Combined Network and High Resolution Mass Spectrometry Analysis of the Formose Reaction Reveals Mechanisms for Emergent Behaviors

Romulo Cruz-Simbron[1,2*], Siddhant Sharma[1,3*], Aayush Arya[1,4], Jessica Ray[1], Alejandro Lozano[1,5], Jakob Lykke Andersen[6], Huan Chen[7], H. James Cleaves II[1,8,9*]

Affiliations

1 Blue Marble Space Institute of Science, Seattle, 98104 WA, United States
2 Department of Chemistry, University of Colorado Boulder, Boulder, 80309 CO, United States
3 Department of Biology, Ashoka University, 131029 Haryana, India
4 Institut für Physik, Johannes Gutenberg-Universität Mainz, 55128 Mainz, Germany
5 Department of Biomedical Data Science, Stanford University, Stanford, CA 94305 United States
6 Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark
7 National High Magnetic Field Laboratory, Florida State University, Tallahassee, 32310 FL, United States
8 Earth-Life Science Institute, Tokyo Institute of Technology, 2-12-1-IE-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
9 Department of Chemistry, Howard University, 20059, Washington, D.C., United States

Abstract

The formose reaction (FR) autocatalytically converts simple plausibly prebiotic feedstocks into molecules of biological interest, including ribose. Autocatalysis is a hallmark of life, thus various studies have explored the formose reaction with respect to the origins of life. The FR is robust under appropriate conditions, occurring readily at low temperatures from a variety of substrates, and has been implicated in the generation of meteoritic organic compounds. We explored the FR here using a combination of *in silico* modeling techniques and high resolution mass spectrometry. The models match experimental results well, and point to the FR being much more complex than previously modeled or measured, and help explain the FR's potential to generate homochirality and primitive compartments, both of which are also hallmarks of life, before the emergence of the complex directed molecular encoding suggested by the RNA World model. These results suggest the FR requires further study with regard to the origins of life, and its importance may lie in the way it enables and coordinates emergent chemistries, rather than the particular products it generates, such as ribose.

Introduction

The formose reaction (FR)[1] has attracted attention since its discovery, especially due to its potential relevance to the origins of life for reasons which reflect changing conceptions of what life fundamentally is and how life may have begun[2]. The FR converts plausibly prebiotic environmentally-supplied reactive zero-valent carbon in the form of HCHO into more complex compounds which are important for biology, for example aldonic acids that

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

can be a storage compound for ribose[3,4]. The FR may provide a way to encode information in the temporal states of molecules and produce molecules that create new phases[5]. The undirected mechanisms of the FR produce compounds of biological interest and may steer products in terms of abundance and homochirality, but there is a gulf between the stochastic steering so far evident in the FR compared with measurable evolutionary contingency.

The concept of entailment is useful here, which describes the idea that the products of complex reaction networks may depend on the efficiency of earlier stochastic reaction steps: early random fluctuations in a reaction network may govern and cause measurable distinct outcomes which may vary over repetitions of the same experiment: for example in the form of different product distributions or different enantiomeric excesses in products. The evolved entailment of complex biochemical reaction networks is highly deterministic compared to what the FR creates, but the FR, by virtue of its complexity, creates opportunities for the emergence of highly entailed outcomes. We explore here whether the crossover between loosely entailed chemical reactions and highly entailed chemical reactions is discernible by examining the potential and observed products of FR reaction products.

The FR converts formaldehyde and an ancillary precursor reactant (generally an enolizable aldose) into a variety of sugars and sugar-related products[6]. The FR has been implicated as having been important for the origin of ribose for RNA World origins of life models[7], and as a source of compounds detected in carbonaceous meteorites[8,9]. The interpretation of the mechanism of the reaction and its import has been discussed for decades[10], but previous discussions have likely missed important nuances of how the reaction proceeds, and it is important to bring new analytical methods to bear to understand how the reaction proceeds and what its outcomes are[11].

We present here a combined experimental and computational analysis of the FR that helps explain its course and the diverse products detected. The perception that the reaction is "prebiotic," implying it produces contemporary biological compounds rather than biological-like processes, has led many researchers to focus on its ability to produce biological sugars, especially ribose, while ignoring the diversity of other products the FR produces[12]. This perception has led research teams to explore ways to make ribose preferentially over other products[13], and sometimes to refer to all "undesired" products of the reaction as "tar"[14], and seek ways to avoid those outcomes[15].

We argue here that this is the wrong way to examine the FR, and that the FR may in fact create the correlated aspects of a chemoton or protocell which have been argued as having been important for the emergence of life previously[16].

Focusing on the production of particular compounds of modern biochemical importance may be the wrong way to understand the OoL, which depends on the cohesive self-amplification of initially random processes. We suggest analyses of the FR should be refocused on analysis of the coordination of processes in terms of compounds and phenomena.

Butlerov discovered that sugar-like substances are produced when formaldehyde (HCHO) is treated with base[1]. The FR is loosely understood as a base-catalyzed autocatalytic condensation of HCHO to yield a complex mixture of carbohydrates, among other compounds[17]. Breslow[18] and Socha *et al.*[19] generally explained how a complex product mixture results when carbohydrates containing an α-hydrogen react with HCHO, allowing for concatenated condensation and elimination reactions to occur via various reactions which

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

are promiscuously activated by the reaction conditions. Previous analyses have focused on the formation of low molecular weight molecules, but it is clear the FR, depending on initial conditions, creates a complex set of products, with a wide range of masses.

The FR does not require the presence of Ca^{2+} or a strong base, nor does it depend on the presence of contaminating carbohydrates like glycolaldehyde (GA)[7,20]. The origin of a complex mixture of carbohydrates from the FR in the absence of a priming aldol catalyst and the generation of a "priming sugar" by a non-obvious conversión of HCHO to glycolaldehyde via an unknown and slow mechanism remains a problem in understanding the FR[21]. Caramelization in the FR remains an enduring phenomenon which also requires explanation.

The chemical diversity generated by the FR may depend on the input species, but since the FR is a collective complex set of reactions, and contains numerous common species and reactions, many starting reagents may produce similar outcomes. Since the reaction is autocatalytic, a key question becomes how "priming compounds" are generated.

Breslow[18] proposed a mechanism by which GA can be created by an autocatalytic cycle, which would place GA in a central location as an enabling molecule of the FR, though other mechanisms might enable the production of GA (Figure 1).

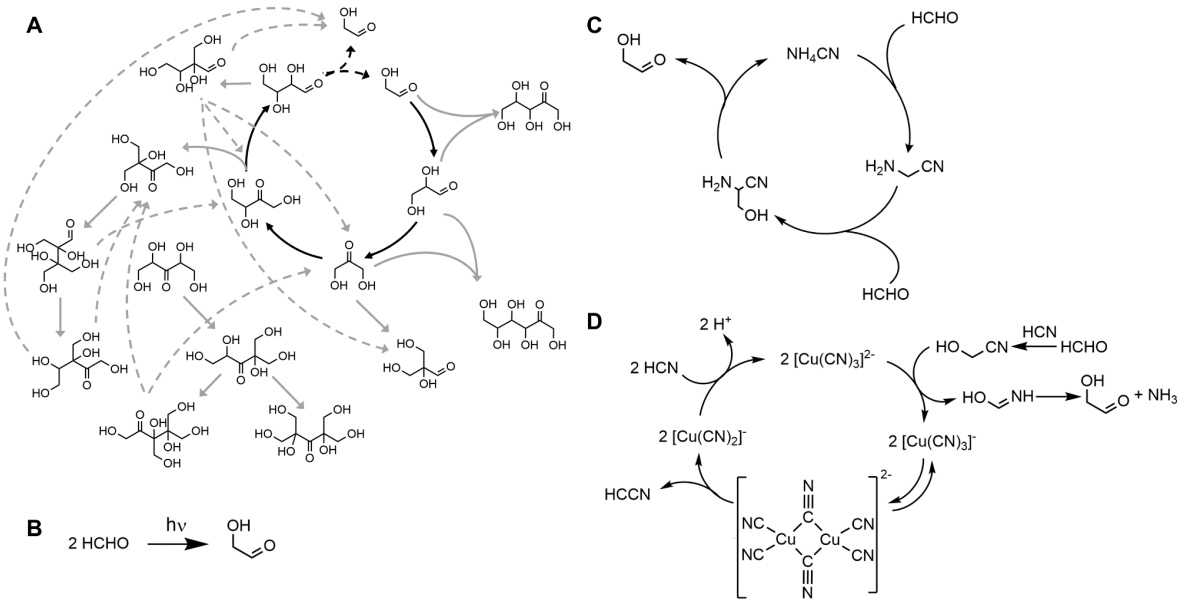


Figure 1. A. Autocatalysis in the FR according to the Breslow[18] (black lines) model overlain with a postulated FR reaction network adapted from[17] (grey lines). Solid arrows represent aldol condensations with HCHO, dashed lines show retro-aldol reactions. B-D suggested reactions for generating GA from HCHO: B. using UV light, C. from the Michael addition of HCHO to amino acetonitrile, and D. via cyclic UV photocatalytic reduction of glycolonitrile[22].

The condensation and polymerization of GA produces tetroses, pentoses, hexoses, heptoses and octoses, along with larger molecules, all of which may act as reactants for numerous new reactions[21]. The GA-initiated cycle is likely but one of many autocatalytic reactions which wax and wane as reactant concentrations change during the course of the reaction. The autocatalytic cycles described by Breslow[18] and Huskey and Epstein[23] are likely only snapshots of how the reaction proceeds.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Huskey and Epstein[23] compared the homogeneous FR under continuous-flow stirred-tank reactor (CSTR) conditions and batch conditions. Their observations and kinetic simulations led them to conclude that catalytic hysteresis loops observed in the CSTR-formose reaction do not arise from chemical instabilities, but rather from chemical reactions that are slow relative to the time scale of the experiment. However, they later constructed new models showing bistability and built a model for the autocatalytic phase of the FR consistent with experimental results (Figure 1A). The aldotetrose mechanism proposed by Breslow[18] as the source of autocatalysis is not part of this mechanism, but other molecules may help seed the retro-aldol reactions which help this reaction to be diversity-generating. It is thus possible that the mechanism of the reaction involves the operation of numerous concurrent autocatalytic cycles, and this may better explain the observed product distribution[24]. While it may be that there are formal intermolecular catalytic interactions in the FR (*e.g.*,[21,25]), the important point is that the reaction network itself appears to be self-propagating on many levels, and in principle could operate indefinitely in a et alflow-through regime. It may thus be useful to recognize the FR's unique inherent complexity to better understand it both experimentally and theoretically[26].

Using 2-hydroxymethylphenylboronate and deuterated substrates Ricardo *et al.*[27] further clarified the mechanism of the FR by showing that the enol intermediates react quickly with HCHO while it is abundant in the early stages of the reaction. This conclusion was based on the observation that the reaction carried out in D_2O with *in situ* generated $Ca(OD)_2$) doesn't produce heavily deuterated C6 or C7 compounds (only 10 % of the C7 species were deuterated), whereas extensive deuteration would be expected if retro aldol reactions of tetrose to give two GA molecules were a major feeder reaction. After most of the HCHO is consumed, widespread deuteration becomes evident.

Shigemasa *et al.*[28] conducted a photochemical FR using UV irradiation. They found that 2-hydroxymethyl glycerol was the major product under such reaction conditions which they proposed was produced via the condensation of glyceraldehyde with FA to produce 2-formyl glycerol, followed by a cross-Cannizzaro reaction. Importantly, 2-formyl glycerol is largely inert to the major operative reactions of the FR and thus accumulates. Of the intermediates detected by Shigemasa *et al.*[28], only 2-formyl glycerol is present in Epstein's mechanism[23], and none of Shigemasa et al.'s suggested intermediates are present in Breslow's mechanism[18]. It is therefore probable that there is a network of chemical reactions which propagates the FR which are not well represented by various previous minimal schemas of the FR.

The FR has also been carried out under hydrothermal conditions[29-32]. Using 0.5 M HCHO solution at 200 °C and 100 bar in a continuous flow reactor, Kopetzksi *et al.*[32] reported that contrary to the reaction at low temperature, hexoses were formed in low concentration and pentoses were formed in high yield compared to lower temperature reactions. 2-Deoxyribose and other deoxy compounds were detected in these reactions, but little effort was made to relate the yields of these compounds to reaction temperature. Additionally, the induction period does not occur under these conditions, which points to a larger issue, that while the FR is robust and obeys Arrhenius behavior, without the required condition that the starting materials be sufficiently concentrated, the reaction does not occur at all, and instead, the reactants degrade via pathways such as the Cannizzaro reaction.

Weber[33] argued based on kinetic and thermodynamic principles that under mild aqueous reaction conditions, the FR may in fact be the only one-carbon condensation

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

reaction capable of generating complex organic compounds for the origin of life. Under neutral pH conditions, HCHO solutions can also produce a series of low molecular weight polyoxymethylene (POM) polymers (of formula $HO(CH_2O)_nH$)[34]. A five-weight percent (~1.3 M) aqueous HCHO solution contains ~14% POM dimer, 3% trimer, and 0.5% tetramer[34], and such compounds are suspected to be major components of comets[35]. Under the basic conditions which allow the FR to proceed, a complex mixture of carbohydrates can arise even at concentrations as low as 10^{-3} M[3]. Nevertheless, it may only be possible to reach such concentrations in specialized early Earth environments[36].

