Exploring the Untargeted Synthesis of Prebiotically-Plausible Molecules

University of Glasgow

A thesis submitted in fulfilment of the requirements for the degree of **Doctor of Philosophy**

School of Chemistry
College of Science and Engineering
University of Glasgow

Presented by

Stephanie Marie Colón-Santos

September 2019

“You know, I mean, like, old-lady science, you know?
She's a real-- You got to hang on tight, you know? Because she-- she'll--
She bucks pretty hard.”

-Rick and Morty, ‘The Ricks Must Be Crazy’

Acknowledgements

The work presented in this thesis was carried out between February 2016 and August 2019 in the group of Prof. Leroy Cronin in the School of Chemistry at the University of Glasgow. Through-out the years, I received the help and support of several group members. In particular, I would like to express my sincere gratitude to the following people:

Prof Leroy Cronin, for taking me on as an intern in June 2015 and giving the opportunity to start a PhD in his research group. It has been a life-changing experience.

Dr Geoffrey J. T. Cooper, for being a great team leader and providing excellent guidance during the writing process of my first-first author paper. Also for proofreading most of this thesis.

Dr Andrew J. Surman, for teaching me about multiple analytical techniques during the first-part of my PhD and providing me with the essential skills for my research.

Dr Irene Suarez-Marina, for her great help during my PhD and the overall savvy advice.

Dr Laia Vila-Nadal, for help being a great role-model and my 'big-sister' away from home.

Dr Dario Caramelli, Dr Vasilis Duros, Dr Eric Jannusson, Dr Davide Angelone, Andrius Bubliauskas, Manuel Kupper, Alex Hammer and Tom Goosens, for being more than colleagues and offering me their friendship during this experience. I will always keep our memories very close to my heart.

Dr Rebecca Turk Macleod, Ms Liva Donina, Dr Jaroslaw Granda, Dr James Taylor, Dr Irene Suarez-Marina, Dr Sergio Martin, Dr Jessica Bame, Mr Hector Soria, Ms Sabrina Galinares, Dr Yousef Abul-Haija, Dr Ommid Anamimoghadam, Dr Jonathan Grizou, Dr Laurie Points and Dr Leanne Bloor for being awesome co-workers, making working hours much more enjoyable and always being up for a science chat.

Jim McIver and Diana Castro-Spencer, for their endless help whenever I had technical problems and their unconditional kindness.

Amanda McGarvey, for her administrative help and support.

Also, I do not want to forget **Cassandra Clarkson, Yaritza Santos and Paul Vincent Ferri**. Although they are not Cronin group members, they made a genuine difference through the most challenging moments of the PhD.

All the members of the Cronin Group, past and present, who have helped to make this experience an unforgettable one.

Table of Contents

Publications

The following articles were published as a result of work undertaken over the course of this PhD programme.

1. *"Miller–Urey Spark-Discharge Experiments in the Deuterium World."* G. J. T. Cooper, A. J. Surman, J. McIver, S. Colón-Santos, P. S. Gromski, S. Buchwald, I. Suárez Marina, and L. Cronin, *Angewandte Chemie Int. Ed.* **56,** 8079–8082 (2017)

2. *"Integrated synthesis of nucleotide and nucleosides influenced by amino acids."* I. Suárez-Marina, Y.M. Abul-Haija, R. Turk-MacLeod, P.S. Gromski, G.J.T. Cooper, A.O. Olivé, S. Colón-Santos and L. Cronin *Communications Chemistry Int. Ed.*, **2**, 28 - 36 (2019)

3. *"Taming the Combinatorial Explosion of the Formose Reaction via Recursion within Mineral Environments."*
   S. Colón-Santos, G. J. T. Cooper, and L. Cronin. *ChemSystemsChem*, **1**, 3, e1900014 (2019) – *(Front cover)*

Abbreviations

In addition to standard notation, the following abbreviations were used in this thesis:

RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
HILIC	Hydrophobic Interaction Liquid Chromatography
EIC	Extracted Ion Chromatogram
ESI-MS	Electrospray Ionisation Mass Spectrometry
FLD	Fluorescent detector
HPLC	High Performance Liquid Chromatography
RT	Retention Time
m/z	Mass to charge ratio (mass-spectrometry)
CAD	Corona Aerosol Detector
DAD	Diode-Array Detector
LC	Liquid Chromatography
UPLC	Ultra Performance Liquid Chromatography
HR- MS	High Resolution Mass Spectrometry
NMR	Nuclear Magnetic Resonance Spectroscopy
MS/MS	Tandem Mass Spectrometry
SEM	Scanning Electron Microscopy
EDS	Energy Dispersive X-ray Spectroscopy
RP	Reverse Phase
NP	Normal Phase
SEC	Size-Exclusion Chromatography
PCA	Principal Component Analysis
GC-MS	Gas Chromatography Mass Spectrometry
UHPLC	Ultra-High Performance Liquid Chromatography
UV/Vis	Ultra-Violet / Visible
FT-ICR	Fourier Transform Ion Cyclotron Resonance
TCA	Tri-Carboxylic Acdd (or Krebs)
ToF	Time-of-Flight
FT-ICR	Fourier Transform Ion Cyclotron Resonance

Abstract

One of the biggest challenges we face when studying the Origins of Life (OoL) is that in the absence of a time-machine, it is not possible to make direct observations about what actually happened on early earth. Recently, a more 'systems' approach has been taken on, which looks for new phenomena and is not constrained by the search of particular products. Investigations of prebiotic complex chemical networks are increasingly tailored towards the elucidation of which environmental conditions are capable of 'tuning' the product distribution towards a greater degree of complexity. For this reason, a series of classic Miller-Urey experiments were conducted alongside with all-deuterated Miller-Urey experiments to explore the effect a 'heavier' isotope in the resulting chemical space of the complex mixture.

Previous work in prebiotic chemistry has demonstrated that the inclusion of mineral surfaces in complex reaction networks, can effectively steer the product distribution into a particular product. In order to address this, we carried out the Formose reaction in a mixture of water and Formamide (50:50 v/v) and investigated how different environmental inputs (such as mineral surfaces and reaction cycling) can affect the reaction, by steering it into a particular outcome. Also, inspired by the metabolomics workflows designed for metabolite discovery, we conducted UPLC-MS/MS in a Data-Dependent fashion, which allows for features to be generated in a confident manner with each one representing a product within the complex product distribution and mapping the resulting chemical space of the products.

Finally, in the case of the Miller-Urey experiment, few versions have been carried out so far (i.e. besides variations within the energy source used in the experiment or the gas mixture employed). Therefore, this prompted us to investigate the effect of reaction cycling in the Miller-Urey reaction. The effect of natural processes such as atmospheric cycling, is an important but not yet addressed variable within the prebiotic broths framework. Therefore, we decided to investigate what effect could this have in the overall product distribution of the famous experiment.

1. Introduction

Origins of Life: *An open question*

The Origin of Life on earth remains one of the most important open questions in science. Around four and a half billion years ago (4.5 Ga), from the gas and dust left over by a newly formed sun, our planet, Earth, came to existence.[1] During the following hundred million years, the young Earth was bombarded by meteorites and comets and had hot nascent oceans and many violent volcanic eruptions.[2] However, within about a billion years, life had arisen.[3,4] The current timeline for *when* life arose is part of an ongoing debate, but it is estimated to be 2.5 to 3.7 billion years ago, depending on who you ask.[5,6] Similarly, *where* life arose, is also still an open question.[7] The specifics of the environment that cradled the first forms of life (e.g. atmospheric/oceanic composition, range of temperatures, etc.) still remains unknown and is highly debated among the scientific community.

Many different theories have been developed as to *where* life could have started, taking a plethora of scenarios into consideration and their plausibility for supporting life.[8] For example, in the theory of drying ponds or wet/drying cycles, abiotically synthetized simple organic compounds concentrate as a pool evaporates and the total volume is reduced. This means that it can promote condensation and polymerization, with the loss of water molecules. The discovery of hydrothermal vents awoke a good deal of interest towards more extreme environments; where the redox potential of a hot and mineral-enriched environment could serve as an energy source to overcome the thermodynamic barriers of making life's building blocks.[9,10,11] Even really cold environments[12] or atmospheric aerosols[13, 14] have been considered. Furthermore, there is always the possibility that life's building blocks already existed in outer space and they reached earth during the Late Heavy Bombardment (LHB) period through a meteorite.[15,16,17]

If the *when* and the *where* of the emergence of life on earth wasn't mysterious enough, *how* life arose on earth is even more troubling. The complexity of even the simplest life forms is astonishing, and consequently the transition of non-living, simple chemical compounds into the molecules of life remains one of the biggest mysteries in science. The uncertainty revolving around almost every aspect about life's origin leaves the door open for a myriad of possibilities. However, we can try to narrow it down.

There are three things that we know about life, which will guide us in the quest to understand life's origins:

1) All of life's building blocks (e.g. proteins, carbohydrates, nucleic acids and lipids) are primarily composed of carbon, hydrogen, nitrogen, oxygen, phosphorus and sulfur also known as 'CHNOPS'. [18]

2) The ability to undergo evolution, to change and adapt, is one (if not the most) important feature of life as we know it. Darwin's work 'On the Origin of Species'[19], initiated a discussion on evolution that still continues to this day, but this time trying to fill the gap between inanimate matter and life. The chemical 'evolution' of simple organic molecules into a higher level of organization and complexity is triggered by the relationship with the environment. Much like the theory of evolution, the individuals (molecules) that are more suited to the environment are more likely to survive; then continue to change, evolve and adapt, necessary for their survival as a consequence of the dynamic environment. A deeper discussion on the relationship between the environment and evolution is discussed in Henderson's book 'The fitness of the environment', or as some call it, 'Darwin's fitness'.[20,21] Currently, the general working definition of life by the National Aeronautics and Space Agency (NASA) is 'a self-sustained chemical system capable of undergoing Darwinian evolution'.[22]

3) Nothing was 'pure' to begin with; the synthesis of biomolecules was not accomplished by nature in pristine laboratory conditions. All the possible pathways for the one-pot synthesis of life's building blocks results in a very messy and complex mixture of products. The transition from a combinatorial explosion of products into constrained reaction networks is necessary for the construction of biomolecules in sufficient yields. However, the energy required to overcome the thermodynamic barriers of this construction can be obtained from the environment, in a process that Schrödinger describes as "feeding from negentropy".[23] This reflects the need for temporal organization, self-replication and auto-catalysis in the complex systems.[24] Nonetheless, exactly how this happened remains unknown and we set out to explore this with the help of modern analytical techniques.

1.1 A (brief) history of the 'Origins of Life' field

A complete overview of the literature regarding this matter would be highly ambitious in any context. Therefore, I must be somewhat selective and will attempt to deliver an unbiased narrative of the main events that constructed the current view of the field, as well as the remaining challenges.

1.1.1 The philosophy of Life: *Vitalism and Spontaneous Generation*

What originated as a rich interplay between philosophical assumptions on the nature of life itself, mainly from a metaphysical and religious stand-point, kick started a discussion on the 'vital' quality of living matter. This notion came to be known as Vitalism and it argues that there is a distinct physical-chemical behaviour governing all living beings that separates them from the non-living. An assumption that can be dated all the way back to ancient Egyptians, who looked at bodily functions as a proof of Vitalism. As well as, the Stoics who described this quality as the 'soul' behind the unique character of living matter.[25] Other ancient Greeks, including Aristoteles, also believed that all living beings had originated from non-living material as a spontaneous and serendipitous process.[26] Later this came to be known as the 'Spontaneous Generation' theory of the Origins of Life, when it was officially introduced by John Needham in a scientific context.[27] However, Vitalism remained a preferred theory until several scientists in the 1600- 1700's (including Needham) conducted experiments that challenged the hypothesis. Amongst these were two note-worthy Italian scientists: Francesco Redi, who discovered that maggots in rotting meat did not grow in the absence of eggs and Lazzaro Spallanzani, who found that microorganisms were present in the air.[28]

A century later, a discovery by German scientist Freidrich Wohler, irreversibly changed the meaning of vitalism in the most remarkable way. He managed the abiotic synthesis of urea (an organic compound present in the urine of all mammals) from inorganic material and sparked a debate that carries on until this day: the Origins of Life from abiotic material.[29] Not long after, Louis Pasteur irrefutably disputed the theory by demonstrating in his experiments that the absence of a pristine environment (i.e. proper controls) was the driving force behind this idea.[30] The theory of 'Spontaneous Generation' never fully recovered from that major blow, but it mutated into something else.

1.1.2 Darwinian evolution: *A game-changer*

Inspired in the principle of evolution developed by Darwin, Pasteur was able to demonstrate the implausibility of spontaneous generation and sparked a general belief in the scientific community that living organisms were a product of a gradual transformation of inanimate matter. Many theories were formulated at the time, but none of them persisted. It was not until the 1920's that two scientists, Alexander Oparin and John B. Haldane, independently revisited the idea and constructed a pathway of chemical evolution that could fit the theory.[31] Their work suggested a series of chemical steps that would increase the molecular complexity and functionality of abiotically produced organic compounds generated in a reducing atmosphere. The sequential accumulation of organic compounds and their eventual polymerization, would had resulted in the generation of aggregates and lead to the formation of coacervates (i.e. protocell), from which the first heterotrophic microbes evolved.[32] A hypothesis that eventually came known as the 'Primordial Soup' theory, see **Figure 1.**

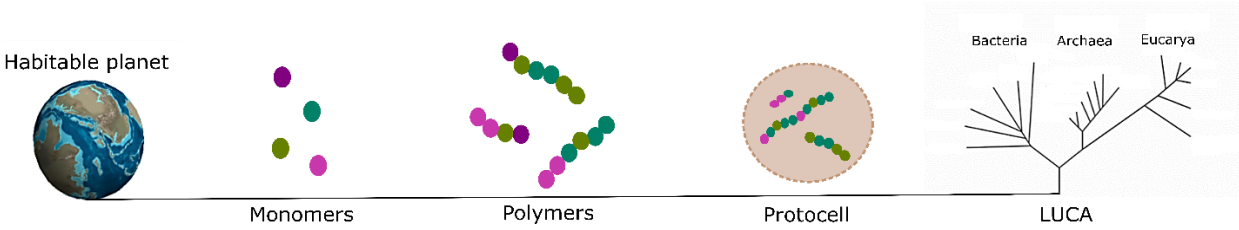


Figure 1. The Primordial Soup theory: Once the planet had become habitable, abiotically produced monomers develop into polymers, consequently leading to the production of the first protocell, which ultimately transitioned towards the Last Universal Common Ancestor (LUCA).

It is hypothesized that the transition of simple chemistry into life's building blocks, must have required some sort of selective process.[33] Potentially, by a mechanism of energy exchange with the environment, since the level of order and complexity needed to make the building blocks is superior to the one we get from the one-pot batch reactions of these molecules. This assumption leads us to believe that there must be a pathway, unknown to us, that resembles a process of natural selection, which allows for the "messy" reactions to converge into a higher order of organization and give rise to a more complex system.[34]

In a letter to a friend (Hooker) written in 1871,[35] Darwin wrote the following sentences: "*It is often said that all the conditions for the first production of a living organism are now present which could ever have been present. But If (and oh what a big if) we could*

conceive in some warm little pond with all sorts of ammonia and phosphoric salts, light, heat, electricity etc. present, that a protein compound was chemically formed, ready to undergo still more complex changes at the present such matter would be instantly devoured, which would not have been the case before living creatures were formed."

1.1.3 Abiogenesis: *The birth of Prebiotic Chemistry*

In 1870 an English biologist named T.H. Huxley, proposed that "living organisms arise only from pre-existing living matter such as simple organic compounds", a concept known as Abiogenesis.[36] About sixty years later in 1938, Oparin's book titled 'Origin of Life' was released.[37] It explained the theory of the 'Primordial Soup' in detail, alongside with a complete description of early Earth's environmental conditions. However, it was not until the 1950's that the theory was put to the test by a couple of American chemists in the University of San Diego, California. An experiment that would mark the beginning of the Prebiotic Chemistry field.[38]

In 1952, a student of Harold Urey, wanted to do an experiment completely unrelated to the research carried out in his lab: he wanted to recreate the early earth environment (as it was envisioned by Oparin and Haldane) with a lab-designed apparatus that involved a round bottom flask filled with water (as a substitute of ocean water), two tungsten electrodes (to simulate lighting) and a mix of ammonia, hydrogen and methane as the atmosphere. The student's name was Stanley Miller and he managed to identify a series of organic compounds relevant to all known life forms, amino acids.[39] The resulting mixture contained many other small organic compounds in low concentration, to which a complete characterization remains a challenge. This mixture of compounds, came to be known as the prebiotic broth (*or soup*) and was the first experimental proof of Oparin and Haldane's theory: the synthesis of 'spontaneously' generated organic material, which are known to be present in living systems. The discovery re-ingnited the ideas of the 19th Century into the theory of 'Abiogenesis'. Contrary to a biogenic origins of living organisms (which can be indirectly related to vitalism), the abiogenic theory of life hinges on a series of evolutionary transitions from this prebiotic mixture that eventually resulted in a proto-organism, which preceded all of us. The study of this 'prebiotic broth' and its chemical evolution paved path for an ongoing field, coined 'Prebiotic Chemistry'.[40]

The results from the Miller-Urey experiment were presented in the very first meeting of the International Conference on the Origins of Life, in Moscow (1957). In the meeting, Miller was able to meet and discuss with other pioneers of the field, such as Leslie Orgel, who had been working closely with Francis Crick. The same year Miller published his findings on the 'Prebiotic Soup' hypothesis, Crick (alongside Watson and Franklin) had managed to complete a molecular model for DNA.[41] Leslie Orgel, inspired by the discovery of the chemical basis of the biological genetic code, started questioning the relation between proteins (polymerized amino acids) and DNA/RNA. He came to the conclusion that the connection between the two biopolymers that control our biological machinery, indicated an early connection between them in the path of chemical evolution; envisioning peptide-nucleic acids polymers as the predecessors of the current form.[42] The theory proposed by Orgel has not been proven to date, mainly due to the unsuccessful polymerization of the abiotically generated monomers in a prebiotic broth. Therefore, a transition into polymers is still the main bottleneck when working directly with the Primordial Soup. As a way to circumvent this, interest has grown on how the environmental conditions could have promoted self-organization and the initial transition into information baring mechanisms, see **Figure 2**, the hypothesis being that once a set of 'reproducible' polymers have been achieved, from any natural source of abiotic material, then the principles of Darwinian evolution would take over and promoting selection and evolution.[24]

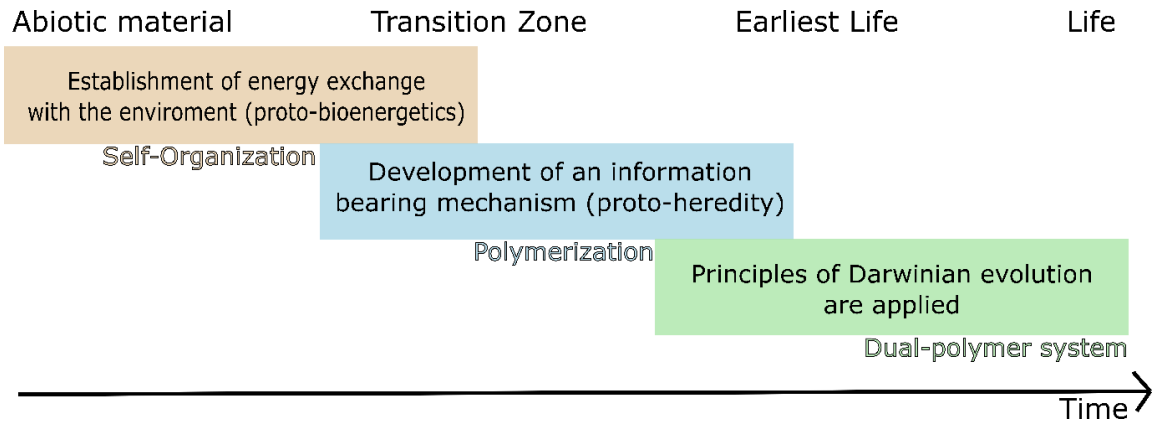


Figure 2. The theory of chemical evolution: A schematic on the hypothetical transition of abiotic material into life, from the self- organization of small molecules to selective polymerization, further subjected to Darwinian evolution principles and eventually leading to modern life

Nonetheless, the problems encountered with the chemical evolution of abiotic material, ignited a debate in the 'Prebiotic Chemistry' community that continues to this day. Opposite views on the significance of either polymer (DNA/RNA or Proteins/peptides) have dominated the discussion. Their prebiotic plausibility is currently a subject of main interest

in the field and whether the interaction of such polymers constructed a symbiotic relationship *early* on the process of chemical evolution, remains an open question. Only recently, in the last 20 years, have we found evidence of the simultaneous synthesis of amino acids and nucleobases in a Miller-Urey type experiment.[43]

1.2 The cradle of life: *A selection of geochemical scenarios*

1.2.1 Warm-little pond

Charles Darwin, in a famous letter, had described "*a warm little pond with all sorts of… (chemicals, in which) …a protein was chemically formed*." as the event that preceded the '*Origin of Species'*. A theory that remains a favourite for many Origins of Life scientists to this day.[35] One of the biggest challenges in the transition from simple building blocks into biopolymers is that this process is not thermodynamically favourable in water, our biological solvent. Therefore, there is an energetic requirement for this transition, which can be addressed by coupling it to a natural process (see **Figure 3**).[44,45] A warm little pond could provide enough energy by the gradual evaporation of water molecules, a process known to drive the polymerization of amino acids into peptide units.[46,47] As well as, providing the energetics for the abiotic formation of RNA monomers, through a wet-dry cycling process.[48,49]

