyl group attacking the neighboring carbonyl group.

(ii) Isomerization (IM): This refers to the proton exchange between a carbonyl and the adjacent hydroxyl groups and the simultaneous H exchange between these two adjacent C atoms.

(iii) Retro-aldol (RA): The retro-aldol reaction that splits a monosaccharide into the aldehyde and enediol fragments occurs when the C−C bond between the α-C and the β-C of a carbonyl group breaks.

(iv) Tautomerization (TA): The switch between keto and enol configurations can occur via the H exchange in between a carbonyl group and its neighboring α-C.

(v) Retro-Michael-addition (RM): This refers to the dehydration reaction by combining a proton of enol hydroxyl and a neighboring β-OH group to generate the α,β-unsaturated carbonyl compound.

(vi) β-H elimination: The β-H elimination (1,2-dehydration) involves the simultaneous α-OH removal and the C_β−H bond cleavage.

The CR reactions often have the lowest energy barrier, ∼1.7 eV with respect to D-glucose provided with the optimal H-bonding network.[29] Then it follows the IM reaction, which has a barrier slightly higher or similar to that of CR reactions. An IM reaction can occur directly in between two neighboring groups, but is also likely between two non-neighboring groups that are spatially close due to the folding of carbon chain. The RA reactions, despite having relatively low barrier (∼0.1 eV higher than that of IM), are usually highly endothermic, e.g., more than 0.5 eV. Next, the TA reactions are difficult in the gas phase, usually with a barrier greater than 2.5 eV. The reaction can however be promoted if the nearby H-bonding, e.g., hydroxyl groups from other molecules, is present, which could assist the proton transfer by stabilizing the TS. The reaction barrier can thus be reduced to ∼2.0 eV in the presence of an appropriate H-bonding network. The product of TA reactions (enol) can often undergo a facile RM reaction to drop out one H_2O molecule in one elementary step with a barrier as low as 1.5 eV. The overall barrier with respect to the keto reactant of TA reaction is ∼1.9 eV. In addition to the dehydration route by RM reactions, it is also likely to remove

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

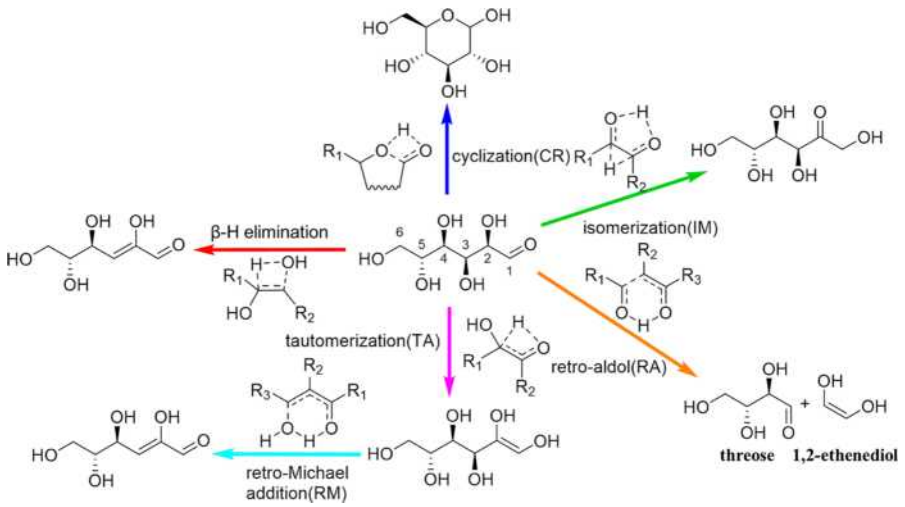


Figure 5. Major low energy reaction patterns identified from the glucose pyrolysis reaction network.

H_2O molecule via the β-H elimination, which is however highly difficult with the barrier ~2.4 eV. Even with the help of H-bondings, e.g., via other nearby hydroxyl groups, the barrier for β-H elimination remains above 2.1 eV.