Solid phase reactions offer another way to increase selectivity in the FR. The FR also proceeds under solvent-free conditions by grinding the reagents together with a solid base. Haas *et al.*[37], and Vinogradoff *et al*.,[38] showed that the formose reaction in the solid state achieves a better selectivity in terms of the production of biological sugars and stability of the sugars produced.

Although ribose can be synthesized by polymerizing HCHO in the FR, its yield is low, and ribose is relatively unstable in aqueous solution[4]. According to Georgelin *et al.*[39] ribose polymorphs are also unstable in aqueous medium so that a stabilization mechanism has to be invoked. Dass *et al.*[40] studied the isomerization of ribose by NMR at different temperatures. Using thermodynamic parameters obtained from NMR experiments, Dass *et al.*[40] showed the enrichment of the more unstable β-furanose isomer in an out-of-equilibrium environment was supported by stationary temperature gradients.

Computational techniques offer an attractive way to understand and predict the products of complex diversity-generating reactions such as the FR[41-45]. These computational models produce what we have termed complex reaction network representations (CRNRs[46]), which can help explain the compositions of complex reaction network products. The products of these models can be vetted via their correspondence with experimental data[47]. Though various analytical techniques have tended to focus on biologically relevant compounds, it is clear the molecular diversity generated by the FR is much higher than typically reported, and the reaction likely proceeds in subtly or significantly different ways under different reaction conditions due to variations in the relative rates of the many underlying reaction mechanisms that comprise it.

We here present a generalized reaction “map” of the FR using graph theory-based simulation, supported by high resolution Fourier-Transform Ion Cyclotron Mass Spectrometry (FT-ICR-MS). Our comprehensive FR model helps explain the FR’s major and minor products as reported in previous literature reports and predicts additional products and potential phase-changes that are either readily observable or potentially measurable .

Results

FR Product Diversity

The FR involves the repetitive application of a small set of reaction mechanisms. These transformations must be comparably energetically favorable since the reaction is typically run at relatively low temperatures but may produce many isomeric products that simple one-dimensional MS analysis cannot discriminate, as there may be many isobaric products.

In contrast to the derivatized mass spectra reported in Decker *et al.*[48] and Ricardo *et al.*[27], the direct injection ESI-FT-ICR-MS spectrum suggests considerably more product diversity which warrants mechanistic explanation. This is in accord with Decker *et al.*’s[48]

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

observation that some 80% of the input HCHO in a FR is converted into non-volatile and/or non-derivitizable products. The *in silico* reaction simulation was limited by an arbitrary MW cutoff of 200 amu to economize computational load. This allowed the simulation to form carbohydrates up to hexoses, though the experimental spectra show that much higher MW products are generated in the FR (Figure 2A). The FR evidently also allows for the formation of a large number of side-products as evidenced by the ESI-FT-ICR-MS spectra presented in Fig. 2A.

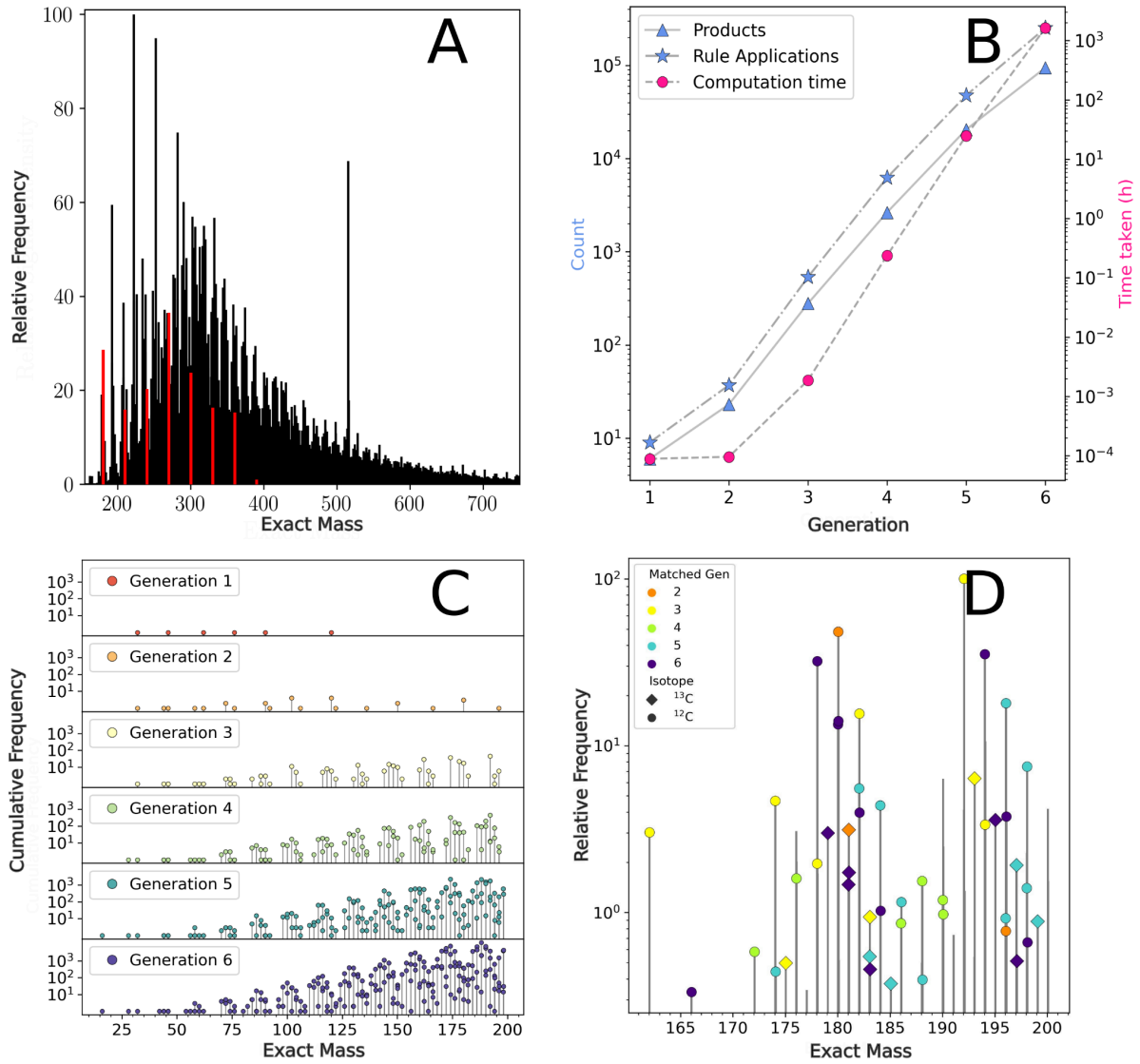


Figure 2. A. Negative mode FT-ICR-MS spectrum of the products of an exemplary FR (conducted at 150° C) over the m/z range from 150 to 750. Note that this MS analysis method has a minimum MW cutoff of ~ 150 amu. Overlain in red are the peaks corresponding to the common carbohydrate formula $(CH_2O)_n$, where n = 6-13. B. The number of new products (blue triangles) and new rule applications (blue stars, y-axis), and the computation time per generation (red circles, right y-axis) in the *in silico* FR simulation. The presented data are not cumulative. C. Lollipop plot of the mass distribution and isomer redundancy produced by the FR CRNR as a function of reaction generation. The Y-axis scale shows the cumulative frequency of unique isobaric constitutional isomers generated in the

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

CRNR. D. The *m/z* 150-200 region of the ESI-FT-ICR mass spectra of FR experiments conducted at 150 °C with mass matches (46 in number) color coded by their generation of first appearance in the *in silico* network. The FR products in Generation 1 (G1) do not reach the 150 amu cutoff. 52 mass matches are seen in the *in silico* network for the FR experiments conducted at 85°C (see Figure SI1). The apparent presence of more than one dot per bar is due to the close spacing of the exact masses of products in some mass regions. Figure S12 shows a closer view of the concordance of the model and measurement.

Running the *in silico* model through six generations produced 94,415 unique constitutional isomer structures and took ~ one month to compute using our computational resources (Table SI1). These computations could likely be considerably improved using more powerful computational resources.

While analytical measurement is the gold standard for determining the abundance of products in a reaction at any given time point, presumably the balance of synthetic pathways which produce a compound and subsequent reactions which consume them govern the abundance of measurable compounds. The computed six generation FR network as a function of modeled reaction generation is shown in Figure 3, where node size represents the in-degree of the network (representing reactions which produce a given compound) or out-degree (representing reactions which consume a given compound). It can readily be appreciated how the FR may produce thousands of products, as evidenced by the FT-ICR-MS spectrum shown in Figure 2A, though potentially in highly variable concentrations.

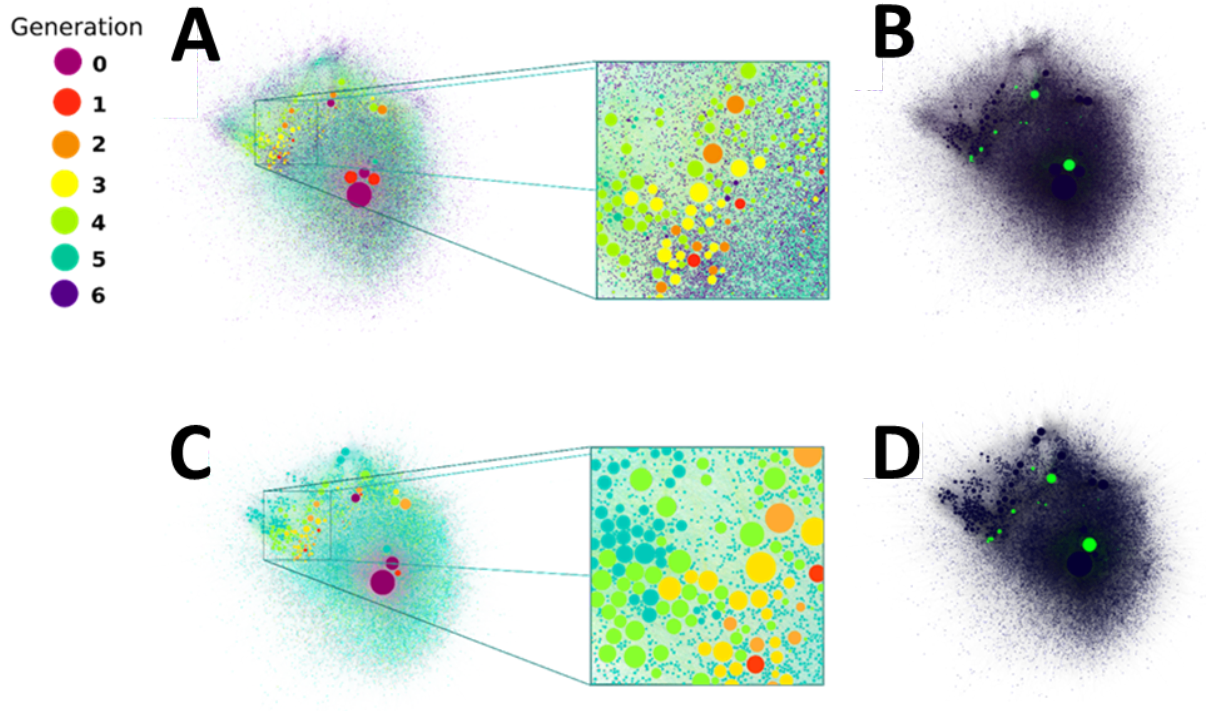


Figure 3. Representation of the FR CRNR after six generations with compounds colored by their first generation of appearance. Node size is proportional to the *in-degree* of compounds. C. A comprehensive visualization of the computed FR network after six

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

generations colored by the *out-degree* of each compound. Zoomed insets in A and C show the fine-scale structure of the FR CRNR, demonstrating the large number of potential by-products which may contribute to the CRN's overall compositional diversity. The heavy weighting of the connections between the 5th and 6th generation products is evident. B and D show the species detected by Omran *et al.*[49] in green.

