THE FORMOSE REACTION

PREBIOTIC SOUP OR RECIPE FOR LIFE?

Peer van Duppen

The Formose Reaction

Prebiotic Soup or Recipe for Life?

Peer van Duppen

The work described in this thesis was supported by Functional Molecular Systems, Gravitation Grant (024.001.035) of the Dutch Ministry of Education, Culture and Science.

ISBN: 978-94-6506-591-5

Design: Peer van Duppen

Printed by: Ridderprint | www.ridderprint.nl

The Formose Reaction

Prebiotic Soup or Recipe for Life?

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit
op gezag van de rector magnificus prof. dr. J. M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op
dinsdag 17 december 2024
om 14:30 uur precies

door
Peer van Duppen
geboren op 30 november 1990
te Helmond

Promotor:
prof. dr. W. T. S. Huck

Manuscriptcommissie:
Prof. dr. J. Roithová
Prof. dr. M. Powner (University College London, Verenigd Koninkrijk)
Dr. ing. S. Y. Wong (Universiteit Twente)

Paranimfen:
Dr. D. P. A. Versteegden
N. M. Ivanov

The Formose Reaction

Prebiotic Soup or Recipe for Life?

Doctoral Thesis

to obtain the degree of doctor
from Radboud Universiteit
on the authority of the Rector Magnificus prof. dr. J. M. Sanders,
according to the decision of the Council of Deans
to be defended in public on
Tuesday 17th of December 2024
at 14:30 hours

by
Peer van Duppen
born on November 30,1990
in Helmond, the Netherlands

Supervisor:
Prof. dr. W. T. S. Huck

Doctoral Thesis Committee:
Prof. dr. J. Roithová
Prof. dr. M. Powner (University College London, United Kingdom)
Dr. ing. S. Y. Wong (University of Twente)

Paranymphs:
Dr. D. P. A. Versteegden
N. M. Ivanov

Table of Content

Chapter 5

Fluctuations in the environment direct the organization of prebiotic reaction networks

Chapter 6

Perspectives – Evolution of prebiotic reaction networks towards the origin of life

Addendum

Chapter 1

From synthetic organic chemistry to out-of-equilibrium reaction networks in a prebiotic context - A literature review

The world we live in hosts plants, animals, fungi, bacteria and all kinds of other life. Although the theory of biological evolution provides a good explanation for how modern-day life has developed, it is not yet understood how the first living entity came into being. We cannot comprehend how simple prebiotic feedstock molecules did react and evolve into ever more complex systems.

In this chapter, I will discuss chemistry studied in a prebiotic context. From prebiotic routes of synthetic organic chemistry, to more complex reaction networks. In the absence of genetic and enzymatic machinery, it is only the inherent chemical reactivity that interacts with the environment and shapes the overall outcome of reaction mixtures. In this thesis, I will use the formose reaction as a prebiotic model system. The combinatorial complexity of this reaction enabled me to study how the environment drives self-organization of the reaction network.

1.1 Introduction – The chemical origin of life

Modern-day life, as we know it, is the product of Darwinian evolution. All traits of living beings are the product of the environmental history their ancestors were subjected to.[1] Life that is fit enough to survive and reproduce in the world it lives in, carries on features to the next generation. Together with a factor of random variation, beneficial traits are optimized for and the less advantageous ones disappear.

We have a good explanation for how life produced a plethora of species, see figure 1.1.[2] This process we understand to the point where the different species on the tree of life converge: our Last Universal Common Ancestor (LUCA).[3] Still, we do not know what it took for this LUCA to come into being. Similar to modern-day life it was the product of a process, probably understood by physical and chemical principles, shaped by its environment.

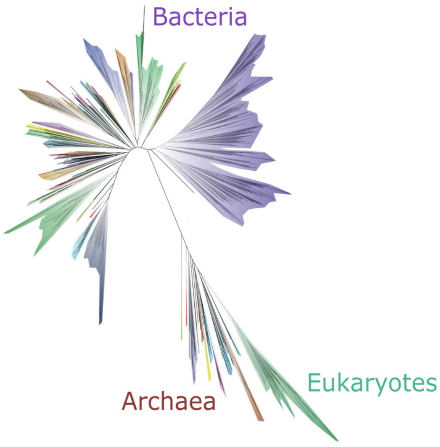


Figure 1.1: Phylogenetic classification of the tree of life, accounting for the total genetic diversity of sequenced genomes. Figure by Hug *et al.* 2016.[2]

This thesis is about prebiotic chemistry, the study of simple aqueous chemistry relevant for the origin of life. In recent years, the main focus of this field of research has been the design of prebiotically plausible synthetic organic chemistry routes to form the building blocks of life.[4,5] Large reaction networks have been drawn up by stitching individually studied, typically high yielding, reactions together.[6] However, incompatibilities in reaction conditions prohibit these routes to operate in 'one pot', as it requires multiple drastic changes in the reaction conditions.

In this literature overview, I will discuss the main strategies for constructing reaction routes or networks to produce biomolecules under prebiotic conditions.

At the advent of prebiotic chemistry, experiments were set up to screen for the production of biomolecules *via* an untargeted approach, where the chemistry was allowed to happen in complex mixtures.[7] This approach is typically low yielding and produces an incomplete variety of essential biomolecules for the origin of life. This shortcoming is circumvented by a more modern approach of aqueous synthetic organic chemistry. Reaction routes of organic chemistry were constructed to increase the yield and to produce more complex biomolecules.[5]

In this chapter, I will firstly discuss the general constrains for experimental chemistry to be prebiotically relevant. Subsequently, I will elaborate on the different approaches in the field of prebiotic chemistry. In the final section, I will give a historical perspective on the formose reaction, both as a prebiotic source of sugar metabolites and its potential role as protometabolic reaction network.

1.2 Setting the stage – Prebiotic chemistry in the lab

It is important to note that there is no consensus on what reaction conditions provide a plausible model for a prebiotic earth. For example, there is even an ongoing debate as to whether prebiotic chemistry should be aqueous, or whether (some) chemistry could have occurred in other solvent systems.[8] I will consider the following chemistry and reaction conditions as relevant to model a prebiotic environment:

1. Solvent system - The chemistry is performed in aqueous systems, since water was abundantly available on prebiotic earth, both in liquid and solid phase.[4]
2. Feedstock molecules - Small organic molecules are used as starting compound, or feedstock molecules, such as HCN, H_2CO, H_2S, CO_2, C_2H_2.[4,9–11] These compounds were likely to be present in the atmosphere of a primordial earth.[12]
3. Catalysts – Different mineral and metal catalysts, such as Fe^{2+}, can be used in prebiotic chemistry, however only if these were available in aqueous environments.[13–15]
4. Energy sources - To drive endergonic reactions under prebiotic conditions light[16] and heat[17] can be used as energy sources.
5. Experimental conditions - The experiment can be designed to simulate specific scenarios for the chemical environment, such as: eutectic phase[18], paste formation[19,20], hydrothermal vents[21] and flow systems (e.g. fluvial)[22,23].
6. Experimental operations - Ideally human intervention is minimized, not to modify or direct the experimental outcome.[24] Experiments of prebiotic chemistry aim to recreate conditions and a scenario relevant

for the prebiotic earth. However, common practices in chemical laboratories, such as a chemical work-up, or multistep synthesis are labor intense processes.

The field of prebiotic chemistry is an experimental discipline and different studies have to be feasible in a laboratory setting. Therefore, I consider these constraints for prebiotic chemistry as an urgent guideline only.

1.2.1 The advent of prebiotic chemistry

Early discoveries in organic chemistry made clear it was not the molecules that make life special *per se*.[25] During the 19th and 20th century, complex biomolecules were synthetically reproduced[26] and even found on meteorites[27]. The molecular building blocks of life might be ubiquitous, yet abiotic production of biomolecules cannot be assumed. Traditional prebiotic chemistry mostly focused on the synthesis of essential biomolecules, which were envisioned to feed into a system capable of metabolic activity or Darwinian evolution.[28]

By mimicking the Hadean atmosphere in their laboratory, Miller and Urey managed to produce a mixture of amino acids.[29,30] In their experimental setup, a mixture of reduced gasses (N_2, H_2 and CH_4) and water was exposed to heat and electrical spark discharges.[29,30] Cyanide, ammonia and carbonyl compounds, such as formaldehyde, were generated *in situ*. Several amino acids were formed *via* a Strecker type reaction pathway, see figure 1.2.[31] Although biomolecules had been produced in an abiotic setting before, this was the first experiment which aimed to recreate the setting of an early earth in the lab. The formation of essential biomolecules was deemed essential for the origin of life *via* a process of abiogenesis.[32–34] The experiment by Miller and Urey produced a plethora of organic compounds, but only formed 8 of the 21 proteinogenic amino acids, some of which in trace amounts.[35]

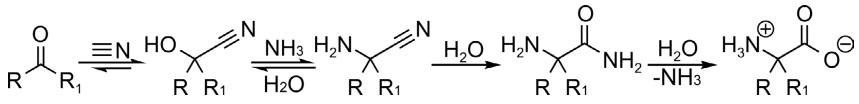


Figure 1.2: Reaction pathway to amino acids in the spark discharge experiment of Urey and Miller.[5,31]

The first experimental approaches to produce amino acids[29,30], sugars[36,37] and nucleobases[38,39] led to the formation of intractable mixtures. Hence, the field of prebiotic chemistry developed a systematic synthetic organic chemistry approach to produce complex biomolecules in high yield.[6,40,41] Large interconnected synthetic routes were created, with chemistry relevant to a prebiotic earth.[6,40,41]

1.2.2 Synthetic prebiotic chemistry

With a small set of reaction types, starting from hydrogen cyanide (**1**) as carbon source, Sutherland and colleagues managed to produce a set of key biomolecular building blocks and metabolites, see figure 1.3.[6] A Kiliani-Fisher type homologation of **1** was used for the synthesis of formaldehyde (**2**) and small sugars glycolaldehyde (**3**), glyceraldehyde (**4**) and also acetaldehyde (**5**), by Cu^{+}-Cu^{2+} photoredox chemistry (fig. 1.3, bold pink arrows).[42] Also, prebiotic synthesis of building blocks cyanamide[43,44] (**6**) and cyanoacetylene[6] (**7**) from **1** have been reported.

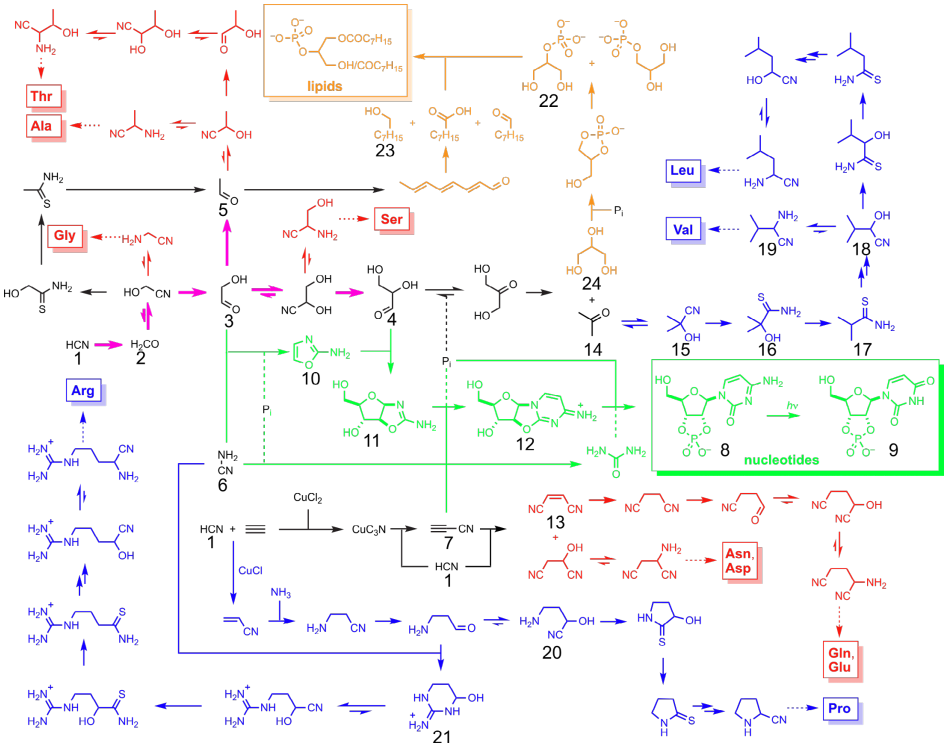


Figure 1.3: **Reaction pathways of cyanosulfidic chemistry**. These pathways start from **1** and produce proteinogenic amino acids, RNA and lipid precursors as essential biomolecules for abiogenesis. Scheme adjusted from Wu and Sutherland, 2019.[10]

With these small building blocks at hand, pyrimidine ribonucleotides (**8**,**9**) were synthesized as RNA building blocks.[40,41,45] Instead of a direct reaction between the nucleobase and ribose, these building blocks were bypassed in a sequential series of addition reactions (fig. 1.3, reactions in green). First, **3** and **6** were reacted to give 2-aminooxazole (**10**).[40,45] The key intermediate arabino

aminooxazoline (**11**) was formed after subsequent addition with **4**.[45–48] The following additions with **7** to form anhydroarabinonucleoside (**12**) and a dry state reaction with a mixture of pyrophosphate and urea gave a cytidine nucleotide derivative (**8**).[40] The uridine nucleotide (**9**) derivative was obtained after UV irradiation of **8**.[40]

With a similar strategy, a prebiotic reaction route from ribose aminooxazoline to **8** and **9** was also found.[41] Moreover, it was shown that the anhydroribonucleoside intermediates in the ribose aminooxazoline pathway provide a pathway to purine deoxynuleosides for DNA synthesis.[49] The group of Sutherland has suggested that both building blocks for RNA and DNA could have coexisted on a prebiotic earth.[49]

A versatile prebiotic mechanism to grow alkyl chains is of crucial importance for the synthesis of proteinogenic amino acids and alkyl alcohols. To this end, the Kiliani-Fisher products **2**, **3** and **5** were not only used as building blocks for ribonucleoside synthesis, but also functioned as amino acid precursors. Photoreduction of **2**, **3** and **5** with hydrogen sulfide and subsequent reaction with ammonia did yield Strecker precursors for glycine, serine, alanine and threonine (fig. 1.3, reactions in red).[50] Maleonitrile (**13**), the product of addition between **1** and **7,** was converted to Strecker precursors of glutamine and glutamic acid *via* a Kiliani-Fisher reaction pathway (fig. 1.3, reactions in red).[6] From **13**, the Strecker precursors of asparagine and aspartic acid were synthesized *via* hydration and a subsequent reaction with urea.[6]

Unfortunately, effective Kiliani-Fisher type homologation was not achieved for cyanohydrins of aldehydic compounds **15**, **18**, **20** and **21**.[6] The equilibrium was not favored for the formation of the respective cyanohydrins. Both **1** and the aldehydic starting compounds underwent a deleterious photoreduction. Effective conversion of starting compounds proceeded after introduction of hydrogen sulfide *via* the Le Chatelier principle. For example, from **14** the respective cyanohydrin **15** was converted to an α-hydroxythioamide **16** (fig. 1.3, with blue arrows).[6] Subsequent irradiation with UV deoxygenated **16** to produce the respective thioamide **17**.[6] Further photoreduction in the presence of firstly hydrogen sulfide and secondly **1** produced cyanohydrin **18**.[6] For amino acid synthesis, the Strecker precursor **19** for valine was produced after reaction with ammonia.[6] From **18**, **20** and **21**, the Strecker precursors for respectively leucine, proline and arginine were synthesized in similar fashion (fig. 1.3, with blue arrows).[6]

For the production of phospholipids, glycerol-2-phosphate (**22**) and alkyl alcohols **23** were coupled together by combining two separate reaction routes (fig. 1.3, reactions in orange).[6,51] The phosphorylated **22** was produced in a formamide solution, where glycerol (**24**) was dissolved with urea and phosphate.[6] For **23**, first **5** was produced by photoreduction of **3** with hydrogen sulfide.[50] Unsaturated C_2, C_4, C_6 and C_8 alcohols were produced *via* a homoenolization reaction of **5** (at elevated pH, where $[Na_2CO_3]$ = 5 mmol) and subsequent reduction with $Ni/H_2PO_2^-$.[51]

1.2.3 Shifting paradigm in prebiotic chemistry – adding reactions to the pot

The cyanosulfidic chemistry proposed by Sutherland and colleagues was coined a protometabolic reaction network.[6,10] Although most chemistry was performed under mild aqueous conditions – at neutral pH and room temperature - the proposed chemistry with high concentration of **1** as feedstock, however versatile in nature, seem an unlikely candidate to produce a protometabolism. Cyanide would interact strongly with all electrophiles in the system, thus preventing the formation of such elaborate reaction network. Most of these reaction pathways were constructed from reactions performed and characterized in isolation. The cyanosulfidic network exemplifies a paradigm of reaction routes *via* multi-step organic synthesis towards a desired biomolecule. Reaction conditions were not compatible to a 'one pot' scenario. A sophisticated flow setup was built to simulate a prebiotic scenario without human intervention.[23] Sequential feedstock recombination did allow for spatiotemporal separation of deleterious reactions to support a prebiotic scenario for cyanosulfidic chemistry.[23]

Life, on the other hand, manages to operate a myriad of chemical reactions simultaneously, where the product of one reaction feeds directly into the next reaction.[3,52,53] Towards the advent of life, the individual chemical reactions were forged together to function as one system. It is questionable how likely the proposed cyanosulfidic networks of chemical reactions could have formed on a prebiotic earth, devoid of human intervention and careful planning. This poses the question: what happens when multiple reactions are run in a 'one pot' environment?

1.3 From soup to structure

Controlled enzyme expression allows living cells to catalyze a multitude of different chemical reactions simultaneously.[3,52,53] In a prebiotic setting, reaction pathways emerge and are selected for, in the absence of enzymatic control. It has been hypothesized that central metabolic reaction pathways were templated by

prebiotic reaction pathways. The reverse TriCarboxylic Acid cycle (rTCA) has been speculated,[33] and disputed,[54] to have originated from a template prebiotic reaction cycle. The rTCA cycle forms a network-autocatalytic loop and hence doubles its input each run through the cycle.[3,52,53] This self-enhancing property is a unique feature of the rTCA and potentially allows the reaction cycle to become self-sustainable in a prebiotic context. The autocatalytic property was the key incentive to scrutinize rTCA chemistry for its potential prebiotic origin.

The rTCA fixates carbon *via* metabolites **24** - **34** and is the starting point towards key biomolecular building blocks: eleven different amino acids, nucleotides and tetrapyrroles, see figure 1.4a. Under prebiotic conditions, a combination of inherent chemical reactivity of the molecules and the environment specifies what chemistry can occur, such as *via* non-specific metal-ion catalysis.[10] This allows for a set of reaction classes to occur (e.g. keto-enol tautomerization, carbonyl reduction, hydrolysis, etc.). Starting from acetate (**35**) as a derivative of **24**, only seven reaction classes are required to recreate the rTCA. However, a multitude of reaction cycles can be constructed from **35** by combining these reaction classes, creating a divergent reaction network of 175 compounds connected by 444 reactions (fig. 1.4b).[55] Does this necessarily mean the chemistry under prebiotic conditions spreads in all directions across the network?[54,55]

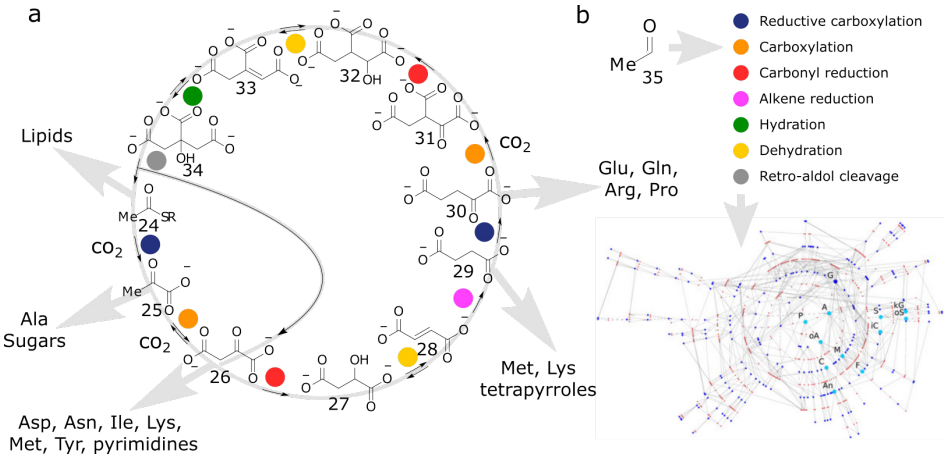


Figure 1.4: The rTCA cycle and the seven reaction classes which occur in the cycle. a) The rTCA cycle with metabolic side branches indicated. b) From acetate (**35**) a reaction network of 175 compounds (blue nodes) can be created by applying the seven reaction classes recursively (red nodes). Scheme adjusted from Muchowska *et al*, 2017 and Zubarev *et al*, 2015.[55,56]

1.3.1 rTCA cycle templates in prebiotic environments

Extant life circumvents deleterious metabolic side reactions *via* enzyme catalysis and allows the rTCA cycle to form a network-autocatalytic loop. Running the cycle in the reverse direction provides an important mechanism to fix carbon atoms and increase the mass of the system. Central metabolic pathways, such as the rTCA cycle, are controlled by enzymes with both high substrate specificity and high catalytic activity.[53,57] Chemically similar intermediates in the rTCA, such as the alkenes fumarate (**28**) and aconitate (**33**) or α-ketoacids pyruvate (**25**), oxaloacetate (**26**), α-ketoglutarate (**30**) and oxalosuccinate (**31**), undergo different types of reactions (fig. 1.4a).[54,58] Simplifying each of the reaction steps in the rTCA to only seven reaction classes, which lack substrate specificity, could lead to deleterious off-cycle reactions (fig. 1.4b). To run the cycle effectively in a prebiotic context and retrieve at least one input molecule per cycle, the different reaction-steps require an average efficiency of at least 90 %.[54] If the rTCA operated in prebiotic waters, off-cycle reactions were likely prevented kinetically, where fast downstream on-cycle reactions allowed a forward reaction flux.[54,59]

To explore the rTCA chemistry under prebiotic conditions, the rTCA metabolites were exposed to different prebiotic conditions. It was shown that ZnS colloidal nanoparticles were able to photocatalyze the reduction of **26** to malate (**27)** and from fumarate (**28**) to succinate (**29**) (fig. 1.4a).[59,60] Also, carboxylation reaction of **29** to **30** was observed in a 2.5 % yield under these conditions.[57] At low temperature (e.g. T = 7 °C) the estimated carboxylation rate for **25** to **26** was estimated to proceed efficiently in conjuncture with the observed photoreduction of **26** to **27**.[58] Similar to the rTCA cycle in living cells, the group of Moran tried to operate the rTCA in a 'one pot' environment. In their initial attempt, two multi-reaction sequences from **26** to **29** and from **31** to citrate (**34**), including reduction and (de)hydration were recreated at 140 °C in a mixture of Cr^{3+}, Zn^{2+} and Fe^{0} (fig. 1.4a).[56]

No effective carboxylation reactions to incorporate CO_2 in the rTCA were discovered under the explored prebiotic conditions. Therefore, an alternative strategy was applied for the growth of a carbon backbone. A reaction mixture of small metabolites was allowed to react with glyoxylate (**36**) and pyruvate (**25**) as feedstock molecules, see figure 1.5.[15] These have been shown to undergo an aldol addition reaction to form α-hydroxy ketoglutarate (**37**).[61] The experimental approach was inspired by studies of the prebiotic origin of the catabolic glycolysis pathway, the pentose phosphate pathway and the forward TCA cycle, described by the Ralser group.[13,14,62] Metabolites were exposed to an anaerobic

Fe^{2+} rich environment, mimicking conditions in primordial oceans.[63] After the reaction was allowed to progress for 48 hours at 70 °C, nine compounds from the rTCA cycle were produced. The proposed reaction scheme was underpinned by aldol addition, oxidative decarboxylation, alkene reduction, carbonyl reduction and dehydration. After formation of **37** and its subsequent dehydration to **38**, the alkene reduction to form **30** is crucial for the production of Fe^{3+}, which in turn drives the oxidative decarboxylation reactions towards **27**, **28** and **29**.[61] In this experimental approach, only rTCA metabolites **26**, **31** and **34** were not observed in the reaction mixture.[61] Formation of **31**, to connect **30** to isocitrate (**32**) was bypassed by an aldol addition reaction between **30** and **36** and subsequent carbonyl reduction. The presence of **26** as an intermediate, was inferred from the presence of both **27** and malonate (**39**).

The creation of a network with nine of the eleven rTCA intermediates is impressive, but the product yields are low and the conversions in the reaction network do not follow a similar reaction pathway to the rTCA cycle. For example, the major reaction towards **27**, **28** and **29** proceeded *via* an oxidative decarboxylation from their α-ketoacid analog (pathway **37** – **38** – **30** in fig. 1.5). In the rTCA, **27** undergoes a dehydration reaction to form **28**, which subsequently undergoes an alkene reduction to form **29** (fig. 1.4a).

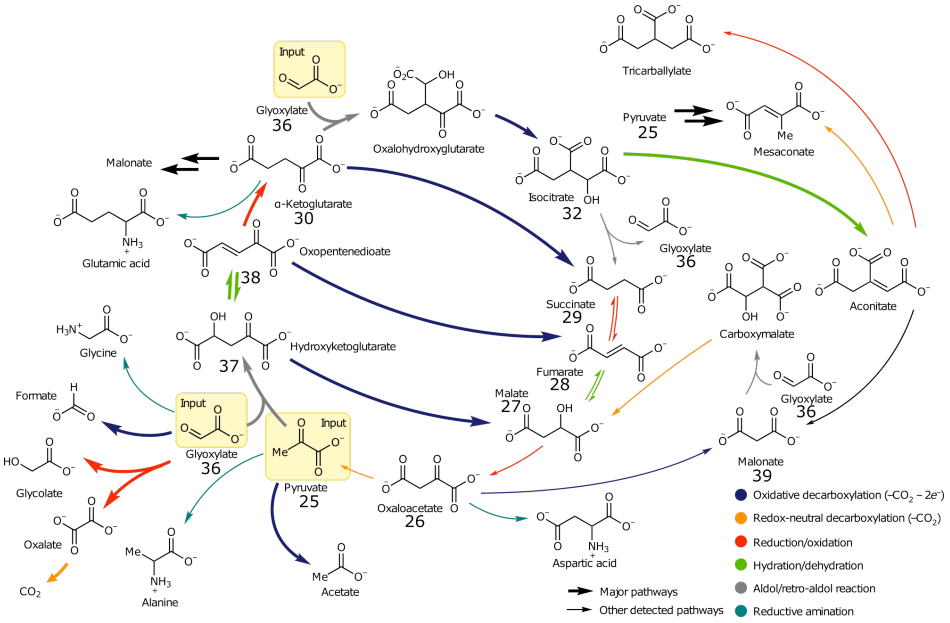


Figure 1.5: **The reaction network for the formation of nine rTCA compounds** in the presence of Fe^{2+}. Scheme adjusted from Muchowska *et al*, 2019.[15]

Inspired by this work, Krishnamurthy and colleagues reconstructed a rTCA cycle analog largely encompassing α-ketoacid metabolite analogs. Starting from **25** and **36**, an analogous pathway with α-ketoacid from **26** to aconitate analog **41** and citrate analog **42** was reproduced, see figure 1.6.[64]

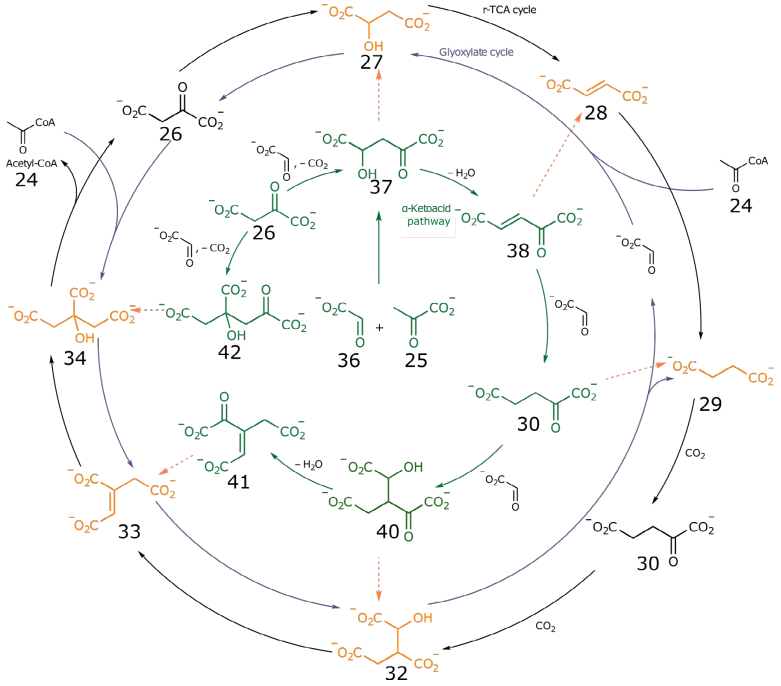


Figure 1.6: The α-ketoacid cycle, encompassing nine rTCA metabolites. Compounds in green were produced in the α-ketoacid pathway, whereas compounds in orange were produced in the oxidative reaction step. Scheme adjusted from Stubbs *et al*, 2020.[64]

The analogous pathway was similar to the Fe^{2+} promoted reaction network, but the reductive and oxidative steps were performed separately.[64] The first part of the reaction pathway proceeded from the α-ketoacid analogs of **37** to **30** (which is also a rTCA intermediate). Remarkably, **30** was produced from alkene reduction of the fumarate α-ketoacid **38** *via* a hydride transfer from the hydrate of **36**.[64] Next, **30** underwent an aldol addition with glyoxylate to form the isocitrate analog **40**. Subsequent dehydration from **40** gave the aconitate analog **41**. Furthermore, citrate α-ketoacid **42** was formed in an aldol addition reaction between **26** and **36**. The rTCA metabolites **27**, **28**, **29**, **32**, **33** and **34** were produced from their α-ketoacid analog in a one-step oxidative decarboxylation with hydrogen peroxide.

1.3.2 Towards protometabolic autocatalytic reaction networks

The reconstructed reaction pathways, similar or analogous to the rTCA cycle, recapitulate some of its reactants. This is an important finding in the field of prebiotic chemistry as a mechanism for carbon fixation, and the intermediates can function as substrates towards key molecular building blocks. For example, it was shown α-ketoacids can react with cyanide for amino acid synthesis.[65] However, it was not for the metabolites *per se* what attracted the prebiotic community in the rTCA cycle. The autocatalytic property can give reaction cycles the ability to become self-sustainable.[3,54] The recapitulated reaction pathways followed different routes, not showing autocatalysis.

Regarding a prebiotic scenario, the metabolites **25** and **36** provided relatively advanced prebiotic feedstock molecules (see 1.2).[66,67] Both molecules were carefully chosen[68] to restrain the chemistry in the emerging reaction pathways.[69] The carboxylate moiety affects the reactivity of the α-carbonyl. The glyoxylate chemistry was not only inspired by the metabolites in the rTCA cycle, but also by the reactivity of formaldehyde.[67] As a carboxylated formaldehyde derivative, glyoxylate limits the expansion of the reaction network. The reactivity of formaldehyde leads to a combinatorial explosion of reaction products in the formose reaction.[70–72] The formaldehyde feedstock molecule allows for the formation of more elaborate reaction pathways and autocatalytic network topologies, hence it has been studied as a protometabolic reaction network.[7,73]

1.4 The formose reaction

The formose reaction, discovered by Butlerow in 1871,[37] is considered as a plausible chemical mechanism for the origin of sugar metabolites on a prebiotic earth.[5] The oligomerization of formaldehyde leads to a complex mixture of both linear and branched sugars, sugar acids and polyol derivatives.[70,74] The reaction took brought interest: from fundamental research into the reaction mechanism, to synthetic approaches to produce sugar metabolites.[72] In the prebiotic community a number of comprehensive review articles which discuss protometabolic autocatalytic reaction pathways[7] and prebiotic ribose synthesis[5] for RNA production have been published. Abiotic processes in the atmosphere, such as the simulated conditions in Miller's electrical discharge experiment or photochemical reduction of CO_2, most likely accounted for formation of formaldehyde on prebiotic earth.[9,29]

In this section, I will elaborate how the formose reaction operates through a recursive set of reaction classes in the presence of different mineral catalysts. I will discuss the role of Ca^{2+} and OH^- and how they are involved in the different reaction classes. Further, I will describe how the different reaction classes

interact to form an expanding reaction network. I will discuss the special interest of the prebiotic community in the formose reaction for its potential provenance of ribose. Finally, I will explain the nature and mechanism of the autocatalytic property of the formose reaction and how it has been regarded as a protometabolic reaction network.

1.4.1 The chemistry of the formose reaction

The formose reaction is driven by oligomerization of formaldehyde under alkaline aqueous conditions in the presence of a carbohydrate initiator, typically glycolaldehyde or dihydroxyacetone.[71,75,76] Different studies suggested the formose reaction could start without initiator sugar.[77,78] However, these results were deemed inaccurate and it was concluded that trace amounts of initiator were likely to be present.[70,79] The product mixture is a result of a set of recursive reaction classes: keto-enol tautomerization, aldol addition, retro aldol cleavage and the Cannizzaro reaction, see figure 1.7. Also, larger linear sugars (> C_3) reversibly form five membered furanose rings and six membered pyranose rings.

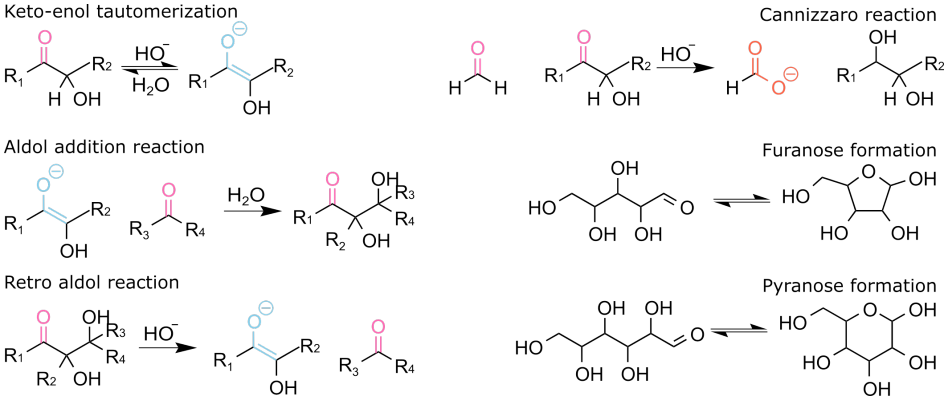


Figure 1.7: Recursive reaction classes governing the formose reaction.

Starting from formaldehyde (**2**) and dihydroxyacetone (**43**) as feedstock molecules, these reaction classes lead to an expanding reaction network, see figure 1.8. In this scheme the stereochemistry is removed and Cannizzaro reactions are omitted for simplicity. Carbohydrates containing an α-proton next to the carbonyl can undergo an enolization reaction. The enol intermediate is the starting point for each addition reaction. Larger carbohydrates can be produced *via* chain elongation with multiple formaldehyde additions, or by direct recombination with a larger carbonyl compound such as glycolaldehyde (**3**), glyceraldehyde (**4**) and **43**. In the reverse aldol reaction, carbohydrates breaks-up to produce two smaller carbohydrates.

Under 1.4.3 I will discuss how autocatalytic reaction pathways can be formed by combining formaldehyde addition chain growth, keto-enol tautomerization and a retro aldol reaction.

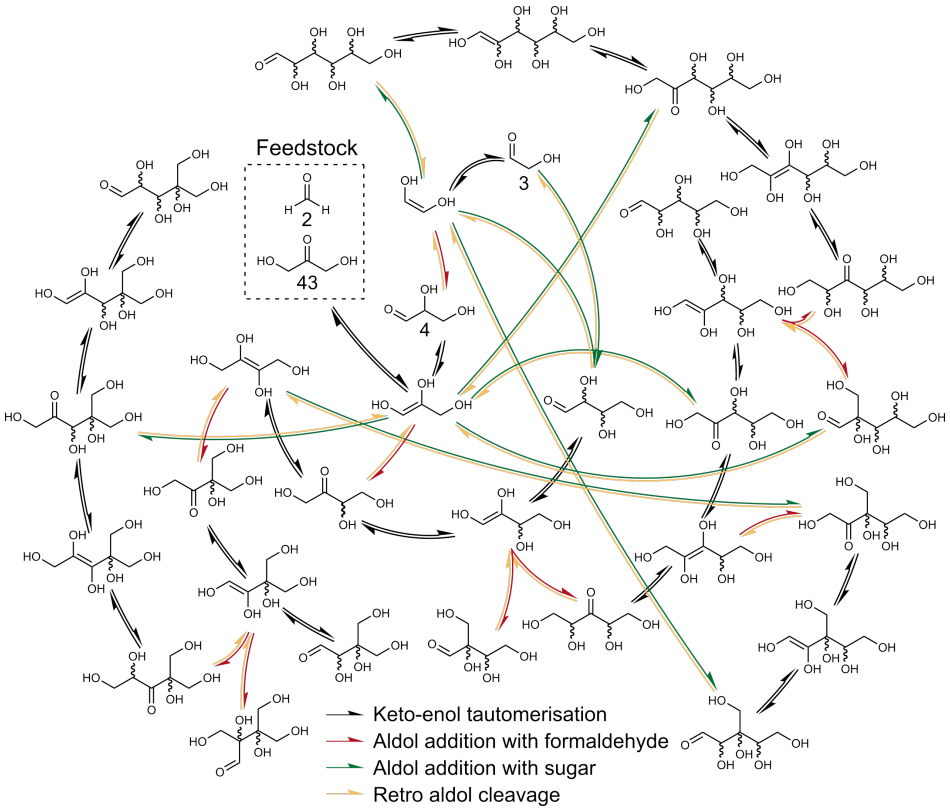


Figure 1.8: A reaction scheme of the formose reaction up to C_6 sugars. The reaction scheme includes common reaction classes: keto-enol tautomerization (black), aldol addition (red/green) and retro aldol cleavage (yellow), Cannizzaro reactions were omitted for simplicity. The reaction scheme separately indicates formaldehyde (red) and sugar (green) additions.

The reaction classes constituting the formose reaction are catalyzed by numerous inorganic catalysts, as summarized by Iqbal and Novarin.[75] Alkaline earth metal hydroxides $Ca(OH)_2$ and $Sr(OH)_2$ were found to be the most active catalysts, together with thallium hydroxide (TlOH) and lead oxide (PbO).[71,80] Most research on the formose reaction was performed with $Ca(OH)_2$ to catalyze the different reaction classes.[70,71,79–86]

The hydroxide ion plays an important catalytic role in the formose reaction.[71,75] For the base catalyzed enolate formation, the hydroxide ion serves to abstract a

proton from the α-carbon, the rate limiting step of enolate formation, see figure 1.9.[87] Formation of the enolate species governs the aldol addition reaction, which proceeds *via* a nucleophilic attack from the enolate on a carbonyl carbon (fig. 1.9b).[88] For the retro aldol cleavage[89], the hydroxide is involved in proton abstraction from the hydroxyl on a β-carbon (fig. 1.9c).[83]

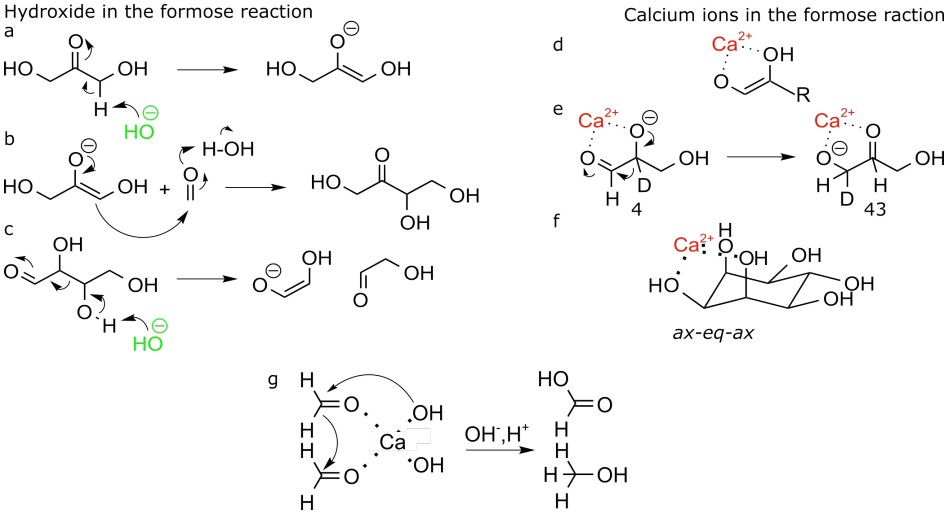


Figure 1.9: The role of OH^- (left) and Ca^{2+} (right) in different reaction classes in the formose reaction. a) Base catalyzed enolate formation. b) Aldol addition between a C_3 enolate and formaldehyde. c) Retro aldol cleavage to produce two smaller sugars. d) Calcium ions stabilized enolates. e) Ca^{2+} catalyzed 1,2-hydride shift. f) Coordination of Ca^{2+} to *ax-eq-ax* hydroxide sequence. g) Cannizarro reaction with $Ca(OH)_2$.