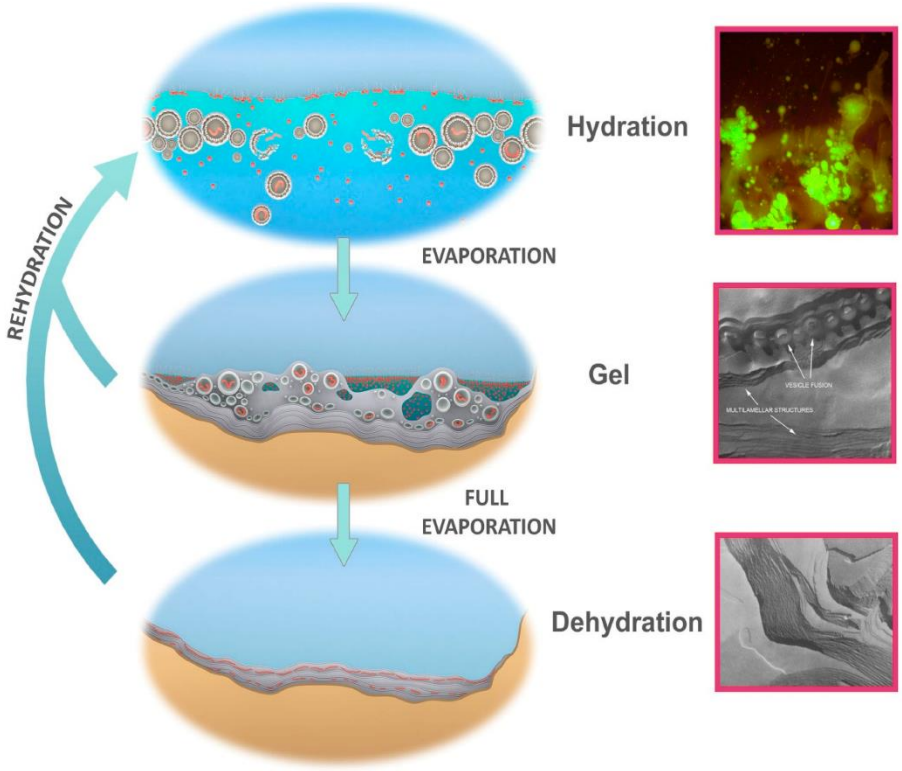


Figure 3. The envisioned scenario for a prebiotic reactor on early earth, a 'warm little pond' that enabled the gradual condensation of abiotic material into polymers that can transcend into a proto-cell. Reproduced from MDPI journal *Life,* 2016, no permissions needed after citation.[44]

How natural processes, such as environmental cycling, can affect the generation of organic compounds and their subsequent polymerization is an important variable within prebiotic studies. The chemical cycling of abiotic material is a bounded characteristic of any environment on planet earth.[50] As well as, all known living systems contain a series of chemical cycles as a way of achieving self-sustenance, while developing the adaptive properties necessary for their survival in an ever-changing environment.[51] This are two very present aspects in any form of life of earth, which has consequently lead most scientists to believe in early relationship between abiotic and proto-biotic chemical cycling processes.[52,53] Furthermore, one main distinction in the definition of anything considered alive, relies in the capacity of "avoiding the rapid decay into the inert state of 'equilibrium" - as mentioned by Schrödinger in his book, 'What is Life?'.[54]

In the context of prebiotic broths, the question of reaction cycling driving the product distribution towards a higher-level of self-organization, has not been addressed to date. Nonetheless, it's been discussed as one of the viable ways to impart selectivity in such mixtures.[55,56,57] Therefore, since all known batch synthesis of Life's building blocks results in a combinatorial explosion of products with no apparent selectivity (often referred as

'asphalt'), there is the possibility that a continuous recursive process would have driven the complexification of abiotic material towards a higher-level of order, eventually leading to a living entity. See **Figure 4**. An approach, which has been seen to drive selectivity in complex mixtures, by recursively enhancing a selection process in a Dynamic Combinatorial Library (DCL).[58] Moreover, there has been an open forum on the duration of chemical cycles on early earth, alongside with other environmental conditions such as temperature. These studies have found that the typical duration of a natural atmospheric cycle, during the period in which Life is envisioned to arise, was of 2 to 6 hours.[59] Adding to the increasing number of considerations, when trying to adapt the prebiotic synthesis of organic material in a laboratory setting towards more realistic far-from-equilibrium conditions, as they would occur in the natural environment of early earth.

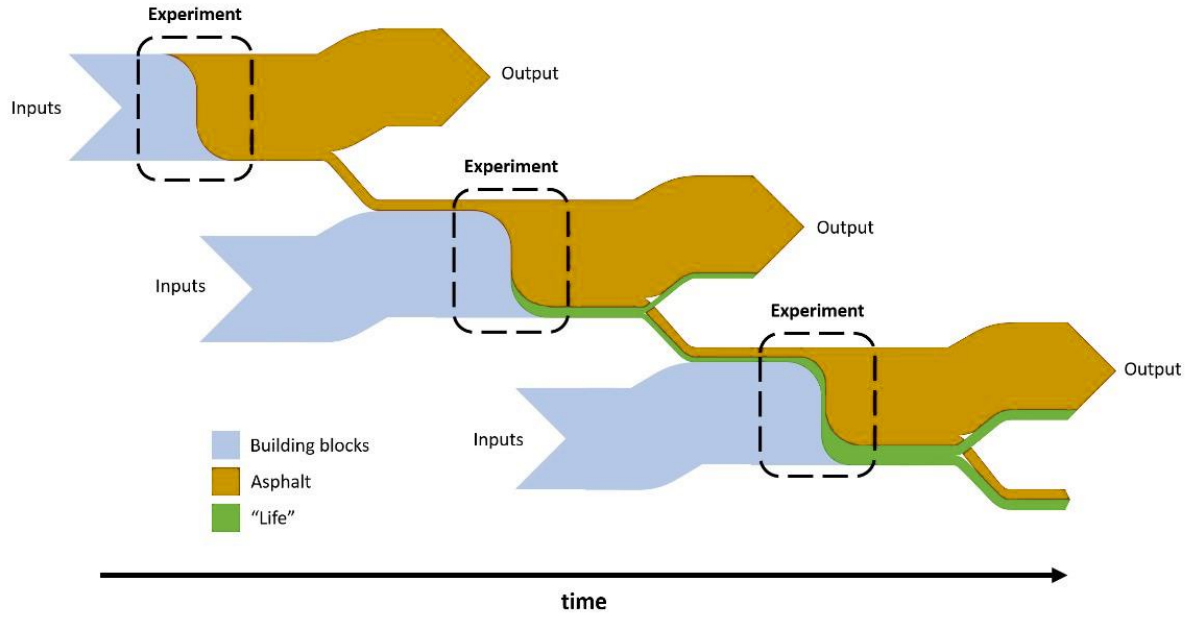


Figure 4. Sankey diagram of a recursive process: Chemical cycling on early earth, mediated by natural processes, could have jump-started the complexification of abiotically-produced material.

1.2.2 Hydrothermal-vents

The transition of monomeric units into life's biopolymers remains a main challenge and therefore alternative theories on the energy source behind this leap have been proposed. In 2007, 20 years after the discovery of hydrothermal vents (see **Image 1**), a publication by Russell *et al.* provided a compelling case on this environment being the cradle for life. He argued that the necessary energetics where provided by a pH gradient generated from the plume releasing H_2S into the relatively basic ocean.[60] Subsequently, Russell demonstrated that alkaline vents created (what he called) an abiogenic 'proton motive force' (PMF) or chemiosmotic gradient,[61] sustaining the ideal conditions to drive the synthesis of complex molecules. This is simultaneously coupled to the probability of generating micro-compartments that provide an abiotic mechanism for concentrating organic molecules, by the action of iron-sulfur minerals, such as pyrite and mackinawite, which translates into a mineral-cell type moiety with multiple catalytic properties, as once predicted by Wächtershäuser.[62] Meanwhile, the movement of ions across the membrane can be achieved by two main mechanisms: (*a*) The diffusion force generated through the concentration gradient, moving from high to low concentrations, and (*b*) a resulting electrostatic force promoted by the electric potential of cations (in this case, protons).

The aforementioned proton motive force can be seen as a measure of the potential energy, coming together from a combination of proton and voltage gradients across a membrane. Furthermore, Pier Luigi Luisi and Jack W. Szostak suggested that the energetics required to overcome the known thermodynamic barriers of chemical evolution are better suited to a hydrothermal vents geothermal scenario, since their abiotic activity provides a better chance for life to arise than the theory of drying ponds, even in the presence of minerals.[63] All by means of taking advantage of the natural chemiosmotic gradient and coupling the associated geo-chemical reactions to the prebiotic synthesis of complex material. Not long after, in 2010, a study demonstrated the increased probability of the first living organisms arising from hot water rich in minerals, by conducting a series of analyses of sea water near hydrothermal vents.[64]

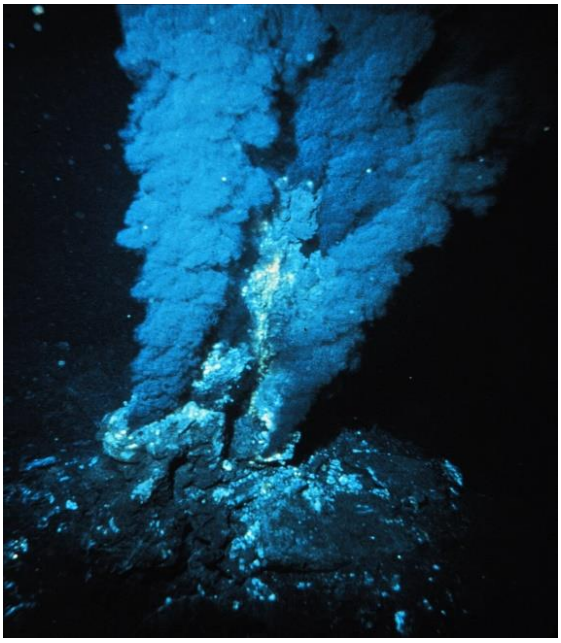


Image 1. A deep hydrothermal vent or black smoker, taken from the NOAA photo library (public domain).

The development of laboratory experiments under hydrothermal conditions, usually involve the heating (200°C or higher) of aqueous solutions of prebiotic building blocks at high pressures, by means of a hydrothermal reaction vessel.[65] An extensive review by Cleaves *et al.* efficiently discusses the plausibility of this scenario providing the necessary energetics to drive the synthesis of a known biopolymer: peptides.[66] He notes that all polymerization experiments of amino acids under hydrothermal conditions start with an unrealistic (high) concentration of amino acids. As well as, the risk towards monomer degradation at such high-temperatures, which might compete with the polymerization process. Therefore, shifting the monomer-polymer equilibrium towards net depolymerisation or breakdown. Nonetheless, several experiments on the condensation of amino acids into short peptides under hydrothermal conditions have been proven successful, particularly in the presence of a certain catalysts, for example: copper (Cu^{+2}) ions[67], alumina (from clays)[68] and fatty acids.[69]

1.2.3 Water-soil interface *(Mineral-clay hypothesis)*

An organic chemist and molecular biologist at the University of Glasgow, Graham Cairns-Smith, developed an alternative theory based on the interaction of clay minerals with the prebiotic soup.[70] He envisioned that the defects on the surface of mineral, could act as a selective and information bearing agent in the process of chemical evolution. If molecules who had a stronger interaction with the silicon surface managed to survive environmental

fluctuations, then a selective process had begun, which eventually would lead to perfect microenvironments for the first living machinery to thrive. His theory has never been experimentally proven *per se*, but clay minerals do selectively catalyse the synthesis and adsorption of certain compounds under prebiotic conditions.[71,72] For example, there is the case of the formamide condensation on clay minerals resulting in synthesis of nucleobases and amino sugars.[73] As well as, the silicate mediated formose reaction, which managed to select and stabilized pentoses-hexoses preferentially, which happen to be the most relevant in modern biology.[74] Understanding how selection could be imparted in the combinatorial explosion of prebiotic broths remains a main driver for the mineral- interaction theory. Since this can act as a dual beneficiary of the chemical evolutionary transition, participating in selectivity and subsequent concentration of certain molecules over others (with non-covalent interactions promoting the preference of higher molecular weight and potentially more complex material), thus achieving a truncation of the promiscuous product distribution.[75]

The minerals used in prebiotic experiments are constrained by a mineral evolution theory developed by Bob Hazen.[76] His work proposes that the incredible variety of minerals present on earth today came into existence only after the great oxygenation event (i.e. a mayor atmospheric compositional transition from neutral or CO rich to having relatively high concentrations of oxygen), which is believe to be a consequence of life arising on earth.[4] This results confirm that only a subset of known minerals were available on early earth.[77] Most work carried out on the mineral catalysis of prebiotic systems are aware of this and adjust their mineral selection accordingly, see **Table 1**.

TABLE 1. Ten stages of mineral evolution of terrestrial planets, with possible timing on Earth, examples of minerals, and estimates of the cumulative number of different mineral species

Stage	Age (Ga)	Examples of minerals	~ Cumulative no. species
The era of planetary accretion (>4.55 Ga)			
1. Primary chondrite minerals	>4.56 Ga	Mg-olivine/pyroxene, Fe-Ni metal, FeS, CAIs	60
2. Planetesimal alteration/differentiation	>4.56 to 4.55 Ga		250
a) aqueous alteration		phyllosilicates, hydroxides, sulfates, carbonates, halite	
b) thermal alteration		albite, feldspathoids, biopyriboles	
c) shock phases		ringwoodite, majorite, akimotoite, wadsleyite	
d) achondrites		quartz, K-feldspar, titanite, zircon	
e) iron meteorites		many transition metal sulfides and phosphates	
The era of crust and mantle reworking (4.55 to 2.5 Ga)			
3. Igneous rock evolution	4.55 to 4.0 Ga		350 to 500
a) fractionation		feldspathoids, biopyriboles (volatile-poor planets)	350
b) volcanism, outgassing, surface hydration		hydroxides, clay minerals (volatile-rich planets)	500
4. Granite formation	4.0 to 3.5 Ga		1000
a) granitoids		quartz, alkali feldspar (perthite), hornblende, micas, zircon	
b) pegmatites		beryl, tourmaline, spodumene, pollucite, many others	
5. Plate tectonics	>> 3.0 Ga		1500
a) hydrothermal ores		sulfides, selenides, arsenides, antimonides, tellurides, sulfosalts	
b) metamorphic minerals		kyanite, sillimanite, cordierite, chloritoid, jadeite, staurolite	
6. Anoxic biological world	3.9 to 2.5 Ga		1500
a) metal precipitates		banded iron formations (Fe and Mn)	
b) carbonates		ferroan carbonates, dolostones, limestones	
c) sulfates		barite, gypsum	
d) evaporites		halides, borates	
e) carbonate skarns		diopside, tremolite, grossularite, wollastonite, scapolite	
The era of bio-mediated mineralogy (>2.5 Ga to present)			
7. Paleoproterozoic atmospheric changes surface oxidation	2.5 to 1.9 Ga	>2000 new oxide/hydroxide species, especially ore minerals	>4000
8. Intermediate ocean	1.9 to 1.0 Ga	minimal mineralogical innovation	>4000
9. Neoproterozoic biogeochemical changes	1.0 to 0.542 Ga		>4000
a) glaciation		extensive ice deposition, but few new minerals	
b) post-glacial oxidation		extensive oxidative weathering of all surface rocks	
10. Phanerozoic Era	0.542 Ga to present		4300+
a) biomineralization		extensive skeletal biomineralization of calcite, aragonite, dolomite, hydroxylapatite, and opal	
b) bio-weathering		increased production of clay minerals, soils	

Note: The timings of some of these stages overlap and several stages continue to the present.

Table 1. Prebiotically plausible mineral surfaces: Minerals identified in Eoarchaen (~4.0–3.6 Ga) mineral deposits. Table reproduced without permission from Science, 2012.[232]

Origins of Life scientists have acknowledged the catalytic capacity of mineral surfaces and hypothesized a non-trivial role of minerals in the chemical evolution theory.[78,79] One of the pioneers was Bernal in the 1950s, whose work was later re-articulated by Orgel as 'Polymerization on the rocks'.[80] Moreover, the catalytic effect of mineral surfaces was recently reviewed in detail by Lambert.[81] More recent studies have focused on the interactions of life's building blocks with minerals, in aims to demonstrate that the potential adsorption of these compounds (particularly nucleotides and amino acids) on a mineral surface could favour polymerisation.[82,83] Nonetheless, quite a variety of different parameters determine the adsorption rate of biological building blocks in mineral surfaces; such as the solubility, molecule size, mineral charge and experimental conditions (e.g. pH and temperature). This in turn gives a very complex array of conditions and consequently, some aspects remain largely unexplored. On the other hand, Lambert's review concludes that even if amino acid polymerisation can be favoured in the absorbed state, the resulting thermodynamic effect is problematic since it makes polymerization an overall slower

process. Conveying yet another 'goldilocks' problem, when considering the decreased activity of water (in which the polymerisation reactions depend on) in the presence of mineral-interfaces.[84]

The polymerization of amino acids on different environments (including mineral surfaces) has been explored systematically, in recent work by Surman. *et al.*[85] The work carried out demonstrated distinct product ensembles arise as an effect of environmental variations within the experiments, see **Figure 5**. By investigating the abiotic condensation of three amino acid monomers (Glycine – G, Alanine – A and Histidine – H), they could observe non-trivial variations in the distribution of the resulting peptides. An assessment in the distribution of the peptides was achieved through their subsequent characterization by HPLC-MS. In order to observe the differences in the non-enzymatic polymerization process, they had to reduce the dimensionality of the mass-spectral data by PC-DFA analysis of the data-sets. Equipped with experimental and instrumental triplicates, they validated a trend that differentiated the peptides depending on the environment employed, by either salts (**Figure 5a**) or mineral surfaces **(Figure 5b**) being introduced in the reaction vessel. The presence of inorganic compounds has been proved (in this sense) to impart a selectivity criterion, probably driven by both thermodynamic and kinetic factors. The alteration of the reaction conditions (i.e. pH and temperature) due to the presence of the inorganic moieties was not studied directly in this work, but it's hypothesised to have played an important role in the dynamics of the polymerization process. Also, whether the mechanism behind the apparent selectivity is primarily based on catalysis or an adsorption mechanism (particularly, in the case of the mineral surfaces), was not discussed.

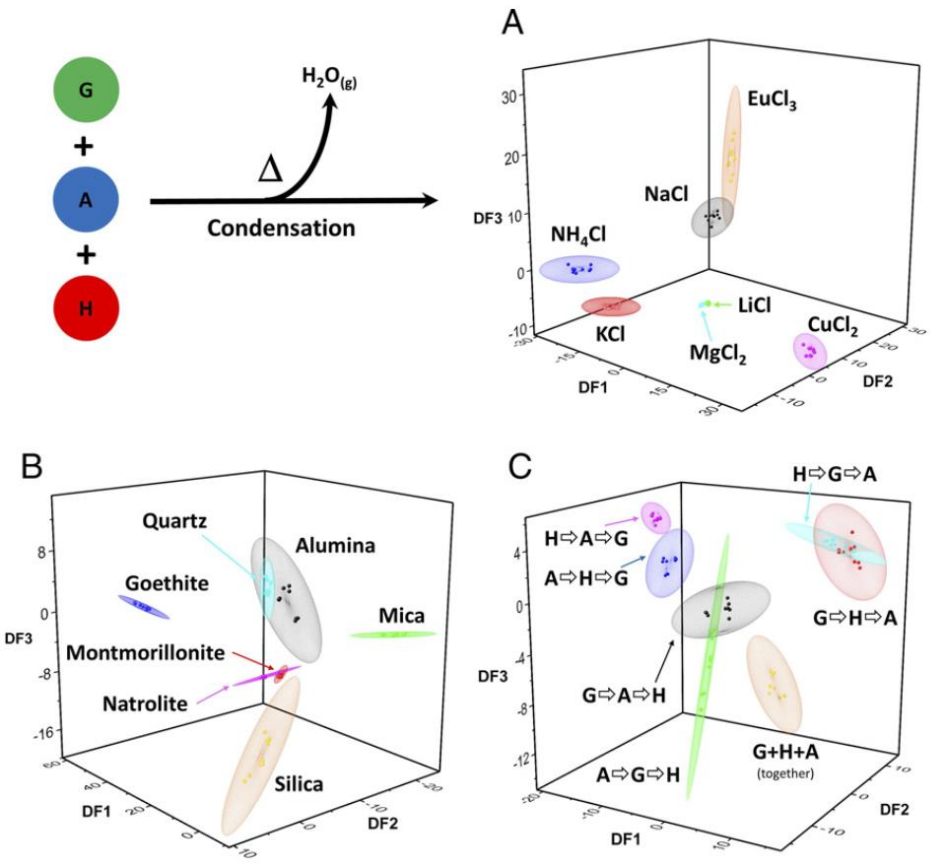


Figure 5. PC-DFA analysis of LC-MS data from condensation of G, A, and H in different environments/conditions: (**A**) different soluble salts, (**B**) different minerals, and (**C**) different mixing orders. Reproduced without permission from PNAS, 2019.[85]

As mentioned previously, there are examples of amino acid condensation on clay minerals (such as montmorillonite), which resulted in better yields and longer peptides than in Rode's experiments.[68] Also discussed, is the potential roles of clay minerals on the polymerization as protection against hydrolysis and improved directionality by surface adsorption.[86] Furthermore, the possible roles of mineral surfaces in the abiotic synthesis of RNA monomers, has also been explored to a similar extent. In this case it starts with minerals aiding the integrated synthesis of the RNA components from an abiotic route. For example, borate minerals aid in the stabilization of ribose in complex mixtures, making it possible for it to co-exist with the prebiotic synthesis of nucleobases and therefore enable nucleoside formation.[87] As well, Martin Ferus demonstrated that the meteorite-catalyzed condensation of formamide can effectively produce nucleosides, although in low yields.[88] In addition, phosphate minerals introduce an abiotic source of phosphorus, providing a pathway for the prebiotic phosphorylation of nucleosides into nucleotides (e.g. ribose, nucleobase and phosphate),[89,90,91] a vital step in the preparation of active RNA units.[92]

The potential role of mineral surfaces on protecting, selecting and catalysing reactions of prebiotic organic molecules remains one of the recurrent themes in the Origins of Life discussions. They are hypothesised to be an important stepping-stone in the transition from simple chemicals towards complex organic molecules, see **Figure 6**. Also, it must be noted that amongst the mentioned minerals, mica and other clays remain the most cited (e.g. worked with),[93] closely followed by various transition metals, such as Fe, Ni, Co and Cu.[94,95] Also, sulfide and borate minerals, have been proposed to have played key catalytic roles in prebiotic organic synthesis, either by a direct catalytic effect or the integration of multiple geo-chemical pathways.[96,97,98]

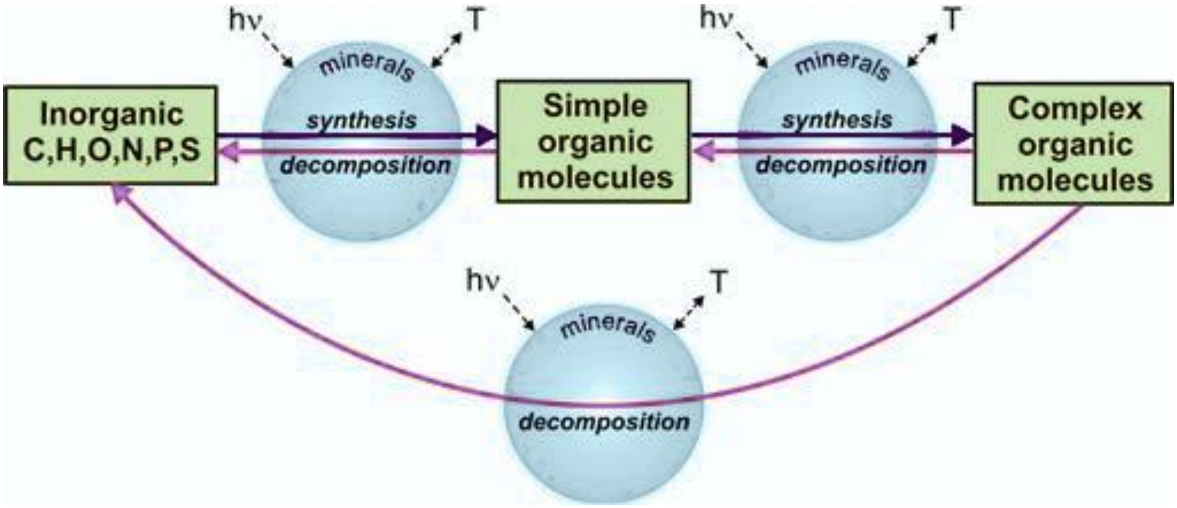


Figure 6. From inorganic matter to complex organic molecules, aided by the interaction with minerals surfaces. Reproduced without permission from AMBIO: A Journal of the Human Environment, 2004.[79]

1.3 Approaches to Prebiotic Chemistry

Putting it all together, as some may say, it's about taking life's building blocks (e.g. amino acids –or peptides-, ribose, nucleobases and phosphates) and combining them (in a specific set of conditions, or set of environmental variables) into life's basic machinery: RNA/DNA and/or proteins.[99] Answering what came first, RNA/DNA or proteins, is a rather difficult question when there is such a big jump from the abiotic synthesis of the building blocks (from simple chemistry) to the way the current metabolic machinery is constructed. For that reason, here is where we should start re-formulating the basic questions.