The FT-ICR-MS data and the CRNR both suggest the FR generates a complex product suite, which is considerably more complex than what has been reported using NMR or GC-MS analysis[50]. The correspondence between the compound structures predicted by the CRNR and the measured MS spectra can be compared using various MS data transforms. For example, Kendrick plots allow normalization using a common mass to show relationships among related series of species, and Van Krevelen diagrams plot product elemental ratios (for example H/C ratios vs. O/C ratios), and offer another way to visualize and interpret complex mass spectra[51]. Figure 4 shows a comparison of the CRNR and FT-ICR-MS FR data superposed and compared using Kendrick plots and Van Krevelen diagrams.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

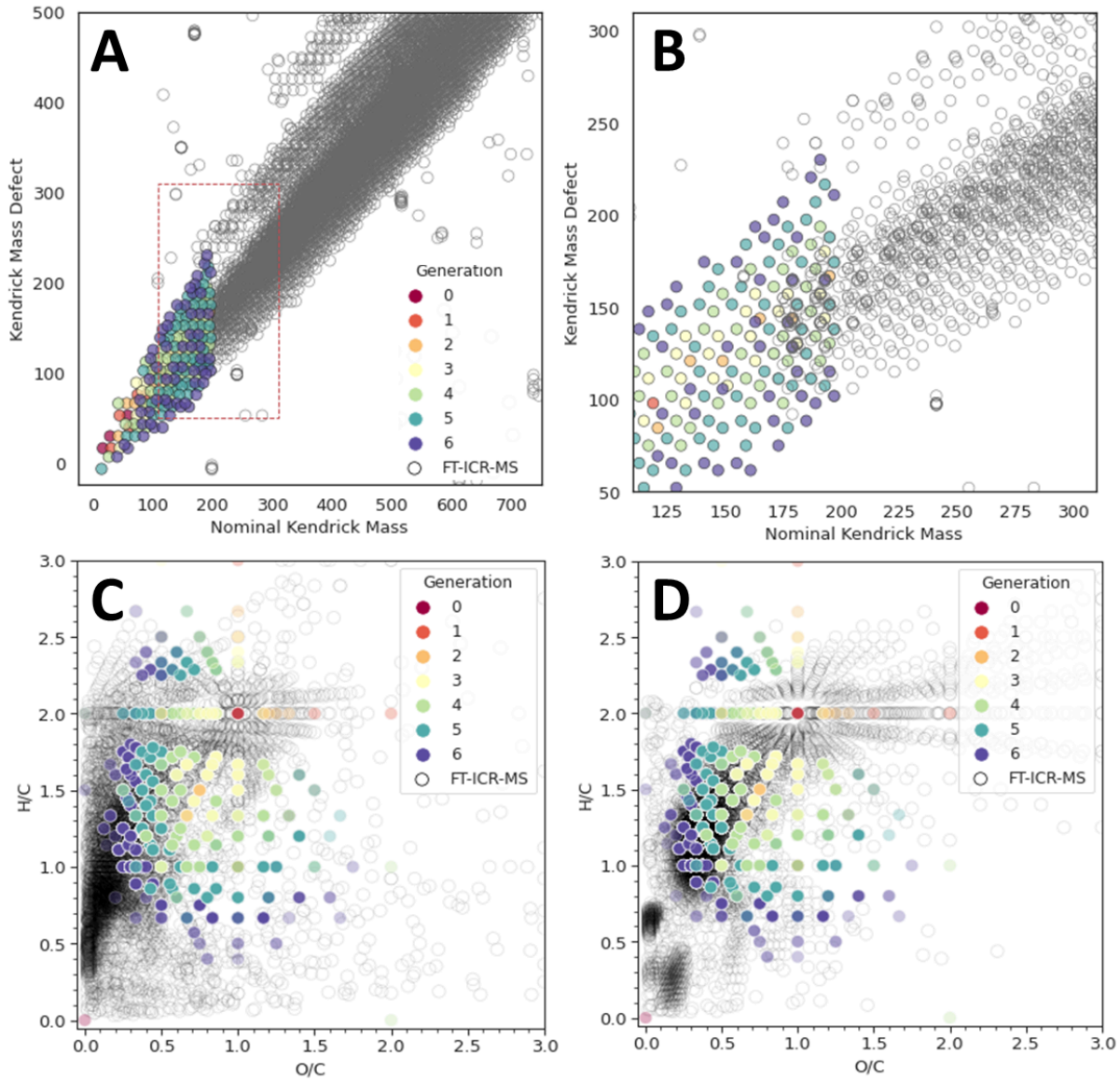


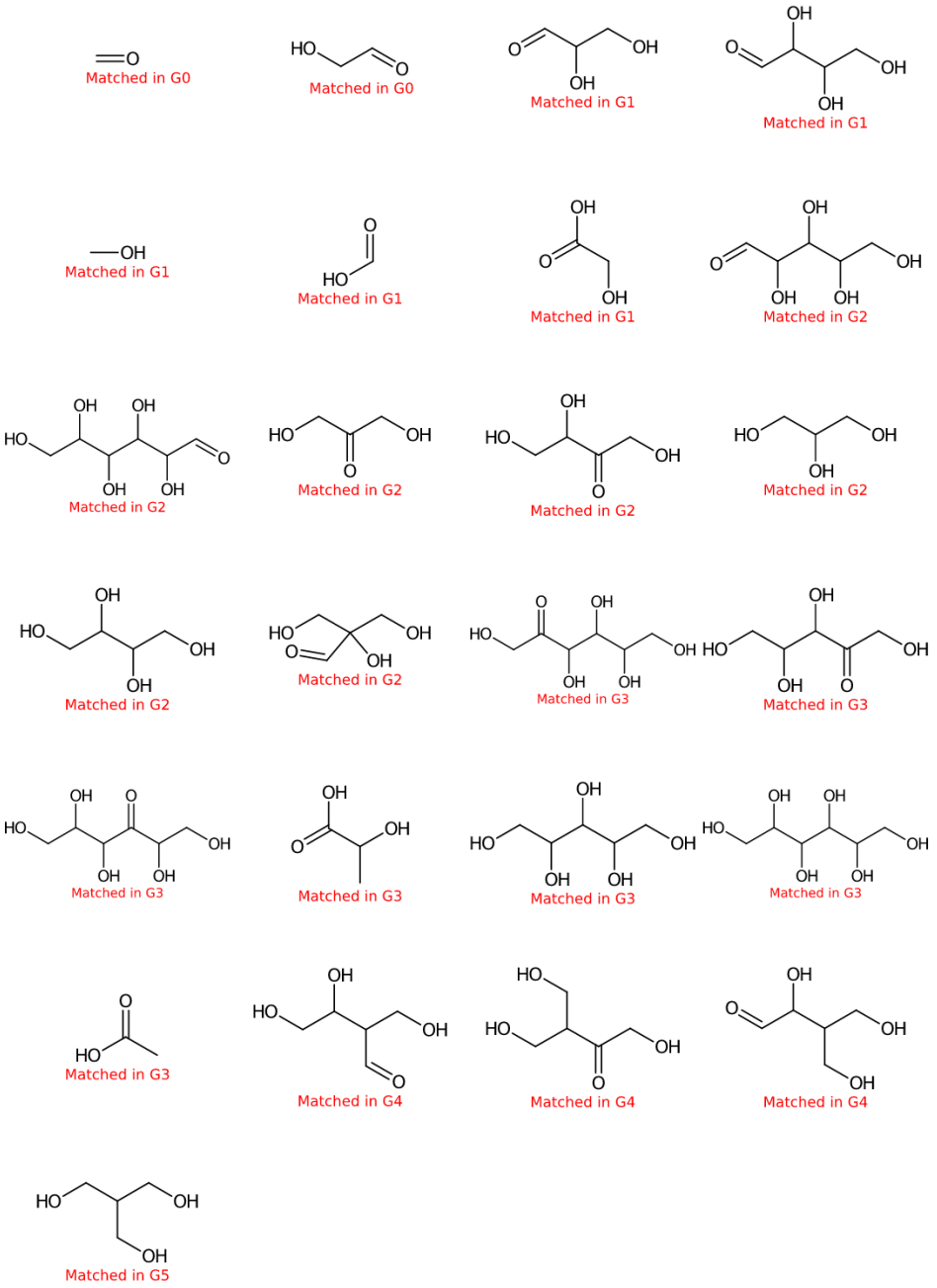
Figure 4. Modeled negative ionization mode Kendrick plot of the FR CRNR (open grey circles) overlaid on measured FR products produced at 150°C as measured using negative mode ESI-FT-ICR-MS (colored by the first generation of appearance in the CRNR). (A) The modeled FR CRNR extends from m/z 14 to 200 while the measured FT-ICR-MS data extends from m/z ~150 to 750. (B) A zoomed view of the box in (A) in which the overlap of the CRNR and measured data is more clearly evident (where colored dots fall inside gray circles). The effect of experimental conditions on the product suite of FR can be seen by comparing the data presented in the Van Krevelen diagram of the computed formose products at 85°C with the Van Krevelen diagram of the computed formose products at 150°C (D). The X and Y axes in C and D denote the atomic ratios of H/C and O/C in the products.

While the CRNR is truncated at *m/z* 200, it can be seen in the Kendrick plot in Figure 4A that the general trend in the experimental data is matched well by the model and in Figures 4A and 4B that the CRNR reproduces the spur of data located above the main diagonal data cluster, and in Figure 4B that there is considerable, but imperfect, overlap between experimental and computed data. The imperfect overlap in Figure 4B may be

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

partially attributed to the inability of the model to take kinetics and yield into account, as well as measurement bias. In the Van Krevelen comparison, while the model has considerable overlap with the measured data, the measured data also produce unique dense patches of data points in the lower left region of the plots, as well as more diffuse patches towards the right of the modeled data, and these measured data show considerable structure. These patches may be attributable to species measured above the MW cutoff of the computational model.

100% of the test set compound masses from the data provided by Decker *et al.*[48], and Omran *et al.*[49] were matched after six generations in the FR CRNR. The complete list of matched compounds annotated by their generation of appearance is presented in Figure 5. The correspondence between the CRNR model and experimental observations shows that the model accurately predicts detected compounds.



https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Figure 5: The compounds identified in Decker *et al.*[48] and Omran *et al.*[49] matched by the computed FR network as a function of generation. Figure was generated using the CDK Depict tool: https://www.simolecule.com/cdkdepict/depict.html

FR Product Properties and Complexity

As the products of the FR increase in number, and as their MW increases, the diversity of their chemical properties also increases. This diversity can be estimated using numerous metrics, including the Hybridization Ratio (Figure 6a), and Estimated Molecular Complexity (Figure 6b), which were estimated using chemoinformatic techniques[52]. As would be expected, as product molecules grow larger, the range of property values they can explore grows. While HCHO and GA are particularly low flexibility molecules, carbohydrates in general are often high flexibility molecules, thus the products of the FR also grow into a region that produces very "floppy" molecules.

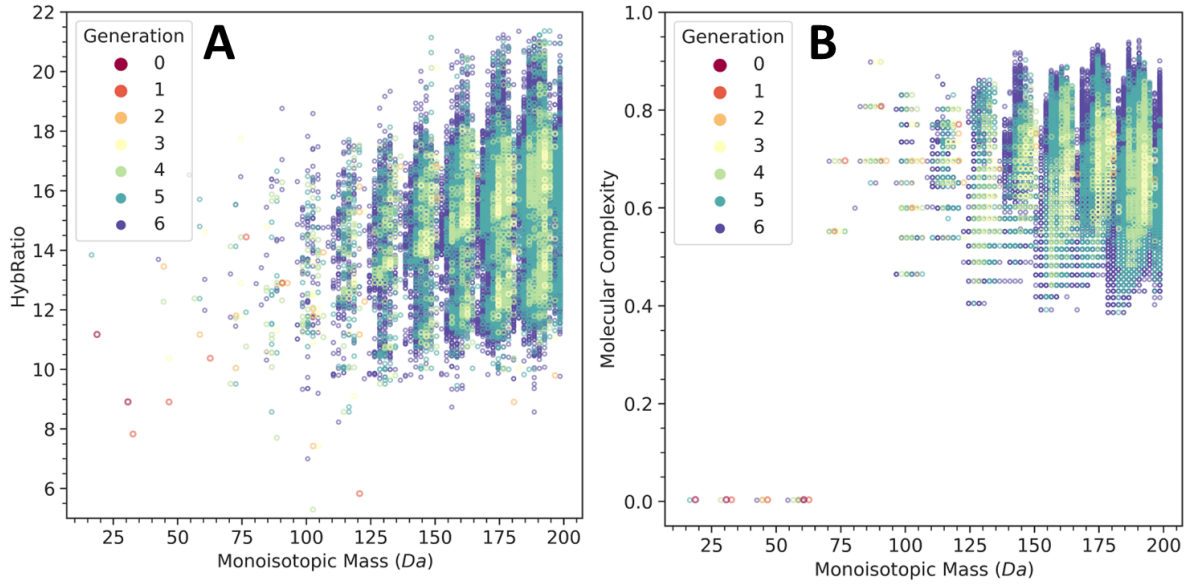


Figure 6. A. The generational diversity of HybRatio as a function of generation in the FR CRNR. HybRatio is defined as the ratio of sp^3 hybridized carbon atoms in a molecule over the sum of sp^2 and sp^3 hybridized carbon atoms ($Nsp^3/(Nsp^2 + Nsp^3)$)[53]. B. Estimated Molecular Complexity as a function of mass and generation in the computed FR products[54].

The progress of the FR likely results in changes in the hybridization ratio of the products, mainly due to elimination and cyclization reactions which reduce the flexibility of molecular bonds, defined as the ratio of sp^2 to sp^3 bond hybridization. One might thus expect the FR to both explore reactions which make floppy molecules as well as more rotationally inflexible ones (see Figure SI3).

Numerous metrics have been proposed for classifying and hierarchically organizing the complexity of chemical compounds (*e.g.,*[55]). Various chemometric algorithms exist for the computation of chemical complexity, which consider metrics such as the number of atoms in a molecule, their connectivity, and properties such as symmetry which allow for the computational compression of their graphic representation. Figure 6b shows the estimated evolution of the chemical complexity of FR CRNR products as a function of reaction generation. There is a general trend for generational product cohort complexity to rapidly

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

increase from the starting materials, then decrease among all generations while allowing for the complexification of rarer species. Notably, there is a small cohort of very low complexity and molecular flexibility species generated in the network, most of which are key species implicated in prebiotic synthesis (see Figure SI4).

Phase Separation via the FR

The previously discussed FR maturation predicts novel mesoscale properties of FR products. Thermal decomposition reactions of carbohydrates are known to produce new phases[56,57], which sometimes resemble those observed in carbonaceous meteorites[58]. This *in silico* exploration of the FR independently suggests that such phases form during the course of the reaction as reaction products become new phase-forming products as a function of their mass, cLogP, and estimated aqueous solubility.