Calcium ions are reported to be the most effective metal ions to catalyze the formose reaction.[72,90,91] The positively charged ion can form complexes with hydroxyl groups, both on the sugar backbone and on the enediolate. The latter complex formation is involved in stabilizing enolate species[85,92] (fig. 1.9d), thus promoting aldol additions and retro aldol cleavage. The C_3 isomerization from **4** to **43** proceeds through a 1,2-hydride shift in the Ca^{2+} complex with the respective enediolate (fig. 1.9e).[93] Interactions of Ca^{2+} with larger linear carbohydrates (> C_4) can potentially stabilize them. Calcium ions can stabilize pyranose and furanose rings, reducing the reactivity of these molecules.[94,95] Especially ring conformations with an axial-equatorial-axial (*ax-eq-ax*) hydroxyl sequence are stabilized by coordinating to Ca^{2+}.[74] Other monosaccharides and polyols do form complexes to some extend with Ca^{2+} as well.[93] The complex formation between calcium ions and larger sugars, such as glucose, increases the solubility limit of $Ca(OH)_2$.[84]

The Cannizzaro reaction is considered a dead end in the formose reaction.[71] In a disproportionation reaction two formaldehyde molecules are converted to formic acid and methanol, removing two reactive carbonyl compounds from the mixture. In the proposed mechanism two carbonyls form a complex with $Ca(OH)_2$.[86] Also, a hydroxide, and thus a catalytic species, is consumed in the reaction. Please note, although the preceding paragraphs describe the key reactions of the formose network, other reactions have been reported, including the oxidation of enediolates by soluble oxygen in the reaction mixture,[92] and Bilik type rearrangement of α-hydroxymethyl aldotetrose to form linear C_5 ribulose or xylulose products.[70]

1.4.2 A prebiotic perspective and the production of D-ribose

Chemists have tried to optimize the formose reaction as a synthesis tool for prebiotic sugar production. In search for a relevant prebiotic model system, the reaction was investigated under milder reaction conditions.[37] The formaldehyde concentration was lowered to the millimolar range and also the pH was lowered to mimic prebiotically plausible environments. With $Ca(OH)_2$ as catalyst, a pH of 11 and higher can facilitate formose chemistry.[70,84]

Attempts to direct the formose outcome requires control over the combinatorial product explosion. Mineral catalysts have been explored to steer the reaction network towards the production of C_5 sugars and ribose.[71,75,80,96,97] It has been shown that this effect can be amplified by cycling a reaction mixture over mineral surfaces.[72] Remarkably, under simulated hydrothermal vent conditions (T = 200 °C, P = 20 bar) mainly carbohydrates smaller than C_6 were produced.[98]

It was hypothesized that a prebiotic formose reaction could function as a source for specific sugar metabolites. For example, by providing a feeding source of glucose into a primordial glycolysis pathway.[99] However, the production of D-ribose, as a building block for RNA nucleotides, has been the main focus.[5,37,100] Formose reaction mixtures definitely contain ribose, but the reaction mechanism lacks an intrinsic selectivity and the product mixture is seemingly intractable.[70,80,101] For initial prebiotic studies the yields of ribose were no higher than 1 %.[102]

Although different studies have tried to optimize the reaction selectivity, increasing the ribose production might not be the main obstacle to overcome. Ribose is not stable under typical formose reaction conditions and degrades readily ($t_{1/2}$ = 5 hours at pH = 12.5).[70,103] Therefore, the conversion of ribose to produce the corresponding nucleotide needs to be immediate or a mechanism is required to quench ribose from the reaction mixture.[5] With a chemical derivatization, different linear C_5 sugars formed a stable reaction product. In a

reaction with cyanamide, a D-ribose aminooxazoline (**44**) was produced as intermediate towards ribonucleotide synthesis, see figure 1.10 (also discussed under 1.2.2).[49,104]

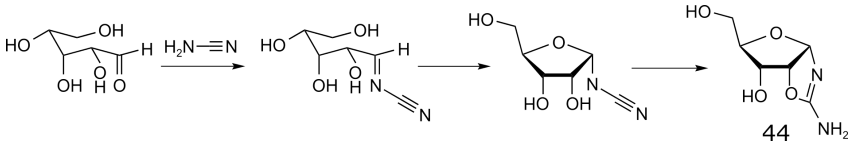


Figure 1.10: Capturing ribose from a reactive mixture as D-ribose aminooxazoline (44) derivative from a reaction between ribose and cyanamide.[104]

Borate complexes were shown to stabilize ribose under formose reaction conditions ($t_{1/2}$ = 45 hours at pH = 12.5), respectively forming 1,2- and 2,3-borate esters, see figure 1.11.[70,105–107] With borate, carbohydrate synthesis with the fomose reaction was geared towards linear and branched C_5 sugars.[70] An initial 1 : 2 ratio of formaldehyde to glycolaldehyde gave high yields of natural pentoses, with ribose, arabinose and xylulose as major reaction products.[70] A similar study was carried out with silicate, where an equimolar mixture of formaldehyde to glycolaldehyde mostly yielded C_5 sugars (30 – 40 %).[97]

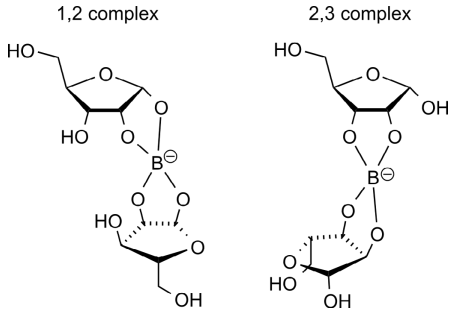


Figure 1.11: Stabilizing ribose *via* the formation of borate complexes, with respectively 1,2- and 2,3-borate esters.[107]

Modifying the initiator sugar is another way to influence formose reactivity.[108,109] The introduction of glycolaldehyde-2-phosphate (**45**) to react with **2** produced a mixture of phosphorylated C_5 and C_6 sugars, see figure 1.12.[108,109] Interestingly, this reaction only produced linear sugars, since the phosphate prohibited formation of keto-sugars.[106]

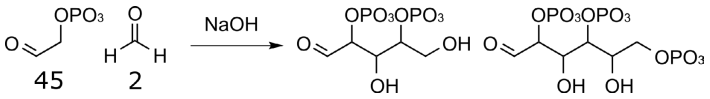


Figure 1.12: The production of linear C_5 and C_6 phosphate sugars, in a formose type of chemistry with reduced reactivity. This behavior was facilitated by **45** as sugar initiator.[108,109]

For nucleotide synthesis, it was shown 5-phosphate ribose (**46**) could be selectively synthesized from a mixture of D-ribose (**47**), phosphate, borate and urea.[110] The urea and borate respectively form protecting groups of the 1- and 2,3-hydroxyl groups on **47**, see figure 1.13. Also, it has been shown that the formation of ribonucleoside-2'-phosphate to ribonucleoside-5'-phosphate was catalyzed in a mixture of borate and urea.[111] Prebiotic condensation reactions for nucleotide synthesis were demonstrated with model wet-dry cycles with both sugars and phosphates[19] and sugars and nucleobases[20].

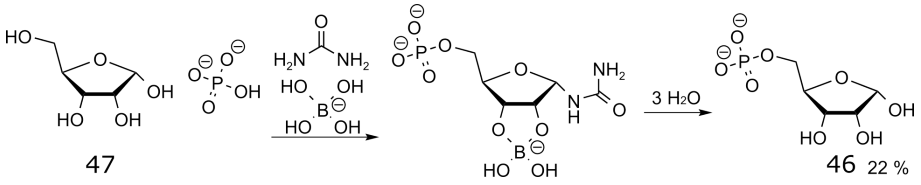


Figure 1.13: Selective synthesis of 5-phospate ribose (46) from D-ribose (47), in the presence of urea and borate.[111]

1.4.3 The formose reaction as a chemical reaction network

The kinetic nature of the formose reaction was suggested to resemble that of metabolic networks.[7,73,112] In typical formose batch reactions the consumption of formaldehyde (**2)** over time exhibits an autocatalytic signature see figure 1.14a.[71,79,81,83] Typically, an induction phase follows after combining the reactants (**2**, a sugar initiator and $Ca(OH)_2$), this is followed by a rapid increase in consumption of **2**.[71,75,79,81,83] Autocatalytic properties of metabolic and genetic networks are of fundamental importance for life to sustain and grow.[52,112,113]

The origin of autocatalytic properties in the formose reaction result from the underlying network topology (fig. 1.14b).[3,113,114] Molecular multiplication occurs *via* homologous chain growth with **2**, followed by a retro aldol reaction. This process creates reaction cycles where a starting molecule is doubled each run through the cycle. The mechanism for autocatalysis in the formose reaction was first proposed by Breslow in 1959 (fig. 1.14c).[115] Breslow's cycle starts with addition of **2** to the C_2-enolate to form glyceraldehyde (**4**). Subsequent isomerization to dihydroxyacetone (**43**) occurs *via* a 1,2-hydride shift (fig. 1.9e).[93]

The C_4, erythrulose (**48**), is produced after formation of the corresponding C_3-enolate and aldol addition with **2**. To complete the cycle, the erythose or threose isomer (**49**) of **48** undergoes a retro aldol cleavage. This results in the duplication of the C_2 starting compound **3** and its enolate.

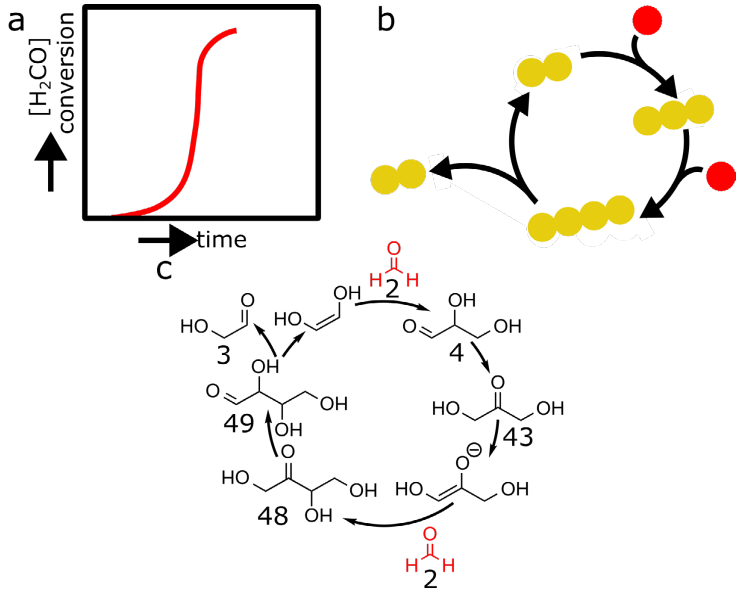


Figure 1.14: Nature of autocatalysis in the formose reaction. a) Representation of typical formaldehyde conversion in a formose batch reaction.[71,75,79,81,83] b) Schematic representation of the autocatalytic network topology in the formose reaction.[3,113,114] c) The Breslow cycle for autocatalysis in the formose reaction.[93,115]

Several reports have suggested the presence of bistability of the formose reaction in a flow reactor.[84,116] Two apparent regimes of both high and low formaldehyde conversion were attained by changing the flow rate in a continuous stirred-tank reactor (CSTR), see figure 1.15.[73] Upon further exploration, however, bistability has not been confirmed. The observed hysteresis loop was speculated to be the result of insufficiently long observation times (30 – 120 minutes to reach steady-state). If the system were to reach its true steady-state, the reaction network would always relax to a regime of high formaldehyde conversion.[73] Though, these results do show a slow relaxation of the reaction network, resulting from slow reactions in the network with respect to the time scale of the experiment.

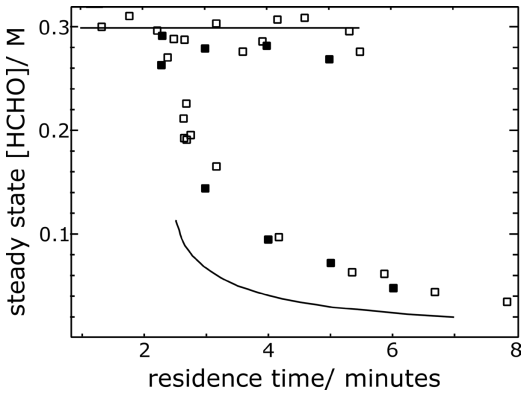


Figure 1.15: Apparent bistability of the formose reaction in a flow reactor. The observed hysteresis loop results from insufficient time to reach steady-state (30 – 120 minutes).[73]

Dynamic behavior, such as bistability, is a trademark of complex networks.[22,117] Cellular response mechanisms are often controlled by bistable dynamics.[118] Together with its autocatalytic property, the apparent bistability in the formose reaction is reminiscent of metabolic network properties.

It has been shown that appropriate balancing of rates in small reaction networks with specific topologies[119] allows for functional behavior, such as bistability and oscillations under out-of-equilibrium conditions.[22,120,121] In principle, it should be possible to determine the rate constants for individual reactions and model the progress of the network.[122] However, accurate modelling with mass-action kinetics becomes impossible for more complicated divergent reaction networks with a high degree of interconnectivity, such as the formose reaction.[73,123] Modern analytical techniques have shown hundreds of different carbohydrate complexes can be present in the formose mixture.[72,124]

The unpredictable behavior of large chemical reaction networks, together with the analytical challenges to characterize complex mixtures, let these systems to be coined as 'soup'. A central challenge I will be addressing in this thesis is the development of methods that lead to a better understanding of these complex reaction network and how they respond to changes in the environment. The preformed studies were unaided and open-ended, contrary to attempts to gear the formose reaction towards the production of D-ribose (see 1.4.2). This approach is required in order to gain insight in how ever more complex reaction pathways can be forged in a prebiotic setting. This very process of forming more complex reaction pathways is crucial for systems that ultimately evolve towards the origin of life.

1.5 Outline of this thesis

The formose reaction provides a system which leads to a divergent reaction network, similar to the large theoretical supernetwork of rTCA chemistry.[55] Defined by a small set of reaction types, it is also capable of producing network-autocatalytic loops with reaction pathways that branch-off in different directions. For a real reaction network, it is not the chemical reactivity alone which selects one pathway over another. Environmental parameters can be tweaked to direct the outcome of the formose reaction. However, there is no fundamental understanding how the interaction between inherent chemical reactivity and the environment translates to self-organization of chemical prebiotic reaction networks.

In this thesis, the formose reaction is employed as a model prebiotic reaction network to study the emergence of reaction pathways. A better understanding of the following mechanisms would provide important insights in the organizational principles of prebiotic reaction networks:

1. Chemical reactivity patterns in a 'one pot' scenario – the formose reaction produces a complex reaction mixture which results from recursively applying a set of reaction classes. Is it possible to control activation and deactivation at the level of different reaction classes?
2. Control over reaction pathway formation – the formose reaction lacks any direction in self-organization from the chemical reactivity patterns alone (fig. 1.8). Is it possible to change and control the formation of reaction pathways?
3. Evolving network states – a prebiotic reaction network requires evolution in network state to allow for the transition from inanimate chemistry towards life. Is it possible to find a mechanism, or extract general principles, for a model prebiotic network to evolve towards new network states?

In **chapter 2**, I will explain the experimental design and how the composition of the formose reaction mixture was analyzed. The formose reaction was carried out in a continuous stirred-tank reactor (CSTR). Input parameters can be varied to screen different types of environments. The compositional outcomes and underlying reaction networks were interpreted from liquid and gas chromatographic traces and mass fragmentation spectra.

Subsequently, in **chapter 3**, different strategies to reconstruct the connectivity of chemical reaction networks will be discussed. The reaction connectivity was deduced by probing the formose reaction with a sinusoidal input modulation. Also, the signature of pseudo-random input concentration of catalyst $Ca(OH)_2$

in different parts of the network were used to reconstruct the underlying reaction network.

With these tools in hand, I will elaborate in **chapter 4** how different environmental parameters affected the composition of the network. The network exhibited unique responses to different input sugars, or changes in formaldehyde or $Ca(OH)_2$ input concentrations. Also, I will discuss how these environments interacted with the chemistry and rewired the underlying reaction structure.

In **chapter 5**, I will discuss how the signal strength and the rate of change of temporal patterns from the environment directed the compositional outcomes in the formose reaction. Further, I will elaborate on how the signal is transferred to different parts of the network.

In the final **chapter 6**, I will give a broader perspective on this work and I discuss the implications for the field of prebiotic chemistry. I will elaborate how new insights from this thesis help to define a scenario for the origin of life.

1.6 References

1. Darwin, C. R. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. (1859).
2. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
3. Braakman, R. & Smith, E. The compositional and evolutionary logic of metabolism. *Phys. Biol.* **10**, 011001 (2012).
4. Cleaves, H. J. Prebiotic Chemistry: What We Know, What We Don't. *Evol. Educ. Outreach* **5**, 342–360 (2012).
5. Islam, S. & Powner, M. W. Prebiotic Systems Chemistry: Complexity Overcoming Clutter. *Chem* **2**, 470–501 (2017).
6. 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).
7. Ruiz-Mirazo, K., Briones, C. & De La Escosura, A. Prebiotic Systems Chemistry: New Perspectives for the Origins of Life. *Chem. Rev.* **114**, 285–366 (2014).
8. Benner, S. A. Prebiotic plausibility and networks of paradox-resolving independent models. *Nat. Commun.* **9**, 5173 (2018).
9. Cleaves II, H. J. The prebiotic geochemistry of formaldehyde. *Precambrian Res.* **164**, 111–118 (2008).
10. Wu, L.-F. & Sutherland, J. D. Provisioning the origin and early evolution of life. *Emerg. Top. Life Sci.* **3**, 459–468 (2019).
11. Wołos, A. *et al.* Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. *Science* **369**, eaaw1955 (2020).
12. Zahnle, K., Schaefer, L. & Fegley, B. Earth's Earliest Atmospheres. *Cold Spring Harb. Perspect. Biol.* **2**, a004895–a004895 (2010).
13. Keller, M. A., Turchyn, A. V. & Ralser, M. Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible A rchean ocean. *Mol. Syst. Biol.* **10**, 725 (2014).
14. Keller, M. A. *et al.* Conditional iron and pH-dependent activity of a non-enzymatic glycolysis and pentose phosphate pathway. *Sci. Adv.* **2**, e1501235 (2016).
15. Muchowska, K. B., Varma, S. J. & Moran, J. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019).
16. Ranjan, S. & Sasselov, D. D. Influence of the UV Environment on the Synthesis of Prebiotic Molecules. *Astrobiology* **16**, 68–88 (2016).
17. Huber, C., Eisenreich, W. & Wächtershäuser, G. Synthesis of α-amino and α-hydroxy acids under volcanic conditions: implications for the origin of life. *Tetrahedron Lett.* **51**, 1069–1071 (2010).
18. Menor-Salván, C. & Marín-Yaseli, M. R. Prebiotic chemistry in eutectic solutions at the water–ice matrix. *Chem. Soc. Rev.* **41**, 5404 (2012).
19. Maguire, O. R., Smokers, I. B. A. & Huck, W. T. S. A physicochemical orthophosphate cycle via a kinetically stable thermodynamically activated intermediate enables mild prebiotic phosphorylations. *Nat. Commun.* **12**, 5517 (2021).
20. Becker, S. *et al.* Wet-dry cycles enable the parallel origin of canonical and non-canonical nucleosides by continuous synthesis. *Nat. Commun.* **9**, 163 (2018).
21. Hudson, R. *et al.* CO_2 reduction driven by a pH gradient. *Proc. Natl. Acad. Sci.* **117**, 22873–22879 (2020).

22. Semenov, S. N. *et al.* Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature* **537**, 656–660 (2016).
23. Ritson, D. J., Battilocchio, C., Ley, S. V. & Sutherland, J. D. Mimicking the surface and prebiotic chemistry of early Earth using flow chemistry. *Nat. Commun.* **9**, 1821 (2018).
24. Richert, C. Prebiotic chemistry and human intervention. *Nat. Commun.* **9**, 5177 (2018).
25. Wöhler, F. Ueber künstliche Bildung des Harnstoffs. *Ann. Phys. Chem.* **88**, 253–256 (1828).
26. Woodward, R. B. The total synthesis of vitamin B12. *Pure Appl. Chem.* **33**, 145–178 (1973).
27. Cronin, J. R. & Pizzarello, S. Amino acids in meteorites. *Adv. Space Res.* **3**, 5–18 (1983).
28. Sasselov, D. D., Grotzinger, J. P. & Sutherland, J. D. The origin of life as a planetary phenomenon. *Sci. Adv.* **6**, eaax3419 (2020).
29. Miller, S. L. A Production of Amino Acids Under Possible Primitive Earth Conditions. *Science* **117**, 528–529 (1953).
30. Miller, S. L. Production of Some Organic Compounds under Possible Primitive Earth Conditions [1]. *J. Am. Chem. Soc.* **77**, 2351–2361 (1955).
31. Miller, S. L. The mechanism of synthesis of amino acids by electric discharges. *Biochim. Biophys. Acta* **23**, 480–489 (1957).
32. Oparin, A. I. Origin of life. *Mosc. Izd Mosk. Rabochii* (1924).
33. Horowitz, N. H. On the Evolution of Biochemical Syntheses. *Proc. Natl. Acad. Sci.* **31**, 153–157 (1945).
34. Miller, S. L. & Urey, H. C. Organic Compound Synthesis on the Primitive Earth: Several questions about the origin of life have been answered, but much remains to be studied. *Science* **130**, 245–251 (1959).
35. Johnson, A. P. *et al.* The Miller Volcanic Spark Discharge Experiment. *Science* **322**, 404–404 (2008).
36. Gabel, N. W. & Ponnamperuma, C. Model for Origin of Monosaccharides. *Nature* **216**, 453–455 (1967).
37. Butlerow, A. Bildung einer zuckerartigen Substanz durch Synthese. *Justus Liebigs Ann. Chem.* **120**, 295–298 (1861).
38. Oró, J. Synthesis of adenine from ammonium cyanide. *Biochem. Biophys. Res. Commun.* **2**, 407–412 (1960).
39. Oró, J. & Kimball, A. P. Synthesis of purines under possible primitive earth conditions. I. Adenine from hydrogen cyanide. *Arch. Biochem. Biophys.* **94**, 217–227 (1961).
40. Powner, M. W., Gerland, B. & Sutherland, J. D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459**, 239–242 (2009).
41. Xu, J. *et al.* A prebiotically plausible synthesis of pyrimidine β-ribonucleosides and their phosphate derivatives involving photoanomerization. *Nat. Chem.* **9**, 303–309 (2017).
42. Ritson, D. & Sutherland, J. D. Prebiotic synthesis of simple sugars by photoredox systems chemistry. *Nat. Chem.* **4**, 895–899 (2012).
43. Schimpl, A., Lemmon, R. M. & Calvin, M. Cyanamide Formation under Primitive Earth Conditions. *Science* **147**, 149–150 (1965).
44. Hulshof, J. & Ponnamperuma, C. Prebiotic condensation reactions in an aqueous medium: A review of condensing agents. *Orig. Life* **7**, 197–224 (1976).

45. Anastasi, C., Crowe, M. A., Powner, M. W. & Sutherland, J. D. Direct Assembly of Nucleoside Precursors from Two- and Three-Carbon Units. *Angew. Chem. Int. Ed.* **45**, 6176–6179 (2006).
46. Sanchez, R. A. & Orgel, L. E. Studies in prebiotic synthesis. *J. Mol. Biol.* **47**, 531–543 (1970).
47. Borsenberger, V. *et al.* Exploratory Studies to Investigate a Linked Prebiotic Origin of RNA and Coded Peptides. *Chem. Biodivers.* **1**, 203–246 (2004).
48. Saewan, N. *et al.* Exploratory Studies to Investigate a Linked Prebiotic Origin of RNA and Coded Peptides. 4th Communication: Further Observations Concerning Pyrimidine Nucleoside Synthesis by Stepwise Nucleobase Assembly. *Chem. Biodivers.* **2**, 66–83 (2005).
49. Xu, J. *et al.* Selective prebiotic formation of RNA pyrimidine and DNA purine nucleosides. *Nature* **582**, 60–66 (2020).
50. Ritson, D. J. & Sutherland, J. D. Synthesis of Aldehydic Ribonucleotide and Amino Acid Precursors by Photoredox Chemistry. *Angew. Chem. Int. Ed.* **52**, 5845–5847 (2013).
51. Bonfio, C. *et al.* Length-Selective Synthesis of Acylglycerol-Phosphates through Energy-Dissipative Cycling. *J. Am. Chem. Soc.* **141**, 3934–3939 (2019).
52. Smith, E. & Morowitz, H. J. Universality in intermediary metabolism. *Proc. Natl. Acad. Sci.* **101**, 13168–13173 (2004).
53. Nam, H. *et al.* Network Context and Selection in the Evolution to Enzyme Specificity. *Science* **337**, 1101–1104 (2012).
54. Orgel, L. E. The Implausibility of Metabolic Cycles on the Prebiotic Earth. *PLoS Biol.* **6**, e18 (2008).
55. Zubarev, D. Y., Rappoport, D. & Aspuru-Guzik, A. Uncertainty of Prebiotic Scenarios: The Case of the Non-Enzymatic Reverse Tricarboxylic Acid Cycle. *Sci. Rep.* **5**, 8009 (2015).
56. Muchowska, K. B. *et al.* Metals promote sequences of the reverse Krebs cycle. *Nat. Ecol. Evol.* **1**, 1716–1721 (2017).
57. Bar-Even, A. *et al.* The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **50**, 4402–4410 (2011).
58. Ross, D. S. The Viability of a Nonenzymatic Reductive Citric Acid Cycle – Kinetics and Thermochemistry. *Orig. Life Evol. Biospheres* **37**, 61–65 (2007).
59. Guzman, M. I. & Martin, S. T. Oxaloacetate-to-malate conversion by mineral photoelectrochemistry: implications for the viability of the reductive tricarboxylic acid cycle in prebiotic chemistry. *Int. J. Astrobiol.* **7**, 271–278 (2008).
60. Zhang, X. V. & Martin, S. T. Driving Parts of Krebs Cycle in Reverse through Mineral Photochemistry. *J. Am. Chem. Soc.* **128**, 16032–16033 (2006).
61. 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).
62. Keller, M. Sulfate radicals enable a non-enzymatic Krebs cycle precursor. Mendeley https://doi.org/10.17632/VGPMNZDZ55.1 (2017).
63. Rouxel, O. J., Bekker, A. & Edwards, K. J. Iron Isotope Constraints on the Archean and Paleoproterozoic Ocean Redox State. *Science* **307**, 1088–1091 (2005).

64. 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).
65. Pulletikurti, S., Yadav, M., Springsteen, G. & Krishnamurthy, R. Prebiotic synthesis of α-amino acids and orotate from α-ketoacids potentiates transition to extant metabolic pathways. *Nat. Chem.* **14**, 1142–1150 (2022).
66. Marshall, S. M., Murray, A. R. G. & Cronin, L. A probabilistic framework for identifying biosignatures using Pathway Complexity. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **375**, 20160342 (2017).
67. Marshall, S. M. *et al.* Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nat. Commun.* **12**, 3033 (2021).
68. Eschenmoser, A. On a Hypothetical Generational Relationship between HCN and Constituents of the Reductive Citric Acid Cycle. *Chem. Biodivers.* **4**, 554–573 (2007).
69. Krishnamurthy, R. & Liotta, C. L. The potential of glyoxylate as a prebiotic source molecule and a reactant in protometabolic pathways—The glyoxylose reaction. *Chem* **9**, 784–797 (2023).
70. Kim, H.-J. *et al.* Synthesis of Carbohydrates in Mineral-Guided Prebiotic Cycles. *J. Am. Chem. Soc.* **133**, 9457–9468 (2011).
71. 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).
72. Colón-Santos, S., Cooper, G. J. T. & Cronin, L. Taming the Combinatorial Explosion of the Formose Reaction via Recursion within Mineral Environments. *ChemSystemsChem* **1**, (2019).
73. Huskey, W. P. & Epstein, I. R. Autocatalysis and apparent bistability in the formose reaction. *J. Am. Chem. Soc.* **111**, 3157–3163 (1989).
74. Weiss, A. H., Socha, R. F., Likholobov, V. A. & Sakharov, M. M. Formose sugars from formaldehyde. *Appl. Catal.* **1**, 237–246 (1981).
75. Zafar Iqbal & Senad Novalin. The Formose Reaction: A Tool to Produce Synthetic Carbohydrates Within a Regenerative Life Support System. *Curr. Org. Chem.* **16**, 769–788 (2012).
76. 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).
77. Shigemasa, Y., Fujitani, T., Sakazawa, C. & Matsuura, T. Formose Reactions. III. Evaluation of Various Factors Affecting the Formose Reaction. *Bull. Chem. Soc. Jpn.* **50**, 1527–1531 (1977).
78. Shigemasa, Y., Nagae, O., Sakazawa, C., Nakashima, R. & Matsuura, T. Formose reactions. 5. A selective formose reaction. *J. Am. Chem. Soc.* **100**, 1309–1310 (1978).
79. Socha, R. Homogeneously catalyzed condensation of formaldehyde to carbohydrates VII. An overall formose reaction model. *J. Catal.* **67**, 207–217 (1981).
80. Khomenko, T. I., Sakharov, M. M. & Golovina, O. A. The Synthesis of Carbohydrates from Formaldehyde. *Russ. Chem. Rev.* **49**, 570–584 (1980).
81. De Bruijn, J. M., Kieboom, A. P. G. & Bekkiun, H. V. Alkaline Degradation of Monosaccharides VI [1]: The Fhucto-Fobmose Reaction of Mixtures of D-Fructose and Formaldehyde. *J. Carbohydr. Chem.* **5**, 561–569 (1986).

82. Mizuno, T. & Weiss, A. H. Synthesis and Utilization of Formose Sugars. in *Advances in Carbohydrate Chemistry and Biochemistry* vol. 29 173–227 (Elsevier, 1974).
83. 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).
84. Weiss, A. Homogeneously catalyzed formaldehyde condensation to carbohydrates. *J. Catal.* **16**, 332–347 (1970).
85. Fujino, K., Kobayashi, J. & Higuchi, I. Complex Formation from Calcium Hydroxide and Carbohydrate in Alkaline Solutions. *NIPPON KAGAKU KAISHI* 2287–2292 (1972).
86. Fujino, K., Kobayashi, J. & Higuchi, I. Homogeneous Reaction of Formose Formation Catalyzed by Complexes of Calcium Hydroxide. *NIPPON KAGAKU KAISHI* 2292–2297 (1972).
87. 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).
88. Gutsche, C. D. *et al.* Base-catalyzed triose condensations. *J. Am. Chem. Soc.* **89**, 1235–1245 (1967).
89. Guthrie, J. P. The Aldol Condensation of Acetaldehyde: the Equilibrium Constant for the Reaction and the Rate Constant for the Hydroxide Catalyzed RetroAldol Reaction. *Can. J. Chem.* **52**, 2037–2040 (1974).
90. Shigemasa, Y., Taji, T., Waki, E. & Nakashima, R. Formose Reactions. XIV. A Selective Formose Reaction in the Presence of a Slight Amount of Calcium Ions. *Bull. Chem. Soc. Jpn.* **54**, 1403–1409 (1981).
91. Shigemasa, Y., Shimao, M., Sakazawa, C. & Matsuura, T. Formose Reactions. IV. The Formose Reaction in Homogeneous Systems and the Catalytic Functions of Calcium Ion Species. *Bull. Chem. Soc. Jpn.* **50**, 2138–2142 (1977).
92. De Wit, G., Kieboom, A. P. G. & van Bekkum, H. Enolisation and isomerisation of monosaccharides in aqueous, alkaline solution. *Carbohydr. Res.* **74**, 157–175 (1979).
93. Appayee, C. & Breslow, R. Deuterium Studies Reveal a New Mechanism for the Formose Reaction Involving Hydride Shifts. *J. Am. Chem. Soc.* **136**, 3720–3723 (2014).
94. Angyal, S. J. Haworth Memorial Lecture. Sugar–cation complexes—structure and applications. *Chem Soc Rev* **9**, 415–428 (1980).
95. Briggs, J., Finch, P., Matulewicz, M. C. & Weigel, H. Complexes of copper(II), calcium, and other metal ions with carbohydrates: Thin-layer ligand-exchange chromatography and determination of relative stabilities of complexes. *Carbohydr. Res.* **97**, 181–188 (1981).
96. Shigemasa, Y. Formose reactions *1IX. Selective formation of branched sugar alcohols in a modified formose reaction and factors affecting the selectivity. *J. Catal.* **62**, 107–116 (1980).
97. 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).
98. Kopetzki, D. & Antonietti, M. Hydrothermal formose reaction. *New J. Chem.* **35**, 1787 (2011).
99. Meléndez-Hevia, E., Montero-Gómez, N. & Montero, F. From prebiotic chemistry to cellular metabolism—The chemical evolution of metabolism before Darwinian natural selection. *J. Theor. Biol.* **252**, 505–519 (2008).

100. Reid, C. & Orgel, L. E. Model for Origin of Monosaccharides: Synthesis of Sugars in Potentially Prebiotic Conditions. *Nature* **216**, 455–455 (1967).
101. 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).
102. Shapiro, R. The prebiotic role of adenine: A critical analysis. *Orig. Life Evol. Biosph.* **25**, 83–98 (1995).
103. Larralde, R., Robertson, M. P. & Miller, S. L. Rates of decomposition of ribose and other sugars: implications for chemical evolution. *Proc. Natl. Acad. Sci.* **92**, 8158–8160 (1995).
104. Springsteen, G. & Joyce, G. F. Selective Derivatization and Sequestration of Ribose from a Prebiotic Mix. *J. Am. Chem. Soc.* **126**, 9578–9583 (2004).
105. Furukawa, Y., Horiuchi, M. & Kakegawa, T. Selective Stabilization of Ribose by Borate. *Orig. Life Evol. Biospheres* **43**, 353–361 (2013).
106. Ricardo, A., Carrigan, M. A., Olcott, A. N. & Benner, S. A. Borate Minerals Stabilize Ribose. *Science* **303**, 196–196 (2004).
107. Šponer, J. E., Sumpter, B. G., Leszczynski, J., Šponer, J. & Fuentes-Cabrera, M. Theoretical Study on the Factors Controlling the Stability of the Borate Complexes of Ribose, Arabinose, Lyxose, and Xylose. *Chem. - Eur. J.* **14**, 9990–9998 (2008).
108. Müller, D. *et al.* Chemie von a-Aminonitrilen. Aldomerisierung von Glycolaldehyd-phosphat zu racemischen Hexose-2,4,6-triphosphaten und (in Gegenwart von Formaldehyd) racemischen Pentose-2,4-diphosphaten: rac-Allose-2,4,6-triphosphat und rac-Ribose-2,4-diphosphat sind die R. *Helv. Chim. Acta* **73**, 1410–1468 (1990).
109. Krishnamurthy, R., Arrhenius, G. & Eschenmoser, A. Formation of Glycolaldehyde Phosphate from Glycolaldehyde in Aqueous Solution. *Orig. Life Evol. Biosph.* **29**, 333–354 (1999).
110. Hirakawa, Y., Kakegawa, T. & Furukawa, Y. Borate-guided ribose phosphorylation for prebiotic nucleotide synthesis. *Sci. Rep.* **12**, 11828 (2022).
111. Kim, H.-J. & Benner, S. A. Prebiotic stereoselective synthesis of purine and noncanonical pyrimidine nucleotide from nucleobases and phosphorylated carbohydrates. *Proc. Natl. Acad. Sci.* **114**, 11315–11320 (2017).
112. Peretó, J. Out of fuzzy chemistry: from prebiotic chemistry to metabolic networks. *Chem. Soc. Rev.* **41**, 5394 (2012).
113. Nghe, P. *et al.* Prebiotic network evolution: six key parameters. *Mol. Biosyst.* **11**, 3206–3217 (2015).
114. Hanopolskyi, A. I., Smaliak, V. A., Novichkov, A. I. & Semenov, S. N. Autocatalysis: Kinetics, Mechanisms and Design. *ChemSystemsChem* **3**, (2021).
115. Breslow, R. On the mechanism of the formose reaction. *Tetrahedron Lett.* **1**, 22–26 (1959).
116. Decker, P. Spatial, chiral, and temporal self-organization through bifurcation in 'bioids', open systems capable of a generalized darwinian evolution. *Ann. N. Y. Acad. Sci.* **316**, 236–250 (1979).
117. Kholodenko, B. N. Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell Biol.* **7**, 165–176 (2006).
118. Tyson, J. J., Chen, K. & Novak, B. Network dynamics and cell physiology. *Nat. Rev. Mol. Cell Biol.* **2**, 908–916 (2001).

119. Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461 (2007).
120. Cafferty, B. J. *et al.* Robustness, Entrainment, and Hybridization in Dissipative Molecular Networks, and the Origin of Life. *J. Am. Chem. Soc.* **141**, 8289–8295 (2019).
121. Novichkov, A. I. *et al.* Autocatalytic and oscillatory reaction networks that form guanidines and products of their cyclization. *Nat. Commun.* **12**, 2994 (2021).
122. Whitesides, G. M. Physical-Organic Chemistry: A Swiss Army Knife. *Isr. J. Chem.* **56**, 66–82 (2016).
123. Whitesides, G. M. & Ismagilov, R. F. Complexity in Chemistry. *Science* **284**, 89–92 (1999).
124. Briš, A. *et al.* Direct Analysis of Complex Reaction Mixtures: Formose Reaction. *Angew. Chem.* **136**, e202316621 (2024).

Chapter 2

Compositional analysis and characterization of the formose reaction

Accurate characterization of the formose reaction mixture is essential to study the chemical selection mechanisms for reaction pathway formation. The formose reaction was employed under out-of-equilibrium conditions to control its kinetic nature. Compositional snapshots of the formose reaction at different timepoints were used to reconstruct these reaction pathways. A snapshot contains information on the identity and quantity of the different compounds present in the network.

In this chapter, I will discuss how the formose reaction was performed in a Continuous Stirred-Tank Reactor (CSTR). The CSTR setup was used in combination with previously established Gas Chromatography-Mass-Spectrometry (GC-MS) and High-Performance Liquid Chromatography (HPLC) methods to create time-resolved compositional snapshots of the formose reaction.

Further, I will elaborate on the process for interpreting the identity of compounds in the reaction mixture from the GC-MS output. The gas chromatogram contained quantitative information on different compounds in the mixture. The compound identity was interpreted after an initial guess, based on retention time in the gas chromatogram, and subsequent interpretation of its mass fragmentation pattern.

Parts of this chapter have been published in:

1. W.E. Robinson, E. Daines, **P. van Duppen**, T. de Jong, W.T.S. Huck, *Nat. Chem.*, **14**, 623-631 (2022).

2. **P. van Duppen**, E. Daines, W.E. Robinson, W.T.S. Huck, *J. Am. Chem. Soc.*, **145**, 7559-7568 (2023).

2.1 Introduction - The formose reaction as a model prebiotic reaction network

Conditions on a prebiotic earth had the potential to facilitate a plethora of chemical reactions.[1,2] These underpin synthetic organic reaction routes towards key biomolecules and even reaction pathways analogous to the core metabolism.[3–5] Chemical selection in reaction systems from small feedstock molecules (e.g. HC≡N, H_2C=O) remains poorly understood. The formose reaction starts from formaldehyde (**1**) and another small feedstock molecule – e.g. dihydroxyacetone (**2**) – to produce a complex mixture of carbohydrates and polyols.[6–8] The chemical reactivity of these feedstock molecules alone does not dictate the formation of one reaction pathway over another.[9] The emerging prebiotic reaction network, however, is embedded in an environment characterized by physicochemical traits.[1] Environmental conditions provide a potential directing force for the self-organization of prebiotic reaction systems.[10,11]

Different studies have used thermodynamic constraints to control the combinatorial explosion in the formose reaction.[8,12,13] Life, on the other hand, operates under out-of-equilibrium conditions,[14] which are relevant for the prebiotic earth as well.[15] Here, kinetic properties, much rather than thermodynamic, define the observed chemical behavior.[16] The exploration of the formose reaction under out-of-equilibrium conditions is limited.[17–21] However, performing the formose reaction under flow conditions enables control over the steady-state outcome of the reaction.[17–21]

The formose reaction products have a high degree of similarity, which compromises accurate characterization of the full product mixture.[22] In literature examples, characterization of the behavior of the formose reaction often relied heavily on monitoring the concentration of formaldehyde.[17–21,23–26] To study the full behavior and pathway selection in the formose reaction network, an accurate and reproducible method of characterization for the full product mixture is required. Different analytical techniques were developed to characterize the formose reaction mixture.[22]

In this chapter, I will discuss why and how the formose reaction was carried out in a Continuous Stirred-Tank Reactor (CSTR). Subsequently, I will elaborate on the workflow for analysis of the reactor output with HPLC and GC-MS. Further, I will explain how the analytical results were interpreted to obtain compositional snapshots of the CSTR.

2.2 The formose reaction in a CSTR

For the experimental studies as discussed in this thesis, the formose reaction was carried out in a CSTR, see figure 2.1a. This setup represents fluvial conditions on

a prebiotic earth. Syringe pumps with aqueous solutions of formaldehyde, initiator sugar, $CaCl_2$, NaOH and H_2O were connected to the input channels of the CSTR (fig. 2.1a and 2.6.1). The chemical environment was varied by changing the concentration of respective input chemicals. The reactor flow rate and temperature were used as physical control parameters. In chapter 4, I will elaborate on how the steady-state of the formose reaction changed for different combinations of environmental conditions. Important to note: the complexation of different sugars with Ca^{2+} increases the solubility limit of $Ca(OH)_2$ (21 mM at 25 ^{o}C)[27] as discussed under 2.6.3. The interaction between Ca^{2+} and formose reaction products therefore increase the input range for $CaCl_2$ and NaOH.

Different experiments were performed with modulated input concentrations into the reactor. For example, the flow rate of initiator sugar was modulated over time (chapter 3 and 4). The input flow rate of a syringe containing water was simultaneously adjusted against the initiator sugar flow rate, to maintain a constant residence time in the reactor. In chapter 3, I will elaborate further on how the differential transfer from the input modulations to the measured compounds was used to reconstruct operational reaction pathways.[28–30]

The CSTR allowed for accurate and reproducible characterization of the formose reaction products. The contents of the CSTR were sampled to obtain time-resolved compositional measurements of the reaction progress, see figure 2.1b.[15,31] Droplets (35 ± 0.1 µL) at the outlet of the CSTR pinched off the spout, dropped into an Eppendorf tube which was transferred directly into a bath of liquid nitrogen. The droplet was freeze-quenched immediately to stop the reaction progress. The sample was prepared for further analysis with HPLC or GC-MS and interpreted as discussed in subsequent sections 2.3 and 2.4.

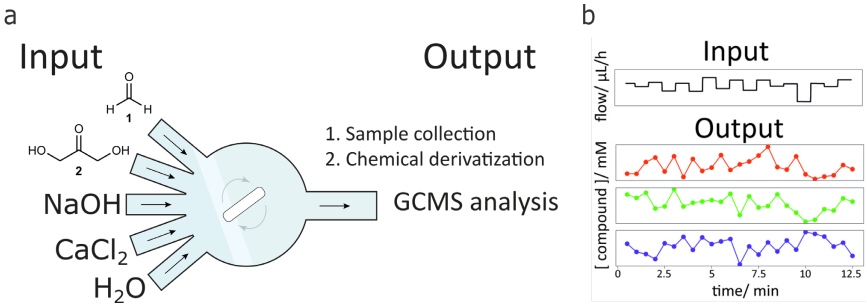


Figure 2.1: The experimental CSTR setup and a typical input and output concentration profile. a) Schematic drawing of a CSTR, used to employ the formose reaction under flow conditions. b) Typical modulated flow input and measured output concentration profile. See 2.6.1 for a detailed schematic of the experimental flow setup.

2.3 Analysis of compositional snapshots of the formose reaction

Untangling the reaction composition of the formose reaction has been a major challenge, since products have a high degree of structural similarity and stereoisomers.[22,32] The mixtures were characterized after chemical derivatization with both gas and liquid chromatographic methods.

The obtained output from the reactor was analyzed by Gas Chromatography-Mass-Spectrometric analysis (GC-MS) and High-Performance Liquid Chromatography (HPLC). The GC-MS analysis provided information on the compounds produced in the formose mixture. The HPLC analysis contained quantitative information on the feedstock molecules **1** and **2**. In this section, I will first discuss how the collected samples were prepared for GC-MS and HPLC analysis by chemical derivatization. Next, I will explain the quantitative analysis from the output chromatogram and the qualitative analysis by GC-MS.

2.3.1 Chemical derivatization of the CSTR output mixture for GC-MS

In the chemical derivatization for GC-MS, the carbohydrates were modified in two chemical reactions. The analysis can be potentially complicated by dynamic formation of open-chain and furanose/pyranose ring structures. Therefore, first the carbonyl substituent was converted to an *O*-ethyl oxime bond, see figure 2.2 (orange).[22] This way, all furanose and pyranose sugars were converted to an open chain formation. In the second step, the hydroxyl groups were converted to a *O*-trimethylsilyl derivative (fig. 2.2, blue).[22,33,34] The derivatization yielded open chain and asymmetric molecules, which had both distinct gas chromatographic and mass fragmentation properties.[22,33,34]

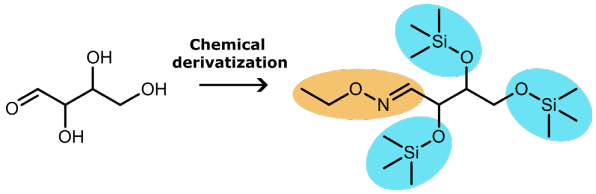


Figure 2.2: Chemical derivatization of carbohydrates. The carbonyl substituent was converted to an *O*-ethyl oxime bond (orange) and the hydroxyl substituents were converted to *O*-trimethylsilyl substituents (blue).

Before derivatization, water was removed from the frozen samples by lyophilization, see figure 2.3a. In step 1 of the derivatization, the aldehydes were converted to corresponding oximes after the addition of 75 µL *O*-ethylhydroxylamine ($EtONH_2$) in pyridine (20 mg/mL) (fig. 2.3b). The samples were heated at 70 °C and shaken at 700 rpm for 30 minutes. After cooling to room temperature, the hydroxyl groups were silylated with 25 µL *N,O*-bis(trimethylsilyl)trifluoroacetamide (BSTFA), see step 2 (fig. 2.3b). The

samples were heated at 70 °C and shaken at 700 rpm for 30 minutes. After cooling to room temperature, in step 3, 100 µL internal standard mixture in pyridine (1.6 mM dodecane and tetradecane) was added to the samples (fig. 2.3b). 100 µL of sample volume was transferred from the derivatized sample mixture to a GC vial, sealed with an airtight cap and put on the GC-MS for analysis (fig. 2.3c).

