nature chemistry

ARTICLES

https://doi.org/10.1038/s41557-022-00956-7

Check for updates

Environmental conditions drive self-organization of reaction pathways in a prebiotic reaction network

William E. Robinson, Elena Daines, Peer van Duppen, Thijs de Jong and Wilhelm T. S. Huck ✉

The evolution of life from the prebiotic environment required a gradual process of chemical evolution towards greater molecular complexity. Elaborate prebiotically relevant synthetic routes to the building blocks of life have been established. However, it is still unclear how functional chemical systems evolved with direction using only the interaction between inherent molecular chemical reactivity and the abiotic environment. Here we demonstrate how complex systems of chemical reactions exhibit well-defined self-organization in response to varying environmental conditions. This self-organization allows the compositional complexity of the reaction products to be controlled as a function of factors such as feedstock and catalyst availability. We observe how Breslow's cycle contributes to the reaction composition by feeding C_2 building blocks into the network, alongside reaction pathways dominated by formaldehyde-driven chain growth. The emergence of organized systems of chemical reactions in response to changes in the environment offers a potential mechanism for a chemical evolution process that bridges the gap between prebiotic chemical building blocks and the origin of life.

The origin of life from the prebiotic environment required extensive development of increasingly sophisticated chemical systems, with only environmental factors and inherent chemical activity as the driving forces[1]. Conditions on early Earth allowed for a wide range of chemical reactions that could have given rise to a diverse range of structurally complex organic molecules[2–7]. Recent work has shown how these simple starting materials may produce many of the components found in extant metabolic networks, hinting at the possibility of a prebiotic origin of core metabolism[4,8,9]. Yet, for a functional genetic or metabolic system to emerge from a mixture of feedstock molecules, increasingly organized and interconnected reaction pathways must be forged between them[10–13]. Chemical reactivity alone is not sufficient to dictate the formation of one pathway over another[14–16]. Therefore, it is difficult to conceive how, on prebiotic Earth, chemical systems became organized to avoid the formation of intractable mixtures. Environmental conditions could provide a directing force for the emergence of (pre) biotic systems[1,17]. There is a dearth of understanding and experimental studies of how reactive and environmental information translates into the self-organization of chemical reaction networks. Therefore, creating a conceptual framework in which chemical reactivity and reaction conditions combine to organize dynamic, out-of-equilibrium reaction networks is key to elucidating how inanimate matter evolved towards life.

Here we demonstrate how simple sets of reactions between chemical feedstocks present in a model prebiotic reaction system form well-defined compositional patterns via the interaction between environmental conditions and chemical reactivity rules. Employing the formose reaction as a model system in a series of flow reactions, we studied the responsiveness of the system to a broad range of environmental factors. By characterizing the propagation of periodic input modulations through chemical reaction networks[18–21], we were able to infer the underlying reaction connectivity between formose products and unravel the self-organizational response of the network to environmental conditions. Our results demonstrate how patterns of chemical reactivity and environmental conditions may give rise to organized systems of chemical reactions, offering the first glimpse of a possible mechanism for chemical evolution.

Results and discussion

The formose reaction as a model prebiotic system. The formose reaction is a prebiotically plausible model system of sugar-forming reactions using formaldehyde as a C_1 building block (Fig. 1a)[22–24]. It broadly consists of five reactions: enolate formation/protonation, aldol addition, retro-aldol reaction and Cannizzaro reaction (Extended Data Fig. 1). Much of its core reactivity is catalysed by hydroxide and divalent metal ions such as Ca^{2+} (ref. [25]). Conceptually, any given monosaccharide or enol(ate) compound in the formose reaction may be converted into another via application of the aforementioned reaction types. Therefore, a range of feedstock monosaccharides may be used to initiate the reaction.

Unconstrained, recursive application of the limited set of reaction classes operative in the formose reaction produces a so-called combinatorial explosion of compounds (Fig. 1a). A number of studies have explored means to contain the potential generation of intractable mixtures of compounds formed in the formose reaction using thermodynamic constraints in batch reactions[26–28]. However, relatively little data have been collected for comprehensively rationalizing the formose reaction under out-of-equilibrium conditions[24,28–32]. Such conditions are a key characteristic of living systems and of great relevance on prebiotic Earth[12], upon which the conditions were dynamic and modulated on a variety of time scales. In out-of-equilibrium chemical reactions, kinetic, rather than thermodynamic, properties govern the reaction behaviour and product distribution[11]. Therefore, molecular reactivity is a prime controller of such systems.

Investigating the compositional response of the formose reaction to environmental conditions. In this study, out-of-equilibrium conditions were induced in the formose reaction using flow

Institute for Molecules and Materials, Radboud University Nijmegen, Nijmegen, Netherlands. ✉e-mail: w.huck@science.ru.nl

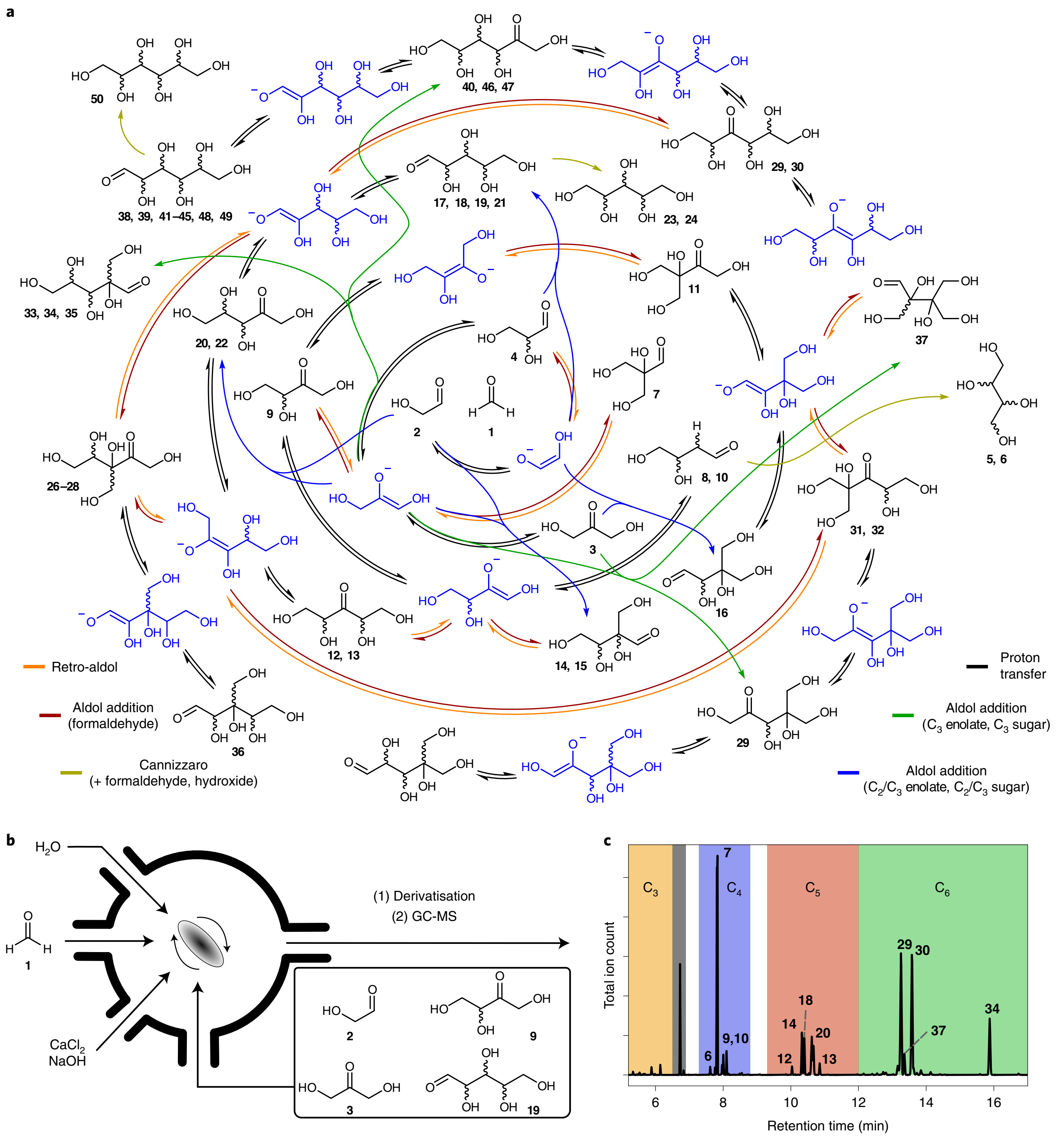


Fig. 1 | Background to this work. a, A schematic overview of the formose reaction. Key reaction pathways are coloured according to the reaction type. Details of reaction types are given in Extended Data Fig. 1. Wavy lines indicate stereogenic centres. **b**, A schematic drawing of the experimental set-up. Syringe pumps containing formaldehyde, an initiator sugar, $CaCl_2$, NaOH and water are connected to the inlets of a CSTR. A more detailed diagram is given in Extended Data Fig. 2. The outlet of the reactor was sampled continually. **c**, An example GC–MS chromatogram indicating reactor composition. Regions of the chromatogram are coloured according to carbon chain length. Peaks are labelled according to the compound numbers in **a** and Extended Data Fig. 3.

conditions in a continuous stirred-tank reactor (CSTR, Fig. 1b). The compositional and reaction connectivity responses of this model system were studied in response to variations in an overarching environment. The environment, a collection of 17 input variables, included concentration variations of formaldehyde, $CaCl_2$ and NaOH, temperature and the nature of the initiating sugar (glycolaldehyde, dihydroxyacetone, erythrulose or ribose) (Supplementary Data 1–4 and Methods).

Experiments were performed to measure steady-state equilibrium compositions of the formose reaction. In addition, investigations were also performed in which the input concentrations of initiating sugars were modulated sinusoidally. Measuring the transfer of input modulation to product compounds (Supplementary Data 4) provided a handle on which estimations of the underlying reaction pathways of the formose reaction could be based[18–20].

The composition of the CSTR was continually sampled from its outflow. Following appropriate derivatization[33–35], samples were analysed by gas chromatography–mass spectrometry (GC–MS) and HPLC (Fig. 1c; see Methods for a comparison of the two methods). Analysis of the chromatographic peaks and mass spectra provided a compositional pattern for each sample (Supplementary Data 3; five examples are shown in Fig. 2a), comprising of varying amounts of the 52 compounds detected within the dataset.

Trends in compositional data are revealed by hierarchical clustering. To visualize the relationships between the average compositions and kinetic signatures generated for each condition, hierarchical clustering was performed using a correlation-based pairwise dissimilarity metric (Methods). The resulting dendrogram (depicted qualitatively in Fig. 2b; see also Supplementary Fig. 1) represents the relative relationships between reaction outcomes. Pie plots placed on the 'leaf' positions represent normalized average product distributions. Longer paths between leaves represent lower similarity.

