# Materials and Methods

# Trajectory-dependent Compositional Outcomes in a Prebiotic Reaction Network

Peer van Duppen[1], William E. Robinson[1], Elena Daines[1], and Wilhelm T. S. Huck[1,*]

[1] Institute for Molecules and Materials, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands

* Corresponding author. Email: w.huck@science.ru.nl

## Table of contents

## 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, ribose, paraformaldehyde, N,O-bis(trimethylsilyl)trifluoroacetamine, O-ethylhydroxylamide hydrochloride and pyridine were purchased from Sigma-Aldrich. Formaldehyde solution was prepared by depolymerization of paraformaldehyde by heating in water at 60 ºC[1]. Ultrapure water (from a Veolia system) was used to prepare all aqueous solutions. All other chemicals were used without further purification.

## Instrumentation

Gas-Chromatography-Mass Spectrometric analysis were performed on a JEOL JMS-100GCv. The gas chromatograph Agilent 7890A GC 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 ºC and for sample analysis (injection volume: 1 µl) split mode was applied (ratio: 1/10). The GC was used 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.

## Experimental methods

A detailed discussion of parameters chosen in this study is included in the Supplementary Information.

## Flow reactions

Flow reactions were performed in a cone shaped Continuous Stirred-Tank Reactor (CSTR, V = 411 µL) with five inlet channels at the bottom and an outlet channel at the top (fig. S2). It was fabricated of polydimethylsiloxane (PDMS) according to previously reported procedure[2]. Five Gastight Hamilton syringes (25 mL) were filled with solutions of $H_2C{=}O$ (198 mM), DHA (198 mM), NaOH (240 mM) and $CaCl_2$ (120 mM) and water. The syringes were mounted on high precision Cetoni NEMESYS syringe pumps and attached to the inlet channels of the CSTR. The syringe pumps were controlled with NEMESYS software in which a flow profile was loaded for each syringe.

## GC-MS derivatization

Samples from the CSTR outlet were derivatized according previously reported protocols[3–5]. Aliquots (35 ± 0.1 µL) were collected from the CSTR outlet, immediately flash frozen in liquid nitrogen then freeze-dried overnight to give white solids. The samples were randomized before proceeding with derivatization. O-ethylhydroxylamine in pyridine (75 µL, 20 mg/mL) was added to each sample. The samples were heated and shaken (70 ºC, 30 minutes, 600 rpm), then left to cool. Bis(trimethylsilyl)trifluoroacetamide (25 µL) was added and the samples were shaken and heated again (T = 70 ºC, 30 minutes, 600 rpm). The samples were allowed to cool down to room temperature and both dodecane and tetradecane in pyridine (100 µL, 1.6 mM) was added. The samples were then centrifuged (3 minutes, 13,000 rpm). The supernatant was decanted into analytical vials for analysis by GC-MS (see Instrumentation).

## GC-MS data processing

GC-MS peak integration and assignment were performed as previously reported[6] on raw chromatographic data using a program written in the programming language Python with the

Numpy (1.22.2)[7] and Scipy (1.8.0)[8] packages. Peak assignment was performed by comparison of peak retention times and mass spectra with reference chromatograms. Peaks which could not be assigned using reference chromatograms were assigned by inferring their identity based on retention time, fragmentation pattern and consideration of the assignment likelihood given the other compounds produced in experiment (see fig. 4a). Quadratic calibration curves (see table S1) were used to convert peak integrals to compound concentrations. For non-commercially available compounds (branched $C_4$ - $C_6$), for which calibrations were unavailable, an average calibration curve for all calibrated sugars with similar molecular weight was used. For compounds with two peaks in the chromatogram, the largest peak was used in the data analysis.

# Data analysis and interpretation

All data analyses were performed using the Numpy[7] (1.22.2.) and SciPy[8] (1.8.0) Python libraries. The python scripts are supplied in the Extended Data, and are available at https://github.com/huckgroup/Formose_2022.git.