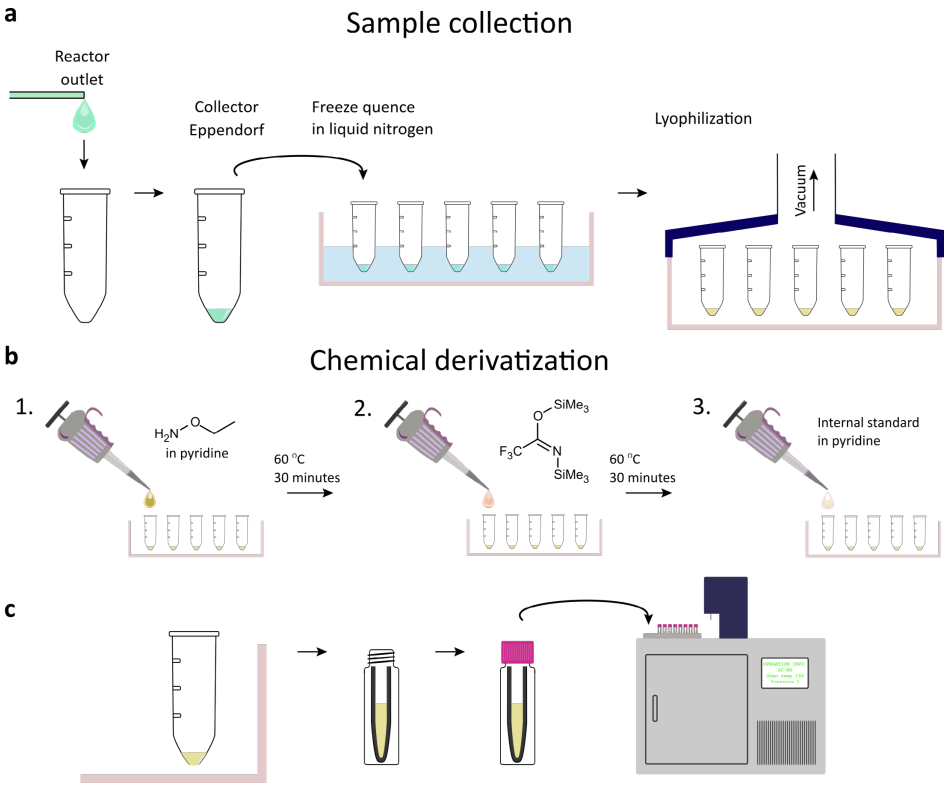


Figure 2.3: Sample preparation from CSTR output to GC-MS analysis *via* chemical derivatization. a) Freeze quenched samples from the CSTR output and lyophilization. b) Chemical derivatization with the subsequent addition of $EtONH_2$ in pyridine (1), BSTFA (2) and internal standard mixture (3). Between reagent addition the samples were shaken for 30 minutes at 60 °C. c) Transfer of derivatized formose mixture into a GC vial, which were placed on the GC-MS for analysis.

2.3.2 Chemical derivatization for HPLC analysis of feedstock molecules

For analysis by HPLC, the compounds in the output mixture were converted to a hydrazone, in a reaction with 2,4-dinitrophenylhydrazone (DNPH), see figure 2.4.[22] The DNPH substituent functioned as a chromophore to detect the reacted sugar with a UV-VIS detector at 364 nm.

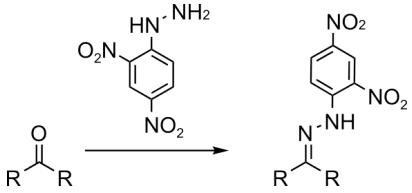


Figure 2.4: Reaction scheme for chemical derivatization of formaldehyde with DNPH for HPLC analysis.

The process of chemical derivatization for 35 µL of reaction mixture was caried out in a mixture of 2.5 µL 2M HCl, 65 µL water, 97.5 µL acetonitrile, 300 µL acetonitrile saturated with DNPH. This was incubated for 30 minutes at room temperature before HPLC analysis.

2.3.3 Sample analysis on GC-MS

Each sample was injected onto the GC column, where different compounds in the mixture were separated. Mass-spectrometry was used to obtain a typical mass fragmentation pattern to identify the compound structure. The temperature gradient in the GC was optimized to separate sugars with different carbon backbone length.[22]

Gas-Chromatography-Mass Spectrometric analysis were performed on a *JEOL JMS-100GCv*. The *Agilent 7890A GC* gas chromatograph was equipped with a *HP-5MS* column (length: 30 m, inner diameter: 0.25 mm, film thickness: 0.25 µm). The injector inlet was set to a temperature of 250 °C and for sample analysis (injection volume: 1 µl) split mode was applied (ratio: 1/10). The GC was operated with the following temperature program: oven temp/ °C: 100, 170, 210, 250, 325, rate/ °C min.$^{-1}$: 0, 14, 4, 15, 60, time/ min.: 2.33, 0, 0, 0, 3.75. Helium was used as a carrier gas (flow rate: 1 mL/min). For the mass analysis a JEOL AccuTOF mass spectrometer was used with an Electron Impact Ionization Mode.

The injection syringe was rinsed before and after each injection with dichloromethane and cyclohexane. At least once per 24 hours, the syringe was manually rinsed with dichloromethane.

2.3.4 Sample analysis on HPLC

The derivatized feedstock molecules **1** and **2** were separated on the HPLC column for quantification. Larger sugar derivatives were not optimally separated for characterization and quantitative analysis.

The HPLC analysis was performed on a *Shimadzu Nexera X2* instrument. Conditions: GIST C18 column (2 µm pore size, 75 x 3.0 mm), 40 °C, 0.8 mL min^{-1},

acetonitrile : water (1 : 1, 0.1% trifluoroacetic acid), 1 µL injection volume, UV-vis detection at 364 nm, or GWS C18 column (5 µm pore size, 250 x 4.6 mm) at 1.0 mL min.$^{-1}$, 40 °C, 1.0 mL min^{-1} , acetonitrile : water (1 : 1, *v/v*, 0.1% trifluoroacetic acid), 1 µL injection volume, UV-vis detection at 364 nm.

2.4 Interpretation of compositional snapshots of the formose reaction and the underlying reaction network

The compositional snapshots of the formose reaction in the CSTR were extracted from the GC-MS output. For chapter 4, also HPLC analysis was used to quantify the feedstock molecules **1** and **2**. The analysis provided both structural and quantitative information on the different compounds present in the reactor output. In this section, I will explain how the GC-MS output was analyzed to reconstruct a snapshot of the formose mixture.

2.4.1 Information in the GC-MS output

The output of the GC-MS contains both a gas chromatography (GC) trace and the corresponding mass fragmentation spectra. Accurate interpretation of the formose mixture was complicated by the structural similarity of the analytes, mainly carbohydrates, sugar acids and polyols.[22] Separating out structurally identical compounds on the GC column was required for characterization of the different analytes. In the GC trace, each peak corresponded with a compound with unique structural properties, see figure 2.5a. Each compound was quantified by integrating the peak area, as elaborated further under 2.4.2.

The GC output was sampled at 10 Hz for mass spectrometric analysis. At each sample point, a mass fragmentation spectrum was acquired (fig. 2.5b). By stacking these mass spectra, bell-shaped curves emerged for each observed mass fragment, see the side view in figure 2.5b. The GC trace was constructed from a time-stack of the total ion count in the acquired mass spectra. The mass spectrum is unique for the different compounds in the formose mixture. Therefore, the underlying mass fragmentation pattern of an observed GC peak is an important factor for compound identification, as will be elaborated further under 2.4.3.

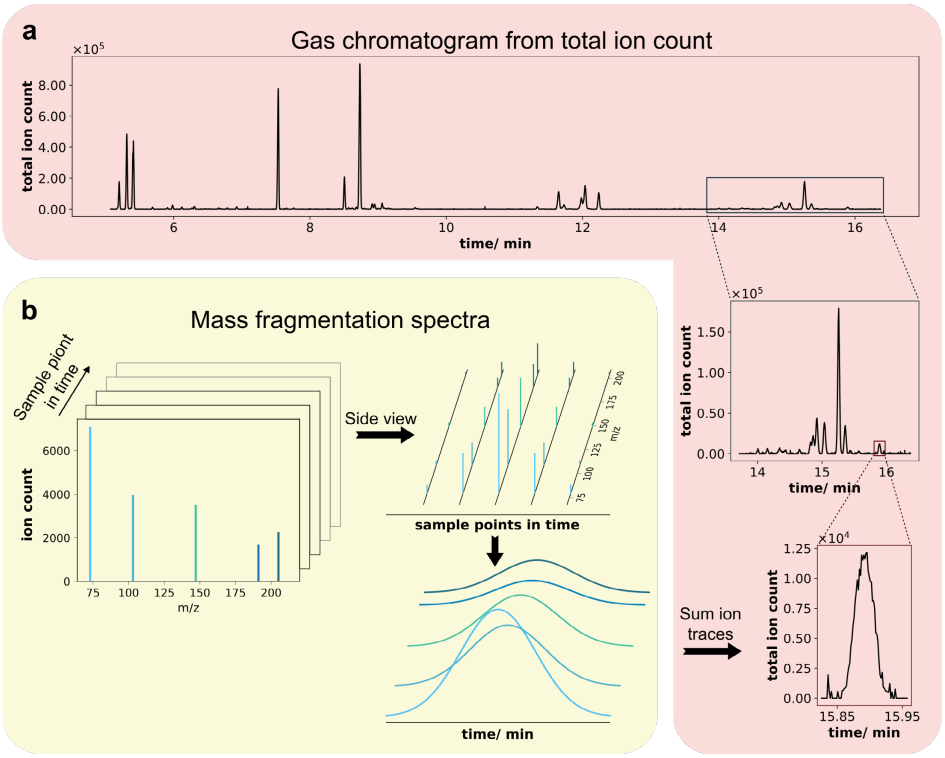


Figure 2.5: Construction of gas chromatogram from total ion count at each MS sample point. a) Constructed gas chromatograph from a time-stack of the total ion count in the acquired mass spectra. b) Mass fragmentation spectra, acquired at each sample time for the gas chromatograph output into the mass spectrometer.

2.4.2 Acquiring quantitative information from the GC output

The concentration of a compound in the formose mixture correlated with the area under the respective peak in the chromatographic output trace, both in HPLC and GC data. In this section, I will discuss the signal processing and peak integration of the GC trace. The interpretation of HPLC chromatograms did proceed in a similar fashion, except for preprocessing step of the output signal (noise filtering). Therefore, I will omit a separate discussion for peak quantification for HPLC data.

From the GC output, noise around the baseline was removed in a preprocessing step to obtain the total ion chromatogram. Subsequently, the boundaries of all peaks were determined and the peak areas were integrated. For precise quantification of the compounds an internal standard was used.[22,35] The compound concentration was calculated with calibration lines, which were based on commercially available carbohydrates.

The high signal to noise ratio conceals peaks from low concentration compounds (peak maximum < 30.000), see figure 2.6a. To reduce the background noise, a preprocessing step was performed on the GC-MS output. A mass-filter was applied on the acquired mass spectra before the total ion chromatogram was calculated. In this step, all fragments with an ion-count below a set threshold (e.g. ion count < 500, see 2.6.4) were removed from the mass spectra (fig. 2.6b). These stripped spectra were used to calculate a new gas chromatogram from total ion count. After the mass-filter was applied, the background noise around the baseline was removed to reveal peaks from low concentration compounds (fig. 2.6c).

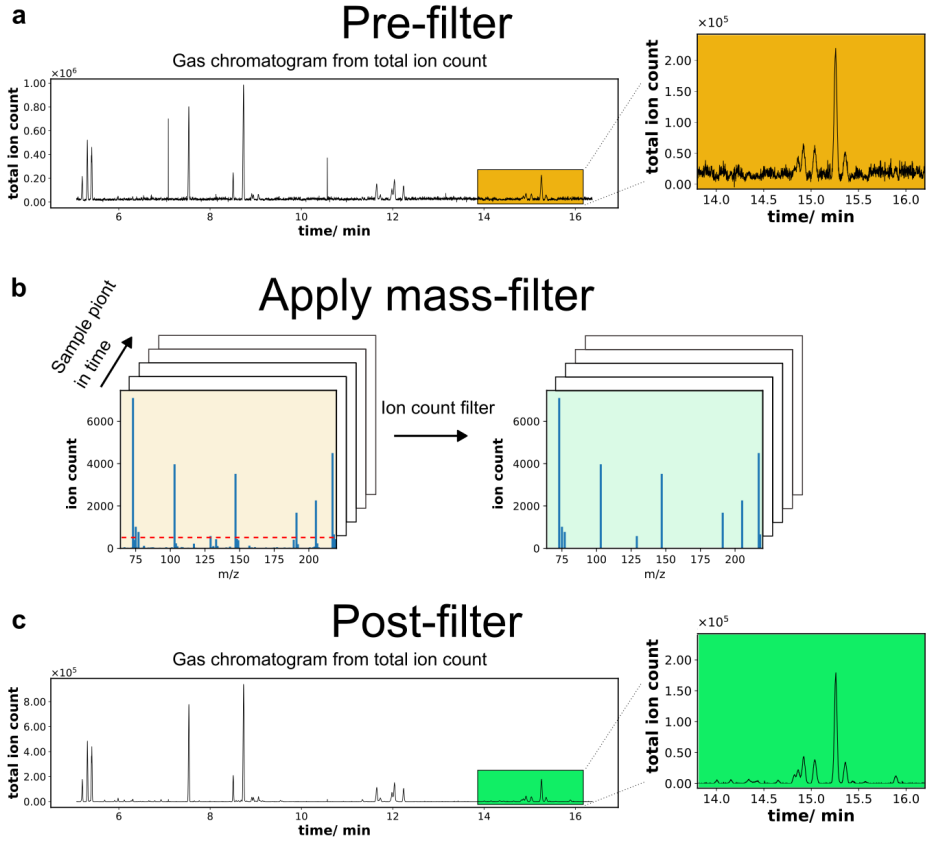


Figure 2.6: **Noise reduction in the GC chromatogram with a mass-filter.** The high noise in the GC chromatogram (a) was reduced by applying a mass-filter in each of the acquired mass spectra (b) to obtain a post-filter chromatogram (c), for quantitative interpretation.

The peaks in the GC trace were defined by determining the peak boundaries in the time domain. A first order differential of the GC signal was calculated to localize the peaks on the time axis, see figure 2.7a. With a simple algorithm the shape of the differential signal was interpreted to identify where the peak starts and finishes. At the peak maximum, the differential of the signal descends through zero (fig. 2.7b). The algorithm moved in both directions from this intersection of the differential with the horizontal axis. It moved left (right) until the differential was no longer larger (smaller) than zero to define the start (end) of the peak. For two overlapping peaks, the border between the peaks was defined where the differential of the two peaks rises through zero (fig. 2.7c).

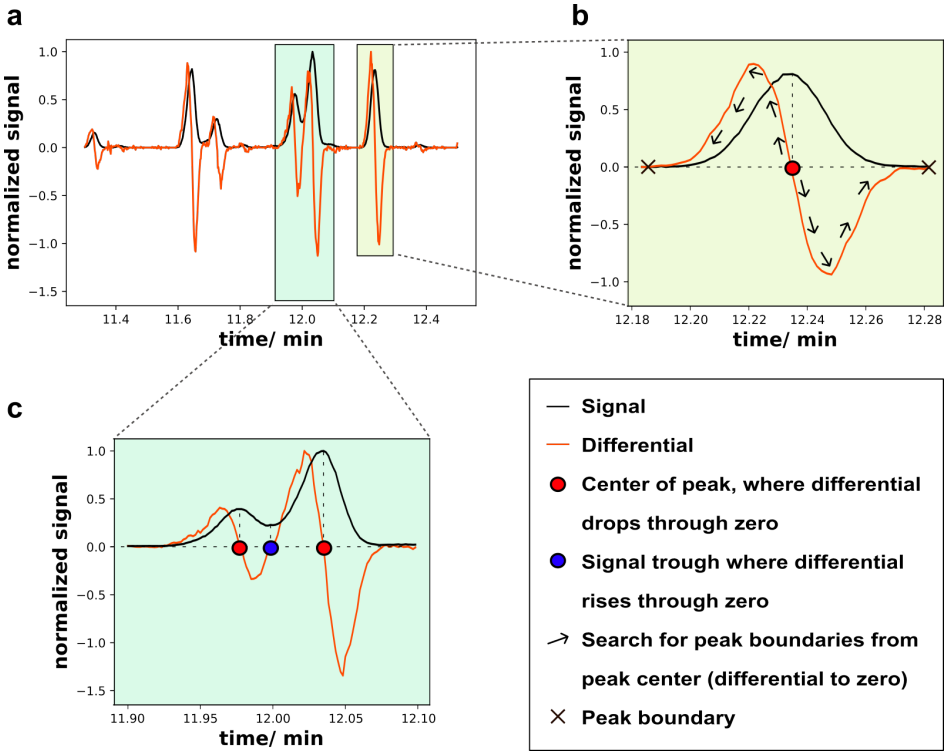


Figure 2.7: Peak identification and determination of peak boundaries with the signal first order differential. a) An overlap of a GC signal and the corresponding first order differential (orange). b) Zoom of an individual GC peak. c) Zoom of two overlapping GC peaks.

Additional conditions were applied for the peak identification algorithm. The differential was only used for peak identification, if the peak maximum crosses a set threshold (e.g. > 1 % intensity of the largest peak) and was no larger (smaller) than a set maximum (minimum) peak intensity (by default an ion count

$< 1 \times 10^{100}$ ($> 1 \times 10^{-100}$)) in the GC signal. Between the peak maxima, a minimum time distance was retained (0.1 minutes).

The peak areas were calculated with the identified peak boundaries, such as in figure 2.8. Each peak was integrated with the numpy.trapz() function.[36] The baseline of the peak was linearly interpolated between the defined peak boundaries (fig. 2.8).

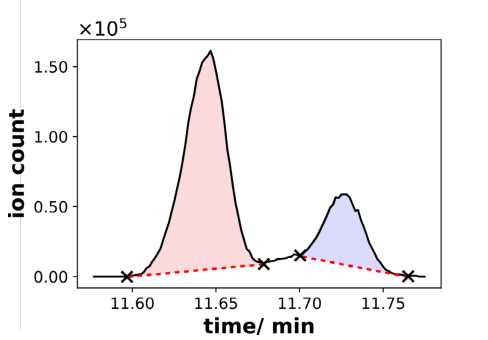


Figure 2.8: Area for peak integration, marked by the red dotted lines. The baseline was a linear interpolation between the defined peak boundaries.

The compound concentration was converted from the peak integral with a calibration curve. For precise quantification of the analytes, an internal standard (mixture of tetradecane (0.8 mM) and dodecane (0.8 mM)) was used to account for variability in sample injection.[22,35] To normalize the peaks in a GC trace, each identified peak area was divided by the tetradecane peak area.

The conversion of the peak integral to the corresponding concentration was different for each analyte. Only after compound identification (see 2.4.3) the integral was converted to a concentration. Quadratic calibration curves were fitted for commercially available sugars, see figure 2.9 for an example calibration curve of arabinose. For identified non-commercially available compounds, an average calibration curve for all calibrated sugars with similar molecular weight was used. The largest peak was used in the data analysis in case a compound had two peaks in the chromatogram, arising from the *syn* and *anti* form of the oxime.

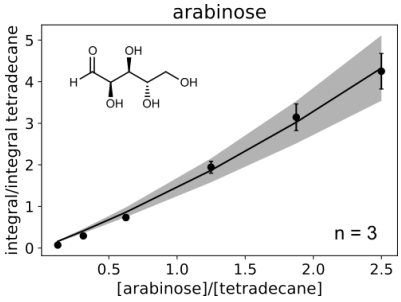


Figure 2.9: Calibration curve example of arabinose. Each timepoint in the calibration curve was measured in triplicate (n = 3).

2.4.3 Compound identification based on GC retention time and mass fragmentation pattern

Structural information of the compounds in the formose mixture was obtained both from the GC retention time and mass fragmentation spectrum. First, the identified peaks were compared to reference GC-MS data of commercially available sugars. Unidentified peaks were assigned by inferring the molecular structure from GC retention time and mass fragmentation pattern.

The peak retention time in the GC contains structural information. Each carbon number (C_3, C_4, C_5 and C_6) of reaction products eluted in a specific time-region of the chromatogram, see figure 2.10. Chemically derivatized carbohydrates (2.3.1) did elute later from the column as the carbon number increased.

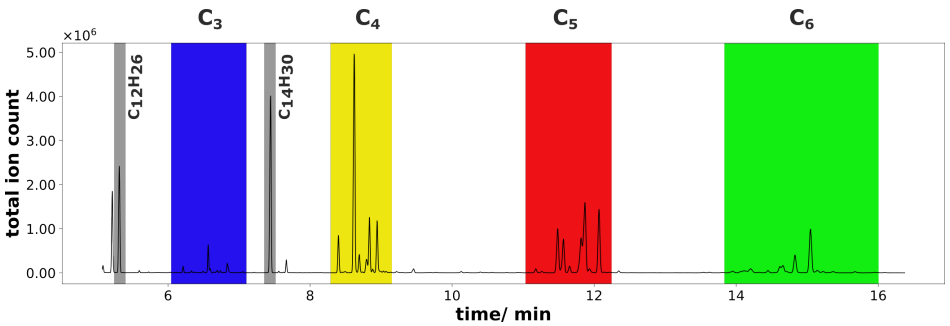


Figure 2.10: Carbon number related to a GC retention time region. Peaks in grey are from the internal standards dodecane ($C_{12}H_{26}$) and tetradecane ($C_{14}H_{30}$).

The mass fragmentation pattern was specific for the structural properties of the corresponding parent compound.[37] Therefore, further structural information was obtained from the mass fragmentation pattern which corresponds to a GC peak. The mass spectra of carbohydrates contain three types of fragments (fig. 2.11). A non-specific category contains molecular ions derived from the trimethylsilyl derivatization reagent (fig. 2.11a). The other two categories contain analyte-specific molecular ions, which contain part of the carbon

backbone. The first category represents molecular ions from silylated fragments (fig. 2.11b). The other category also contains the ethyl oxime bond (fig. 2.11c), which has formed on the sugar carbonyl (2.3.1).

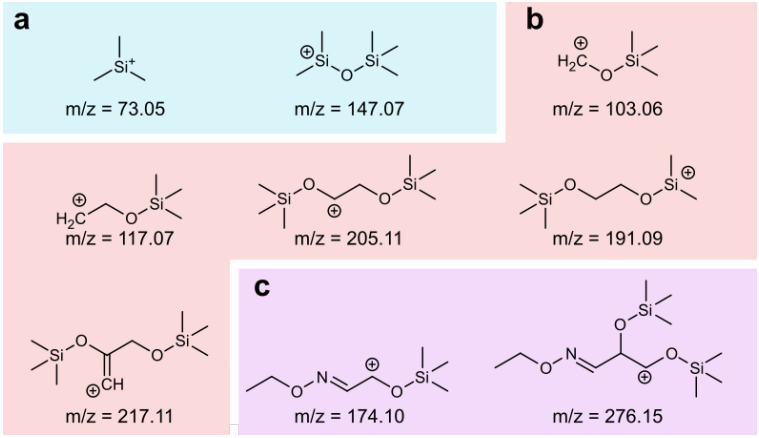


Figure 2.11: **Mass fragment structures** of a) non-specific derivatization reagent, b) silylated analyte fragments and c) oximated analyte fragments.

The analyte-specific molecular ions revealed information on the respective molecular structure. The difference in bond stability in the carbon backbone dictates what molecular ions were formed.[37] For example, the carbon backbone breaks preferably not directly adjacent to the oxime moiety, see figure 2.12.[38] A remarkable similarity in fragmentation pattern was found between both the aldotetrose erythrose and β-ketopentose xylulose (fig. 2.12b,e), and both the aldopentose xylose and the β-ketohexose fructose (fig. 2.12d,f). Similar fragmentation patterns arose from a molecular structure which had an analogous molecular structure next to the carbonyl, irrespective of the stereochemistry. Molecular ions with the oxime bond were scarce and only present for dihydroxyacetone (*m/z* = 174.10) (fig. 2.12a).

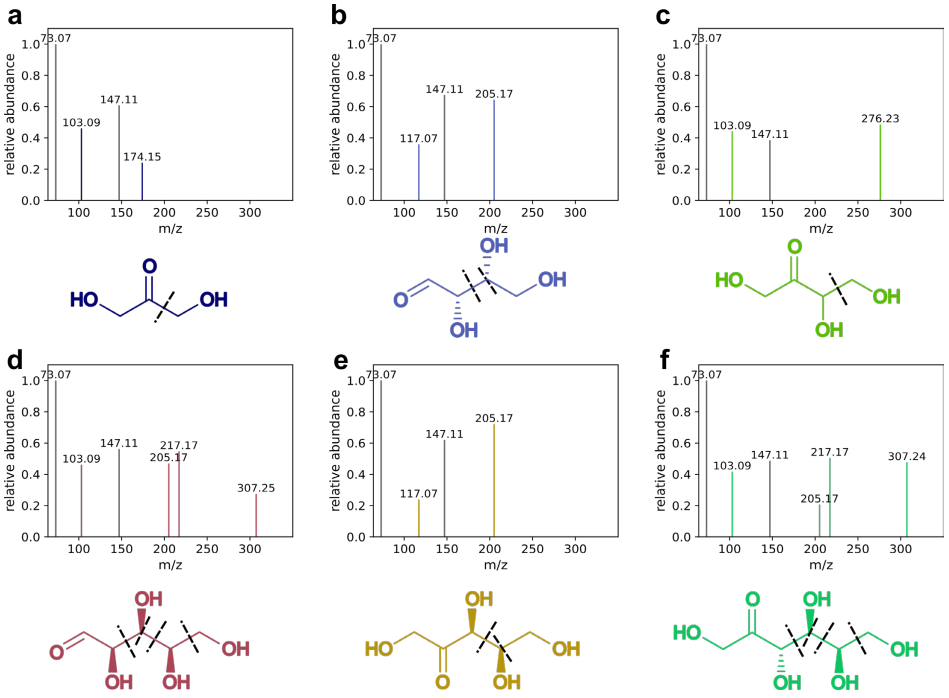


Figure 2.12: **Mass fragmentation patterns** for a) dihydroxyacetone b) threose c) erythrulose d) xylose e) xylulose and f) tagatose. The analyte specific molecular ions were depicted in the same color as the respective molecular structure. Peaks at m/z = 75.05 and 143.10 were omitted from the mass spectra for simplicity and only peaks with a higher relative abundance than 0.2 were displayed.

The fragmentation patterns of commercial compounds were used to infer the molecular structure for unidentified peaks. The GC retention time was used to find the carbon number of the analyte. For example, all peaks between 11.0 and 12.5 minutes were assumed C_5 sugars. An unknown C_5 sugar had mass fragments of *m/z* = 117.07 and 205.17, see figure 2.13a, similar to threose. Therefore, this compound was assumed to have a linear C_3 fragment next to a carbonyl. Since it was a C_5 compound the hydroxymethyl substituent was located at the α position, the compound was assigned as α-hydroxymethyl-*R,R*-aldotetrose. Similarly, with the commercially available ketotetrose erythrulose, a C_5 sugar was assigned (fig. 2.13b). Both the unassigned C_5 sugar and erythrulose only had mass fragment *m/z* = 103.09 as silyl molecular ion. The unassigned C_5 sugar also had a hydroxymethyl substituent located at the α position. The compound was therefore assigned as α-hydroxymethyl-2-ketotetrose. The molecular structure of unassigned C_6 sugars was inferred in a similar fashion. For example, for the α-hydroxymethyl-*R,R,R*-aldopentose, which had a similar fragmentation pattern

as the aldopentose xylose (fig. 2.13c). The previous assignments for unknown compounds were also used to aid the assignment unidentified peaks. For example, the α-hydroxymethyl 2-ketotetrose (fig. 2.13b) was used to identify C_6 sugar α-hydroxymethyl 3-ketopentose (fig. 2.13d).

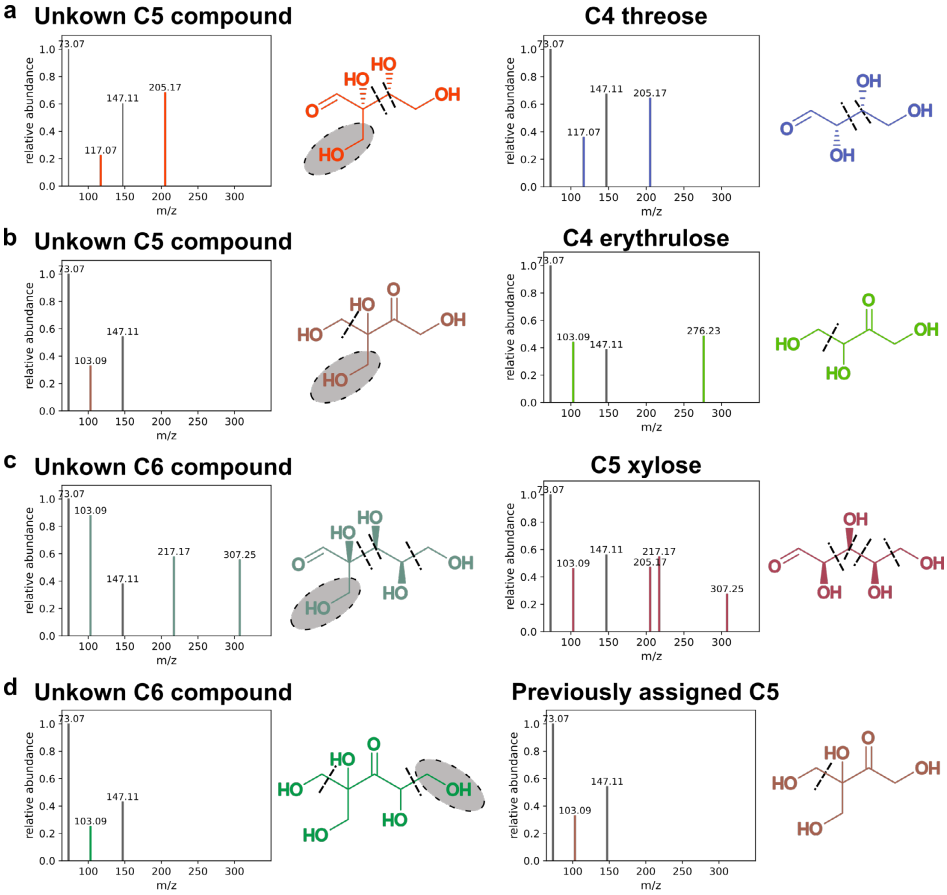


Figure 2.13: **Mass fragmentation patterns for non-calibrated, branched carbohydrates** for a) α-hydroxymethyl-*R,R*-aldotetrose based on threose fragmentation b) α-hydroxymethyl-2-ketotetrose based on erythrulose c) α-hydroxymethyl-*R,R,R*-aldopentose based on xylose and d) α-hydroxymethyl-3-ketopentose based α-hydroxymethyl-2-ketotetrose. Peaks at m/z = 75.05 and 143.10 were omitted from the mass spectra for simplicity and only peaks with a higher relative abundance than 0.2 were displayed.

From GC retention time and mass fragmentation patterns alone, peak assignment was not unambiguous for non-calibrated sugars. Different molecular structures can correspond to a particular mass fragmentation pattern. Therefore, the compound assignment is an iterative process aided by

connectivity of reconstructed reaction networks. In other words, once more is known about the reaction pathways that connect a certain unknown compound to other known compounds, unambiguous assignments can often be made. This process will be discussed further in chapter 3.

2.5 Conclusion

The formose mixture was deemed intractable as it produced a plethora of the structurally similar products.[22,32] The CSTR provides an out-of-equilibrium setup to contain the combinatorial explosion in the formose reaction. This platform allows the study of compositions of the formose reaction mixture both in constant (steady-state) and modulated environments. The composition of samples taken from the reactor were studied using both HPLC and GC-MS analysis. Before analysis, the output samples were chemically derivatized with DNPH, or $EtONH_2$ and BSTFA, respectively.[22]

The output of the HPLC analysis gives quantitative information on input feedstock molecules in the CSTR. Quantitative and structural analysis of the reaction products was obtained by GC-MS. The GC trace contained quantitative information on concentrations of compounds present in the reactor, and on the carbon chain length. The structural information was interpreted from mass fragmentation patterns that correspond to the respective GC peak. Identification of non-calibrated compounds was based on fragmentation patterns of calibrated compounds.

In chapter 3, different approaches to reconstruct reaction pathways from experimental data are explored. The reconstructed reaction networks provide a mechanistic insight in how the formose reaction network interacts with different chemical environments, as discussed in chapter 4 and 5.

2.6 Supplementary information

2.6.1 Detailed scheme of the reaction setup

The scheme in figure S2.14 indicates how five inlet syringes ($H_2C{=}O$, DHA, $CaCl_2$, NaOH, H_2O) were connected to the bottom of a CSTR (side view), constructed from polydimethylsiloxane (PDMS), as discussed under 2.6.2. The cone shaped reactor was temperature controlled and had the outlet at the top, which was connected to a droplet spout. The 35 ± 0.1 µL droplets were collected and immediately freeze-quenched in liquid nitrogen.

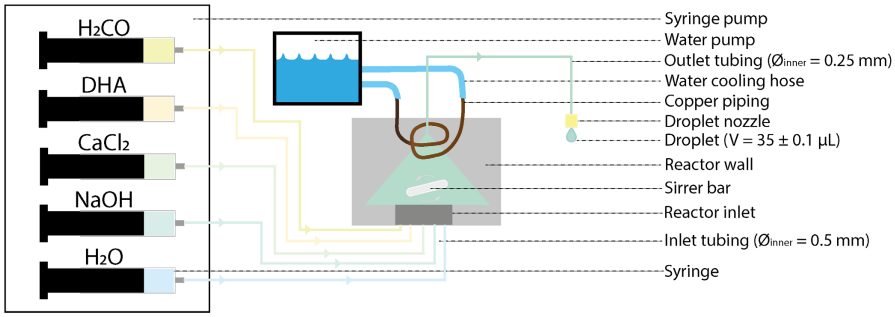


Figure S2.14: **A detailed schematic representation of the flow reactor setup.**

2.6.2 CSTR construction

The CSTR was constructed from a PDMS body on a glass slide, see figure S2.15. The reactor cavity was formed around a cone shaped mold. For temperature control, a curled copper pipe connected *via* a temperature-controlled water pump, was added around the reactor cavity. Inlet tubing was inserted into channels in the PDMS, entering the bottom of the reactor. The outlet tubing was attached to the top of the reactor.

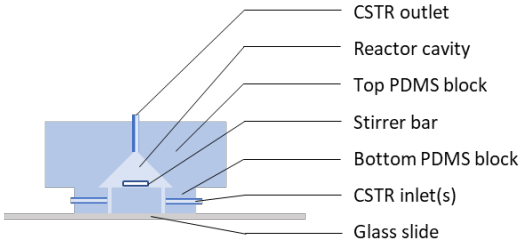


Figure S2.15: **Schematic CSTR constructed from PDMS on a glass slide.**

Short procedure for fabricating CSTR: The silicone elastomer crosslinker (4.9 mg [14 wt %], Slygrad ® 184) was added to the silicone elastomer base (30 mg, Sylgard ® 184). After mixing, the elastomer mixture was placed in a vacuum desiccator for degassing until no air bubbles emerged from the mixture.

The reactor base was prepared by pouring the degassed crosslinker/elastomer mixture into a Petri dish (35 x 10 mm^2) to a depth of 5 – 6 mm. The mixture and Petri dish were placed in a vacuum desiccator to remove air bubbles, before being placed in an oven (65 ^oC, 2 hours). From the hardened PDMS was cut a 20 x 20 mm^2 square. The desired number of inlet holes were punched through the PDMS with a 1.5 mm biopsy puncher. The holes were placed such that the inlets emerge at the edge of the reactor cone. From the side of the PDMS square, holes punched to the inlet hole with a 1.0 mm biopsy punch to meet the inlets. Inlet tubing was inserted into these holes.

A cone shaped brass mold with the desired dimensions (d = 14 mm, h = 10 mm) was placed in a Petri dish (35 x 10 mm^2) and a few drops of PDMS were placed at the side of the cone to form a thin layer at the bottom of the Petri dish. Air bubbles were removed in the vacuum desiccator and the Petri dish was placed in an oven (65 ^oC, 2 hours). The coiled copper pipe was laid around the brass cone on the thin PDMS layer, such that it did not touch the side of the cone. The Petri dish was filled with elastomer mixture, air bubbles were removed in the desiccator and the Petri dish was again placed in an oven (65 ^oC, 2 hours). The hardened PDMS was removed from the Petri dish and the brass cone was removed from the PDMS. A hole was punched from the tip of the cone to the outside of the reactor body with a 1.5 mm biopsy puncher.

Before assembly, the reactor base, body, stirring bead/bar and glass slide were washed with 2-isopropanol and dried under N_2 gas flow. The components were treated in a plasma cleaner (3 minutes oxygen flow prior to a 20 second bonding cycle at 100 W). The PDMS base was pressed firmly onto the glass slide (1 minute). The reactor body was bonded to the reactor base in a similar manner, with the stirring bar placed inside the reactor cavity. The assembled reactor placed in an oven (100 ˚C, 2 hours). Teflon tubing (OD = 1 mm, ID = 0.5 mm) was pushed in the inlet holes of the reactor to connect to the syringes. Teflon tubing with (OD = 1.63 mm, ID = 0.25 mm) was pushed in the top of the reactor as outlet. Optionally the whole reactor with tubing can be covered in PDMS and cured to reduce the risk of leakage.

2.6.3 Concentration limits of Ca^{2+} and OH^- in the CSTR

$Ca(OH)_2$ has a limited solubility in water of 21 mM at 25 ^oC (1.58 g L^{-1}) and at higher concentrations a white precipitate is formed.[27] In the performed experiments, this concentration limit was exceeded, as the upper limit of concentration of $CaCl_2$ and NaOH in the performed experiments were 52.0 mM and 96.0 mM, respectively. Over the experimental conditions, the ratio between $CaCl_2$ and NaOH was varied and within the studied combinations, no precipitation of $Ca(OH)_2$ was observed. Complex formation of Ca^{2+} with sugar diols or enolate species, were likely increasing the upper concentration limit before $Ca(OH)_2$ precipitates.[39] For example the presence of D-fructose, the

solubility of $Ca(OH)_2$ was shown to increase from 16 mM to at least 340 mM at 60 °C.[19] Therefore, we assume that all of the reactions occurred in homogeneous mixtures.

2.6.4 Interpreting and summarizing the GC-MS output with ChromProcess

The Python program for the following described data analysis is available at https://github.com/Will-Robin/ChromProcess. The data analysis in this thesis was performed with older versions of ChromProcess (version from 21-07-2021 and older).

For the data analysis, ChromProcess obtained information about the experiment and parameters for the data analysis.

Information about the experiment was provided in a file *EXP000_conditions.csv*, which contained all information about the experiment and the conditions. For the experimental information the following was provided:

1. Series values for the collected samples (e.g. the exact sampling time corresponding to the flow profile, or the sample number).
2. The series unit (e.g. sample_time/ s or sample_number/ n).

For the experimental conditions the following was provided, respective concentrations corresponded to the syringe concentration:

1. reactor volume/ µL
2. $[sugar]_{syringe}$/ M
3. $[formaldehyde]_{syringe}$ / M
4. $[NaOH]_{syringe}$ / M
5. $[CaCl_2]_{syringe}$ / M
6. $[water]_{syringe}$ / M (always 0)
7. Flow profile/ s (time axis for following flow profiles)
8. Sugar flow rate/ µL/h
9. formaldehyde flow rate/ µL/h
10. NaOH flow rate/ µL/h
11. $CaCl_2$ flow rate/ µL/h
12. water flow rate/ µL/h
13. residence time/ s

The analysis details were provided in a file *EXP000_analysis_details.csv*:

1. Method (GCMS or HPLC).
2. Defined chromatogram regions for peak integration/ minute (alternating start – end point of the different regions, figure 2.10).
3. Internal reference region/ (start - end point).
4. Internal reference concentration/ M.

5. Dilution factor for GC-MS or HPLC derivatization.
6. Extract mass spectra (TRUE or FLASE).
7. Mass spectra filter (cut-off below certain ion count (fig. 2.6).
8. Pick peak threshold (as a fraction of the largest peak (1))

To account for local variability, peak borders for GC-MS or HPLC analysis were defined in: *EXP000_local_assignments.csv*. For each adjusted compound region, a start and end boundary was defined:

'name compound'; start time/ minutes; end time/ minutes.

For calculating the concentration of the peak integrals a calibration file was provided. For each assigned compound, respectively A and B were provided to calculated the concentration *via:*

$$[compound] = A \times \left(\frac{peak\ integral}{internal\ standard\ integral}\right)^2 + B \times \frac{peak\ integral}{internal\ standard\ integral}$$

A calibration curve was estimated for compounds which were not commercially available by taking the average calibration curve for the calibrated sugars of similar length.

Now, ChromProcess was allowed to analyze HPLC and GC-MS output data as respectively .csv and .cdf file formats with the following script '2_scrape_single_folder.py'. For each sample the respective chromatograms, peak table (with peak boundaries) and mass spectra were stored in separate folders as .csv files. The report with peak integrals and the report with compound concentrations were stored in a separate folder.

The HPLC and GC-MS analysis output is available for all data for chapter 4 at https://github.com/huckgroup/formose-2021/tree/main/DATA.

The GC-MS analysis output is available for all data for chapter 5 at https://github.com/huckgroup/Formose_2022/tree/main/Extended_data_GH.

2.7 References

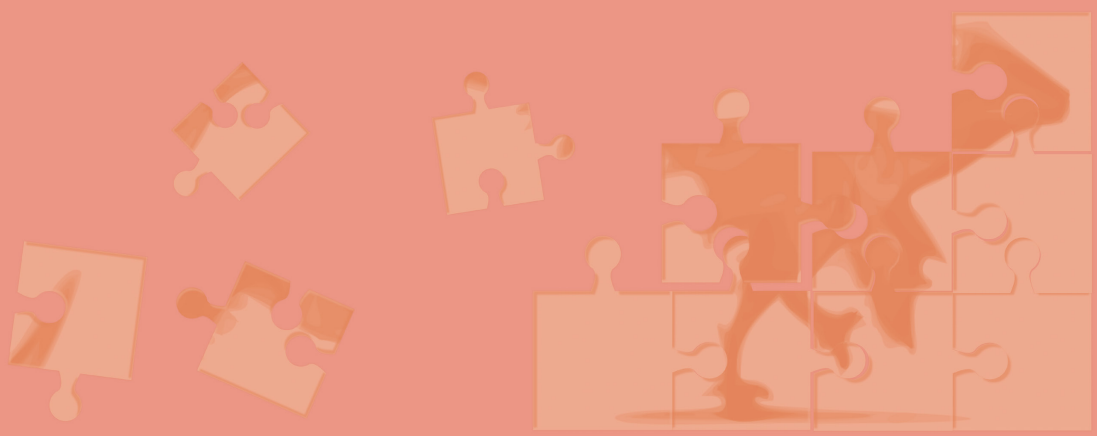
1. Sasselov, D. D., Grotzinger, J. P. & Sutherland, J. D. The origin of life as a planetary phenomenon. *Sci. Adv.* **6**, eaax3419 (2020).
2. Islam, S. & Powner, M. W. Prebiotic Systems Chemistry: Complexity Overcoming Clutter. *Chem* **2**, 470–501 (2017).
3. Patel, B. H., Percivalle, C., Ritson, D. 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).
4. Muchowska, K. B., Varma, S. J. & Moran, J. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019).

5. 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).
6. 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).
7. Kim, H.-J. *et al.* Synthesis of Carbohydrates in Mineral-Guided Prebiotic Cycles. *J. Am. Chem. Soc.* **133**, 9457–9468 (2011).
8. Colón-Santos, S., Cooper, G. J. T. & Cronin, L. Taming the Combinatorial Explosion of the Formose Reaction via Recursion within Mineral Environments. *ChemSystemsChem* **1**, (2019).
9. Zubarev, D. Y., Rappoport, D. & Aspuru-Guzik, A. Uncertainty of Prebiotic Scenarios: The Case of the Non-Enzymatic Reverse Tricarboxylic Acid Cycle. *Sci. Rep.* **5**, 8009 (2015).
10. Cronin, L. & Walker, S. I. Beyond prebiotic chemistry. *Science* **352**, 1174–1175 (2016).
11. Surman, A. J. *et al.* Environmental control programs the emergence of distinct functional ensembles from unconstrained chemical reactions. *Proc. Natl. Acad. Sci.* **116**, 5387–5392 (2019).
12. Ricardo, A., Carrigan, M. A., Olcott, A. N. & Benner, S. A. Borate Minerals Stabilize Ribose. *Science* **303**, 196–196 (2004).
13. 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).
14. Self-assembling life. *Nat. Nanotechnol.* **11**, 909–909 (2016).
15. Semenov, S. N. *et al.* Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature* **537**, 656–660 (2016).
16. 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.
17. 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).
18. Huskey, W. P. & Epstein, I. R. Autocatalysis and apparent bistability in the formose reaction. *J. Am. Chem. Soc.* **111**, 3157–3163 (1989).
19. Weiss, A. Homogeneously catalyzed formaldehyde condensation to carbohydrates. *J. Catal.* **16**, 332–347 (1970).
20. Weiss, A. H., Socha, R. F., Likholobov, V. A. & Sakharov, M. M. Formose sugars from formaldehyde. *Appl. Catal.* **1**, 237–246 (1981).
21. Kopetzki, D. & Antonietti, M. Hydrothermal formose reaction. *New J. Chem.* **35**, 1787 (2011).
22. 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).
23. Socha, R. Homogeneously catalyzed condensation of formaldehyde to carbohydrates VII. An overall formose reaction model. *J. Catal.* **67**, 207–217 (1981).
24. Zafar Iqbal & Senad Novalin. The Formose Reaction: A Tool to Produce Synthetic Carbohydrates Within a Regenerative Life Support System. *Curr. Org. Chem.* **16**, 769–788 (2012).

