# Complex Chemical Reaction Networks from Heuristics-Aided Quantum Chemistry

# Complex Chemical Reaction Networks from Heuristics-Aided Quantum Chemistry

Dmitrij Rappoport,*,† Cooper J Galvin,‡ Dmitry Yu. Zubarev,† and Alán Aspuru-Guzik*,†

*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA, and Pomona College, 333 North College Way, Claremont, CA 91711, USA*

E-mail: rappoport@chemistry.harvard.edu; aspuru@chemistry.harvard.edu

**Abstract**

While structures and reactivities of many small molecules can be computed efficiently and accurately using quantum chemical methods, heuristic approaches remain essential for modeling complex structures and large-scale chemical systems. Here we present heuristics-aided quantum chemical methodology applicable to complex chemical reaction networks such as those arising in metabolism and prebiotic chemistry. Chemical heuristics offer an expedient way of traversing high-dimensional reactive potential energy surfaces and are combined here with quantum chemical structure optimizations, which yield the structures and energies of the reaction intermediates and products. Application of heuristics-aided quantum chemical methodology to the formose reaction of prebiotic evolution reproduces the experimentally observed reaction products, major reaction pathways, and autocatalytic cycles.

*To whom correspondence should be addressed
†Harvard University
‡Pomona College

# 1 Introduction

Complex reaction mechanisms, in which many competing reaction steps combine to form a network of chemical reactions, are increasingly recognized as a common pattern in chemistry.[1,2] Characteristic features of complex reactions include branching and interference of reaction pathways, autocatalysis, and product inhibition and are observed in systems as varied as transition-metal catalysis,[3] cell metabolism,[4,5] and polymerization.[1,6,7] A better understanding of the network effects in these complex reactions offers means for influencing their dynamics and product composition. Useful contributions to this effort can be expected from theoretical works, which are capable of providing accurate predictions of molecular structures and reactivities. Theory and computation of kinetics of elementary reactions from first principles have made enormous progress;[8–10] nonetheless, complex reaction mechanisms continue to pose significant methodological and algorithmic challenges.[2,11,12]

Encouragingly, heuristic approaches have proven useful for solving complex and large-scale problems across diverse fields such as graph search,[13] sequence alignment,[14] and cheminformatics.[15] One does not have to look far to find heuristic methods: The classical force fields of molecular mechanics[16,17] may well be viewed as heuristic rules of classical chemical structure theory enforced by penalty functions and thus made amenable to computation. In the field of chemical reactivity of organic compounds, a similarly successful set of heuristic rules exists that regards chemical transformations as flows of electrons and is known under the moniker "arrow pushing" to students of organic chemistry.[18] The existence of simple yet predictive "arrow pushing" heuristics for polar organic reactions strongly indicates that a useful heuristic scheme may be developed from these rules. Rule-based systems have been successfully used for development and optimization of organic syntheses since the pioneering work of Corey and Wipke over 40 years ago[11,19–25] and have been recently developed into a broad-spectrum synthetic tool by Grzybowski and coworkers.[26,27]

Guided by the above expectation, we propose a computational framework of heuristics-aided quantum chemistry (HAQC) suitable for exploring complex and large-scale reaction mechanisms.

In the proposed methodology, chemical heuristics such as the "arrow pushing" rules serve to quickly navigate across high-dimensional reactive potential energy surfaces and are complemented by quantum chemical structure optimizations to locate stable reaction intermediates and products. The utility of chemical heuristics lies in their capability to map the potential energy surface onto individual chemical species and reactive trajectories into stepwise transformations, facilitating large-scale moves. Compared to generic heuristic optimization techniques such as simulated annealing and evolutionary computation,[28–30] chemical heuristics offer the advantage of having some empirical chemical knowledge built in.

Furthermore, the discretization of the reactive potential energy surface into individual chemical species gives rise to a network model of the complex reaction mechanism composed of chemical species as network nodes and chemical transformations between them as network edges.[31] Albeit a stark simplification, the network representation is convenient for studying global reaction dynamics and for exploring complex reaction properties such as reaction path interference and autocatalysis. A rich body of work addresses abstract reaction network models or the best-studied, but in many ways exceptional, reaction network of cell metabolism.[4,5,32,33] With this work, we wish to provide a methodology for constructing detailed models of arbitrary chemical reaction networks amenable to study of their global structures and dynamics.

Simple and efficient methods for describing the thermodynamic and kinetic reaction parameters are necessary to link the molecular-level description of reactive dynamics and the systems-level view of reaction networks. While quantum chemistry has the tools for computing both thermodynamic and kinetic parameters of elementary reaction steps, the associated computational cost and algorithmic challenges differ substantially. Predictions of reaction thermodynamics depend only on energy minimizations, while reaction kinetics calculations in addition require first-order saddle-point (transition state) searches within the standard treatment of transition-state theory.[8–10] We are interested in computationally inexpensive methods applicable to large reaction networks and thus replace kinetic reaction parameters by heuristic functions of the energies of the reactants, intermediates, and products along the reaction path. Our approach is motivated by Hammond's postulate,

which holds that transitions states of reactions involving unstable intermediates resemble the intermediates themselves[34] or, alternatively, that the reaction energy and the height of the activation barrier are correlated with each other.[35] We show below that even simple heuristic kinetic parameters lead to useful predictions of reaction products and pathways.

We applied the HAQC methodology to the reaction network of the *formose* reaction, a well-studied organic reaction occurring in alkaline solutions of formaldehyde and resulting in a complex mixture of aldose and ketose sugars.[36,37] More than 40 compounds were experimentally identified as products of the formose reaction[38,39] and major pathways are known,[37,40] however many mechanistic details remain obscure. The formose reaction is one of the simplest organic reaction exhibiting autocatalysis[41] and was early conjectured as a potential route to sugars in the course of prebiotic evolution.[40,42–44] We present models of the formose reaction in different stoichiometries obtained using a combination of chemical heuristics and semiempirical quantum chemistry (Section 3). The formose reaction models contained formose sugars up to $C_5$ known from experiments[38,39] and major reaction pathways postulated in the literature.[37,41] Furthermore, the reaction models obtained using heuristics-aided quantum chemistry permit analyses of chemical composition, energetics, and network structure, which are detailed in our companion publication.[45]

This paper is organized as follows. Section 2 develops the framework of the HAQC methodology and heuristic thermodynamic and kinetic reaction feasibility criteria. Models of the formose reaction network in different stoichiometries are constructed and their chemical compositions are analyzed in Section 3. A discussion and outlook are given in Section 4.