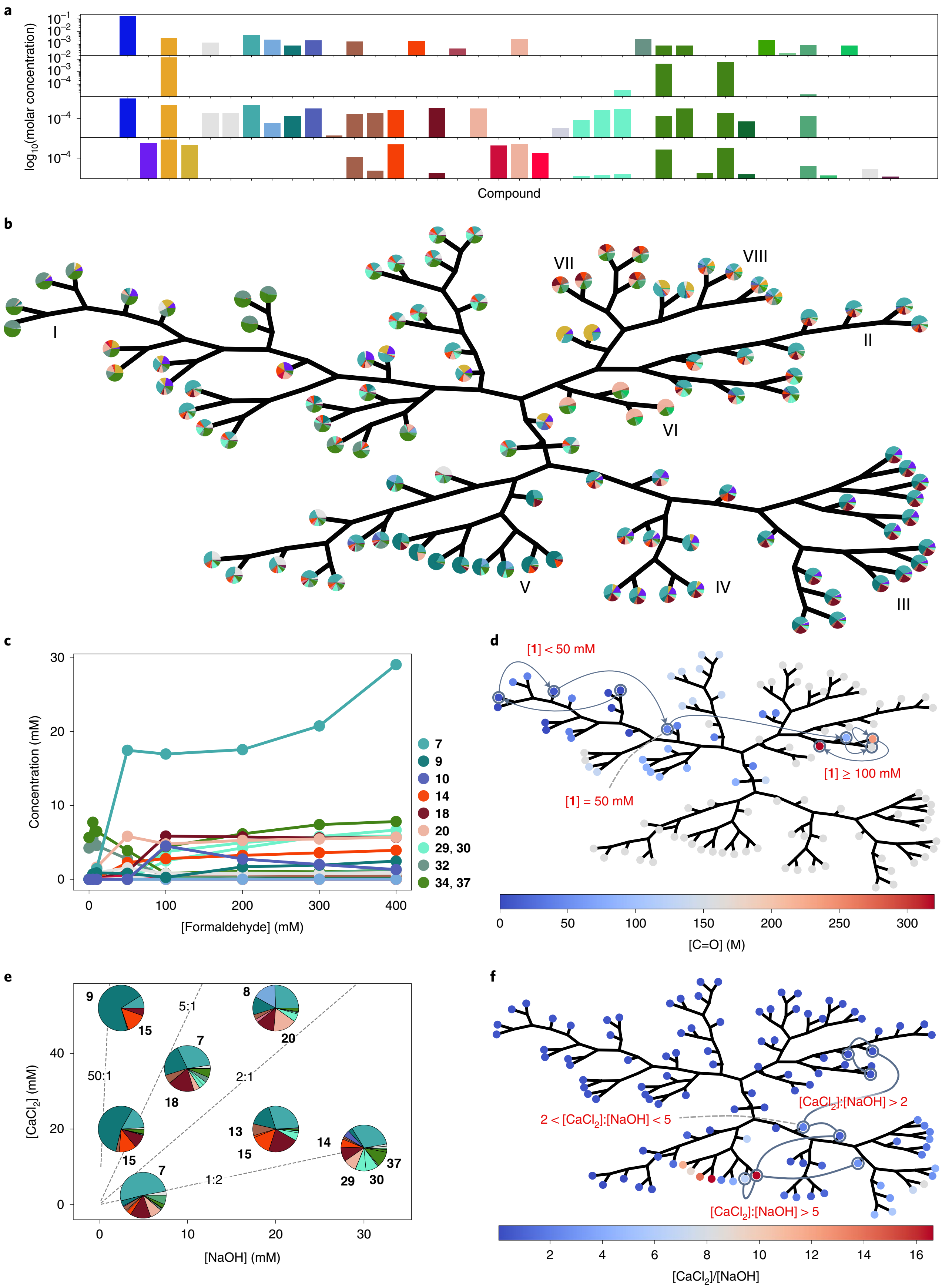
Each branch of the dendrogram arises as a result of dominant environmental factors, or combinations thereof (see Extended Data Fig. 4 for how key conditional variations map to the dendrogram). Fine-tuning of compositions within branches results from the mixing of additional condition variables. For instance, branch I is characterized by relatively low concentrations of formaldehyde (**1**, ≤50 mM) and inputs combining glycolaldehyde (**2**) dihydroxyacetone (**3**) and erythrulose (**9**). Its constituent compositions are distinguished by relatively high amounts of C_6 compounds (green sector hues). Following branch I from its tip towards the centre of the tree, more diverse sets of compounds are produced, including both branched and linear C_4 (denoted by hues of blue) and C_5 compounds (denoted by hues of red). Interestingly, varying the concentration of **1** (with fixed inputs of **3**, $CaCl_2$ and NaOH of 50, 15 and 30 mM, respectively) results in a series of compositional transitions (Fig. 2c), manifesting as a series of 'jumps' of varying magnitude across the dendrogram (Fig. 2d). Beginning in branch I ([**1**] ≤ 50 mM), the compositions consist mostly of an α-hydroxymethyl-aldohexose (**32**) and an α,β-(hydroxymethyl)-aldotetrose (**37**). Within the series of experiments with [**1**] ≤ 50 mM, the composition remains in branch I but the molecular diversity increases as [**1**] increases. The contributions of **32** and **37** in the reaction mixture decrease in favour of the generation of α-hydroxymethyl-glyceraldehyde (**7**), **9** and ribulose (**20**). Increasing [**1**] to 50 mM results in a jump towards the centre of the tree, suggesting the beginning of a more notable compositional transition. Compounds **7**, **20** and an α-hydroxymethyl-aldotetrose (**14**) become prominent. Further increasing [**1**] to above 50 mM results in a substantial compositional transition from branch I to branch II, highlighting a transition in the molecular complexity of the system. Threose (**10**), **18**, two 3-ketohexoses (**29**, **30**) and a new α-hydroxymethyl-aldopentose (**34**) are added to the composition. Compounds **32** and **37** are almost completely depleted. Thus, the concentration of feedstock molecule **1** controls a thresholded compositional transition whose dynamic range is in the region [**1**] = 0–100 mM.

Ca^{2+} and hydroxide are involved in several reaction types of the formose reaction. $Ca(OH)_2$ has been noted to have greater activity in catalysing the formose reaction than Sr^{2+} or Ba^{2+} hydroxides[36]. $Fe(OH)_3$ has also been reported to be active in catalysing formose reactions[37]. Although definitive characterization is not yet available, it is generally understood that Ca^{2+} binds divalently to enolates[36,38], and may participate in organizing intermediates in aldol addition reactions. The principal role of hydroxide is in α-proton abstraction and enolate formation from monosaccharides, which has been noted to be a key rate-limiting step in the reactions of this class of compounds[39–41]. Hydroxide is also involved as a reactant in Cannizzaro reactions.

A range of [Ca^{2+}]:[NaOH] input ratios (remaining below the solubility limit of $Ca(OH)_2$) were explored in the dataset, maintaining fixed concentrations of **1** and **3**. A demonstrative subset of the data (Fig. 2e) crosses three branches of the dendrogram (II, III and IV, Fig. 2f). Beginning at low [Ca^{2+}]:[NaOH] (in branch II), compositions similar to those previously described in the high [**1**] regime are found. Increasing the [Ca^{2+}]:[NaOH] ratio induces a jump from branch II to III due to increases in the relative proportions of **9** and **18** in comparison to **7**. Further increasing the ratio lowers the population of **7** and **18**, eventually creating a composition dominated by **9**. Notably, compositions recorded for [NaOH] = 2.5 mM and [Ca^{2+}] = 20–52 mM are compositionally very similar, mainly dominated by **9**. Furthermore, at [$CaCl_2$] = 52 mM and [NaOH] = 20 mM, erythrose (**8**) and **20** are particularly prominent.

Other environmental conditions, such as varying the initiator sugar identity, lead to distinctive compositions. Branches V, VI and VII result from using **9**, ribose (**19**) and the dimer of **2**, respectively, as initiators. When the temperature in the reactor is increased from 10 to 40 °C (branch II), the reaction composition remains remarkably unaltered (Extended Data Fig. 5). At 10 °C there is a relatively lower concentration of **10** in comparison to higher temperatures. There is also a slight divergence of the concentrations of **18** and **20** with increasing temperature. Therefore, the influence of temperature on the steady-state composition of the formose reaction is modest in the range investigated. Varying the residence time between 1 and 8 min (Extended Data Fig. 6) led to compositions rich in lyxose (**18**) and ribulose (**20**) at low retention times. The broadest spread of product carbon chain lengths was observed for a residence time of 2 min, in which **9**, **10** (threose), **14** (an α-branched aldotetrose), **37** (and α-branched aldopentose), and 3-ketohexoses **29** and **30** are particularly prominent. As the residence time was increased beyond 2 min, the amount of **18** observed was substantially reduced.

Fig. 2 | Description of the reaction composition dataset collected in this work. a, Example product distributions demonstrating the compositional diversity covered by the datasets. Bars are coloured according to compound, with C_4 alcohols in grey, C_4 sugars in blue, C_5 compounds in red and C_6 compounds in green. **b**, The relationships between product compositions and compound response profiles in the dataset when clustered as described in the Methods. Pie plots represent experimental product composition, coloured similarly to **a**. Roman numerals indicate branches as discussed in the main text. **c**, The variation in composition of the formose reaction in response to increasing formaldehyde concentration. The numbers in the legend correspond to compounds. The data were obtained for the formaldehyde concentrations indicated on the *x* axis, each with a residence time of 2 min, and inputs of dihydroxyacetone (amplitude, 25 mM; offset, 50 mM, period, 6 min), $CaCl_2$ (15 mM) and NaOH (30 mM), 21 °C. Input concentrations are quoted as the initial concentration of compounds upon entering the CSTR. **d**, The compositional path taken through the dendrogram as a function of formaldehyde concentration. The underlying leaf nodes are coloured according to the formaldehyde concentration. **e**, The variation in composition of the formose reaction in response to varying $CaCl_2$ and NaOH concentrations. Pie charts are presented similarly to those in **b** and [$CaCl_2$]:[NaOH] ratios are indicated as annotated dashed lines. The data were obtained using the $CaCl_2$ and NaOH concentrations given on the *x* and *y* axes, each with a residence time of 2 min, and inputs of dihydroxyacetone (amplitude, 25 mM; offset, 50 mM; period, 6 min) and formaldehyde (200 mM), 21 °C. Input concentrations are quoted as the initial concentration of compounds upon entering the CSTR. **f**, The compositional path taken through the dendrogram as [$CaCl_2$] and [NaOH] are varied. The underlying leaf nodes are coloured according to [$CaCl_2$]:[NaOH].