25. De Bruijn, J. M., Kieboom, A. P. G. & Bekkiun, H. V. Alkaline Degradation of Monosaccharides VI [1]: The Fhucto-Fobmose Reaction of Mixtures of D-Fructose and Formaldehyde. *J. Carbohydr. Chem.* **5**, 561–569 (1986).
26. Weiss, A. Homogeneously catalyzed formaldehyde condensation to carbohydrates III. Concentration instabilities, nature of the catalyst, and mechanisms. *J. Catal.* **32**, 216–229 (1974).
27. *Water Chemicals Codex*. 159 (National Academies Press, 1982). doi:10.17226/159.
28. Arkin, A., Shen, P. & Ross, J. A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements. *Science* **277**, 1275–1279 (1997).
29. Roszak, R., Bajczyk, M. D., Gajewska, E. P., Hołyst, R. & Grzybowski, B. A. Propagation of Oscillating Chemical Signals through Reaction Networks. *Angew. Chem.* **131**, 4568–4573 (2019).
30. Urmès, C. *et al.* Periodic reactor operation for parameter estimation in catalytic heterogeneous kinetics. Case study for ethylene adsorption on Ni/Al2O3. *Chem. Eng. Sci.* **214**, 114544 (2020).
31. Ritson, D. J., Battilocchio, C., Ley, S. V. & Sutherland, J. D. Mimicking the surface and prebiotic chemistry of early Earth using flow chemistry. *Nat. Commun.* **9**, 1821 (2018).
32. Wu, L.-F. & Sutherland, J. D. Provisioning the origin and early evolution of life. *Emerg. Top. Life Sci.* **3**, 459–468 (2019).
33. 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).
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. Dolan, J. W. When Should an Internal Standard be Used? *LCGC N. Am.* **30**, 474–480.
36. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
37. 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).
38. Laine, R. A. & Sweeley, C. C. O-methyl oximes of sugars. Analysis as O-trimethylsilyl derivatives by gas-liquid chromatography and mass spectrometry. *Carbohydr. Res.* **27**, 199–213 (1973).
39. Fujino, K., Kobayashi, J. & Higuchi, I. Complex Formation from Calcium Hydroxide and Carbohydrate in Alkaline Solutions. *NIPPON KAGAKU KAISHI* 2287–2292 (1972).

Chapter 3

Strategies for reaction pathway reconstruction with perturbations in the reactor input



The identified compounds from the reactor output are connected by a hidden myriad of reactions and short-lived intermediates. From the reactor output at a single timepoint there is little information on what pathways have formed. Multiple routes, built up from a set of recursive reaction types, can lead from the feedstock molecules to the carbohydrates produced in the network.

In this chapter, I will elaborate on two strategies for identifying underlying connectivity between the identified compounds. The reaction network was probed with temporal signatures in the input concentration profiles. Compound concentrations couple to the input *via* cascading reaction pathways. The first strategy is based on sinusoidal oscillations in concentration of input sugar. Since all compounds in the network originate from this compound, the amplitude decay was used to obtain an estimate for the distance in reaction pathway from the origin. The second pathway reconstruction utilized pseudo-random fluctuations in $Ca(OH)_2$ concentration in the input. Pathways were constructed by grouping compounds with a similar temporal signature together and finding the shortest pathway between them.

Parts of this chapter have been published in:

1. W.E. Robinson, E. Daines, **P. van Duppen**, T. de Jong, W.T.S. Huck, *Nat. Chem.*, **14**, 623-631 (2022).

2. **P. van Duppen**, E. Daines, W.E. Robinson, W.T.S. Huck, *J. Am. Chem. Soc.*, **145**, 7559-7568 (2023).

3.1. Introduction - A systems approach in prebiotic chemistry for reaction pathway reconstruction

In the field of prebiotic chemistry, strategies for construction of reaction pathways or functional systems often rely on a bottom-up approach.[1–4] Experiments are typically performed under abiotic conditions in the absence of enzymatic catalysis to mimic metabolic reaction routes.[1,5] With this strategy, individually performed reactions are then stitched together to recreate the desired pathway.[1,2,6–13] On a prebiotic earth, however, it is questionable how likely a scenario of spatiotemporal separation of individual reactions is. Natural reaction pathways, whether part of a living system or not, are embedded in an environment with certain physicochemical properties.[14,15] There are no obvious barriers for molecules to react with one another in every possible way. Under prebiotic conditions the total set of chemical reactions is defined by the interplay between environment and the different chemical species present. Thus, systems left to their own devices could potentially generate very complex mixtures. Although conventional chemical analytical techniques can capture the composition of different chemical species in the system,[14,16–18] the chemical conversions that yielded the different species are not necessarily obvious. To map out and study reaction pathways in a system demands a top-down approach. Therefore, a shift in experimental design in prebiotic chemistry is required to reveal these reaction pathways.

Prebiotic chemistry has already dealt with pathway reconstruction for chemical reaction networks. Studies of prebiotic chemistry to reconstruct carbon core pathways have deduced relevant reaction pathways from more complex reaction systems.[14,16] For example, Moran and coworkers have observed a small reaction pathway analogous to the rTCA cycle.[16] The reaction progress was monitored using time-resolved GC-MS data and reaction pathways were followed by ^{13}C-labeled pyruvate.[16] Also, studying reactions in isolation is a popular approach for top-down network reconstruction in prebiotic systems. Although specific addition reactions[19–21], keto-enol tautomerizations[22] or cleavage reactions[23] have been studied for formose pathway reconstruction. The combinatorial complexity of the formose reaction impedes a similar approach to Moran's group beyond the early stages of the reaction.

We aim to reconstruct the formose reaction network to capture the most essential characteristics of the reaction system, similar to reaction pathway reconstruction in biochemical systems or electrical engineering.[24–26] To guide the reaction pathway reconstruction, a chemical search space is defined from a small set of recursive reaction types as discussed under 1.4.1. In this chapter, I will discuss

how we searched for reaction pathways within a 'global' reaction network. The 'global' reaction network is generated *in silico*, as discussed under 3.2. In section 3.3 and 3.4, I will discuss different experimental strategies to probe the formose reaction network and how these were used in a search strategy to prune the 'global' reaction network and obtain the respective experimental reaction network.

3.2 Generating a 'global' formose reaction network

The pathway search is caried out within a 'global' reaction network space to rationalize the experimental outcome. The 'global' reaction network was generated *in silico*, using the rule-based pathway reaction network generation strategy. A predefined set of reaction rules was iteratively applied on the network compounds, starting from a set of initial compounds (formaldehyde (**1**), glycolaldehyde, dihydroxyacetone (**2**), H_2O and OH^-).[27–29] The predefined set of reaction rules is represented by transformations of molecular substructures (see 3.7.2). *Via* rules specified with SMILES (Simplified Molecular Input Line Entry Specification), these transformations are then converted into coding language (e.g. Python) with strings of so-called SMARTS (SMILES Arbitrary Target Specification).[30] The Python code was written by Dr. W.E. Robinson and is available on GitHub (see 3.6.1). Starting from the set of initial compounds, after each round of iteration the generated products were filtered before feeding into the next round. In this step, each compound was assessed for chemical correctness, and also compounds larger than C_6 carbon number were removed. The resulting reaction network was converted into a bipartite graph, with separate compound and reaction nodes. The 'global' network serves as a 'map' to assist in pathway identification for both the signal transduction and correlation strategy (respectively discussed under 3.3 and 3.4).

The pathway search can be explained *via* a model network graph, see figure 3.1a, where the nodes represent compounds and the edges represent chemical conversions between compounds. Before the *in silico* network search for an experiment, the 'global' network was pruned and all nodes from non-detected carbonyl and polyol compounds were removed. The dominant pathways were inferred with agnostic search approaches. For example, connecting the observed compounds *via* the shortest pathway from the input sugar provides a useful initial search strategy (fig. 3.1b). However, guided by the experimental outcomes, new search strategies will be proposed in this chapter for more accurate and complete approximation of the reaction pathways that are actually in operation (see 3.3 and 3.4).

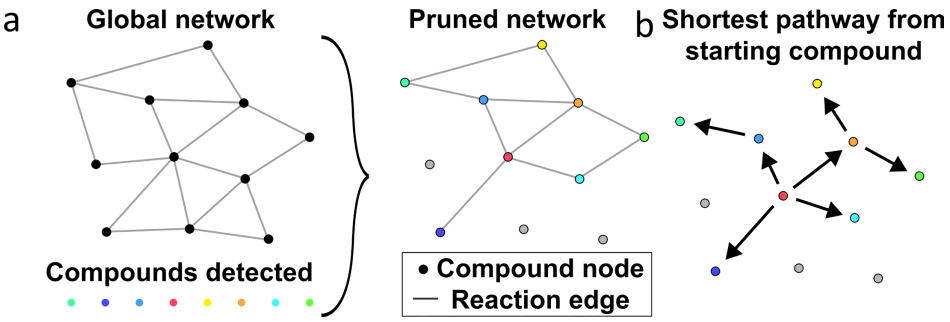


Figure 3.1: A model reaction network, which will be used to explain pathway reconstruction from experimental data. a) Before the pathway search, the 'global' network is pruned to only contain experimentally produced compounds. b) The reaction network can be reproduced from the pruned network. For example, by connecting all the compound nodes *via* the shortest pathway from the feedstock molecule (pink node).

3.3 Reconstructing reaction pathways with sinusoidal oscillation of the input sugar

The pathway search in the 'global' reaction network can be guided by modulating the CSTR input. In this section I will discuss experiments which used sinusoidal oscillations of the input sugar. This can be utilized to probe the network, since all products in the formose mixture originate from this input. The signal transduction through the network holds key information about the connectivity between observed compounds.

3.3.1 Signal transduction from sinusoidal concentration variation of the input sugar through the reaction network

Propagation of sinusoidal input variation through a chain or network of chemical reactions has been explored both experimentally and in theoretical papers.[31–34] The rate of a chemical reaction is only sensitive to changes in the environment if it happens on a similar or slower time scale.[31,32] Therefore, a chemical reaction functions as a low-pass filter, and only allows sufficiently low frequencies to pass, see figure 3.2. The signal is transduced by the chemical reaction, if the rate of the reaction is fast compared to the signal frequency (e.g. $2\pi f/\text{rate} > 1$)(fig. 3.2a).[31,32] The signal amplitude is dampened for a slower rate of the reaction ($2\pi f/\text{rate} < 1$)(fig. 3.2b). Reversible reactions allow the signal to pass in a similar fashion.

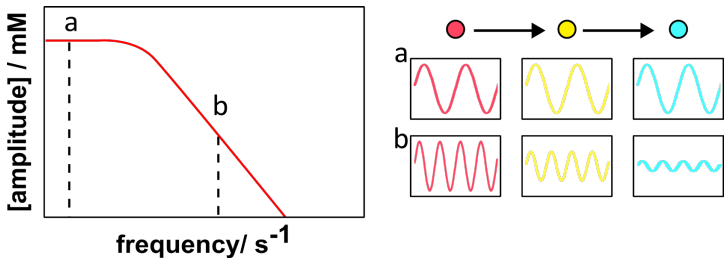


Figure 3.2: Chemical reactions as low-pass filter for signal transduction in a chain of reactions. A chemical reaction allows low frequencies (a) to pass, e.g. $2\pi f/\text{rate} < 1$. Signal amplitude is dampened towards high frequencies (b), e.g. $2\pi f/\text{rate} > 1$.

For more complicated systems of chemical reactions, comprising of bimolecular reactions, band-pass filter behavior is also possible and attenuation occurs for low frequency input signals, see figure 3.3. A middle range of frequencies, between low and high, pass optimally through the system. Different types of chemical systems, described by rate equations with nonlinear terms, have been shown to exhibit band-pass frequency filtering.[31]

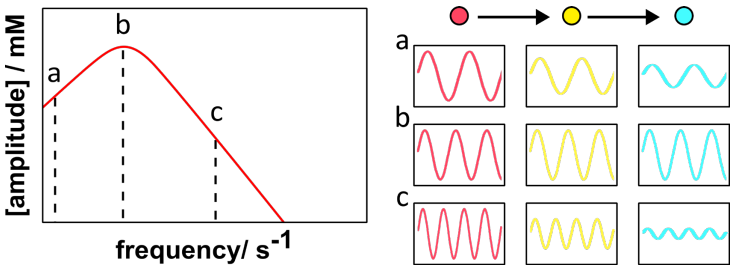


Figure 3.3: Chemical reactions as band-pass filter for signal transduction in a chain of reactions. The system attenuates the signal at low and high frequency (a,c) and allows the signal to pass optimally in a middle frequency range (b).

Chemical reaction systems attenuate input signals in the high frequency domain ($2\pi f/\text{rate} > 1$), for both the low-pass and band-pass filter. This property can be exploited to gain mechanistic insight in the reaction network structure. In a linear reaction chain, the amplitude can decay (but not increase) in each reaction step, cascading away from the signal source. The amplitude of a species decays as the pathway length increases away from the signal source (fig. 3.2b). In the formose reaction network, all observed species in the network are produced from the input sugar. To probe this network, a sinusoidal concentration profile for the input sugar (e.g. dihydroxyacetone) is applied. The amplitude decay in the time traces of detected compounds is used to find reaction pathways in the 'global' formose reaction network.

3.3.2 Reconstructing the network with concentration amplitude decay

The pathway search in the 'global' formose reaction network relies on concentration series of detected compounds. For each modulated experiment, the sine wave was sampled in triplicate and with > 5 equally spaced samples per cycle. The amplitude for each of the detected compounds was obtained from a Fourier transform (see 3.6.2). The compounds were ordered from high to low amplitude, starting from the initiator feedstock.

First, the compounds were listed in order of decreasing amplitude, see figure 3.4a. The 'global' reaction network was pruned to contain experimentally detected compounds only (fig. 3.1a and 3.4b). The pathway search was performed with the networkX module for Python, with the shortest pathway search (see 3.6.2).[35] For each detected compound with an amplitude above the threshold, a set of shortest reaction pathways was reconstructed towards compounds with a lower amplitude. These pathways were summarized as a directed graph between a source node (high amplitude) towards a target node (low amplitude), since the amplitude decay has a downstream direction (fig. 3.4c). For each detected compound a directional subnetwork was reconstructed. These reaction pathways were superimposed to get a directed graph to represent the experimental reaction network (fig. 3.4d).

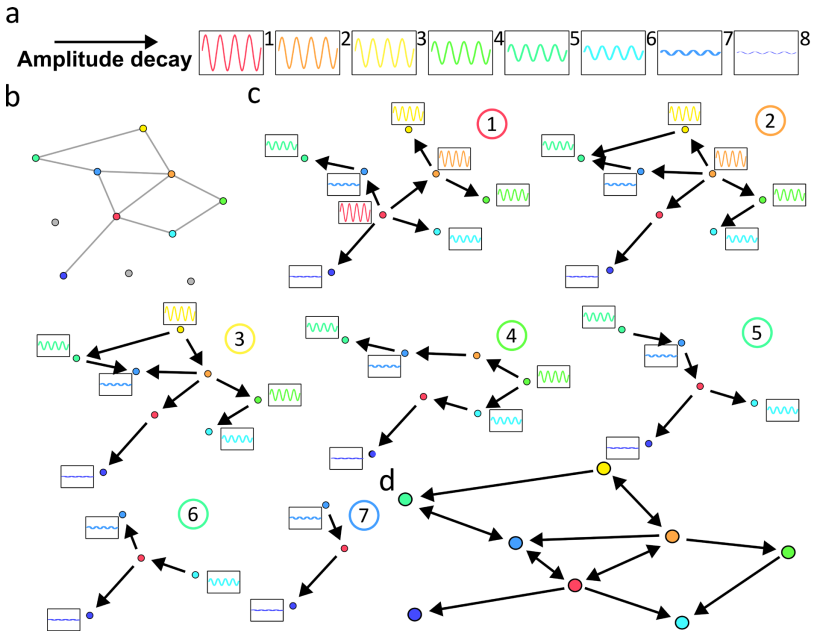


Figure 3.4: Schematic representation of pathway reconstruction with amplitude decay in concentration profiles for compounds in the formose reaction. a) Compound time traces sorted by amplitude decay. b) The pruned 'global' network to only contain

experimentally produced compounds. c) Pathway reconstruction performed from each compound in the amplitude decay towards compounds with smaller amplitude. d) A superimposed reaction network from all reconstructed reaction pathways.

Figure 3.5 shows an example of a reaction network reconstruction using the amplitude decay from real data (reaction conditions: $[formaldehyde]_{in}$ = 200 mM, $[NaOH]_{in}$ = 30 mM, $[CaCl_2]_{in}$ = 15 mM and probed by modulating input concentration of the initiator $[erythrulose]_{in}$ = 50 mM ± 25 mM with a 300 second period). The time traces of the observed compounds were sorted from high to low amplitude (fig. 3.5a). The observed reaction species were used to prune the 'global' reaction network only to include reactions involving compound **3** to **17** and their enolates. The first set of shortest pathways was assembled from the sugar initiator erythrulose (**3**), which was listed at the top of the amplitude decay, to all other compounds **4 – 17** (fig 3.5b). Subsequently, from each of the other observed compounds, a set of reaction pathways was reconstructed towards all compounds with lower amplitude. For example, a set of reaction pathways was found to start from compound **6** towards downstream compounds **7 - 17** (fig. 3.5c). After all sets of reaction pathways were reconstructed, these were superimposed to obtain the reconstructed reaction network, see figure 3.6.

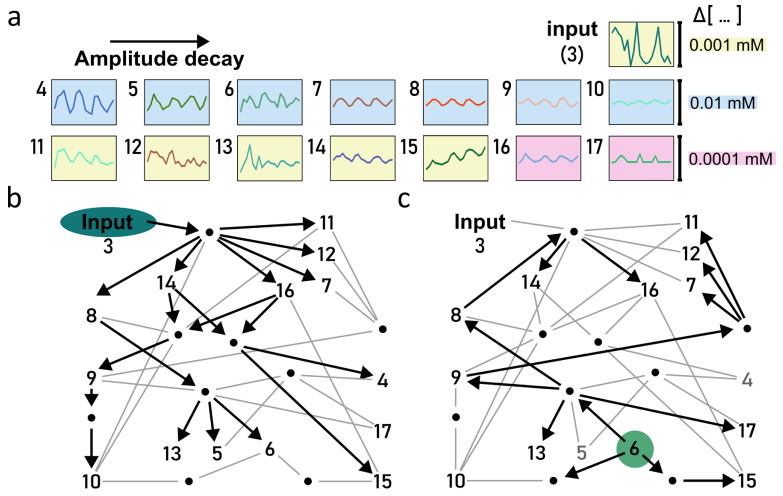


Figure 3.5: The concentration series for reaction pathway reconstruction with the amplitude decay strategy. a) Series of amplitude decay of the observed compounds, starting from source compound **3**. b) The shortest pathways connecting the input (**3**) to all downstream compounds *via* the shortest pathway. c) The network produced **6** connected to all downstream compounds *via* the shortest pathway. Reaction conditions (code: FRN087D): $[formaldehyde]_{in}$ = 200 mM, $[erythrulose]_{in}$ = 50 mM (amplitude: 25 mM, period: 300 seconds), $[NaOH]_{in}$ = 30 mM, $[CaCl_2]_{in}$ = 15 mM, residence time = 120 seconds.

In the network, aldol addition reactions with formaldehyde were deemed irreversible. Most bimolecular reactions were, however, found to be reversible in this experiment (fig. 3.6). For example, psicose (**6**) did undergo four out of five bimolecular reactions reversibly.

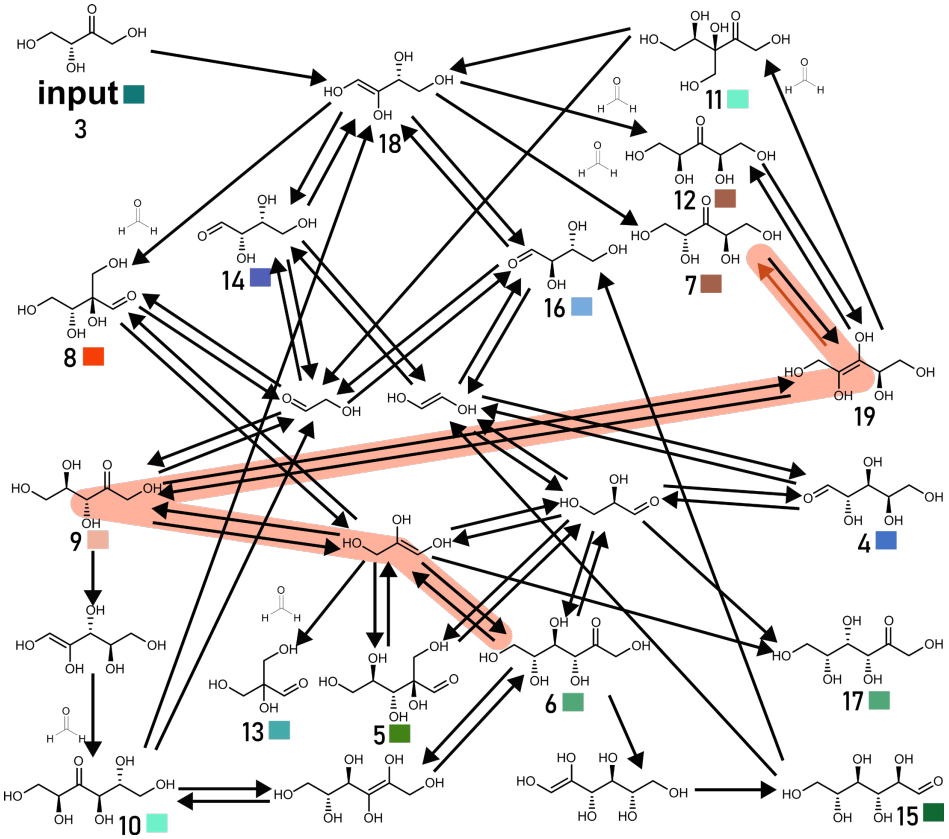


Figure 3.6: Reconstructed reaction network with amplitude decay from real data. The orange marked reaction pathway illustrates the shortest pathway to connect the high amplitude **6** to the lower amplitude **7**. Reaction conditions (code: FRN087D): $[formaldehyde]_{in}$ = 200 mM, $[erythrulose]_{in}$ = 50 mM (amplitude: 25 mM, period: 300 seconds), $[NaOH]_{in}$ = 30 mM, $[CaCl_2]_{in}$ = 15 mM, residence time = 120 seconds. The reaction network hosts two types of reversible reactions: keto-enol tautomerization and aldol addition/cleavage reactions. The amplitude guided pathway search determined whether a reaction in the reconstructed network was reversible. The shortest pathway from **6** to **7** (marked in orange) was used to find the reversible reaction from **19** to **7** (red arrow).

The keto-enol tautomerization reaction with 3-ketopentose sugar (**7**) illustrated how both reversible reactions were included from the search strategy. **7** was

listed as fifth compound in the amplitude series (fig. 3.5a and 3.7.1: fig. S3.9). From **7**, an isomerization reaction functioned as an outgoing reaction *via* a C_5-2,3-enolate towards target compounds (**8** – **17**) with a lower amplitude. However, **7** was produced in two reactions. From the enolate of **3** (C_4-1,2-enolate, **18**) it was produced in an aldol addition reaction with formaldehyde. It was produced also in the reverse isomerization reaction from its C_5-2,3-enolate (**19**) neighbor (red arrow in fig. 3.6). The shortest pathway ending at this reaction roots from a higher amplitude sugar in the network. From two C_6 source sugars (**5**, **6**), different pathways lead back to **7** as target. From the high amplitude psicose (**6**), the pathways *via* **18** to **7** is comprised of five reaction steps. An alternative pathway *via* **9** and **19** included only four reaction steps and was therefore chosen as shortest pathway (marked in orange). The reverse isomerization reaction from **19** to **7** (red arrow, fig 3.6), as well as many other reverse reactions, was introduced by a source compound produced in the network.

Remarkably, in this network, input **3** was only converted to its C_4-1,2-enolate (**18**), 1,3,4-trihydroxybut-1-en-2-olate. It was ranked as first compound for the pathway search since it was only a source compound for the pathway search towards downstream target molecules (fig. 3.5a). Hence, the search algorithm did not include the reverse reaction and also it was not consumed nor produced in the network in an addition or cleavage reaction.

This experimental framework provides a tool to access the underlying reaction network structure. Of course, the retrieved network structure is constrained by a set of predefined reaction rules and the specific search strategy. Therefore, the reconstructed reaction network forms a hypothesis for the mechanism that is most likely to be in operation to generate the detected compounds. It becomes especially powerful in large datasets, for example, in a cumulative analysis of different reactions present over a large scope of experimental conditions, as explored in Chapter 4.

3.4 Finding pathways in reaction networks perturbed with $Ca(OH)_2$ input

The formose reaction network can also be probed by modulation of the effective concentration input of $Ca(OH)_2$. The input concentration profile of $Ca(OH)_2$ was constructed by sampling from a normal distribution around the steady-state input concentration. Different reaction types interacted with catalytic species Ca^{2+} and OH^- (see 1.4.1). In this section, I will discuss how the signal transfer from $[Ca(OH)_2]_{in}$ to different compounds was interpreted using a cluster analysis to reconstruct the underlying reaction pathways.

3.4.1 Cluster analysis to find groups of compounds with similar response

The modulated $Ca(OH)_2$ input in each of the explored set of experimental conditions leads to a unique response for each of the observed compounds. A hierarchical clustering analysis was performed on each perturbed reaction condition, to identify different parts of the formose networks with a unique response to the fluctuations of Ca^{2+} and OH^-. For each experimental condition, a linkage matrix was calculated based on correlation distance for the detected concentration profiles (see 3.6.3). The resulting dendrogram revealed partitioning of different groups of compounds, see figure 3.7a. Within these groups, the concentration profiles share a high degree of similarity in their response to the input signal. Different groups in the dendrogram were marked in a dotted circle and were assigned a roman numeral (**I**, **II**). The behavior of the identified groups was governed by the interaction of the input signal and the chemistry in the underlying reaction network. The pathways in the reaction network were reconstructed by identifying the shortest pathway from feedstock molecules to the compounds in a group (fig. 3.7b).

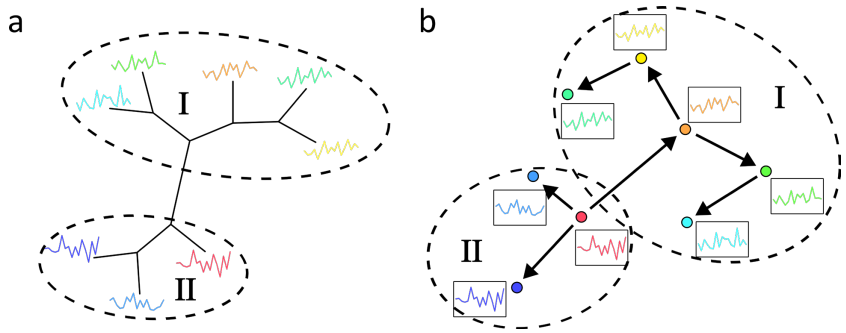


Figure 3.7: **Schematic clustering of compound time traces.** a) A dendrogram with the different time traces, where two groups of high similarity are identified (**I**, **II**). b) A shortest set of reaction pathways reconstructed from the input sugar to connect the compounds in the identified groups.

3.4.2 Network reconstruction based on identified clusters of compounds

For the experiments with modulated $[Ca(OH)_2]_{in}$, 50 samples were collected to obtain a statistical representation of the compound concentration distribution. The compound time traces were used to construct a dendrogram separating out the detected compounds (see 3.7.1: fig. S3.10a). From this dendrogram, compounds were assigned to a group marked with a roman numeral - **I**, **II**, **III** – (see 3.7.1: fig. S3.10b). The network reconstruction was started from initiator sugar **2**.

In the reconstructed network (fig. 3.8), **2** first isomerized to form its C_3-enolate. Subsequent addition reactions with either formaldehyde or glycolaldehyde led to the production of **3**, **13** and xylulose (**20**) in group **I**. In this group, **3** was involved in the production of **14/16, 7** and **21**, respectively *via* isomerization, addition with formaldehyde or a C_2-enolate. Finally, also 3-ketohexose **23** was produced in group **I**, from a formaldehyde addition reaction which involved the enolate of **20**.

The C_2-enolate was involved in the production of xylose (**24**), ribose (**25**) and 3-hydroxymethyl-aldotetrose (**26**) in group **II**. Further, **27** and **28** were produced in this group *via* formaldehyde addition reactions with the enolate of **26**. The C_6 compounds **5** and **17** were produced in an addition reaction of glyceraldehyde and a C_3-enolate, 2,3-dihydroxypropen-1-olate and 1,3-dihydroxypropen-2-olate respectively.

The carbohydrates in group **III** were all produced in addition reactions with a C_3-enolate. Compounds **29**, **30, 31, 32** were produced from an addition reaction with glyceraldehyde, whereas **33** and **8** were produced from an addition with glycolaldehyde.

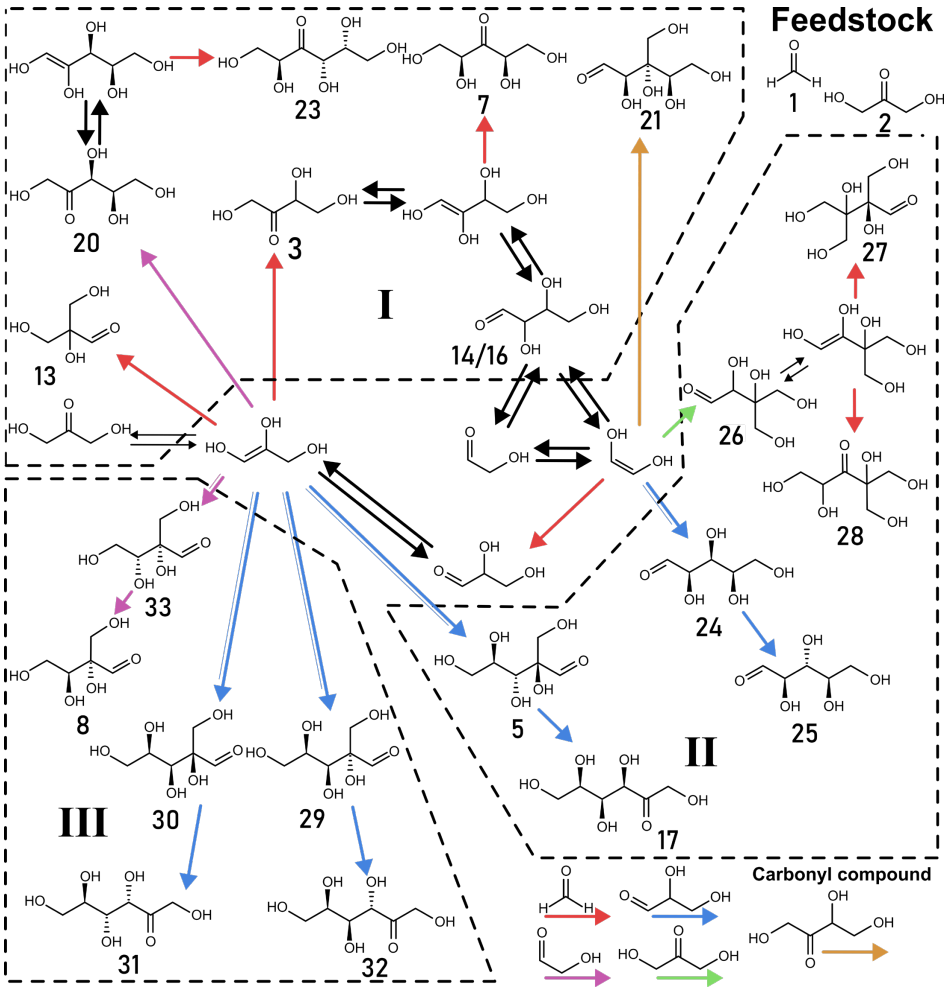


Figure 3.8: A reconstructed network of a $Ca(OH)_2$ modulated experiment (EXP003). The assigned compound groups (**I**, **II**, **III**) are based on a cluster analysis (see 3.7.1: fig. S3.10).

This strategy for pathway reconstruction was utilized to reconstruct networks in chapter 5. Together with the strategy as discussed under 3.3, the reconstructed reaction networks aid accurate compound assignment across all experiments (see diagram in 3.7.3).

3.5 Conclusion

The recursive reaction classes in the formose reaction produce a combinatorial explosion. A top-down approach is required to study the reaction networks that

are formed under different reaction conditions structure. By probing steady-state conditions in a flow reactor, a differential response of the observed compounds allowed for an estimation of the underlying reaction connectivity. Chemical reactions exhibit low-pass filter behavior, and are only sensitive to environmental changes at similar or slower time-scales than the reaction rate ($2\pi f/\text{rate} > 1$). This chemical property was used to probe the reaction network by modulating different input variables.

In this chapter, I have discussed two different approaches to reconstruct chemical reaction networks. For the first strategy, the input sugar is the root node in the reaction network from which all observed compounds can be derived. The respective input concentration is modulated sinusoidally to probe the reaction network. Amplitude decay in the observed compounds is a measure for the 'distance' in the reaction network with respect to the root compound. A 'global' reaction network created *in silico*, containing all reactions theoretically possible, serves as an overall 'map' to guide the pathway reconstruction. The observed compounds were sorted based on amplitude decay. With a 'shortest pathway algorithm', reaction pathways were obtained from high amplitude to lower amplitude compounds. A superposition of all retrieved reaction pathways within one experiment did yield the reconstructed reaction network.

In the second experimental approach, the formose reaction network was probed with a pseudo-random input flow profile of the effective $Ca(OH)_2$ input concentration. Different parts of the network with a collective response to the $Ca(OH)_2$ input modulation were identified with a hierarchical cluster analysis. The observed grouping was the result of the interaction between the input modulation with different parts of the reaction network. For each group of compounds in a cluster, a shortest pathway was reconstructed from feedstock molecules.

Additionally, the network reconstruction aids compound assignment, see scheme in 3.7.3. A compound in the network is assumed to be connected to at least one other compound in the network *via* one enolate intermediate. Hence, if the network reconstruction contains a non-calibrated compound without contiguous neighbor, it was reassigned to fulfill this criterium. If this compound has been assigned already in other experimental datasets, the new assignment was applied to previously reconstructed reaction networks.

3.6 Method summary

Python programs for the following described data analysis are available at https://github.com/huckgroup/formose-2021.git (authored by Dr. W.E. Robinson).

3.6.1 Generation of the formose reaction space *in silico*

A set of reaction pathways in line with the expected reaction types of the formose reaction is generated using The RDKit (RDKit: Open-source cheminformatics; http://www.rdkit.org, date of access: June 2021).[29] The reactions outlined in 3.7.2 are translated into reaction SMARTS which are iteratively applied to a seed set of compounds (glycolaldehyde, formaldehyde, hydroxide and water). Products of each reaction operation are fed into the next iteration. Compounds with a chain length of greater than 6 carbon atoms and the reactions leading to them are removed after every iteration. The resulting network corresponds to a hypothetical case of the formose reaction in which all pathways possible according to the constructing reaction rules are taken. This set of reactions is used as a framework for determining reaction pathways from data.

3.6.2 Reaction pathway search with amplitude decay

The generated list of reactions for the formose reaction is converted into a networkX DiGraph object.[36] Nodes corresponding to compounds are connected by directed edges to nodes for reactions. The edge direction indicates 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 are removed from the graph.

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

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

The construction of a reaction pathway starts 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 are then found between consecutive members of the amplitude-ordered compound list. The pathways are 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 is found *via* the shortest path to the compound from a set of those with higher amplitudes.

3.6.3 Calculation of distance matrix and cluster analysis

For the cluster analysis of each experimental condition, pairwise distances based on the 'correlation' metric are calculated between time traces using scipy.spatial.distance.pdist(x) (Equation 1) to create a distance matrix.

Eq. 1: $$distance = \frac{(u - \bar{u}) \cdot (v - \bar{v})}{||(u-\bar{u})||_2 \; ||(v-\bar{v})||_2}$$

Where u and v are vectors of the concentration time profiles for compounds, $\bar{u}$ and $\bar{v}$ are the means of u and v, respectively, and $x \cdot y$ is the dot product of x and y, and $||(v - \bar{v})||_2 = \sqrt{(v_1 - \bar{v})^2 + \cdots + (v_n - \bar{v})^2}$.

A hierarchical cluster analysis is performed on the *distance matrix*, by creating a *linkage matrix* with scipy.cluster.hierarchy.linkage(*distance matrix*), using the '*average*' method. At each step of the clustering the nearest two clusters A and B are combined in a higher-level cluster X. The height of X in the dendrogram is calculated as $\delta(_{A,X}) = \delta_{(B,X)} = d_{(A,B)}/2$. The new distance matrix is calculated, where distances between newly formed cluster $A \cup B$ and cluster Y is calculated with Equation 2.

Eq. 2: $$d_{(A \cup B),Y} = \frac{|A| \cdot d_{A,Y} + |B| \cdot d_{B,Y}}{|A| \cdot |B|}$$

Where, $|A|$ and $|B|$ are the number of elements in A and B. Thus, the newly calculated distance is a proportional average between distance $d_{A,X}$ and $d_{B,X}$. Clusters are assigned *via* inspection of figure S3.10 (in 3.7) with reference to the underlying chemistry.

3.7 Supplementary information

3.7.1 Supplementary figures

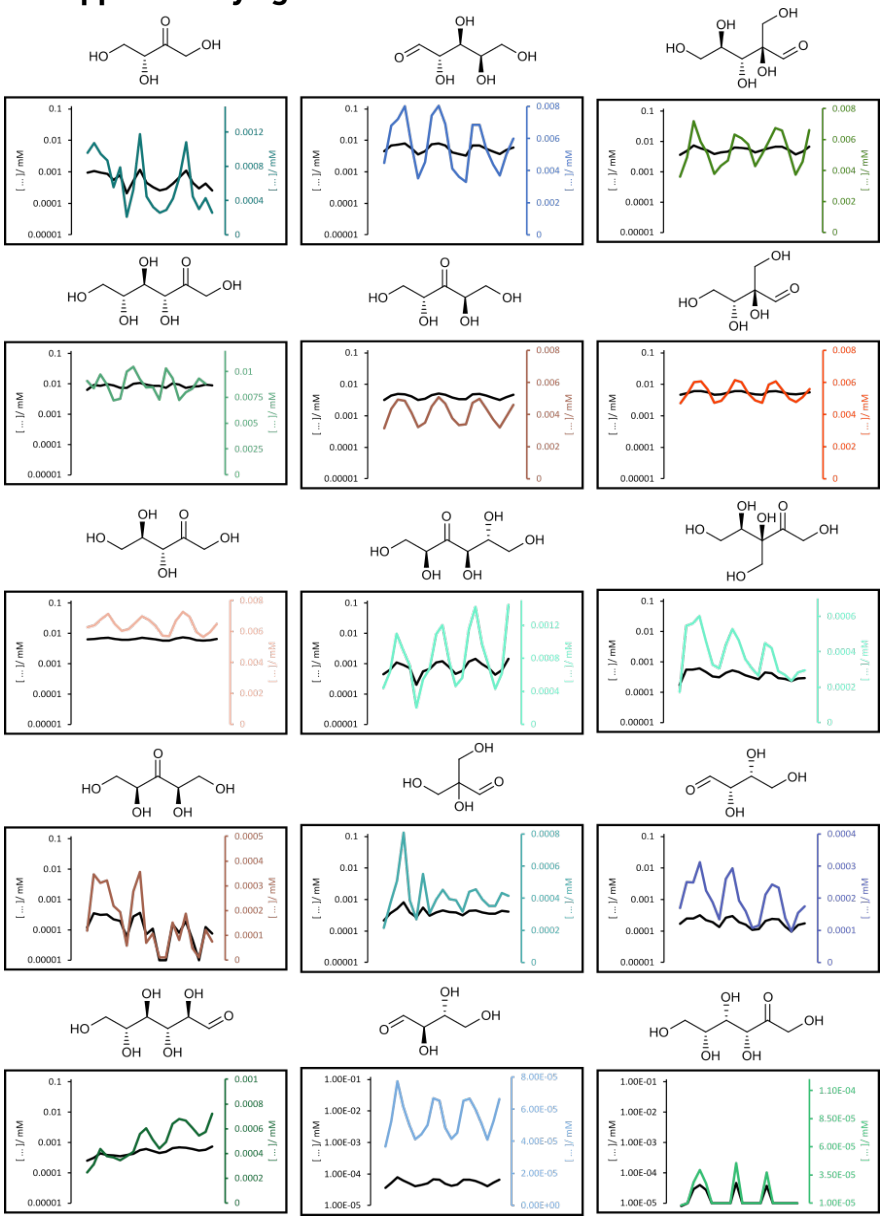


Figure S3.9: Amplitude decay for a series of detected compounds, from which reaction network in figure 3.6 was reconstructed (FRN087D). Erythrulose was used as initiator sugar, hence it was listed first.

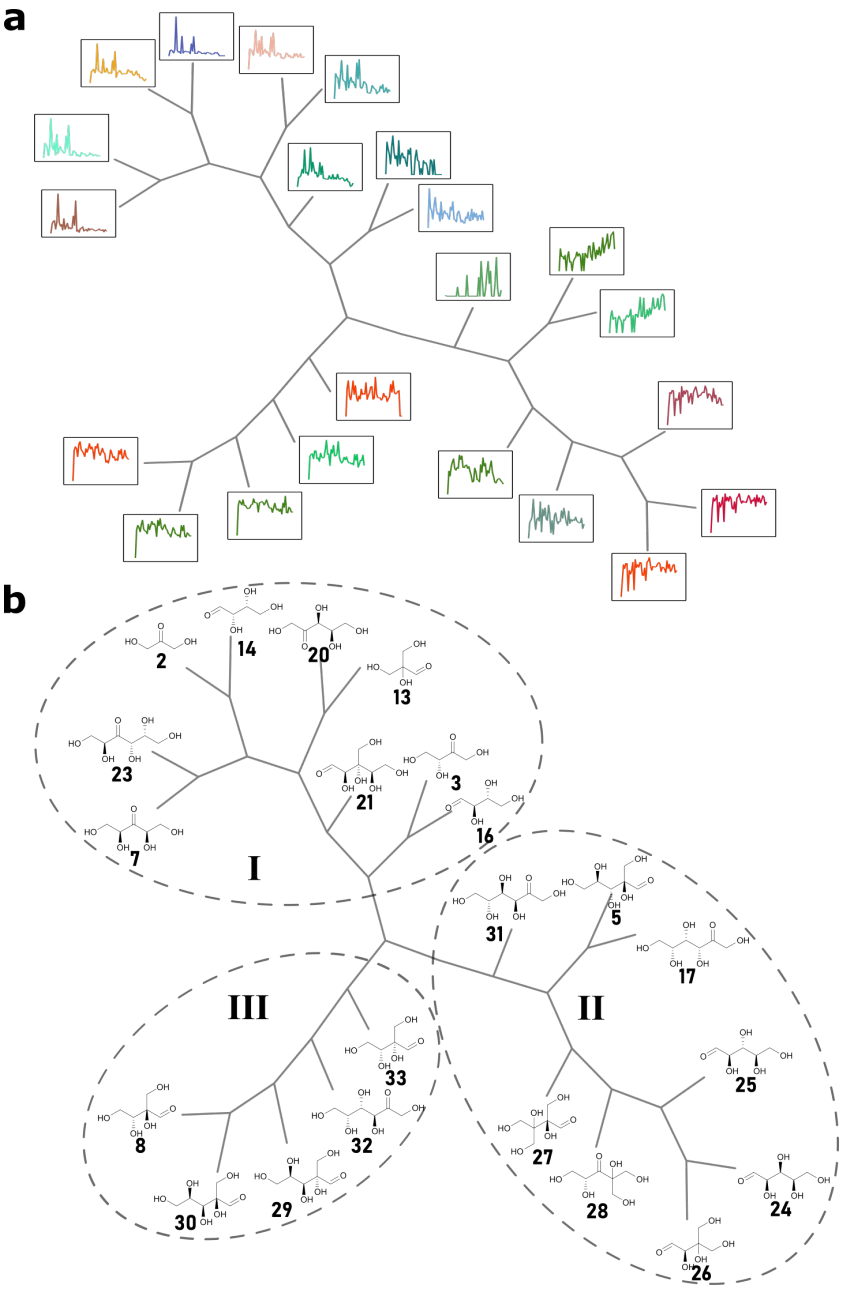


Figure S3.10: Cluster analysis real data (EXP003), which is used for network reconstruction in figure 3.8. Dendrogram for the cluster analysis on time traces (a) of detected compounds. The corresponding compounds are grouped (b) according to partitioning of the cluster analysis, the different groups are assigned a roman numeral.