We have carefully analyzed the pathways in our reaction database to identify the pathways to different products mentioned in the experiment, e.g., LG, HAA, different furan products, and formic acid. We found that the LG, HMF, FF, and HAA, the major products in the experiment, are the products with the lowest energy pathways among the known products. In the following we describe the lowest energy pathways to three key products, HMF, FF, and AA, in detail (the path to LG is straightforward and is left in the SI, together with the pathway to the other products, i.e., 2-furanmethanol and formic acid). As mentioned in the Introduction, the reactions of D-glucose have been discussed in many previous literatures. For example, the major reaction route to HMF, the most concerned product, was suggested to follow the fructose path (see Figure 1). From our reaction database, this fructose path has a high overall barrier, 2.1 eV, occurring at the β-H elimination step (e.g., **5** → **6** in Figure 1), which are the highest barrier reaction pattern in our reaction database (Figure 5). In addition, the β-H elimination reaction has no site selectivity and can occur between any adjacent −COH and −CH. This is obviously contradictory to the observed major selectivity to HMF. Instead, our reaction database provides lower-barrier alternative routes to generate HMF, HAA, and FF, as shown in Figure 6, where the molecular intermediates are also labeled by numbers (**1**, **2**, ...) following those in Figure 1. Importantly, the two value-added furan products, HMF and FF, share the same key reaction pattern in the dehydration step, namely, a retro-Michael addition step and no β-H elimination reaction is involved. We now elaborate them below.

3.3.1. HMF Pathway. β-D-Glucopyranose first undergoes the ring opening and isomerization to D-fructose (**14** → **1** → **2**). D-Fructose then transforms to an enediol structure (**2** → **10**) with the help of an internal H-bonding network (also see Figure 7). For the enediol (**10**) configuration, the subsequent 1,3-dehydration (RM reaction) is facile which generates an α,β-unsaturated carbonyl compound, 3-DGE (**7**), which is a reverse reaction of 1,4-michael addition. The dehydration product **7** can then go through a similar 1,3-dehydration to intermediate **11**, which can isomerize to **15** with an H atom transferring from C5 to the carbonyl oxygen on C2. The HMF can finally be produced by following the cyclization and 1,5-dehydration (**15** → **16** → **6**). The 1,5-dehydration is also an RM type reaction, similar to the previous 1,3-dehydration. Overall, the rate-determining step belongs to the enol-keto TA reaction (**2** → **10**), with a barrier of 1.91 eV (with respect to the most stable configuration of β-D-glucopyranose hereafter), which is 0.19 eV lower than the previous pathways (2.10 eV in β-H elimination). Although such enol-keto TA reactions are higher in barrier than the CR, IM, and RA reactions, the mechanism is overall favored due to the opening of the RM route in the subsequent dehydration reactions and the avoidance of the direct β-H elimination. It should be mentioned that D-glucose can also tautomerize to 3-DGE (**7**) via **1** → **10** directly with a slightly higher overall barrier (by 0.08 eV).

3.3.2. FF and HAA Pathway. These two common products share the same the first step, that is, the carbonyl group at the C1 position (as labeled in glucose molecule of Figure 5) of D-glucose (**1**) isomerizes to the C4 position (**17**,3-ketohexose) (**1** → **17**). This is also the rate-determining step for both pathways with an overall barrier of 1.95 eV (Figure 6B). **17** can then produce HAA directly after an RA reaction (**17** → **18**). After HAA is produced, the enediol molecule (but-1-ene-1,2,3,4-tetraol) may further undergo 1,3-dehydration to produce enal (2,4-dihydroxybut-2-enal) that is thermodynamically more stable (**18** → **19**).

The pathway to FF bifurcates after the intermediate **17**, from which a formaldehyde is removed to generate **20** through a retro-aldol reaction. After that, the pathway is similar to the HMF pathway from the intermediate **10**: two consecutive 1,3-dehydration can happen facilely first from the enediol product (pent-1-ene-1,2,3,4,5-pentaol) of **20** (**20** → **21** → **22**), which are followed by the isomerization, cyclization, and dehydration reactions (**22** → **23** → **24** → **25**). Since the rate-determining step to FF and to HAA is the same, these two products are in competition. It is expected that at high temperatures the HAA route with more fragments produced are preferred due to the increased entropy gain.

We would like to further emphasize the critical role of the H-bonding network in the β-D-glucose pyrolysis. While this effect has been shown previously,[29] here with the large reaction