## 2 Chemical Heuristics for Complex Reaction Mechanisms

Many, if not most, hard problems in chemical structure and reactivity may be traced back to the high dimensionality of the quantum chemical models for electrons and nuclei. This is particularly true for complex reaction networks, which are chracterized by having complicated potential energy surfaces with numerous energy minima. While stable energy minima of medium-sized and large

molecules can be located in an efficient and robust way, enumerating first-order stationary points (transition states) on reactive potential energy surfaces is still a challenging task, despite notable progress.[9,10] We wish to characterize both thermodynamic and kinetic properties of all reactions of a complex reaction network, which requires us to develop simple and robust approximations. In our approach to this problem, we draw inspiration from molecular mechanics that successfully transformed heuristic rules of chemical bonding into efficient computational schemes.[16,17,46] As was the case with early classical force fields, we proceed by introducing a number of heuristic but physically motivated propositions that allow us to tackle complex reaction networks with hundreds or thousands of distinct chemical species and transformations. The heuristics-aided quantum chemistry (HAQC) approach is based on the following assumptions.

1. Reaction products and pathways are obtained by a set of heuristic *transformation rules*, which are recursively applied to structure formulas of molecules. We encode molecular structures by their SMILES (simplified molecular-input line entry system) representations.[47] The transformation rules used in this work are given in Scheme 1, where X, Y, and Z represent arbitrary atoms.

| | | |
|---|---|---|
| (a) | $\mathrm{X{=}Y} + \mathrm{Z}^+ \rightarrow \mathrm{X}^+{-}\mathrm{Y}{-}\mathrm{Z}$ | Electrophilic addition |
| (b) | $\mathrm{X{=}Y} + \mathrm{Z}^- \rightarrow \mathrm{X}^-{-}\mathrm{Y}{-}\mathrm{Z}$ | Nucleophilic addition |
| (c) | $\mathrm{X}^+{-}\mathrm{Y}^- \rightarrow \mathrm{X{=}Y}$ | Double bond depolarization |
| (d) | $\mathrm{X}{-}\mathrm{Y} \rightarrow \mathrm{X}^+ + \mathrm{Y}^-$ | Single bond breaking |
| (e) | $\mathrm{X}^+ + \mathrm{Y}^- \rightarrow \mathrm{X}{-}\mathrm{Y}$ | Single bond formation |
| (f) | $\mathrm{X}^+ \smile \mathrm{Y}^- \rightarrow \underset{\smile}{\mathrm{X}{-}\mathrm{Y}}$ | Ring closure |

Scheme 1: Heuristic transformation rules for polar reactions used in this work.

We wish to stress that these primitive transformations are not required to describe genuine elementary reactions. Rather, they provide a simple device for constructing elementary reactions in an unbiased fashion and should capture the electron flow in polar organic reactions in aqueous solutions. The primitive transformations (a), (b), (d)–(f) correspond to actual elementary reactions, while depolarization of multiple bonds (c) does not have an equivalent in quantum chemistry and is energy neutral.

2. The SMILES representations of the reaction intermediates and products obtained by way of heuristic transformations are mapped onto the corresponding three-dimensional structures and are subject to quantum chemical structure optimizations. In order to obtain a consistent description of the chemical structures that are part of the complex reaction network, a robust equivalence should be enforced between the structure formulas (given by SMILES) and the three-dimensional optimized structures from quantum chemistry. Therefore, we exclude all molecules, for which structure optimization does not preserve heavy-atom connectivity.

3. The heuristic transformation rules operate on molecular collections, which we refer to as *flasks* $\mathscr{F}_K = \{M_{K1}, \ldots, M_{Km_K}\}$ in the following. $K$ is the flask index and $M_{Kk}$ denotes the constituent molecules of flask $K$. We consider the molecular collection as a closed system and keep its stoichiometry constant across flasks. As a consequence, flask energies are directly comparable to each other. Further, we assume that interactions between the molecules are negligible and thus flask energies are well approximated by sums of the energies of its constituent molecules $E_K = \sum_k \varepsilon_{Kk}$, which may be computed using any suitable quantum chemical method.

4. We distinguish between neutral and charged constituent molecules and label the flasks containing only neutral constituent molecules as *product flasks*. Assuming that the overall flask stoichiometry is conserved and the total charge is zero, we can expect the neutral forms of all constituent molecules to form in a sufficiently large number of transformation steps. Therefore, we may represent all stable reaction products as constituent molecules of product flasks without limiting the generality of the procedure. We utilize that polar reactions involve movement of electric charges between reaction participants, producing charged compounds as intermediates and, following Hammond's postulate, we make the additional assumption that the sequence of flasks containing one or more charged constituent molecules (*intermediate flasks*) may be considered as approximations to the instantaneous configurations along the reaction trajectory.

5. The recursive application of heuristic rules produces an auxiliary network representation containing both product flasks and intermediate flasks. (Fig. 1(a)) The root node of the network is the initial flask $\mathscr{F}_1$, which is referred to as *generation* 0 of the network, and the *generation* $g > 0$ is obtained by combinatorially applying heuristic rules of Scheme 1 to all flasks of generation $g-1$. Incidentally, the generation number $g$ may serve as a coarse-grained time variable indicating the progress of the reaction. Since multiple paths may lead to the same flask, the auxiliary network representation is not a true tree graph. The *reaction network* is obtained from the auxiliary network representation by retaining only product flasks as *network nodes* and adding *network edges* based on the threshold criteria for thermodynamic and kinetic reaction parameters developed below. (Fig. 1(b))

6. We employ flask energies of product flasks and intermediate flasks to define thermodynamic and kinetic reaction parameters for transformations between product flasks $\mathscr{F}_K \to \mathscr{F}_L$. Energy differences between initial and final product flasks, $\Delta E_{K\to L} = E_L - E_K$, are natural choices for thermodynamic parameters and are independent of possible multiple pathways between $\mathscr{F}_K$ and $\mathscr{F}_L$. In addition, we develop heuristic kinetic reaction parameters, which take into account the flask energies of initial and final product flasks as well as the flask energies of the intermediate flasks connecting them. The heuristic kinetic reaction parameter $W_{K\to L}$ of a $N$-step transformation $\mathscr{F}_K \to \mathscr{F}_L$ should be a function of the flask energies $\{E_{K_i},\ i = 0,\ldots,N\}$ of the sequence of flasks $\{\mathscr{F}_{K_0} = \mathscr{F}_K, \mathscr{F}_{K_1}, \ldots, \mathscr{F}_{K_N} = \mathscr{F}_L\}$, which is non-negative, additive for concatenated reaction sequences, and physically reasonable. We suggest *climb* parameter $W_{c,K\to L}$ and *arc* parameter $W_{a,K\to L}$ as heuristic kinetic parameters and assess their performance below. If multiple paths exist for the transformation $\mathscr{F}_K \to \mathscr{F}_L$ are present, we choose the *most feasible* path among them, which is defined by having the smallest heuristic kinetic reaction parameter.

7. Simple threshold criteria serve to determine thermodynamic and kinetic feasibility of transformations between product flasks. Only transformations $\mathscr{F}_K \to \mathscr{F}_L$ with $\Delta E_{K\to L} \leq \Delta E^{\max}$

and $W_{K\to L} \leq W^{\max}$ are added as network edges to the reaction network, where $\Delta E^{\max}$ and $W^{\max}$ are the thermodynamic and kinetic threshold constants, respectively.