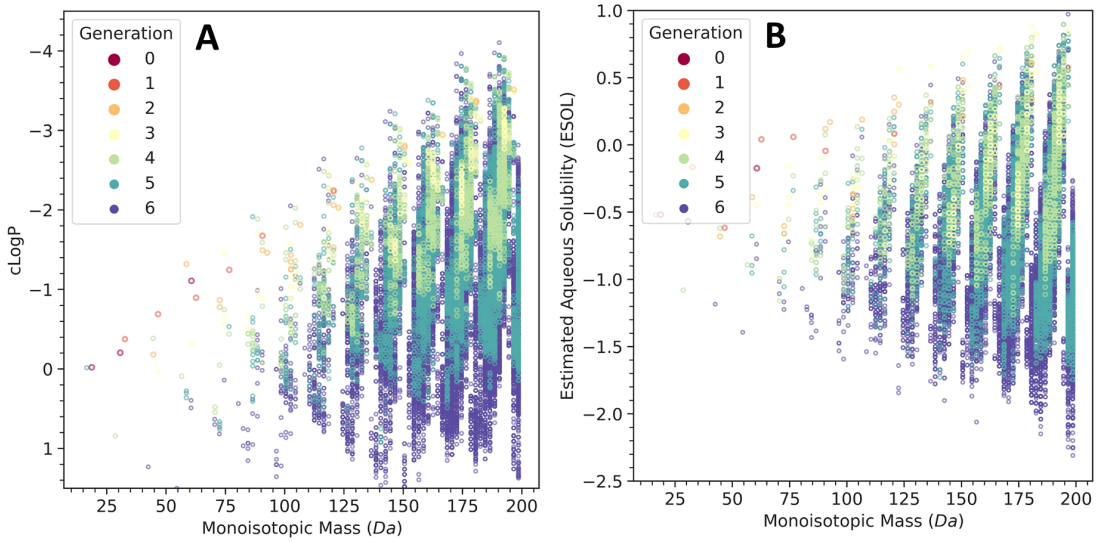


Figure 7. A. Computed cLogP values as a function of molecular mass, colored by the generation in which each species is first produced. B. Estimated Aqueous Solubility (ESOL) in mol/L as a function of mass and generation in the computed FR products.

cLogP is a proxy for the tendency of compounds to partition between hydrophobic and hydrophilic phases, and potentially form new phases. As can be seen, the FR CRNR produces more hydrophobic species as it evolves (Figure 7a). It might be expected that as general HCHO-derived structures dehydrate, they become more hydrophobic and thus may self-aggregate due to hydrophobic, or other, effects. ESOL (Estimated SOLubility) is a measure of the estimated aqueous solubility of a molecule inferred from its structure based on a QSAR (Quantitative Structure-Activity Relationship) regression model calculated from logP, molecular weight, the proportion of heavy atoms in aromatic systems, and the number of rotatable bonds[59]. Figure 7b shows the estimated evolution of the ESOL trends of the products of the FR CRNR. Higher ESOL values indicate higher aqueous solubility. FR products evolve towards both lower and higher ESOL values as a function of reaction generation in the CRNR.

Hydrogen-bonded self-organization may also play a role in the formation of observed self-organizing structures during the course of the FR. The tendency of FR-derived structures

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

to form cell-sized structures (Figure 8) is often noted, this is likely a scale under which various molecular forces act in concert to drive self-assembly[60]. This adds novel commentary to the discussion of the presence of cell-sized organic structures which have been observed in carbonaceous meteorites for 60 years (*e.g.*,[61]), and sometimes interpreted as biogenic (*e.g.*,[62]).

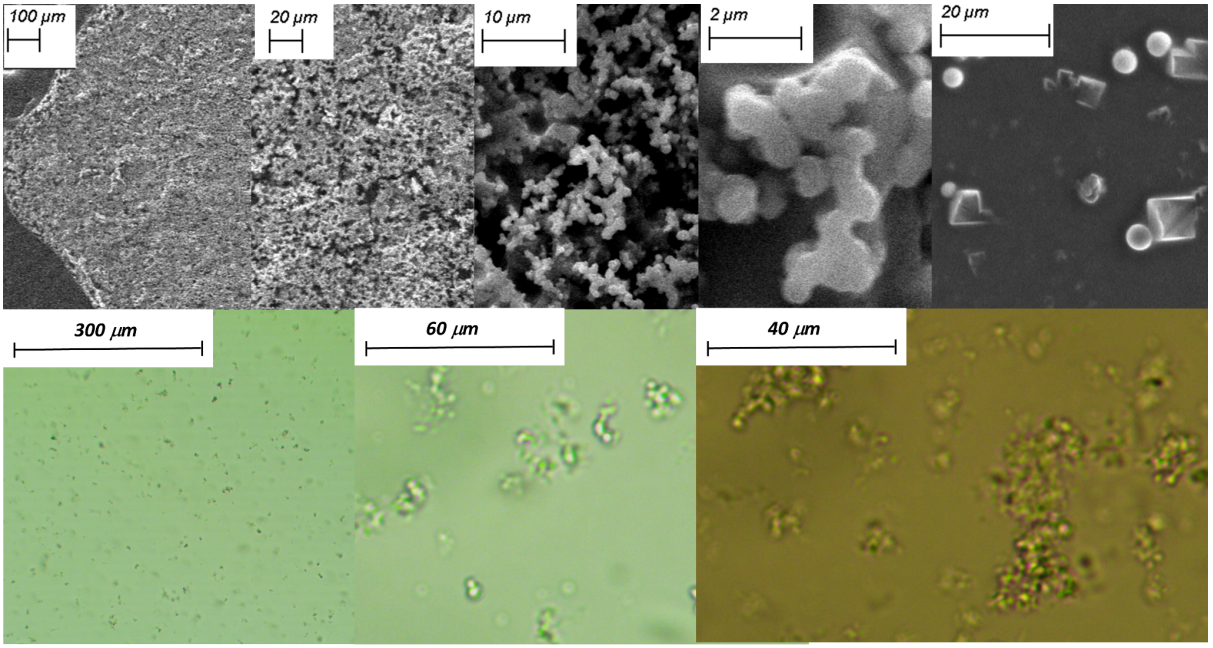


Figure 8. SEM (top row) and brightfield (bottom row) images of structures formed from the reaction of 1 M HCHO with 10 mM GA in the presence of $Ca(OH)_2$. The top row shows progressive zooms of simple dried solutions, the right-most image shows the same reaction diluted 1/100x with milliQ water and then dried. The bottom image series show that aggregates are common and not necessarily mutually attractive as they are often dispersed, and that furthermore the structures are not salt crystals since the images show the soluble fraction. They have similar size and morphology as the structures shown in the top row indicating the same structures are stable to the vacuum drying required for SEM imaging. In the rightmost top row, spherical structures were shown using Energy-Dispersive X-ray Spectroscopy (EDAX) to be C-rich, indicating they are carbonaceous, while the angular products were shown to be NaCl crystals (see Figures SI7 and 8).

We further investigated the morphology of the structures produced in these reactions using scanning electron microscopy (SEM), and Energy-Dispersive X-ray Spectroscopy (EDAX). The data (Figures SI7 and 8) suggest the structures observed are derived from FR solids rather than from the inorganic salts resulting from the reaction and its drying for SEM analysis, though there is some variation in the correspondence between morphology and composition, which deserves attention at this scale for this type of analysis.

SEM analysis shows that the FR produces regular spherical particles with diameters on the order of less than one to a few μm. Undiluted and dried directly on SEM stubs, the self-organized products are imageable by SEM in the dry state. When diluted, dried and visualized using SEM products, singular spheres are visible, suggesting their structural

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

cohesion. Observed directly in an aqueous solution using light microscopy, they appear as both singular and aggregated spheres.

EDAX analysis allows for elemental composition assignment to the observed morphologies. Multimicron-scale inspection of the product materials reveals several morphologies, dominated by dendritic growths composed of 1-2 um diameter spheroids which have EDAX-determined compositions compatible with proposed formose reaction mechanisms, but also showing evidence of various salts (NaCl, $CaCl_2$, etc.) likely produced during drying on the SEM sample stubs. Salt crystals with readily identifiable elemental compositions are observable.

It would be expected that insoluble formose solids that include carboxylic acid groups, potentially derived from Cannizarro reactions, would be able to chelate cations such as Na^+ and Ca^{2+} present in the reaction mixture. Carbonate salts should have EDAX measured Na:C or Ca:C ratios ranging between 2:1 or 1:1, while the pairing of observed morphologies and element ratios is often much lower (Figure SI8), suggesting the observed regions in the spherical structures are dominated by organic chelates.

Figures SI7 and 8 present analyses of typical formose reaction products created here analyzed using SEM and EDAX. Three types of structures can be readily identified: square crystals, isolated spherical units and dendritic structures. Dendritic and square crystals are easily identified as NaCl crystals using EDAX (see spectra 76 and 80, respectively in Figure SI8 below). The isolated spherical structures (see spectra 72 and 77, Figure SI7), in addition to their unique rounded morphology, these display a roughly 1:1 C:O ratio typical of carbohydrate-like material.

The Generation of Stereochemical Coherence in FR-related Reactions

While the basic implementation of the FR CRNR here does not specifically consider stereochemistry, if the FR plays a role in the organic complexification of meteoritic organics, there may be nodes in the network which could amplify stereochemistry and make it more coherent over time. The raw CRNR does not constrain stereochemistry, which, depending on model implementation, could amplify or dampen enantiomeric excesses (Figure 9). Opportunities for stereochemical enrichment are widely distributed in the CRNR.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

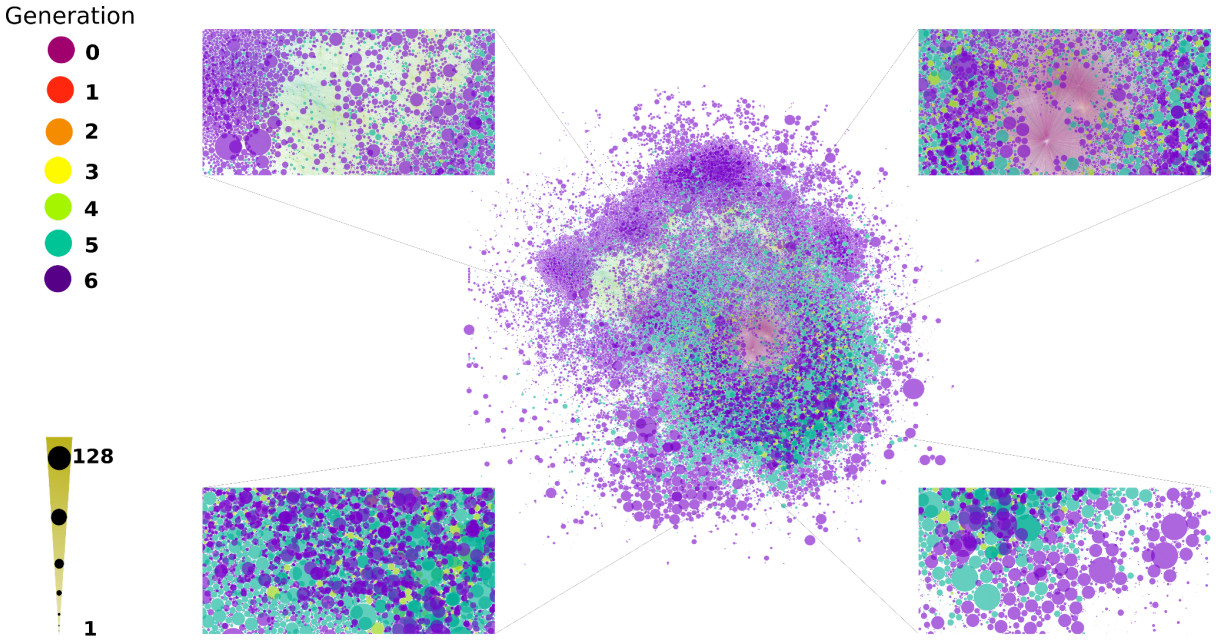


Figure 9. The number of potential stereoisomers as a function of reaction generation of first appearance (upper left color scale) for the computed FR network. Node size is linearly proportional to stereoisomer number (lower left size scale). A Circle Pack Layout in Gephi with generational hierarchy as an attribute was used for this figure, and the scale size and magnifications are consistent for the zoomed out portions.

To our knowledge, enantiomeric enrichment has not yet been investigated experimentally in the FR., and there may be an excellent opportunity to understand how small-scale causes create large scale effects. Many studies assume FR products are essentially racemic (*e.g.*, [48]). However, studies of meteoritic carbohydrates suggest there is an increasing stereochemical concordance among detected polyols which increases with increasing MW[8]. A similar phenomenon has been reported with respect to non-racemizable amino acids (*e.g.*,[4]), suggesting network effects exist which enhance coherent stereochemistry in meteoritic organic chemistry[63]. Why these observed stereochemical coherences seem to steer in the same direction as the terrestrial stereochemical bias remains to be vigorously explained. As can be seen in Figure 9, and as would be expected, as molecular products become larger as a function of generation time, the potential for the numerosity of stereoisomers increases. It can reasonably be assumed that early steering and kinetic feedbacks which selectively amplify certain stereoisomers could significantly influence the stereochemistry of higher MW species[64,65].