3.7.2 Reaction classes in the formose reaction

The 'global' formose reaction network is constructed from a set of reaction rules (fig. S3.11). This network is generated by recursively applying these reaction rules on a set of starting compounds (formaldehyde, glycolaldehyde, dihydroxyacetone, H_2O and OH^-).

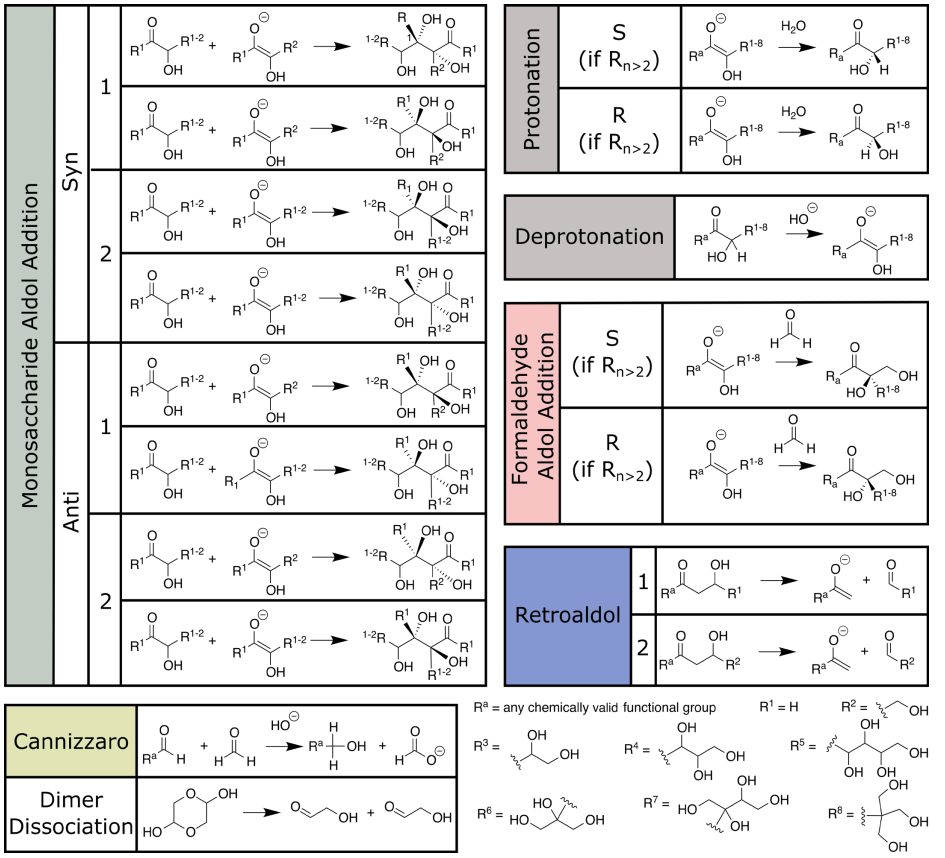


Figure S3.11: Reaction rules for pathway search. These reactions are translated to SMARTS and iteratively applied to a set of seed compounds: formaldehyde, glycolaldehyde, dihydroxyacetone, water and hydroxide. Figure prepared by Dr. W.E. Robinson.

3.7.3 Scheme for network aided compound assignment

The compound assignment is reassessed for non-calibrated compounds, if it did not have any contiguous neighbour in the network reconstruction, see figure S3.12.

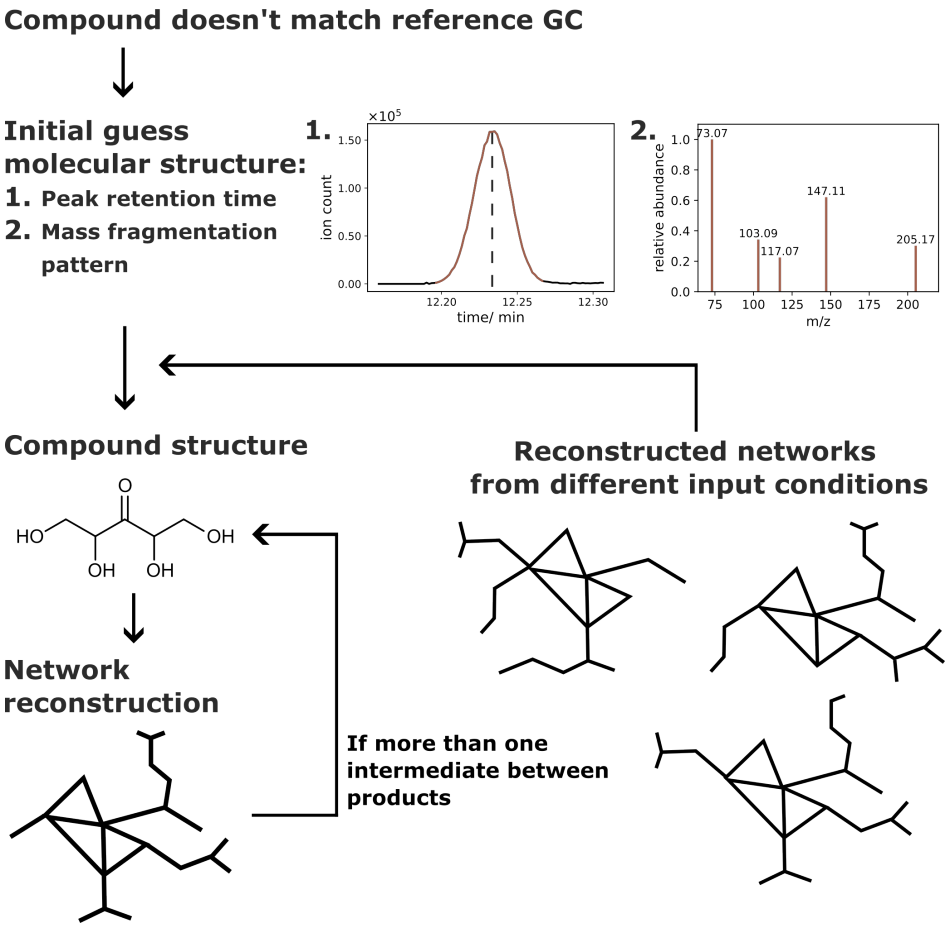


Figure S3.12: Scheme for the assignment of unknown compounds, aided by reaction network reconstruction. To find the molecular structure of an unknown compound the GC retention time and mass spectrum is combined with knowledge from previous experiments. Network reconstruction is used to assess the assignment.

3.8 References

1. Islam, S. & Powner, M. W. Prebiotic Systems Chemistry: Complexity Overcoming Clutter. *Chem* **2**, 470–501 (2017).
2. 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).
3. 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).
4. Semenov, S. N. *et al.* Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature* **537**, 656–660 (2016).

5. Meléndez-Hevia, E., Montero-Gómez, N. & Montero, F. From prebiotic chemistry to cellular metabolism—The chemical evolution of metabolism before Darwinian natural selection. *J. Theor. Biol.* **252**, 505–519 (2008).
6. Powner, M. W., Gerland, B. & Sutherland, J. D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459**, 239–242 (2009).
7. Mullen, L. B. & Sutherland, J. D. Simultaneous Nucleotide Activation and Synthesis of Amino Acid Amides by a Potentially Prebiotic Multi-Component Reaction. *Angew. Chem. Int. Ed.* **46**, 8063–8066 (2007).
8. Ritson, D. & Sutherland, J. D. Prebiotic synthesis of simple sugars by photoredox systems chemistry. *Nat. Chem.* **4**, 895–899 (2012).
9. Ritson, D. J. & Sutherland, J. D. Synthesis of Aldehydic Ribonucleotide and Amino Acid Precursors by Photoredox Chemistry. *Angew. Chem. Int. Ed.* **52**, 5845–5847 (2013).
10. Coggins, A. J. & Powner, M. W. Prebiotic synthesis of phosphoenol pyruvate by α-phosphorylation-controlled triose glycolysis. *Nat. Chem.* **9**, 310–317 (2017).
11. 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).
12. Xu, J. *et al.* A prebiotically plausible synthesis of pyrimidine β-ribonucleosides and their phosphate derivatives involving photoanomerization. *Nat. Chem.* **9**, 303–309 (2017).
13. Xu, J. *et al.* Selective prebiotic formation of RNA pyrimidine and DNA purine nucleosides. *Nature* **582**, 60–66 (2020).
14. Keller, M. A. *et al.* Conditional iron and pH-dependent activity of a non-enzymatic glycolysis and pentose phosphate pathway. *Sci. Adv.* **2**, e1501235 (2016).
15. Sasselov, D. D., Grotzinger, J. P. & Sutherland, J. D. The origin of life as a planetary phenomenon. *Sci. Adv.* **6**, eaax3419 (2020).
16. Muchowska, K. B., Varma, S. J. & Moran, J. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019).
17. Hold, C., Billerbeck, S. & Panke, S. Forward design of a complex enzyme cascade reaction. *Nat. Commun.* **7**, 12971 (2016).
18. 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).
19. Gutsche, C. D. *et al.* Base-catalyzed triose condensations. *J. Am. Chem. Soc.* **89**, 1235–1245 (1967).
20. Kim, H.-J. *et al.* Synthesis of Carbohydrates in Mineral-Guided Prebiotic Cycles. *J. Am. Chem. Soc.* **133**, 9457–9468 (2011).
21. 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).
22. 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).

23. Guthrie, J. P. The Aldol Condensation of Acetaldehyde: the Equilibrium Constant for the Reaction and the Rate Constant for the Hydroxide Catalyzed RetroAldol Reaction. *Can. J. Chem.* **52**, 2037–2040 (1974).
24. Milo, R. *et al.* Network Motifs: Simple Building Blocks of Complex Networks. *Science* **298**, 824–827 (2002).
25. Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461 (2007).
26. Oster, G. F. & Desoer, C. A. Tellegen's theorem and thermodynamic inequalities. *J. Theor. Biol.* **32**, 219–241 (1971).
27. 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).
28. 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).
29. Simm, G. N. & Reiher, M. Context-Driven Exploration of Complex Chemical Reaction Networks. *J. Chem. Theory Comput.* **13**, 6108–6119 (2017).
30. Landrum, G. *et al.* RDKit: Open-source cheminformatics. (2021) doi: https://doi.org/10.5281/zenodo.591637.
31. Samoilov, M., Arkin, A. & Ross, J. Signal Processing by Simple Chemical Systems. *J. Phys. Chem. A* **106**, 10205–10221 (2002).
32. Roszak, R., Bajczyk, M. D., Gajewska, E. P., Hołyst, R. & Grzybowski, B. A. Propagation of Oscillating Chemical Signals through Reaction Networks. *Angew. Chem.* **131**, 4568–4573 (2019).
33. Urmès, C. *et al.* Periodic reactor operation for parameter estimation in catalytic heterogeneous kinetics. Case study for ethylene adsorption on Ni/Al2O3. *Chem. Eng. Sci.* **214**, 114544 (2020).
34. Mettetal, J. T., Muzzey, D., Gómez-Uribe, C. & Van Oudenaarden, A. The Frequency Dependence of Osmo-Adaptation in *Saccharomyces cerevisiae*. *Science* **319**, 482–484 (2008).
35. Robinson, W. E., Daines, E., van Duppen, P., de Jong, T. & Huck, W. T. S. Environmental conditions drive self-organization of reaction pathways in a prebiotic reaction network. *Nat. Chem.* **14**, 623–631 (2022).
36. Hagberg, A. A., Schult, D. A. & Swart, D. J. Exploring network structure, dynamics, and function using NetworkX. *Proc 7th Python Sci. Conf. SciPy 2008*, 11–15 (2008).

Chapter 4

Environmental conditions drive self-organization of prebiotic reaction networks

Different prebiotic chemical synthetic routes towards biomolecules have been established. As of yet, it is still unclear how elaborate reaction systems were shaped to evolve towards life. In the absence of genetic and enzymatic control, only the environmental driving forces remain.

In this chapter, I will discuss how the formose reaction self-organizes in response to different physicochemical environments. Variation in input concentration of feedstock or catalyst, residence time and temperature provide environmental control parameters for the reaction system. Remarkably, distinct compositional patterns emerge from changes in these environmental conditions. The formose reaction is probed with temporal modulation of the input sugar to reconstruct the underlying reaction pathways. The observed shifts in compositional outcome are the result of rewiring of the reaction network. The environment controls the expression of different reaction classes which constitute the formose reaction. Interestingly, the Breslow cycle not only provides the system with autocatalytic properties. It is also an important source of C_2 and C_3 building blocks for the system. These network generated building blocks feed a chain growth mechanism alongside the formaldehyde chain growth.

The interactions between the environment and inherent chemical reactivity control prebiotic network organization. This insight provides a potential mechanism for the process of chemical evolution towards the origin of life.

Parts of this chapter have been published in:

1. W.E. Robinson, E. Daines, **P. van Duppen**, T. de Jong, W.T.S. Huck, *Nat. Chem.*, **14**, 623-631 (2022).

4.1 Introduction - Environmental conditions and prebiotic model reaction networks

The origin of life required extensive development to ever more complex chemical systems. On a prebiotic earth, only the physicochemical properties of the environment function as driving force, interacting with the inherent chemical reactivity.[1] I have already discussed in chapter 1, how prebiotic reaction conditions allow for a wide range of chemical reactions. These can be utilized for the production of a diverse range of structurally complex organic molecules.[2–8] By combining simple metabolites under prebiotic conditions, chemical reaction networks are formed that resemble the carbon core metabolic pathways.[9–13] For a protometabolic system to emerge from a mixture of prebiotic feedstock molecules, chemical reactivity alone is not sufficient to dictate the formation of one pathway over another.[14–16] It is difficult to conceive how to prevent uncoordinated growth of disorderly reaction pathways. However, prebiotic reactions were embedded in their environments, thus presenting a potential directing force for emerging systems.[1,17] The role of the environment on the self-organization of prebiotic chemical reaction networks remains poorly understood and explored in experimental systems. I will discuss how the interaction between out-of-equilibrium reaction networks and the environment provide fundamental insights for the understanding of chemical evolution towards life.

In this chapter, I will elaborate how a simple set of recursive reactions in the formose reaction forms well-defined compositional patterns. In a series of flow reactions, the formose reaction was employed as a model prebiotic reaction network. Experiments were performed to measure steady-state equilibrium compositions of the formose reaction. I will discuss how the environmental variables control the compositional outcome of the network. In addition, the networks were probed by modulating the input concentrations of initiating sugars sinusoidally, as discussed under 3.3.2, see figure 4.1. The transfer of input modulation to product compounds was utilized to reconstruct the underlying reaction pathways of the formose reaction.[18–21] I will explain how the observed compositional trends correspond to underlying reaction pathways. Furthermore, I will discuss how the environment controlled the expression of different types of reactions in the formose reaction.

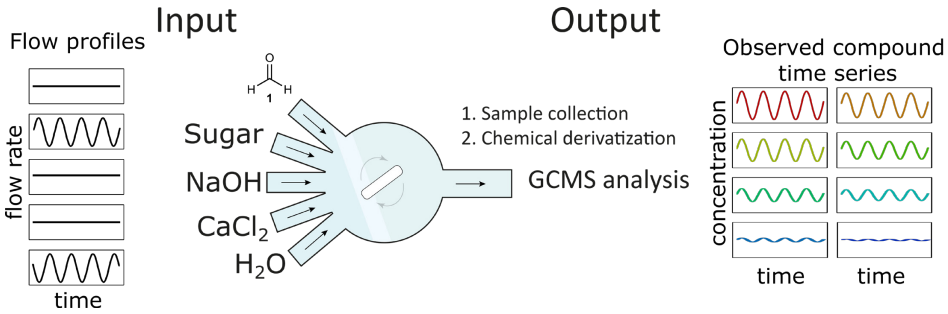


Figure 4.1: **Scheme of experimental setup for modulated flow inputs.** The flow rate of the input sugar was modulated to probe the reaction network. The flow rate of the water syringe was simultaneously adjusted against the flow rate of the input sugar to maintain a constant residence time.

4.2 Compositional outcomes were controlled by the environment

The formose reaction was carried out in a CSTR, with formaldehyde (**1**), an initiator sugar, $CaCl_2$, NaOH and H_2O in the input syringes, see figure 4.1. The compositional signature of the formose reaction was examined in 112 environmental conditions (see 4.7.1). Analysis of the chromatographic peaks and mass spectra provided a snapshot of the CSTR content (see 2.2 – 2.4). The compositional and reaction connectivity outcomes 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 (**2**), dihydroxyacetone (**3**), erythrulose (**9**) or ribose (**19**)). For the dataset, a default reaction network was used with $[\mathbf{1}]_{in}$ = 200 mM, $[\mathbf{2/3/9/19}]_{in}$ = 50 mM, $[CaCl_2]_{in}$ = 15 mM, $[NaOH]_{in}$ = 30 mM, r.t. = 2 minutes and T = 21 °C. Within the dataset discussed in this chapter, a total of 52 different compounds were used to account for the data, see figure 4.2.

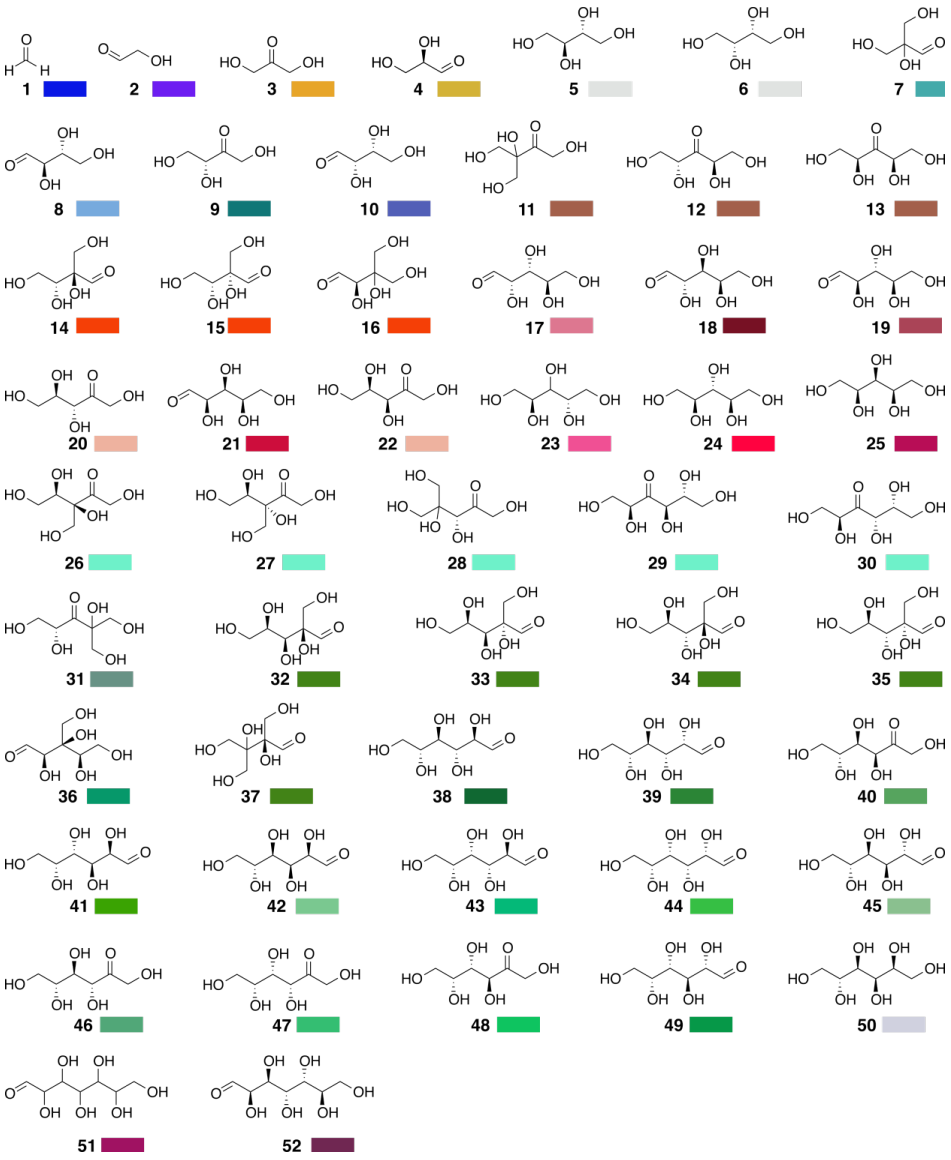


Figure 4.2: **The molecular structure of the 52 different compounds** in the dataset discussed in this chapter. Each compound has a color hue assigned. Alcohols in grey and carbohydrates of C_4 in blue, C_5 in red and C_6 in green.

Each environmental condition resulted in a unique compositional outcome, see for example figure 4.3a. A hierarchical clustering was performed on these time-averaged compositions using a correlation-based pairwise dissimilarity metric (see 4.6.1), see figure 4.3b. The resulting dendrogram 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. In the dendrogram, different branches were assigned a roman numeral (**I – VIII**).

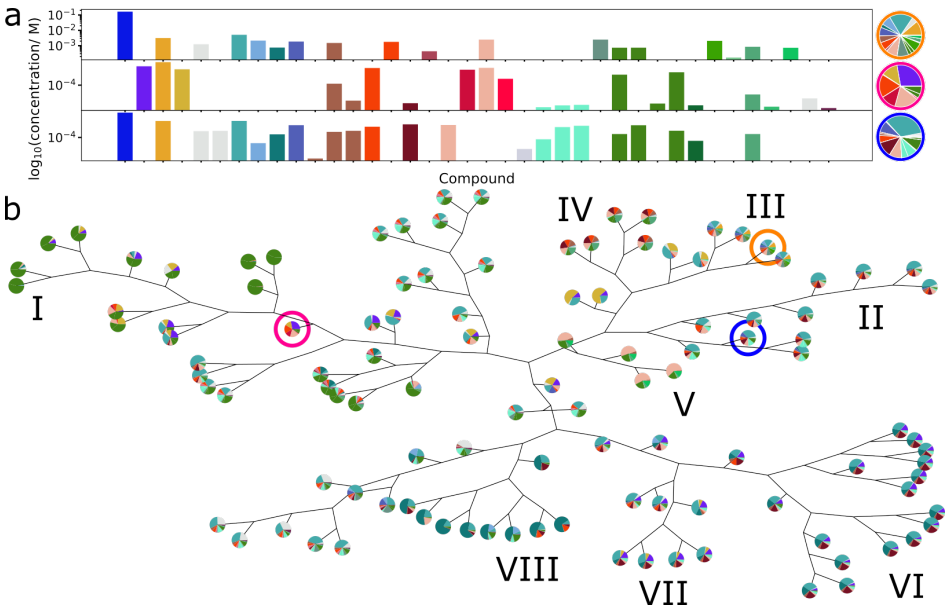


Figure 4.3: The compositions observed for changes in an overarching environment. a) Illustrative compositional bar charts which result from unique environmental conditions. Corresponding pie plots end up in branch **I**, **III** and **VI** of the cluster analysis. b) A hierarchical cluster tree, which separates out the compositional outcome of the different environments within the dataset.

A very rough interpretation of the branches in the dendrogram suggests that they are associated with one (or two) dominant factor(s) in the environment, see figure 4.4. Fine-tuning of compositions within branches results from the mixing of additional variables present in the experimental conditions. Branch **I** corresponds to relatively low formaldehyde input concentrations ($([\mathbf{1}]_{in}$ < 50 mM, fig. 4.4a) and variation in residence time (r.t. = 1 – 8 minutes, fig. 4.4b). Branch **II** corresponds to high formaldehyde concentration ($[\mathbf{1}]_{in}$ = 100 – 400 mM, fig. 4.4a), variation in temperature (T = 10 – 40 °C, fig. 4.4c) and variation in $[Ca(OH)_2]_{in}$ = 2.5/15/50 mM. The input sugars **2**, **9** and **19** correspond respectively with branches **III**, **IV** and **V** (fig. 4.4d). The combination of input concentration and ratio between $CaCl_2$ and NaOH results were expressed in branch **VI**, **VII** and **VIII** (fig. 4.4e,f,g).

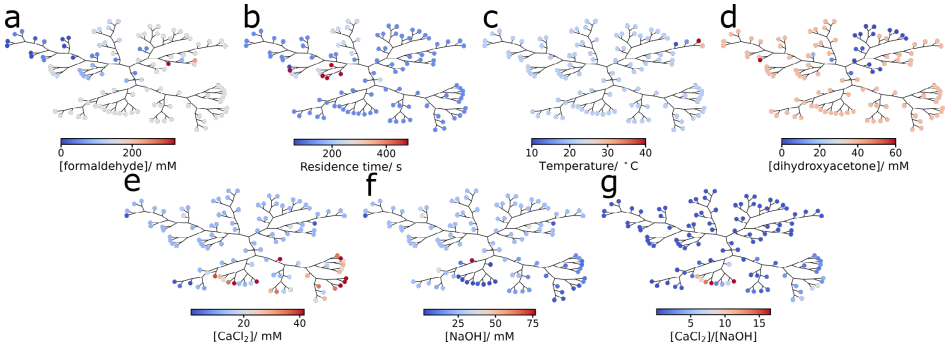


Figure 4.4: Respective conditional variation corresponds with dominantly modified environmental factor: a) formaldehyde, b) residence time, c) temperature, d) initiator sugar, e) $CaCl_2$, f) NaOH, g) Ca^{2+}/OH^-.

4.2.1 Sharp compositional transitions in gradual increase of formaldehyde feedstock

Varying in input concentration of **1** resulted in a drastic compositional shift, see figure 4.5a, especially in the low-concentration regime of **1** (fig. 4.5b). At low input concentration ($[\mathbf{1}]_{in}$ < 50 mM), the composition was dominated by relatively high concentration of C_6 compounds α-hydroxymethyl-aldohexose (**32**) and α,β-(hydroxymethyl)-aldotetrose (**37**). An increase of input concentration **1** towards 50 mM resulted in a concentration drop in **32** and **37**, alongside with a concentration rise of α-hydroxymethyl-glyceraldehyde (**7**), α-hydroxymethyl-aldotetrose (**14**), lyxose (**18**), ribulose (**20**), xylose (**21**), and an increased compositional complexity. This trend continued towards $[\mathbf{1}]_{in}$ = 100 mM, where **7**, threose (**10**), **18**, **20**, 3-ketohexose (**29**, **30**) and the newly formed α-hydroxymethyl-aldopentose (**34**) reached the highest concentration ([**7**/**10**/**18**/**20**/**29**/**30**/**34**] > 2 mM). Interestingly, **32** and **37** were almost completely depleted in this reaction network. Although individual concentrations did alter at the input range of **1** between 100 – 400 mM, no drastic changes in compositional complexity were observed here.

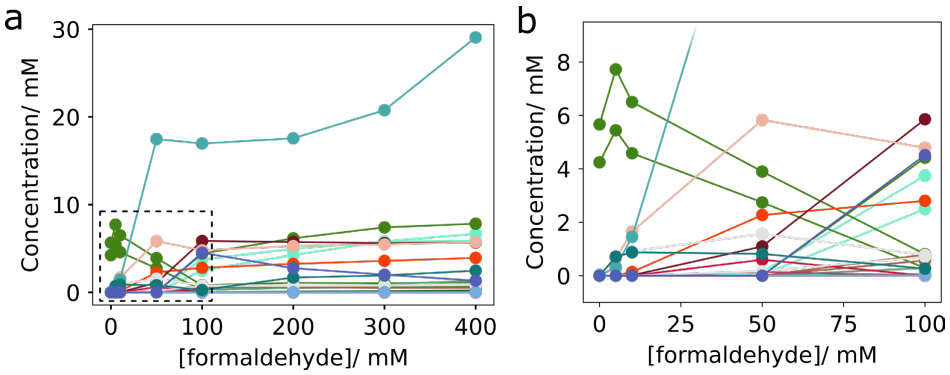


Figure 4.5: Compositional changes for changes in formaldehyde concentration. a) Variation in composition over a range from $[\mathbf{1}]_{in}$ = 0 – 400 mM. b) A zoom of the compositional transition between $[\mathbf{1}]_{in}$ = 0 – 100 mM.

The concentration of formaldehyde controls a thresholded compositional transition whose dynamic range is in the region $[\mathbf{1}]_{in}$ = 0 – 100 mM. Interestingly, this separates out branch **I** and **II** in the hierarchical cluster analysis, see figure 4.6. Following branch **I** from its tip towards the centre of the tree, more diverse sets of compounds were produced. This corresponds with an increase in concentration of ($[\mathbf{1}]_{in} \leq$ 50mM), where $[\mathbf{1}]_{in}$ = 50 mM at the root of **I**. Towards $[\mathbf{1}]_{in}$ = 100 mM the composition passed a threshold towards branch **II**, where **10**, **18**, **29**, **30** and **34** were added to the composition. This is the maximal compositional complexity which the network acquired in the **1** concentration series.

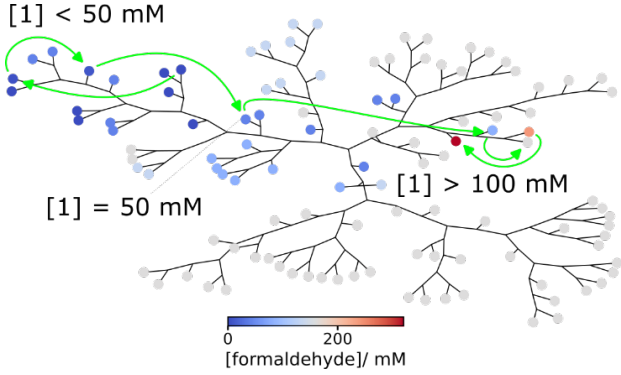


Figure 4.6: The trajectory through the composition dendrogram for changes in formaldehyde input from $[\mathbf{1}]_{in}$ = 0 – 400 mM.

4.2.2 Varying [Ca^{2+}] : [OH^-] ratios lead to high compositional variation

Variation in both Ca^{2+} and OH^- concentration and their respective ratio, had great effect on the compositional outcome. A range of [Ca^{2+}] : [OH^-] input ratios (remaining below the solubility limit of $Ca(OH)_2$ as discussed under 2.2 and 2.6.3) were explored in the dataset. A demonstrative subset of the data crosses four branches of the dendrogram (**II**, **VI**, **VII** and **VIII**), see figure 4.7. Branch **II**, with $[\mathbf{1}]_{in}$ > 100 mM, covers two experimental conditions with an altered input concentration of $CaCl_2$ and NaOH, which remain at a 1 : 2 ratio. The compositional complexity followed the trend as discussed under 4.2.1 for branch **II**. Lowering the ratio to 1 : 1, resulted in a jump from branch **II** to **VI**, with an increase in relative proportions of **9**, **13**, **15** and **18** in comparison to **7**, **14**, **29**, **30** and **37**. Further increasing the ratio lowered the the compositional diversity, with **9** becoming more prominent and the composition crossed to branch **VII** and further to **VIII**. Notably, in branch **VIII**, compositions recorded for $[CaCl_2]_{in}$ = 20 or 52 mM, and $[NaOH]_{in}$ = 2.5 mM were compositionally very similar and mainly dominated by **7**, **9** and **15**. Furthermore, at $[CaCl_2]_{in}$ = 52 mM and $[NaOH]_{in}$ = 20 mM, **7**, **8** and **20** were particularly prominent.

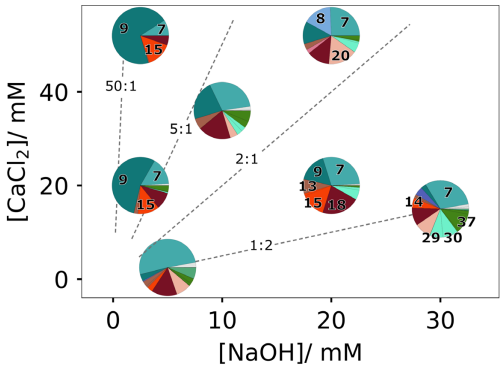


Figure 4.7: Pie plots for the compositional variation of different $[CaCl_2]_{in}$ and $[NaOH]_{in}$. The remainder experiment conditions were: $[\mathbf{1}]_{in}$ = 200 mM, r.t. = 2 minutes and T = 21 °C.

4.2.3 Distinct compositional outcomes for varying other environmental factors

Variation in residence time between 1 and 8 minutes led to compositional shifts which remain in the low formaldehyde branch **I**. The respective composition was rich in lyxose (**18**) and ribulose (**20**) at low retention times, and the broadest spread of product carbon chain lengths was observed for the default residence time of 2 minutes. Here, compound **9**, **10**, **14**, **37**, **29** and **30** were particularly prominent. After an increase in residence time beyond 2 minutes, the amount of

18 was substantially reduced. The concentration of **20** and **30** continued to have a strong contribution to the composition. At residence times above 2 minutes, the concentration of threitol (**6**) and **14** increased together with **32** and **37**, which dominated the low formaldehyde ($[\mathbf{1}]_{in}$ < 50 mM) reaction network.

Remarkably, the composition of the formose reaction remained very similar for both the temperature series in branch **II** and the initiator series with **2**, **9** and **19**, respectively in branch **III**, **IV** and **V**. The observed influence of temperature between 10 – 40 °C on the steady-state composition was modest, and future work should include a broader range. The impact of different sugar initiators on the compositional outcome was much more significant, but unpredictable.

4.3 Rewired reaction pathways governed compositional transitions

The observed compositional variations were a direct result of the translation of the various input conditions through the underlying formose reactions. The input concentration of initiating sugar was modulated to probe the reaction connectivity in product compositions, as discussed under 3.3.2. A 'global' formose reaction network was generated *in silico* (see 3.2 and 3.6.1) to guide the pathway search with the observed amplitude decay in reaction products. This framework provides a direct translation of the experimental data into a descriptive set of reactions responsible for the compositions observed.

4.3.1 Network reorganization over the formaldehyde input threshold

The search procedure was used to estimate the self-organizational of the formose reaction pathways in response to various environmental changes. The thresholded reorganization of the network upon increasing formaldehyde concentration can be rationalized, see figure 4.8a. In this experimental series, $[\mathbf{3}]_{in}$ was modulated sinusoidally between 25 – 75 mM with a 6 minutes period. Figure 4.8a represents a comparative network of the most prominent reaction pathways ([…] > 2 mM), which rationalizes the observed compositional shift. The green flux arrows represent the underlying structure of the formose network for $[\mathbf{1}]_{in}$ = 5mM. The operative reaction pathways are mainly accounted for by a small set of reactions between C_3 species to form C_6 compounds **32** and **37** (fig 4.8a). Increasing $[\mathbf{1}]_{in}$ to 100 mM triggered an expansion of the repertoire of reactions, as depicted by the purple flux arrows (fig. 4.8a). In the high concentration of **1** network, the number of possible pathways for formaldehyde addition increases, with a corresponding increase in the number of proton transfer pathways. Interestingly, the observed reaction pathways are not only accounted for by formaldehyde addition chain growth. The production of C_5

compounds **14**, **18** and **20** is attributed to C_2 addition reactions.[22] These C_5 compounds coupled strongly to the modulated input **3**, with A_{14} = 0.35 mM, A_{18} = 2.1 mM and A_{20} = 0.46 mM (fig. 4.8b), whereas intermediate **9** for the formaldehyde chain growth pathway did not (A_9 = 0.09 mM). The coupling to the input of the reaction pathways with **2** instead of **1** is remarkable, since **1** is the strongest electrophile fed into the network at a high concentration of 100 mM. Meanwhile, formaldehyde addition can build on top of such enolate–monosaccharide pathways. For example, **29** was produced after formaldehyde addition to the enolate species of **20**.

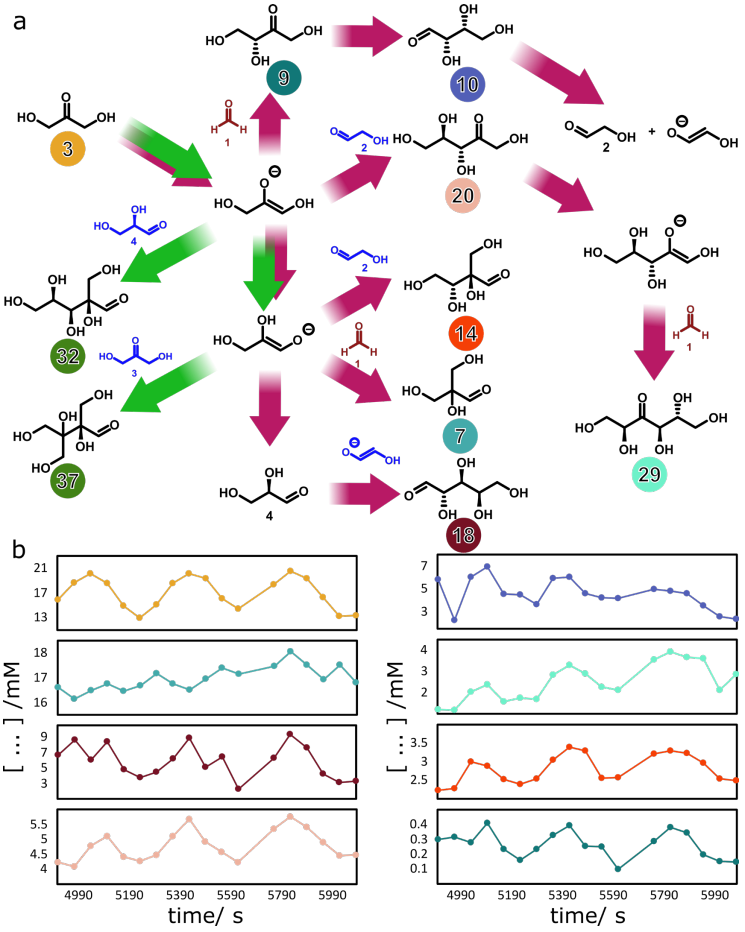


Figure 4.8: **Rewiring of reaction pathways in the network in the formaldehyde series.** a) The green arrows correspond to the flux through the reaction network at low formaldehyde input concentration ($[\mathbf{1}]_{in}$ = 5 mM). Purple arrows represent the reaction

network at high formaldehyde input concentration ($[\mathbf{1}]_{in}$ = 100 mM). b) Concentration series of compounds observed in the $[\mathbf{1}]_{in}$ = 100 mM network.

4.3.2 Shifts in reaction types governed shifts in compositional outcomes

Following the formaldehyde input series from $[\mathbf{1}]_{in}$ = 0 – 400 mM revealed key reaction characteristics that govern the various reaction outcomes, see figure 4.9. An important feature of the formose network at low formaldehyde input concentration ($[\mathbf{1}]_{in} \leq$ 10 mM) 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 were responsible for creating products **32** and **37**.[23,24] Moving to the high formaldehyde input regime ($[\mathbf{1}]_{in} \geq$ 100 mM), the repertoire of reactions expanded, and aldol addition reactions involving **1** were added to the network. In particular, reactions in which the α-carbon is bound to a hydrogen (R_1) or glycol group (R_3) became more prevalent. A range of protonation/deprotonation reactions were promoted as well. Deprotonation was favored at less sterically hindered positions (where the α-carbon is bound to a hydrogen or hydroxymethyl group ($R_{1/2}$)). Protonation was favored at α-carbons bound to hydroxymethyl groups (R_2). Interestingly, the number of monosaccharide – enolate reactions also increased, suggesting that some monosaccharide products interact with other members of the network as reactants. The expression of different reaction classes (see 3.7.1, fig. S3.11) in the $[\mathbf{1}]_{in}$ = 50 mM network was a clear mix of the expressed reaction classes between both low and high **1** input.

The compositional shift from a $[Ca^{2+}]$: $[OH^-]$ ratio of 1 : 2 in branch **II** to 1 : 20 in branch **VIII**, as discussed under 4.2.2, followed an opposite trend for the expressed reaction classes. Branch **II** accommodated both experiments with a $[Ca^{2+}]$: $[OH^-]$ ratio of 1 : 2, with $[Ca(OH)_2]$ = 2.5 / 15 mM and $[\mathbf{1}]_{in}$ = 200 mM. The reactivity in the lowest $[Ca^{2+}]$: $[OH^-]$ ratio appears similar to the high input concentration regime of **1**. Increasing the $[Ca^{2+}]$: $[OH^-]$ ratio (7 : 2, $[Ca^{2+}]$ = 35 mM, $[OH^-]$ = 10 mM), as part of branch **VI**, resulted in a smaller contribution of formaldehyde-controlled pathways and the contribution of (de)protonation was lower. The expression of reaction classes in branch **VI**, followed a similar pattern as the network of $[\mathbf{1}]_{in}$ = 50 mM. Compared to branch **II**, the population of C_5 species and instances of Cannizzaro reactions decreased. As the ratio was increased further in branch **VIII**, the conditions and reaction pathways began to resemble those found for the low $[\mathbf{1}]_{in}$ regime (≤ 10 mM). At $[Ca^{2+}]_{in}$ = 52 mM and $[OH^-]_{in}$ = 2.5 mM in branch **VIII**, a limited set of pathways was in operation, the majority of which may be accounted for by formaldehyde addition and proton

transfer reactions terminated at **15** *via* **9**. Interestingly, the linear C_5 compounds **12** and **13** were not formed in appreciable quantities, so the reaction hits the 'dead end' branched **15**. The system unexpectedly had a strong preference for formaldehyde addition product **9** instead of **7** under these conditions.

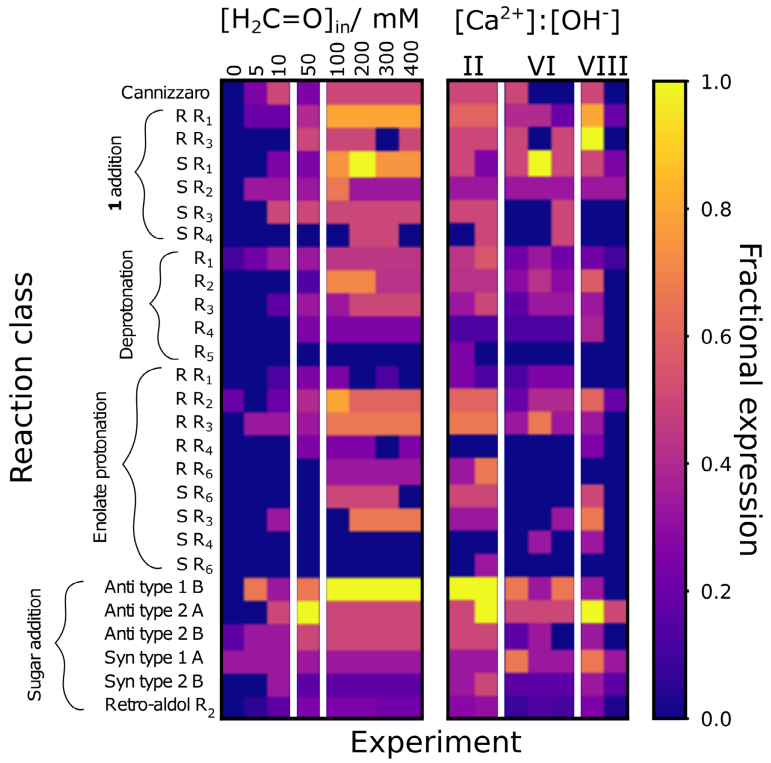


Figure 4.9: Expression of the determined reaction types vary between different environmental conditions. Fractional expression refers to the counts of the reaction class (fig. S3.11) observed compared to the number of reactions in the class in a reaction network representing the union of the reactions found for all datasets.

4.3.3 The Breslow cycle and glycolaldehyde production have a key role in the self-organization of the network

In contrast to prevailing views on the formose reaction, the presented experimental data indicates that formaldehyde-based chain growth pathways do not completely account for the observed behavior. Rather, reactions between C_2 and C_3 compounds are key chain-building reactions.[22,25–27] Interestingly, glycolaldehyde (C_2) is produced in the network with **3** as input sugar through the emergence of a self-organized cyclic set of reactions, where **2** is created from retro-aldol reactions. This corresponds with Breslow's proposed mechanism for

autocatalysis in the formose reaction, see fig. 4.10.[28,29] Although usually seen as an autocatalytic mechanism, the 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 4.8, C_6 compounds, such as **29**, may also act as sources of C_2 and C_4 monosaccharides and enolates which reinforces 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 network in which chain growth occurs *via* formaldehyde addition. Therefore, the Breslow cycle can be envisaged as a source of new reaction pathways through which monosaccharides are built. These reactions between formose reaction products are an excellent example of how underlying patterns in chemical reactivity define reaction outcomes. The molecular diversity is promoted by activating a class of reactions in which longer carbon chains are synthesized from building blocks larger than **1**.