1.3.1 Autotrophic versus Heterotrophic Origins of Life

A main division in the Prebiotic Chemistry community lies in the answer of this question: what came first, RNA/DNA or proteins? This is also known as 'the chicken or egg problem'. The DNA/RNA hypothesis gained momentum primarily due to the necessity of heredity and a simple mechanism of replication,[93] which is countered by the catalytic capacity of proteins and their role on the regulation of metabolic processes, making a compelling case for peptides.[94] It is known that the storage of genetic information is the main function of DNA, but it does not exhibit functional or catalytic properties in modern cells. Furthermore, it requires highly complex and specific proteins to support its replication and in turn, proteins are synthesised according to the information stored in DNA. Therefore, it is not clear which one was more important in the path of chemical evolution.

The progress made during the 1950s on the abiotic synthesis of some of life's precursors enabled the idea of a bottom's up Origins of Life. A huge effort was then placed on the synthesis of the building blocks of life under prebiotic conditions, which will be discussed in greater detail in the following section. The initial success of these experiments was combined with recently acquired knowledge of the modern biological machinery and several theories on the Origin of Life were proposed. These theories can be reduced to two main categories: 'Heterotrophic' or 'Autotrophic', see **Figure 7**. Initially, a heterotrophic Origin of Life became generally accepted, as Stanley Miller himself supported it. Nonetheless, the resistance of complex prebiotic mixtures towards a higher degree of complexity, either by truncation of the combinatorial explosion or the spontaneous generation of polymeric material, was imminent in all following studies. This prompted the need of an alternative approach, which was introduced by Wächtershäuser in 1988.

He proposed an 'Autotrophic' Origins of Life, where the necessary bio-energetics to evolve the prebiotic mixtures into a higher degree of complexity, came from the integration of geochemical cycles with the abiotic synthesis of compound.[59] Not long after, the hydrothermal vents were discovered and Mike Russell developed a convincing theory behind the coupling of such mechanisms (see **Section 1.2.2).** The new way of driving chemical complexity, foresees the role of an evolving metabolism as a main requirement in the process of generating self-sustaining proto-organisms (i.e. autopoetic units, as Maturana and Varela described)[95], opposed to a Heterotrophic Origins of Life, where the spontaneous formation of polymers capable of selective self-replication, preceded the need of encapsulation. In this sense, the autotrophic versus heterotrophic Origins of Life are mainly distinguished by the

order and importance they give to specific mechanisms that drive the chemical evolution: one argues for a metabolism-first approach (e.g. autotrophic), where another puts self-replication at the beginning of the process (e.g. heterotrophic).[96]

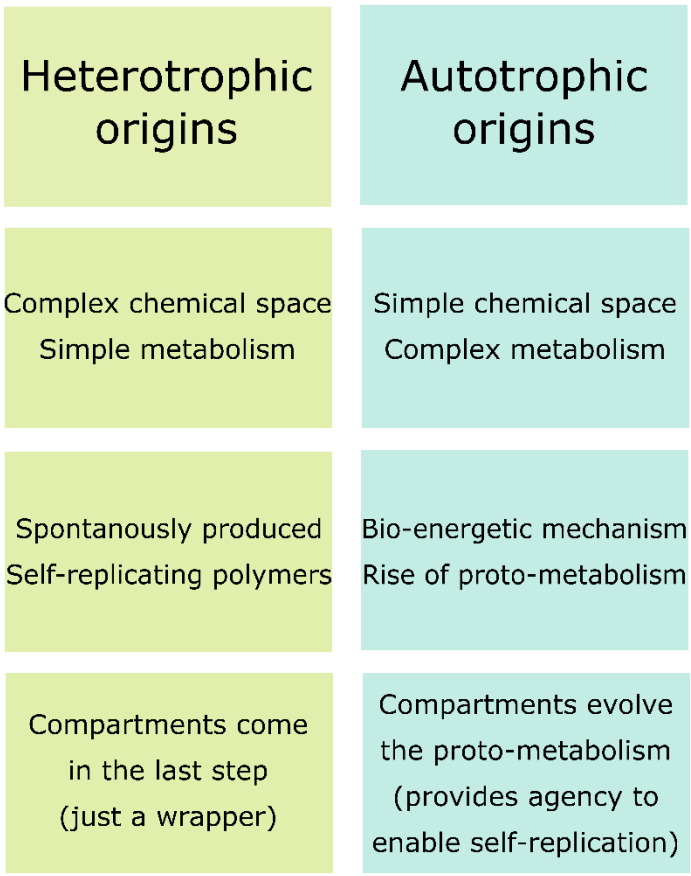


Figure 7. Main controversies on the Origins of Life: Was there a Heterotrophic or autotrophic Origins of Life? Adapted without permission from International Microbiology, 2005.[97]

1.3.1.1 Replication-first hypothesis

1953 was a good year for the Prebiotic Chemistry field, in the same year that the Miller-Urey experiment was published, Watson and Crick discovered the molecular structure of DNA,[41] a breakthrough in the elucidation of the mystery behind our genetic code. Not long after, Crick, Wöese and Orgel, proposed a theory for the origin of the genetic code.[100,101] Inspired in a recent discovery of tertiary structures could be made by single-stranded RNA oligomers, they argued that RNA could be the predecessor that stored both the genetic information and carry out metabolic functions. However, it took the discovery of the ribozymes and the capability of small-RNA units to act as enzymes, to give rise to the 'RNA world' theory as one of the most accepted theories for the Origins of Life.[102] Consequently, this provided the necessary evidence to suggest that RNA played a key role in the transition

towards the first living organisms, see **Figure 8** for a schematic of the 'RNA world' theory.[103]

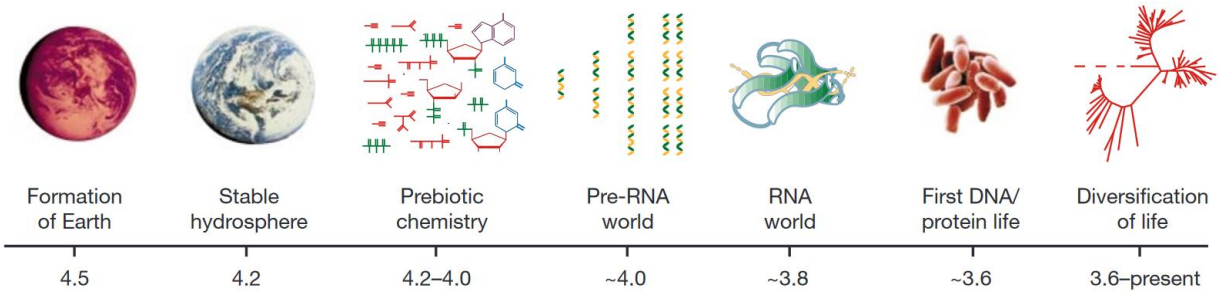


Figure 8. A historical time-line for the 'RNA-world' hypothesis. Reproduced without permission Nature, 2002.[233]

The theory has been experimentally challenged by organic chemists in the field and continues to be favoured by Origins of Life researchers with a biological background.[79] Nonetheless, the central dogma of biology does not account for the mayor hurdles that the non-enzymatic polymerization and synthesis of nucleic acid monomers has encountered. A problem of regio-selectivity, monomer concentration and phosphorylation has been a main driver of the disillusioned state of the RNA theory.[102,104] However, with every problem there has been a possible solution presented, maintaining the theory through the experimental hardships. For example, in work carried out by Braun *et al.*, they propose a plausible mechanism to fill the gap in the abiotic polymerization of RNA monomers by subjecting them to a spatially-confined thermal gradient,[105] see **Figure 9**. The escalated polymerization of nucleotides is made possible by their accumulation in the confinements, as well as, driven by thermophoresis and convection.

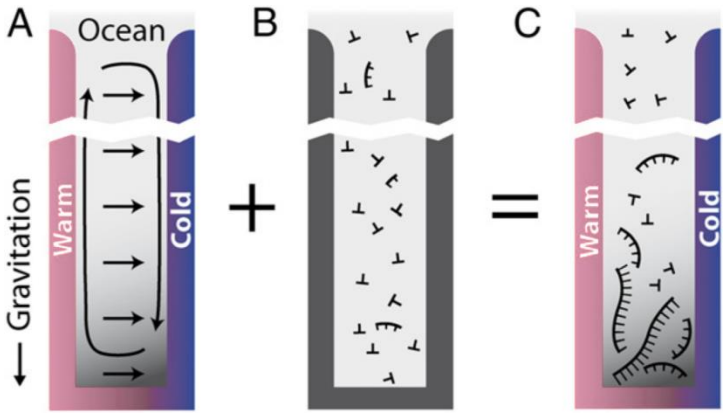


Figure 9. Thermal-cycling enables the abiotic (non-enzymatic) polymerization of RNA building blocks. Reproduced without permission from PNAS, 2013.[105]

In other words, despite being one of the most accepted theories, the 'RNA world' still faces fundamental problems that remain unsolved. For the last 50 years, many scientists have tried to synthesise RNA using a bottom-up approach, finding a rough patch even in the synthesis of polymers with comparable complexity.[106,107] An attempt to synthesise their monomeric building blocks alongside with their simultanous polymerization has not been accomplished yet, fuelling many arguments about the plausibility of these reactions and its current relevance in the Origins of Life theories.[104,108]

1.3.1.2 Metabolism-first hypothesis

All metabolism-first hypotheses revolve around the idea that chemical networks with a high degree of mutual catalysis between its components, allow for adaptation and evolution without the need of molecular recognition and replication.[109] An important 'Metabolism-first' hypothesis was proposed by Wächtershäuser in 1988 under the name of the 'Iron-Sulfur (Fe-S)' theory.[62] It differs from the 'Primordial Soup' theory, in that the main building blocks of life are not synthesised with external sources of energy (e.g. UV radiation, lightning), but with iron-sulfur clusters from the redox reactions of metal sulphides. The ideal scenario where these processes could have taken place is in deep-sea hydrothermal vents, as discussed by Russell *et al.*[61]

Also, it was later discovered that an iron sulfide mineral was capable catalysed a series of chemical reactions that create a reverse citric acid cycle, in which carbon monoxide is reduced to form complex organic molecules (e.g. acetate, pyruvate and others) that are central in the metabolism of current living systems. Moreover, it has also been shown that amino acids and dipeptides can result from this reaction.[110,111] As well, it connects back with

the autotrophic approach through the generation of Fe-S complexes capable of forming microscopic bubbles upon precipitation.[112]

Whilst this hypothesis has many associated problems (for example, it does not explain how these metabolic cycles could reproduce or evolve in the absence of genetic material), it provided an alternative framework for understanding the origin of life in the absence of polymeric species with a high-degree of molecular complexity, such as RNA.[113] A significant amount of research has focused on the reactions that occur in hydrothermal systems, not only for the development of proto-metabolic cycles, but also for understanding the missing-link between metabolism-first scenarios and replication-based life forms.[114,115] Also, in a recent communication by Muchowska et al., a non-enzymatic reaction network with remarkable similarities to the Tri-Carboxylic Acid (TCA) cycle, was demonstrated (see **Figure 10)**.[116]

The abiotic reaction network was able to replicate 7 of the 11 reactions in the TCA cycle and 9 of the 11 intermediates (e.g. only oxalosuccinate and citrate are missing). If that was not impressive enough, the same work presented that much of the glyoxylate cycle (e.g. another central metabolic pathway in living systems) was also achieved, including 8 of its 9 intermediates (e.g. only citrate is missing) and 5 of its 8 reactions. These results give rise to a set of abiotic chemical pathways that resemble the core of carbon biochemistry, that have been exclusively promoted by ferrous iron. Moreover, it has been experimentally proven that sulfate radicals generated from peroxydisulfate, can interconvert the Krebs cycle precursors.[117] Demonstrating, yet again, the beneficial coupling of a ‘prebiotic’ chemical pathway with abiotic sources.

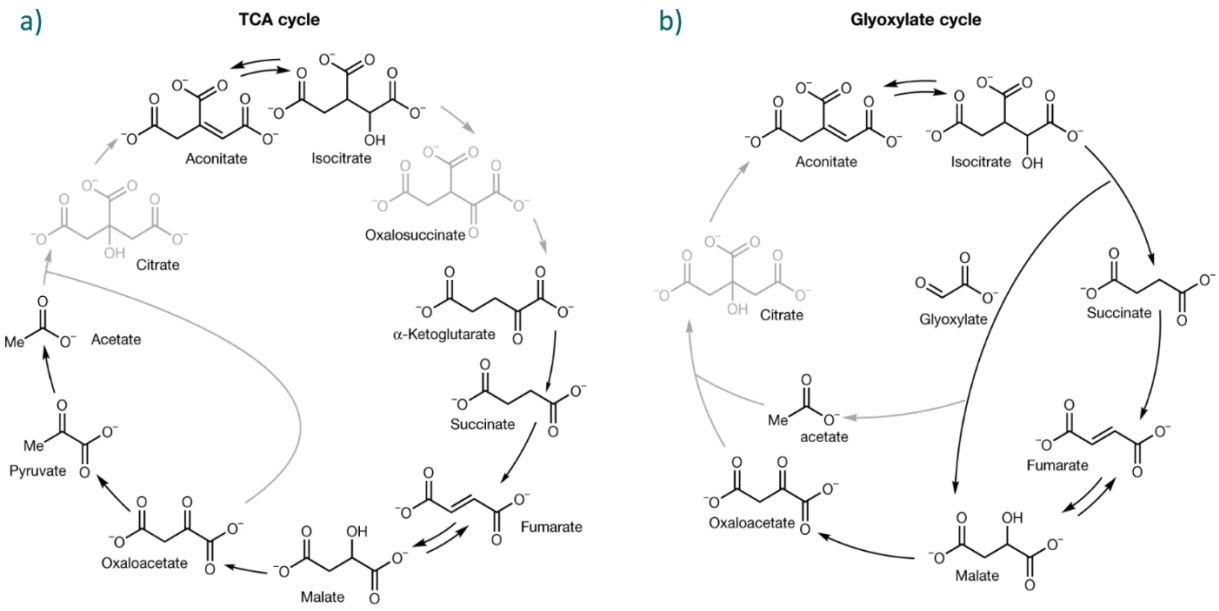


Figure 10. A comparison across the TCA (**a**) and glyoxylate cycles (**b**): Highlighted in grey are the compounds only found in the biological cycle. Abiotically reproduced material is shown in black. Reproduced without permission from Nature, 2019.[116]

1.4 One-pot synthesis of building blocks

The great majority of biological building blocks can be synthetized by electric-discharge experiments and reactions of simple precursors. A source of carbon and nitrogen under the right conditions are enough to promote the formation of amino acids, sugars and nucleobases as seen in several experiments described in this section. However, the synthesis of biopolymers from their monomeric units is achieved by the loss of a water molecules to form glyosidic bonds in the case of sugars and amino acids, or ester bonds in the case of nucleic acids. In our biological pathway, this condensation reaction happens in aqueous media despite it being thermodynamically unfavourable. Through the course of evolution, specialized enzymes where developed, which provided the energy necessary (e.g. by the hydrolysis of pyrophosphate bonds) to achieve the polymerization.

However, chances are that the modern enzymatic process was perfected over billions of years and therefore it points towards the existence of an alternative pathway in the chemical evolution of these systems, which allowed the polymerization of the monomeric species generated from prebiotic broths. This mechanism remains unknown and connects back to our limited knowledge on the greater part of the resulting product space of the primordial soup. The following sections will discuss what we have learn on the composition of the highly convoluted mixtures, by the application of different analytical approaches. For a (rough) timeline of the abiotic synthesis of known biological units, see **Figure 11** below.

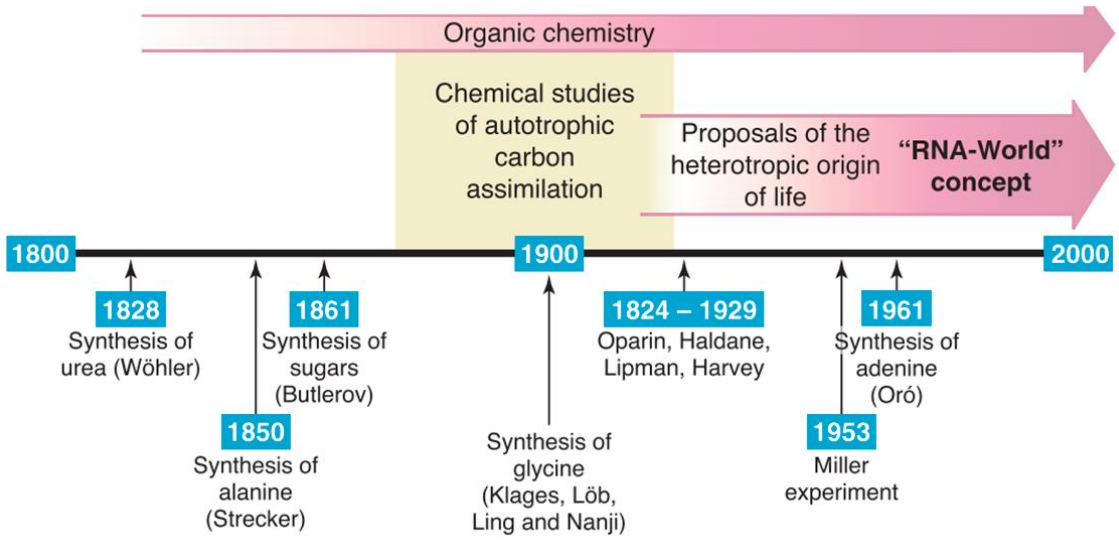


Figure 11. Historical milestones in the prebiotic synthesis of biological building blocks. Reproduced without permission from Science, 2003.[234]

1.4.1 Miller-Urey experiment: *Amino acids*

Stanley Miller and Harold Urey from the University of San Diego, California, decided to put the Oparin-Haldane theory to the test by assembling an apparatus that allowed them to emulate the early earth atmosphere and primitive 'ocean' containing a mixture of simple gases (CH_4, NH_3 and H_2), hot water and an electrical discharge as an energy source, in what came to be known as the Miller-Urey experiment, see **Figure 12**. [18]

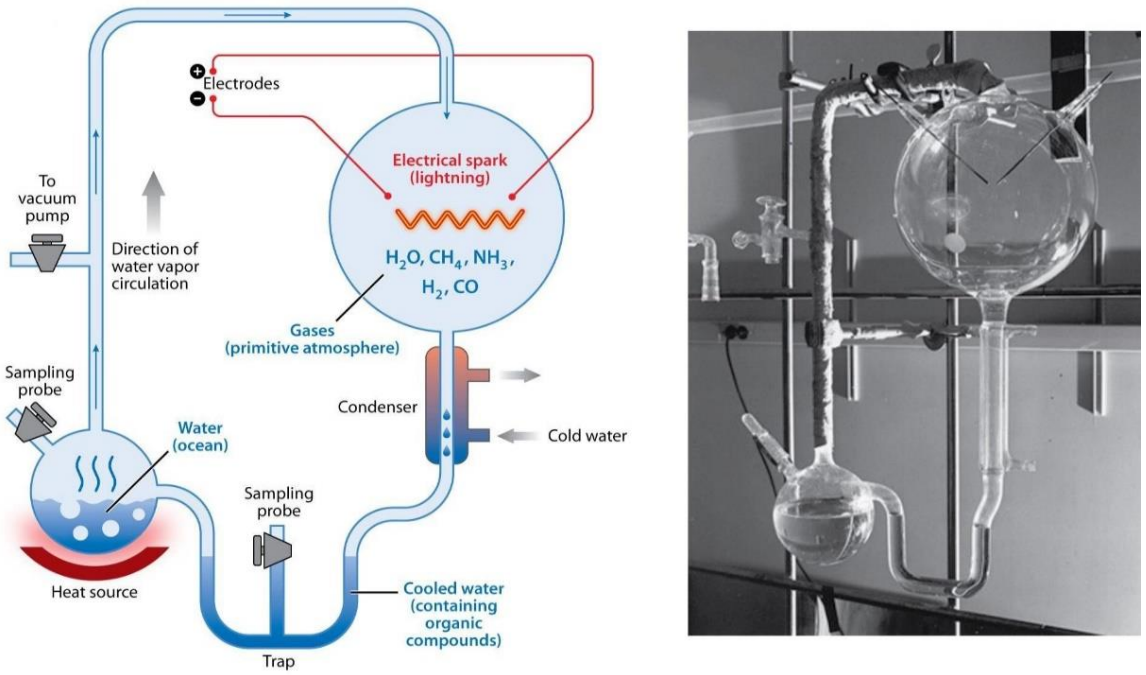


Figure 12. The spark-discharge apparatus used in the Miller-Urey experiments. (**a**) Schematic drawing of the apparatus. (**b**) Photo of the apparatus taken by Stanley Miller Reproduced without permission from Evolution: Education and Outreach, 2012.[235]

Their results demonstrated that a successful synthesis of small organic compounds under a primitive earth environment was possible, a great example of the inorganic to organic transition, mediated by natural geological processes. However, the greatest significance of the experiment relies in the compounds that they managed to identify through (a now rudimentary analytical technique) paper chromatography: amino-acids. The amino acids are known to be the constituent of proteins, an essential part of all known living systems. Proteins drive our metabolic machinery and are formed through elongated peptide units.