Autocatalytic Motif Search

Using the imperative approach detailed previously[47], possible autocatalytic cycles for which all reactions were spontaneous according to thermochemical calculations were searched for using eQuilibrator API[66] under the basic conditions which allow the formose reaction to proceed (*e.g.*, for which the free energy for each reaction, ΔG_{rxn}, is negative at pH 10). We restricted the search to cycles that don't include formaldehyde or methanol as cycle components (but which allow these molecules to be feeders or products) because these molecules are among the most connected molecules. Figures SI5 and SI6 show the

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

distribution of ΔG_{rxn} vs the total number of reactions as a cumulative function of generation. We found more than 5600 thermodynamically spontaneous cycles for catalytic nodes containing glycolaldehyde, 1,3-dihydroxyacetone, glyceraldehyde, 2,3,4-trihydroxybutanal, glycolic acid, propanedial, pyruvic aldehyde, 3-hydroxybutanal, 3-hydroxypropanal, hydroxyacetone, 3-oxopropanoic acid, pyruvic acid and glyoxal. Some catalytic nodes do not show autocatalytic cycles, but it is possible a more intensive search could yield still more autocatalytic cycles. Four examples of autocatalytic cycles are presented in Figure 10. Each cycle is accompanied by a histogram that displays the frequency of similar spontaneous cycles of different cycle lengths.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

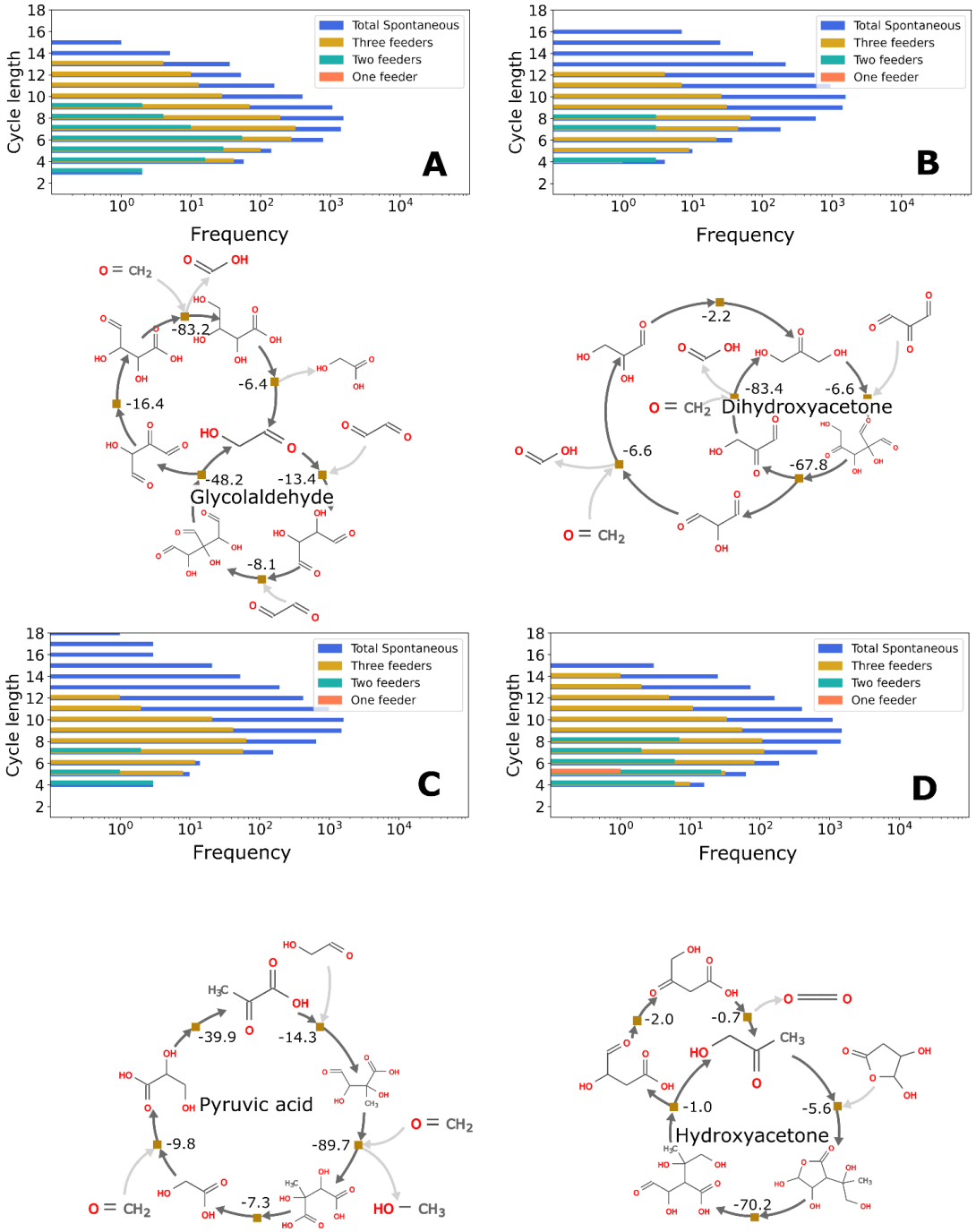


Figure 10. Representative topologically identified autocatalytic cycles that were found in the FR CRNR fed by (A) glycolaldehyde, (B) dihydroxyacetone, (C) pyruvic acid, and (D) hydroxyacetone. Bar charts show the frequency distributions of spontaneous cycles, in which each reaction has a negative ΔG_{rxn} at pH 10 as computed using eQuilibrator API[55], as a function of cycle length, and showing the contribution by cycles containing one, two or three unique feeder molecules. Specific examples of autocatalytic cycles that use the featured feeder molecules are shown, the numerical values superposed on the reaction diagrams represent the computed ΔGr_{rxn} in kJ mol^{-1}.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Network Hub Species

The modeled FR CRNR allows tracking of the most highly connected, or "hub," species in the reaction graph, which may be pivotal species in real-world chemistry. The most highly connected in-degree molecules are water, formic acid, methanol, formaldehyde, acetaldehyde, CO_2, glycolaldehyde, acrolein, glyoxal and methylglyoxal, in that order (Figure 11 top). The highest connected out-degree node molecules are water, formaldehyde, acetaldehyde, acrolein, glycolaldehyde, CO_2, propionaldehyde, acetone, methane and crotonaldehyde (Figure 11 bottom). Notably, with the exception of crotonaldehyde which has E/Z isomers, none of these species contain stereocenters, though most of them offer an opportunity for stereospecific addition or condensation reactions.

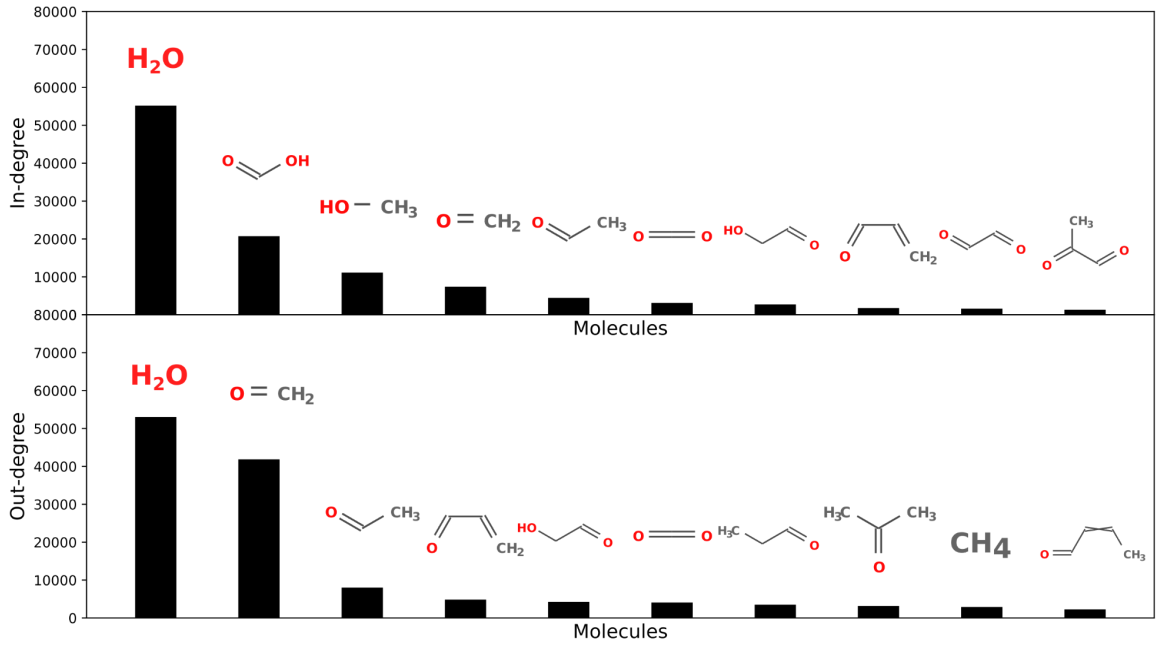


Figure 11. Water is the most connected component over the six reaction network generations explored here with respect to both in- and out-degree. **(Top)** The remaining highest connected in-degree node molecules are formic acid, methanol, formaldehyde, acetaldehyde, CO_2, glycolaldehyde, acrolein, glyoxal and methylglyoxal. **(Bottom)** The remaining highest connected out-degree node molecules are the input molecule formaldehyde, acetaldehyde, acrolein, glycolaldehyde, CO_2, propionaldehyde, acetone, methane and crotonaldehyde.

Out of the total ~350,000 reactions in the computed FR CRNR, there were only 78 reactions of the type A + B → 2 C or A → 2 B, with 70% being retro aldol reactions, and 30% being Knoevenagel C or H inverse reactions (an inverse reaction rule is simply the corresponding reverse reaction). 21 of these 78 reactions were assessed as thermodynamically spontaneous at pH 10 (with $\Delta G_{rxn} < 0$). Reactions which depend on the symmetry of the products and reactants are likely especially important for autocatalytic motifs and may represent "linchpin motifs" in autocatalytic reactions.

221 formols (compounds containing the $ROCH_2OH$ motif) were also found in the network, as was just one with a methylenedioxy motif ($R_1OCH_2OR_2$). Such formols form readily from reactions of HCHO with various nucleophiles, but the formation of amine and amide formols was not included in the reaction model as these compounds often form

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

reversibly and with low equilibrium constants, and there is little available data to benchmark how common such structures are in real reactions.

Since the reaction simulation is capped by both molecular weight and the number of generations that can be explored for computational reasons, the analysis in Figure 11 is biased in various ways. Furthermore, since kinetics are not considered in the CRNR, the relative importance of highly connected species cannot be assessed using these methods. This ranking thus only illustrates that small achiral aldehyde and ketone functional group-containing intermediates likely have the greatest power to enable stereochemical amplification in the FR.

Conclusions

The FR is a complex diversity-generating reaction, which has been largely examined in the context of prebiotic chemistry as a source of ribose[67]. It is however a much more complex reaction than previously appreciated and likely leads to numerous autocatalytic processes which can amplify underappreciated phenomena, including the generation of specific types of molecules and stereochemical enrichment.

The products of the FR CRNR presented here overlap well with experimental FT-ICR-MS measurements of actual FR products within the range of overlap of the model and the measurement, though not completely. Nevertheless, this overlap points to places where the model and observation can be brought into a better agreement by modification of the CRNR and the use of alternative measurement techniques.

This FR CRNR suggests that there may be numerous points and mechanisms for stereochemical amplification in the FR, suggesting this may be the point of stereochemical enrichment. The CRNR presented here can likely be modified to predict places where enantiomeric symmetry breaking can occur. Such symmetry-breaking effects have been observed in carbonaceous chondrites with respect to products which are either direct products of FR-like chemistry or derived from reactions which use FR products as inputs[8,64], though the existence and physical source of such possible enantio-enrichment remains to be explored (see also[63]).

This FR CRNR also predicts the creation of novel phases during the course of the reaction, which agrees with the observations of Weber[5] and Rand *et al.*[56] regarding the formation of phase-assembled microspherules, which are also observed in carbonaceous chondrites[68]. Such phase-assembled structures may be important for the origins of life for reasons described previously (*e.g.*,[33, 69]).

There has been a recent spate of papers[16,70-72] examining the FR noting the reaction's ability to integrate various aspects of life-like functions that are sometimes compared to Ganti's chemoton model[73], which takes an integrative view of life and its origins. This recent work resurrects and explores more deeply older work making some of the same points[74], but the present work suggests the creation of compartments may be a concomitant aspect of the FR under some conditions which may help facilitate the development of integrative chemistries which lead to self-reproducing chemical unities. There is clearly a good deal of work to be done to explore and expound how FR chemistry can lead to the ideas articulated by Dekker decades ago, which may require the development of novel experimental techniques.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Overall, the FR CRNR presented here suggests the FR can create unappreciated molecular diversity, and that as yet unspecified autocatalytic reactions may contribute to the creation of phase separation and enantiomeric symmetry breaking that may have contributed to producing the primordial organics observed in CCs, which may be widespread phenomena in protostellar disks and primitive planetary environments throughout the Universe. How these systems might complexify to give rise to living systems remains to be elucidated, but it is evident there are numerous places for coordinated self-reinforcing emergent behaviors to arise[75].

Materials and Methods

Reaction Modelling

In silico modeling of the FR was conducted using previously published methods[47] which potentially allows for the identification of potential autocatalytic pathways based on the open-source Python MØD platform[76,77], a graph theory-based chemical reaction modeling software package that can generate CRNRs using user-defined reaction rules (see SI File 1). The FR network explored here was computed using MØD run on a Hewlett-Packard Z820 computer with 128 GB RAM and 16 cores in two Intel® Xeon® E5-2670 2.60 GHz processors, running Ubuntu 20.04 in a virtual machine hosted on a Windows 10 system. MØD uses state-of-the-art graph canonicalization algorithms which are efficient at dealing with large collections of molecular graphs[78,79].