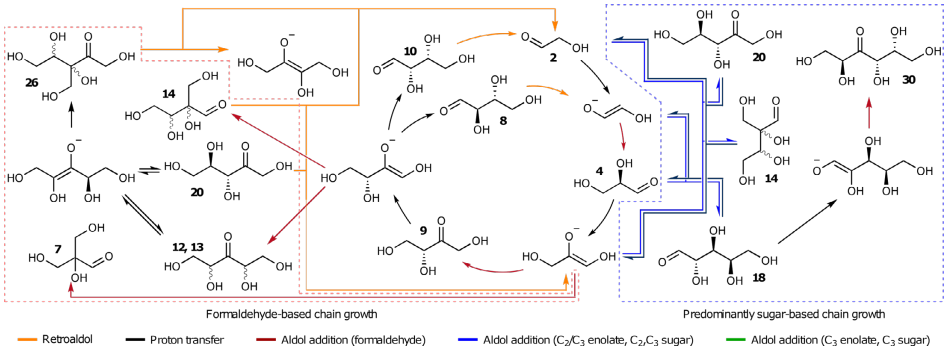


Figure 4.10: Network generated 2 as part of the Breslow's cycle. The new reactivity introduced by **2** alters the network reactivity (blue arrows).

4.4 Conclusion

This study has demonstrated how the environment directs the self-organization of reaction pathways and creates well-defined product compositions in the formose reaction. In the absence of a directing force, the recursive application of a limited set of chemical reactions operative in the formose reaction potentially leads to a wide range of possible reaction pathways and compositions. However, in the diverse set of investigated environments, the formose reaction did use only a subset of these pathways. Local conditions define the self-organization of the reaction network and the resulting reaction pathways produced well-defined product mixtures. Interestingly, over a range of formaldehyde input concentrations, a complex compositional change was observed. Similarly,

changes in Ca^{2+} to OH^- ratio did result in various characteristic compositional outcomes. The organization of the underlying reaction routes is estimated *via* analysis of the time-resolved propagation of periodically changing inputs. Remarkably, the thresholded compositional change in the formaldehyde input series cannot be simply explained by formaldehyde addition reactions. In high formaldehyde concentration networks, the most significant reaction pathways are governed by C_2 and C_3 compounds generated within the network. By summarizing the different reaction classes underpinning the reconstructed reaction networks, a similar reactivity trend was observed for changes decreasing the ratio of Ca^{2+} to OH^-. It is likely that the Breslow cycle, an autocatalytic reaction pathway, provided the reaction network with these key reactive species. Consequently, it introduced an alternative and emergent set of reactions that allows sugars to be built from C_2 and C_3 building blocks, alongside addition reactions with **1**.

Spontaneous fine tuning of chemical systems is directed by a forcing from the environment. This property of prebiotic reaction systems was already predicted in theoretical models.[30,31] The interaction between the environment and inherent chemical reactivity controls the self-organization of the model prebiotic formose reaction. Abiotic molecular systems can adapt to the environment in a controlled fashion, consequently yielding well-defined reaction pathways and compositional outcomes.

4.5 Method summary

4.5.1 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 form Carbosynth Ltd, L-(−)-sorbose and D-tagatose, 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. All other chemicals were used without further purification.

4.5.2 Instrumentation

GC-MS experiments were carried out on *JEOL JMS-100GCv* attached to an *Agilent 7890A GC* using the procedure as described under 2.3.3.

A *Shimadzu Nexera X2* was used for HPLC experiments, as described under 2.3.4.

4.5.3 Methods

Flow reactions were carried out in a CSTR as described under 2.2 and 2.6.1.

Derivatization for GC-MS and HPLC analysis was performed as described under 2.3.1 and 2.3.2.

For chromatographic data processing we use the Python program ChromProcess, authored by W.E. Robinson, as described under 2.4 and 2.6.4.

The *in silico* generation of the 'global' reaction network was carried out as described under 3.2 and 3.6.1 with reaction classes in 3.7.1. Estimation of reaction pathways from experiments with modulated input sugar is performed as described under 3.3.2 and 3.6.2.

4.6 Data analysis methods

Python programs for the following described data analysis are available at https://github.com/huckgroup/formose-2021.git (authored by W.E. Robinson). The extracted data from GC-MS and HPLC analysis were available under 'DATA'.

All data analyses were performed using the Numpy[32] (1.22.2.) and SciPy[33] (1.8.0) Python libraries.

4.6.1 Hierarchical clustering of data

The average compositions and amplitudes for each experiment set are combined into an array. The pairwise dis-similarity between each data set is then determined using a correlation-based metric (Eq 1., scipy.spatial.distance.pdist() using the 'correlation' metric)

Eq 1. $1 - PCC = 1 - \frac{(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{(u_i - \overline{u})^2}.\sqrt{(v_i - \overline{v})^2}}$

Where u and v are both one-dimensional vectors (arrays) of average compound concentrations and amplitudes determined for a given experiment.

4.7 Supplementary information

4.7.1 Table reaction conditions of the performed experiments.

Exp code	[1]/ M	Initiator sugar	[sugar]/ M	[NaOH]/ M	[CaCl2]/ M	amp/ M	freq/ s	rt/ s	T/ °C
FRN051A	0.1	DHA	0.05	0.03	0.015	0	0	480	21
FRN051B	0.1	DHA	0.05	0.03	0.015	0.025	960	480	21
FRN055B	0.1	DHA	0.05	0.03	0.015	0.025	120	120	21
FRN059A	0.1	DHA	0.05	0.03	0.015	0	0	240	21
FRN059B	0.1	DHA	0.05	0.03	0.015	0.025	240	240	21
FRN060A	0.1	DHA	0.05	0.03	0.015	0	0	480	21
FRN060B	0.1	DHA	0.05	0.03	0.015	0.0	480	480	21
FRN061A	0.166	DHA	0.05	0.03	0.015	0	0	240	21
FRN061B	0.166	DHA	0.05	0.03	0.015	0.025	240	240	21
FRN062A	0.166	DHA	0.05	0.03	0.015	0.025	120	240	21
FRN062B	0.166	DHA	0.05	0.03	0.015	0.025	240	240	21
FRN062C	0.166	DHA	0.05	0.03	0.015	0.025	360	240	21
FRN062D	0.166	DHA	0.05	0.03	0.015	0.025	480	240	21
FRN063A	0.166	DHA	0.05	0.03	0.015	0	0	240	21
FRN063B	0.166	DHA	0.05	0.03	0.015	0	0	240	21
FRN063C	0.166	DHA	0.05	0.03	0.015	0	0	240	21
FRN063D	0.166	DHA	0.05	0.03	0.015	0	0	240	21
FRN063E	0.166	DHA	0.05	0.03	0.015	0	0	240	21
FRN067A	0.2	DHA	0.05	0.03	0.0015	0	0	120	21
FRN067B	0.2	DHA	0.05	0.03	0.003	0	0	120	21
FRN067C	0.2	DHA	0.05	0.03	0.006	0	0	120	21
FRN067D	0.2	DHA	0.05	0.03	0.012	0	0	120	21
FRN067E	0.2	DHA	0.05	0.03	0.024	0	0	120	21
FRN067F	0.2	DHA	0.05	0.03	0.048	0	0	120	21
FRN067G	0.2	DHA	0.05	0.003	0.015	0	0	120	21
FRN067H	0.2	DHA	0.05	0.006	0.015	0	0	120	21
FRN067I	0.2	DHA	0.05	0.012	0.015	0	0	120	21
FRN067J	0.2	DHA	0.05	0.024	0.015	0	0	120	21
FRN067K	0.2	DHA	0.05	0.048	0.015	0	0	120	21
FRN067L	0.2	DHA	0.05	0.096	0.015	0	0	120	21
FRN071A	0.2	ribose	0.05	0.03	0.015	0.025	120	120	21
FRN071B	0.2	ribose	0.05	0.03	0.015	0.025	180	120	21
FRN071C	0.2	ribose	0.05	0.03	0.015	0.025	240	120	21
FRN071D	0.2	ribose	0.05	0.03	0.015	0.025	300	120	21
FRN077A	0.2	2 $C_2H_4O_2$	0.025	0.03	0.015	0.0125	120	120	21
FRN077B	0.2	2 $C_2H_4O_2$	0.025	0.03	0.015	0.0125	180	120	21
FRN077C	0.2	2 $C_2H_4O_2$	0.025	0.03	0.015	0.0125	240	120	21
FRN077D	0.2	2 $C_2H_4O_2$	0.025	0.03	0.015	0.0125	300	120	21
FRN087A	0.2	erythrulose	0.05	0.03	0.015	0.025	120	120	21
FRN087B	0.2	erythrulose	0.05	0.03	0.015	0.025	180	120	21
FRN087C	0.2	erythrulose	0.05	0.03	0.015	0.025	240	120	21
FRN087D	0.2	erythrulose	0.05	0.03	0.015	0.025	300	120	21
FRN087E	0.2	erythrulose	0.05	0.03	0.015	0.025	360	120	21
FRN088A	0.2	DHA	0.05	0.005	0.0025	0.025	360	120	21
FRN088B	0.2	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN088C	0.2	DHA	0.05	0.05	0.025	0.025	360	120	21
FRN089A	0.1	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN089B	0.2	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN089C	0.3	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN089D	0.4	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN090A	0	DHA	0.05	0.005	0.0025	0.025	360	120	21
FRN090B	0	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN090C	0	DHA	0.05	0.04	0.02	0.025	360	120	21
FRN090D	0	DHA	0.05	0.05	0.025	0.025	360	120	21
FRN091	0.2	DHA	0.05	0.03	0.015	0.025	360	240	21
FRN092A	0.2	DHA	0.05	0.03	0.015	0.025	960	480	21

FRN092B	0.2	DHA	0.05	0.03	0.015	0.025	120	60	21
FRN093A	0.005	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN093B	0.01	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN093C	0.05	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN094A	0	DHA/2 $C_2H_4O_2$	0.05/0.0025	0.03	0.015	0.025/0	360/0	120	21
FRN094B	0	DHA/2 $C_2H_4O_2$	0.05/0.005	0.03	0.015	0.025/0	360/0	120	21
FRN094C	0	DHA/2 $C_2H_4O_2$	0.05/0.025	0.03	0.015	0.025/0	360/0	120	21
FRN097A	0.2	DHA	0.05	0.0025	0.02	0	0	120	21
FRN097B	0.2	DHA	0.05	0.005	0.02	0	0	120	21
FRN097C	0.2	DHA	0.05	0.01	0.02	0	0	120	21
FRN097D	0.2	DHA	0.05	0.015	0.02	0	0	120	21
FRN097E	0.2	DHA	0.05	0.02	0.02	0	0	120	21
FRN097F	0.2	DHA	0.05	0.0025	0.028	0	0	120	21
FRN097G	0.2	DHA	0.05	0.005	0.028	0	0	120	21
FRN097H	0.2	DHA	0.05	0.01	0.028	0	0	120	21
FRN097I	0.2	DHA	0.05	0.015	0.028	0	0	120	21
FRN097J	0.2	DHA	0.05	0.02	0.028	0	0	120	21
FRN097K	0.2	DHA	0.05	0.0025	0.036	0	0	120	21
FRN097L	0.2	DHA	0.05	0.005	0.036	0	0	120	21
FRN097M	0.2	DHA	0.05	0.01	0.036	0	0	120	21
FRN097N	0.2	DHA	0.05	0.015	0.036	0	0	120	21
FRN097O	0.2	DHA	0.05	0.02	0.036	0	0	120	21
FRN098A	0.2	DHA	0.05	0.0025	0.044	0	0	120	21
FRN098B	0.2	DHA	0.05	0.005	0.044	0	0	120	21
FRN098C	0.2	DHA	0.05	0.01	0.044	0	0	120	21
FRN098D	0.2	DHA	0.05	0.015	0.044	0	0	120	21
FRN098E	0.2	DHA	0.05	0.02	0.044	0	0	120	21
FRN098F	0.2	DHA	0.05	0.0025	0.052	0	0	120	21
FRN098G	0.2	DHA	0.05	0.005	0.052	0	0	120	21
FRN098H	0.2	DHA	0.05	0.01	0.052	0	0	120	21
FRN098I	0.2	DHA	0.05	0.015	0.052	0	0	120	21
FRN098J	0.2	DHA	0.05	0.02	0.052	0	0	120	21
FRN099A	0.05	DHA	0.00225	0.03	0.015	0	0	120	21
FRN099B	0.05	DHA	0.003	0.03	0.015	0	0	120	21
FRN099C	0.05	DHA	0.006	0.03	0.015	0	0	120	21
FRN099D	0.05	DHA	0.01	0.03	0.015	0	0	120	21
FRN099E	0.05	DHA	0.02	0.03	0.015	0	0	120	21
FRN099F	0.05	DHA	0.05	0.03	0.015	0	0	120	21
FRN099G	0.05	DHA	0.075	0.03	0.015	0	0	120	21
FRN100A	0.05	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN100B	0.05	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN100C	0.05	DHA	0.05	0.03	0.015	0.025	360	120	21
FRN103	0.2	DHA	0.05	0.01	0.036	0.025	360	120	21
FRN104A	0.2	DHA	0.05	0.0025	0.02	0.025	360	120	21
FRN104B	0.2	DHA	0.05	0.0025	0.052	0.025	360	120	21
FRN105A	0.2	DHA	0.05	0.02	0.02	0.025	360	120	21
FRN105B	0.2	DHA	0.05	0.02	0.052	0.025	360	120	21
FRN106A	0.2	DHA/2 $C_2H_4O_2$	0.05/0.005	0.03	0.015	0.025/0	360/0	120	21
FRN106B	0.2	DHA/2 $C_2H_4O_2$	0.05/0.01	0.03	0.015	0.025/0	360/0	120	21
FRN106C	0.2	DHA/ 2 $C_2H_4O_2$	0.05/0.015	0.03	0.015	0.025/0	360/0	120	21
FRN106D	0.2	DHA/2 $C_2H_4O_2$	0.05/0.02	0.03	0.015	0.025/0	360/0	120	21
FRN106E	0.2	DHA/2 $C_2H_4O_2$	0.05/0.025	0.03	0.015	0.025/0	360/0	120	21
FRN107A	0.2	DHA	0.05/0.05	0.03	0.015	0	0	120	10
FRN107B	0.2	DHA	0.05	0.03	0.015	0	0	120	15
FRN107C	0.2	DHA	0.05	0.03	0.015	0	0	120	30
FRN107D	0.2	DHA	0.05	0.03	0.015	0	0	120	40

4.7.2 GC-MS calibration for quantitative analysis

Constants for fitted quadratic GC calibration curves for C_4, C_5 and C_6 compounds (fig. S4.11-13). The same calibration curves have been used as previously reported[1], based on: $[compound] = A \times (\frac{peak\ integral}{internal\ standard\ integral})^2 + B \times \frac{peak\ integral}{internal\ standard\ integral}$. A calibration curve was estimated for non-calibrated compounds by taking the average calibration curve for the calibrated sugars of similar length.

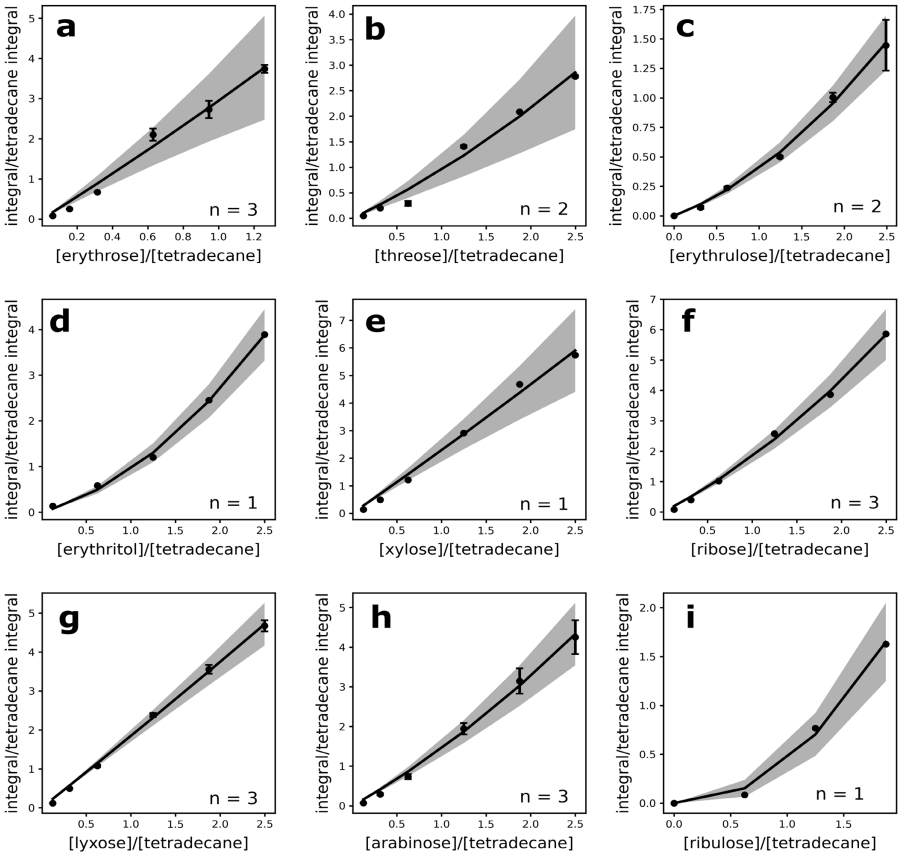


Figure S4.11: **GC-MS calibration lines** for a) erythrose, b) threose, c) erythrulose, d) erythritol, e) xylose, f) ribose, g) lyxose, h) arabinose, i) ribulose. Dots: data, lines: fitted polynomial curves, shaded areas denote confidence interval of the fitted lines. Concentrations and peak areas are reported relative to the tertradecane internal standard. Due to the presence of cis/trans- isomers of the ethoxime group, some compounds give two peaks in GC-MS chromatograms. The peak with the highest integral of the two was

used for calibration. The peak integral reported is the highest of Error bars are ± one standard deviation (n given in panel).

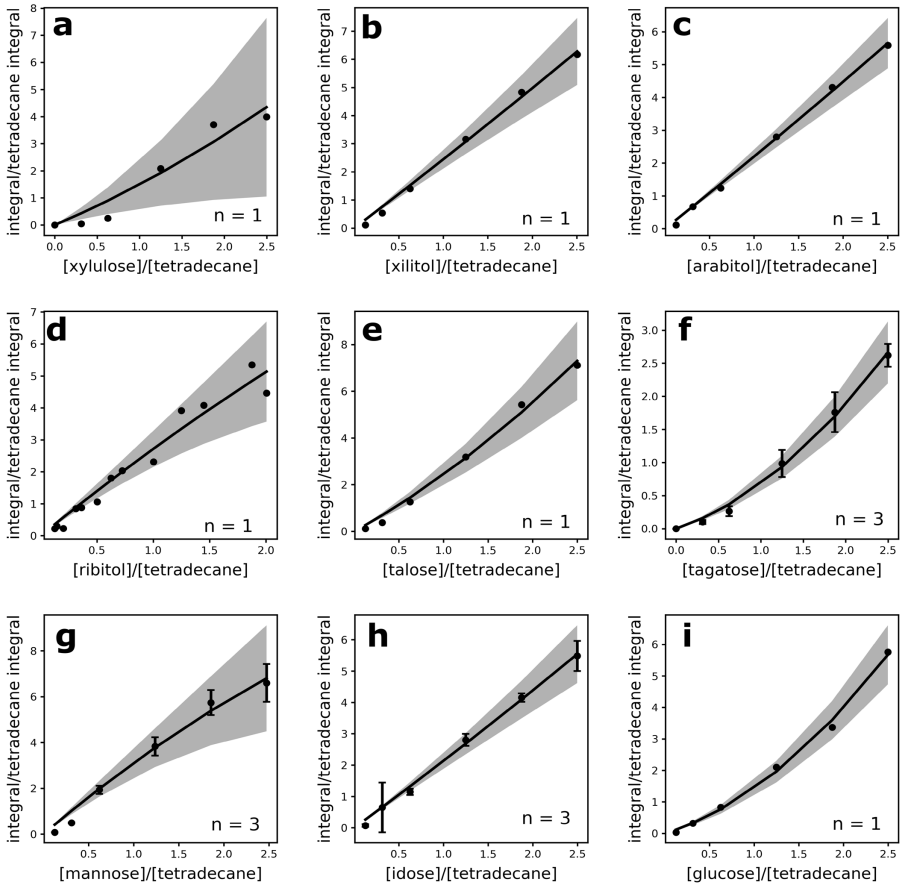


Figure S4.12: **GC-MS calibration lines** for a) xylulose, b) xylitol, c) arabitol, d) ribitol, e) talose, f) tagatose, g) mannose, h) idose, i) gulose. Dots: data, lines: fitted polynomial curves, shaded areas denote confidence interval of the fitted lines. Concentrations and peak areas are reported relative to the tertradecane internal standard. Due to the presence of cis/trans- isomers of the ethoxime group, some compounds give two peaks in GC-MS chromatograms. The peak with the highest integral of the two was used for calibration. Error bars are ± one standard deviation (n given in panel).

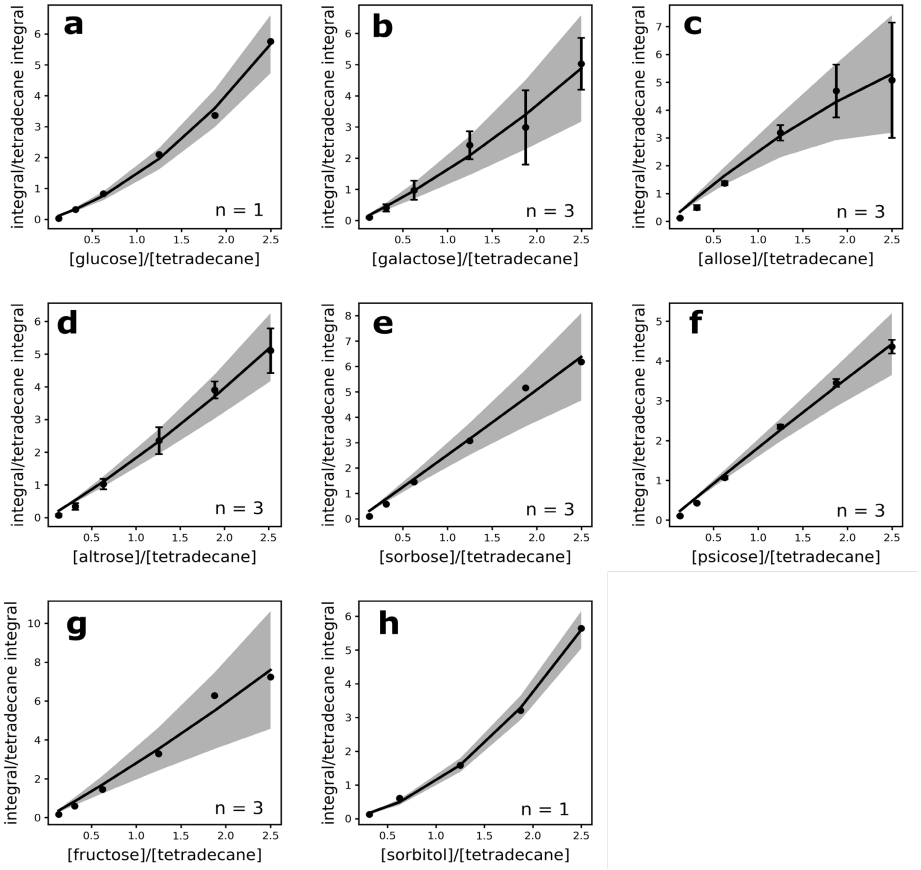


Figure S4.13: **GC-MS calibration lines** for a) glucose, b) galactose, c) allose, d) altrose, e) sorbose, f) psicose, g) fructose, h) sorbitol. Dots: data, lines: fitted polynomial curves, shaded areas denote confidence interval of the fitted lines. Concentrations and peak areas are reported relative to the tertradecane internal standard. Due to the presence of cis/trans- isomers of the ethoxime group, some compounds give two peaks in GC-MS chromatograms. The peak with the highest integral of the two was used for calibration. Error bars are ± one standard deviation (n given in panel).

4.7.3 Selection of formaldehyde concentrations

Within the conditions explored in this work, formaldehyde concentration was varied in the range 0 to 400 mM. This range was chosen based on the series of experiments depicted in figure 4.5. Increasing the formaldeyde concentration beyond 200 mM had little effect on the composition of the formose reactions. 200 mM was chosen as benchmark for the formaldehyde concentration, under these conditions a broad range of products was formed.

4.7.4 Selection of calcium hydroxide concentrations

The conditional variations explored cover $CaCl_2$ variation in the range 1.5 to 52.0 mM and NaOH variation in the range 2.5 to 96.0 mM. Within these studied combinations, no precipitation of $Ca(OH)_2$ was observed. Applying higher concentrations beyond the ranges used for both compounds would likely result in significant precipitate formation. Increasing the NaOH concentration to above 30 mM resulted in a higher amount of formaldehyde consumption, without concurrent increases in product concentration. This effect is attributed to increased levels of Cannizzaro reaction, by disproportionation of formaldehyde into formate and methanol. 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. 4.7). 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.

4.7.5 Selection of residence times

A variety of residence times was explored in the range 60 to 480 s. At a residence time of 120 s, the product distribution is 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 formed. Above 120 s residence time, the amount of C_4 products, as well as lyxose, was significantly diminished. The observation of compounds with a variety of chain lengths is a vital component in searching for reaction pathways connecting the various formose products. Furthermore, keeping the formose reaction at conditions under which the reaction is far from completion renders it in a state which is more sensitive to changes in other reactions conditions. For these reasons, a benchmark residence time of 120 s was chosen to fulfil these criteria.

4.7.6 Selection of temperatures

Temperatures in the range 10 – 50 °C were explored, finding little impact on the reaction composition. The experimental setup does not allow temperatures outside this range. A temperature of 21 °C was used as a benchmark temperature for the experiments in table 4.7.1.

4.7.7 Selection of amplitudes

An amplitude of 25 mM was applied for the modulation of each input sugar, whilst steady-state experiments were performed with an input modulation of 0 mM. Since glycolaldehyde was used as a dimer, the amplitude concentration was lowered to 12.5 M. Other magnitudes of amplitudes were not explored, as such investigations were 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 likely prevent the observation of the transfer of the input modulation to products, as the induced amplitudes may drop below the signal-to-noise ratio of the analysis workflow employed. Significantly increasing the amplitude beyond 25 mM may have introduced strong non-linear effects on the reaction. With the intent of modulation providing only a characterization handle for reaction connectivity, the amplitude is attempted to maintain moderate perturbations of the reaction.

4.7.8 Selection of periods

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.[34] Therefore, the modulation periods are benchmarked at three times the residence time of the experiment (360 s for a benchmark residence time of 120 s).

4.8 References

1. Surman, A. J. *et al.* Environmental control programs the emergence of distinct functional ensembles from unconstrained chemical reactions. *Proc. Natl. Acad. Sci.* **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. Ritson, D. & Sutherland, J. D. Prebiotic synthesis of simple sugars by photoredox systems chemistry. *Nat. Chem.* **4**, 895–899 (2012).
4. Ritson, D. J. & Sutherland, J. D. Synthesis of Aldehydic Ribonucleotide and Amino Acid Precursors by Photoredox Chemistry. *Angew. Chem. Int. Ed.* **52**, 5845–5847 (2013).
5. 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).
6. Xu, J. *et al.* A prebiotically plausible synthesis of pyrimidine β-ribonucleosides and their phosphate derivatives involving photoanomerization. *Nat. Chem.* **9**, 303–309 (2017).
7. Becker, S. *et al.* Wet-dry cycles enable the parallel origin of canonical and non-canonical nucleosides by continuous synthesis. *Nat. Commun.* **9**, 163 (2018).
8. Becker, S. *et al.* Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides. *Science* **366**, 76–82 (2019).

9. Keller, M. A., Turchyn, A. V. & Ralser, M. Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible A rchean ocean. *Mol. Syst. Biol.* **10**, 725 (2014).
10. Keller, M. Sulfate radicals enable a non-enzymatic Krebs cycle precursor. (2017) doi:10.17632/VGPMNZDZ55.1.
11. 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).
12. 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).
13. Muchowska, K. B., Varma, S. J. & Moran, J. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019).
14. Jinich, A. *et al.* A thermodynamic atlas of carbon redox chemical space. *Proc. Natl. Acad. Sci.* **117**, 32910–32918 (2020).
15. Orgel, L. E. Self-organizing biochemical cycles. *Proc. Natl. Acad. Sci.* **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.* **131**, 4568–4573 (2019).
20. Urmès, C. *et al.* Periodic reactor operation for parameter estimation in catalytic heterogeneous kinetics. Case study for ethylene adsorption on Ni/Al2O3. *Chem. Eng. Sci.* **214**, 114544 (2020).
21. Mettetal, J. T., Muzzey, D., Gómez-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. Gutsche, C. D. *et al.* Base-catalyzed triose condensations. *J. Am. Chem. Soc.* **89**, 1235–1245 (1967).
24. Berl, W. G. & Feazel, C. E. The Kinetics of Hexose Formation from Trioses in Alkaline Solution1. *J. Am. Chem. Soc.* **73**, 2054–2057 (1951).
25. 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).
26. 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).
27. 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).

28. Breslow, R. On the mechanism of the formose reaction. *Tetrahedron Lett.* **1**, 22–26 (1959).
29. Appayee, C. & Breslow, R. Deuterium Studies Reveal a New Mechanism for the Formose Reaction Involving Hydride Shifts. *J. Am. Chem. Soc.* **136**, 3720–3723 (2014).
30. 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).
31. Horowitz, J. M. & England, J. L. Spontaneous fine-tuning to environment in many-species chemical reaction networks. *Proc. Natl. Acad. Sci.* **114**, 7565–7570 (2017).
32. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
33. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
34. 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).

Chapter 5

Fluctuations in the environment direct the organization of prebiotic reaction networks

Prebiotic environments were dynamic, containing a range of periodic and aperiodic variations in reaction conditions. However, the impact of temporal environmental changes upon prebiotic chemical reaction networks has not been investigated.

This chapter demonstrates how the magnitude and rate of temporal fluctuations of the catalysts Ca^{2+} and OH^- control the product distributions of the model prebiotic formose reaction. Surprisingly, the product compositions of the formose reaction under dynamic conditions deviated significantly from those under steady-state conditions. These compositional changes are attributed to the non-uniform and non-linear propagation of fluctuations through the network, thereby shaping reaction outcomes. An examination of temporal concentration patterns shows that collections of compounds respond collectively to perturbations. This indicates that key 'gating' reactions branching from the Breslow cycle are important responsive features of the formose reaction.

These findings illustrate how the compositions of prebiotic reaction networks are shaped by dynamic environmental events, illustrating the necessity for considering the temporal traits of prebiotic environments that supported the origin of life.

Parts of this chapter have been published in:

1. **P. van Duppen**, E. Daines, W.E. Robinson, W.T.S. Huck, *J. Am. Chem. Soc.*, **145**, 7559-7568 (2023).

5.1 Introduction - Dynamic environments on a prebiotic earth

Understanding how prebiotic chemical systems evolved toward primitive life remains a major scientific challenge. Many prebiotically plausible synthetic pathways towards key building blocks of living systems have been reported.[1–6] Together they provide an outline of the compositional landscape accessible from prebiotic precursors.[7] During the evolutionary process toward life, these building blocks were incorporated in available prebiotic chemical reaction networks destined to form the various metabolic and replicative networks necessary for primitive life.[4,8–13] Prebiotic chemical reactions are typically performed in a series of well-defined steps at constant pH, temperature, or mineral composition which mimic a prebiotic environment.[1–3,6,9–11,14,15] In chapter 4, it was shown that the environment is a key directing force for the self-organization of the model prebiotic formose reaction network. However, prebiotic environmental conditions would have been dynamic, just as they are on modern earth. These varied on different of timescales, such as the diurnal cycle or more erratic weather patterns. This work represents the first steps in understanding how fluctuations in the environment impact on prebiotic chemical reaction networks.

As discussed in chapter 4, product compositions of the formose reaction are the result of inherent interactions between chemical reactivity and the environment.[16] From prebiotic feedstock molecules formaldehyde (**1**) and dihydroxyacetone (**3**), together with catalytic Ca^{2+} and OH^-, a wide range of compositions are formed consisting of mostly monosaccharides.[16–21]

In this chapter the effect of time-dependent input variables on the composition and reaction connectivity in the formose reaction was studied by continually varying fluctuations in the concentrations of Ca^{2+} and OH^-. I will discuss the effect of varying the magnitude and the rate of change in these dynamics and how these propagate through the network. Different groups of compounds responded in unison to dynamic inputs. I will elaborate how the presence of responsive 'subnetworks' exhibit a striking similarity to biochemical reaction pathways,[22] thus hinting at how the environment could have played a role in the evolution of network structures in prebiotic systems.

5.2 Compositional shifts forced by environmental dynamics

To provide an out-of-equilibrium system to which environmental dynamics can be applied, the formose reaction was performed in a CSTR, see figure 5.1.[16] The

experimental conditions explored in this investigation are summarized in table 5.8.4. Stepwise varying input patterns of NaOH and $CaCl_2$ were applied to the system by changing their input flow rates. Flow rates were varied to maintain a 1 : 2 ratio of Ca^{2+} to OH^-. The applied environmental dynamics will be referred to as the net fluctuation of $Ca(OH)_2$. Flow rates were sampled from normal distributions around averages of $[NaOH]_{in}$ = 30 mM and $[CaCl_2]_{in}$ = 15 mM with varying magnitudes ($\sigma_{[Ca(OH)2]in}$ = 0.00, 2.89, and 5.75 mM). Two time-intervals between changes in the flow rate were used (45 and 120 s). The flow rate of a syringe containing water was simultaneously adjusted against the flows of $CaCl_2$ and NaOH solutions to maintain a constant residence time of 2 min.

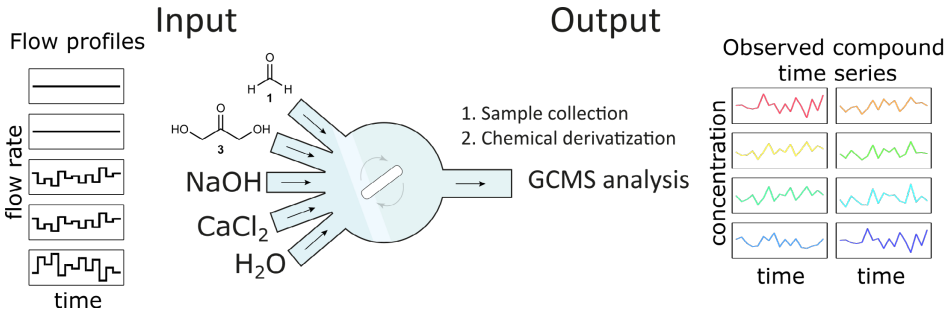


Figure 5.1: Examples of an input for Ca^{2+} and OH^- concentration profile and output carbohydrate concentration profiles (in color). For each experimental condition 50 datapoints were collected in which the concentration of 28 compounds was determined from GC-MS chromatograms.

The influence of temporal fluctuations on formose reactions was studied at a fixed input concentration of **3** (50 mM) and three different input concentrations of **1** (20, 50, and 100 mM). These conditions cover previously identified organizational patterns which are dominated by either aldol addition reactions of C_2 and C_3 carbohydrates ($[\mathbf{1}]_{in}$ = 20 mM)[16,17,21,23] or by addition reactions of **1** to sugar enolates in the network ($[\mathbf{1}]_{in}$ = 100 mM)[16,18] as well as a condition intermediate between the two ($[\mathbf{1}]_{in}$ = 50 mM), see 4.2.1.[16] For each experiment, 50 samples were collected from the output of the CSTR at 40.8 s intervals for EXP001−EXP009 and at 30.6 s intervals for EXP010−EXP013. Samples were analyzed by GC-MS, as discussed previously (chapter 2.2 – 2.3). Each peak was assigned to a compound *via* a combination of retention time and mass spectral fragmentation pattern (see 2.4) and using authentic sample calibrations for linear chain sugars (see 4.7.3 and 5.8.5). In total, 28 compounds were identified, see figure 5.2.

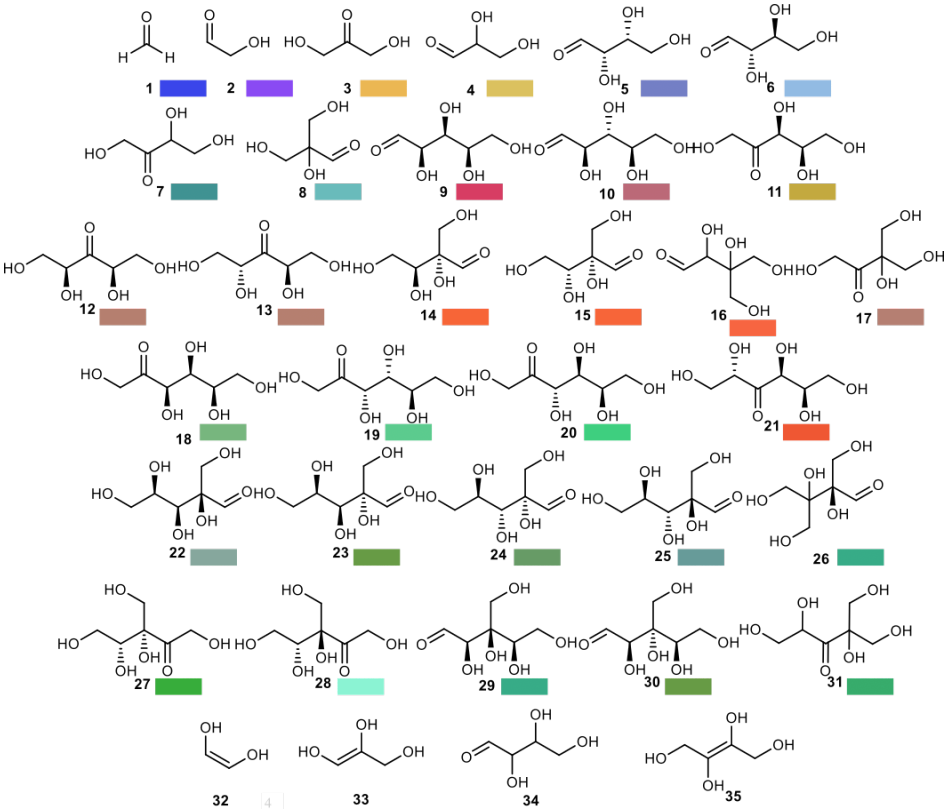


Figure 5.2: **The molecular structure of the different compounds** in the dataset discussed in this chapter. Each compound corresponds has a color hue assigned, enolate species (**32 – 35**) were not assigned a color hue.

5.2.1 Effect of the magnitude in variation of [Ca(OH)₂] upon composition of the formose reaction

The effect of amplitude in the variation of $[Ca(OH)_2]$ ($\sigma_{[Ca(OH)2]}$) on the concentrations of a number of formose products is shown in figure 5.3 (for all compounds, see 5.8.7: fig. S5.10−S5.15). Surprisingly, despite the system having the same average input concentration of $Ca(OH)_2$ compared to an unperturbed reaction, the measured concentration distributions vary significantly (at least $p \leq 1 \times 10^{-4}$ in each series for compounds **3**, **5**, **7**, and **11**) in response to fluctuations in $[Ca(OH)_2]_{in}$. Please note that in the absence of applied input dynamics, some compounds (for example, **3**, **5**, and **7**, fig. 5.3a−c, respectively; $[\mathbf{1}]_{in} = 50$ mM) also exhibited a relatively broad distribution of concentrations (for example, **11**, fig. 5.3d). This variation is due to experimental noise introduced by factors which include sample handling (time taken to freeze-quenching or different steps in

the chemical derivatization process). This variance is constant across the data set.

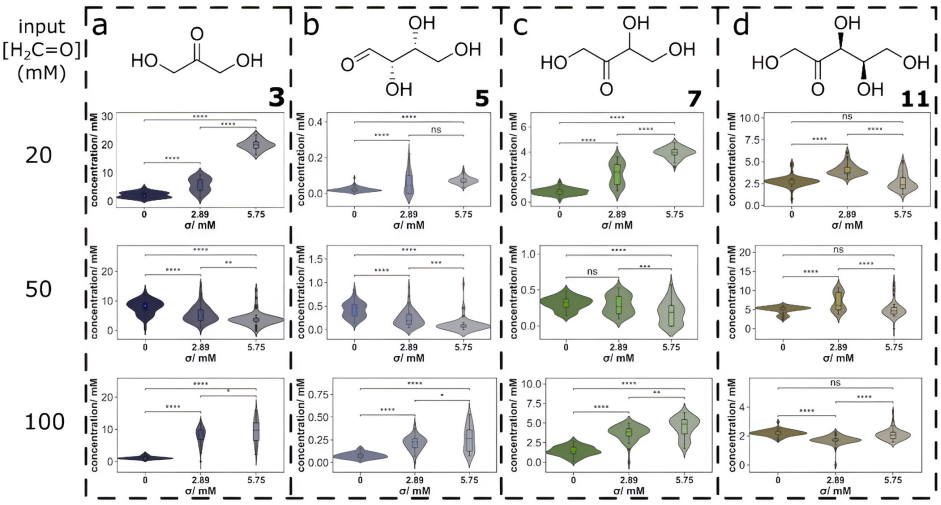


Figure 5.3: Concentration distributions of C_3, C_4 and C_5 sugars in the formose reaction with varying $Ca(OH)_2$ fluctuations. The distributions of a: DHA (**3**), b: threose (**5**), c: erythrulose (**7**) and d: xylulose (**11**) for varying concentrations of $[formaldehyde]_{in}$ perturbed with $[Ca(OH)_2]_{in}$ of various amplitudes (σ = 0, 2.89, 5.75 mM). p-values were calculated as indicators for the significance of the difference between means and are annotated as follows: ns ($5\times10^{-2} < p \leq 1$), * ($1\times10^{-2} < p \leq 5\times10^{-2}$), ** ($1\times10^{-3} < p \leq 1\times10^{-2}$), *** ($1\times10^{-4} < p \leq 1\times10^{-3}$),**** ($p \leq 1\times10^{-4}$).

The applied perturbations altered not only the average concentration of each compound, but also the distribution and the range of concentrations that can be accessed by each compound. Furthermore, the sensitivity and trends in response to formose reaction product concentrations to the applied dynamics were dependent on the concentration of **1**. A range of responses were observed across the conditions investigated (see 5.8.7: fig. S5.10–S5.15), hinting at the nonlinear behavior of the underlying formose reaction pathways. Figure 5.3 depicts four case studies which illustrate the breadth of this behavior. At $[\mathbf{1}]_{in}$ = 20 mM, dihydroxyacetone (**3**, fig. 5.3a) and erythrulose (**7**, fig. 5.3c) exhibited significant increases in the average concentration in response to the increasing $\sigma_{[Ca(OH)2]}$. On the other hand, the distribution of threose (**5**, fig. 5.3b) only responded significantly to increasing $\sigma_{[Ca(OH)2]}$ from 0.0 to 2.89 mM, but not between 2.89 and 5.75 mM. Xylulose (**11**, fig. 5.3d) responded significantly to $\sigma_{[Ca(OH)2]}$ = 2.98 mM, but its concentration distributions were similar between $\sigma_{[Ca(OH)2]}$ = 0.00 and 5.75 mM. Increasing the concentration of **1** to 50 mM lowered

the significance in the changes induced by increasing $\sigma_{[Ca(OH)2]}$ for **3** and **7**, while **5** became more responsive to higher magnitudes of input dynamics. **11**'s responses remained relatively similar to those observed with $[\mathbf{1}]_{in}$ = 20 mM. Further increasing the concentration of **1** to 100 mM again resulted in little effect on the response of **11** to varying $\sigma_{[Ca(OH)2]}$. The responses of **3** and **5** became more similar than those at lower concentrations of **1**. Significant changes in their mean values were observed when dynamics were applied, but little change in their means is seen on increasing $\sigma_{[Ca(OH)2]}$ to 5.75 mM due to their high variance. **7** regained its sensitivity which was absent at $[\mathbf{1}]_{in}$ = 50 mM albeit with lower sensitivity than when the concentration of $[\mathbf{1}]_{in}$ = 20 mM.