## Mass distribution t-test analysis

To test if two concentration distributions from different experimental conditions come 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. 2, S6-S13) we used an annotation legend to indicate the statistical significance:

ns: $5 \times 10^{-2} < p \leq 1$, *: $1 \times 10^{-2} < p \leq 5 \times 10^{-2}$, **: $1 \times 10^{-3} < p \leq 1 \times 10^{-2}$,

***: $1 \times 10^{-4} < p \leq 1 \times 10^{-3}$, ****: $p \leq 1 \times 10^{-4}$.

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

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

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

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. Where 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 eq. 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 $d_{B,X}$.

In order to visualize *linkage matrices*, a dendrogram was plotted with scipy.cluster.hierarchy.dendrogram(*linkage matrix*). The clusters were assigned via visual inspection (see figure 3**b**).

## Time-interval correlation analysis

Correlation analyses were performed on discrete differentials at varying time intervals (150 s, 120 s, 90 s, 60 s and 30 s) of concentration time traces of the $Ca(OH)_2$ input against the measured outputs (compound: **3**, **5** – **21**, **23**, **25**, **26**, **28** – **31**) (fig. S14).

First, a vector was created from the input $Ca(OH)_2$ concentration at time points where samples were collected. For each input and measured compound we calculated the mean concentration over sample bins covering different time intervals (150 s, 120 s, 90 s, 60 s and 30). Subsequently, we calculated the concentration difference between consecutive bin means for the different time intervals – 150 s, 120 s, 90 s, 60 s and 30 s – to create a 'differential' vector *x*. The values in *x* were means-centered (*x* – np.mean(*x*)) and normalized by dividing by the standard deviation with numpy.std(*x*), (eq 3).

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

Here, *v* is the means-centred and scaled differential vector at a specific time interval, $\bar{x}$ is the mean of vector *x* and $\sigma_x$ the standard deviation of *x*.

The Pearson correlation coefficient for each time interval (150 s, 120 s, 90 s, 60 s, 30 s) was calculated between the respective input differential $v_{in}$ and each of the corresponding output differentials $v_{out}$ using scipy.stats.pearsonr($v_{in}$ , $v_{out}$) (eq. 4).

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

The correlation value r ranges between -1 and 1, $v_{in}$ is an input vector for a specific time interval and $v_{out}$ an output vector for the corresponding time interval. Here, $\overline{v_{in}}$ and $\overline{v_{out}}$ represent the mean value of respectively $v_{in}$ and $v_{out}$. A positive linear correlation is indicated by an r value between 0 – 1 and an inverse linear correlation is indicated an r value

between -1 – 0. A correlation value of r = 0 indicates that there is no linear correlation between the input and output.

## Relative concentration changes from steady state

The relative concentration difference was calculated between each of the identified compounds in the perturbed state with respect to the steady state.

First, we constructed a vector for each of the observed compounds for the relative change of the average concentration in a perturbed state with respect to its steady state (eq. 5).

Eq. 5: Relative change = $\frac{\overline{perturbed\ state} - \overline{steady\ state}}{\overline{steady\ state}}$

For compounds in each of the identified clusters **I** – **V** the relative concentration changes were means centred and normalised by dividing by the largest relative difference from steady state. If a compound was not observed in a steady state condition, the relative difference for the perturbed state was set to 1 after the normalization step.

# References


1. Ellis, E. A. A Fast, Simple, and Safe Way to Prepare Paraformaldehyde Solutions. *Microsc. Today* **17**, 48–49 (2009).

2. Semenov, S. N. *et al.* Rational design of functional and tunable oscillating enzymatic networks. *Nat. Chem.* **7**, 160–165 (2015).

3. 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).

4. Becker, M. *et al.* Evaluation of different derivatisation approaches for gas schromatographic–mass spectrometric analysis of carbohydrates in complex matrices of biological and synthetic origin. *J. Chromatogr. A* **1281**, 115–126 (2013).

5. 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).

6. 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).

7. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

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