Reactant molecules were loaded into MØD using SMILES format[80]. In MØD, graphs provide a framework for representing chemical reactions where molecules can be treated as nodes in a graph, while the edges connecting them represent reactions. Reactions are processed by application of reaction rules which provide templates describing which transformations substructures in reactant molecules can undergo (Figure 12). Products of such reactions then become reactants for subsequent iterations of this process, leading to a concatenated chain of reactions (hereafter termed "generations"). The product graphs are then output in an internally canonicalized SMILES format. It should be noted that information allowing reconstruction of the entire network can be exported from MØD to other desired formats and parsed by other software for further analysis.

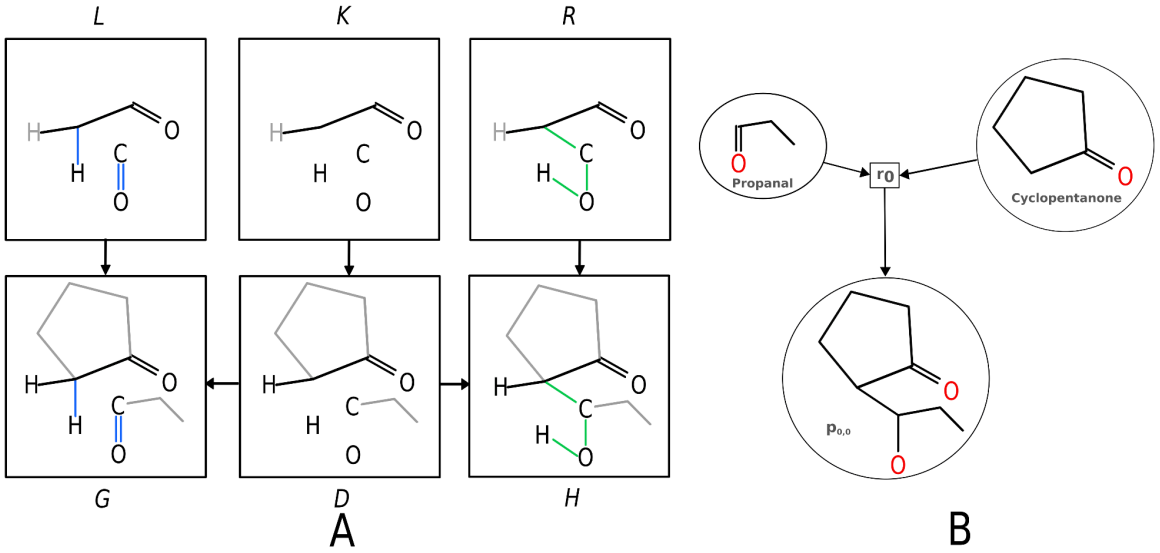


Figure 12. A: The aldol condensation reaction (the rule $L \xleftarrow{l} K \xrightarrow{r} R$) applied to cyclopentanone (the graph *G*) as implemented in MØD. The blue highlighted bonds in L are broken and the green highlighted bonds in R are created in the course of the reaction, bonds

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

in black do not change. The graph *H* is the product of the reaction (2-(1-hydroxypropyl)cyclopentan-1-one), while *D* is an intermediary graph in the transformation algorithm. This reaction normally proceeds through the action of a base (*e.g.*, $Ca(OH)_2$, KOH or NaOH), the availability of such bases in the medium is assumed. B: The same reaction shown as a fragment of a reaction network, formally a directed multi-hypergraph[81]. Reactions between species store the information about which reaction rule was used to generate the products.

The initial reactants, formaldehyde, glycolaldehyde, and H_2O were input in SMILES. A set of defined reaction rules (summarized in Table SI2) developed based on literature precedent was then loaded. Stereochemistry is undefined in the formalism employed here, e.g., DL-glyceraldehyde is treated as a single "flattened" isomer (*R,S*-2,3-dihydroxypropanal), threose and erythrose are treated as equivalent, etc. Reaction rules are templates for reactions and are formally Double Push-Out (DPO) graph transformation rules[82,83]. A DPO rule consists of three graphs (or molecular patterns): the left graph 'L' represents the reactant patterns to search for, the right graph 'R', the product pattern that will be substituted in place of "L", and the rule context graph 'K' which is needed along with the two arrows to formally encode which atoms of L map to which in R. If the reactant pattern L can be found in a molecule, then the corresponding edges of the graph (the chemical bonds of the molecule(s)) can be changed by removing all edges defined by L that are not in K, and adding all of the edges of R. The initial step in reaction rule application is thus the search for the reactant pattern L.

Rule Selection

The rule selection approach used in our workflow does not allow for automated consideration of the effects of pH, temperature, or other kinetic considerations, or specifically consider kinetic effects except as they are latently represented by the selection of reaction rules. Reaction choice thus involves user-defined biases, and considers all user-defined mechanistically plausible reactions. The implemented reactions were however vetted by their ability to create products that can explain the compounds reported by the comprehensive studies of Decker *et al.*[48], and Omran *et al.*[49], and selected to be chemically reasonable in terms of their reversibility, for example reactions known to be essentially reversible in basic medium, e.g., ester hydrolysis or decarboxylation, were encoded here as only proceeding in the forward direction.

Post-Generation Filtering

During application of reaction rules, structures deemed to be chemically unstable were sometimes produced which are unlikely to appear in real chemistry. The exponential growth in the number of species produced with each generation also quickly made the computation resource heavy. To ameliorate these issues, each computed product was examined for sterically problematic or "bad" substructures, and a molecular weight (MW) cutoff filter was applied (herein 200 amu). Any species containing a bad substructure or exceeding the prescribed MW limit created during reaction generation was discarded from the network. This helped to both clean bad products and limit the size of the product space input to the next generation, making further generation computation more tractable The forbidden "bad" structures were selected based on literature precedent (*e.g.,* [84]), as well as those included in MolGen 5.0[85], see SI File 2), *e.g.*, three- and four-membered rings were excluded as they are typically unstable due to their large ring strain[86]. Some chemically implausible substructures (*e.g.*, structures with double bonds at bridgehead carbons, which violate Bredt's rule[87] were also specifically forbidden. This filtering reduces the computational load for succeeding generations, and since each product must be computed

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

before it is checked, limiting mass growth helps avoid "combinatorial explosions", which allows computation of more iterations than would otherwise be possible.

Treatment of Tautomers

Tautomers are compounds with identical molecular formulas, which are structural (*i.e.*, constitutional) isomers of compounds that readily interconvert, often by proton migration, among other mechanisms[88]. Many organic compounds containing conjugated bonds can be represented by multiple human-interpretable structures, though chemists often prefer one tautomeric representation. Compounds with unique anhydrous crystal structures often interconvert into multiple bonding states in solution, and even relatively simple compounds such as guanine ($C_5H_5N_5O$), with completely defined core bonding, may have as many as 36 valid tautomeric representations[89]. MØD treats different tautomeric forms as distinct chemical species; it was thus necessary to suppress the generation of multiple tautomeric forms for the same molecule. Tautomerism is generally acknowledged to be a difficult representational problem in computational chemistry (see[90,91]), although resolutions for this issue have been proposed (*e.g.*, [92]). Tautomeric structures are capable of undergoing the same reactions because they are interconvertible, and denoting tautomers as unique in this context is meaningful in some ways (*e.g.*, with respect to precise reaction mechanisms or hydrogen bonding), but not in others (*e.g.*, the overall outcome of reactions). We handled tautomerism at the level of encoded reaction mechanisms, such that reaction schemes would provide tautomeric species that a chemist would recognize as "catalog representations" of tautomers[93], and such that only one unique tautomer of a given compound would be produced anywhere in the network, *e.g.*, enol forms were forced into keto or aldo forms, and reaction mechanisms were built to infer the existence of transient reactive species such as enolates. This simplification of tautomers reduces the number of unique species representations generated, reducing computational cost and assisting in output interpretation.

Comparison with literature analyses

The predictive validity of the generated FR network was evaluated by checking if compounds detected in references [48] and [49] could be explained by this workflow. To do so, the molecules detected were rendered into sdf format. These reported species (henceforth referred to as the "test set") were then imported into our workflow using RDKit and matched against the compounds generated in our workflow by performing isomorphism tests using the VF2 algorithm in MØD[94]. Molecules are isomorphic if all the atoms in one molecule are bonded in exactly the same fashion as another, despite their initial arbitrary labeling. We did not represent aromaticity explicitly in our reaction rules, aromatic compounds would need to be identified from their Kekulé forms. Since there are multiple Kekulé representations for a single aromatic species, compounds in the test set were kekulized to a canonical representation using RDKit before comparison.

Cheminformatic Descriptor Computation

The compounds produced in these networks represent real chemical compounds with real chemical properties, and it was of interest to understand how real chemical reactions explore chemical property space for emergent behaviors (*e.g.*, [95]). Energy minimization optimizations of output SMILES were carried out using OpenBabel[96]. Physical and topological chemoinformatic descriptors were computed using DataWarrior[52]. Two-dimensional descriptors were calculated and benchmarked using Mordred[97].

Detection of Autocatalytic Motifs

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Autocatalysis is a phenomenon in which a reaction or sequence of reactions produces a compound catalyzed by one or more products of the reaction(s) in a system (*e.g.*, [73, 98-102]). Kinetically, autocatalysis leads to super-linear growth of a species or network. However, as discussed in [103] for a CRN, computationally identifying whether a compound produces more than one copy of itself is not straightforward. Although formalized definitions of autocatalysis that can be generally applied to CRNs remains to be defined, we here searched for topologically formal autocatalytic motifs, i.e., for those in which a cyclic set of reactions produces two or more copies of one of the cycle component compounds, as inferred from the network topology. We used an imperative programming approach to identify autocatalytic cycles in which a graph algorithm was written to find motifs, using the Ford-Fulkerson algorithm[104]. A data loader routine was written in Python to import the MØD-generated chemical network data into a graph query database, Neo4j so that pattern match queries could be built. The loader routine ensured that no distinct molecule or reaction node was repeated, which gave a high-level representation on the molecule and reaction node (bipartite, or two-label graph per Neo4j) graph. Ensuring node distinctness allowed the network model to be lighter to improve computational performance since the pathways were being examined, and not the raw quantity of molecules in the network. The network model could then be thought of as a model of every distinct pathway extracted from the simulated chemical network from MØD (whose network model is physically closer to the actual chemical network representation, since this model is a level of abstraction on the MØD network). As the database populates, additional information, such as when the molecule was first generated, is stored in the .txt files. If a molecule is found again in a subsequent generation, that molecule's node is not re-recorded. Similarly, if a reaction of a certain rule already exists between molecules, that edge is not repeated. Based on the rules of the loader script, the Neo4j graph model is a species-only graph, rather than the bipartite molecule and reaction graph. The bipartite graph is useful for describing complex patterns that require constraints on the relationships between reactions and molecules, while the species-only graph is useful for network centrality measures.

The pattern match query is user-defined, and thus modifiable. The query template can further be improved to allow modification of the queried length of the shunt pathway. The pattern matches were exported to a tabulated CSV file where several columns of information, e.g., the contents of each substructure path, the SMILES string of each molecule in the pattern, reaction information on each node, etc., were included for better visualization. Neo4j can also be used to compute various network statistics, including centrality measures for example: eigenvector and betweenness centrality, random-walk betweenness and node-degree metrics example: count of incoming or outgoing edges and node degree rank by generation, and the change in node degree by generation which may help assess when certain molecules become most influential in affecting network development.

Thermochemical Calculations

We used the eQuilibrator API to calculate each compound's standard free energy of formation, $\Delta G_f'^{\circ}$, and the change in free energy for each reaction, $\Delta G_r'^{\circ}$ [66], estimated at standard conditions, meaning that all species are at 1 M concentration at 25° C and 1 bar (see Figure SI5). We used the eQuilibrator API default value for the Mg^{2+} concentration (pMg, in this case pMg = 3.0), which is required for these calculations, but which mostly affects reactions involving ATP. We explored the eQuilibrator API calculations at pH 10. The eQuilibrator software uses component contribution methods to estimate thermodynamic values. Temperature and pressure cannot be varied, because the model calibration relies on experimental measurements for ΔfG conducted at standard T and P. The Neo4J defined search pattern doesn't guarantee a pathway identified as topologically autocatalytic is energetically favorable, since the reactions derived from reaction expansion do not

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

specifically encode energetic information. To address this problem, we merged computed thermochemical information derived from the eQuilibrator API onto the CRNR nodes so that network queries could constrain energetic favorability. For example, sorting the pattern match results by the minimum aggregated energy across the reactions in the ring path and returning the lowest energy paths should yield the most energetically favorable reaction motifs.

Network Data Visualization

The Gephi software package was used to visualize the generated reaction networks. In Gephi plots, node size can be made proportional to any graph property for emphasis, for example the in-degree or out-degree of the nodes. Plot layouts were rendered using Gephi's ForceAtlas2 function by setting the gravity parameter to 0.05 and scaling parameter to 4.0, and running this for 10 minutes without avoiding overlapping and then 20 minutes avoiding overlapping[105].

Laboratory Formose Reaction

Formose reactions were prepared in pre-ashed glass ampoules sealed under vacuum according to the general methods of Kebukawa *et al.*[68]. Reactions contained 180 mg paraformaldehdye (Sigma-Aldrich, 6 mmoles), 180 mg glycolaldehyde (Sigma-Aldrich, 3 mmoles) and 45 mg $Ca(OH)_2$ (Fischer Scientific, 0.6 mmoles) dissolved in 3 mL 18 MΩ·cm water from a Waters milliQ purification system. Sealed reaction tubes were placed in parallel at 85 or 150° C for 8 days.

At the end of the reaction period, the tubes were opened, and homogenized by rapid pipetting and an aliquot was removed and centrifuged briefly using a benchtop centrifuge. The resulting supernatant was dried under vacuum and then analyzed by FT-ICR-MS as documented previously[106].