5.2.2 The Rate of Variation in $[Ca(OH)_2]$ Affects the Composition of the Formose Reaction

Fluctuations in the environment can occur at multiple timescales (for example, the terrestrial diurnal or tidal cycles are slower than precipitation or wind). Therefore, the response of the formose reaction was explored for varying perturbation frequencies. The compositional changes of the formose reaction, using $[\mathbf{1}]_{in}$ = 50 mM under dynamic conditions with varying rates of $Ca(OH)_2$ input fluctuations (45 and 120 seconds, see figure 5.4), were compared to the composition at steady-state input. The average input concentration and magnitude of variation ($\sigma_{[Ca(OH)2]}$) were fixed. Again, the average concentrations of compounds deviated from steady-state averages (**3**, **5**, **7**, and **11**; $p \leq 1 \times 10^{-4}$, in all compound series; see 5.8.7: fig. S5.16−S5.17). Varying the rate of environmental change also has marked effects on the concentration distributions of formose products (fig. 5.4). The concentration of erythrulose (**7**) has the same standard deviation, but lower average, when the input of $Ca(OH)_2$ was varied at 45 s ([**7**] = 4.8 ± 0.5 mM) intervals than at a steady-state ([**7**] = 6.0 ± 0.5 mM). However, when the frequency of variation was 120 s, the average concentration of **7** was lower relative to sample at 45 s variation albeit with a higher standard deviation (4.2 ± 0.7 mM).

These compositional trends show how the formose reaction is sensitive to the rate of fluctuations in $Ca(OH)_2$ input concentrations, leading to significant, but non-uniform, changes in concentrations of reaction products.

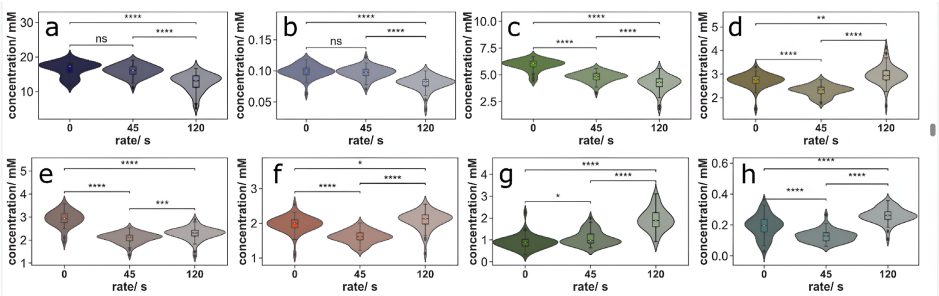


Figure 5.4: **The distributions of compounds a: 3, b: 5, c: 7, d: 11, e: 13, f: 15, g: 23** and **h: 25** for the [formaldehyde]$_{in}$ = 50 mM network at steady-state and perturbed with at a 45 s or 120 s rate of change ([Ca(OH)$_2$]$_{in}$ σ = 5.75 mM). p-values were calculated as indicators for the significance of the difference between means and are annotated as follows: ns ($5\times10^{-2} < p \leq 1$), * ($1\times10^{-2} < p \leq 5\times10^{-2}$), ** ($1\times10^{-3} < p \leq 1\times10^{-2}$), *** ($1\times10^{-4} < p \leq 1\times10^{-3}$) **** ($p \leq 1\times10^{-4}$).

5.3 Collective responses in the network relate to time scales in the input signal

The grouped responses of compounds (for example, compounds **3**, **5**, and **7** in fig. 5.4) for different Ca(OH)$_2$ dynamics indicate that the formose reaction network may behave as a collection of subnetworks. To explore the transfer of environmental fluctuations through the network in greater depth, the formose reaction was perturbed using a superposition of input dynamics. A dynamic input was constructed from the combination of three signals made of values selected at random from normal distributions ($\mu_{[Ca(OH)2]}$ = 15 mM, $\sigma_{[Ca(OH)2]}$ = 5.75 mM) which varied at three different frequencies (30, 60, and 120 seconds), [**1**]$_{in}$ = 50 mM, see figure 5.5.

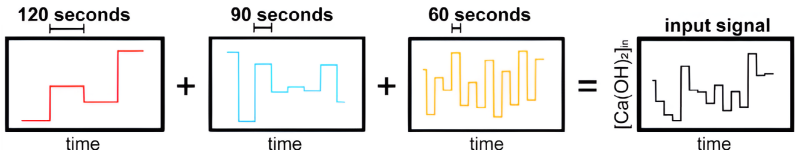


Figure 5.5: **The superposition of three timescales of variation in Ca(OH)$_2$ input reveal collective responses of compounds to conditional dynamics in the formose reaction.** The input consisted of variations on the scales of 60, 90 and 120 sseconds

5.3.1 Hierarchical clustering reveals partitioning of different parts of the network

A hierarchical clustering analysis was performed on the time−concentration traces from this experiment. The cluster analysis revealed the partitioning of the formose reaction products into five (**I**−**V**) clusters, see figure 5.6. Within each

cluster, compounds responded to the changes in the environment in a similar way.

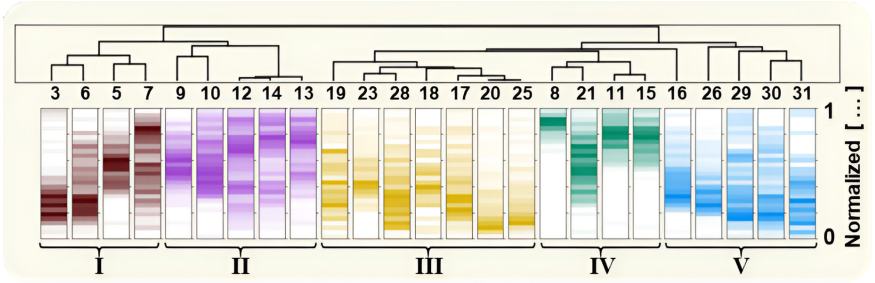


Figure 5.6: **Hierarchical clustering analysis of the concentration-time traces** reveals partitioning into five groups of compounds (**I – V**) as indicated.

5.3.2 Different groups of compounds had a unique response to the embedded time scales in the input signal

A time averaged-correlation analysis was performed to uncover which embedded timescales in the input were transferred to the various compounds. Clear differences in correlation to the input at varying timescales were observed between the clusters, see figure 5.7 (see 5.7.3 and figure 5.9 for a detailed description of the time interval averaging analysis). The correlations were superimposed on the five identified clusters (**I**–**V**, fig. 5.7). Positive (negative) correlations indicate that changes in concentrations of products move in the same (opposite) direction as the input fluctuation.

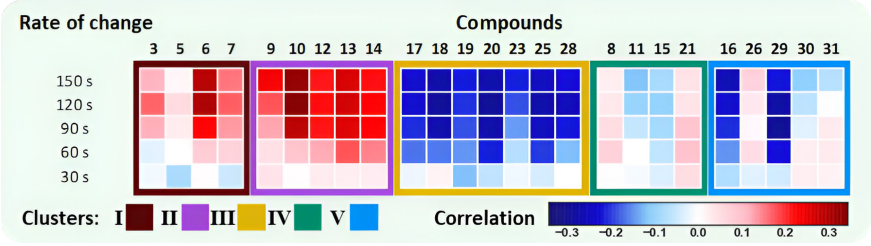


Figure 5.7: **The correlation of each compound trace to input dynamics on five different time scales.** For each compound in this network the time-interval correlation of the input and output differential was calculated over concentration difference at different time scales in the $[Ca(OH)_2]_{in}$ (150 s, 120 s, 90 s, 60 s, 30 s).

The compounds in cluster **I** (**3**, **5**, **6**, and **7**) have a strong positive correlation to the input signal on longer (150 s) input timescales (fig. 5.7). Conversely, an increasingly weaker correlation is seen for the shorter time scales. Cluster **II** exhibits a similar trend, with even stronger positive correlations to the longer

input timescales. A strong negative correlation to the input is seen for cluster **III** at all but the shortest timescales (30 s). Compounds in clusters **IV** and **V** broadly show a weaker overall correlation with the input dynamics in comparison to the other clusters. Compounds **16** and **29** are exceptions to this trend, especially on longer timescales.

5.4 The transfer of dynamics is governed by the reaction network structure

A closer inspection of the clusters identified in figure 5.6, reveals that compounds typically associated with the Breslow cycle[24,25] (**3**, **5**, **6**, and **7**) remain within cluster **I** across all the variations in input conditions described above, see figure 5.8a. These compounds also show a strong positive correlation with the input fluctuations and are also most sensitive to environmental perturbations, having higher standard deviations with respect to their average concentration compared to compounds in other clusters (see 5.8.7: fig. S5.10−S5.17).

Key to understanding the dynamics in the other clusters observed in the data are four enolate species (C_2-enolate **32**, C_3-enolate **33**, C_4-1,2-enolate **34**, and the off-cycle C_4-2,3-enolate **35**). These enolates control the 'gating' reactions that connect the central cycle to the identified clusters of compounds (**II**−**V**) (fig. 5.8a). The C_4-1,2-enolate **34** is involved in reactions which gate access to cluster **II**, *via* aldol addition reactions with **1** (fig. 5.8b). The products in cluster **III** form in addition reactions which involve the C_3-enolate **33** and the off-cycle C_4-enolate **35** (fig. 5.8c). Similar to cluster **III**, access to cluster **IV** occurs *via* the C_3-enolate **33** which reacts with **2** to produce xylulose (**11**) (fig 5.8d). Cluster **V** is connected to the central cycle *via* aldol addition reactions between C_2-enolate **32** and **3** or **7** (fig. 5.8e).

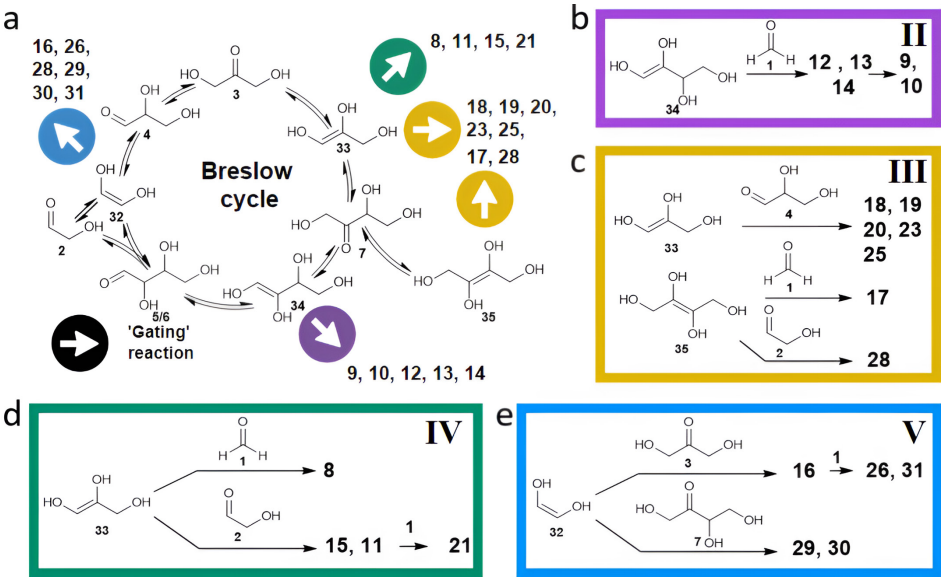


Figure 5.8: The identified clusters can be rationalized with 'gating' reactions stemming from the central Breslow cycle. a) The $[\mathbf{1}]_{in}$ = 50 mM network, perturbed with multiple embedded rates (120 s, 90 s, 30 s), represented around the core Breslow cycle. The 'gating' reactions are indicated with colored arrows. b – e) The 'gating' reactions to produce the products in the respective cluster (**II** – **V**).

The varying connectivity of the clusters to the central cycle also explains the difference in the correlation strength for the different clusters. 'Gating' reactions have varying degrees of sensitivity to $[Ca(OH)_2]$ fluctuations. As a result, they mediate the transfer of environmental dynamics to varying degrees to the compounds created downstream from them. Here, the Breslow cycle behaves as a central hub for the propagation of environmental fluctuations. Reactions with higher sensitivities allow for the transfer of faster conditional fluctuations compared to those with lower sensitivities, which only allow the propagation of fluctuations on longer timescales.

5.5 Conclusion

In summary, this study has shown how fluctuations in the environment are translated into product distributions *via* dynamic propagation through the formose reaction network. Rather than each compound responding uniquely to applied dynamics, collections of compounds respond together due to the structure of the underlying reaction network. Interestingly, this structure is attributed to four 'subnetworks' of the formose reaction, which connect to a central Breslow cycle *via* as many 'gating' reactions. These 'gating' reactions

control the propagation of environmental fluctuations through the network, thus producing the observed grouped behavior. It was found that for the evolution of metabolic modules in extant life is enhanced by environmental variation a species is subjected to.[26] The grouped response in different parts of the formose reaction resembles the modular build-up in extant metabolic networks and might be a hint of its primordial origin.[22]

The adaptation of reaction networks to fluctuations in the environment has important implications for the field of prebiotic chemistry.[27,28] Dynamic conditions must be considered in modelling prebiotic chemical processes and offer a mechanism through which complex chemical reactions may inherit information from the environment. The temporal signatures of reaction conditions offer an extra level of control on top of their average magnitudes in guiding the reaction outcomes of prebiotic reaction networks. Multiple reaction outcomes may be accessed from a single set of precursors, dependent on the rate and magnitude of variation of each parameter affecting the reaction. Time-dependent conditional variations can be used as a tool in controlling the compositions of prebiotic chemical reactions by manipulating the operational reaction pathways.

These results and interpretation demonstrate that there is a mechanistic basis to the compositional changes induced by dynamic environmental conditions. Reaction modularity offers a design principle for future investigations to target specific reactions and compounds. Thus, future experimental designs can be based on the promotion or inhibition reaction pathways in a mechanistic manner. Such investigations will provide a stronger conceptual union between the chemical properties of prebiotic reaction networks and environmental dynamics, thus unveiling mechanistic pathways to the evolution of life's first biochemical processes.

5.6 Method summary

5.6.1 Materials

1,3-Dihydroxyacetone was purchased from Carbosynth Ltd., $CaCl_2$, NaOH, paraformaldehyde, O-ethylhydroxylamine hydrochloride, N,O-bis(trimethylsilyl)trifluoroacetamide 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. All other chemicals were used without further purification.

5.6.2 Instrumentation

GC-MS analysis was performed on a *JEOL JMS-100GCv* attached to an *Agilent 7890A GC* with the procedure as described under 2.3.3.

5.6.3 Methods

Flow reactions were performed in a CSTR as described under 2.2 and 2.6.1.

Derivatization for GC-MS analysis was performed as described under 2.3.1.

Chromatographic data processing was performed with the Python program ChromProcess, authored by W.E. Robinson, as described under 2.4 and 2.6.4.

Estimation of reaction pathways from experiments with fluctuations of $[Ca(OH)_2]$ is performed as described under 3.4.2.

5.7 Data analysis methods

Python programs for the following described data analysis are available at https://github.com/huckgroup/Formose_2022.git. Here, also the extracted data from the GC-MS analysis was available under Extended_Data_GH.

All data analyses were performed using the Numpy[29] (1.22.2.) and SciPy[30] (1.8.0) Python libraries.

5.7.1 Mass distribution t-test analysis

To test if two concentration distributions from different experimental conditions came from the same concentration distribution, a t-test was used. This analysis compares the mean of the two samples. The time concentration traces x of a compound from two distinct experimental conditions were compared (using scipy.stats.ttest_ind($x_{cond\ 1.}$, $x_{cond\ 2.}$)). In the violin plots (fig. 5.3-4, S5.9-S5.17) we used an annotation legend to indicate the statistical significance: ns: $5 \times 10^{-2} < p \leq 1$, *: $1 \times 10^{-2} < p \leq 5 \times 10^{-2}$, **: $1 \times 10^{-3} < p \leq 1 \times 10^{-2}$, ***: $1 \times 10^{-4} < p \leq 1 \times 10^{-3}$, ****: $p \leq 1 \times 10^{-4}$.

5.7.2 Hierarchical clustering analysis dynamic response and reaction pathway reconstruction

The hierarchical cluster analysis was performed on all experimental conditions that were perturbed with $Ca(OH)_2$ fluctuations (fig. 5.6). For the cluster analysis of an experimental condition, an array of vectors (x) was constructed consisting of the time series data for each compound. Pairwise distances based on the '*correlation*' metric were calculated from this array (x) to construct a *distance matrix* with scipy.spatial.distance.pdist(x) (Equation 1).

Eq. 1: $distance = \frac{(u - \bar{u}) \cdot (v - \bar{v})}{||(u-\bar{u})||_2 \; ||(v-\bar{v})||_2}$

Where u and v are vectors of the concentration time profiles in array x for compounds, $\bar{u}$ and $\bar{v}$ are the means of u and v, respectively, and $x \cdot y$ is the vector dot product of x and y.

A hierarchical cluster analysis was performed on the *distance matrix*, by creating a *linkage matrix* with scipy.cluster.hierarchy.linkage(*distance matrix*), using the '*average*' method (Equation 2).

Eq. 2: $d(a,b) = \sum_{ij} \frac{d(a[i],b[j])}{(|a| \times |b|)}$

For all points i and j where $|a|$ and $|b|$ are the number of elements in each of the clusters a and b respectively.

In order to visualize the *linkage matrix*, a dendrogram was plotted with scipy.cluster.hierarchy.dendrogram(*linkage matrix*).

The clusters were assigned after visual inspection, see figure 5.6.

5.7.3 Time-interval correlation analysis

Correlation analyses were performed on discrete differentials at varying time intervals (30, 60, 90, 120, 150 s) of concentration time traces of the $Ca(OH)_2$ input against the measured outputs (compound: **3**, **5 – 21**, **23**, **25**, **26**, **28 – 31**). First, the input of $Ca(OH)_2$ was resampled by linear interpolation so that it had the same time axis as the output data. Sliding time windows (30, 60, 90, 120, 150 s) were passed along the time progresses of the input and compound time-concentration profiles, see figure 5.9. The mean of the values within each time window was subtracted from the mean of the values in the previous window to create a vector of values indicating how each signal changes on average over the scale of the time window (x). One such vector was created for each time window width, for each input and compound time-concentration profile. This process is illustrated in figure 5.9. The values in x were means-centered (x – np.mean(x)) and normalized by dividing by the standard deviation with numpy.std(x) (Equation 3).

Eq. 3: $v = \frac{x - \bar{x}}{\sigma_x}$

Where v is the means-centered and scaled differential vector at a specific time interval, $\bar{x}$ is the mean of vector x and σ_x the standard deviation of x.

For each time window (30, 60, 90, 120, 150 s), the Pearson correlation coefficient was calculated between the averaged and differentiated input (v_{in}) and each of the corresponding output averaged differentials (v_{out}) using scipy.stats.pearsonr(v_{in} , v_{out}) (Equation 4).

Eq. 4: $$r = \frac{\Sigma(v_{in} - \overline{v_{in}})(v_{out} - \overline{v_{out}})}{\sqrt{\Sigma(v_{in} - \overline{v_{in}})^2 \Sigma(v_{out} - \overline{v_{out}})^2}}$$

Where $\overline{v_{in}}$ and $\overline{v_{out}}$ represent the mean value of v_{in} and v_{out}, respectively. A positive linear correlation is indicated by a Pearson correlation coefficient between 0 and 1 and an inverse linear correlation is indicated Pearson correlation coefficient between –1 and 0. A Pearson correlation coefficient 0 indicates that there is no linear correlation between the input and output.

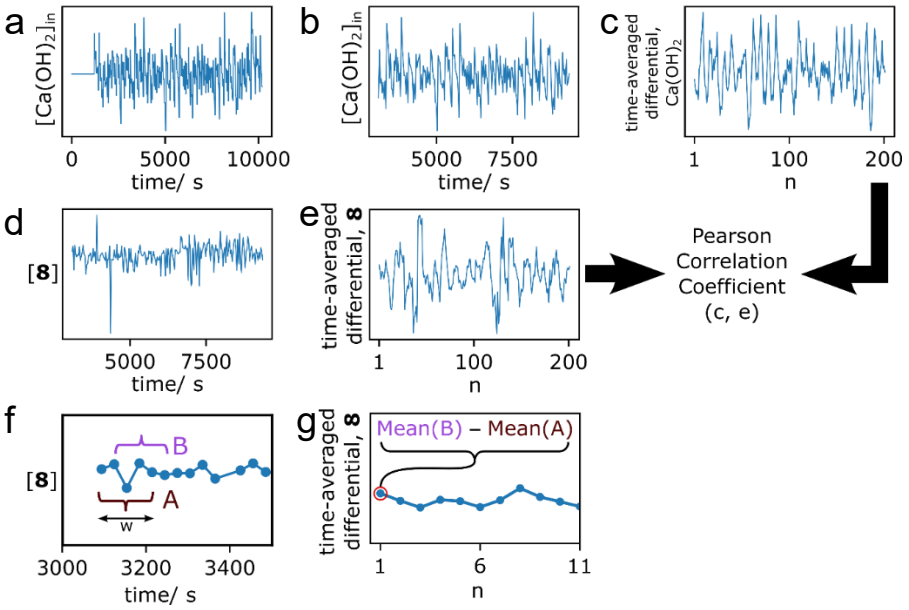


Figure 5.9: Illustration of the calculation of the time-averaged correlation of a compound (**8**) time course to the dynamic input in EXP013 (see table in 5.8.4). This process produced one cell of figure 5.7. The dynamic input of $Ca(OH)_2$ (**a**) is resampled by linear interpolation (**b**) to share the same time points as the experimentally determined concentration time courses (**d**). A sliding time-window (width, w = 1 - 5 time points, see **f**) is passed across the input (**b**) and concentration time course (**d**). The mean of each window is subtracted from that of the previous window (**g**) to produce a time-average, differentiated signal for the input and concentration time course (**c** and **e**, respectively). The Pearson correlation coefficient is calculated between the input (**c**) and concentration (**e**) traces to yield a value which can be mapped to the grid in figure 5.7.

5.8 Supplementary information

5.8.1 Selection of reaction conditions

Previous work has shown that the input concentration of formaldehyde controls a compositional transition between 0 – 100 mM with $[DHA]_{in}$ = 50 mM, $[NaOH]_{in}$ = 30 mM and $[CaCl_2]_{in}$ = 15 mM.[16] The average input concentrations of formaldehyde, dihydroxyacetone, NaOH and $CaCl_2$ are selected from within this space of parameters. Thus, the chosen conditional ranges provide access to a range of compositional outcomes.

5.8.2 Selection of $Ca(OH)_2$ amplitude series

Dynamic inputs of $[Ca(OH)_2]$ were generated by sampling from normal distributions (μ = 15 mM, σ = 2.89 mM or 5.75 mM) generated using Numpy[29] (1.22.2.) with function numpy.random.normal(). For the amplitude series $[NaOH]_{in}$ and $[CaCl_2]_{in}$ were varied around the steady-state average (15 mM for $CaCl_2$ and 30 mM for NaOH) with varying amplitude randomly sampled from a normal distribution, (σ_{in} = 2.89 mM or 5.75 mM) of concentrations, varied via syringe flow rates as described in the Experimental Methods section Main Text. The magnitudes of variation were chosen such that they would perturb the reaction mainly via the applied dynamics, rather than via a bulk change in concentration. Flow rates where switched every 45 seconds, thus not allowing the catalyst or reaction to reach steady-state to maintain a constant compositional variation. Note that NaOH and $CaCl_2$ syringes were not allowed to produce negative flow rates.

5.8.3 Selection of multiple frequency input signal

To construct the input signal containing variations on three timescales, the range 30 -120 s was chosen. The residence time of the experiment was chosen as an upper bound (120 s), with a lower bound of 30 s to ensure that the applied dynamics could be sampled on a similar timescale.

5.8.4 Table experimental conditions

All experiments were performed with fixed input concentrations of formaldehyde ([formaldehyde]$_{in}$) as indicated and dihydroxyacetone (50 mM). Input concentrations of $CaCl_2$ were sampled from a Gaussian distribution with average 15 mM and standard deviations as indicated ($\sigma_{Ca(OH)2,\ in}$) and applied via modulation of the input flow rate of $CaCl_2$ solution with a frequency indicated by the step rate. The input concentration (flow rate) of NaOH was varied in unison with the input of $CaCl_2$ to maintain a constant 1 : 2 ratio of Ca^{2+} : HO^-. The flow rates applied to EXP013 were created from a linear combination of concentrations selected from three independent Gaussian distributions (mean = 15 mM), which varied over three different time scales as indicated. The flow rate of a separate water input was varied to offset the varying flows of $CaCl_2$ and NaOH to maintain a constant residence time (120 s). Temperature: 21 °C.

Experiment	$\sigma_{Ca(OH)2,\ in}$/ mM	Step rate/ s	$[1]_{in}$ / mM
EXP001	0.00	0	50
EXP002	2.89	45	50
EXP003	5.75	45	50
EXP004	0.00	0	20
EXP005	2.89	45	20
EXP006	5.75	45	20
EXP007	0.00	0	100
EXP008	2.89	45	100
EXP009	5.75	45	100
EXP010	0.00	0	50
EXP011	5.75	120	50
EXP012	5.75	45	50
EXP013	5.75	30;60;120	50

5.8.5 Table GC-MS calibration for quantitative analysis

Constants for fitted quadratic GC calibration curves for C_4, C_5 and C_6 compounds. The same calibration curves have been used as previously reported[1], based on: $[compound] = A \times (\frac{peak\ integral}{internal\ standard\ integral})^2 + B \times \frac{peak\ integral}{internal\ standard\ integral}$. A calibration curve was estimated for non-calibrated compounds (**8**, **12 – 16**, **21 – 31**) by taking the average calibration curve for the calibrated sugars of similar length.

Compound	A	B
DHA, 3	0.028749	0.118714
threose, 5	0.143626	0.789532
erythrose, 6	-0.38468	2.13775
erythrulose, 7	0.143519	0.222204
ribose, 9	0.339712	1.487932
xylose, 10	0.048971	2.23855
xylulose, 11	0.695714	-0.17935
fructose, 17	0.589499	2.049516
sorbose, 18	0.08	2.354559
tagatose, 19	-0.10622	1.017795
8	0.071759	0.111102
12 to 16	0.229793	0.618952
21 to 31	0.250408	1.677321

5.8.7 Concentration distribution patterns

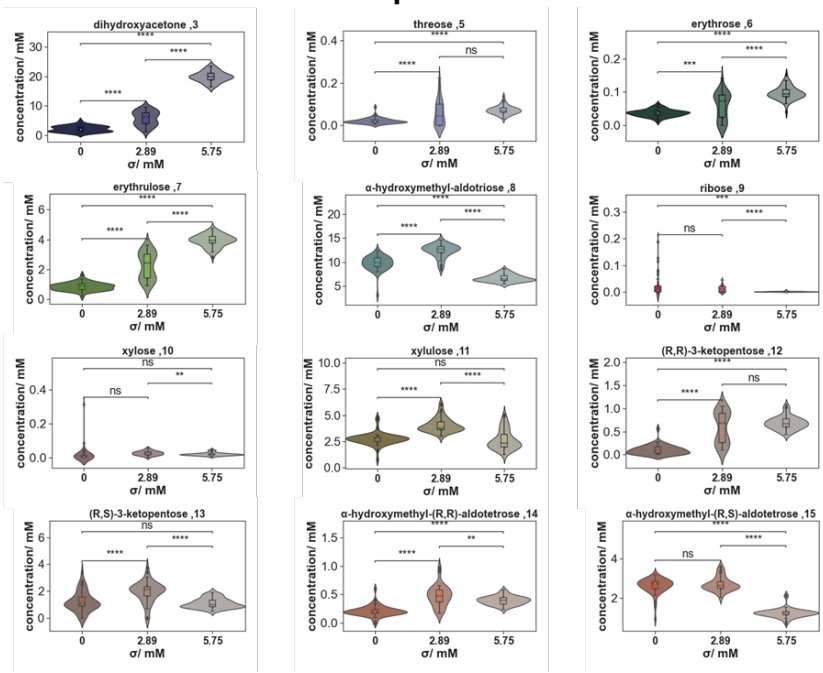


Figure S5.10: **Violin plots of concentration** distributions for identified compounds (**3**, **5**, **6**, **7**, **8**, **9**, **10**, **11**, **12**, **13**, **14**, **15**) at steady-state, low amplitude and high amplitude ($\sigma_{[Ca(OH)2]in}$ = 0, 2.89 and 5.75 mM). Conditions: 20 mM $H_2C{=}O$, 50 mM DHA, 30 mM NaOH, 15 mM $CaCl_2$.

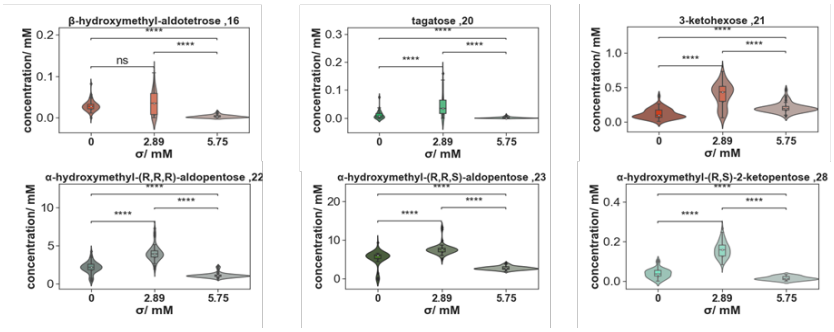


Figure S5.11: Violin plots for concentration distributions for the identified compounds (**16**, **20**, **21**, **22**, **23**, **28**) at steady-state, low amplitude and high amplitude($\sigma_{[Ca(OH)2]in}$ = 0, 2.89 and 5.75 mM). Conditions: 20 mM $H_2C{=}O$, 50 mM DHA, 30 mM NaOH, 15 mM $CaCl_2$.

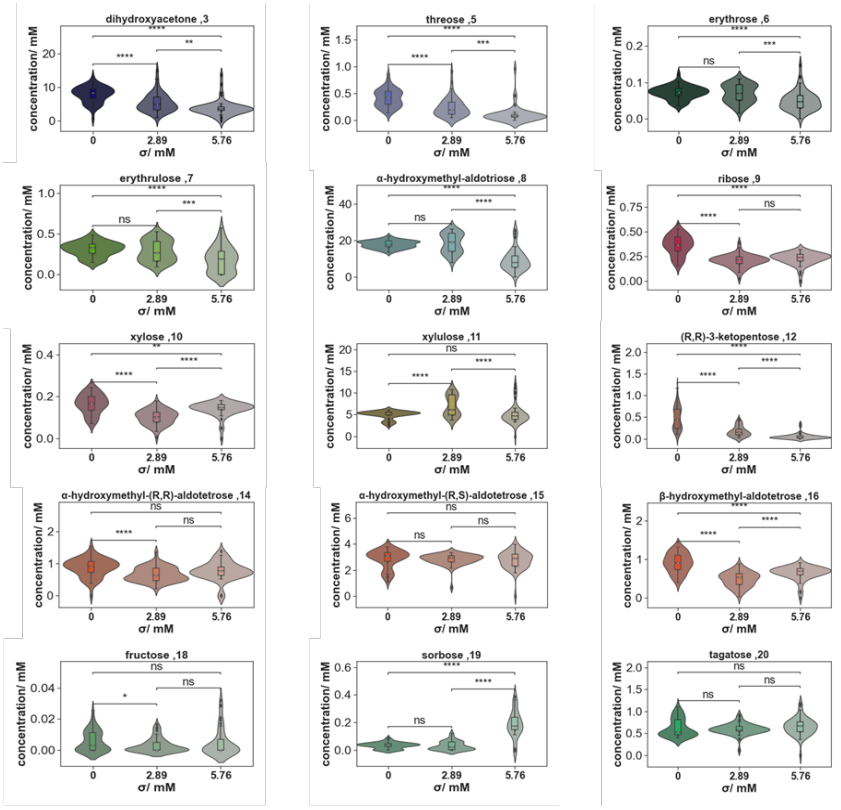


Figure S5.12: Violin plots of the concentration distributions for identified compounds (**3**, **5**, **6**, **7**, **8**, **9**, **10**, **11**, **12**, **14**, **15**, **16**, **18**, **19**, **20**) at steady-state, low amplitude and high amplitude ($\sigma_{[Ca(OH)2]in}$ = 0, 2.89 and 5.75 mM). Conditions: 50 mM $H_2C{=}O$, 50 mM DHA, 30 mM NaOH, 15 mM $CaCl_2$.

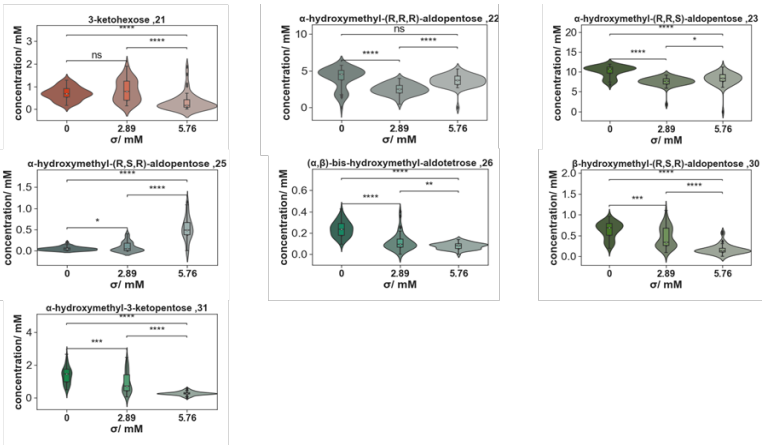


Figure S5.13: **Violin plots for concentration distributions** of the identified compounds (**21**, **22**, **23**, **25**, **26**, **30**, **31**) at steady-state, low amplitude and high amplitude($\sigma_{[Ca(OH)2]in}$ = 0, 2.89 and 5.75 mM). Conditions: 50 mM $H_2C{=}O$, 50 mM DHA, 30 mM NaOH, 15 mM $CaCl_2$.

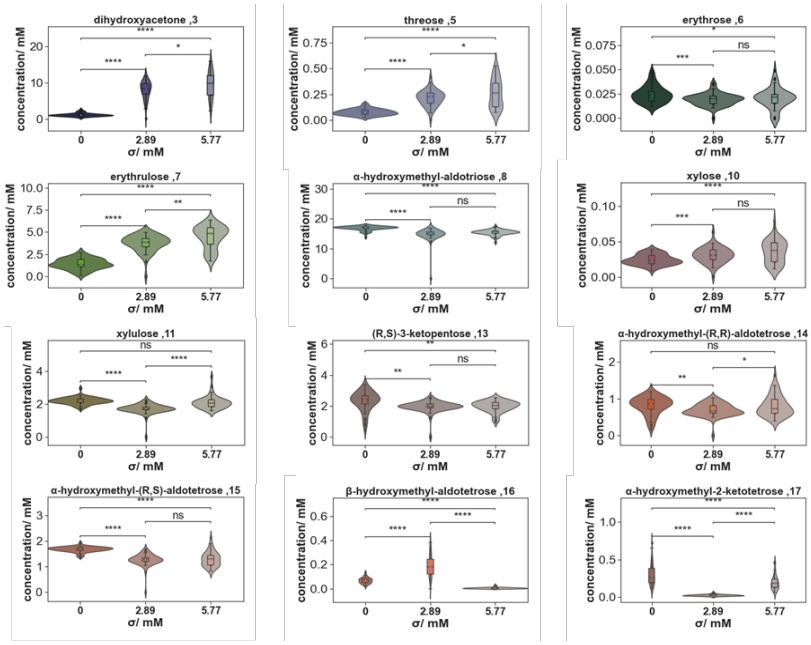


Figure S5.14: **Violin plots for concentration distributions** for the identified compounds (**3**, **5**, **6**, **7**, **8**, **10**, **11**, **13**, **14**, **15**, **16**, **17**) at steady-state, low amplitude and high amplitude ($\sigma_{[Ca(OH)2]in}$ = 0, 2.89 and 5.75 mM). Conditions: 100 mM $H_2C{=}O$, 50 mM DHA, 30 mM NaOH, 15 mM $CaCl_2$.

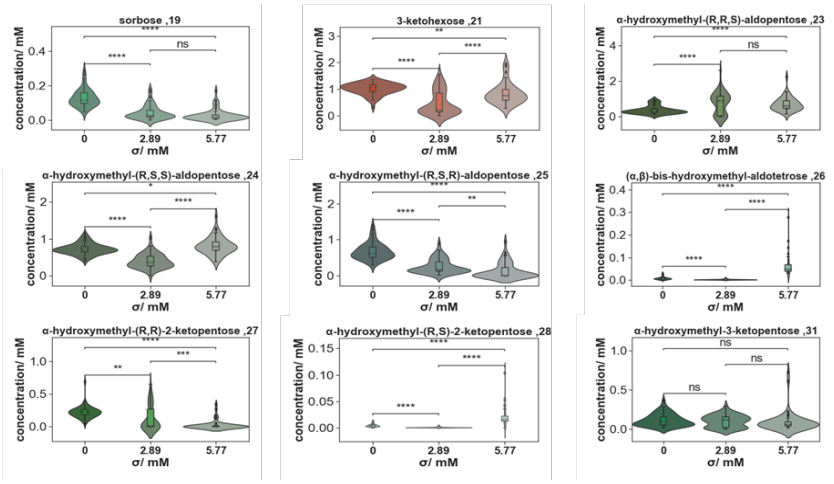


Figure S5.15: **Violin plots for concentration distributions** for the identified compounds (**19**, **21**, **23**, **24**, **25**, **26**, **27**, **28**, **31**) at steady-state, low amplitude and high amplitude ($\sigma_{[Ca(OH)2]in}$ = 0, 2.89 and 5.75 mM). Conditions: 100 mM $H_2C{=}O$, 50 mM DHA, 30 mM NaOH, 15 mM $CaCl_2$.

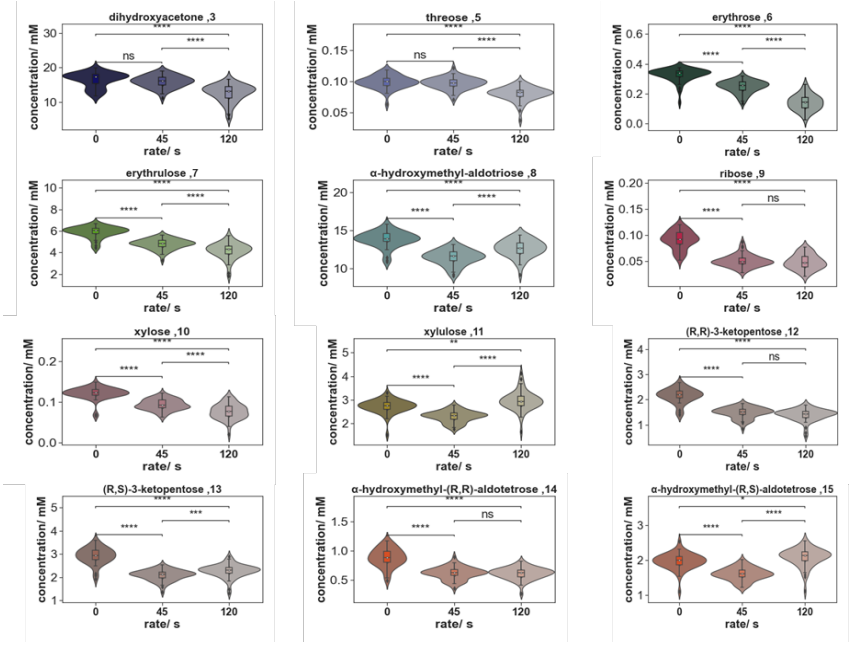


Figure S5.16: **Violin plots for concentration distributions** for the identified compounds (**3**, **5**, **6**, **7**, **8**, **9**, **10**, **11**, **12**, **13**, **14**, **15**) at steady-state, 45 s and 120 s rate of change ($\sigma_{[Ca(OH)2]in}$ = 0, 5.75 and 5.75 mM). Conditions: 50 mM $H_2C{=}O$, 50 mM DHA, 30 mM NaOH, 15 mM $CaCl_2$.

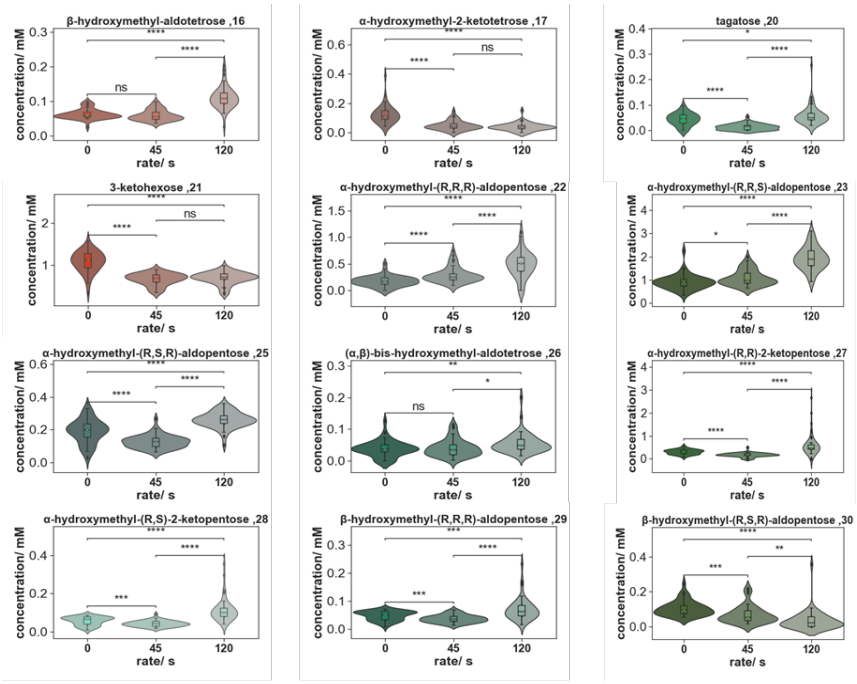


Figure S5.17: Violin plots for concentration distributions for the identified compounds (**16**, **17**, **20**, **21**, **22**, **23**, **25**, **26**, **27**, **28**, **29**, **30**) at steady-state, 45 s and 120 s rate of change ($\sigma_{[Ca(OH)2]in}$ = 0, 5.75 and 5.75 mM). Conditions: 50 mM $H_2C{=}O$, 50 mM DHA, 30 mM NaOH, 15 mM $CaCl_2$.

5.9 References

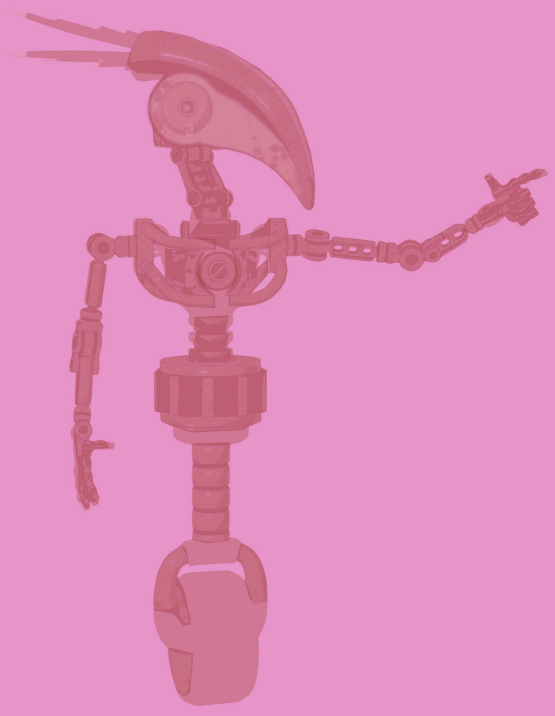
1. Powner, M. W., Gerland, B. & Sutherland, J. D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459**, 239–242 (2009).
2. Ritson, D. J., Mojzsis, S. J. & Sutherland, John. D. Supply of phosphate to early Earth by photogeochemistry after meteoritic weathering. *Nat. Geosci.* **13**, 344–348 (2020).
3. Becker, S. *et al.* Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides. *Science* **366**, 76–82 (2019).
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. Wołos, A. *et al.* Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. *Science* **369**, eaaw1955 (2020).
6. Foden, C. S. *et al.* Prebiotic synthesis of cysteine peptides that catalyze peptide ligation in neutral water. *Science* **370**, 865–869 (2020).
7. Wu, L.-F. & Sutherland, J. D. Provisioning the origin and early evolution of life. *Emerg. Top. Life Sci.* **3**, 459–468 (2019).
8. Braakman, R. & Smith, E. The compositional and evolutionary logic of metabolism. *Phys. Biol.* **10**, 011001 (2012).