a
log$_{10}$(molar concentration)
10^{-1}
10^{-2}
10^{-3}
10^{-2}
10^{-3}
10^{-4}
10^{-4}
10^{-4}
Compound
b
I
II
III
IV
V
VI
VII
VIII
c
Concentration (mM)
30
20
10
0
0
100
200
300
400
[Formaldehyde] (mM)
7
9
10
14
18
20
29, 30
32
34, 37
d
[1] < 50 mM
[1] = 50 mM
[1] ≥ 100 mM
0
50
100
150
200
250
300
[C=O] (M)
e
[CaCl$_2$] (mM)
40
20
0
0
10
20
30
[NaOH] (mM)
9
15
5:1
8
20
7
50:1
18
2:1
13
15
14
37
15
7
1:2
29
30
f
[CaCl$_2$]:[NaOH] > 2
2 < [CaCl$_2$]:[NaOH] < 5
[CaCl$_2$]:[NaOH] > 5
2
4
6
8
10
12
14
16
[CaCl$_2$]/[NaOH]

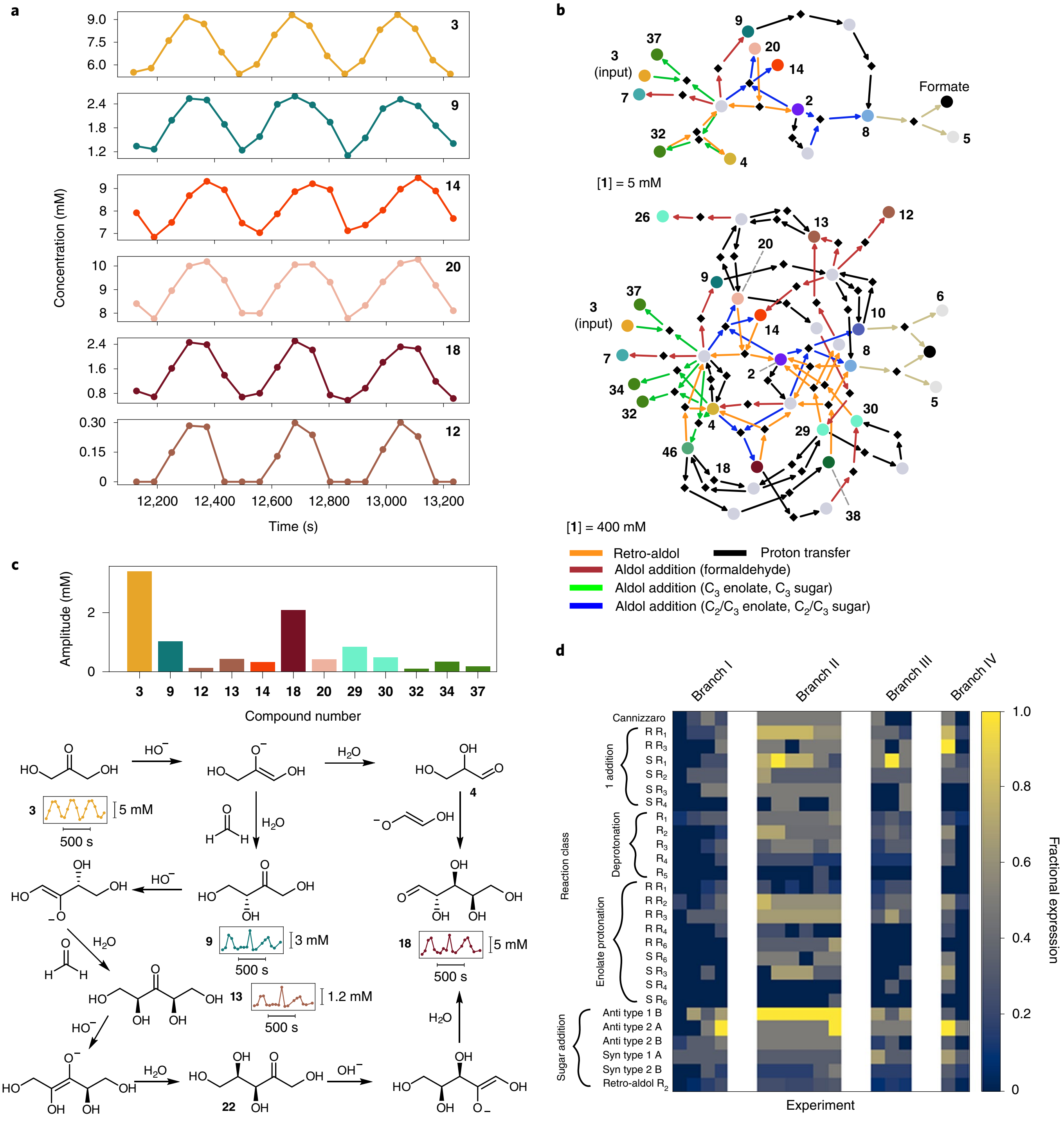


Fig. 3 | Elucidation of how the formose reaction network reorganizes in response to varying conditions. **a**, Selected time courses from a single experiment. The traces are numbered according to the structures give in Fig. 1a. The data were obtained from a flow reaction with a residence time of 2 min, 21 °C, with inputs of dihydroxyacetone (amplitude, 25 mM; offset, 50 mM; period, 6 min), formaldehyde (200 mM), $CaCl_2$ (25 mM) and NaOH (50 mM). Input concentrations are quoted as the initial concentration of compounds upon entering the CSTR. **b**, Bipartite graph representations of how the reaction pathways of the formose reaction vary with increasing formaldehyde concentration. Reactions (black diamonds) connect by directed edges to reactants and products. The dots are coloured according to compound (compound numbers are annotated). The key indicates the colour code for arrows connecting compounds and reactions. **c**, Top: a bar chart of the amplitudes determined for compounds in a flow reaction between **1** and **3**. The amplitude data were determined from a flow reaction with a residence time of 2 min, 21 °C, with inputs of dihydroxyacetone (amplitude, 25 mM; offset, 50 mM; period, 6 min), formaldehyde (300 mM), $CaCl_2$ (15 mM) and NaOH (30 mM). Input concentrations are quoted as the initial concentration of compounds upon entering the CSTR. Bottom: reaction scheme illustrating two possible routes to **18** from **3**. Panels **a**,**b** and **c** use similar colouring schemes for compounds. **d**, A heat map showing how the determined reaction types vary between branches of the dendrogram in Fig. 2b. Fractional expression refers to the counts of the reaction class observed compared to the number of reactions in the class in a reaction network representing the union of the reactions found for all datasets.

The concentrations of **20** and **30** continued to contribute strongly to the overall composition. The concentration of **14** and **6** (threitol) increased alongside **32**, an epimer of **37**.

Inference of the formose chemical reaction network structure. The observed compositional variations are a direct result of the translation of the various input conditions through the underlying formose reactions. To elucidate how the network reorganizes in response to changes in input conditions, we performed experiments in which the input flow of the initiating sugar was modulated to probe the reaction connectivity in product compositions (Fig. 3a). First, a 'global' reaction network was generated using a rule-based reaction network generation strategy[42–44] in which a set of reaction rules were applied iteratively to an initial set of compounds (for example, **1**, the dimer of **2** and NaOH) in silico, thereby creating a combinatorial explosion of compounds and reaction pathways connecting them (Methods). Second, it has been shown that the amplitudes of the variation in product concentration are a function of the input amplitude and the number and rate of reactions between the input and each product[18–21]. We used the number of reactions between the initiator sugar and the various products as a basis for guiding searches of an overarching reaction network generated in silico for subsets of reactions that describe the data. To perform the pathway search for each modulated experiment, detected compounds were listed in order of decreasing amplitude. The generated reaction network was converted into a directed graph representation, with nodes corresponding to compounds and reactions[45,46]. Within this graph, the shortest paths between consecutive members of the amplitude-ordered list (omitting compounds that were not present in the product composition) were found and combined (Methods). Examples of two pathways are given in Fig. 3b, which depicts the formose reaction as bipartite graphs in which dots represent compounds and diamonds represent reactions. This choice of representation is beneficial in depicting complex systems of chemical reactions because reactions involving multiple reactants and products can connect compounds in a much simpler manner than traditional reaction schemes. These sets of reactions offer a hypothesis for the mechanism which created the observed product composition based on the chemical reaction types outlined in the reaction rules used to generate the global reaction network. As such, this framework provides a direct translation of the experimental data into a descriptive set of reactions responsible for the compositions observed.

The search procedure was used to estimate the self-organizational response of the formose reaction pathways in response to increasing formaldehyde concentration (initiated by **3**; Fig. 3b). This depiction represents the underlying structure of the formose network when [**1**] is 5 mM. The operative reaction pathways are mainly accounted for by a small set of reactions between C_3 species to form C_6 compounds. Increasing [**1**] to 400 mM triggers expansion of the repertoire of reactions. The number of possible pathways for formaldehyde addition increases, with a corresponding increase in the number of proton transfer pathways. However, **1**-based chain growth pathways do not appear to completely account for the observed behaviour. Although compounds **12** and **13** (3-ketopentoses) have lower amplitudes than **9** (Fig. 3c, bar chart), consistent with chain growth via formaldehyde addition, **18** couples with comparatively more strength to the input modulation. A shortest formaldehyde-based chain growth pathway to **18** would proceed via **13** (the sequence of reactions passing from **3** to **18** via **9**, **13** and **21**; Fig. 3c). Therefore, we attribute the production of **18** to the reaction between the enolate of **2** and **4** (Fig. 3c, right-hand pathway)[47]. It is also possible for **14**, **15** and **20** to be generated from similar pathways between the enolate and carbonyl forms of **2** and **3** or **4**. Indeed, increasing the $Ca(OH)_2$ concentration can induce strong coupling of **14** and **20** to the input modulation (Fig. 3a), which also implies that **12** and **13** would be bypassed. Formaldehyde addition can build on top of such enolate–monosaccharide pathways to produce **29** and **30** as a product of the formaldehyde addition to enolates derived from **18** and **20**.

Chemical reaction network structure can be viewed in terms of the activation of specific reaction types. The formose reaction can also be viewed in terms of the types of reaction activated in response to varying conditions. Each reaction was assigned a class based on the reaction rule used to generate it during construction of the global network in silico (as described above). Recasting the lists of reactions as counts of each reaction class provides a condensed view of how formose reactivity adapts to environmental conditions (Fig. 3d).