However, even if the Miller-Urey experiment succeeded in the prebiotic synthesis of amino acids, their conversion into peptides is still a challenge. Since then, several Miller-Urey type experiments have been conducted, with different atmospheric compositions and energy sources, in order to satisfy the variety of prebiotic possibilities.[118,119,120] See Table 2 for a list of detected amino acids in a selection of Miller-Urey type experiments.

Author(s)	Reactants	Energy source	Results reported	*
Miller [81]	CH_4, NH_3, H_2O, H_2	Electric discharges	Simple amino acids organic compounds	(−)
Abelson [82]	CO, CO_2, N_2, NH_3, H_2, H_2O	Electric discharges	Simple amino acids HCN	(−)
Garrison et al. [83]	CO_2, H_2O	40 MeV Helium ions	Formic acid, formaldehyde	(+)
Bar–Nun et al. [84]	CH_4, NH_3, H_2O	Shock wave	Simple amino acids	(+)
Harada and Fox [85]	CH_4, NH_3, H_2O	Thermal energy (900–1200°C)	14 of the 'essential' amino acids of proteins	(+)
Lawless and Boynton [86]	CH_4, NH_3, H_2O	Thermal energy	Glycine, alanine, aspartic acid, β-alanine, *N*-methyl-β-alanine and β-amino-*n*-butyric ac.	(+)
Groth and Weyssenhoff [87]	CH_4, NH_3, H_2O	Ultraviolet light (1470 and 1294 Å)	Simple amino acids (low yields)	(+)
Sagan and Khare [88]	CH_4, C_2H_6, NH_3, H_2O, H_2S	Ultraviolet light (>2000 Å)	Simple amino acids (low yields)	(+)
Oro [34]	CH_4, NH_3, H_2O	Plasma jet	Simple amino acids	(−)
Yoshino et al. [28]	H_2, CO, NH_3, montmorillonite	Temperature of 700°C	Glycine, alanine, glutamic acid, serine, aspartic acid, leucine, lysine, arginine	(−)
Yanagawa et al. [89]	Various sugars, hydroxylamine, inorganic salts	Temperature of 105°C	Glycine, alanine, serine aspartic acid, glutamic acid	(+)
Pavolvskaya and Pasynskii [90]	Formaldehyde, nitrates	High pressure Hg lamp (photolysis)	Simple amino acids	+
Bahadur et al. [91]	Formaldehyde, molybdenum oxide	Sunlight (photosynthesis)	Simple amino acids	+
Kobayashi et al. [92]	CO, N_2, H_2O	Proton irradiation	Glycine, alanine, aspartic acid, β-alanine, glutamic acid, threonine, α-aminobutyric acid, serine	+
Palm and Calvin [93]	CH_4, NH_3, H_2O	Electron irradiation	Glycine, alanine, aspartic acid	(+)
Hanic et al. [7]	CO_2, N_2, H_2O	Electric discharges	Several amino acids	+

* The probability to occur according to recent geochemical knowledge.
(−), little or no probability; (+), possible under special circumstances, +, possible.

Table 2. . Amino acids synthetized under assumed prebiotic Earth conditions. [17]

Throughout the years of Origins of Life research, we have learned that there are two main challenges for the one-pot synthesis of amino acids and their development into more complex molecules (e.g. peptides). First, it appears to be the limited amount of hydrogen cyanide (HCN) that is formed, which limits the amount of amino acids that can be produced. Since HCN is a central intermediate in the Strecker amino acid synthesis and also an important precursor for the synthesis of nucleobases, it has therefore been considered of great importance in the one-pot synthesis of life's building blocks.[121] Miller had proposed that aldehydes and hydrogen cyanide were synthesised in the gas phase by exposure to the spark. Then, these would further react in the aqueous solution in the presence of ammonia to produce α-aminonitriles and cyanohydrins (see **Figure 13**). The slow hydrolysis of these products would finally yield a mixture of α-amino acids and α-hydroxy acids.[122]

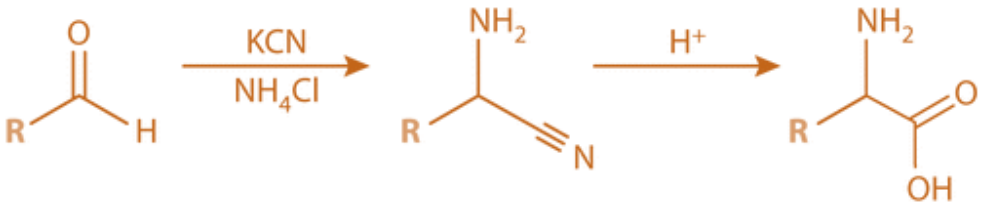


Figure 13. Reaction pathway for formation of α-amino acids by the Strecker synthesis: An aldehyde reacts with ammonium and cyanide (e.g. potassium cyanide) to form an aminonitrile, which is then hydrolysed and consequently oxidized, to form an amino acid.

Furthermore, a recent study was demonstrated that Miller–Urey experiments in a neutral atmosphere are capable of producing RNA nucleobases, as well as, identifying formamide (i.e. the hydrolysed form of HCN) as one of the main atmospheric products in discharge experiments, when laser-driven plasma impact simulations were carried out.[43] This research addresses the chemistry of a mildly-reducing/neutral atmosphere (NH_3 + CO + H_2O) and the role of formamide as an intermediate of nucleobase formation in Miller–Urey experiment. Therefore, expanding the amount of biomolecules deemed to be possible in these type of prebiotic experiments, unlocking the curiosity of prebiotic chemist on how much more can we get from prebiotic broths? Particularly, how complex are the molecules synthesised by these experiments and which are the main players in the game of chemical evolution that are yet to be discovered within the complex mixtures. A re-analysis of the original Miller-Urey experiment, alongside newly found samples from an old experiment of Miller's which took into consideration volcanic emissions, was carried out with state-of-the-art analytical techniques (e.g. HPLC-FLD-MS/MS) .[123] The results unravelled a difference in the yields obtained for the detected compounds, based on the initial differences of the atmospheric compositions (e.g. gas composition used in the experiment) and where compared to those obtained for an extract of the Murchinson meteorite, see **Figure 14**.

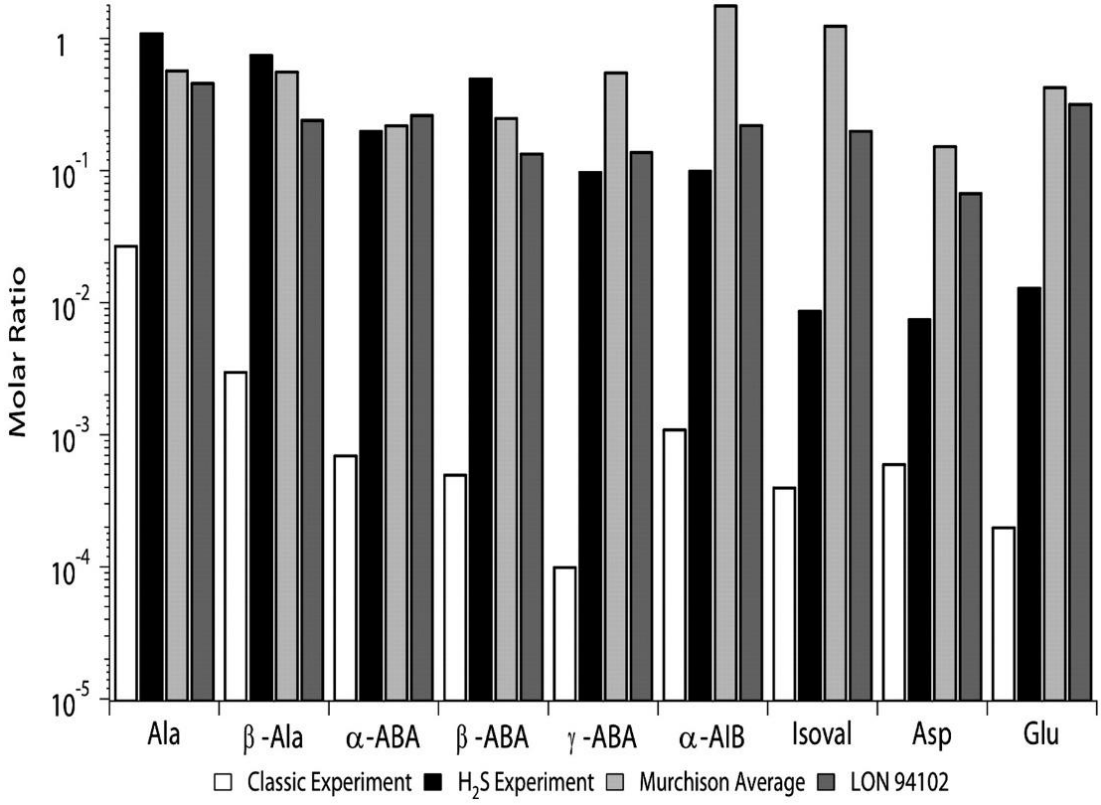


Figure 14. Comparison of amino acid molar ratios (*relative to glycine = 1*) found in Miller's H_2S and classic spark discharge experiment and the Murchison meteorite. Reproduced without permission from Science, 2008 [123]

Additionally, in a more recent communication by Parker *et al.*, a Miller-Urey experiment was carried out with the inclusion of cyanimide. Through the application of modern analytical techniques and a targeted mass-spectral acquisition, they could expand the variety of amino acids detected (i.e. finding proline and histidine), several dipeptides and diketopiperazines. Also, a study by Rodriguez *et al.* reported the reactivity of numerous N-heterocycles in a Miller-Urey spark discharge experiments (with either a reducing or neutral atmosphere), by adding the isolated compounds directly into the prebiotic broth as a way to investigate how N-heterocycles are modified under plausible prebiotic conditions.[124] Modern analytics save the day again, this time through tandem mass spectrometry and nuclear magnetic resonance spectroscopy, which allowed them to generate plausible reaction pathways for the newly formed products, as well as, the identification of a Peptide Nucleic Acids (PNAs) amongst the detected products.

Yet another remarkable example of advanced analytical techniques aiding in the elucidation of the complex product mixture of the Miller-Urey (MU) broth, was provided by Wollrab *et al.* Through the application of High Performance Liquid Chromatrography – High Resolution Mass Spectrometry (HPLC-Orbitrap) coupled to a computational model and time –resolved sampling, they were able to demonstrate that the mass-density in the resulting product distribution of the MU experiment is comparable with the one of the Beilstein database, which is the database that compiles all known organic compounds known to man,[125] driving the conclusion that the prebiotic broth is quite a messy system and encompasses a broader distribution of products than originally thought.

On the other hand, recent development has been focussed on finding an abiotic route to peptide formation in water. The abiotic synthesis of peptides is thermodynamically unfavoured in aqueous solutions and therefore is represents another unsurmountable challenge. However, by coupling the production of amino acids to wet-dry cycles that resemble a natural environmental cycling process (e.g. the water cycle), we could assume the potential of peptide formation. See **Figure 15**. Several approaches have been taken to tackle this problem, which include the use of a variety of catalysts (e.g. clays, metal oxides) and a volcanic gas (carbonyl sulfide, COS).[126,127,128] However, a recent publication by Rodriguez *et al.* managed to present an even simpler solution.[129] They discovered that the abiotic synthesis of peptides from isolated amino acids, by simple wet-dry cycles in relatively mild conditions, yielded oligopeptide chains in around 50% yield. A digital recursive reactor system was developed to investigate the process and allowing for a good variety of parameters to be explored: temperature, number of cycles, cycle duration, initial monomer concentration and pH. The length of the peptides –*up to 20*- was longer than ever reported for such simple conditions and particularly, in the absence of any catalyst. As well as, consistent with many other types of amino –acids (e.g. Ala, Asp, Glu, His, Lys, Pro, Thr and Val).

Figure 15. Example of an amino acid condensation, or peptide formation by water removal. Adapted from without permission from Nature Communications, 2015.[129]

An alternative approach to peptide synthesis in aqueous solutions was presented by Powner *et al.*, where they were capable of bypassing α-peptide ligation in water, which is known to be problematic for some of the 20 proteinogenic amino acids. All by developing a chemoselective, high-yielding α-aminonitrile ligation that exploits only prebiotically plausible molecules: hydrogen sulfide, thioacetate, and ferricyanide (or cyanoacetylene) to yield α-peptides in water. See **Figure 16**. To make things even better, the ligation was shown to be extremely selective for α-aminonitrile coupling and tolerates all of the 20 proteinogenic amino acid residues.

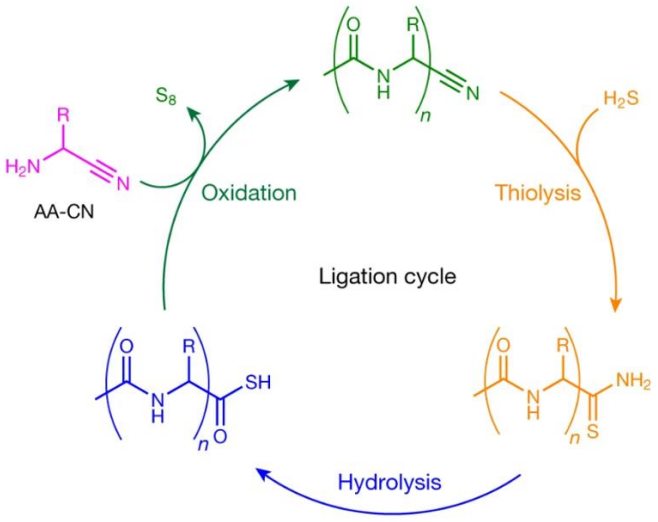


Figure 16. Sulfide-mediated α-aminonitrile ligation: Iterative AA-CN (*pink*) ligation to give N-acetyl peptide nitriles (*green*) by sequential thiolysis, hydrolysis and AA-CN ligation. Reproduced without permission from Nature Chemistry, 2019.[236]

Other mechanisms for the formation of polymeric species from the Miller-Urey products have been discussed recently. In a series of publications from the Nicholas Hud lab, an investigation on the formation and characterization of oligoesters in far-from-equilibrium conditions was carried out.[130] A general procedure for the synthesis of these polymers consists in the one-pot reaction of L-malic acid monomers subjected to a series of wet-dry cycles.[131,132] The synthesis of ester-based polymers holds through prebiotic plausibility, since the monomers used have been characterized previously amongst variations of Miller-Urey type experiments. In addition, recent work presented by Jia *et al.* demonstrated the spontaneous generation of membraneless compartments, by the abiotic condensation (e.g. wet-dry cycles) of α-hydroxy acids (αHAs),[133] compounds which are generally co-produced along with α-amino acids via the Strecker synthesis.

The polymerization process is simple and rapid, yielding both homo and hetero-polyester within a week at 80 °C. Most notably, they found that upon addition of 4:1 (v/v) water/acetonitrile to the dried polyester samples and sonication followed by vortexing, it resulted in turbid solutions. The solutions were then analysed by optical microscopy (see **Figure 17**), finding spherical microdroplets in four out of the 5 αHAs studied. The microdroplets were resistant towards repeated dilution (e.g. water addition), suggesting that they could withstand natural processes such as periodic rainfalls. These results are the first of their kind, proving how a simple procedure can lead to the formation of suspended individuals from prebiotically plausible compounds, a transition needed for the development of autopoietic units.

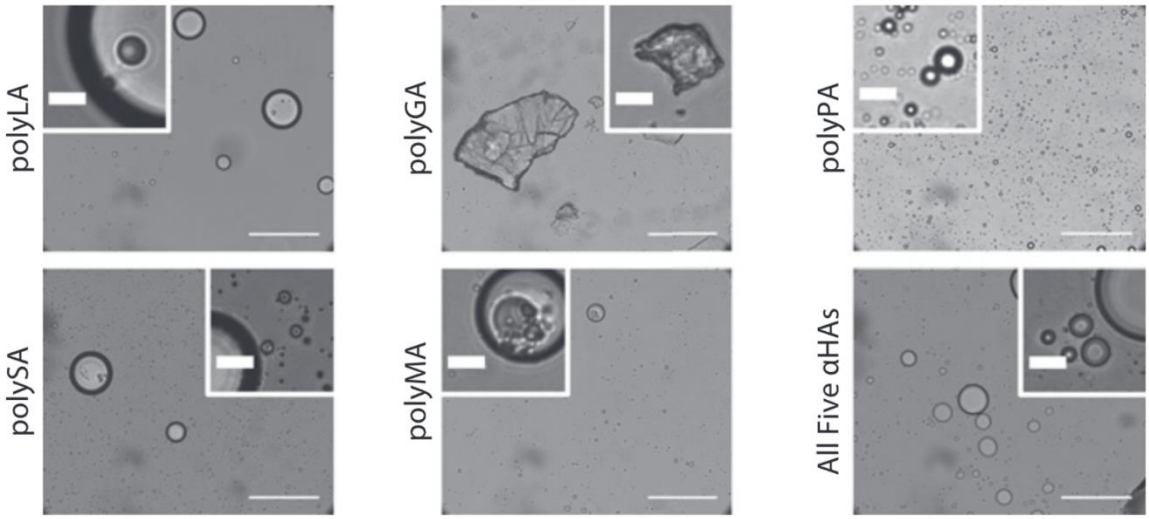


Figure 17. Optical microscopy images of microdroplets generated from abiotically synthetized polyesters. Reproduced without permission from PNAS, 2019.[133]

Finally, beyond the many experiments executed to investigate under which conditions amino acids and other relevant monomers might be produced endogenously, the discovery of a rich-variety of organic compounds in meteorites or carbonaceous chondrites unravelled the possibility of an extra-terrestrial delivery.[134] Analysis conducted on samples of the Murchinson meteorite, resulted in the identification of non-racemic amino acids, mono-carboxylic acids, hydrocarbons and high-molecular weight insoluble material (e.g. tholin analogues).[135,136] Even more so, nucleobases have also been identified within the complex distribution of organic compounds found in the most studied meteorite in the world (e.g. Murchinson, after the Australian city where it originally landed).[137]

Also, the possibility that amino acids could be synthesised in an interstellar medium has been discussed. Through the Ultra-Violet (UV) irradiation of interstellar grain-dust analogues, a variety of biological relevant compounds have been successfully produced: sugars, aldehydes and amino acids.[138,12,139] The possibility of complex organic material generated in outer space being delivered to the early earth during the late-heavy bombardment (LHB) period, has been called the theory of 'Panspermia'.[140]

1.4.2 Formose Reaction: *Sugars*

In 1861, a Russian chemist named Alexander Butlerow, made a 'sweet' discovery. He boiled formaldehyde with calcium hydroxide in water, which then turned into a yellowish-brown mixture that smelled like 'caramel' and tasted like liquorise.[141] The reaction has come to be known as the formose reaction and is significant due its capacity to produce a combinatorial explosion of sugars (carbohydrates). Beginning with two formaldehyde molecules, the reaction condenses into glycolaldehyde (1) which then further reacts (in an aldol reaction) with another formaldehyde molecule, making glyceraldehyde (2). Next, an aldol-ketose isomerization of the glyceraldehyde, forms dihydroxyacetone (3), which can react with a formaldehyde molecule and produce tetrulose, followed by aldotetrose, that can also split into glyceraldehyde (2) in a retro-aldol condensation. See **Figure 18**, for a scheme of the reaction mechanism.