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

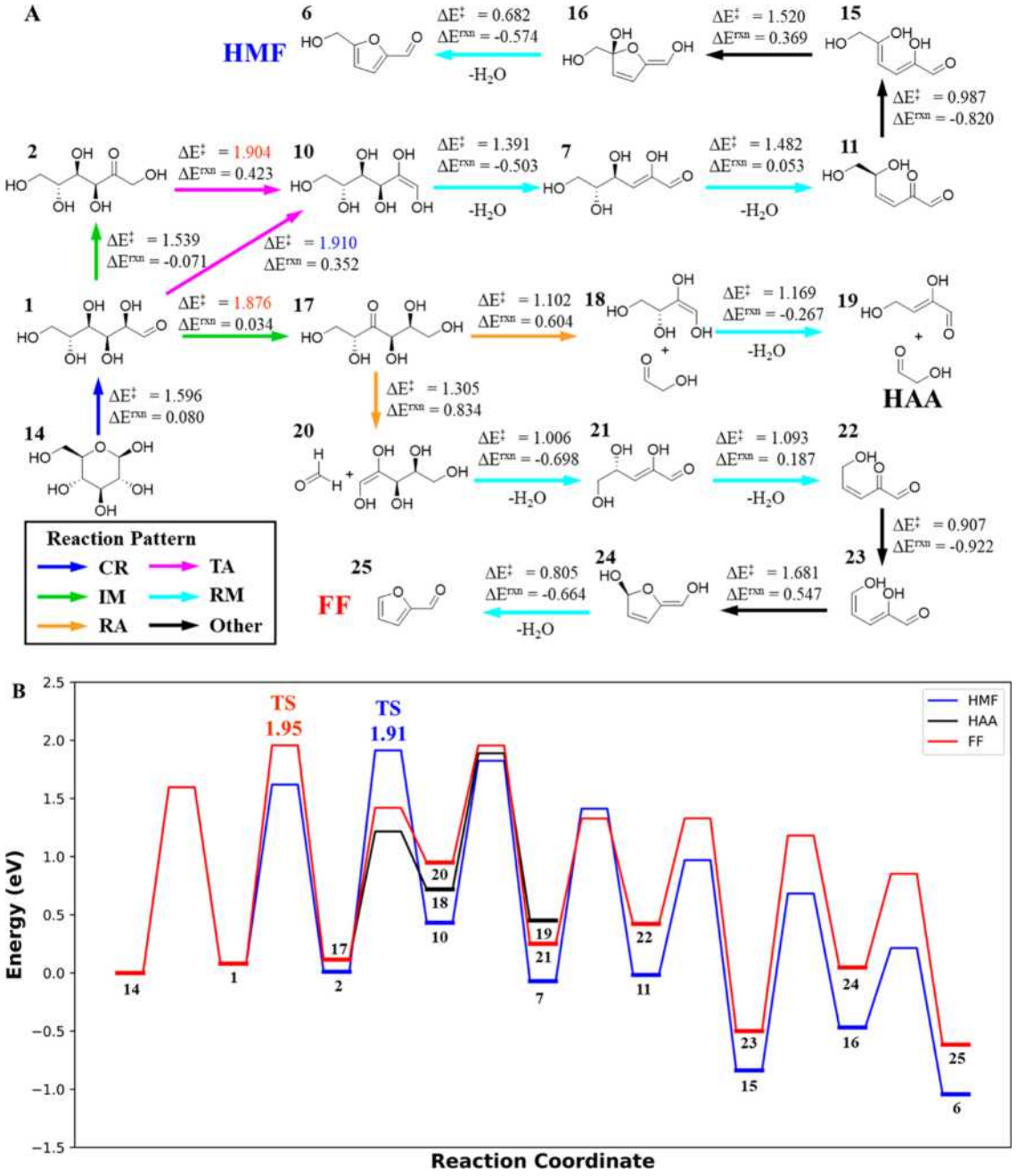


Figure 6. (A) The lowest energy pathways to three major products, HMF, FF, and AA, identified from the glucose reaction network; and (B) the energy (zero-point-energy corrected) profile for the lowest energy pathways to HMF, HAA, and FF. The energy zero refers to the most stable configuration of β-D-glucopyranose. In (A), $\Delta E^{\ddagger}$ and ΔE^{rxn} correspond to the barrier and the reaction energy of the elementary reaction, respectively.

database we can better quantify the promotional effect of the H-bonding network to each class of reactions. As glucose contains multiple hydroxyl groups, nearly all reactions are found to be affected by the nearby H-bonding. For example, IM reactions often occur in between two neighboring C, but becomes likely in between non-neighboring C under specific spatial configurations with the appropriate H-bonding network, such as **1** → **17** (Figure 6) where the molecule folds to allow close contact in between C1 and C4. In addition, the CR and TA reactions need the H-bonding network to transfer proton, which is critical to decrease the contact distance between the transferring H and the accepting O. We found that the barrier reduction can be as large as ∼1 eV. As illustrated in Figure 7, in the rate-determining step of the HMF pathway, **2** → **10**, the H-bonding network as identified by SSW-NN significantly facilitates the TA reaction that involves the removal of the H atom of C1 and the hydrogenation of the carbonyl group at C2. The hydroxyl groups at C6 and C5 and the carbonyl group at C2 all participate in the H atom transfer, forming a sequential H-bonding chain. In this way, the highly twisted configuration via direct H transfer from C1 to the carbonyl at C2 can be avoided, which reduces the barrier by ∼1 eV!