Following the branches of the dendrogram traversed in Fig. 2d (variation in [**1**]) reveals key reaction characteristics that govern the various reaction outcomes. An important feature of branch I is the relatively low proportion of formaldehyde aldol addition reactions. The majority of the reactivity is accounted for by monosaccharide–enolate reactions between C_3 compounds, which are responsible for creating products **32** and **37** (Fig. 4a, branch I)[40,48]. Moving to branch II (higher [**1**]), the repertoire of reactions is expanded, and aldol addition reactions involving **1** are added to the network. In particular, reactions in which the α-carbon is bound to a hydrogen or glycol group are promoted. A range of protonation/deprotonation reactions are promoted in branch II. Deprotonation is favoured at less sterically hindered positions (where the α-carbon is bound to a hydrogen or hydroxymethyl group, for example, following the sequence **3**, **9**, **12**, **20**, **29** in Fig. 4a, branch II). Protonation is favoured at α-carbons bound to methoxy groups. Interestingly, the number of monosaccharide–enolate reactions also increases, suggesting that some monosaccharide products interact with other members of the network as reactants.

The formose reaction reorganizes in a different manner when the [Ca^{2+}] and [NaOH] are varied. At high [Ca^{2+}] (52 mM) and low [NaOH] (2.5 mM) a limited set of pathways is formed, the majority of which may be accounted for via formaldehyde addition and proton transfer reactions terminating at **15** via **9** (the pathway connecting **3**, **9** and **15** in Fig. 4a). Interestingly, the linear C_5 compounds **12** and **13** are not formed in appreciable quantities, so the reaction hits the 'dead end' branched **15**. The system unexpectedly avoids the formation of the branched C_4 compound **7** under these conditions. Decreasing the [Ca^{2+}]:[NaOH] ratio (35 mM:10 mM) leads to a larger contribution of formaldehyde-controlled pathways. The branched C_4 **7** is created, while the branched C_5 **15** is demoted. Simultaneously, the population of C_5 species and instances of Cannizzaro reactions are increased. As the ratio is further decreased, the conditions and reaction pathways begin to resemble those found for the high [**1**] region, as described above.

Explanation of the concentration ratios of some diastereoisomers. Across the data collected, there is a general trend of selectivity towards **20** over **22** (xylulose) and **14** over **15**. These selectivities can be rationalized using the Traxler–Zimmerman model of aldol addition reactions (Extended Data Fig. 7). During addition of **2** to the enolate of **3** (there are two possible positions), six-membered ring transition states are formed. Their lowest-energy conformations, in which the α-hydroxymethyl group of **2** is axial and the enolate approaches in a Z-conformation, results in the preferred generation of **14** (for aldol addition at the central carbon of the enolate; Extended Data Fig. 7a) and **20** (aldol addition to the terminal carbon of the enolate; Extended Data Fig. 7b).

The varying ratio between **12** and **13** as a function of [Ca^{2+}]:[NaOH] suggests a second role for Ca^{2+}. At a ratio of 1:2, the two diastereoisomers are close in concentration, as expected due to the non-specific proton-transfer or aldol-addition reactions

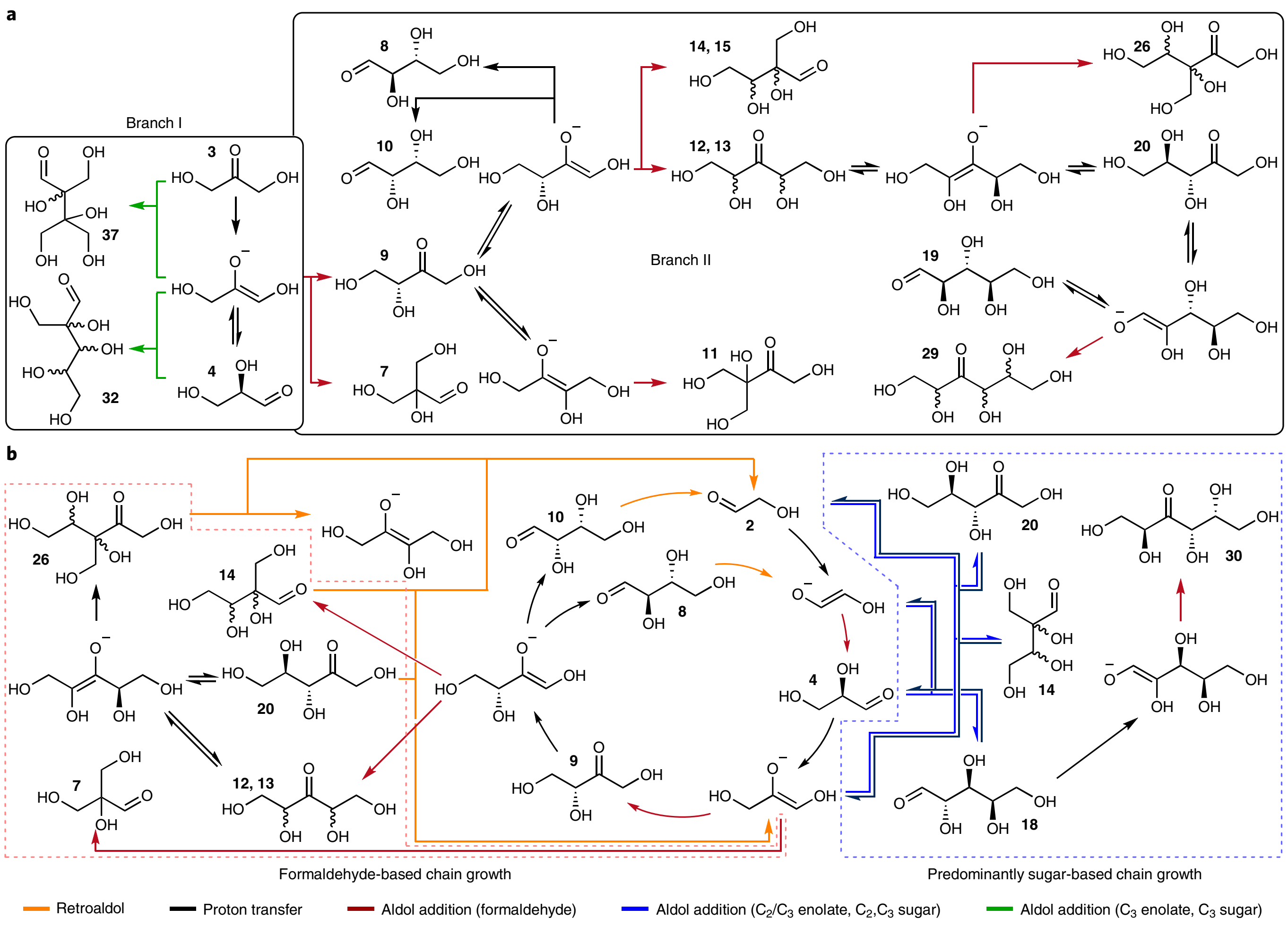

a
Branch I
Branch II
37
3
4
32
8
10
9
7
14, 15
12, 13
11
26
20
19
29
b
26
14
20
7
12, 13
10
2
8
9
4
20
14
30
18
Formaldehyde-based chain growth
Predominantly sugar-based chain growth
Retroaldol
Proton transfer
Aldol addition (formaldehyde)
Aldol addition (C_2/C_3 enolate, C_2,C_3 sugar)
Aldol addition (C_3 enolate, C_3 sugar)

Fig. 4 | A summary of key reaction pathways governing the behaviour of the formose reaction. a, The shift in reaction pathways from branch I to branch II is controlled by formaldehyde concentration. **b**, Relation of the apparent reaction pathways to the Breslow cycle[51], indicating how the cycle may act as a source of C_2 building blocks for the formose reaction while its constituents may be involved in off-cycle formaldehyde chain-growth reaction pathways. The reaction arrows are coloured according to the legend below the figure.

that lead to them. High concentrations of Ca^{2+} seem to favour the formation of **13** over **12**. This selectivity can be explained by considering that the 3-ketopentoses' two α-hydroxymethyl groups may coordinate to form six-membered rings (Extended Data Fig. 8). In the case of **12**, this ring contains an axial hydroxymethyl group, whereas for **13**, both groups are equatorial. Thus, **13** may be a more thermodynamically favourable product than **12** at sufficiently high [Ca^{2+}]. This observation highlights that Ca^{2+} can play a role both as catalyst and as a thermodynamic directory of product distribution via selective binding.

The Breslow cycle is a key contributor to formose reaction network structure and composition. In contrast to prevailing views on the formose reaction, our data indicate that formaldehyde-based chain growth pathways do not completely account for the observed behaviour. Rather, reactions between C_2 and C_3 compounds are key chain-building reactions[22,47,49,50]. Surprisingly, we observe the emergence of a self-organized cyclic set of reactions that explain how the C_2 monosaccharide **2** must be created from retro-aldol reactions, as described in Breslow's proposed mechanism for autocatalysis in the formose reaction (Fig. 4b)[51]. Although usually seen as an autocatalytic mechanism, our results show how this cycle of reactions directs the composition of the formose reaction by generating **2**, **3** and **4** (and their enolates). Furthermore, as suggested by the retro-aldol reaction pathways shown in Figure 3b, C_6 compounds such as **29** and **30** may also act as sources of C_2 and C_4 monosaccharides and enolates which bolster the Breslow cycle's constituents. The C_2 building blocks emerge from the formose reaction to create an alternative chain-growth mechanism embedded in another in which chain growth occurs via formaldehyde addition. As such, the Breslow cycle can be envisaged as a source of new reaction pathways through which monosaccharides may be built. These reactions between formose reaction products are an excellent example of how underlying patterns in chemical reactivity define reaction outcomes. Thus, we propose that reinforcement of molecular diversity in the formose reaction does not necessarily occur via promotion of an initiating species (glycolaldehyde). Rather, diversity may be promoted by the activation of a class of reactions in which longer carbon chains are synthesized from building blocks with units of greater than one carbon.

Conclusion

We have demonstrated how an environment directs the self-organization of reaction pathways in the model prebiotic formose reaction network to create well-defined product compositions. Our analysis is based on detailed HPLC and GC–MS characterization of the formose reaction under out-of-equilibrium conditions, combined with graph pathway analysis informed by chemical knowledge.

In the absence of a directing force, the recursive application of the small number of chemical reactions operative in the formose reaction potentially leads to a wide range of possible reaction pathways and compositions. However, we have found that in the diverse environments studied, the formose reaction is induced into using only a subset of these pathways, depending on the local conditions. This refinement of reaction pathways results in well-defined product compositions, in contrast to the intractable mixtures expected of a combinatorially detonated reaction route. We were able to estimate the organization of the underlying reaction pathways via analysis of the time-resolved propagation of periodically changing inputs. This analysis revealed that sets of reactions can respond collectively to environmental conditions. Furthermore, we have shown how the Breslow cycle, an autocatalytic reaction pathway, is an important director of composition. It provides an alternative, emergent set of reactions that allow sugars to be built from C_2 and C_3 building blocks, alongside those that utilize reactions between formaldehyde and C_{2-6} compounds.