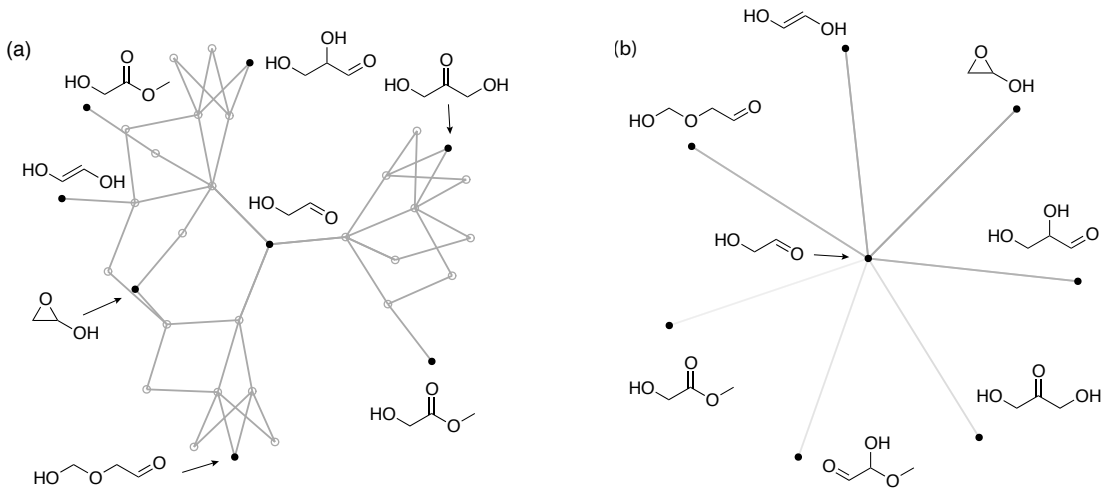


Figure 1: (a) Auxiliary network and (b) reaction network $\mathsf{T}_3$ of formose reaction after 3 generations. Neutral flasks are indicated by black solid circles, intermediate flasks are shown by open circles. Chemical formulas denote the largest constituent molecule of each flask. Line intensities signify kinetic arc parameters of individual reaction steps; smaller arc values (more feasible reactions) are denoted by darker lines.

The heuristic kinetic reaction parameters are motivated by Hammond's postulate and are designed to approximate reaction activation barriers. In the framework of transition state theory,[8] the activation barrier of a reaction is given by the energy of the highest point along the reaction energy profile relative to the preceding energy minimum and may be approximated by the highest-energy reaction intermediate. For multistep reactions, the elementary reaction with the highest activation barrier determines the overall kinetics as the rate-limiting step. A convenient functional form for heuristic kinetic parameters is suggested by the following analogy: In thermal equilibrium, the abundance of flask $\mathscr{F}_K$ is given by the Boltzmann distribution, $c_K \propto \exp(-\beta E_K)$, in which $\beta = 1/(k_B T)$ with Boltzmann constant $k_B$ and absolute temperature $T$. By analogy, we define heuristic kinetic parameters $W_{K\to L}$ for the reaction $\mathscr{F}_K \to \mathscr{F}_L$ in such a manner that the corresponding reaction rate may be represented as $k_{W\to L} \propto \exp(-\beta W_{K\to L})$.

The simplest approximation for the kinetic reaction parameter follows if we assume that the energy of the highest-energy intermediate flask approximates the activation barrier of the rate-

limiting step. The corresponding kinetic *climb* parameter $W_{c,K\to L}$ for the $N$-step reaction $\mathscr{F}_K \to \mathscr{F}_L$ is given by

$$W_{c,K\to L} = \sum_{i=0,\ldots,N-1} \max(E_{K_{i+1}} - E_{K_i}, 0)\,, \tag{1}$$

where we use the flask energies $\{E_{K_i},\ i = 0,\ldots,N\}$ as defined above. By definition, $W_{c,K\to L}$ yields the highest activation barrier of a multistep reaction relative to the initial flask $\mathscr{F}_K$.

In complex reaction mechanisms, a further consideration are branching points in reactive trajectories, which reduce the yield of each individual reaction product. Assuming that trajectory bifurcations occur with a constant rate at each intermediate flask, the probability of reaching a given product flask decreases exponentially with the number of steps. Hence, it appears reasonable to use an energetic parameter that increases roughly linearly with the number of transformation steps, and we are led to define the kinetic *arc* parameter $W_{a,K\to L}$ for the reaction $\mathscr{F}_K \to \mathscr{F}_L$ as

$$W_{a,K\to L} = \sum_{i=0}^{N-1} \left((E_{K_{i+1}} - E_{K_i})^2 + \alpha^2\right)^{\frac{1}{2}} \tag{2}$$

where $\alpha$ is an empirical parameter and the flask energies $\{E_{K_i},\ i = 0,\ldots,N\}$ are defined as above. We can consider $\alpha$ as a penalty factor for long paths and set $\alpha = 1$ eV for the purposes of the following discussion.

We can calibrate the kinetic climb and arc parameters and assess their performance using experimental knowledge of constituent processes of the formose reaction. We employ the heuristic rule set of Scheme 1 and use the OpenBabel structure builder to convert SMILES strings to three-dimensional models.[48–50] The energies are determined throughout this work by structure optimizations using the PM7 semiempirical method within the MOPAC package.[51] Solvation effects in water are included using the conductor-like solvation model (COSMO)[52] with an effective dielectric constant of $\varepsilon = 78.4$. We consider reactions involving one molecule glycolaldehyde and one formaldehyde molecule ($\mathscr{F}_1$ ={O=CHCH$_2$OH, CH$_2$=O}, Fig. 2). The predicted reaction mechanisms include several well-established reaction routes: (i) enolization of glycolaldehyde to ethene-1,2-diol (product indicated by blue circle in Fig. 2),[53,54] (ii) aldol addition of glyco-

laldehyde and formaldehyde to form glyceraldehyde[55,56] (product in red), and (iii) hemiacetal formation (product in green).[57] As suggested by Hammond's postulate, the intermediate flasks, shown by empty circles in Fig. 2, trace the movement of charge in reactions (i)–(iii) in fairly good approximation. The last step of the enolization describes a fictitious depolarization of the C=C double bond and is energy neutral. The reaction paths (i) and (ii) share the enolate anion as the highest-energy intermediate flask and thus have the same climb parameter $W_c = 1.66$ eV, while their arc parameters are different: $W_a = 4.70$ eV (enolization) and $W_a = 5.13$ eV (aldol addition). An additional reaction, (iv) a C−C coupling reaction via an aldehyde anion is predicted to occur at larger values of kinetic parameters ($W_c = 2.75$ eV, $W_a = 7.45$ eV). While the reaction product of (iv), dihydroxyacetone (shown in red in Fig. 2), is more stable than the products of reactions (i)–(iii), the larger values of kinetic parameters reflect the experimental finding that deprotonation of an aldehydic proton is unfavorable and requires *umpolung* techniques.[58] In contrast, we expect the enolate-based reactions (i) and (ii) as well as the hemiacetal formation (iii) ($W_c = 1.81$ eV, $W_a$ = 5.33 eV) to be feasible in aqueous solution.