4. DISCUSSIONS

It is of significance to discuss our new mechanism in the context of previous experimental findings. As for the likely

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

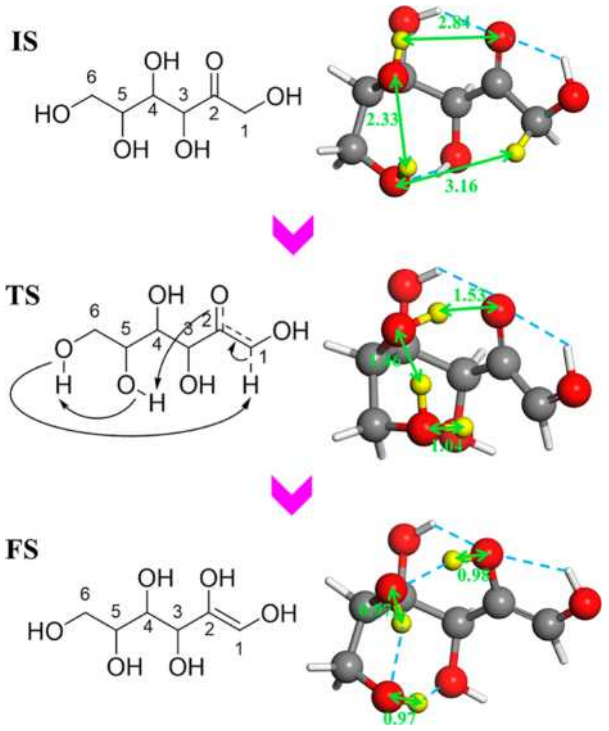


Figure 7. Reaction snapshots of the rate-limiting step in HMF path, **2** → **10** (see also Figure 6), which illustrates the critical catalytic role of the H-bonding network. Gray balls: C; red balls: O; white stick: H; and yellow balls: the reacting H and the H atoms involved in the H-bonding chain.

intermediates, fructose has been detected by the experiments of Ponder and Richard in glucose pyrolysis tar.[64] Paine et al. further confirmed that the pyrolysis from fructose results in a higher yield of HMF (~4 times) than the pyrolysis from glucose, as would be expected if fructose is the key intermediate.[65] On the basis of this, the fructose pathway to HMF was proposed in the literature (Figure 1).[23,25,65] In addition, the 3-DG intermediate, captured as its (2,4-dinitrophenyl)-osazone derivative, was confirmed in the acid-catalyzed formation of HMF from fructose by Anet,[1,18] which inspired the 3-DG path (Figure 1). These two pathways however are not consistent with each other on the key intermediates. Nevertheless, they do share a key reaction pattern, i.e., β-H elimination in dehydration (**3** → **4**, **9** → **5**, **5** → **6**, **1** → **12**), which turns out to be the rate-limiting step for both pathways from recent DFT calculations.[20,21,23,25] Since β-H elimination has no obvious selectivity on removing the neighboring H of CH, it is puzzling that HMF is a major product in glucose pyrolysis.

Obviously, our new mechanism for 5-HMF does confirm both fructose and 3-DGE (enol form of 3-DG) as intermediates, but does not require β-H elimination. Instead, the 1,3-dehydration or 1,5-dehydration by RM reaction to produce α,β-unsaturated carbonyl compounds occurs, which require only O−H bond cleavage in all three dehydration steps (**10** → **7**, **7** → **11**, and **16** → **6**). Such dehydration reactions can greatly reduce the reaction barrier from our data and are consistent with the fact that 1,4-Michael addition for α,β-unsaturated carbonyl compounds is a well-known reaction in organic chemistry. In fact, we note that the difficulty in β-H elimination has been recognized by Mayer et al, who proposed that 1,3-dehydration reactions may occur in the second dehydration step with a lower barrier in the fructose pathway (Figure 1, **4** → **5**), although the β-H elimination remains in the first and third dehydrations of the fructose pathway.[23]