The self-organization we observe is reminiscent of theoretically predicted fine-tuning of reaction network behaviour to environments[52,53]. Our work represents an important advance in understanding how molecular systems adapt abiotically to the environment. Environmental conditions control reaction types to varying degrees in complex abiotic reaction networks, leading to well-defined reaction pathways and product compositions. Such translation of information from the environment into its embedded chemical reaction networks hints at how reaction networks of biological importance may be the result of the abiotic self-organization of systems of reactions in the absence of specific catalysis or genetic inheritance. Applying our methodology to other prebiotically relevant reaction networks that include cyanosulfidic[54] or α-ketoacid[8,9] reactivity could shed new insight into an environment-driven formation of a primitive core metabolic networks furnishing the building blocks of life.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41557-022-00956-7.

Received: 2 August 2021; Accepted: 26 April 2022;
Published online: 6 June 2022

References

1. Surman, A. J. et al. Environmental control programs the emergence of distinct functional ensembles from unconstrained chemical reactions. *Proc. Natl Acad. Sci. USA* **116**, 5387–5392 (2019).
2. Powner, M. W., Gerland, B. & Sutherland, J. D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459**, 239–242 (2009).
3. Foden, C. S. et al. Prebiotic synthesis of cysteine peptides that catalyze peptide ligation in neutral water. *Science* **370**, 865–869 (2020).
4. Springsteen, G., Yerabolu, J. R., Nelson, J., Rhea, C. J. & Krishnamurthy, R. Linked cycles of oxidative decarboxylation of glyoxylate as protometabolic analogs of the citric acid cycle. *Nat. Commun.* **9**, 91 (2018).
5. Becker, S. et al. Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides. *Science* **366**, 76–82 (2019).
6. Ritson, D. J., Mojzsis, S. J. & Sutherland, J. D. Supply of phosphate to early Earth by photogeochemistry after meteoritic weathering. *Nat. Geosci.* **13**, 344–348 (2020).
7. Wołos, A. et al. Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. *Science* **369**, eaaw1955 (2020).
8. Muchowska, K. B., Varma, S. J. & Moran, J. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019).
9. Stubbs, R. T., Yadav, M., Krishnamurthy, R. & Springsteen, G. A plausible metal-free ancestral analogue of the Krebs cycle composed entirely of α-ketoacids. *Nat. Chem.* **12**, 1016–1022 (2020).
10. Wu, L.-F. & Sutherland, J. D. Provisioning the origin and early evolution of life. *Emerging Top. Life Sci.* **3**, 459–468 (2019).
11. Pascal, R., Pross, A. & Sutherland, J. D. Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open Biol.* **3**, 130156 (2013).
12. Semenov, S. N. et al. Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature* **537**, 656–660 (2016).
13. Semenov, S. N. et al. Rational design of functional and tunable oscillating enzymatic networks. *Nat. Chem.* **7**, 160–165 (2015).
14. Jinich, A. et al. A thermodynamic atlas of carbon redox chemical space. *Proc. Natl Acad. Sci. USA* **117**, 32910–32918 (2020).
15. Orgel, L. E. Self-organizing biochemical cycles. *Proc. Natl Acad. Sci. USA* **97**, 12503–12507 (2000).
16. Shapiro, R. Prebiotic ribose synthesis: a critical analysis. *Orig. Life. Evol. Biosph.* **18**, 71–85 (1988).
17. Sasselov, D. D., Grotzinger, J. P. & Sutherland, J. D. The origin of life as a planetary phenomenon. *Sci. Adv.* **6**, eaax3419 (2020).
18. Samoilov, M., Arkin, A. & Ross, J. Signal processing by simple chemical systems. *J. Phys. Chem. A* **106**, 10205–10221 (2002).
19. Roszak, R., Bajczyk, M. D., Gajewska, E. P., Hołyst, R. & Grzybowski, B. A. Propagation of oscillating chemical signals through reaction networks. *Angew. Chem. Int. Ed.* **58**, 4520–4525 (2019).
20. Urmès, C. et al. Periodic reactor operation for parameter estimation in catalytic heterogeneous kinetics. Case study for ethylene adsorption on Ni/Al_2O_3. *Chem. Eng. Sci.* **214**, 114544 (2020).
21. Mettetal, J. T., Muzzey, D., Gomez-Uribe, C. & van Oudenaarden, A. The frequency dependence of osmo-adaptation in *Saccharomyces cerevisiae*. *Science* **319**, 482–484 (2008).
22. Kim, H.-J. et al. Synthesis of carbohydrates in mineral-guided prebiotic cycles. *J. Am. Chem. Soc.* **133**, 9457–9468 (2011).
23. Delidovich, I. V., Simonov, A. N., Taran, O. P. & Parmon, V. N. Catalytic formation of monosaccharides: from the formose reaction towards selective synthesis. *ChemSusChem* **7**, 1833–1846 (2014).
24. Simonov, A. N. et al. Selective synthesis of erythrulose and 3-pentulose from formaldehyde and dihydroxyacetone catalyzed by phosphates in a neutral aqueous medium. *Kinet. Catal.* **48**, 550–555 (2007).
25. Iqbal, Z. & Novalin, S. The formose reaction: a tool to produce synthetic carbohydrates within a regenerative life support system. *Curr. Org. Chem.* **16**, 769–788 (2012).
26. Lambert, J. B., Gurusamy-Thangavelu, S. A. & Ma, K. The silicate-mediated formose reaction: bottom-up synthesis of sugar silicates. *Science* **327**, 984–986 (2010).
27. Ricardo, A., Carrigan, M., Olcott, A. & Benner, S. Borate minerals stabilize ribose. *Science* **303**, 196–196 (2004).
28. Colón-Santos, S., Cooper, G. J. T. & Cronin, L. Taming the combinatorial explosion of the formose reaction via recursion within mineral environments. *ChemSystemsChem* **1**, e1900014 (2019).
29. Huskey, W. P. & Epstein, I. R. Autocatalysis and apparent bistability in the formose reaction. *J. Am. Chem. Soc.* **111**, 3157–3163 (1989).
30. Weiss, A. H., LaPierre, R. B. & Shapira, J. Homogeneously catalyzed formaldehyde condensation to carbohydrates. *J. Catal.* **16**, 332–347 (1970).
31. Weiss, A. H., Socha, R. F., Likholobov, V. A. & Sakharov, M. M. Formose sugars from formaldehyde. *Appl. Catal.* **1**, 237–246 (1981).
32. Kopetzki, D. & Antonietti, M. Hydrothermal formose reaction. *New J. Chem.* **35**, 1787 (2011).
33. Haas, M., Lamour, S. & Trapp, O. Development of an advanced derivatization protocol for the unambiguous identification of monosaccharides in complex mixtures by gas and liquid chromatography. *J. Chromatogr. A* **1568**, 160–167 (2018).
34. Becker, M., Liebner, F., Rosenau, T. & Potthast, A. Ethoximation–silylation approach for mono- and disaccharide analysis and characterization of their identification parameters by GC/MS. *Talanta* **115**, 642–651 (2013).
35. Becker, M. et al. Evaluation of different derivatisation approaches for gas chromatographic–mass spectrometric analysis of carbohydrates in complex matrices of biological and synthetic origin. *J. Chromatogr. A* **1281**, 115–126 (2013).
36. Ziemecki, S., LaPierre, R. B., Weiss, A. H. & Sakharov, M. Homogeneously catalyzed condensation of formaldehyde to carbohydrates VI. Preparation and spectroscopic investigation of complexes active in formaldehyde condensation. *J. Catal.* **50**, 455–463 (1977).
37. Weber, A. L. Prebiotic sugar synthesis: hexose and hydroxy acid synthesis from glyceraldehyde catalyzed by iron(III) hydroxide oxide. *J. Mol. Evol.* **35**, 1–6 (1992).
38. Khomenko, T. Homogeneously catalyzed formaldehyde condensation to carbohydrates IV. Alkaline earth hydroxide catalysts used with glycolaldehyde co-catalyst. *J. Catal.* **45**, 356–366 (1976).
39. Kooyman, C., Vellenga, K. & De Wilt, H. G. J. The isomerization of D-glucose into D-fructose in aqueous alkaline solutions. *Carbohydr. Res.* **54**, 33–44 (1977).