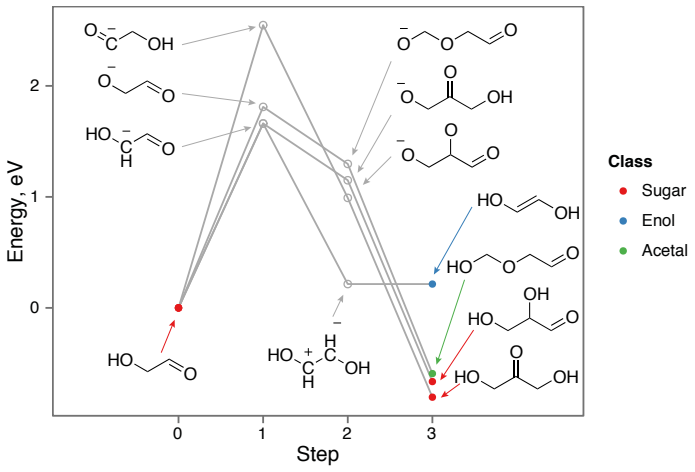


Figure 2: Selected energy profiles along 3-step chemical reactions between glycolaldehyde and formaldehyde ($\mathscr{F}_1$ ={O=CHCH$_2$OH, CH$_2$=O}). Product flasks are represented by solid circles, intermediate flasks by empty circles. Color coding and chemical formulas denote the largest constituent molecule of each flask (see legend).

In order to investigate the performance of kinetic climb and arc parameters in more detail, we consider the predicted formose reaction products after 3 and 6 generations starting from the

flask $\mathscr{F}_1$ ={O=CHCH$_2$OH, CH$_2$=O, CH$_2$=O} (tetrose stoichiometry). We denote the resulting reaction networks as $\mathsf{T}_3$ and $\mathsf{T}_6$, respectively. Using suitable threshold values for either kinetic climb parameter ($\Delta E^{\max} = 0.75$ eV, $W_c^{\max} = 2.00$ eV) or kinetic arc parameter ($\Delta E^{\max} = 0.75$ eV, $W_a^{\max} = 5.50$ eV), we are able to generate a classification of feasible / unfeasible reactions for the $\mathsf{T}_3$ network that agrees with empirical expectations outlined above (Fig. 3).

The differences between the threshold criteria based on the kinetic climb and arc parameters are noticeable in the $\mathsf{T}_6$ network (Fig. 4). A large number of compounds can be reached directly from the initial flask via a comparatively low barrier (small values of $W_c$); however, for many of them this is possible only by way of a long sequence of intermediate flasks. A simple threshold criterion using kinetic climb parameter does treat this problem adequately and therefore does not permit a simple feasible / unfeasible classification for multistep reactions. The kinetic arc parameter exhibits a more desirable behavior: By penalizing long transformation sequences, it spreads the parameter distribution such that a consistent set of threshold criteria ($\Delta E^{\max} = 0.75$ eV, $W_a^{\max} = 5.50$ eV) remains useful over longer sequences of steps. These criteria are used throughout this work. The detailed results are given in Section S1 of the Supporting Information.

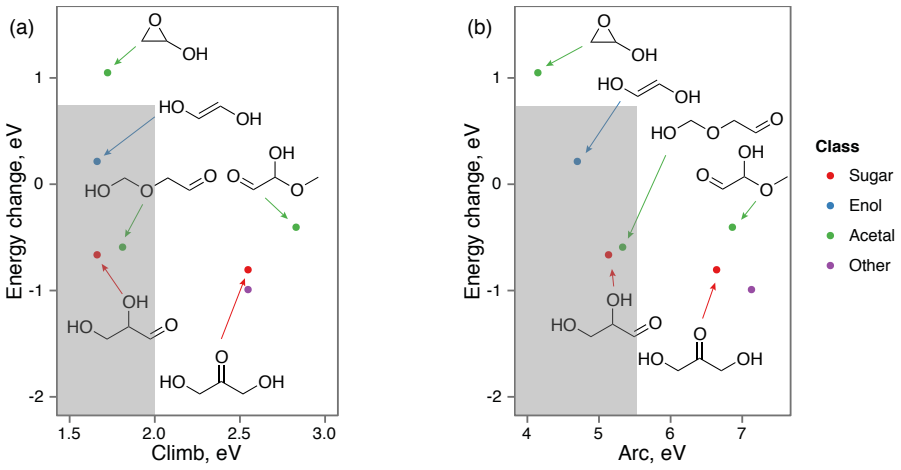


Figure 3: Thermodynamic and kinetic reaction parameters for formose reaction products in the $\mathsf{T}_3$ network using (a) kinetic climb parameter $W_c$, (b) kinetic arc parameter $W_a$. Filled circles represent product flasks; color coding and chemical formulas denote the largest constituent molecule of each flask (see legend). The dark shaded areas depict the range of feasible reactions given by the threshold criteria for thermodynamic and kinetic reaction parameters.

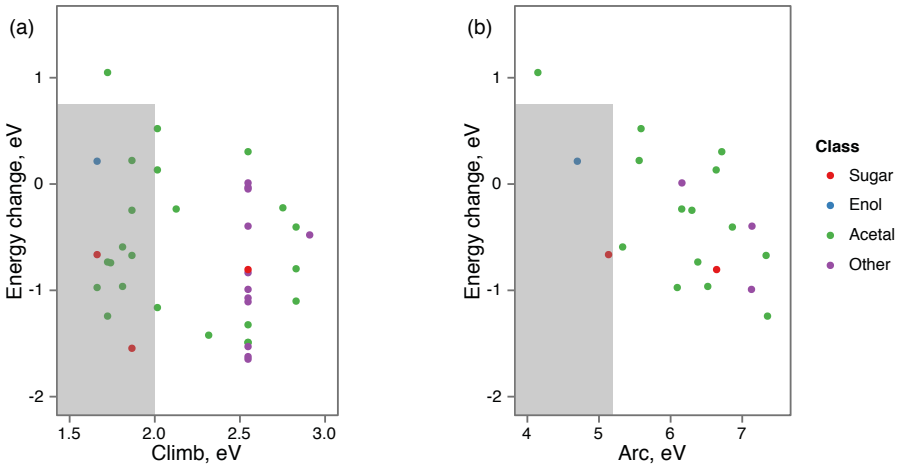


Figure 4: Thermodynamic and kinetic reaction parameters for formose reaction products in the $T_6$ network using (a) kinetic climb parameter $W_c$, (b) kinetic arc parameter $W_a$. See Fig. 3 for details.