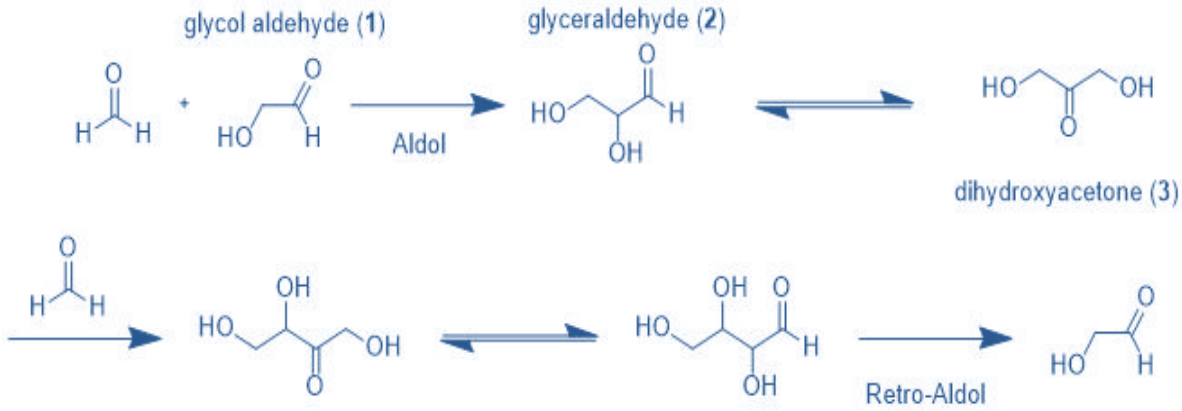


Figure 18. Original mechanism for the Formose Reaction, as proposed by R. Breslow.[32]

The reaction intermediates, glycoaldehyde and glyceraldehyde, have been known to catalyse the reaction by means of an autocatalytic cycle, see **Figure 19.**[142] This makes the first condensation of two formaldehyde molecules, the slow step responsible for the induction period. Consequently, in many reported experiments, this is circumvented by introducing directly any of the first two intermediates. However, the reaction networks in the complex mixture of carbohydrates are (in many ways) still unknown. For example, a recent deuterium study indicated that the original mechanism for the Formose reaction proposed by Breslow was not exactly correct.[143] Besides, the reaction quickly turns to 'tar', challenging its capacity to create biologically relevant sugars such as ribose. As an integral part of DNA, an abiotic source of ribose would be of great importance in the emergence of life as we know it. Many efforts have been made to drive this combinatorial explosion towards the preferred product (ribose), the most successful being the addition of borate minerals.[144] Nonetheless, the inclusion of silicates[74] and freeze-thaw cycles[145], have also accomplished satisfactory results.

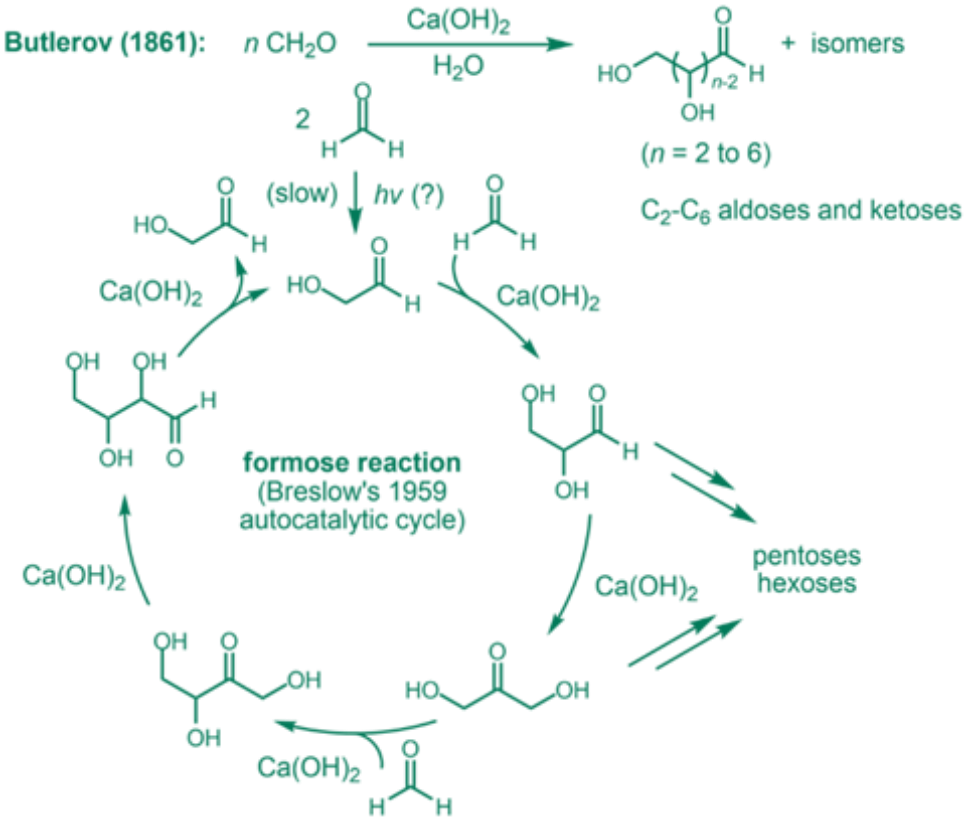


Figure 19. Autocatalytic cycle in the Formose reaction. Reproduced from Wikipedia Commons Creative Common (CC) 4.0, Public Domain.

Furthermore, as underlined by Orgel in a 2008 communication, the only (currently) known prebiotic example of an autocatalytic system is the formose reaction.[101] Another example is provided by Weber *et al.* with a triose-ammonia reaction, where a mixture of glyceraldehyde with ammonia produced pyruvaldehyde and a complex mixture of nitrogen-containing compounds.[147] As well, Schwartz *et al.* demonstrated the formation of an HCN tetramer from the reaction of formaldehyde in the presence of HCN.[146]

1.4.3 HCN and Formamide condensation: *Nucleobases*

The synthesis of adenine from ammonium cyanide by Juan Oró in 1961, paved the way to explore plausible synthetic pathways in the synthesis of nucleobases (e.g. a main constituent of DNA/RNA monomers) from very simple compounds.[148] However, the synthesis of such monomeric units, nucleotides (nucleobase, sugar and phosphate) or even nucleosides (nucleobase and sugar) in a one-pot fashion proved to be technically challenging.[149] This motivated a series of prebiotic chemists sought to find alternative ways to make those compounds, which would not necessarily agree with the pre-established ideas on prebiotic plausibility.[150]

HCN is one of the most studied chemical precursors of biomolecules: it is involved in the formation of amino acids *via* the Strecker synthesis, as well as in the formation of nucleobases.[151] However, in recent years, there has been an increasing interest in the chemistry of formamide. Unlike HCN, formamide can act as a solvent as well as a reactant for the synthesis of a variety of biochemical compounds.[152] Formamide results from the hydrolysis of HCN and recent studies suggest that it is a ubiquitous molecule in the Universe.[153] Initially, the potential role of formamide in prebiotic chemistry was considered to be limited since preliminary studies only showed the formation of a small number of heterocycles including adenine from UV irradiation.[154] However, studies by Saladino *et al.* have demonstrated that heating formamide in the presence of different catalysts of terrestrial and meteoritic origin yields complex combinations of nucleobases, amino acids, sugars, amino sugars and condensing agents,[88,149,155] as well as demonstrating a selective synthesis of certain nucleobases when mineral surfaces are added to the reaction, see **Figure 20**. Also, this was followed by an even more interesting observation of acylonucleoside formation, which was achieved by simply heating formamide in the presence of titanium oxide (e.g. TiO_2).[156] The resulting acyclonucleosides can be further phosphorylated in the presence of a phosphate source to yield 2',3'-and 3',5'-cyclic nucleotides. Despite obtaining overly-

convoluted mixtures of acyclonucleotides along with nucleobases, the simplicity of this-one pot reaction differs strikingly with the multi-step syntheses described by Sutherland and Carell.[157,106] Moreover, a recent study illustrates how an extremely rich variety of relevant prebiotic compounds can be obtained when liquid formamide is irradiated by high-energy proton beams in the presence of powdered meteorites.[158] The products obtained were amino acids, carboxylic acids, nucleobases, sugars and most notably, the four nucleosides: adenosine, thymidine, cytidine and uridine.

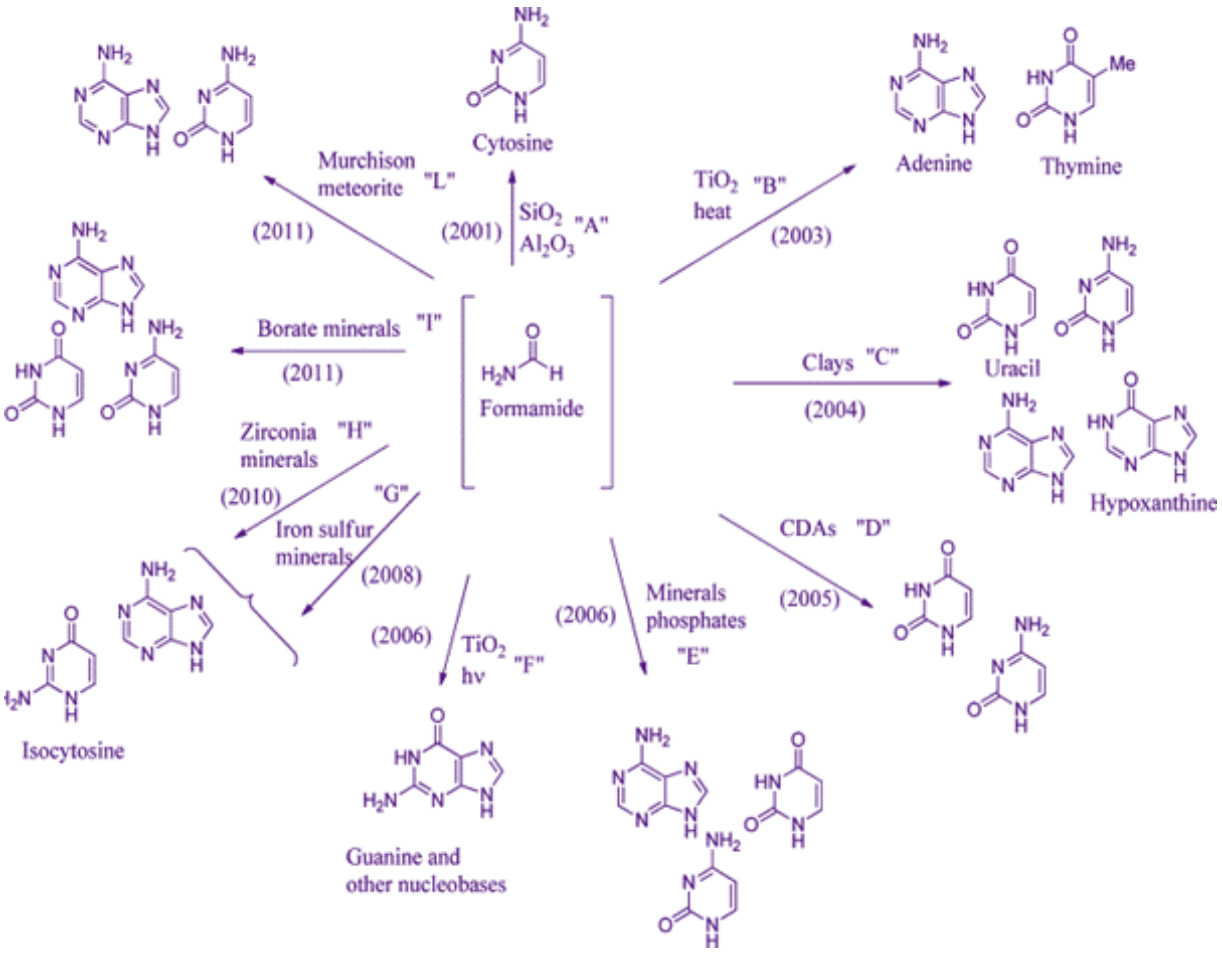


Figure 20. The basic prebiotic chemistry of Formamide into nucleobase synthesis, in the presence of different catalysts. Reproduced without permission from Biochimie, 2017.[237]

A large amount of evidence supports that numerous biochemical monomers can be abiotically synthesised and that similar processes could have occurred, either under early earth or space-wise conditions. However, a question remains: How might abiotically available monomers react to produce the two types of polymers that are essential to life's genetic code (RNA/DNA), in a one-pot fashion?

Nearly 40 years ago, Schoffstall and his co-workers used formamide as a solvent to permit the phosphorylation of nucleosides by inorganic phosphate to give nucleoside phosphates, which cannot be easily created in water, due to their thermodynamic instability with respect to hydrolysis,[159] namely, the ‘‘water problem’’ in prebiotic chemistry. More recently, work by Benner *et al.* also showed that borate could stabilize certain sugars against degradation (i.e. the ‘‘asphalt problem’’).[99] This was followed by work from Furukawa *et al.*, where they combine the two concepts to show that borate can work in formamide to guide the reactivity of nucleosides under conditions where they are phosphorylated.[91] In particular, this work showed that the reaction of adenosine in formamide with inorganic phosphate and pyrophosphate in the presence of borate gives adenosine-5-phosphate as the only detectable phosphorylated product.

In another recent work, carried out by Powner and Sutherland, the synthesis of RNA monomers was achieved by prebiotically plausible reagents.[92] However, this has generated quite a controversy amongst researchers that believe in the abiotic formation of RNA as a continuous process from simple precursors, without human intervention (e.g. not a multi-step procedure). Whilst Sutherland’s work does presents an elegant solution to the problem, it creates a conflicting scenario considering the low complexity of the system when compared to a realistic product distribution of the prebiotic mixtures that enclose the selected starting materials. The reagents used in their synthetic approach have been proved to be present in prebiotic broths but not in isolation, but rather as a minor constituent of a large chemical space. This does not address whether such units could be achieved by a sequential process of chemical selection driven by natural processes, but rather a highly restrictive environment with questionable prebiotic relevance. However, we must acknowledge that this approach has led Sutherland to successfully show that ribonucleic acids, amino acids and lipids can all be simultaneously produced as a consequence of hydrogen cyanide and hydrogen sulphide photochemistry.[160] Sutherland’s work involves a complex network of independent reactions and cannot be performed in a ‘one-pot’ fashion, but the author has proposed a geochemical model in which the three-atom precursors (HCN and H_2S) would react in different micro-environments to yield the main components behind biological metabolism, genetics and spacial-separation. Nonetheless, these assumptions can be problematic for a *‘true’* bottom’s up approach to prebiotic synthesis. With that in mind, see **Figure 21**, for a summary of all the multi-step efforts made on the abiotic synthesis of RNA molecules.

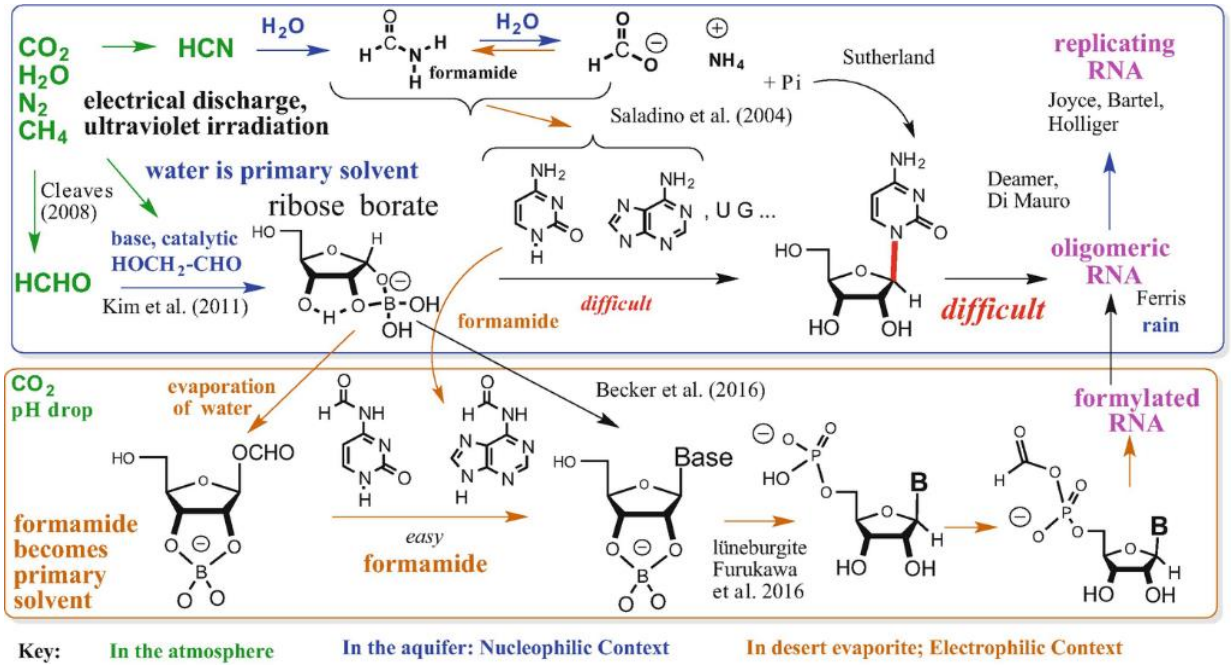


Figure 21. A schematic description of the main interconnections in different environments for the prebiotic synthesis of RNA monomers Reproduced without permission from Prebiotic Chemistry and Chemical Evolution of Nucleic Acids, 2018.[215]

1.5 Systems Chemistry

Since the pioneering studies in prebiotic chemistry, very specific conditions and starting materials were selected for the reactions. These studies were always trying to prove or follow a certain hypothesis or theory, which limited the possibilities for the synthesis of precursors and their polymerization. Systems Chemistry is the science of studying networks of interacting molecules and aspires to create new functions from an ensemble of molecular components, at different hierarchical levels and with emergent properties.[161] In the context of prebiotic chemistry, it aims to extend the studies of different environmental conditions for the synthesis of abiotic material in a systematic way.[162]

The "systems chemistry approach" was first described and openly used by J. Sutherland *et al.* in 2015 towards the synthesis of life's building blocks using a bottom-up approach.[163] These group of researchers used HCN and H_2S as simple reactants in a series of interconnected chemical reactions aimed at the formation of life's precursors, such as amino acids, lipids and mononucleotides. However, the prebiotic plausibility in this work was criticised due to the purification steps included in between synthetic steps. Despite criticism, Sutherland's research is a guide towards the use of heterogeneous mixtures for the formation

of more complex molecules inside the reaction media, which could present interesting physical properties (i.e. aggregation) or reactivity (i.e. polymerization).

Collaborative effects in the synthesis of life's building blocks have also been studied in the past. For example, it has been demonstrated that the addition of amine and ammonia into a mixture of glycolaldehyde and formaldehyde catalyses the formation of sugar.[164] The use of compartmentalization (vesicles) promotes the synthesis of peptides in the presence of dipeptide, and also changes the product distribution in formose reaction.[69,165] These examples show how systems chemistry can lead to the formation of interesting prebiotic molecules when a mixture of simple reactants is used as the starting material, which otherwise would not occur if the same reaction was accomplished with just pure reactants. In the last years, "prebiotic systems chemistry" has become the answer to the problems posed by the classical approach, where prebiotic chemistry is studied in very specific and constraint conditions by plausibility.[166] This 'new' field studies multiple reaction parameters and reactants simultaneously, being able to explore new properties that arise from chemical interactions and cooperative effects. Also, in the last decade, analytical techniques (such as multidimensional NMR and coupling of HPLC-MS) have experienced considerable improvement, allowing "prebiotic systems chemistry" to detect smaller changes of individual chemicals inside a complex mixtures.

1.6 Prebiotic mixtures: *An analytical challenge*

The continued development of analytical techniques and methods, enabled us to improve our understanding of the chemical processes (and stochasticity) in the complex chemical systems involved in life's origins. The one-pot synthesis of life's building blocks, in all its possible scenarios, produces a highly convoluted chemical mixture. However, the rigorous chemical specificity needed for life's building blocks (e.g. D-sugars, L-aminoacids), presents a big challenge when countered with the thousands of compounds generated in prebiotic soups. How did a complex chemical system that produces a large variety of compounds, where the precursors to life's building blocks are in relatively low yields, evolve to give life?

1.6.1 Preferred analytical techniques

Due to the intrinsic complexity of analysing prebiotic mixtures, a selective process on the characterization of the material became almost a necessity. The products generated usually need to be separated into fractions based on their properties prior to analysis. For example, volatile and non-volatile material must be addressed on different instrumentation and/or with the aid of derivatisation reagents: GC-MS for volatiles and semi-volatiles or LC-MS for the non-volatile material. Furthermore, there is the insoluble fraction (mainly composed of polymerized material) whose complete characterization remains an analytical challenge. Alternatives to bypass this problem include analysis by NMR in solid-state or after dissolution of the dried-*concentrated*- material with a deuterated solvent.[167,168] However, this approach is countered by difficulties on the concentration range of the organic fraction, as NMR is roughly one-thousand times less sensitive than GC or LC-MS.[169] Also, the derivatisation reactions required for a reproducible quantification of the material through spectroscopic techniques, conventionally targets known biological building blocks such as amino acids or nucleobases, giving a secondary importance to the remainder of the chemical space.

Continuous interest in finding biological building blocks within prebiotic broths have concluded, in all its possible scenarios, that they are minor constituent of the resulting chemical space. Indeed, known biological material can be synthesized but the yields are very low and the mechanism of concentration for them unknown. This highlights that the greater part of the product distribution remains a mystery. However, modern technologies have helped introduce an ever-growing array of analytical strategies that further our understanding and characterization of prebiotic mixtures,[170] see **Figure 22**. In order to achieve an experimental framework that tackles all levels in the transition from inorganic matter to polymerized material, an increasing development in separation (e.g. from paper chromatography to gas or high-performance liquid chromatography), detection by Fourier-transformed high resolution mass-spectrometers (e.g. FT-ICR-MS and Orbitrap-MS) and the unequivocal characterization of compounds by specialized NMR experiments (e.g. bi-dimensional and non-uniform sampling), has unlocked the possibility for such studies.