40. Gutsche, C. David et al. Base-catalyzed triose condensations. *J. Am. Chem. Soc.* **89**, 1235–1245 (1967).
41. Nagorski, R. W. & Richard, J. P. Mechanistic imperatives for aldose−ketose isomerization in water: specific, general base- and metal ion-catalyzed isomerization of glyceraldehyde with proton and hydride transfer. *J. Am. Chem. Soc.* **123**, 794–802 (2001).
42. Andersen, J. L., Flamm, C., Merkle, D. & Stadler, P. F. Inferring chemical reaction patterns using rule composition in graph grammars. *J. Syst. Chem.* **4**, 4 (2013).
43. Andersen, J. L., Flamm, C., Merkle, D. & Stadler, P. F. Rule composition in graph transformation models of chemical reactions. *MATCH Commun. Math. Comput. Chem.* **80**, 661–704 (2018).
44. Simm, G. N. & Reiher, M. Context-driven exploration of complex chemical reaction networks. *J. Chem. Theory Comput.* **13**, 6108–6119 (2017).
45. Bajczyk, M. D., Dittwald, P., Wołos, A., Szymkuć, S. & Grzybowski, B. A. Discovery and enumeration of organic-chemical and biomimetic reaction cycles within the network of chemistry. *Angew. Chem. Int. Ed.* **57**, 2367–2371 (2018).
46. Kowalik, M. et al. Parallel optimization of synthetic pathways within the network of organic chemistry. *Angew. Chem. Int. Ed.* **51**, 7928–7932 (2012).
47. Harsch, G., Bauer, H. & Voelter, W. Kinetik, Katalyse und Mechanismus der Sekundärreaktion in der Schlußphase der Formose-Reaktion. *Liebigs Ann. Chem.* **1984**, 623–635 (1984).
48. Berl, W. G. & Feazel, C. E. The kinetics of hexose formation from trioses in alkaline solution. *J. Am. Chem. Soc.* **73**, 2054–2057 (1951).
49. Simonov, A. N., Pestunova, O. P., Matvienko, L. G. & Parmon, V. N. The nature of autocatalysis in the Butlerov reaction. *Kinet. Catal.* **48**, 245–254 (2007).
50. Delidovich, I. V., Simonov, A. N., Pestunova, O. P. & Parmon, V. N. Catalytic condensation of glycolaldehyde and glyceraldehyde with formaldehyde in neutral and weakly alkaline aqueous media: kinetics and mechanism. *Kinet. Catal.* **50**, 297–303 (2009).
51. Breslow, R. On the mechanism of the formose reaction. *Tetrahedron Lett.* **1**, 22–26 (1959).
52. Pross, A. & Pascal, R. How and why kinetics, thermodynamics, and chemistry induce the logic of biological evolution. *Beilstein J. Org. Chem.* **13**, 665–674 (2017).
53. Horowitz, J. M. & England, J. L. Spontaneous fine-tuning to environment in many-species chemical reaction networks. *Proc. Natl Acad. Sci. USA* **114**, 7565–7570 (2017).
54. Patel, B. H., Percivalle, C., Ritson, D. J., Duffy, C. D. & Sutherland, J. D. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat. Chem.* **7**, 301–307 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Materials. D-Threose, L-erythrose, L-erythrulose, D-xylulose, D-ribulose (aqueous solution), D-talose, L-idose (aqueous solution), L-gulose, D-allose, D-altrose and 1,3-dihydroxyacetone were purchased from Carbosynth. L-(−)-Sorbose, D-tagatose and D-psicose were purchased from TCI Europe. $CaCl_2$, NaOH, dihydroxyacetone, paraformaldehyde, glycolaldehyde, glyceraldehyde, ribose, *O*-ethylhydroxylamine hydrochloride, *N*,*O*-bis(trimethylsilyl)trifluoroacetamide, acetonitrile, trifluoroacetic acid and 2,4-nitrotrophenylhydrazine were purchased from Sigma Aldrich. Pyridine was purchased from Fluorochem. Water was obtained from a Millipore system. Formaldehyde solutions were prepared via sublimation of paraformaldehyde or depolymerization of paraformaldehyde in water via heating at 60 °C. The concentrations of the resulting formaldehyde solutions were determined by HPLC. All other chemicals were used without further purification.

Instrumentation. HPLC analyses were performed on a Shimadzu Nexera X2 instrument. Conditions: GIST C18 column (2 µm pore size, 75 × 3.0 mm), 40 °C, 0.8 ml min^{-1}, acetonitrile:water (1:1, 0.1% trifluoroacetic acid), 1 µl injection volume, ultraviolet–visible detection at 364 nm; or GWS C18 column (5 µm pore size, 250 × 4.6 mm) at 1.0 ml min^{-1}, 40 °C, 1.0 ml min^{-1}, acetonitrile:water (1:1, 0.1% trifluoroacetic acid), 1 µl injection volume, ultraviolet–visible detection at 364 nm.

GC–MS analyses were performed on a JEOL JMS-T100GCv. Gas chromatography conditions: Agilent 7890A gas chromatograph; HP-5MS column, 30 m length, 0.250 mm diameter, 0.25 mm film thickness, helium carrier gas, 1 ml min^{-1}, 1 µl split injection (1/10), injector temperature 250 °C. Temperature programme: oven temperature, 100, 170, 210, 250, 325 °C; rate, 0, 14, 4, 15, 60 °C min^{-1}; time, 2.33, 0, 0, 0, 3.75 min. Mass spectrometer conditions: JEOL AccuTOF mass spectrometer, electron impact ionization mode, ionization voltage 2,300 V, sampling rate 10 Hz. The injection syringe was rinsed before and after each injection with dichloromethane and cyclohexane. The syringe was periodically rinsed with dichloromethane manually.

Experimental methods. *Flow reactions.* CSTRs (volume, 411 or 439 µl) with five inlets and an outlet were fabricated from polydimethylsiloxane as previously reported[13]. Cetoni Nemesys syringe pumps with Hamilton syringes were used to control input flow rates.

Derivatization. Derivatization for HPLC was performed similarly to a previously reported method[33]. Samples from the CSTR outlet (35 µl) were dropped directly into a solution consisting of DNPH saturated acetonitrile (300 µl), acetonitrile (97.5 µl), water (65 µl) and HCl solution (2 M, 2.5 µl). The solutions were incubated for at least 30 min before HPLC analysis. Derivatization for GC–MS analysis was performed according to reported procedures[33–35]. Samples from the flow reactor outlet (35 µl) were flash-frozen in liquid nitrogen and freeze-dried overnight to give dry to oily residues. To each sample was added a solution of *O*-ethylhydroxylamine hydrochloride in pyridine (75 µl, 20 g l^{-1}). A solution of dodecane and tetradecane (100 µl, 1.6 mM each in pyridine) was then added to each sample. The samples were then shaken at 70 °C for 30 min. After cooling to room temperature, *N*,*O*-bis(trimethylsilyl)trifluoroacetamide (25 µl) was added to each sample. The samples were again shaken at 70 °C for 30 min. The samples were then cooled to room temperature, followed by centrifugation (3–5 min, 10,000 r.p.m.). The supernatants were decanted into sample vials for analysis by GC–MS (Instrumentation).

Chromatographic data processing. Peak integration and assignment of raw chromatographic data was performed using a program written in the programming language Python with the packages NumPy[55] and Scipy[56]. Peaks were detected using the first derivative of chromatograms and their integrals were determined using the NumPy function trapz() with subtraction of a baseline linearly interpolated between the beginning and the end of the peak. Peak assignments were performed via comparison to reference samples (Supplementary Figs. 2–4), or via interpretation of peak mass spectra (calibrated samples match known fragmentation patterns; Supplementary Figs. 5–7)[57] and retention times, aided by inference from experimental data. Integrals were converted to concentrations using quadratic calibration lines (Supplementary Figs. 8–10). When authentic samples were not available, calibrations were estimated by averaging the calibration factors of compounds of similar carbon chain length to the uncalibrated compound. In cases where two peaks were observed for a compound, the calibration for the peak with the larger integral was used.

Comparison of reaction characterizations by HPLC and GC–MS. HPLC analysis was predominantly employed for the quantification of dihydroxyacetone, glycolaldehyde and formaldehyde, which could not be quantified by GC–MS. In HPLC chromatograms, peaks attributable to longer-chain monosaccharides were poorly resolved (Supplementary Fig. 11a). Conversely, compounds with carbon chain lengths >4 could be observed and better resolved using GC–MS (Supplementary Fig. 11b). Therefore, the two methods are complementary, each providing a view of the reaction composition that the other cannot.

Discussion of the parameters chosen in this study and their magnitudes. *Selection of formaldehyde concentrations.* Within the conditions explored in this work, formaldehyde concentration was varied in the range 0–400 mM. This range was chosen based on the series of experiments depicted in Fig. 2c. Here, increasing the formaldehyde concentration beyond 200 mM had little effect on the composition of the formose reactions (with fixed conditions of 21 °C, [dihydroxyacetone] = 50 mM, [$CaCl_2$] = 15 mM, [NaOH] = 30 mM and residence time = 2 min). We chose 200 mM as a benchmark formaldehyde concentration as under these conditions a broad range of products was formed, and the system was on the edge of the dynamic range of the reaction's response to formaldehyde concentration.

Selection of calcium hydroxide concentrations. The conditions explored covered $CaCl_2$ variation in the range 1.5–52.0 mM and NaOH variation in the range 2.5–96.0 mM. Within the combinations of concentrations studied, no precipitation of $Ca(OH)_2$ was observed. Therefore, we assume that all of the reactions occurred homogeneously. Applying higher concentrations beyond the ranges used for both compounds would probably result in precipitate formation. Therefore, such regions of concentration were avoided. Increasing the NaOH concentration to above 30 mM resulted in a higher amount of formaldehyde consumption, without concurrent increases in product concentration. This effect can be attributed to the increased rate of disproportionation of formaldehyde into formate and methanol caused by Cannizzaro reactions. On the other hand, lowering the NaOH concentration below 30 mM would slow the rate of reaction, leading to less diverse collections of compounds (Fig. 3e). Therefore, a benchmark of 30 mM was chosen across the dataset. To maintain a 2:1 Ca^{2+}:HO^{-} concentration of 2:1, the benchmark concentration of 15 mM $CaCl_2$ was chosen.

Selection of residence times. We explored a variety of residence times in the range 60–480 s. At a residence time of 120 s, the product distribution was best balanced between C_4, C_5 and C_6 products. At a residence time of 60 s, only a small amount of C_6 compounds was formed. Above 120 s residence time, the amount of C_4 products, as well as lyxose, were diminished. The observation of compounds with a variety of chain lengths was a vital component in searching for reaction pathways connecting the various formose products. Furthermore, keeping the formose reaction under conditions in which the reaction was far from completion rendered it in a state that was more sensitive to changes in other reaction conditions. For these reasons, a benchmark residence time of 120 s was chosen to fulfil these criteria.

Selection of temperatures. As described in the main text, we explored temperatures in the range 10–50 °C, finding little impact on the reaction composition. We did not explore temperatures outside of this range due to the constraints of the apparatus used. A temperature of 21 °C was as a benchmark temperature for our experiments.

Selection of amplitudes. An amplitude of 25 mM was applied for the modulation of each input sugar, while more traditional steady-state experiments were performed with an input modulation of 0 mM. We did not explore other magnitudes of amplitudes as such investigations are outside the scope of this study. An amplitude of 25 mM was large enough to induce concentration modulations in the majority of observed products. Lowering the amplitude would probably prevent the observation of the transfer of the input modulation to products because the induced amplitudes may drop below the signal-to-noise ratio of the analysis workflow employed. Increasing the amplitude beyond 25 mM may have introduced strong nonlinear effects on the reaction. With the intent of modulation providing only a characterization handle for reaction connectivity, we sought to maintain moderate perturbations of the reaction.

Selection of periods. We found that varying the period of the input modulation for any of the initiators had little effect on the product composition. It has been shown that flow reactors have a frequency cut-off similar to their residence time[58]. Therefore, we employed benchmark modulation periods of three times the residence time of the experiment (360 s for a benchmark residence time of 120 s).

Data analysis methods. Python programs for the data analysis are available at https://github.com/huckgroup/formose-2021.git.

Hierarchical clustering of data. The average compositions and amplitudes for each experiment set were combined into an array. The pairwise dissimilarity between each dataset was then determined using a correlation-based metric (equation (1), scipy.spatial.distance.pdist() using the 'correlation' metric):

$$\textit{distance metric} = 1 - \frac{\sum\left((\mathbf{u}_i - \bar{\mathbf{u}})\,(\mathbf{v}_i - \bar{\mathbf{v}})\right)}{\sqrt{\sum\left(\mathbf{u}_i - \bar{\mathbf{u}}\right)^2} \cdot \sqrt{\sum\left(\mathbf{v}_i - \bar{\mathbf{v}}\right)^2}} \qquad (1)$$

where **u** and **v** are both one-dimensional vectors (arrays) of average compound concentrations and amplitudes determined for a given experiment. Overbars indicate the average of the elements of a vector.