In summary, our new mechanism for 5-HMF production supports the known experimental findings. First, while fructose is believed to be a key intermediate to HMF,[65] the other channels without fructose participation appear to also be likely. The direct dehydration from glucose to HMF appears to be important for 5-HMF production in the experiment with VCl_3, $GaCl_3$, and $InCl_3$ catalysts that exhibit low rates of fructose formation.[9] In our pathways, the TA reaction from fructose to enediol (**2** → **10**) is 0.08 eV lower than that directly from D-glucose (**1** → **10**). This small energy difference is expected to vary under different catalytic conditions, leading to the switch of the pathways with or without fructose. Second, HMF formation from 3-DG proceeds at a significantly higher rate than that from fructose.[21] In our mechanism, 3-DG is a keto isomer of 3-DGE, which obviously has a shorter and lower barrier path to HMF compared to that from fructose. Third, *in situ* ^{13}C NMR showed that C1 of fructose corresponds to the carbonyl carbon of HMF, and the C6 is the hydroxymethyl carbon of HMF.[66] This atom-to-atom correspondence also agrees with the prediction from our pathway. Last but not least, our results show that the HMF product has a lower barrier (1.91 eV) compared to that of FF and HAA products (1.95 eV). This suggests that HMF is more favorable at low temperatures, which is indeed observed in experiment (Table 1): HMF selectivity can drop from 20% in 300 °C to 12% in 500 °C.

For the HAA product, the isotopic labeling experiment has shown that the main source for HAA comes from C1−C2 (39.5% in 500 °C) of glucose, and those from C3−C4 or C5−C6 are lesser (18.8% and 34.1% in 500 °C). This distribution varies at different reaction temperatures.[24] At the low temperature (350 °C), the C1−C2 ratio is much higher (58.5%), which suggests that the C1−C2 pathway is kinetically favored. Two explanations may be likely for the C1−C2 product selectivity. (i) The HAA is produced from HOCH═CHOH (1,2-ethenediol) via the direct RA of D-glucose (see Figure 5 RA product) as proposed by Lu et al.[24] (ii). The HAA is produced from the decomposition of **17** (carbonyl on C4, **17** → **18**) as shown in our mechanism (see Figure 6A). The explanation (i) must involve the TA reaction of 1,2-ethenediol to HAA, which has a very high barrier (>2.40 eV) from the previous computational results (also confirmed by us).[24] However, the explanation (ii) has a much lower overall barrier of 1.95 eV (see Figure 6B). It is interesting that none of the previous work has considered **17** as an intermediate to HAA, possibly due to the fact that the 1,4-shift of the IM reaction (**1** → **17**) identified by SSW sampling is not intuitive. We emphasize that our mechanism to HAA with the C1−C2 selectivity has the lowest barrier according to our reaction database. The other routes to HAA are kinetically more difficult with a barrier of at least 2.19 eV, e.g., from the RA product of threose[24] with C5−C6 selectivity (see Figure 5).

For FF product, the isotopic labeling experiment shows that more than 90% FF keeps C1−5 with C6 of glucose being removed.[65] On the basis of this experimental fact, different mechanisms with the removal of the C6 were proposed, but all have relatively high reaction barriers (>2.43 eV),[25] being contradictory with the major yields of FF and HMF in experiment (see Table 1). We show that the RA reaction from intermediate **17** can readily drop out one formaldehyde

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

molecule from C6 (**17** → **20**), and the remaining enol product can lead to FF after dehydration and cyclization reactions, similar to those in the HMF path. The whole path has a similar barrier (1.95 eV) as our pathway to HMF (1.91 eV).

Finally, our mechanism not only explains the observed glucose pyrolysis phenomena, but also provides important insights into to the catalytic glucose conversion. The two key reaction patterns in our pathway, the TA and the RM reaction, are known to be facile in acid or metal chloride catalysis from organic chemistry. This indicates that (i) once the glucose transforms to the chain configuration, it can follow the low barrier pathway, **1** → **10** → 7... → **6** without passing fructose; and (ii) if fructose is the starting reagent, then the pathway **2** → **10** → 7... → **6** is similarly promoted by catalysts. This is consistent with the observed high conversion rate and selectivity of the HMF product with and without fructose in catalytic experiments at low temperatures: for example, Zhao et al. reported 70% of HMF yield from glucose at 100 °C catalyzed by $CrCl_2$; Dumesic's group reported 80% HMF selectivity at 90% fructose conversion at 180 °C in acid catalytic conditions.[2,3,9] Under these conditions, fructose can be in fast equilibrium with the chain form of glucose via the rapid TA interconversion, **1**↔**10**↔**2**. Compared to the RM reaction, the β-H elimination for dehydration requires strong acids, which is not often utilized in experiment.