A potential weakness of the kinetic arc parameter is that it does not distinguish between the forward and backward reactions. However, since we only apply simple threshold selection criteria, this is unlikely to significantly affect our conclusions. Nevertheless, it is desirable to develop kinetic parameters that are irreversible and show linear increase with number of steps. This effort will require considering a wider range of chemical reactions and is reserved for future work. The accuracy of our predictions are limited by the choice of the heuristic kinetic parameters as well as systematic errors of quantum chemical calculations. Furthermore, we disregard the stereochemistry and conformation equilibria of the formose products in this work. We expect the effects of the latter approximations to be small compared to the errors related to heuristic kinetic parameters and simple threshold criteria. The deviation from experimental results due to these challenges will be addressed in future work.

# 3 Probing the Chemistry of the Formose Reaction Network

The formose reaction is a self-condensation of formaldehyde in alkaline solutions[36,37] and at surfaces of various minerals.[40,59,60] The presence of autocatalytic cycles[41] and the mechanistic parallels to sugar metabolism led to conjectures that it played an important role in the prebiotic forma-

tion of sugars.[42–44] The product mixture of the formose reaction was analyzed by multiple groups and more than 40 reaction products were identified to date.[37–39]

We investigated the structures and properties of formose reaction networks obtained after 9 generations starting from the initial flasks $\mathscr{F}_1$ ={O=CHCH$_2$OH, CH$_2$=O, CH$_2$=O} (tetrose stoichiometry, denoted by $\mathsf{T}_9$) and $\mathscr{F}_1$ ={O=CHCH$_2$OH, CH$_2$=O, CH$_2$=O, CH$_2$=O} (pentose stoichiometry, $\mathsf{P}_9$). Since the heuristic heuristic transformation rules used in the network construction preserve flask stoichiometry (Scheme 1), the nodes of the the resulting network representation correspond to the possible *states* of the reactive system with fixed stoichiometry. We refer to this network representation as the *finite-state representation*, by analogy with finite-state machines,[61] and contrast it with the commonly used representations of metabolic networks as *interaction networks*, in which individual metabolites are network nodes and network edges connect all reaction participants with each other.[5,32,62,63] The finite-state representation of the reaction network is a directed network with edge weights given by thermodynamic and / or kinetic parameter values. In the following, we only consider the out-component of the reaction network reachable from the initial flask $\mathscr{F}_1$.

The $\mathsf{T}_9$ network contained a total of 149 nodes, including 146 distinct neutral molecules, and 445 edges. The chemical composition of the $\mathsf{T}_9$ network is shown in Table 1. The graphic representation of the network was created by the open-source Cytoscape program[64] using a force-directed algorithm followed by minimal manual adjustments (Fig. 5). The product flasks were characterized by the chemical class of the largest constituent molecule as sugars, enols / enediols, acetals / hemiacetals, or other. (Table 1) The predicted formose products included 2 trioses (glyceraldehyde and dihydroxyacetone) and 3 tetroses (aldotetrose, ketotetrose, and the branched tetrose 2,3-dihydroxy-2-(hydroxymethyl)propanal), which were experimentally identified in the formose reaction mixture.[38,39]

Major reaction pathways of sugar formation were determined by minimizing the sum of the kinetic arc parameters along the path from $\mathscr{F}_1$ to the sugar-containing product flasks (Fig. 5). Pathways predicted in this way included mechanisms previously postulated for the formose reac-

Table 1: Chemical composition of $T_9$ and $P_9$ networks.

| | $T_9$ | $P_9$ |
|---|---|---|
| Sugars | 6 | 11 |
| Acetals | 78 | 235 |
| Enols | 9 | 9 |
| Other | 53 | 99 |
| Total | 146 | 354 |

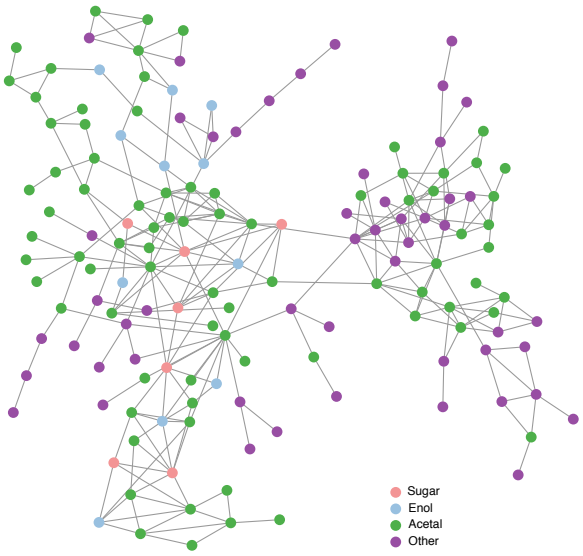


Figure 5: Finite-state representation of the $T_9$ network. Filled circles represent product flasks. Color coding and chemical formulas denote the largest constituent molecule of the respective flask (see legend). Black solid lines indicate major pathways of sugar formation.

tion.[37,40] The central pathway of carbon-chain elongation was found to involve sequences of aldol additions[55,56] and aldose–ketose isomerizations[54] (Fig. 6). As discussed above, glyceraldehyde was formed by aldol condensation of glycolaldehyde and formaldehyde, while subsequent aldol condensation with another molecule of formaldehyde yielded the branched tetrose 2,3-dihydroxy-2-(hydroxymethyl)propanal. Unbranched carbon-chain elongation involved an isomerization of glyceraldehyde to dihydroxyacetone via an enediol intermediate (Lobry de Bruyn–van Ekenstein isomerization),[54] followed by another aldol condendation reaction, which produced ketotetrose. The isomerization of ketotetrose via an enediol intermediate produced aldotetrose. Notably, the aldose–ketose isomerizations involve endothermic steps and appear as the slow steps of sugar formation. (Fig. 5)

In addition, several unexpected reaction pathways were obtained involving three-membered and four-membered cyclic tetrose hemiacetals. (Fig. 6) These reaction pathways involve fewer reactions and appear to provide a shortcut to tetrose sugars. However, the strained three-membered and four-membered hemiacetal structures have not been experimentally characterized, and it is undetermined if they occur as reaction intermediates in aqueous solutions. The favorable flask energies associated with these structures are possibly an artifact of the semiempirical PM7 method and may be corrected by more accurate quantum chemical methods. The full list of reaction products of the $\mathsf{T}_9$ network is given in Section S2 of the Supporting Information. The details of the sugar formation pathways can be found in Section S3 of the Supporting Information.