Generation of the formose reaction space in silico. A set of reaction pathways in line with expected the expected reaction types of the formose reaction was generated using the RDKit (open-source cheminformatics; http://www.rdkit.org, June 2021). The reactions outlined in Extended Data Fig. 1 were translated into reaction SMARTS (Supplementary Data 5) which were iteratively applied to a seed set of compounds (glycolaldehyde, formaldehyde, hydroxide and water). Products of each reaction operation were fed into the next iteration. Compounds with a chain length of greater than six carbon atoms and the reactions leading to them were removed after every iteration. The resulting network corresponded to a hypothetical case of the formose reaction in which all pathways possible according to the contracting reaction rules are taken. This set of reactions was used as a framework for determining reaction pathways from data.

Pathway analysis. The generated list of reactions for the formose reaction was converted into a networkx DiGraph object[59]. Nodes corresponding to compounds were connected by directed edges to nodes for reactions. The edge direction indicated the role of the compound as either a reactant or a product in the reaction to which it is connected. Nodes corresponding to formaldehyde, hydroxide and water were removed from the graph. Data corresponding to compounds' responses to input modulations (Supplementary Data 4) were used as a guide in searching for reaction pathways as described in the main text.

To obtain lists of reactions for each dataset of compound concentration amplitudes, the following process was applied. The list of detected compounds was sorted in order of decreasing amplitude. Additional compounds, such as enolates, which could not be detected by the methods used, were appended to the bottom of the list.

From the set of generated formose reactions, those whose reactants were not present in the list of compounds were removed. Reactions whose products were inputs to the system (for example, dihydroxyacetone or formaldehyde) were also removed.

The construction of a reaction pathway began by determining single shortest paths between each carbon input into the reaction (other than formaldehyde) to a compound from the set of reaction products. Shortest paths were then found between consecutive members of the amplitude-ordered compound list. The pathways were determined in the direction of high to low amplitude. The resulting list of reactions was checked to make sure all product compounds had reactions leading to them. For each compound without an inbound reaction pathway, a connection to the rest of the reaction scheme was found by finding the shortest path to the compound from a set of those with higher amplitudes.

Data availability

All data supporting the findings of this study are available within the paper and Supplementary Data and Supplementary Information. Source data are provided with this paper.

Code availability

All programs used to analyse and plot the data are available on GitHub (https://github.com/huckgroup/formose-2021.git).

References

55. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
56. SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
57. Laine, R. A. & Sweeley, C. C. Analysis of trimethylsilyl *O*-methyloximes of carbohydrates by combined gas-liquid chromatography–mass spectrometry. *Anal. Biochem.* **43**, 533–538 (1971).
58. Meyer, D., Friedland, J., Kohn, T. & Güttel, R. Transfer functions for periodic reactor operation: fundamental methodology for simple reaction networks. *Chem. Eng. Technol.* **40**, 2096–2103 (2017).
59. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. in *Proc. 7th Python in Science Conference (SciPy 2008)* (eds Varoquaux, G., Vaught, T., & Millman, J.) 11–15 (2008).

Acknowledgements

This work was supported by funding from the Simons Collaboration on the Origins of Life (SCOL; award 477123, W.T.S.H., W.E.R.) and the Dutch Ministry of Education, Culture and Science (Functional Molecular Systems Gravity programme 024.001.035, W.T.S.H., W.E.R., E.D., P.D.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank P van Galen, R de Graaf and J del Pozo Mellado for support with GC–MS analysis.

Author contributions

W.E.R. and W.T.S.H. conceived the research and acquired funding. W.E.R. and E.D. developed the experimental methodology. W.T.S.H. administered the project and resources. All authors contributed to the design of experiments. W.E.R., E.D., P.v.D. and T.d.J. performed the experiments. W.E.R. curated and analysed the data. W.E.R. wrote the data analysis software and developed data visualizations. W.T.S.H. supervised the investigation. W.E.R. and W.T.S.H. wrote the original draft and all authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

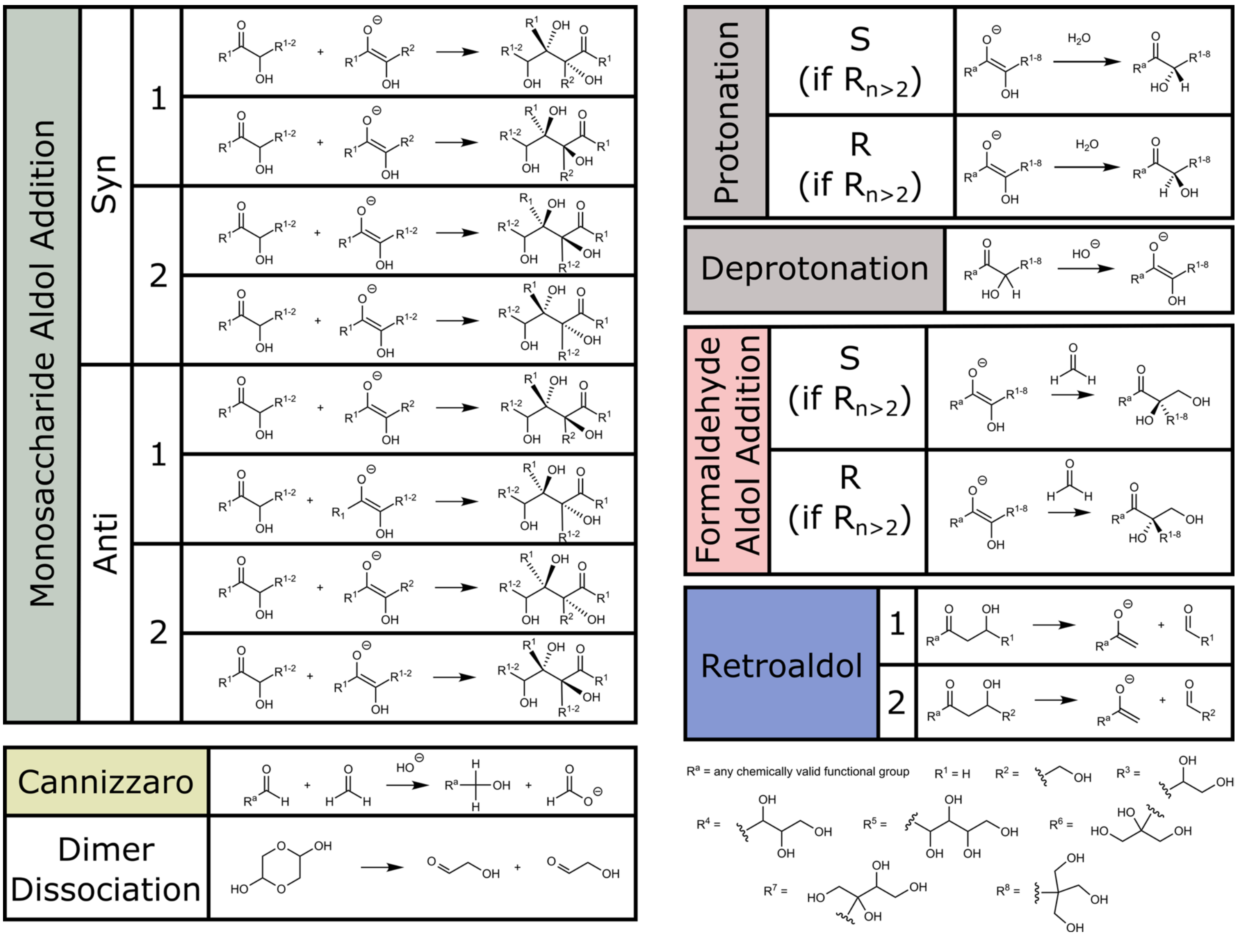
Extended data is available for this paper at https://doi.org/10.1038/s41557-022-00956-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41557-022-00956-7.

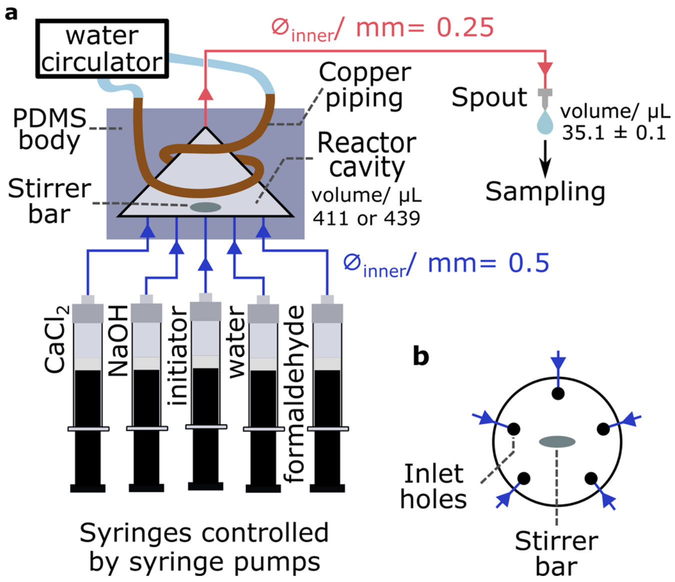
Correspondence and requests for materials should be addressed to Wilhelm T. S. Huck.