5. CONCLUSIONS

For understanding and predicting biomass chemistry, this work designs a systematic procedure to generate a G-NN potential for organic reactions that can explore the vast reaction space with up to C−H−O−N four elements. By fitting 94 854 global PES structures obtained from SSW-NN global sampling, our final NN PES can describe nearly all functional groups (78) that satisfy the octet rule, together with 410 structural patterns that break the octet rule. A reaction library for glucose chemistry is thus established by SSW-NN reaction sampling, allowing the classification of the basic reaction types and the clarification for the whole reaction network. This finally leads to a settling down of the long-standing selectivity puzzles of glucose pyrolysis. Our findings on the glucose pyrolysis are outlined below.

(i) Our HMF pathway involves both fructose and 3-DGE intermediates and a site-selective dehydration pattern via retro-Michael-addition mechanism. The direct dehydration via the β-H elimination mechanism proposed in previous literature is excluded due to the higher barrier from theory and the inconsistencies in the site-selectivity with known experimental findings.

(ii) Our HAA and FF pathways share the same initial IM step, followed by different RA reactions to produce HAA and FF, respectively. The mechanism prefers the HAA production from the C1−C2 of glucose, and the C6 removal in FF production.

(iii) The pathways to HMF, HAA, and FF have similar overall reaction barriers, i.e., 1.91 eV for HMF and 1.95 eV for HAA and FF, and can be significantly promoted under acidic conditions.

While the glucose chemistry is focused on here, the theoretical framework, i.e., the machine learning techniques in combination with SSW global optimization for generating reactive NN potential and creating reaction database, is rigorous and readily applicable to other reaction systems in general. For example, with the current G-NN potential, other organic reactions can be first explored facilely using SSW-NN, and the G-NN potential can then be improved iteratively by retraining the new reaction data. This new way of reaction exploration should significantly speed up our search of complex chemical reactions. Of further interest, SSW-NN also enables the extensive reaction sampling in heterogeneous catalysis,[47,52] and the generation of G-NN potentials for molecules on surfaces are now in progress (see the LASP Web site, www.lasphub.com).We thus believe that the reaction prediction for complex catalytic reactions is within reach with modern computing facilities.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.9b11535.

SSW-NN method for data set generation and NN potential training; data set analysis according to bonding patterns; benchmark for common molecules and reactions; automated SSW-RS to resolve the reaction network; benchmark for glucose pyrolysis reaction network; reaction pathways to other products; and atomic coordinates for HMF, HAA, and FF pathways in Figure 6 (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*cshang@fudan.edu.cn

*zpliu@fudan.edu.cn

ORCID

Cheng Shang: 0000-0001-7486-1514

Zhi-Pan Liu: 0000-0002-2906-5217

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2018YFA0208600) and the National Science Foundation of China (21533001 and 91745201).

■ REFERENCES

(1) Anet, E. 3-DEOXYHEXOSONES. *J. Am. Chem. Soc.* **1960**, *82* (6), 1502−1502.

(2) Román-Leshkov, Y.; Chheda, J. N.; Dumesic, J. A. Phase modifiers promote efficient production of hydroxymethylfurfural from fructose. *Science* **2006**, *312* (5782), 1933−1937.

(3) Zhao, H.; Holladay, J. E.; Brown, H.; Zhang, Z. C. Metal chlorides in ionic liquid solvents convert sugars to 5-hydroxymethylfurfural. *Science* **2007**, *316* (5831), 1597−1600.

(4) Corma, A.; Iborra, S.; Velty, A. Chemical routes for the transformation of biomass into chemicals. *Chem. Rev.* **2007**, *107* (6), 2411−2502.

(5) Binder, J. B.; Raines, R. T. Simple chemical transformation of lignocellulosic biomass into furans for fuels and chemicals. *J. Am. Chem. Soc.* **2009**, *131* (5), 1979−1985.

(6) 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* (2), 323−330.

(7) Pidko, E. A.; Degirmenci, V.; van Santen, R. A.; Hensen, E. J. Glucose activation by transient Cr2+ dimers. *Angew. Chem., Int. Ed.* **2010**, *49* (14), 2530−4.