The $\mathsf{T}_9$ network contained the prominent autocatalysis feature of the formose reaction suggested by Breslow.[41] Breslow's mechanism includes the formation of aldotetrose via a sequence of aldol additions and isomerizations, followed by the retroaldol cleavage of aldotetrose into two glycolaldehyde molecules.[41] Note that in the finite-state representation of the reaction network, the initial and the final flasks of autocatalytic processes are not identical and thus do not form closed cycles. Instead, product flasks arising from autocatalytic processes can be recognized by the doubling of the number of glycolaldehyde molecules per flask (Fig. 5). In addition, autocatalytic cycles involving strained three- and four-membered hemiacetals were found in the $\mathsf{T}_9$

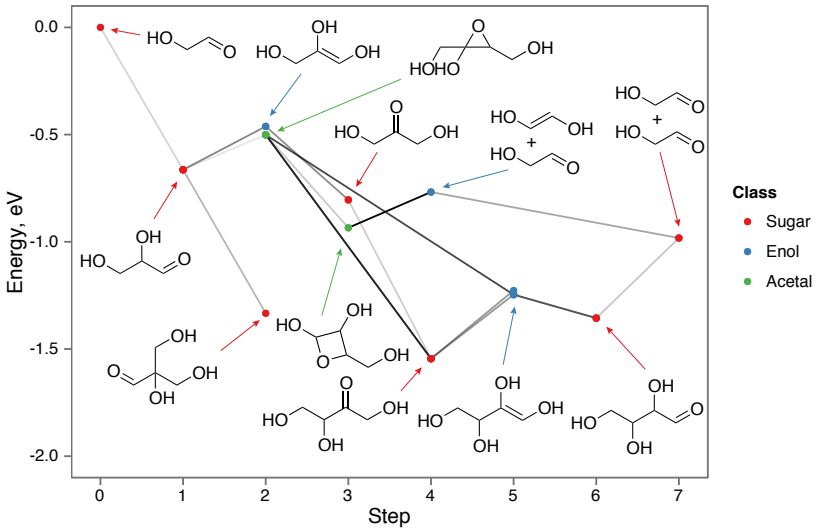


Figure 6: Major reaction pathways of sugar formation in the $T_9$ network. See Fig. 5 for details. Line intensities signify kinetic arc parameters of individual reaction steps; smaller arc values (more feasible reactions) are denoted by darker lines.

network and were favored by shorter reaction sequences (Fig. 6). The key step of these pathways involved an oxetane ring cleavage to glycolaldehyde and ethene-1,2-diol. Along with the four-membered aldotetrose hemiacetal, this mechanism might be rejected on thermodynamic grounds by more accurate quantum chemical methods.

The $P_9$ network consisted of 371 neutral flasks (354 distinct molecules) connected by 1114 reactions (Fig. 7). The reaction mixture contained 11 sugars including 3 pentoses: 3-ketopentose, 2,3,4-trihydroxy-2-(hydroxymethyl)butanal and 1,3,4-trihydroxy-3-(hydroxymethyl)butan-2-one. The formation of 2-ketopentose and aldopentose is expected after 12 and 15 generations, respectively. The subgraph of the $P_9$ network containing sugars, enols, and enediols was found to be qualitatively similar to that of the $T_9$ network but exhibited a larger set of concurrent reaction pathways as well as several Breslow-type autocatalytic processes involving higher sugars, e. g., dihydroxyacetone, as catalysts for condensation of formaldehyde (Fig. 7). The full list of reaction products of the $P_9$ network is given in Section S2 of the Supporting Information.

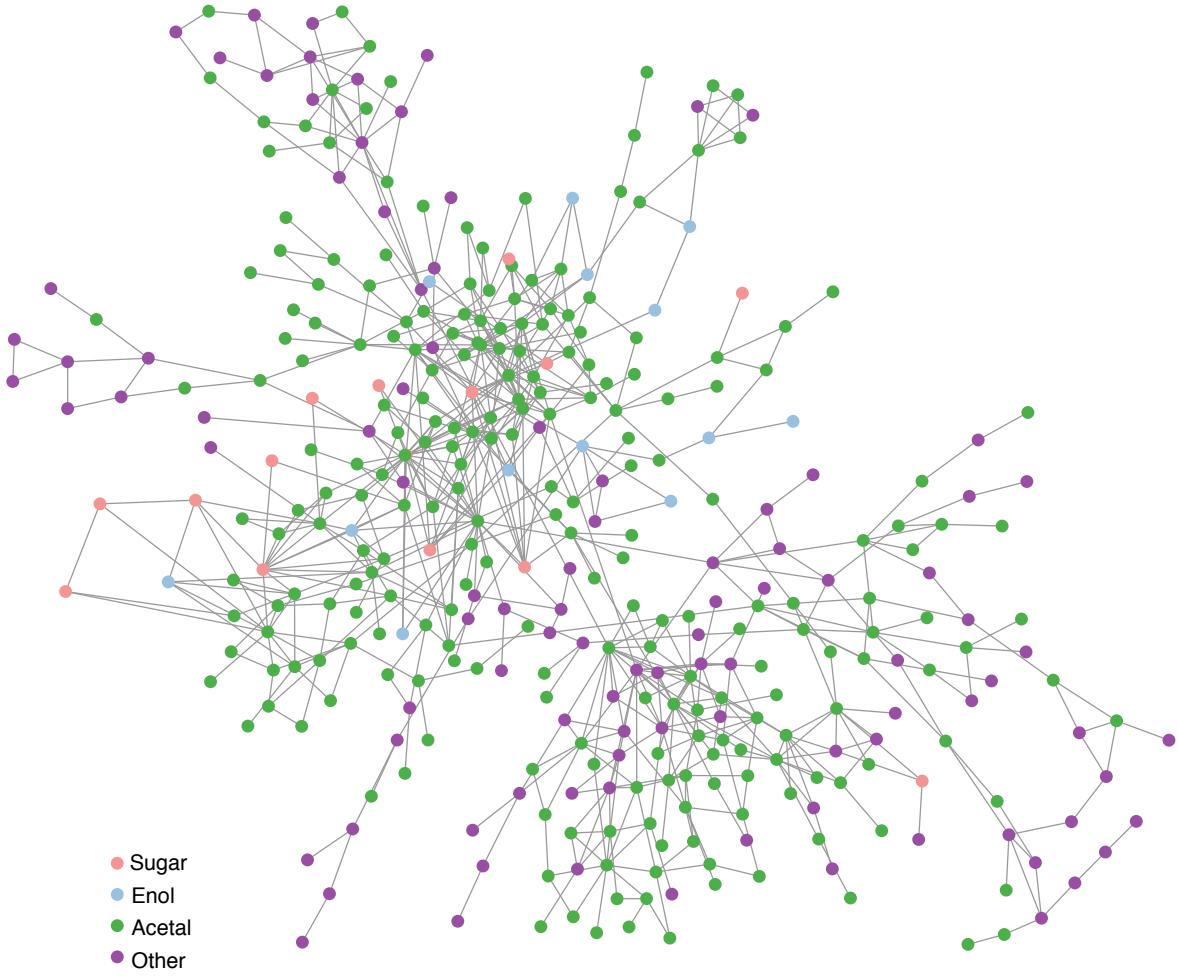


Figure 7: Finite-state representation of the $P_9$ network. See Fig. 5 for details.