Peer review information *Nature Chemistry* thanks Subhabrata Maiti and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

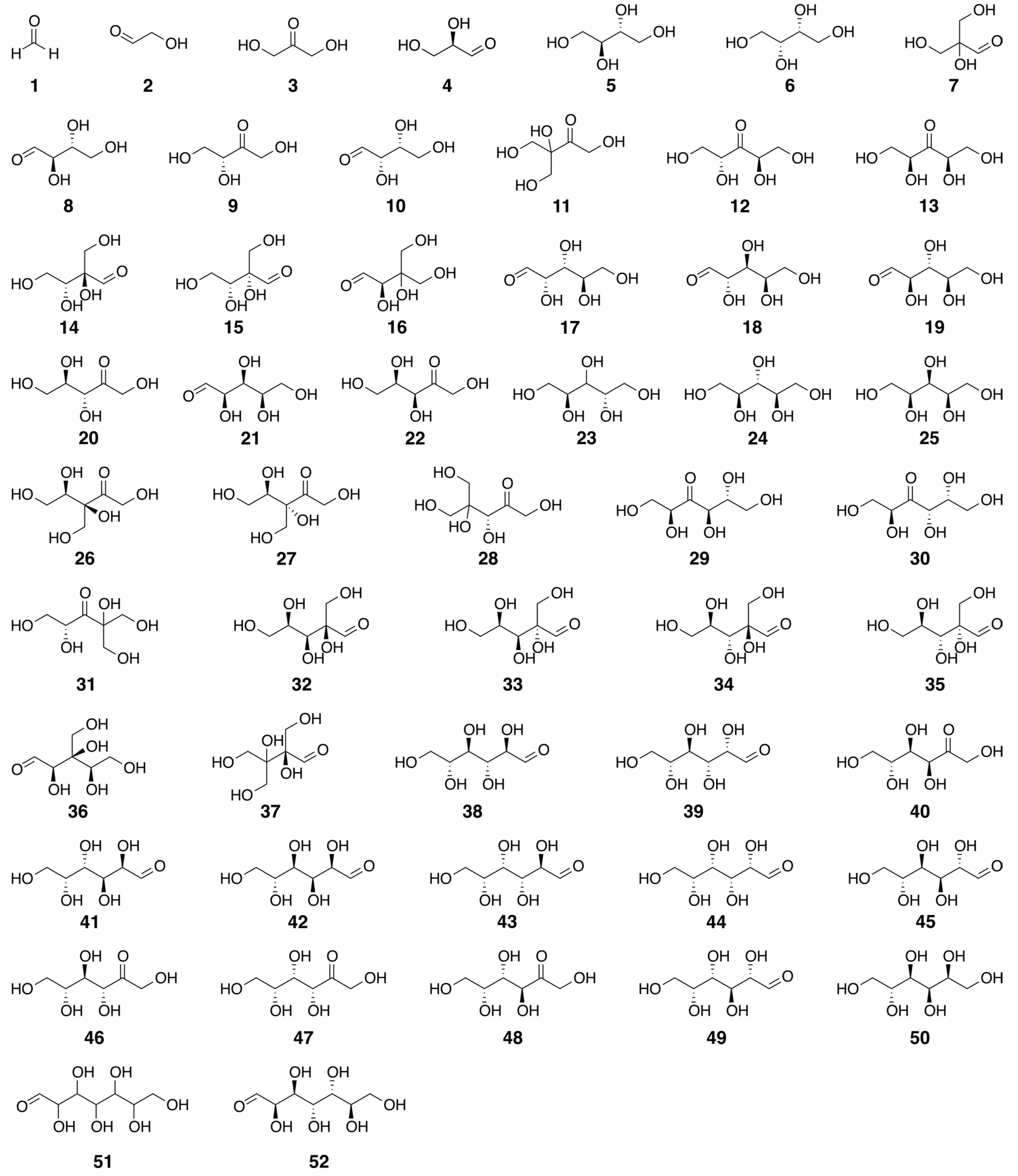
Reprints and permissions information is available at www.nature.com/reprints.



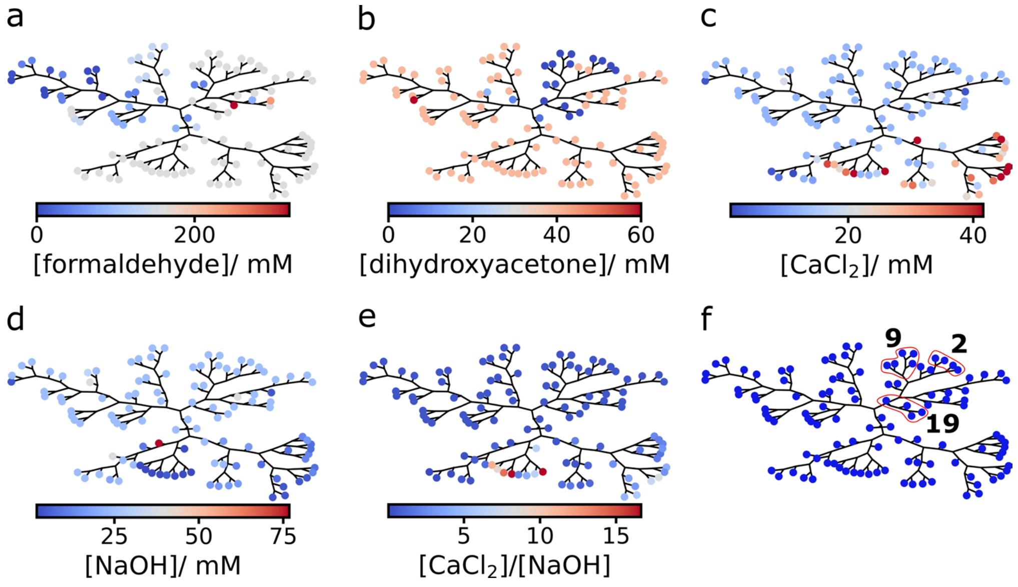
Extended Data Fig. 1 | Reaction types of the formose reaction. Detailed reaction types which describe the transformations shown in Fig. 1a (main text).



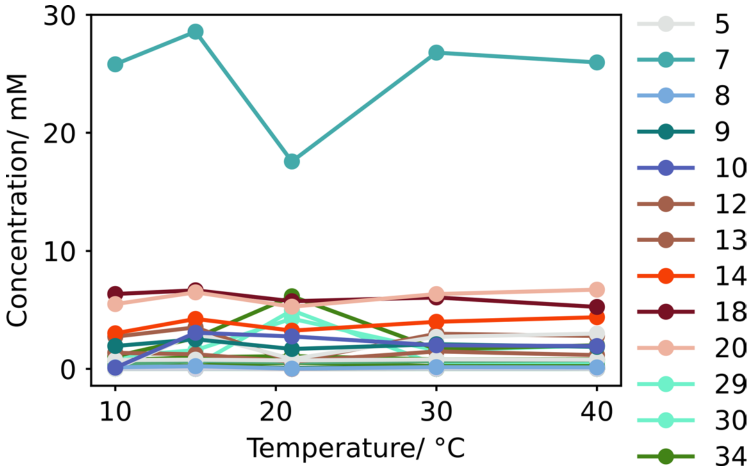
Extended Data Fig. 2 | A detailed schematic of the continuous stirred-tank reactor used in this work. a Side schematic of the reactor depicting how inputs and the outlet were connected to the reactor and how the temperature was controlled. **b** Bottom view of the reactor showing the geometry of the inlet holes into the reactor.



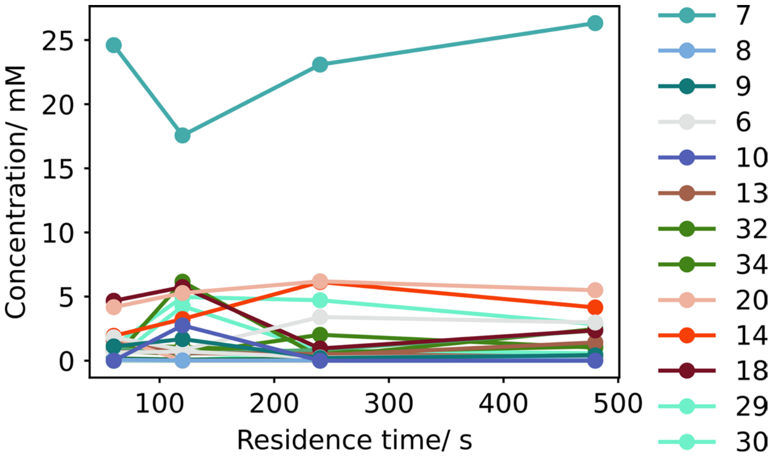
Extended Data Fig. 3 | The structures of compounds assigned in this work and their corresponding numbering scheme.



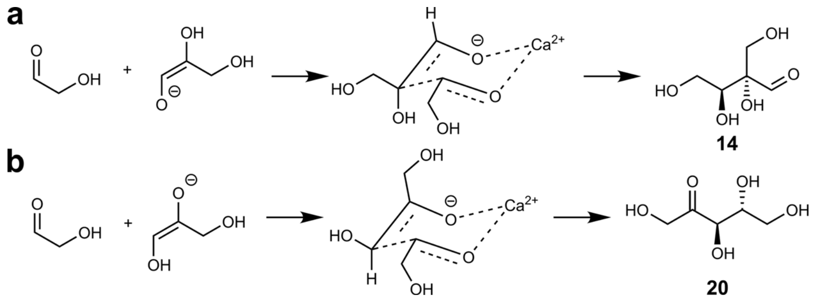
Extended Data Fig. 4 | Mappings of key conditionals variations across data sets to the leaves of the dendrogram. **a** Formaldehyde, **b** dihydroxyacetone, **c** $CaCl_2$, **d** NaOH, **e** the ratio of $CaCl_2$:NaOH, **f** The location of glycolaldehyde (**2**), erythrulose (**9**) and ribose (**19**) initiated reactions. The colour bars below each dendrogram indicate mapping of the colour to the value of each condition.



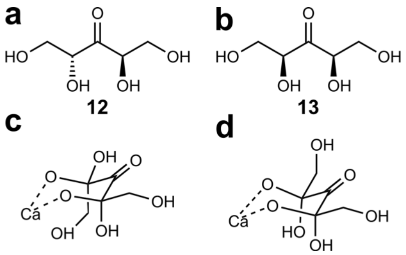
Extended Data Fig. 5 | Variation of the composition of the formose reaction with temperature. Conditions: formaldehyde (200 mM), dihydroxyacetone (25 mM amplitude, 50 mM offset, period 6 min.), $CaCl_2$ (15 mM), NaOH (30 mM).



Extended Data Fig. 6 | Variation of the formose reaction's composition with residence time. The data were determined from flow reactions at 21 °C, with inputs of dihydroxyacetone (25 mM amplitude, 50 mM offset, period three times the residence time), formaldehyde (200 mM), $CaCl_2$ (15 mM), and NaOH (30 mM). Input concentrations are quoted as the initial concentration of compounds upon entering the continuous stirred-tank reactor.



Extended Data Fig. 7 | A proposed mechanism for the selectivity between 14 and 20. a The C_2-C_3 reaction to create **14** via a six-membered ring transition state in which α-hydroxymethyl groups adopt lower energy equatorial positions. **b** A similar reaction and transition state as show in panel **a** from which compound **20** is formed. Dashed bonds indicate those formed and broken over the course of the reaction.



Extended Data Fig. 8 | A proposed mechanism for the selectivity between 12 and 13. a The open-chain structure of **12**. **b** The open-chain structure of **13**. The likely conformation of the six-membered ring formed via coordination of Ca^{2+} to **12** (**c**) and **13** (**d**). Charges (Ca^{2+}, O^{-}) are omitted for clarity.