For microscopic analysis, reactions were prepared by mixing 1 M HCHO and 10 mM GA over a $Ca(OH)_2$ slurry according to the methods described above. Reactions were allowed to proceed at room temperature for approximately one month before imaging.

FT-ICR Mass Spectrometry and Spectral Processing

Samples were analyzed as previously described[106] by a custom-built FT-ICR mass spectrometer equipped with a 9.4 Tesla horizontal 220 mm bore diameter superconducting solenoid magnet[107] and a modular ICR data station (Predator)[108]. FT-ICR MS files are publicly available via the Open Science Framework (https://osf.io/my7ku/) at DOI 10.17605/OSF.IO/MY7KU.

Proton masses were added to the FT-ICR-MS spectra to compare them with the simulated spectra by adding 1.007276. In the FR mass spectra, 63 out of 228 peaks (27.6%) in the < 200 MW range had a molecular formula hit. Over six generations, 6886 out of 8250 (83.4%) unique mass products were matched, though it should be borne in mind that each of these unique masses typically corresponds to multiple unique isomeric species generated by the *in silico* network. Peaks that were unassignable, potentially due to the presence of unconsidered elements, or that were were excluded from this analysis. We also examined ^{13}C matches, by adding 1.003355. In the raw processed data, several peaks could not be assigned a molecular formula by the spectrometer. These peaks could possibly have been inorganic salt clusters. Therefore, we discarded all data entries with "No Hit" for an assigned

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

formula. An exception was made for the peaks in the m/z 150-200 range which we found to be consistent with containing a ^{13}C isotope within a precision of 4 decimal places. Next, given that our intended sample and modeled chemistry (Formose Reaction) involves only C, H and O, we further filtered out data points containing nitrogen and sulfur atoms.

Microscopy

A. Brightfield Microscopy

~2 uL samples of the room temperature reaction described above were aliquoted onto glass slides fitted with tape with holes punched in it before placing a glass coverslip, affording an ~0.05 mm gap between the coverslip and slide surface. Samples were visualized using an AmScope T600A-PCS-8M Digital Professional Phase-Contrast Compound Trinocular Microscope, with imaging conducted using an 8MP CCD Camera.

B. Scanning Electron Microscopy (SEM)

~1 uL samples of the reactions described above were spotted on carbon tape-covered aluminium SEM stubs. These were dried under air overnight, then under vacuum (~ 10 um for five minutes), then sputter-coated with iridium. Samples were then visualized using a Zeiss Auriga Field Emission SEM.

Data availability

The open-source pipeline along with the supporting documentation, mostly written in Python for the simulation can be accessed at https://github.com/Reaction-Space-Explorer/reac-space-exp.

Author contributions

RC conducted the FR computations, performed the searches for spontaneous autocatalytic cycles and created the Gephi visualizations. SS conducted the descriptor computations and visualizations. AA conducted the majority of the MØD simulations and coded most of the reaction mechanisms. JR designed and wrote the Neo4J queries. AL conducted the database comparison studies. JLA developed MØD and helped implement its use in this study. HJC designed and oversaw this study and conducted laboratory experiments (microscopy, and FT-ICR-MS analysis). All authors contributed to the writing of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

RC, SS, AA JR, AL, and HJC would like to thank the Blue Marble Space Institute of Science (BMSIS) for organizational support. JA and HJC would like to thank the Earth-Life Science Institute (ELSI) and the ELSI Origins Network (EON) for financial support during the early development of this work. EON was supported by a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. RC wishes to acknowledge FONDECYT (Convenio 208-2015-FONDECYT) for his Masters scholarship. He would also like to thank Miguel Miní for his suggestion and support in the imperative search code for

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

autocatalytic cycles. JA is also supported by the Novo Nordisk Foundation grants NNF19OC0057834 and NNF21OC0066551, and by the Independent Research Fund Denmark, Natural Sciences, grant DFF-0135-00420B. A portion of this work was funded by the National Science Foundation Division of Chemistry and Division of Materials Research through NSF DMR-1644779, and the state of Florida.

Conflicts of Interest

The authors declare no conflicts of interest.

Keywords
Combinatorial Chemistry - Mass spectrometry - Molecular Diversity - Computational Chemistry - Prebiotic Chemistry - Origins of Life

References

1. Butlerow, A. Formation synthétique d'une substance sucrée. *Comptes Rendus* 1861, 53, 145–147.
2. Martínez, R.; Cuccia, L.; Viedma, C.; Cintas, P. On the origin of sugar handedness: Facts, hypotheses and missing links - A review. *Origins Life. Evol. Biosph.* 2022, *52*, 21–56.
3. Yi, R., Mojica, M., Fahrenbach, A. C., James Cleaves, H., Krishnamurthy, R., & Liotta, C. L. Carbonyl Migration in Uronates Affords a Potential Prebiotic Pathway for Pentose Production. *J. Am. Chem. Soc. Au*. 2023, *3*(9), 2522-2535.
4. Larralde, R.; Robertson, M.; Miller, S. Rates of decomposition of ribose and other sugars: implications for chemical evolution. *Proc. Nat. Acad. Sci. USA* 1995, *92*(18), 8158-8160.
5. Weber, A. Growth of organic microspherules in sugar-ammonia reactions. *Origins of Life Evol. Biospheres* 2005, *35*(6), 523-536.
6. Mizuno, T.; Weiss, A. Synthesis and utilization of formose sugars. *Adv. Carb. Chem. & Biochem.* 1974, *29*, 173–227.
7. Schwartz, A.; De Graaf, R. The prebiotic synthesis of carbohydrates: a reassessment. *J. Mol. Evol.* 1993, *36*(2), 101-106.
8. Cooper, G.; Rios, A. Enantiomer excesses of rare and common sugar derivatives in carbonaceous meteorites. *Proc. Nat. Acad.Sci.* 2016, *113*(24), E3322-E3331.
9. Paschek, K.; Kohler, K.; Pearce, B. Lange, K.; Henning, T.; Trapp, O.; Pudritz, R.; Semenov, D. Possible ribose synthesis in carbonaceous planetesimals. *Life*. 2022, *12*(3), 404.
10. Delidovich, I.; Simonov, A.; Taran, O.; Parmon, V. Catalytic formation of monosaccharides: From the formose reaction towards selective synthesis. *ChemSusChem*. 2014, *7*(7), 1833-1846.
11. OoLEN; Asche, S.; Bautista, C.; Boulesteix, D.; Champagne-Ruel, A.; Mathis, C.; Markovitch, O.; Peng, Z.; Adams, A.; Dass, A. V.; et al. What It Takes to Solve the Origin(s) of Life: An Integrated Review of Techniques. 2023, *arXiv*.
12. Schwartz, AW. Intractable mixtures and the origin of life. *Chem. Biodivers.* 2007, *4*(4), 656-64.
13. Springsteen, G.; Joyce, FG. Selective derivatization and sequestration of ribose from a prebiotic mix. *J. Am. Chem. Soc.* 2004, *126.31*, 9578-9583.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

14. Benner, S. A., Kim, H. J., & Carrigan, M. A. Asphalt, water, and the prebiotic synthesis of ribose, ribonucleosides, and RNA. Acc. Chem. Res. 2012, *45*(12), 2025-2034.
15. Ricardo, A., Carrigan, M. A., Olcott, A. N., & Benner, S. A. Borate minerals stabilize ribose. Science, 2004. *303*(5655), 196-196.
16. Q. P. Tran, R. Yi, A. C. Fahrenbach, Towards a prebiotic chemoton – nucleotide precursor synthesis driven by the autocatalytic formose reaction. *Chem. Sci*. 2023.
17. Simonov, A.; Pestunova, O.; Matvienko, L.; Parmon, V. The nature of autocatalysis in the Butlerov reaction. *Kinet. Catal.* 2007, *48*, 245–254.
18. Breslow, R. On the mechanism of the formose reaction. *Tet. Letts*. 1959, *1*(21), 22-26.
19. Socha, R.; Weiss, A.; Sakharov, M. Autocatalysis in the formose reaction. *Reaction Kinetics and Catalysis Letts.* 1980, *14(2)*, 119-128.
20. Lambert, J.; Gurusamy-Thangavelu, S.; Ma, K. The silicate-mediated formose reaction: bottom-up synthesis of sugar silicates. *Science* 2010, *327*(5968), 984-986.
21. Appayee, C.; Breslow, R. Deuterium studies reveal a new mechanism for the formose reaction involving hydride shifts. *J. Am. Chem. Soc.* 2014, *136*(10), 3720-3723.
22. Ritson, D.; Sutherland, J. Prebiotic synthesis of simple sugars by photoredox systems chemistry. *Nat. Chem.* 2012, *4*(11), 895-899.
23. Huskey, W.; Epstein, I. Autocatalysis and apparent bistability in the formose reaction. *J. Am. Chem. Soc.* 1989, *111*(9), 3157-3163.
24. Peng, Z.; Linderoth, J.; Baum, D. The hierarchical organization of autocatalytic reaction networks and its relevance to the origin of life, *PLOS Comp. Bio.* 2022*, 18*(9), e1010498.
25. Cheng, L., Doubleday, C., & Breslow, R. (2015). Evidence for tunneling in base-catalyzed isomerization of glyceraldehyde to dihydroxyacetone by hydride shift under formose conditions. *Proc. Nat. Acad. Sci. USA*, 112(14), 4218-4220.
26. Heald, C.; Kroll, J. The fuel of atmospheric chemistry: Toward a complete description of reactive organic carbon. *Sci. Advances* 2020, *6*(6), eaay8967.
27. Ricardo, A.; Frye, F.; Carrigan, M.; Tipton, J.; Powell, D.; Benner, S. 2-Hydroxymethylboronate as a reagent to detect carbohydrates: Application to the analysis of the formose reaction. *J. Org. Chem.* 2006, *71*(25), 9503-9505.
28. Shigemasa, Y.; Matsuda, Y.; Sakazawa, C.; Matsuura, T. Formose reactions. II. The photochemical formose reaction. *Bull. Chem. Soc. Japan* 1977, *50*(1), 222-226.
29. Omran, A. Plausibility of the Formose Reaction in Alkaline Hydrothermal Vent Environments. *Space Life Sci.* 2020, 1–13.
30. Omran, A.; Gonzalez, A.; Menor-Salvan, C.; Gaylor, M.; Wang, J.; Leszczynski, J.; Feng, T. Serpentinization-Associated Mineral Catalysis of the Protometabolic Formose System. *Life*, 2023, *13,* 1297.
31. Feng, S.; Tian, G.; He, C.; Yuan, H.; Mu, Y.; Wang, Y.; Wang, L. Hydrothermal biochemistry: from formaldehyde to oligopeptides. *J. Mat. Sci.* 2008, *43*(7), 2418-2425.
32. Kopetzki, D.; Antonietti, M. Hydrothermal formose reaction. *New J. Chem.* 2011, *35*(9), 1787-1794.
33. Weber, A. Chemical constraints governing the origin of metabolism: the thermodynamic landscape of carbon group transformations under mild aqueous conditions. *Orig. Life Evol. Biosph.* 2002, 32, 333–357.
34. Kua, J.; Avila, J.E.; Lee, C.G.; Smith, W.D. Mapping the kinetic and thermodynamic landscape of formaldehyde oligomerization under neutral conditions. *J. Phys. Chem. A*. 2013, *117, 12658–12667*.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

35. Fray, N.; Bénilan, Y.; Cottin, H.; Gazeau, M. New experimental results on the degradation of polyoxymethylene: Application to the origin of the formaldehyde extended source in comets. *J Geophys. Res. Planets.* 2004, *109*(E7).
36. Cleaves, H. The prebiotic geochemistry of formaldehyde. *Precamb. Res.* 2008, *164*(3-4), 111-118.
37. Haas, M.; Lamour, S.; Christ, S.; Trapp, O. Mineral-mediated carbohydrate synthesis by mechanical forces in a primordial geochemical setting. *Comm. Chem.* 2020, *3*(1), 1-6.
38. Vinogradoff, V.; Leyva, V.; Mates-Torres, E.; Pepino, R.; Danger, G.; Rimola, A.; Cazals, L.; Serra, C.; Pascal, R.; Meinert, C. Olivine-Catalyzed Glycolaldehyde and Sugar Synthesis under Aqueous Conditions: Application to Prebiotic Chemistry. *Earth and Planetary Science Letters*, 2024, *626*, 118558.
39. Georgelin, T.; Jaber, M.; Fournier, F.; Laurent, G.; Costa-Torro, F.; Maurel, M.; Lambert, J. Stabilization of ribofuranose by a mineral surface. *Carbohydrate Res.* 2015, *402*, 241-244.
40. Dass, A.; Georgelin, T.; Westall, F.; Foucher, F.; De Los Rios, P.; Busiello, D.; Liang, S.; Piazza, F. Equilibrium and non-equilibrium furanose selection in the ribose isomerisation network. *Nature Commun.* 2021, *12*(1), 1-10.
41. Liu, Y.; Sumpter, D. Mathematical modeling reveals spontaneous emergence of self-replication in chemical reaction systems. *J. Biol. Chem.* 2018, *293*(49), 18854-63.
42. Simm, G.; Reiher, M. Context-driven exploration of complex chemical reaction networks. *J. Chem. Theory Comput.* 2017, *13*, 6108–6119.
43. D. Rappoport, C. J. Galvin, D. Yu. Zubarev, A. Aspuru-Guzik, Complex Chemical Reaction Networks from Heuristics-Aided Quantum Chemistry. J. Chem. Theory Comput. 2014, *10*, 897–907
44. Robinson, W.; Daines, E.; van Duppen, P.; de Jong, T.; Huck, W. Environmental conditions drive self-organization of reaction pathways in a prebiotic reaction network. *Nature Chem.* 2022, *1-9*.
45. Venturini, A.; González, J. Prebiotic Synthesis of Glycolaldehyde and Glyceraldehyde from Formaldehyde: A Computational Study on the Initial Steps of the Formose Reaction. *ChemPlusChem*, 2023.
46. Sharma, S.; Arya, A.; Cruz, R.; Cleaves, H. Automated exploration of prebiotic chemical reaction space: Progress and perspectives. *Life*. 2021, *11*(11), 1140.
47. Arya, A.; Ray, J.; Sharma, S.; Cruz-Simbron, R.; Lozano, A.; Smith, H.; Andersen, J.; Chen, H.; Meringer, M.; Cleaves, H. An open source computational workflow for the discovery of autocatalytic networks in abiotic reactions. *Chem. Sci.* 2022, *13*(17), 4838-4853.
48. Decker, P.; Schweer, H.; Pohlmann, R. Bioids: X. Identification of formose sugars, presumable prebiotic metabolites, using capillary gas chromatography/gas chromatography—mass spectrometry of *n*-butoxime trifluoroacetates on OV-225. *J. Chrom. A* 1982, *244*(2), 281-91.
49. Omran, A.; Menor-Salvan, C.; Springsteen, G.; Pasek, M. The messy alkaline formose reaction and its link to metabolism. *Life.* 2020, *10*(8), 125.1.
50. Briš, A.; Baltussen, M. G.; Tripodi, G. L.; Huck, W. T. S.; Franceschi, P.; Roithová, J. Direct Analysis of Complex Reaction Mixtures: Formose Reaction. *Angewandte Chemie*, 2023.
51. Wu, Z.; Rodgers, R.; Marshall, A. Two- and three-dimensional Van Krevelen diagrams: A graphical analysis complementary to the kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband fourier transform ion cyclotron resonance mass measurements. *Anal. Chem.* 2004, *76*(9), 2511-6.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