# 4 Discussion and Outlook

Studies of complex reaction networks, their properties, and dynamics are a central theme in cell metabolism and chemical process modeling. A considerable amount of experimental data and chemical experience are required to identify the relevant chemical species and reaction pathways. An even more uncertain picture presents itself in the field of origins of life as both the chemical composition of the primordial mixture and the external conditions are the subject of substantial debate. This work presents the first step towards construction of *global* models of complex reaction networks from quantum chemistry. We seek to overcome the main challenge of complex reaction networks—the high dimensionality of the reactive potential energy surfaces—by using chemical heuristics borrowed from organic chemistry. Quantum chemical methods can then employed to explore the *local* structure of potential energy surfaces, relying on well-established and efficient computational procedures.

The chemical heuristics used in this work (Scheme 1) were chosen to be generic representations of polar organic reactions and to introduce as little bias as possible. Using chemical heuristics and semiempirical quantum chemistry, sugars up to $C_5$ emerge naturally as formose reaction products, and aldol condensations and aldose–ketose isomerizations are predicted as favorable reaction mechanisms, in line with expectations from experiment. However, the presence of strained three- and four-membered cyclic hemiacetals (Fig. 6) indicates that a number of improvements can be expected: (i) More accurate quantum chemical methods than the PM7 semiempirical method and COSMO solvation model used in this work (Mean unsigned error of the PM7 method for reaction energies of simple organic reactions is 4 kcal/mol[51]); (ii) Improvements in thermodynamic and kinetic reaction parameters and more sophisticated classification approaches for feasible / unfeasible reactions; (iii) Refinement and extension of rules of chemical transformation beyond "arrow pushing" rules of polar organic reactions; and (iv) Combinations with existing methods of global potential energy surface (PES) exploration.[31] An important extension is the development of heuristic rule sets for more challenging classes of chemical reactions such as radical reactions, photochemical processes, and reactions involving organometallic compounds. Methods of statisti-

cal inference may help in *deriving* new rule sets specific to these domains from the existing body of experimental data or quantum chemical calculations.

The formose reaction is a convenient testbed for the HAQC approach since many formose products have been identified and mechanistic proposals for major reaction pathways exist. A host of other complex reaction networks have been described but little is known about their product compositions and mechanisms. Complex chemical reactions of relevance to prebiotic chemistry include selective formose reactions catalyzed by phosphate,[65] borate,[40] or silicate;[66] condensations of hydrogen cyanide and formamide to nitrogen heterocycles;[67] the triose–ammonia reaction;[68,69] and the nucleoside synthesis recently suggested by Sutherland and co-workers.[70,71] Detailed studies of these and other abiotic reaction networks may help to elucidate common properties of reaction networks and differences from networks formed by evolution. Work along these lines is described in our companion publication.[45]

Finally, the combination of heuristic rules and quantum chemical calculations might be viewed as an expedient tool for exploring chemically accessible regions of chemical space.[31,72–74] Coupled with efficient quantum chemical methodology and high-throughput computation, it holds promise for novel approaches for molecular design and optimization.

## Acknowledgement

This work was supported the Cyberdiscovery Initiative Type II ($CDI^2$) grant of the National Science Foundation (NSF), grant number OIA-1125087. CJG was supported by the Research Experience for Undergraduates (REU) summer research program of the NSF.

## References

(1) Helfferich, F. G. *Kinetics of Multistep Reactions*, 2nd ed.; Comprehensive Chemical Kinetics; Elsevier: Amsterdam, 2004; Vol. 40.

(2) Vinu, R.; Broadbelt, L. J. *Annu. Rev. Chem. Biomol. Eng.* **2012**, *3*, 29–54.

(3) Hartwig, J. F. *Organotransition Metal Chemistry. From Bonding to Catalysis*; University Science Books: Sausalito CA, 2010.

(4) Palsson, B. *Systems Biology: Simulation of Dynamic Network States*; Cambridge University Press: Cambridge UK, 2011.

(5) Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabási, A. L. *Nature* **2000**, *407*, 651–654.

(6) Ludlow, R. F.; Otto, S. *Chem. Soc. Rev.* **2007**, *37*, 101.

(7) Li, J.; Nowak, P.; Otto, S. *J. Am. Chem. Soc.* **2013**, *135*, 9222–9239.

(8) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. *J. Phys. Chem.* **1996**, *100*, 12771–12800.

(9) Henkelman, G.; Jóhannesson, G.; Jónsson, H. In *Theoretical Methods in Condensed Phase Chemistry*; Schwartz, S. D., Ed.; Kluwer: Dordrecht, 2002; pp 269–302.

(10) Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514–1527.

(11) Broadbelt, L. J.; Pfaendtner, J. *AIChE J.* **2005**, *51*, 2112–2121.

(12) Green, Jr, W. H. In *Advances in Chemical Engineering*; Marin, G. B., Ed.; Elsevier, 2007; Vol. 32; pp 1–50.

(13) Russell, S. J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Prentice Hall: Upper Saddle River NJ, 2010.

(14) Sung, W.-K. *Algorithms in Bioinformatics: A Practical Introduction*; CRC Press: Boca Raton FL, 2011.

(15) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. *Nature Chem.* **2012**, *4*, 90–98.

(16) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.

(17) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Nat. Acad. Sci.* **2005**, *102*, 6665–6670.

(18) Levy, D. E. *Arrow-Pushing in Organic Chemistry. An Easy Approach to Understanding Reaction Mechanisms*; Wiley: Hoboken NJ, 2008.

(19) Corey, E. J.; Wipke, W. T. *Science* **1969**, *166*, 178–192.

(20) Corey, E. J.; Long, A. K.; Rubenstein, S. D. *Science* **1985**, *228*, 408–418.

(21) Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M.; Reitsam, K.; Stein, N. *Angew. Chem. Int. Ed.* **1993**, *32*, 201–227.

(22) Jorgensen, W. L.; Laird, E. R.; Gushurst, A. J.; Fleischer, J. M.; Gothe, S. A.; Helson, H. E.; Paderes, G. D.; Sinclair, S. *Pure Appl. Chem.* **1990**, *62*, 1921–1932.

(23) Todd, M. H. *Chem. Soc. Rev.* **2005**, *34*, 247.

(24) Ihlenfeldt, W.-D.; Gasteiger, J. *Angew. Chem. Int. Ed.* **1996**, *34*, 2613–2633.

(25) Van Geem, K. M.; Reyniers, M.-F.; Marin, G. B.; Song, J.; Green, Jr, W. H.; Matheu, D. M. *AIChE J.* **2006**, *52*, 718–730.

(26) Gothard, C. M.; Soh, S.; Gothard, N. A.; Kowalczyk, B.; Wei, Y.; Baytekin, B.; Grzybowski, B. A. *Angew. Chem. Int. Ed.* **2012**, *51*, 7922–7927.