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

(8) 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* (12), 9808.
(9) Pagán-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* (6), 930−934.
(10) Wang, J.; Ren, J.; Liu, X.; Xi, J.; Xia, Q.; Zu, Y.; Lu, G.; Wang, Y. Direct conversion of carbohydrates to 5-hydroxymethylfurfural using Sn-Mont catalyst. *Green Chem.* **2012**, *14* (9), 2506.
(11) van Putten, R. J.; van der Waal, J. C.; de Jong, E.; Rasrendra, C. B.; Heeres, H. J.; de Vries, J. G. Hydroxymethylfurfural, a versatile platform chemical made from renewable resources. *Chem. Rev.* **2013**, *113* (3), 1499−597.
(12) Teong, S. P.; Yi, G.; Zhang, Y. Hydroxymethylfurfural production from bioresources: past, present and future. *Green Chem.* **2014**, *16* (4), 2015.
(13) Mariscal, R.; Maireles-Torres, P.; Ojeda, M.; Sádaba, I.; López Granados, M. Furfural: a renewable and versatile platform molecule for the synthesis of chemicals and fuels. *Energy Environ. Sci.* **2016**, *9* (4), 1144−1189.
(14) Kreissl, H. T.; Li, M. M. J.; Peng, Y. K.; Nakagawa, K.; Hooper, T. J. N.; Hanna, J. V.; Shepherd, A.; Wu, T. S.; Soo, Y. L.; Tsang, S. C. E. Structural Studies of Bulk to Nanosize Niobium Oxides with Correlation to Their Acidity. *J. Am. Chem. Soc.* **2017**, *139* (36), 12670−12680.
(15) 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* (22), 11023−11117.
(16) Jing, Y.; Guo, Y.; Xia, Q.; Liu, X.; Wang, Y. Catalytic Production of Value-Added Chemicals and Liquid Fuels from Lignocellulosic Biomass. *Chem.* **2019**, *5* (10), 2520−2546.
(17) Toftgaard Pedersen, A.; Ringborg, R.; Grotkjær, T.; Pedersen, S.; Woodley, J. M. Synthesis of 5-hydroxymethylfurfural (HMF) by acid catalyzed dehydration of glucose−fructose mixtures. *Chem. Eng. J.* **2015**, *273*, 455−464.
(18) Anet, E. F. L. J. 3-Deoxyglycosuloses (3-Deoxyglycosones) and the Degradation of Carbohydrates. *Adv. Carbohydr. Chem.* **1964**, *19*, 181−218.
(19) Moreau, C.; Durand, R.; Razigade, S.; Duhamet, J.; Faugeras, P.; Rivalier, P.; Ros, P.; Avignon, G. Dehydration of fructose to 5-hydroxymethylfurfural over H-mordenites. *Appl. Catal., A* **1996**, *145* (1−2), 211−224.
(20) Shen, D. K.; Gu, S. The mechanism for thermal decomposition of cellulose and its main products. *Bioresour. Technol.* **2009**, *100* (24), 6496−504.
(21) Jadhav, H.; Pedersen, C. M.; Solling, T.; Bols, M. 3-Deoxyglucosone is an intermediate in the formation of furfurals from D-glucose. *ChemSusChem* **2011**, *4* (8), 1049−51.
(22) Seshadri, V.; Westmoreland, P. R. Concerted reactions and mechanism of glucose pyrolysis and implications for cellulose kinetics. *J. Phys. Chem. A* **2012**, *116* (49), 11997−2013.
(23) 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* (6), 1461−1473.
(24) 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.
(25) 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* (2), 486−501.
(26) 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-(2H)-Furanone. *Energy Fuels* **2018**, *32* (9), 9519−9529.
(27) Easton, M. W.; Nash, J. J.; Kenttamaa, H. I. Dehydration Pathways for Glucose and Cellobiose During Fast Pyrolysis. *J. Phys. Chem. A* **2018**, *122* (41), 8071−8085.
(28) Friesner, R. A. Ab initio quantum chemistry: methodology and applications. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (19), 6648−53.
(29) Zhang, X. J.; Liu, Z. P. Reaction sampling and reactivity prediction using the stochastic surface walking method. *Phys. Chem. Chem. Phys.* **2015**, *17* (4), 2757−69.
(30) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8* (2),e1354.
(31) Grambow, C. A.; Jamal, A.; Li, Y. P.; Green, W. H.; Zador, J.; Suleimanov, Y. V. Unimolecular Reaction Pathways of a gamma-Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *J. Am. Chem. Soc.* **2018**, *140* (3), 1035−1048.
(32) Margraf, J. T.; Reuter, K. Systematic Enumeration of Elementary Reaction Steps in Surface Catalysis. *ACS Omega* **2019**, *4* (2), 3370−3379.
(33) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of Reaction Pathways and Chemical Transformation Networks. *J. Phys. Chem. A* **2019**, *123* (2), 385−399.
(34) Truhlar, D. G.; Gordon, M. S. From force fields to dynamics: Classical and quantal paths. *Science* **1990**, *249* (4968), 491−498.
(35) Mackerell, A. D., Jr. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **2004**, *25* (13), 1584−604.
(36) Ponder, J. W.; Case, D. A. Force fields for protein simulations. *Adv. Protein Chem.* **2003**, *66*, 27−85.
(37) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98* (14), 146401.
(38) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134* (7), 074106.
(39) 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* (9), 6327−6337.
(40) Huang, S. D.; Shang, C.; Kang, P. L.; Liu, Z. P. Atomic structure of boron resolved using machine learning and global sampling. *Chem. Sci.* **2018**, *9* (46), 8644−8655.
(41) Kolb, B.; Zhao, B.; Li, J.; Jiang, B.; Guo, H. Permutation invariant potential energy surfaces for polyatomic reactions using atomistic neural networks. *J. Chem. Phys.* **2016**, *144* (22), 224103.
(42) Gastegger, M.; Marquetand, P. High-Dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm. *J. Chem. Theory Comput.* **2015**, *11* (5), 2187−98.
(43) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical science* **2017**, *8* (4), 3192−3203.
(44) Herr, J. E.; Yao, K.; McIntyre, R.; Toth, D. W.; Parkhill, J. Metadynamics for training neural network model chemistries: A competitive assessment. *J. Chem. Phys.* **2018**, *148* (24), 241710.
(45) Fink, T.; Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47* (2), 342−353.
(46) Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Da. *Angew. Chem., Int. Ed.* **2005**, *44* (10), 1504−8.
(47) Ma, S.; Shang, C.; Liu, Z.-P. Heterogeneous catalysis from structure to activity via SSW-NN method. *J. Chem. Phys.* **2019**, *151* (5), 050901.
(48) Huang, S. D.; Shang, C.; Kang, P. L.; Zhang, X. J.; Liu, Z. P. LASP: Fast global potential energy surface exploration. *Wiley*