52. Sander, T., Freyss, J., von Korff, M. and Rufener, C. DataWarrior: An open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* 2015, *55*(2):460–473.
53. Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., Torrance, G., Evelo, C. T., Guha, R. & Steinbeck, C. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminf.* 2017, *9*(1):1–19.
54. von Korff, M., & Sander, T. Molecular complexity calculated by fractal dimension. *Sci. Rep.* 2019, *9*(1), 1-8.
55. Bonchev, D.; Rouvray, D. (Eds.) Complexity in Chemistry, Biology, and Ecology. Springer US, 2005.
56. Rand, D.; Belenky, M.; Herzfeld, J. Microspherules from sugars in the absence of nitrogen. *Origins Life Evol. Biospheres* 2011, *41*(1): 17-22.
57. Kebukawa, Y.; Cody, G. A kinetic study of the formation of organic solids from formaldehyde: Implications for the origin of extraterrestrial organic solids in primitive Solar System objects. *Icarus* 2015, *248*, 412-423.
58. Furukawa, Y., Iwasa, Y., & Chikaraishi, Y. (2021). Synthesis of ^{13}C-enriched amino acids with 13C-depleted insoluble organic matter in a formose-type reaction in the early solar system. *Science advances*, *7*(18), eabd3575.
59. Delaney, J. ESOL: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* 2004, *44*, 1000–1005.
60. Dusenbery, D. Living at Micro Scale: The Unexpected Physics of Being Small. Harvard University Press, 2009.
61. Urey, H. Discussion of the identity of the organized elements. *Ann. NY Acad. Sci.* 1963, *108*(2), 606-615.
62. Anders, E.; Fitch, F. Search for organized elements in carbonaceous chondrites. *Science* 1962, *138*(3548), 1392-1399.
63. Cleaves, H. J. A hypothesis for a unified mechanism of formation and enantioenrichment of polyols and aldaric, aldonic, amino, hydroxy and sugar acids in carbonaceous chondrites. In Origins of Life: The Primal Self-Organization (pp. 37-55). Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.
64. Burton, A.; Berger, E. Insights into abiotically-generated amino acid enantiomeric excesses found in meteorites. *Life* 2018, *8*(2), 14.
65. Laurent, G., Lacoste, D., & Gaspard, P. Emergence of homochirality in large molecular systems.*Proc. Nat. Acad. Sci. USA*. 2021, *118*(3), e2012741118.
66. Beber, M. E., Gollub, M. G., Mozaffari, D., Shebek, K. M., Flamholz, A. I., Milo, R. and Noor, E. eQuilibrator 3.0: a database solution for thermodynamic constant estimation. *Nucleic Acids Res.* 2022, *50*(D1):D603–D609.
67. Banfalvi, G. Prebiotic Pathway from Ribose to RNA Formation. Int. J. Mol. Sci. 2021, *22,* 3857
68. Kebukawa, Y.; Kilcoyne, A.; Cody, G. Exploring the potential formation of organic solids in chondrites and comets through polymerization of interstellar formaldehyde. *Astrophys. J.* 2013, *77(1)*, 19.
69. Jia, T.; Chandru, K.; Hongo, Y.; Afrin, R.; Usui, T.; Myojo, K.; Cleaves, H. Membraneless polyester microdroplets as primordial compartments at the origins of life. *Proc. Nat. Acad. Sci. USA*. 2019, *116*(32), 15830-15835.
70. van Duppen, P., Daines, E., Robinson, W. E., & Huck, W. T. Dynamic Environmental Conditions Affect the Composition of a Model Prebiotic Reaction Network. *J. Am. Chem. Soc.* 2023, *145*(13), 7559-7568.
71. Lu, H., Blokhuis, A., Turk-MacLeod, R., Karuppusamy, J., Franconi, A., Woronoff, G., Jeancolas, C., Abrishamkar, A., Loire, E., Ferrage, F. and Pelupessy, P.. Small-molecule

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

autocatalysis drives compartment growth, competition and reproduction. *Nat. Chem*. 2023, pp.1-9.
72. Roszak, R., Wołos, A., Benke, M., Gleń, Ł., Konka, J., Jensen, P., Burgchardt, P., Żądło-Dobrowolska, A., Janiuk, P., Szymkuć, S., & Grzybowski, B. A. Emergence of Metabolic-like Cycles in Blockchain-Orchestrated Reaction Networks. *Chem,* 2024.
73. Gánti, T. Organization of chemical reactions into dividing and metabolizing units: The chemotons. *Biosystems*. 1975, *7*(1):15–21.
74. Decker, P. Possible resolution of racemic mixtures by bistability in “bioids”, open systems which can exist in several steady states. J. Mol. Evol. 1973, *2*, 137-143.
75. Horowitz, J. M., & England, J. L. Spontaneous fine-tuning to environment in many-species chemical reaction networks. *Proc. Nat. Acad. Sci. USA*. 2017, *14*(29), 7565-7570.
76. Andersen, J.; Flamm, C.; Merkle, D.; Stadler, P. A software package for chemically inspired graph transformation. In *International Conference on Graph Transformation* 2016, pp. 73-88.
77. Andersen, J. L., Andersen, T., Flamm, C., Hanczyc, M. M., Merkle, D., and Stadler, P. F. Navigating the chemical space of HCN polymerization and hydrolysis: Guiding graph grammars by mass spectrometry data. *Entropy*. 2013, *15*(10):4066–4083.
78. Andersen, J. L. and Merkle, D. A generic framework for engineering graph canonization algorithms. *ACM J. Exp. Algorithmics*. 2020, *25*:1–26.
79. Andersen, J. L.; Flamm, C.; Merkle, D.; Stadler, P. F. Chemical Transformation Motifs—Modelling Pathways as Integer Hyperflows. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2019, *16*, 510–523.
80. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 1988, *28*(1):31–36.
81. Andersen, J. L., Flamm, C., Merkle, D., and Stadler, P. F. Maximizing output and recognizing autocatalysis in chemical reaction networks is NP-complete. *J. Syst. Chem.* 2012, *3*(1):1–9.
82. Ehrig, H., Ehrig, K., Prange, U., and Taentzer, G. Fundamentals of Algebraic Graph Transformation (Monographs in Theoretical Computer Science. An EATCS Series). 2006, SpringerVerlag, Berlin, Germany.
83. Andersen, J. L., Fagerberg, R., Flamm, C., Fontana, W., Kolčák, J., Laurent, C. V. F. P., Merkle, D., & Nøjgaard, N. Representing Catalytic Mechanisms with Rule Composition. J. Chem. Inf. Model. 2022, *62*, 5513–5524.
84. Weber, A. L. Kinetics of organic transformations under mild aqueous conditions: implications for the origin of life and its metabolism. *Origins Life Evol. Biosphere*. 2004, *34*(5):473–495.
85. Gugisch, R., Kerber, A., Kohnert, A., Laue, R., Meringer, M., Rücker, C., and Wassermann, A. Chapter 6 - Molgen 5.0, a molecular structure generator. In Basak, S. C., Restrepo, G., and Villaveces, J. L., editors, Advances in Mathematical Chemistry and Applications, 2005, pages 113–138. Bentham Science Publishers.
86. Wiberg, K. B. The concept of strain in organic chemistry. *Angew. Chem. Int. Ed. Engl.* 1986, *25*(4):312–322.
87. Bredt, J. Über sterische Hinderung in Brückenringen (Bredtsche Regel) und über die mesotrans-Stellung in kondensierten Ringsystemen des Hexamethylens. Justus Liebigs Ann. Chem. 1924, *437*(1):1–13.
88. Kochev, N. T., Paskaleva, V. H., and Jeliazkova, N. Ambit-Tautomer: An open source tool for tautomer generation. *Mol. Inf.* 2013, *32*(5-6):481–504.
89. Sabio, M., Topiol, S., and Lumma, W. C. An investigation of tautomerism in adenine and guanine. *J. Phys. Chem.* 1990, *94*(4):1366–1372.
90. Sayle, R. A. So you think you understand tautomerism? *J. Comput.-Aided Mol. Des.* 2010, *24*(6):485–496.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

91. Dhaked, D. K., Ihlenfeldt, W.-D., Patel, H., Delannée, V., and Nicklaus, M. C. Toward a comprehensive treatment of tautomerism in chemoinformatics including in InChI V2. *J. Chem. Inf. Model.* 2020, *60*(3):1253–1275.
92. Warr, W. A. Tautomerism in chemical information management systems. *J. Comput.-Aided Mol. Des.* 2010, *24*(6):497–520.
93. Taylor, P. J., van der Zwan, G., and Antonov, L. Tautomerism: Introduction, History, and Recent Developments in Experimental and Theoretical Methods. *In Tautomerism*. 2013, pages 1–24. John Wiley & Sons, Ltd.
94. Cordella, L. P., Foggia, P., Sansone, C., and Vento, M. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 2004, *26*(10):1367–1372.
95. Jain, S. and Krishna. A model for the emergence of cooperation, interdependence, and structure in evolving networks. *Proc. Natl. Acad. Sci. U.S.A.* 2001, *98*(2):543–547.
96. O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* 2011, *3*(1):1–14.
97. Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* 2018, *10*(1):1–14.
98. Blokhuis, A., Lacoste, D., and Nghe, P. Universal motifs and the diversity of autocatalytic systems. *Proc. Natl. Acad. Sci. U.S.A.* 2020, *117*(41):25230–25236.
99. Peretó, J. Out of fuzzy chemistry: from prebiotic chemistry to metabolic networks. *Chem. Soc. Rev.* 2012, *41*(16):5394–5403.
100. Preiner, M., Xavier, J. C., do Nascimento Vieira, A., Kleinermanns, K., Allen, J. F., and Martin, W. F. Catalysts, autocatalysis and the origin of metabolism. *Interface Focus*. 2019, *9*(6):20190072.
101. Peng, Z.; Adam, Z. R.; Fahrenbach, A. C.; Kaçar, B. Assessment of Stoichiometric Autocatalysis across Element Groups. *J. Am. Chem. Soc*. 2023, *145*, 22483–22493.
102. Q. P. Tran, Z. R. Adam, A. C. Fahrenbach, Prebiotic Reaction Networks in Water. *Life. 2020, 10,* 352 .
103. Andersen, J. L., Flamm, C., Merkle, D., and Stadler, P. F. Defining autocatalysis in chemical reaction networks. 2021, arXiv.
104. Ford, L. R. and Fulkerson, D. R. Maximal flow through a network. *Can. J. Math.* 1956, *8*:399–404.
105. Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*. 2014, *9*(6):e98679.
106. Guttenberg, N.; Chen, H.; Mochizuki, T.; Cleaves, H. Classification of the biogenicity of complex organic mixtures for the detection of extraterrestrial life. *Life*. 2021, *11*(3), 234.
107. Kaiser, N.K.; Quinn, J.P.; Blakney, G.T.; Hendrickson, C.L.; Marshall, A.G. A novel 9.4 Tesla FTICR mass spectrometer with improved sensitivity, mass resolution, and mass range. *J. Am. Soc. Mass Spectrom.* 2011, *22,* 1343–1351.
108. Blakney, G.T.; Hendrickson, C.L.; & Marshall, A.G. Predator data station: A fast data acquisition system for advanced FT-ICR MS experiments. *Int. J. Mass. Spectrom.* 2011, *306(2-3)*, 246-52.

https://doi.org/10.26434/chemrxiv-2024-nj0p6 **ORCID:** https://orcid.org/0000-0003-1768-1802 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0