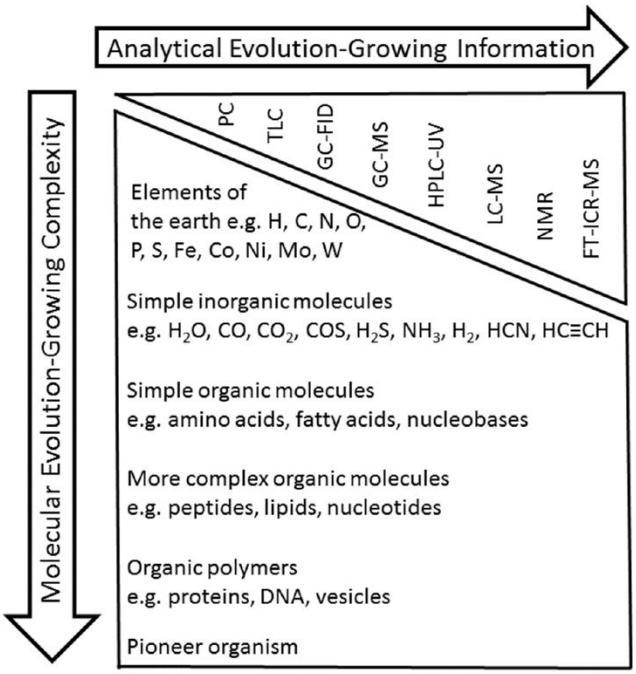


Figure 22. Evolution of molecular complexity in a bottoms-up "Origins of Life" scenario goes side-by-side with the ever-growing progress in analytical techniques and the amount of information retrieved from them. Reproduced without permission from Life, 2019.[170]

1.6.2 GC-MS

The chemical analysis of prebiotic broths will have almost as a prerequisite a separation step, due to the large variety of similar compounds synthetized in the one-pot reaction of small building blocks. A good approach for the simultaneous separation and characterization of the resulting products is Gas Chromatography coupled to a Mass Spectrometer (GC-MS), a general scheme of the instrument is presented in **Figure 23**. The technique was developed for commercial use in the late 60's and it is perfectly suitable to analyse the volatile fraction of a chemical mixture. In order to use this technique with semi-volatile to non-volatile material, the chemical compounds must undergo a derivatisation reaction. This approach would enable an overview of the chemical variety within prebiotic complex mixtures in an unprecedented manner.[171,172,173] In fact, it became a widely used analytical technique for analysing compounds in highly complex matrices, such as environmental samples and blood-serum analysis.[174,175]

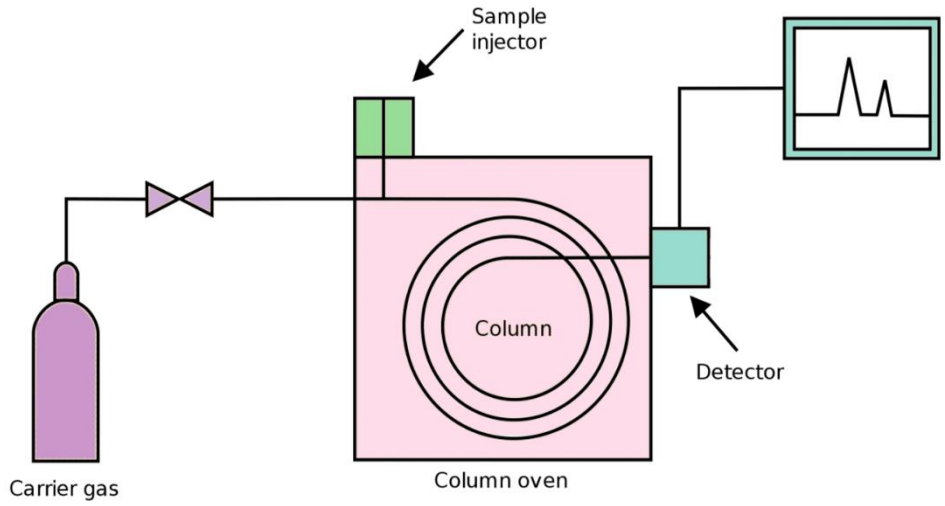


Figure 23. A schematic representation of a GC-MS instrument. A Helium mobile phase (*purple*) complemented to a temperature controlled column oven (*pink*). Samples are injected by a microliter syringe (*green*) and the detection system (*blue*) generates a chromatogram, based on either FID or MS. Adapted from Wikipedia Commons, Creative Commons (CC) 1.0 Public Domain.

In a series of papers conducted by the Wollrab lab, GC-MS and multidimentional GC analysis was carried out in a Miller-type prebiotic broth.[176] This approach was able to give an insight into the chemical variety in the resulting product distribution in electric-discharge experiments. Their results argue for a wide range of saturation in the identified products, which can be either hydrophobic, hydrophilic or amphiphilic in nature. However, their results were inconclusive on the relative abundance of the different types of material, finding no real preference for any reaction product.

In 2017, an important discovery on the non-amino acid product space of the Miller-Urey reaction was made by Ferus *et al.*, by carrying out the Miller-Urey experiment under the premise of a neutral atmosphere (e.g. with carbon monoxide instead of methane, and nitrogen instead of ammonia), being now widely accepted as the atmospheric composition of early earth.[43] The analysis of the samples was carried out with GC-MS, where the analytes are derivatized with MTBSTFA (in a silylation reaction) prior to the analysis.[177] This enabled the detection of several amino acids and other small organic compounds, finding Glycine to be present in all of the experiments. Most notably, the detection of nucleobases within the reaction mixture. Up to this moment, it was not known whether one of the main building blocks for the monomers of DNA and RNA was possible in this reaction. The resulting chromatograms were compared to a blank and a glycine standard, for an ultimate validation. As well as, the detection of the three of the canonical nucleobases: Adenine, Guanine and Cytosine (see **Figure 24**). It must be noted that this was achieved by Single Ion

Monitoring (SIM) of the masses corresponding to the sylilated products of the nucleobases. In other words, they were 'looking' for them in a targeted fashion, which might explain why they had not been previously found in other Miller Urey experiments analysed by GC-MS.

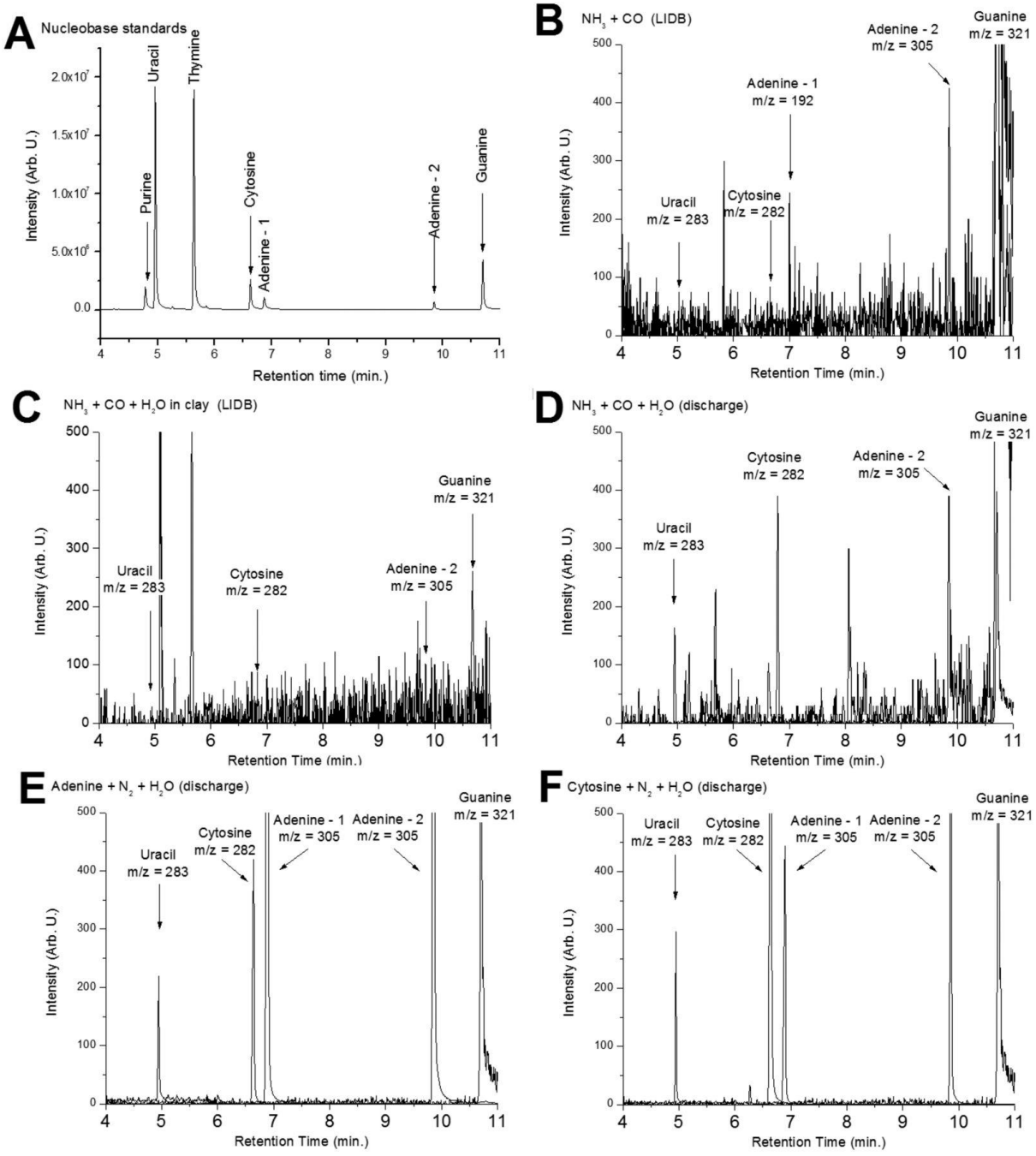


Figure 24. GC-MS detection of nucleic bases in electric discharge experiment with different gas mixtures (**B-D**). The chromatograms where compared with a mixture of the pure standards, by means of external (**A**) and internal (**E,F**) validation. Masses where identified in their trimethyl-silyl derivative form. Reproduced without permission from PNAS, 2017.[43]

Moreover, in recent work presented by Mompeán *et al.*, the GC-MS analysis of a series of Miller-Urey type experiments (e.g. electric discharge) under two different atmospheres: reducing (e.g. methane/ammonia/hydrogen) and neutral (e.g. methane/nitrogen/hydrogen), was conducted.[178] Also, they investigated the effect of alkaline aqueous aerosols in both experiments. The analysis was tailored to include the insoluble material arising from such mixtures, conventionally known as *tholins*. Their results confirm the formation of key elements of the rTCA cycle (e.g. pyruvate, glyoxylate, etc.), amino acids, carboxylic acids and N-heterocycles, see **Figure 25**. Ultimately, adding to the narrative that there are many more biologically-relevant compounds enclosed in the Miller-Urey prebiotic broth, yet to be discovered.

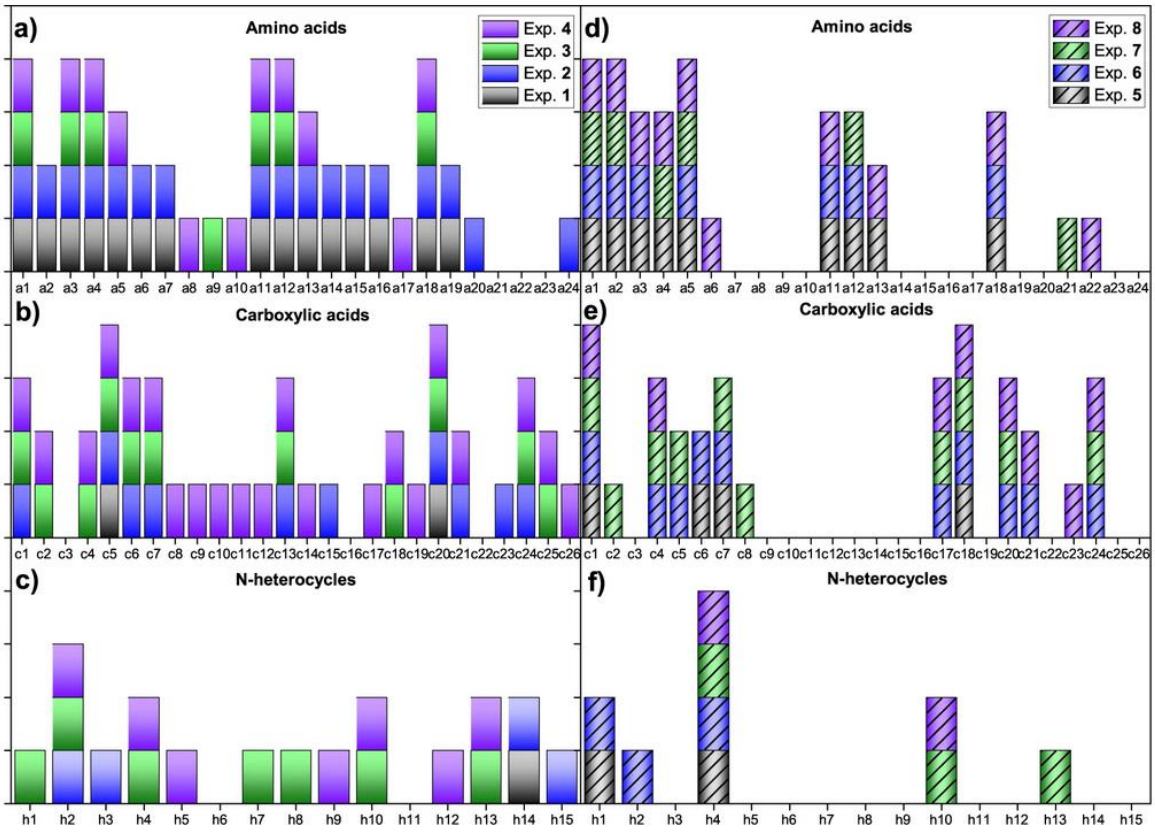


Figure 25. GC-MS analysis of tholins generated in electric discharge experiments under two different atmospheric compositions: Reduced atmosphere (**a,b,c**) and Neutral atmosphere (**c,d,f**). Experiments carried out with water (i.e. classic-setup) are numbered **3-4** and **7-8**; where experiments that were executed in the presence of alkaline aqueous aerosols are numbered **1-2** and **5-6**. Differences across the production of amino-acids, carboxylic acids and N-heterocycles are presented. Reproduced without permission from Scientific Reports, Science, 2019.[178]

1.6.3 LC-UV and HR-MS

Another technique known for its capacity to elucidate the identity of small-organic compounds in highly complex matrices through a separation step coupled to a powerful detection system is Liquid Chromatography coupled to a Mass Spectrometer (LC-MS). This analytical approach is also conventionally coupled to Ultra Violet or Fluorescence spectroscopy, as a mean to quantify the concentration of analytes in solution, by means of the Beer-Lambert equation.

A recent study presents a liquid chromatography–ultraviolet-mass spectrometry (LC–UV-MS) method for the accurate analysis of formose reaction products, a known combinatorial explosion that entails hundreds of compounds in solution.[179] The limitations encountered by gas chromatography, such as the thermal stability of analytes, were avoided through this technique. The targeted analytes in the Formose reaction are sugars (as mentioned in **Section 1.4.2**), which do not have a strong chromophore and needed to be derivatized to their O-benzoylated form prior the analysis. This allowed the use of UV as means to quantify their relative concentration after being subjected to reversed-phase chromatography, see **Figure 26**. This was followed up by electrospray ionization-mass spectrometry (ESI-MS) for validation purposes. After comparison with pure standards, the accuracy of results from LC–UV is dramatically improved due to enhanced chromatographic resolution and validation by ESI-MS. [180]

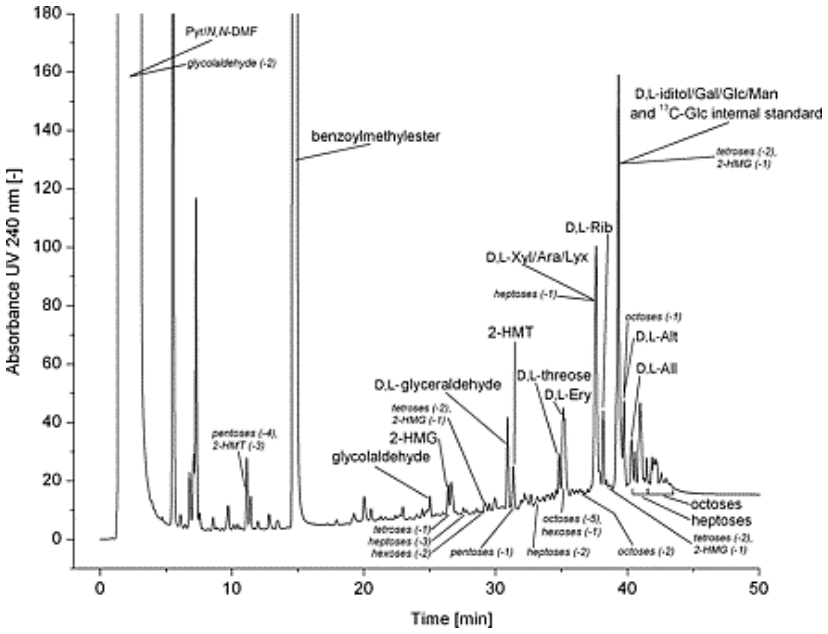


Figure 26. LC–UV chromatogram obtained from a formose sample (reaction initiated with glycoaldehyde). All carbohydrates were analysed as their *O*-benzoylated alditols. Reproduced without permission from the Journal of Chromatographic Science, 2014.[179]

The products synthesized in electric discharge experiments can be quite complex, and there are numerous analytical approaches that can be used to study them. Some of the more commonly used techniques in the literature for analysing amino acids in complex matrices are based on chromatographic and mass spectrometric methods,[181,182] which are highly informative techniques for analysing the complex chemical mixtures produced by Miller-Urey type spark discharge experiments. The quantitative amino acid analyses can be conducted with the aid of a derivatisation reaction, *O*-phthaldialdehyde/N-acetyl-L-cysteine (OPA/NAC)[183], a chiral reagent pair that tags primary amino groups, yielding fluorescent derivatives that can be separated on an achiral stationary phase. Therefore, allowing for the simultaneous spectroscopic (FLD) detection and quantification of the compounds, although they lack a strong chromophore and their highly convoluted matrix, see **Figure 27** below. This method has been adopted as the standard approach for the characterization of amino-acids generated in a Miller-Urey experiment.

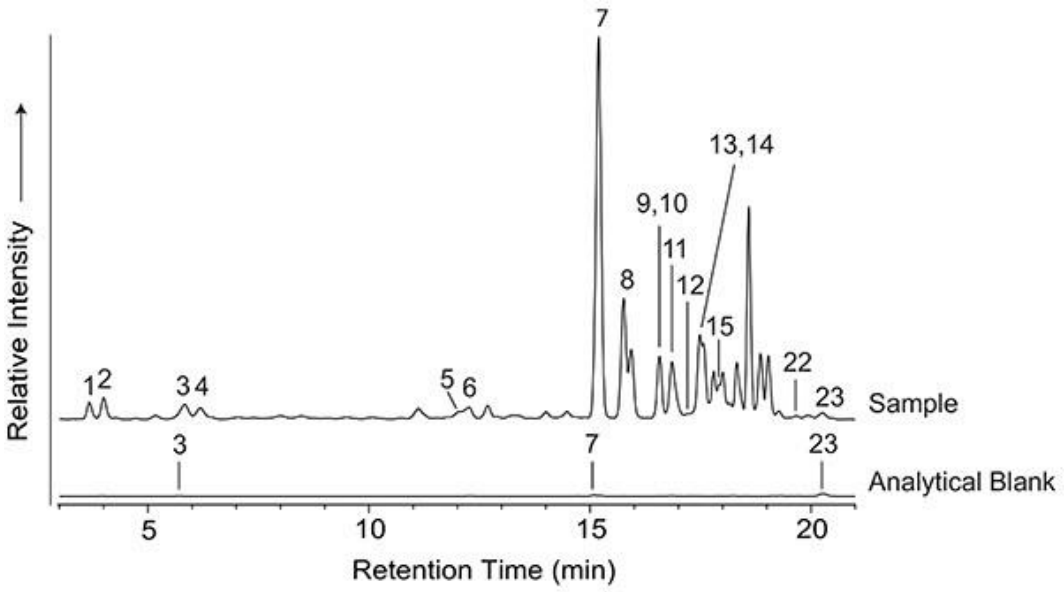


Figure 27. A chromatogram of an OPA/NAC-derivatized amino acid standard in a real sample compared to an analytical blank. Both are obtained by HPLC coupled to fluorescence detection and mass spectrometry. The amino acids contained in the standard include those typically produced in Miller-Urey type spark discharge experiments. Reproduced without permission from the Journal of Visualized Experiments, 2014.[183]

In 2011, samples from a 'volcanic' electric-discharge experiment (e.g. in the presence of hydrogen sulfide) carried out in 1958, were found in the back of a fridge in Miller's old lab.[119,123] The samples were analysed through a modern analytical method, High-Performance Liquid Chromatography coupled to fluorescence and mass-spectrometry detection (HPLC-FLD-MS), finding a diverse array of primary amine compounds. The

identified fraction mainly consisted of amino-acids, which were independently verified using internal and external standards of the compounds. For some of the detected species with identical retention time, a small but non-negligible loss in quantitative accuracy was observed when calculating compound abundances using only HPLC-FLD chromatograms. However, the inclusion of mass spectrometry (e.g. ToF) was shown to overcome co-elution interferences, with the assumption that none of the compounds had *both* identical masses and chromatographic retention times. Consequently, the mass spectrometry data was used to provide a more accurate estimate of the concentrations of the target compounds identified.

Also, the characterization of abiotically produced peptides for the Rodriguez *et al.* work presented in **Section 1.4.1**, was carried out with Size-Exclusion Chromatography coupled to tandem Mass-Spectrometry (HPLC-SEC-MS/MS), see **Figure 28**. The analysis of peptides requires tandem mass-spectrometry, particularly in the presence of multiple monomers (e.g. amino acids), as the exact mass of a polymeric unit can match multiple compositions. This prompted the authors to manually verify the resulting fragments from the MS/MS, as a way to validate the composition of the identified peptides, as seen in **Figure 28b**. Also, the application of HPLC-SEC aided in the efficient separation of resulting oligomers by size in a chromatographic profile that elutes the largest material first, finding peptides as long as 11 units. See **Figure 28a** below.