(27) Kowalik, M.; Gothard, C. M.; Drews, A. M.; Gothard, N. A.; Weckiewicz, A.; Fuller, P. E.; Grzybowski, B. A.; Bishop, K. J. M. *Angew. Chem. Int. Ed.* **2012**, *51*, 7928–7932.

(28) Pearl, J. *Heuristics. Intelligent Search strategies for Computer Problem Solving*; Addison Wesley: Reading MA, 1984.

(29) Fogel, D. B., Bäck, T., Michalewicz, Z., Eds. *Evolutionary Computation*; Taylor and Francis: New York NY, 2000; Vol. 1.

(30) Blum, C.; Roli, A. *ACM Comput. Surv.* **2003**, *35*, 268–308.

(31) Wales, D. J. *Energy Landscapes. Applications to Clusters, Biomolecules and Glasses*; Cambridge University Press: Cambridge UK, 2003.

(32) Barabási, A.-L.; Oltvai, Z. N. *Nature Rev. Gen.* **2004**, *5*, 101–113.

(33) Bar-Even, A.; Flamholz, A.; Noor, E.; Milo, R. *Nature Chem. Biol.* **2012**, *8*, 509–517.

(34) Hammond, G. S. *J. Am. Chem. Soc.* **1955**, *77*, 334–338.

(35) Evans, M. G.; Polanyi, M. *Trans. Faraday Soc.* **1938**, *34*, 11.

(36) Boutlerow, A. *C. R. Acad. Sci.* **1861**, *53*, 145–147.

(37) Mizuno, T.; Weiss, A. H. In *Advances in Carbohydrate Chemistry and Biochemistry*; Tipson, R. S., Horton, D., Eds.; Academic Press, 1974; Vol. 29; pp 173–227.

(38) Decker, P.; Schweer, H.; Pohlmann, R. *J. Chromatogr.* **1982**, *244*, 281–291.

(39) Zweckmair, T.; Bohmdorfer, S.; Bogolitsyna, A.; Rosenau, T.; Potthast, A.; Novalin, S. *J. Chromatogr. Sci.* **2013**, DOI:10.1093/chromsci/bmt004.

(40) Kim, H.-J.; Ricardo, A.; Illangkoon, H. I.; Kim, M. J.; Carrigan, M. A.; Frye, F.; Benner, S. A. *J. Am. Chem. Soc.* **2011**, *133*, 9457–9468.

(41) Breslow, R. *Tetrahedron Lett.* **1959**, *1*, 22–26.

(42) Oparin, A. I. *The Origin of Life on the Earth*, 3rd ed.; Academic Press: New York, 1957.

(43) Orgel, L. E. *Crit. Rev. Biochem. Mol. Biol.* **2004**, *39*, 99–123.

(44) Cairns-Smith, A. G.; Walker, G. L. *Biosystems* **1974**, *5*, 173–186.

(45) Rappoport, D.; Zubarev, D. Y.; Galvin, C. J.; Aspuru-Guzik, A. Structure and Properties of Complex Chemical Reaction Networks Obtained from First Principles. 2013; In preparation.

(46) Rappé, A. K.; Casewit, C. J. R. *Molecular Mechanics Across Chemistry*; University Science Books: Sausalito CA, 1997.

(47) Weininger, D. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(48) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011**, *3*, 33.

(49) OpenBabel package, version 2.3.1. 2013; `http://openbabel.org`.

(50) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. *Chem. Cent. J.* **2008**, *2*, 5.

(51) Stewart, J. J. P. *J. Mol. Model.* **2013**, *19*, 1–32.

(52) Klamt, A.; Schüürmann, G. *J. Chem. Soc. Perkin Trans. 2* **1993**, 799–805.

(53) Keeffe, J. R.; Kresge, A. J. In *The Chemistry of Enols*; Rappoport, Z., Ed.; Wiley: Chichester, 1990; Chapter 7, pp 399–480.

(54) Angyal, S. J. *Top. Curr. Chem.* **2001**, *215*, 1–14.

(55) Heathcock, C. H. In *Comprehensive Organic Synthesis*; Trost, B. M., Fleming, I., Eds.; Pergamon Press: Oxford UK, 1991; Vol. 2; Chapter 1.5, pp 133–179.

(56) Braun, M. In *Modern Aldol Reactions*; Mahrwald, R., Ed.; Wiley-VCH: Weinheim, 2004; Vol. 1; Chapter 1, pp 19–61.

(57) Schmitz, E.; Eichhorn, I. In *The Ether Linkage*; Patai, S., Ed.; Wiley: Chichester UK, 1967; pp 309–351.

(58) Seebach, D. *Angew. Chem. Int. Ed.* **1979**, *18*, 239–258.

(59) Gabel, N. W.; Ponnamperuma, C. *Nature* **1967**, *216*, 453–455.

(60) Schwartz, A. W.; de Graaf, R. M. *J. Mol. Evol.* **1993**, *36*, 101–106.

(61) Hopcroft, J. E.; Motwani, R.; Ullman, J. D. *Introduction to Automata Theory, Languages, and Computation*, 3rd ed.; Addison Wesley: Reading MA, 2007.

(62) Ravasz, E.; Somera, A. L.; Mongru, D. A.; Oltvai, Z. N.; Barabási, A.-L. *Science* **2002**, *297*, 1551–1555.

(63) Newman, M. E. J. *Networks. An Introduction*; Oxford University Press: Oxford UK, 2009.

(64) Shannon, P. *Genome Res.* **2003**, *13*, 2498–2504.

(65) Müller, D.; Pitsch, S.; Kittaka, A.; Wagner, E.; Wintner, C. E.; Eschenmoser, A. *Helvet. Chim. Acta* **1990**, *73*, 1410–1468.

(66) Lambert, J. B.; Gurusamy-Thangavelu, S. A.; Ma, K. *Science* **2010**, *327*, 984–986.

(67) Roy, D.; Najafian, K.; von Ragué Schleyer, P. *Proc. Nat. Acad. Sci.* **2007**, *104*, 17272–17277.

(68) Weber, A. L. *Orig. Life Evol. Biosph.* **2007**, *37*, 105–111.

(69) Eschenmoser, A. *Chem. Biodivers.* **2007**, *4*, 554–573.

(70) Powner, M. W.; Gerland, B.; Sutherland, J. D. *Nature* **2009**, *459*, 239–242.

(71) Powner, M. W.; Sutherland, J. D.; Szostak, J. W. *J. Am. Chem. Soc.* **2010**, *132*, 16677–16688.

(72) Blum, L. C.; Reymond, J.-L. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.

(73) von Lilienfeld, O. A. *Int. J. Quant. Chem.* **2013**, *113*, 1676–1689.

(74) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303.

## Supporting Information Available

Kinetic Selection Criteria for Formose Reaction; Reaction Products of the Formose Reaction; Major Sugar Formation Pathways of the Formose Reaction. This material is available free of charge via the Internet at `http://pubs.acs.org/`.