9. Muchowska, K. B., Varma, S. J. & Moran, J. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019).
10. Keller, M. A. *et al.* Conditional iron and pH-dependent activity of a non-enzymatic glycolysis and pentose phosphate pathway. *Sci. Adv.* **2**, e1501235 (2016).
11. Sasselov, D. D., Grotzinger, J. P. & Sutherland, J. D. The origin of life as a planetary phenomenon. *Sci. Adv.* **6**, eaax3419 (2020).
12. Busiello, D. M., Liang, S., Piazza, F. & De Los Rios, P. Dissipation-driven selection of states in non-equilibrium chemical networks. *Commun. Chem.* **4**, 16 (2021).
13. Coggins, A. J. & Powner, M. W. Prebiotic synthesis of phosphoenol pyruvate by α-phosphorylation-controlled triose glycolysis. *Nat. Chem.* **9**, 310–317 (2017).
14. Maguire, O. R., Smokers, I. B. A. & Huck, W. T. S. A physicochemical orthophosphate cycle via a kinetically stable thermodynamically activated intermediate enables mild prebiotic phosphorylations. *Nat. Commun.* **12**, 5517 (2021).
15. Becker, S. *et al.* Wet-dry cycles enable the parallel origin of canonical and non-canonical nucleosides by continuous synthesis. *Nat. Commun.* **9**, 163 (2018).
16. Robinson, W. E., Daines, E., van Duppen, P., de Jong, T. & Huck, W. T. S. Environmental conditions drive self-organization of reaction pathways in a prebiotic reaction network. *Nat. Chem.* **14**, 623–631 (2022).
17. Kim, H.-J. *et al.* Synthesis of Carbohydrates in Mineral-Guided Prebiotic Cycles. *J. Am. Chem. Soc.* **133**, 9457–9468 (2011).
18. 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).
19. 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).
20. 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).
21. Gutsche, C. D. *et al.* Base-catalyzed triose condensations. *J. Am. Chem. Soc.* **89**, 1235–1245 (1967).
22. Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461 (2007).
23. Berl, W. G. & Feazel, C. E. The Kinetics of Hexose Formation from Trioses in Alkaline Solution1. *J. Am. Chem. Soc.* **73**, 2054–2057 (1951).
24. Breslow, R. On the mechanism of the formose reaction. *Tetrahedron Lett.* **1**, 22–26 (1959).
25. Appayee, C. & Breslow, R. Deuterium Studies Reveal a New Mechanism for the Formose Reaction Involving Hydride Shifts. *J. Am. Chem. Soc.* **136**, 3720–3723 (2014).
26. Kreimer, A., Borenstein, E., Gophna, U. & Ruppin, E. The evolution of modularity in bacterial metabolic networks. *Proc. Natl. Acad. Sci.* **105**, 6976–6981 (2008).
27. England, J. L. Dissipative adaptation in driven self-assembly. *Nat. Nanotechnol.* **10**, 919–923 (2015).
28. Horowitz, J. M. & England, J. L. Spontaneous fine-tuning to environment in many-species chemical reaction networks. *Proc. Natl. Acad. Sci.* **114**, 7565–7570 (2017).
29. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

30. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

Chapter 6

Perspectives – Evolution of prebiotic reaction networks to the origin of life



Chemical evolution towards life has probably proceeded *via* adaptation of ever more complex systems. The environment provides a key directing force for the self-organization of prebiotic reaction networks. In this thesis, I have discussed how reaction pathways in the formose reaction adapt to different environments. We have seen also how imposed dynamic patterns result in unique responses of reaction pathways. These findings help to formulate key prerequisites to underpin a scenario of chemical evolution towards the origin of life.

In this perspective, I will elaborate on the implications of this study for the field of prebiotic chemistry. The adaptivity of prebiotic reaction networks in dynamic environments might provide important mechanisms for chemical evolution. I will discuss also how new reaction networks can be constructed with the prebiotic formose reaction as a template. Finally, I will conclude by proposing a study for the evolution of network states with an expanded formose reactivity.

6.1 From the formose reaction to the origin of life

In the field of prebiotic chemistry bottom-up strategies are common practice. Synthetic organic chemistry pathways, but also functional reaction systems, have been designed carefully.[1–4] However, this approach fails for larger 'one pot' chemical systems with a high degree of connectivity.[5] Simple prebiotic molecules cover a vast space of chemical reactivity.[6] To study, understand and manipulate these complex prebiotic systems a top-down approach is required.

The formose reaction provides a model prebiotic reaction system. There is no obvious direct relation between reconstructed reaction pathways in the formose reaction and the origin of life. However, this study does provide crucial insights in how recursive prebiotic reaction networks self-organize, directed by interactions with the environment. Furthermore, the addition of time-dependent input conditions adds an extra layer of complexity upon the quest for understanding the origin of life.

This thesis shows the potential role of the environment and its dynamics for controlling chemical reactivity patterns and pathway formation towards unique compositional outcomes of the formose reaction. The Breslow cycle plays a central role in the self-organization of formaldehyde directed reaction networks. The autocatalytic property of the formose reaction under batch conditions was already attributed to this cycle.[9] However, with the production of highly reactive C_2 and C_3 sugars it controls the network reactivity. The respective intermediate enolate species were shown to control signal transduction from dynamic patterns in the environment.

The observed directing forces from the environment will translate especially well to reaction networks with a hierarchical structure. This means, reaction networks constituting a central reactivity where important building blocks are formed from feedstock molecules. The network production of highly reactive molecules could provide an important interface between the chemistry and its environment to drive the process of self-organization. Interaction of these 'core' reactions with dynamics in the environment additionally have the potential to control expression of sprouting reaction pathways. Specific regimes of time-dependent conditional variations may be required for life to evolve from prebiotic environments. The observed modular response of the formose reaction to environmental dynamics might hint at the origin of the modularity in extant biochemical pathways.[10] Environmental dynamics have been proposed as mechanism for evolution of the core carbon metabolism in the absence of genetic inheritance or enzymatic catalysis.[2,3,6–8]

6.2 Network states from unique reaction trajectories

The surprising compositional dependence in the formose reaction to fluctuations in the [$Ca(OH)_2$], as discussed in chapter 5, points at a reorganization of the network. The observed compositional shifts indicate that certain pathways become strengthened relative to others. Although the concentration shift remained within the compositional complexity of the steady-state network, it would be interesting to see if the reaction network can be modified and evolve towards new reactivity patterns.

The reorganization occurs at each change in the [$Ca(OH)_2$], thus providing a different compositional starting point from which the network reacts during the next change in [$Ca(OH)_2$]. As a result, each sequence of changes in the environment lead to a different trajectory and associated compositional outcome. For different network states to evolve from a single steady-state, the order of environmental changes can govern a history-dependent collection of end states, see figure 6.1. The product distribution is highly dependent on the trajectory of fluctuations, or in other words, on the sequence of events.[11,12]

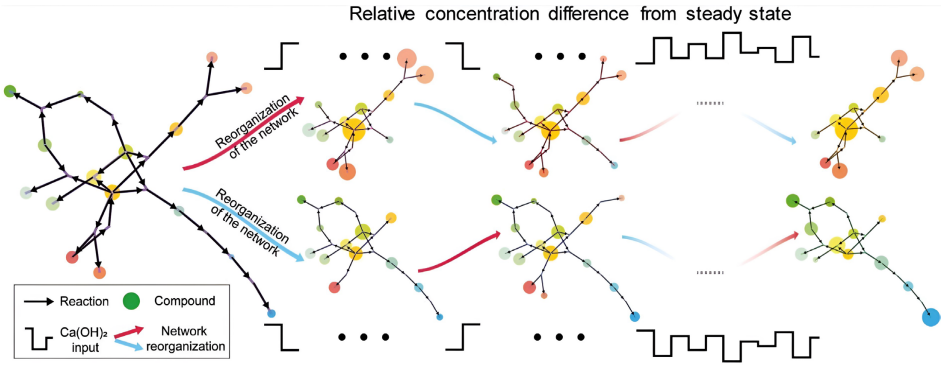


Figure 6.1: State dependence of the formose reaction is governed by dynamics from the environment. The network state defines how it interacts with the environment. The order of events let the system evolve to a transient, but unique end state.

Within the current formose reaction setup, the range of accessible network states is probably low within the range of temporal [$Ca(OH)_2$] dynamics. In chapter 4 it was discussed how each ratio of Ca^{2+} : OH^- exhibit similar compositional and reactivity patterns (fig. 4.7 and fig. 4.9). To acquire more unique starting points at every change of the environment, a more drastic shift is required in the expressed reaction types. Instead of moving over one conditional variable, adding more environmental variation could introduce a higher degree of network evolvability.

Furthermore, it would be interesting to study the robustness these newly acquired network states. For example, by studying the average decay time to the original steady-state. The formose reaction is speculated to exhibit bistable behavior under flow conditions.[13] Possibly, newly formed network states have acquired reactivity *via* feedback loops which allow them to persist for longer.[12,13]

6.3 Expanding the network – introducing new reaction types

Compared to life, the formose reaction has a low level of complexity. For the system to evolve and adopt new functionality, the introduction of new reactivity patterns are a key next step.[14] For biochemists, an increased network complexity translates to an expansion of reactions in the network. This means, for example, new enzymes are introduced to the system to catalyze a specific new reaction, see figure 6.2a.[15] Although enzyme activity is regulated not only by substrate flux (but also by pH dependence or allosteric effects), this approach is rather linear, with a stepwise increase in complexity.[16]

Traditional prebiotic organic chemistry approaches adhere a similar strategy.[1,2] Prebiotic reaction routes were designed towards the synthesis of desired biomolecules. Like the microenvironment in an enzymatic reaction, each chemical conversion requires specific reaction conditions.[1,2] Conflicting reaction conditions prevent these prebiotic networks to run in a 'one pot' system. The concept of reaction classes can be utilized as a tool to predict cross-reactivity and comprehend the interaction of a new reactive molecular structure on a system-level. Therefore, the reaction network is summarized in a higher-level 'reaction class network' (fig. 6.2b). The introduction or modification of a reactive molecular structure can add, remove or modify reaction classes (fig. 6.2c).

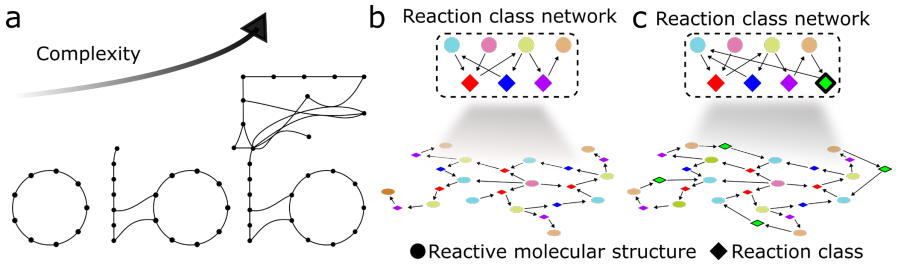


Figure 6.2: Increasing complexity in biological and prebiotic reaction networks. a) Step-wise increase in network complexity in biological systems, for example with the introduction on new enzymes. b) Prebiotic reaction network where a new reaction class (green diamond) is added for the introduction of different feedback loops in the network.

The prebiotic model formose reaction network, as discussed in this thesis, can be tweaked at different reactive groups. The introduction of new chemical species can potentially introduce new reactivity to the network. However, it is not obvious how to add interesting new reactivity to the formose reaction, whilst avoiding deleterious reaction pathways.

The input molecules – formaldehyde (**1**), sugar initiator and catalyst – can be modified to alter the formose reactivity. In different examples from literature the reactivity of either **1** or the sugar initiator was hampered in the formose reaction.[17,18] As discussed under 1.3.2, the aldehydic glyoxylate is a strong electrophile, similar to **1**. The presence of the carboxyl group, however, limits the reactivity of the aldol addition reaction products. Similarly, phosphorylation of the input sugar (glycolaldehyde-2-phosphate) was used for exclusive production of linear sugar phosphates.[17,19] Both these approaches, do only decrease the number of available reaction types in the network.

Different mineral catalysts are known to affect the compositional outcome in the formose reaction.[20] $Ca(OH)_2$ was shown to be the most active formose catalyst.[21,22] Hence, switching formose catalysts will alter the expression of reaction types, but probably not expand the network reactivity. Potentially, a combination of different catalysts might introduce new reactivity to the formose reaction.

The organization of reaction pathways can be modified by interfering with specific reactive groups, such as carbonyls or enolates. Together with **1**, it was advocated cyanide is likely to have played an important role as prebiotic feedstock molecule.[23] In a formose reaction mixture, different cyanohydrins will be formed with the available carbonyls upon the introduction of cyanide.[24] The equilibrium of this reaction with **1** is directed strongly to the cyanohydrin ($K_{eq} = 2.2 \cdot 10^{-6}$ at T = 25 ^{o}C).[24] Hence, cyanide is likely to react with **1** and strongly interfere with the formose reactivity, see figure 6.3a. Oligomerization of glycolonitrile, leads to the formation of an orange solution consisting mostly of glycolonitrile dimers and trimers (fig 6.3b).[25,26] It is questionable how much of the formose reactivity, as discussed in this thesis, remains upon introduction of cyanide.

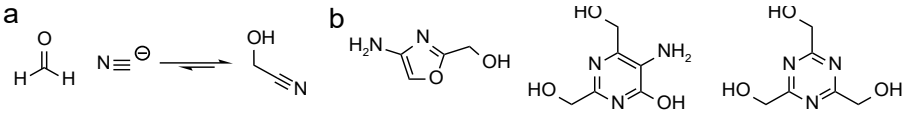


Figure 6.3: Reactivity of a mixture with formaldehyde and cyanide. a) Direct formation of glycolonitrile. b) Dimer and trimer products of oligomerization of glycolonitrile.

6.4 Dynamic interference with formaldehyde availability

Towards a more diverse expression of different reaction classes in a fluctuating environment of $[Ca(OH)_2]_{in}$, I will propose new chemistry that allows for a dynamic interference with different carbonyl substituents in the system. This will provide a potential mechanism to evolve the expressed reaction classes in the formose reaction network.

The introduction of the thiol compound 2-mercaptoethano sulfonate (**2**) allows to control the compositional outcome of the formose reaction, see figure 6.4a. $[\mathbf{2}]_{in}$ was varied between 0 and 200 mM, whereas $[\mathbf{1}]_{in}$ = 50 mM, $[\text{dihydroxyacetone}]_{in}$ = 50 mM, $[CaCl_2]_{in}$ = 15 mM, $[NaOH]_{in}$ = 30 mM, r.t. = 2 minutes and T = 21 °C. Interestingly, the shift in composition upon an increase in thiol concentration ($[\mathbf{2}]_{in}$ = 0 – 200 mM) was similar to a decrease in input **1**, from 50 – 0 mM.

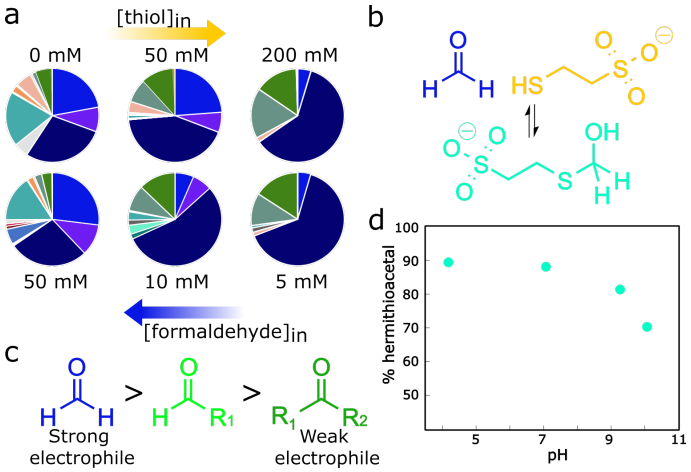


Figure 6.4: Introducing a thiol in the formose reaction, to selectively create a 'sink' for aldehydes, such as formaldehyde. a) The compositional outcome upon increasing thiol concentration, follows an opposite trend to an increasing concentration of **1**. b) The thiol reacts with **1** to form a hemithioacetal species. c) As a carbonyl species **1**

is the strongest electrophile and has a strong interaction with the thiol. d) The pH dependence of the hemithioacetal equilibrium.

The thiol provides a control mechanism for the availability of **1** probably *via* the formation of a hemithioacetal between **1** and **2** (fig. 6.4b).[27,28] The thiol does strongly interfere with reactivity of **1** in the network, since **1** is the strongest electrophile in the system (fig. 6.4c). Therefore, the hemithioacetal can be used as a dynamic 'sink' for **1**.

Interestingly, the equilibrium for hemithioacetal formation is pH dependent. In a ^{1}H-NMR study, **2** was let to react with glyoxylate and form a hemithioacetal (fig. 6.4d). Towards higher pH (> 9 for this example) the equilibrium was directed towards the dissociated state, thus releasing the aldehyde.

The pH dependence of the hemithioacetal equilibrium, could serve as a network-controlled mechanism to release **1** into the system. The pH would control two regimes characterized by the availability of **1** (fig. 6.5a). At low availability of **1**, the network is controlled by reactions between feedstock C_3 to form different C_6 sugars. Together with an increase in pH, **1** is released in the network and introduces new reactivity *via* the formation of reactive C_2 and C_3 species in the Breslow cycle. Subsequent lowering of the pH will introduce these C_2 and C_3 species in a regime with limited availability of **1**. This will result in a shift of network connectivity, since the formation of C_5 sugars is promoted (C_2 + C_3). With the C_5 sugars feeding back into the high pH regime, the composition will be directed away from larger C_6 products. Temporal modulation of the pH provides a thiol-controlled switch between two evolving regimes of reactivity (fig. 6.5b).

We have seen in chapter 5, only temporal modulation of [$Ca(OH)_2$] will not drastically rewire the network reactivity towards new patterns of reactivity. With the newly introduced reactivity of **2**, a potential mechanism is provided for more drastic network reorganization, as discussed under 6.2. The expressed reaction types would undergo a trajectory-dependent shift and allow access to unique dynamic network states (fig. 6.5c). This study could provide useful insight into how prebiotic reaction networks can be controlled dynamically and evolve towards new networks states. We speculate that ever more complex network states are milestones towards the 'final' complex state: life. Presently, we do not oversee the directionality of this gradual complexification of network states. Instead, this study has shed some light on the mechanisms by which networks form more complex network states, and how the environment is a dominant factor.

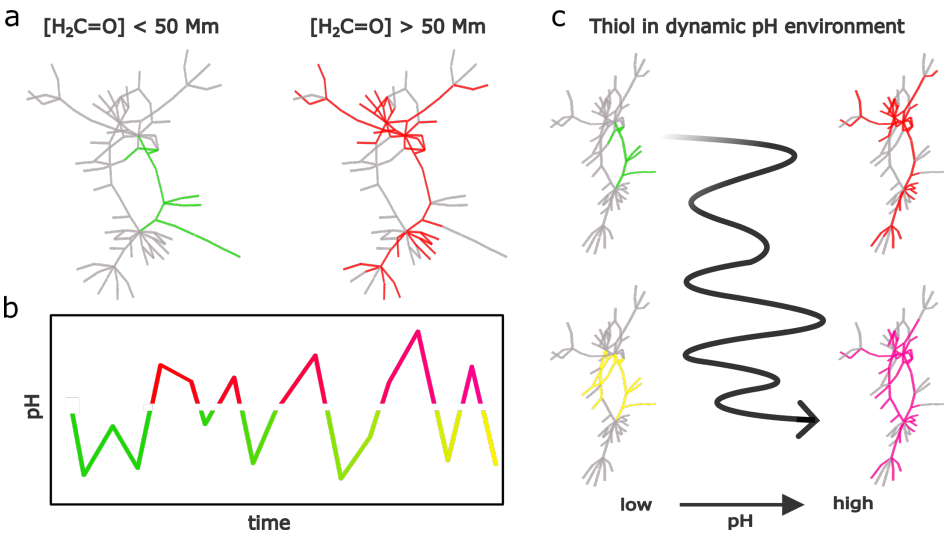


Figure 6.5: Towards new formose reactivity with temporal pH modulation in a thiol-controlled network. a) The availability of **1** controls a thresholded transition of network states ([**1**] = 0 – 100 mM). b) Temporal pH modulation for a new transition of the expressed thiol formose networks. c) Schematic representation of the trajectory which governs the thiol-controlled network transition.

6.5 Conclusion

I have discussed new strategies to study and control the self-organization of a model prebiotic reaction network. Both static and dynamic environmental factors function as directing force for the self-organization of the formose reaction. The observed reactivity patterns adapt and rewire reaction pathways in different environments. Additional dynamic patterns allow to modify the expressed reaction pathways and provide a tool to evolve dynamic network states.

In this perspective, I have argued it is not straightforward to expand the formose reactivity *via* the introduction of new reaction classes. Within this experimental system, the inherent reactivity is ideally maintained upon introduction of a new chemistry. This could be effectuated by introducing a dynamic feedstock 'sink'. However, life did not evolve from the model prebiotic formose reaction. The findings in this thesis provide prerequisites to formulate a scenario for the evolution of prebiotic reaction systems. To understand the origin of life, future research should prioritize the study of chemical evolutionary processes in ever more complex reaction networks.

6.6 References

1. Islam, S. & Powner, M. W. Prebiotic Systems Chemistry: Complexity Overcoming Clutter. *Chem* **2**, 470–501 (2017).
2. 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).
3. 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).
4. Semenov, S. N. *et al.* Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature* **537**, 656–660 (2016).
5. Whitesides, G. M. & Ismagilov, R. F. Complexity in Chemistry. *Science* **284**, 89–92 (1999).
6. Wołos, A. *et al.* Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. *Science* **369**, eaaw1955 (2020).
7. 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).
8. Liu, Y. *et al.* Exploring and mapping chemical space with molecular assembly trees. *Sci. Adv.* **7**, eabj2465 (2021).
9. Breslow, R. On the mechanism of the formose reaction. *Tetrahedron Lett.* **1**, 22–26 (1959).
10. Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461 (2007).
11. England, J. L. Dissipative adaptation in driven self-assembly. *Nat. Nanotechnol.* **10**, 919–923 (2015).
12. Horowitz, J. M. & England, J. L. Spontaneous fine-tuning to environment in many-species chemical reaction networks. *Proc. Natl. Acad. Sci.* **114**, 7565–7570 (2017).
13. Huskey, W. P. & Epstein, I. R. Autocatalysis and apparent bistability in the formose reaction. *J. Am. Chem. Soc.* **111**, 3157–3163 (1989).
14. Kholodenko, B. N. Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell Biol.* **7**, 165–176 (2006).
15. Nam, H. *et al.* Network Context and Selection in the Evolution to Enzyme Specificity. *Science* **337**, 1101–1104 (2012).
16. Hold, C., Billerbeck, S. & Panke, S. Forward design of a complex enzyme cascade reaction. *Nat. Commun.* **7**, 12971 (2016).
17. Müller, D. *et al.* Chemie von a-Aminonitrilen. Aldomerisierung von Glycolaldehyd-phosphat zu racemischen Hexose-2,4,6-triphosphaten und (in Gegenwart von Formaldehyd) racemischen Pentose-2,4-diphosphaten: rac-Allose-2,4,6-triphosphat und rac-Ribose-2,4-diphosphat sind die R. *Helv. Chim. Acta* **73**, 1410–1468 (1990).
18. Krishnamurthy, R. & Liotta, C. L. The potential of glyoxylate as a prebiotic source molecule and a reactant in protometabolic pathways—The glyoxylose reaction. *Chem* **9**, 784–797 (2023).
19. Krishnamurthy, R., Arrhenius, G. & Eschenmoser, A. Formation of Glycolaldehyde Phosphate from Glycolaldehyde in Aqueous Solution. *Orig. Life Evol. Biosph.* **29**, 333–354 (1999).

20. Zafar Iqbal & Senad Novalin. The Formose Reaction: A Tool to Produce Synthetic Carbohydrates Within a Regenerative Life Support System. *Curr. Org. Chem.* **16**, 769–788 (2012).
21. 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).
22. Khomenko, T. I., Sakharov, M. M. & Golovina, O. A. The Synthesis of Carbohydrates from Formaldehyde. *Russ. Chem. Rev.* **49**, 570–584 (1980).
23. Wu, L.-F. & Sutherland, J. D. Provisioning the origin and early evolution of life. *Emerg. Top. Life Sci.* **3**, 459–468 (2019).
24. Schlesinger, Gordon. & Miller, S. L. Equilibrium and kinetics of glyconitrile formation in aqueous solution. *J. Am. Chem. Soc.* **95**, 3729–3735 (1973).
25. Arrhenius, G., Bladridge, K. K., Richards-Gross, S. & Siegel, J. S. Glycolonitrile Oligomerization: Structure of Isolated Oxazolines, Potential Heterocycles on the Early Earth. *J. Org. Chem.* **62**, 5522–5525 (1997).
26. Arrhenius, T., Arrhenius, G. & Paplawsky, W. Archean geochemistry of formaldehyde and cyanide and the oligomerization of cyanohydrin. *Orig. Life Evol. Biosph.* **24**, 1–17 (1994).
27. Kallen, R. G. & Jencks, W. P. The Mechanism of the Condensation of Formaldehyde with Tetrahydrofolic Acid. *J. Biol. Chem.* **241**, 5851–5863 (1966).
28. Lienhard, G. E. & Jencks, W. P. Thiol Addition to the Carbonyl Group. Equilibria and Kinetics [1]. *J. Am. Chem. Soc.* **88**, 3982–3995 (1966).

Addendum

Research data management

This research has been performed in accordance to the research data management policy of the Institute of Molecular Materials (IMM) of Radboud University, the Netherlands.

Research Data Management Policy:

https://www.ru.nl/rdm/vm/policy-documents/policy-imm/

All data as discussed in this thesis are stored on a server maintained by CNCZ:

\\huckdfs-srv.science.ru.nl

The raw data from HPLC and GC-MS analysis as described in chapter 4 and 5 were processed with ChromProcess which is available at:

https://github.com/Will-Robin/ChromProcess (authored by W.E. Robinson)

The output was used for further analysis in chapter 4 and 5.

For data analysis in chapter 4, Python programs are available at:

Python programs for the following described data analysis are available at https://github.com/huckgroup/formose-2021.git (authored by W.E. Robinson). Here, the extracted data from GC-MS and HPLC analysis were available under 'DATA'.

For data analysis in chapter 5, Python programs are available at:

https://github.com/huckgroup/Formose_2022.git. Here, also the extracted data from the GC-MS analysis was available under 'Extended_Data_GH'.

List of publications

Described in this thesis:

1. W.E. Robinson, E. Daines, **P. van Duppen**, T. de Jong, W.T.S. Huck, Environmental conditions drive self-organization of reaction pathways in a prebiotic reaction network, *Nat. Chem.*, **14**, 623-631 (2022).

2. **P. van Duppen**, E. Daines, W.E. Robinson, W.T.S. Huck, Dynamic environmental conditions affect the composition of a model prebiotic reaction network, *J. Am. Chem. Soc.*, **145**, 7559-7568 (2023).

Related to PhD, but not included in this thesis:

3. O.R. Maguire, A.S.Y Wong, M.G. Baltussen, **P. van Duppen**, A.A. Pogodaev, W.T.S. Huck, Dynamic environments as a tool to preserve desired output in a chemical reaction network, *Chem. Eur. J.*, **26 (7)**, 1676-1682 (2020).

Summary

The origin of life was a process underpinned by evolving molecular systems. The field of prebiotic chemistry is inspired by the production of biomolecules under primordial reaction conditions. To increase the yield and selectivity of prebiotic synthesis, individually studied reactions are stitched together for the production of biomolecules. Conflicting reaction conditions in these constructed networks prevent them to be run in a 'one pot' system. The formose reaction provides a model prebiotic reaction system. However, due to its combinatorial complexity the resulting product mixture was coined an intractable 'soup'. The reaction starts from small feedstock molecules formaldehyde and an initiator sugar (e.g. dihydroxyacetone). These react in aqueous solution in the presence of dissolved $Ca(OH)_2$ as inorganic catalyst. The resulting product mixture consists of a plethora of carbohydrates, sugar acids and polyols. This is the result of a set of operational reaction classes: keto-enol tautomerization, aldol additions, retro aldol cleavage and Cannizzaro reactions. By recursively applying these reactions upon the feedstock molecules, formaldehyde and dihydroxyacetone, results in a theoretical divergent and highly interconnected network. Interestingly, the central reaction structure forms an autocatalytic loop (the Breslow cycle), thus resembling metabolic network reactivity.

In this thesis, the formose reaction is used as a prebiotic model reaction network to study the emergence of reaction pathways and how they can be selected for.

Life operates under out-of-equilibrium conditions, relevant for the prebiotic earth as well. I discuss in **chapter 2** how the formose reaction is employed in a continuously stirred-tank reactor (CSTR) to control the steady-state outcome. From the reactor outlet, samples are collected and immediately freeze quenched. The product mixture is processed with an established chemical derivatization protocol for either HPLC or GC-MS analysis. These analysis techniques provide a compositional snapshot of the CSTR content. The feedstock molecules can be quantified and monitored *via* HPLC analysis. The other formose reaction products were characterized from the GC-MS output. Each peak in the GC trace corresponds with a unique compound. The peak area is integrated to quantify each compound. A combination of peak retention time and corresponding mass spectrum was used to deduce the respective molecular structure.

From the compositional outcome of the formose reaction alone, there is little information on which reaction pathways are operational. Therefore, in **chapter 3** I explain how the steady-state composition in the CSTR is probed *via* temporal changes in the input concentration profiles. The differential coupling of the observed compounds provides information of reactivity patterns in the network.

In the first approach, a sinusoidal oscillation of the input sugar is transferred to all species in the network. For each species the observed amplitude decay is used to estimate the distance from the input sugar. In the second strategy, the effective input concentration of $Ca(OH)_2$ is modulated. Here, different groups of compounds, which form small subnetworks, couple together to the input.

In a prebiotic setting, averse of genetic and metabolic control mechanisms, the chemistry interacts directly with the environment. In **chapter 4**, I will discuss how the environment directs the self-organization the formose reaction. Here, the reaction network is probed with varying the input feedstock and catalyst concentration, residence time and temperature. Remarkably, distinct compositional patterns emerge from changes in these environmental conditions. The operational reaction pathways are with temporal modulation of the input sugar and the compositional shifts appeared as a result of rewired reaction pathways. The expression of different reaction classes in the formose reaction is controlled by the environment. Interestingly, the Breslow cycle also provides an important source of C_2 and C_3 building blocks for the system. These network generated building blocks feed a chain growth mechanism alongside the formaldehyde chain growth. The environmental functions as a directing force for the self-organization of prebiotic reaction networks.

Conditions on a prebiotic earth were dynamic, such as day-night cycles and weather patterns. The impact of environmental fluctuations from Ca^{2+} and OH^- is investigated in **chapter 5**. Surprisingly, both the magnitude and rate of fluctuations force a significant deviation in product distribution from the steady-state composition. This is attributed to non-uniform propagation of the environmental fluctuations through the network. Guided by the intermediates in the Breslow cycle, separate sprouting reaction pathways show a collective response. This indicates that environmental dynamics might have an important role in the evolution of prebiotic reaction systems towards the origin of life. Furthermore, the grouped response of different 'subnetworks' might hint at the origin of modular organization in metabolic networks.

The formose reaction is shown to adapt to different environmental traits, both static and dynamic. It can reorganize and utilize different reaction pathways. **Chapter 6** explores the possibility for adaptivity of the formose reaction network and how dynamic environments might provide an important mechanism to underpin the evolution from chemical reaction systems to the origin of life. These findings provide prerequisites to formulate a scenario for the evolution of prebiotic reaction systems.

Samenvatting

Deze samenvatting is geschreven voor de leek, voor de wetenschappelijke versie verwijs ik graag naar de Engelstalige samenvatting.

We begrijpen allemaal hoe je van een levende kip een bord soep maakt. Toch lijkt het onmogelijk om van deze soep weer een levende kip te maken. Hoewel de chemische stoffen uit de kip ook aanwezig zijn in de soep, is het web waarin ze verbonden waren – het metabolisme – kapot. Ooit is op deze aarde het leven ontstaan in een soort oersoep (prebiotic soup), met daarin een primitieve versie van het metabolisme. Vanuit dit eerste leven heeft biologische evolutie het werk overgenomen, met onder andere de kip als resultaat. De omgeving stuurt het ontstaan van nieuwe biologische soorten via aanpassingen in het genetisch bouwplan (het DNA). Zo heeft iedere soort een uniek bouwplan dat is gevormd door de verschillende leefomgevingen waarin zijn voorouders leefden.

Wat het recept is om leven te maken, blijft een raadsel. Het was waarschijnlijk een proces waarin chemische reacties een belangrijke rol speelden. Dit zijn de mechanismes waarmee een nieuwe stof wordt gevormd uit een bestaande stof. Het leven is waarschijnlijk ontstaan vanuit een netwerk van aaneengeschakelde chemische reacties. Uiteindelijk werd dit netwerk steeds complexer. Het netwerk paste zich hiervoor steeds opnieuw aan zijn omgeving aan. In de oersoep gebeurde dit via een ander mechanisme dan dat we kennen in de biologische evolutie. Een genetisch bouwplan dat beschrijft hoe het netwerk eruit komt te zien ontbrak namelijk. De individuele chemische reacties werden daarom direct beïnvloed en gestuurd door de omgeving.

In dit proefschrift beschrijf ik een relatief eenvoudig chemisch netwerk dat ontstaat uit twee bouwstoffen: formaldehyde (CH_2O) en een klein suiker (dihydroxyaceton, $C_3H_6O_3$). Dit is de formose reactie. Wanneer deze stoffen onder de juiste omstandigheden bij elkaar komen, onstaat er een overvloed aan nieuwe stoffen (producten). In het water waarin de reactie plaatsvindt, moeten ook calcium (Ca^{2+}) en hydroxide (OH^-) aanwezig zijn. De geproduceerde stoffen in de formose reactie zijn met elkaar verbonden via een groot netwerk van chemische reacties. Met de formose reactie probeer ik te onderzoeken hoe een eenvoudig chemisch netwerk zich ontwikkelt, gestuurd door zijn omgeving.

In **hoofdstuk 2** beschrijf ik hoe bepaald wordt welke stofjes er in de formose reactie voorkomen en in welke hoeveelheid. Hiervoor worden twee methodes gebruikt: HPLC en GC-MS. Door middel van HPLC kan worden bekeken hoeveel van de beginstoffen aanwezig zijn. Met GC-MS worden de nieuw geproduceerde stoffen in kaart gebracht. Het GC onderdeel bepaalt de

hoeveelheid van iedere stof, terwijl deze verder wordt geïdentificeerd met een moleculaire vingerafdruk uit het MS deel.

De verschillende geïdentificeerde stoffen zijn verbonden via een netwerk van chemische reacties. In **hoofdstuk 3** bespreek ik twee soorten experimenten waarmee het onderliggende netwerk blootgelegd kan worden. Door een signaal door het netwerk heen te sturen kunnen we zien welke stoffen dichtbij elkaar liggen, en welke verbonden zijn op grotere afstand.

In **hoofdstuk 4** wordt bestudeerd wat de invloed is van verschillende omgevingsfactoren op het reactienetwerk dat gevormd wordt. Zo veranderen we bijvoorbeeld de hoeveelheid formaldehyde die wordt toegevoegd. Als deze hoeveelheid een bepaalde drempelwaarde passeert, zien we dat een nieuw soort reactie aan het netwerk wordt toegevoegd. Deze reactie valt te verklaren door een stofje dat gemaakt wordt in de bekende 'Breslow cyclus'. Ook zien we dat het netwerk verandert afhankelijk van de aanwezige hoeveelheid calcium en hydroxide. Dit komt doordat verschillende reacties worden aan- en uitgezet.

De omstandigheden op de vroege aarde waren continue aan verandering onderhevig, zoals de dag- en nachtcyclus en het weer. In **hoofdstuk 5** wordt het gevolg van schommelingen in hoeveelheid calciumhydroxide op het reactienetwerk onderzocht. Tot onze verrassing veranderde hierdoor de hoeveelheid van de aanwezige stoffen significant. De schommelingen in de omgeving verspreiden zich niet gelijkmatig over het reactienetwerk. De ‘Breslow cyclus’ regelt hier als een kringloop van reacties de toegangswegen naar verschillende delen van netwerk. Zodoende lijkt de organisatie van het metabolisme zoals we het nu kennen, welke is opgebouwd uit kleine reactienetwerken, terug te komen in de opbouw van het formose reactienetwerk. Ook laat dit zien hoe een continu veranderende omgeving een belangrijke rol kan spelen in de evolutie van een oerreactienetwerk voor het ontstaan van leven.

We hebben met de formose reactie laten zien hoe een eenvoudig reactienetwerk zich aanpast aan de omgeving en kenmerken begint te vertonen van een eenvoudig metabool netwerk. In **hoofdstuk 6** verken ik de mogelijkheid om de formose reactie te laten evolueren naar een complexer reactienetwerk. Door een nieuwe stof (thiol) toe te voegen kan het reactienetwerk op een andere manier reageren op veranderingen uit de omgeving. Zo hoop ik dat we met de formose reactie meer kunnen leren over de evolutie van chemische reactienetwerken, zoals deze miljoenen jaren geleden plaatsvond in de oersoep. De nieuwe bevindingen in deze thesis vormen een belangrijke grondslag om een scenario uit te tekenen voor het ontstaan van het leven op aarde.

Acknowledgements

Ik ben erg trots op mijn proefschrift! Veel wetenschappelijk onderzoek spreekt wat mij betreft onvoldoende tot de verbeelding. Als tiener leek onderzoek doen naar de oerknal, buitenaards leven of het ontstaan van leven iets onvoorstelbaars. Misschien was dit weggelegd voor een kleine groep hele slimme mensen of iets uit het verleden. Toch is, wonderbaarlijk genoeg, mijn studie op de Radboud Universiteit uitgekomen bij onderzoek naar de 'oersoep'. Veel mensen hebben hieraan bijgedragen, direct of indirect. In het volgende stuk wil ik jullie hier graag voor bedanken!

Wilhelm, jij hebt mij aangenomen voor dit onderzoek, eerst als student en daarna als promovendus. In ons eerste gesprek vroeg ik of de vragen die je stelde niet te groot waren om te onderzoeken. Hier reageerde je stellig op. Juist dit zijn de vragen die de moeite waard zijn om te onderzoeken. Tja, wat doen we hier inderdaad anders? Het enthousiamse dat je hebt voor de wetenschap werkt aanstekelijk. Na de inductiefase van mijn onderzoek, merkte ik dat jouw belangstelling groeide naarmate de resultaten interessanter werden. Ik ben erg trots op onze publicaties. Bedankt voor het vertrouwen en success met je toekomstige wetenschappelijke avonturen!

I would like to express my gratitude to the members of the manuscript committee: **Jana Roithová**, **Matt Powner** and **Albert Wong**. Hopefully it was an enjoyable read, it was my pleasure. I very much appreciate you took the time and effort to review this thesis.

Dirk, we agree to disagree. Desalniettemin is het me altijd een waar genoegen om je weer te zien en nieuw anekdotisch material op te doen. Je uitspraken kunnen inmiddels gebundeld tot een zelfhulpgids voor een prettig leven. Bij voorbaat citeer ik hier graag uit. Ik denk graag aan alle schitterende dingen die we hebben meegemaakt. Je bent een mooi mens en een hele goede vriend. Bedankt!

Nikita, we got introduced from behind a computer screen at the start of the corona pandemic. Odd times. From the moment we switched to real life conversations, it is always a pleasure talking to you. Often this is a bit all over the place, from some science topic, to strange dog behavior, to hiking in Kyrgistan. Thank you for being there and all the conversations we had.

For every PhD student, a partner in crime is an absolute necessity. **Elena**, I am very glad you were so kind to take on this job! Of course, the topic of your research was dear to me, however it wouldn't have worked if you weren't a great

person. Thank you for collaborating (most of the time), all the coffee breaks in and out of the lab and other extracurricular activities.

The first flow experiments with the formose reaction took off rather swiftly. However, **Will**, you were the driving force for giving shape and form to the project as presented in this thesis. Our brainstorm sessions, formal and at the HPLC machine, were both an important source of ideas and also a great joy to me. Many thanks for all your help and the patience with my coding exercises.

Oliver, upon your arrival in the Huck group you were assigned a master student immediately. It was me. With your help I gained confidence in the lab and you helped me to become a more independent researcher. Your knowledge of chemistry and science has always struck me. Thank you for being a colleague and teacher.

De vraag 'wat is leven?' heeft me altijd bezig gehouden, zowel professioneel als daarbuiten. **Albert**, je hebt me hiermee enorm geholpen. Je hebt me geadviseerd stage te lopen op Wilhelm's afdeling. Je dacht dat ik hier goed zou passen. Bedankt voor je vertrouwen en alle boeiende gesprekken.

Alex, thanks for always keeping your head cool whilst doing surgery on a laboratory computer, also you definitely were my greatest neighbor.

Lía, I still very much appreciate you managed to bring sunshine to the lab and brighten days there.

Further, I would like to thank all the members of the Physical Organic Chemistry department and especially **Thijs**, **Mathieu**, **Jeroen**, **Michael**, **Rianne**, **Miglė** and **Souvik** from the joint prebiotic and enzymatic SUPERsubgroup.

I mentioned the co-initiators of the Late Night Conference with Wilhelm Huck before. However, as my favorite extracurricular activity, my gratitude again to **Elena**, **Will**, **Lía**, **Wilhelm** and also to **Glenn** for setting up this magnificent YouTube show.

Desiree van der Wey en **Peter van Dijk**, bedankt voor alles wat jullie op organisatorisch en infrastructureel vlak gedaan hebben. Het was altijd erg fijn om bij een van jullie twee langs te gaan of een praatje te maken op de gang. Namaste.

Een promotietraject tot een succesvol einde brengen kan natuurlijk niet zonder goede vrienden. Ik wil jullie allemaal bedanken, voor de gezelligheid en op zijn tijd voor een beetje mentale ondersteuning. **Bas**, **Jurre**, **Daniël** en **Bart**, lieve

mensen uit Helmond, en **Lars**, **Willem**, **Max**, **Willem**, **Joey**, en **Dirk** ik wil jullie hierbij graag bedanken. Ik vind het erg knap hoe jullie alles voor elkaar hebben, vaak ook met uitgebreide studieloopbaan. Wie had dat gedacht toen we als beginnend student met enige regelmaat in de kroonluchters hingen? Inmiddels lopen de eerste kinderen rond en drinken we ook koffie als we geen tentamen hebben.

Lieve familie, uiteraard was dit werk ook niet zonder jullie van de grond gekomen. **Pap** en **mam**, jullie hebben me alle vrijheid gegeven om uit te zoeken en te doen wat ik wil. Ik voel jullie steun in praktisch alle avonturen waar ik me in stort. **Jaap** en **Okki**, onze fijne jeugd heeft ons veel moois gebracht, maar ook de dingen waar we tegenaan lopen komen vaak overeen. Het is altijd erg fijn om elkaar weer te zien en over al deze dingen te praten. **Oma**, natuurlijk ben jij mijn grootste supporter van het eerste uur. Ik vind het ontzettend bijzonder dat je nog steeds aan de zijlijn staat om me aan te moedigen! Familie Ariaans, in het bijzonder **Leonne** en **Wil**, bedankt dat bij jullie de deur altijd open staat.

Allerliefste **Maud**, je houdt dit soort aangelegenheden graag zakelijk. Mijn liefdesbrief aan jou volgt daarom apart. Op deze plek wil ik je graag heel erg bedanken. Allereerst voor het eindeloos doorlezen van deze thesis en het kijken of de nummering wel consistent is. Je hebt altijd het vertrouwen in me uitgesproken en daarnaast me een spiegel voorgehouden met een menselijke maat. Ik ben heel blij dat je altijd naast me staat en ik kijk heel erg uit naar alle avonturen die we tegemoet gaan samen met onze lieve **Pelle** en **Pip**.

About the author

Peer van Duppen was born on the 30th of November 1990, in Helmond. He graduated his Bachelor Biology (2015) and Master Molecular Life Sciences (2018) at Radboud University. He had his first encounter with science in 2012 during an internship at Karolinska Institutet, Sweden. Although the results from behavioral studies with mice were promising, he decided to pursue a career in the Molecular Sciences. After his internship in the group of professor Wilhelm Huck, he was appointed a PhD position from 2018 – 2022. The most important scientific findings of this period are described in this thesis. He continues his career as a teacher in Organic Chemistry at Wageningen University and Research.