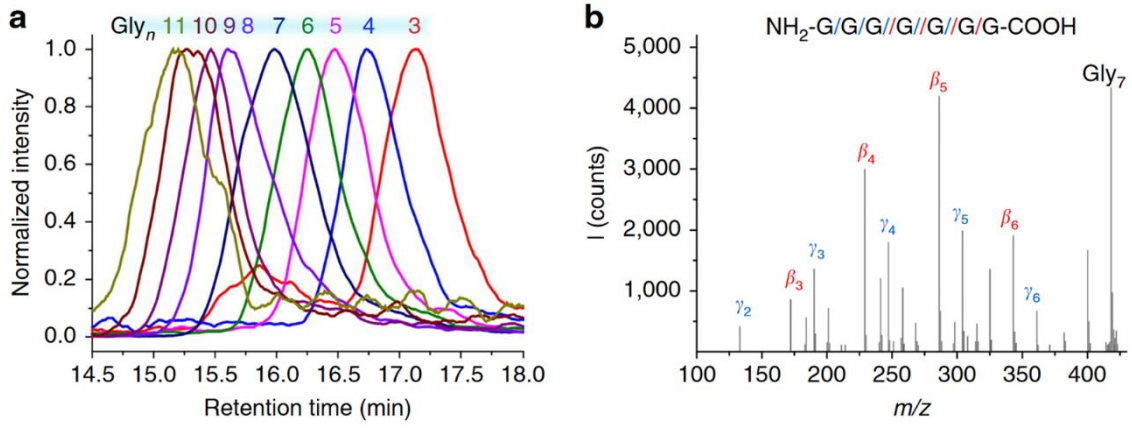


Figure 28. *Analysis by SEC-MS:* (**a**) Extracted ion chromatograms (EICs) for a series of glycine peptides (e.g. Gly_3 to Gly_{11}) and (**b**) Putative assignment of the resulting fragments by MS/MS from the parent ion of a glycine heptamer: 418.17 m/z, [M+H]. Reproduced without permission from Nature Communications, 2015.[47]

Furthermore, Part 1 of the Scherer and Wollrab *et al.* paper series on the Miller-Urey experiment (mentioned previously in **Section 1.4.1**), was partially focussed on analysis by High Resolution Mass-Spectrometry (ToF-MS).[176] This method allowed them to detect 12 different oligomeric species and use the exact mass, assuming an [M+H] adduct, to calculate a plausible chemical formula for each of the polymers found. The observed distribution for all the detected compounds was centralized (e.g. most abundant) around the 300 Da range. Their results also discuss the possibility of an ionization competition that limits the amount detectable compounds through this analytical approach, particularly in the absence of a separation step prior to their detection. Nonetheless, the number of individual compounds detected was in the thousands and covers a wide range of chemical speciation.

Moreover, in a recent publication by Wollrab *et al.*, time-resolved sampling of the Miller-Urey experiment was carried out.[125] The samples were analysed by HPLC-Corona Aerosol Detector (CAD) and HPLC-High Resolution Mass-Spectrometry (Orbitrap), at three different time-points: 8.5, 79.0 and 163.5 hours. These approach allowed for complementarity of two detection systems, giving rise to observations that indicate the production of more and increasingly different substances during the experiment.[125] The HPLC-Orbitrap analysis displayed a large number of organic substances in the range of 100–500 m/z. Also, by applying a blank subtraction 'on the fly', they were able to confidently observe the mass –peak- density distribution of the samples. Finding that over time, an increasing number of compounds appeared in the spectra, but their distribution was shifted towards smaller masses. In the last sample (e.g. 163.5 hours), the distribution of mass densities was found to be comparable to that of the Beilstein database, when limited to C, N,O and H elements), see **Figure 29**.

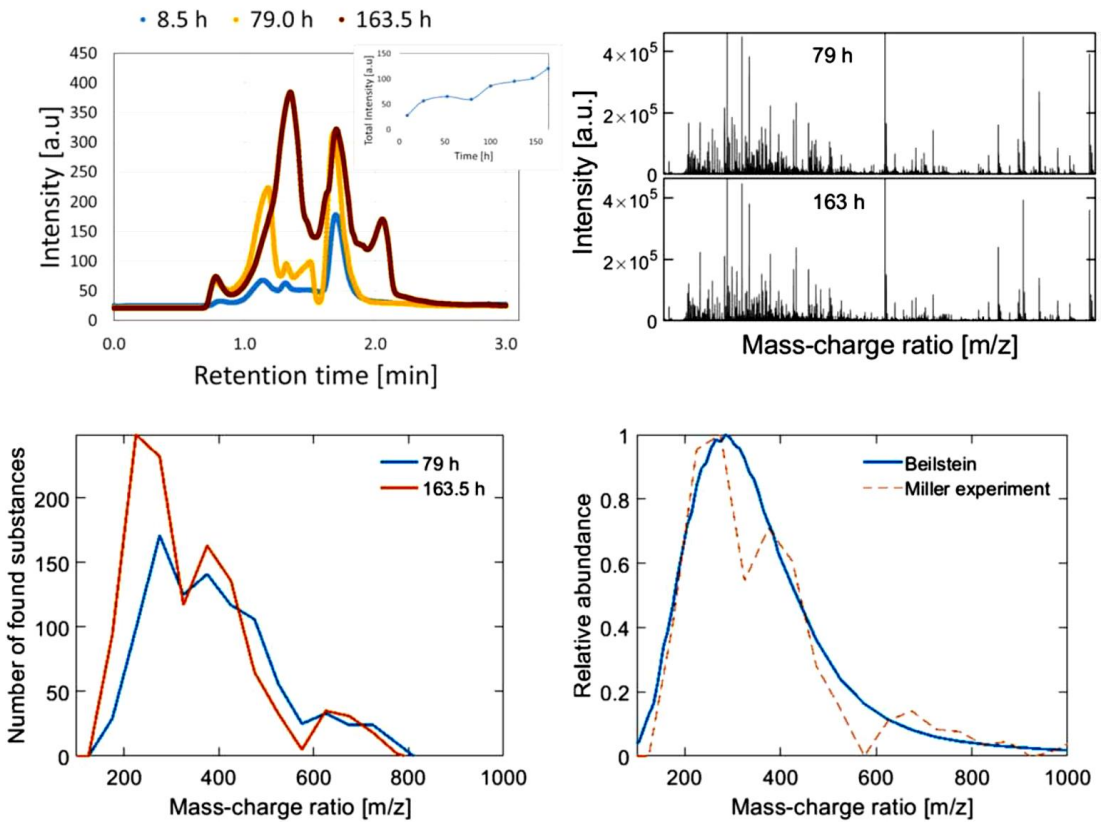


Figure 29. Time-resolved HPLC-CAD analysis (*Top-left*): Three samples were taken at different time points: 8.5 (*blue*), 79.0 (*orange*) and 163.5 hours (*red*). The samples were also analysed offline by direct injection to a HR-MS (i.e. Thermo LTQ Orbitrap) (*Top-right*).and the resulting peak density is presented for two different reaction times, *79* and *163.5* hours (*Bottom-left*) Distribution of the mass densities from the Beilstein database presented in *blue* were limited to C, N,O and H, and compared to the Miller-Urey experiment in *orange* (*Bottom-right*). Reproduced without permission from New Journal of Physics, 2018.[125]

1.6.4 FT-ICR-MS

The highest resolution instrumentation for the detection of small organic compounds in an untargeted fashion from overly-convoluted matrices, is without doubt Fourier Transformed Ion Cyclotron Resonance Mass Spectrometry (FT-ICR-MS).[184] For this reason, it has been used to analyse the most challenging complex mixtures, such as petroleum and soil samples.[185] This justifies why it was selected for the analysis of the composition of small organics in the most analysed meteorite, Murchison.[137,186]

In a publication by Schmitt-Kopplin *et. al*, the FT-ICR-MS analysis on three portions of the Murchinson meteorite was carried out. The analysis was designed to extract as many organic compounds as possible, under mild conditions, with a small amount

of different solvents: apolar aprotic (e.g. toluene and chloroform), polar aprotic (e.g. DMSO and acetonitrile) and polar protic (e.g. ethanol, methanol and water). These fractions where then analysed in an untargeted manner, in (*both*) positive and negative mode ESI. The untargeted approach allowed for the observation of a remarkable molecular complexity, encompassing tens of thousands of different molecular compositions. For an accessible visualization of the chemical diversity encountered, the data was plotted into van Krevelen diagrams, as demonstrated in **Figure 30**. The mass-spectral analysis was focussed on the methanol extract with an acquisition method in negative mode ESI, as it showed to attain a larger number of compounds (e.g. 31,554), than any other solvent and positive mode ESI. Ultimately, their findings indicate a higher level of complexity in the product distribution of the meteorite than those seen for the terrestrial (biological and bio-geological) chemical space.

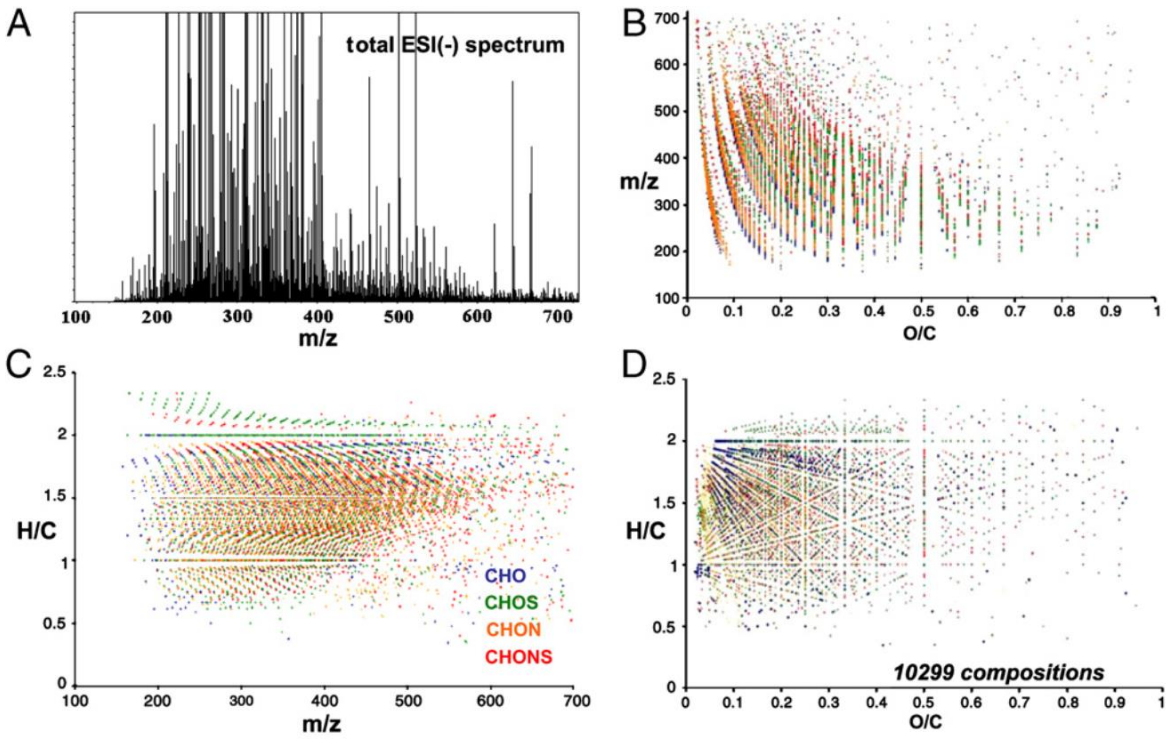


Figure 30. *FT-ICR-MS analysis of the Murchinson meteorite:* The methanol extracted fraction was analysed in negative mode ESI in an m/z range of 150-700, spectrum shown in (**A**). The van Krevelen plots display the elemental ratios of the compounds detected in the mass-spectrum (**B-D**). Reproduced without permission from PNAS, 2014.[134]

This concept was extended even further to examine a system containing the prebiotic condensation of five alpha-hydroxy acids (αHAs) monomers, which generates a possible product space of hundreds to thousands of different oligomer sequences.[168] The analysis

performed for the Mamajanov, *et.al* publication, found that the accurate identification of Isobaric αHA's products or those with multiple condensable functional groups was not possible by this approach. However, high-resolution Fourier transform ion cyclotron resonance MS (FT-ICR-MS) and MS/MS, did conclude that a large proportion (if not all) of the possible sequences were formed, by detecting about 4,300 unique mass peaks between the *m*/*z* range of 210 to1400. This range (>4,000) theoretically includes 10,450 unique linear sequences, ranging from 2–20-mer oligomeric species (see **Figure 31**).

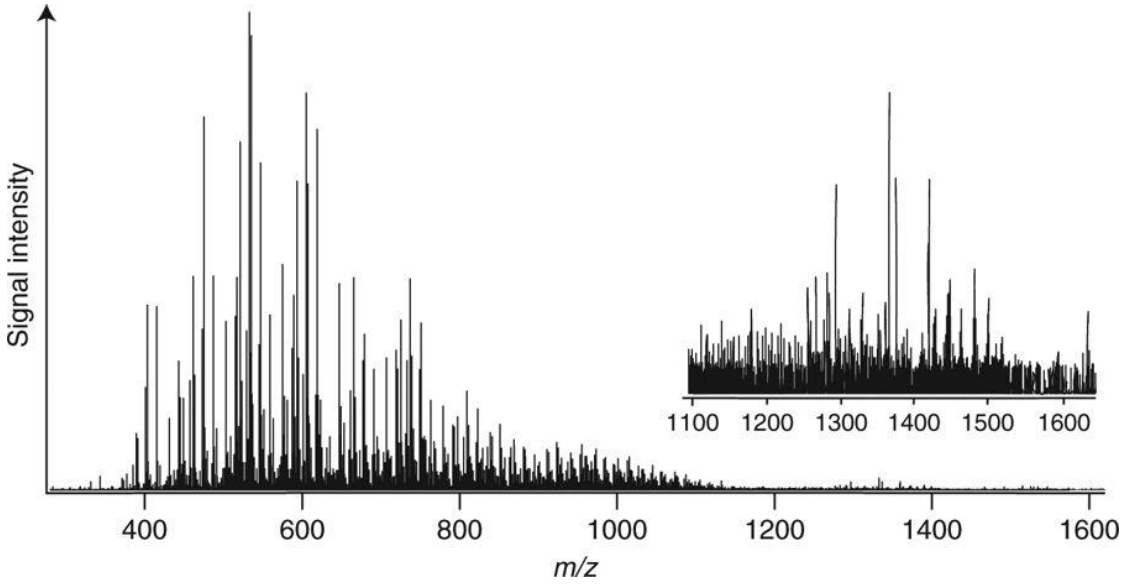


Figure 31. Positive-mode FT-ICR mass spectrum of five αHA mix. The magnification of the m/z 1100–1650 range, was shown as to indicate the density within the detected masses. Reproduced without permission from Nature Communications, 2018.[238]

1.6.5 Towards an untargeted approach

Although significant progress has been made in sample pre-treatment and separation techniques over the last years, there are still considerable limitations when it comes to overcoming complexity and dynamic range problems associated the analysis of complex prebiotic mixtures.[187] Recently, an overview of different techniques and methods currently used for reducing a sample's complexity and concentrating low abundant compounds in biological datasets was presented, detailing advances in database matching and in silico prediction of their molecular composition by tandem MS with comparison of their fragmentation pattern.[188,189,190] These methods are based on two distinct strategies within metabolomics workflows: the targeted or untargeted acquisition of the datasets. The targeted approach allows for a reliable quantification of compounds in complex matrices, but it limits the scope of the study to a particular biological pathway and a specific question (i.e.

hypothesis). An untargeted approach aims to detect and identify as many compounds as possible, broadening the scope of the analysis and achieving a global profile of the samples, thus enabling the discovery of compounds relevant to multiple pathways in biological networks. Ideally, both analytical strategies can be complementary, unleashing the possibility of global profiling across samples and a robust comparison of the relevant compounds. This process is currently underway, resulting in an array of ever growing techniques that aim to overcome the difficulties of compound identification and product distribution resolution in overly-convoluted reaction mixtures. Thus opening the path for new analytical approaches in prebiotic chemistry, particularly in a bottom-up approach.[191,192]

Aims

The transition of simple chemistry into life's building blocks must have required some kind of energy exchange with the environment, since the level of order (and complexity) needed to make the building blocks is more than we get from simple one-pot batch reactions of these molecules. This assumption leads us to believe that there must be a mechanism that allows for the "messy" reactions to converge into a higher order, ready to interact with the other life's building blocks to assemble life. However, the details of this are still unknown to us, providing an interesting open question.

The main objective of this thesis was to analyse and explore complex chemical systems that are relevant to the Origins of Life. First, a series of Miller-Urey experiments were conducted alongside with an all-deuterated version, in order to explore the effect a 'heavier' isotope in the resulting chemical space of the complex mixture. The GC-MS and HPLC-FLD analysis of the samples was carried out, looking for differences between the product distributions in deuterated and protonated Miller-Urey experimental replicates.

While batch reactions are commonplace, the possibility that 'seeding' the reaction with the outcome of a previous one could steer the reaction into a higher degree of order, has been explored far less. It has been demonstrated that the interaction of simple molecules with the environment will create a different outcome or product populations, depending on the environmental input (e.g. minerals, dehydration/hydration cycles, UV radiation, etc.); which can then be used to steer the original reaction into a higher order (or more complex products). The aim of the second part of this work was to explore the effect of reaction cycling and the presence of different mineral environments, in a model system of a complex prebiotic mixture, the formose reaction in formamide.

The third part of this thesis brings these two concepts together and we will discuss how recursive cycles can alter the outcomes and product distribution of a Miller-Urey experiment when run over several weeks.

2. Results and Discussion

2.1 The Miller Urey experiment in a 'Deuterium' world

One of the biggest challenges we face when studying the Origins of Life (OoL) is that in the absence of a functional time-machine, it is not possible to make direct observations about what actually happened on early earth. As a way to circumvent this, efforts have shifted into understanding the potential sources of abiotic compounds that might be relevant to extant life.[193] We model this after our own knowledge of the building blocks that are common across all life forms. Particularly, into the non-biological synthesis of the monomeric units behind the biopolymers we deem responsible for all metabolic processes in living systems. The collection of 'necessary' building blocks for life's polymers can be reduced (roughly) to the constituents behind proteins (e.g. peptides/amino-acids), our genetic code (e.g. nucleic acids, such as DNA and RNA) and a membrane component that encapsulates it all (e.g. glycerol phosphate phospholipids).

In light of this, the Miller-Urey experiment in the 1950s became one the most important experiments in the Origins of Life field.[194] The laboratory simulation carried out by Miller showed that under a reducing atmosphere, with water and an energy source (in this case, spark-discharge to simulate lighting), amino acids could be formed. Since then, several other studies have explored similar systems with the aid of more advanced analytical techniques, in aims to tackle the complexity of the resulting product mixture.[170]

However, all the following studies have inevitably encountered the same problems: analytical intractability. The number of individual compounds generated in this type of 'bottom up' experiment is quite huge and diverse, believed to be indicative of a combinatorial explosion driven by simple uncontrolled reactions.[170] Therefore, the great majority of prebiotic experiments are tailored to look for 'biologically relevant' molecules. Which tackles only a subset of the resulting chemical space, leaving a majority of the abiotically generated material uncharacterized. To address this, a major change in the analytical approach must be considered. Recently, a more 'systems' approach has been taken on, looking for new phenomena, such as unforeseen patterns and new mechanisms of self-organization.[162] Instead of searching for particular products, recent investigations of prebiotic complex chemical networks, have focussed in finding which environmental conditions are capable of 'tuning' the product distribution towards a greater degree of

complexity (for example, the formation of polymers from the abiotically generated monomers).[47,85,151] Following this train of thought, a series of classic (protonated) Miller-Urey experiments were conducted alongside with all-deuterated Miller-Urey experiments to explore the effect a 'heavier' isotope in the resulting chemical space of the complex mixture, with hope that the differences between the product distributions in deuterated and protonated (control) samples may give an insight into the chemical pathways of the system.

Also on a lighter (historical) note, a scientific coincidence regarding the location of these experiments is that the isotopic substitution we chose (deuterium) was discovered by Urey (the same one from the Miller-Urey experiment) in 1931,[195] becoming one of his most important contributions outside of the OoL field. A finding that would have not been possible without the discovery of isotopes by Frederick Soddy in 1912,[196] while working at the University of Glasgow, in the same lab in which the experiments were carried out.

2.1.1 Experimental setup

The experiment was carried out with two spark discharge apparatus, placed right next to each other. Three (duplicate) experiments were executed for two atmospheres: Deuterated or 'Classic' (see **Figure 32**). All were controlled by LabView,[197] which allowed us to initiate the experiment safely and monitor remotely. Initially, all glassware was cleaned and dried, to avoid any contamination, followed by the addition of 400 mL of deuterium oxide or deionized water, before sealing the system. Both rigs where degassed (3 times each) to make sure there was no air left after the water was added. Then they were pressurized to one atmosphere (1 atm), with a gas mixture of: 40% methane, 40% ammonia and 20% hydrogen, or their deuterated counterparts. Finally, the system was heated by a heating mantle to the boiling point of water, before turning on the 24 kV (DC) spark discharge in a 10 sec alternating ON / OFF duty-cycle. The spark discharge is generated by two tungsten electrodes, which degrades over time, as an effect of the harsh experimental conditions and the type of energy source (e.g. AC vs DC voltage) (See **Image A2).** Once the experiment has started (i.e. the system is sealed, water is boiling and the spark is turned on), the water vapour mixes with the gases and passes through the spark, before entering the condenser and becoming liquid again. This process is repeated continuously through the duration of the experiment.

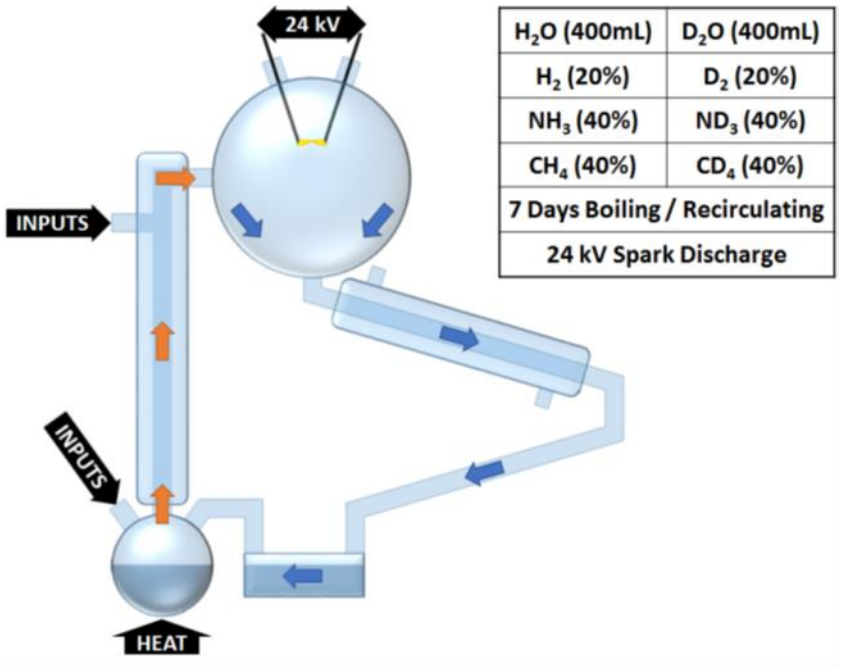


Figure 32. . Scheme of the spark-discharge experimental setup. Arrows demonstrate the flow of water vapour (*orange*), condensate (*blue*), as well as the inputs location (*black*).