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536

Interdiscip. Rev.: Comput. Mol. Sci. **2019**, *9* (6) DOI: 10.1002/wcms.1415.
(49) Ma, S.; Huang, S.-D.; Liu, Z.-P. Dynamic coordination of cations and catalytic selectivity on zinc−chromium oxide alloys during syngas conversion. *Nature Catalysis* **2019**, *2* (8), 671−677.
(50) Shang, C.; Liu, Z. P. Stochastic Surface Walking Method for Structure Prediction and Pathway Searching. *J. Chem. Theory Comput.* **2013**, *9* (3), 1838−45.
(51) Shang, C.; Zhang, X. J.; Liu, Z. P. Stochastic surface walking method for crystal structure and phase transition pathway prediction. *Phys. Chem. Chem. Phys.* **2014**, *16* (33), 17845−56.
(52) Zhang, X. J.; Shang, C.; Liu, Z. P. Stochastic surface walking reaction sampling for resolving heterogeneous catalytic reaction network: A revisit to the mechanism of water-gas shift reaction on Cu. *J. Chem. Phys.* **2017**, *147* (15), 152706.
(53) Zhang, X. J.; Shang, C.; Liu, Z. P. Double-Ended Surface Walking Method for Pathway Building and Transition State Location of Complex Reactions. *J. Chem. Theory Comput.* **2013**, *9* (12), 5745−53.
(54) Zhang, X. J.; Liu, Z. P. Variable-cell double-ended surface walking method for fast transition state location of solid phase transitions. *J. Chem. Theory Comput.* **2015**, *11* (10), 4885−94.
(55) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **1996**, *6* (1), 15−50.
(56) Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50* (24), 17953.
(57) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59* (3), 1758.
(58) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865.
(59) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision a.02; Gaussian, Inc.: Wallingford, CT, 2009.
(60) Zhao, Y.; Gonzalez-Garcia, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109* (9), 2012−2018.
(61) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555* (7698), 604−610.
(62) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.
(63) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864−75.
(64) Ponder, G. R.; Richards, G. N. Pyrolysis of inulin, glucose and fructose. *Carbohydr. Res.* **1993**, *244* (2), 341−359.
(65) Paine, J. B.; Pithawalla, Y. B.; Naworal, J. D. Carbohydrate pyrolysis mechanisms from isotopic labeling. *J. Anal. Appl. Pyrolysis* **2008**, *83* (1), 37−63.
(66) Zhang, J.; Weitz, E. An in Situ NMR Study of the Mechanism for the Catalytic Conversion of Fructose to 5-Hydroxymethylfurfural and then to Levulinic Acid Using 13C Labeled d-Fructose. *ACS Catal.* **2012**, *2* (6), 1211−1218.

DOI: 10.1021/jacs.9b11535
J. Am. Chem. Soc. 2019, 141, 20525−20536