All the experiments were run for seven days, resulting in a brown solution with visible insoluble material at the bottom of the flask. Samples were taken after careful cooling of the system and stored in 500 mL Duran® bottles. Three experimental replicates were carried out in duplicate for both deuterated and non-deuterated atmospheres, labelled as: X1-D1, X1-D2, X1-D3, X2- D1, X2-D2, X2-D3 and X1-H1, X1-H2, X1- H3, X2-H1, H2-H2, X2-H3, respectively. Observable differences can be seen in the resulting solutions of the experimental replicates, they vary from light brown/yellow to a dark brown colour, as seen in **Image 2** below.

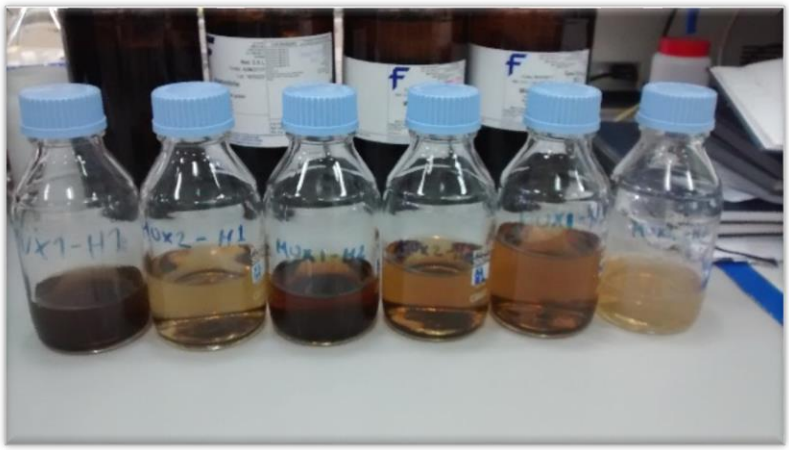


Image 2. Image of the resulting Miller-Urey 'broths': An array of different shades of brown can be observed for the experimental replicates of the 'classic' experiment.

Furthermore, it must be noted that the colour variation within the broths seen in **Image 2** above, results from both the experimental apparatus replicates, as well as the experimental triplicates. The observable differences between the two experimental setups of the same experiments strongly suggests that even if the two spark apparatuses are constructed equally, the subtle inconsistencies they might contain will be translated into non-trivial variations between the systems. Therefore, when comparing Miller-Urey experiments carried out in different laboratories, even through careful consideration of pre-established methodologies, there will be an unavoidable influence of these variations. The level of which the inconsistencies in the experimental setup affect the overall product distribution of the reactions, is not a matter of discussion at the moment, but it should be taken into consideration for future experiments; especially if an effective comparison across experiments carried out in different places and time-lines is to be achieved.

2.1.2 Gas-Chromatography coupled to Mass Spectrometry (GC-MS)

The Miller-Urey experiment is known to produce a complex mixture of small organic compounds, with a wide-array of chemical properties, also, having a broad dynamic range of concentrations in the product distribution. These two aspects, when combined, result in an analytical challenge, since there is no analytical instrumentation with enough resolution and sensitivity to simultaneously detect *and* identify all the generated products. For this reason, many different techniques have been applied. However, most involve a separation step prior to detection, as a requirement for the accurate identification of compounds within the mixture. Initially, the use of paper and thin-layer chromatography was employed. But with the development of more advanced separation techniques and detection systems, gas chromatography (and liquid chromatography) coupled to a mass-spectrometer became easily available for laboratory studies. Following the success of complex mixture analysis by GC-MS for samples with a matrix complexity similar to the Miller-Urey 'soup' (such as environmental, water or soil samples), it quickly became a favoured technique for these types of studies. Therefore, the analysis of the (deuterated and 'classic') Miller-Urey samples was done through Gas Chromatography coupled to Mass Spectrometry (GC-MS), a generally preferred analytical technique for complex mixtures. The sample preparation, derivatisation reaction and GC program (described in **Section 4.1.4**) are a variation of the method described by Molnár-Perl *et al.*,[177] adjusted for the Miller-Urey samples.

The Miller-Urey samples can be analytically challenging for several reasons. First, the complex mixture is made up of both volatile and non-volatile material, which requires a derivatisation reaction. For which we chose a silylation reaction with N-tert-Butyldimethylsilyl-N-methyltrifluoroacetamide (MTBSTFA), since it is a powerful tool for increasing analyte volatility, thermo-catalytic stability, and chromatographic mobility of polar and unstable organic compounds (details in **Section 4.1.3.1**), see **Figure 33**. However, the derivatisation reagent reacts very easily with water, which is a problem for the water-mediated Miller-Urey samples. Consequently, each sample needed to be completely dried before derivatisation, so they were freeze-dried and taken out of the lyophilizer moments before adding the reactants and initializing the derivatisation reaction, in order to minimize the amount of moisture to which the sample is exposed.

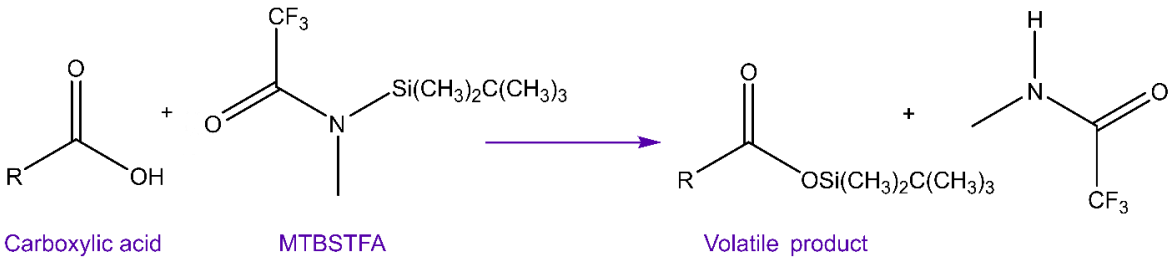


Figure 33. A general reaction scheme of the MTBSTFA derivatisation.

The sample preparation steps prior to the water removal had a non-trivial effect on the amount of material recovered from the samples. Due to the heterogeneous nature of the Miller-Urey mixture, a filtration step prior to lyophilisation was required. For this, we tried a series of different filtration techniques. Then weighted the amount of solid retrieved after lyophilisation using the different filtering techniques (see **Figure A4**) with the same amount of sample volume. As to be expected, the smaller 'cut-off' achieved by syringe filtration (with a 0.22 µm filter) retrieved less material, as more of the insoluble material was retained on the filter itself. Furthermore, filtration prior to analysis (post-derivatisation reaction) is recommended to avoid non-volatiles from accumulating in the GC inlet liner and to conserve chromatographic resolution, as seen in by comparing the resulting TIC's for a filtered versus unfiltered sample (see **Figure A7**) the intensity of some peaks can be affected. Therefore, all samples and analytical standards where filtrated by syringe filter (0.22 µm), after dilution with MS grade acetonitrile (1:10).

The GC-MS analysis was performed using an Agilent 7890 GC coupled to a 5975 MS detector. An Agilent HP-5MS capillary column (95% dimethylpolysiloxane, 5% diphenyl; 30m x 0.25m x 0.25 mm) was selected for the chromatographic separation of the compounds. Also, in order to make sure that all compounds where volatilized, the injector was set at 250 °C. As well, the injector was set to split mode (1:10) as a way to minimize the amount of material introduced into the GC column, without having to dilute the sample any further. The split-mode prevented the over-loading of the column which often results in peak distortion. Several adjustments to the chromatographic method were done to maximize the number of unique peaks detected in the GC chromatogram. The thermal gradient used to elute the compounds from the GC column, was also optimized to give enough time for co-eluting peaks to resolve, resulting in the following program: a starting temperature at 75 °C, then held for 3 minutes, followed by an increase of 3 °C/per minute to 140 °C, then held for another 3 minutes, followed by an increase 3 °C/per minute to 200 °C, then held for a 1 minute, before finally ramping at 5 °C/per minute to 230 °C. Furthermore, the MS method was adjusted so the detector window was opened only after the solvent peak has eluted from the GC column (Retention time: 4.5 min), in order to avoid a temporary saturation of the mass spectrometer and reduction of the filament lifetime. This was done after noticing how the solvent peak significantly reduced the intensity of the peaks eluting after it, in some cases even below the detection limit. The ion polarity for all MS scans was positive, with an EMV mode of 1.0 (or 2448 V) gain factor; acquisition mode in scan, at a normal speed. All of these considerations resulted in the method used to analyse the Miller-Urey samples.

The resulting GC chromatograms are visibly different, but do retain some obvious similarities. The ‘deuterium world’ experiments have more peaks per chromatogram, when compared to the protonated version. In **Figures 35 - 36**, we can see that there are distinct new peaks at later retention times in the deuterated version when compared to the classical. The differences across the peaks was only assessed qualitatively because we were unable to establish the identity of the compound to which the new peaks correspond to, since it did not match any database records or co-eluted with any known standards. See below for an example of the resulting chromatograms for the H experiments (**Figure 34**) and the ‘classic’ version (**Figure 35**).

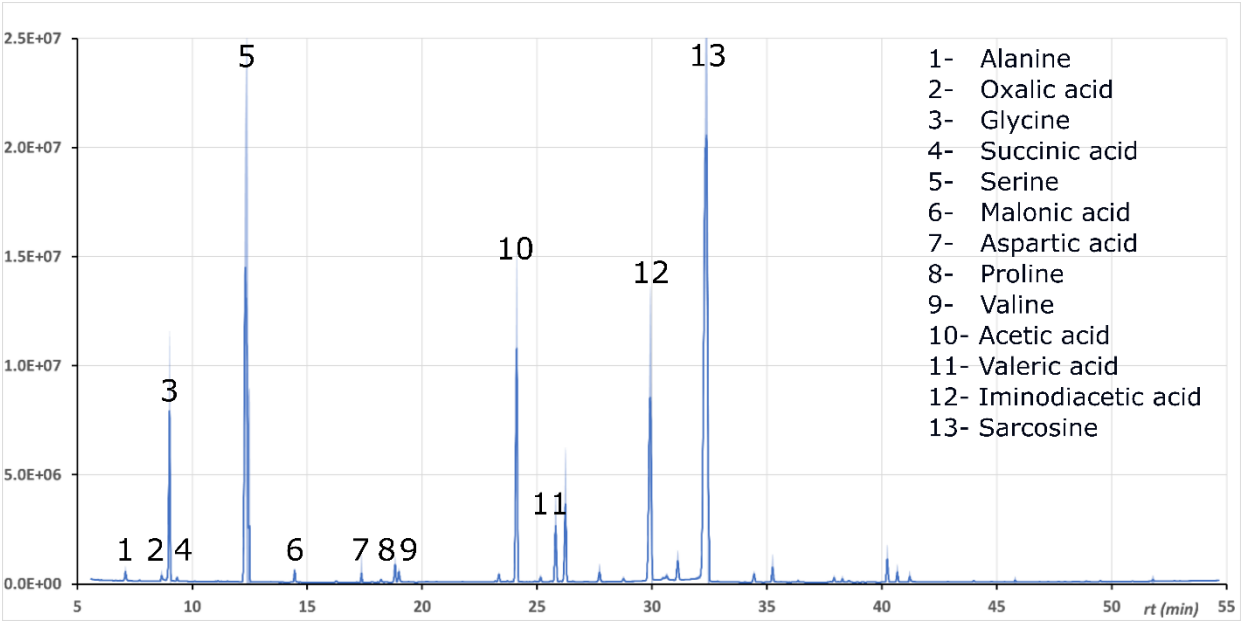


Figure 34. GC-MS Total Ion Chromatogram (TIC) for the 'classic' experiment (MTBSTFA derivatisation). The coloured areas around the traces represent the standard deviation over six experimental replicates and three analytical repeats.

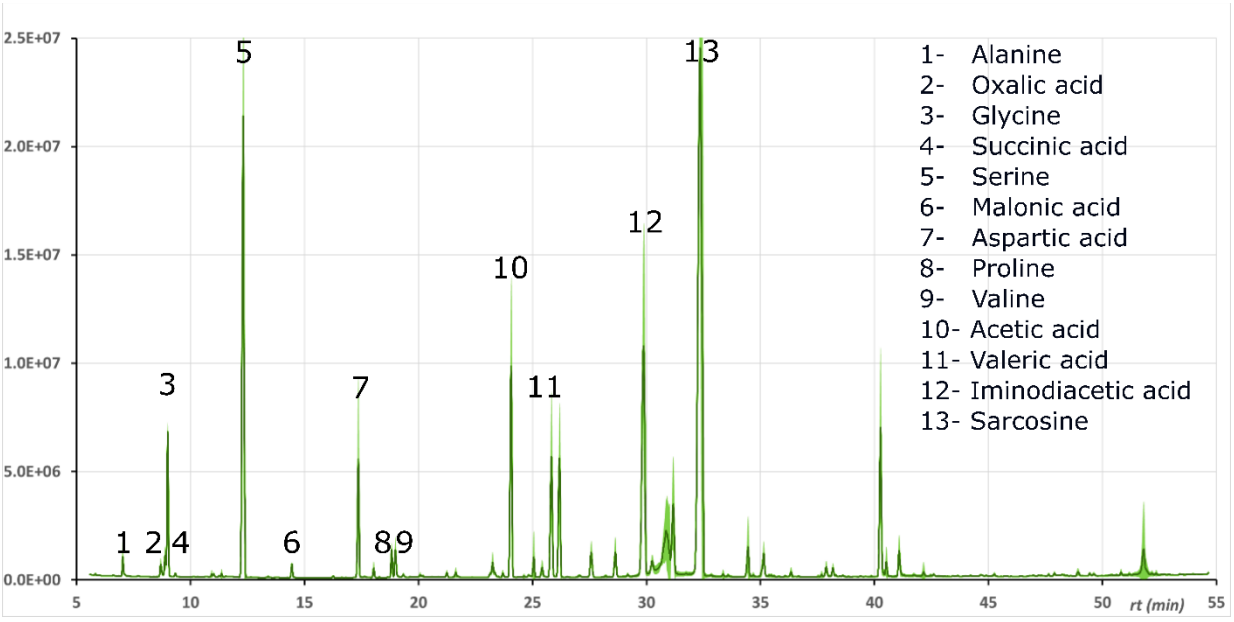


Figure 35. GC-MS Total Ion Chromatogram (TIC) for the deuterated experiments (MTBSTFA derivatisation). The coloured areas around the traces represent the standard deviation over six experimental replicates and three analytical repeats.

In order to identify the compounds, we used the experimental MS spectrum appended for each peak to look for matches in the NIST database. The database search generates a matching score (Quality, %) based on the similarity of the resulting EI fragmentation pattern with the ones available in the database. The databases also take into consideration the use of derivatisation reagents, making compound identification easier. Furthermore, the database matching for the GC-MS benefits from having a more reliable match and mass-spectral pattern, when compared to LC-MS database matching; which needs to take into consideration multiple adducts (e.g. [M+H], [M+Na], etc.) in the exact mass calculation to account for different mobile phases and ionization energies, in order to analyse the experimental MS spectra. However, it does happen that false positives come out of the database search and consequently, real standards are needed to confirm the species. The false positives where seen as matches with a high quality percentage that correspond to compound that by no means could be present in the product ensemble, like for example, any sulfur containing molecule (an element outside the available CHNO chemical space of the 'classic' Miller-Urey experiment). The false positives have been seen before in published results, as in the case the work of Scherer *et al*., where they report species identified by the same data processing method that fell below the threshold of an acceptable quality (%) matching.[176] An acceptable quality in the NIST database matching system is above 70% for most cases, but a manual assessment of the resulting mass-spectral pattern is required for a full-validation in unexpected products. On a further note, the LC-MS database matching is also vulnerable to false positives, but this can be improved by carrying out MS/MS fragmentation and using it as an extra dimension of confidence, by means of pattern matching, alongside with the exact mass.

For a list of the identified compounds through the NIST database matching of GC-MS samples, see **Table 3** below. From the subset of products we were able to identify, we did not observe any notable differences. Identified compounds are present in both experiments (deuterated and protonated) and all experimental replicates, with no significant differences between the intensity of the corresponding peaks in the Total Ion Chromatogram (TIC).

Library ID	Chemical formula	Quality (%)
Hexanoic acid	$C_5H_{11}COOH$	99
L- Valine	$C_5H_{11}NO_2$	98.1
Butanedioic acid (Succinic acid)	$C_4H_6O_4$	98
Tridecane	$C_{13}H_{28}$	96
Urea	CH_4N_2O	96
Hydantoin	$C_3H_4N_2O_2$	93.5
Glycine	NH_2CH_2COOH	93
Aminomalonic acid	$C_3H_4O_4$	92.1
Acetic acid	CH_3COOH	91
Ethanedioic acid (Oxalic acid)	$C_2H_2O_4$	91
L-alanine	$CH_3CH(NH_2)COOH$	90
Oxalate	$C_2O^{2-}{}_4$	90
Silanamine	$C_6H_{19}NNaSi_2$	87
L-Serine	$C_3H_7NO_3$	84.5
Sarcosine	$CH_3CH(NH_2)COOH$	83
Urea	CH_4N_2O	83
Propanoic acid	$C_3H_6O_2$	80
Parabanic acid	$C_3H_2N_2O_3$	77.1
Formamide	CH_3NO	76
N-(2-Acetamido)iminodiacetic acid	$C_6H_{10}N_2O_5$	74.9
L - Proline	$C_5H_9NO_2$	72.8
L- Aspartic acid	$HOOCCH(NH_2)CH_2COOH$	72
2,4(1H,3H)-Pyrimidinedione	$C_4H_6N_2O_2$	68.8
Ethane	C_2H_6	64
2-Butenoic acid (Crotonic acid)	$C_4H_6O_2$	64
Pentanoic acid (Valeric acid)	$C_5H_{10}O_2$	64

Table 3. A list of the Identified products through NIST mass spectral database matching for GC-MS analysis data of the Miller-Urey samples. The compounds highlighted in *green* represent a match quality above 70%, where the *blue* coloured compounds are below the quality threshold, but expected to be present in the mixture (according to previously published results).[178]

The identified products include amino acids, small carboxylic acids, as well as other small nitrogen-containing species such as urea. The formation of succinic and oxalic acid, compounds relevant to the Krebs cycle highlight the ubiquity of the compounds generated in the Miller-Urey prebiotic broth in the metabolic processes of current lifeforms. This goes along side amino acids and urea, for example. It seems that the chemical composition of living systems does in some extent relate to the material 'we get for free' from an uncontrolled abiotic synthesis of small organic compounds. Also, having a product distribution that coincides with that of meteorites, as it was observed in the analysis of the Murchinson meteorite by this technique.[136] This emphasizes that the capacity of forming biologically relevant compounds by abiotic processes is not difficult part, but rather is shifting the product distribution into a more ordered state, where the number of products that changes over time initiate unique patterns, which in turns promotes the complexity of the system without upsetting Schrödinger. A mechanism that would allow for the system to become more complex, would be the initiation of an autocatalytic cycle (or several of them). The results for the GC-MS analysis are not able to suggest the initiation of any autocatalytic process, potentially due to the restricted overview of the product distribution; which arises from the selective effect of having a derivatisation step prior the analysis. This can be seen as the main compromise we took when analysing the samples through this technique, including the complementary and conventionally used sample preparation procedures (e.g. the use of derivatisation reactions). On the other hand, the products we managed to identify, correlate with previous results of Miller-Urey broths. This is indicative of an overall consistency in the composition of the observable product space by GC-MS.

2.1.3 HPLC-FLD

Another preferred analytical technique for the targeted analysis of amino-acids was employed.[198] Previous studies of the Miller-Urey experiments have focused in the detection and quantification of the amino acids generated by the experiment.[199] This is done as to follow up Miller's initial finding of amino acids in the Miller-Urey mixture, while simultaneously enabling an effective comparison across the experiments. The resulting product mixture is known for its highly convoluted state and therefore it is necessary to do a chromatographic separation, in order to identify the amino-acids. Therefore, high performance liquid chromatography (HPLC) was conducted. Also, amino-acid standard solutions are prepared from pure standards and analysed through the same chromatographic method, since they are necessary to identify the peaks in the resulting chromatogram. The

peaks detected in each chromatographic run are then compared to those of the amino acid standards and if the retention time of the eluting peak matches within a considerable time-window (+/- 30s) then we can confidently assign the peak. The retention time window instead of an exact match is used due to the possibility of sample matrix effects altering the elution time for the compounds, especially in highly convoluted mixtures such as the Miller-Urey prebiotic broth.

The HPLC method is conventionally coupled to a spectroscopic technique, which allows for an accurate quantification across samples. The spectroscopic method enabled a direct comparison of their relative abundance by focussing in the resulting intensities of the characterized peaks in a given chromatogram, by means of the Beer-Lambert equation. The concentrations can be calculated if the extinction coefficient of the compounds is known, by integrating the area under the curve for the corresponding chromatographic peak. Furthermore, even in the absence of a calibration curve constructed by external amino-acid standards of known concentrations, a difference in the intensity of the resulting chromatograms can be seen as an indication of the relative concentration of product formed in the samples. We make the assumption that if all samples are prepared equal, then a difference in intensity for any given amino acid (*peak*) would mean that the amino acid is present in higher or lower amounts, when compared to the same amino acid/peak in the other samples.

The spectroscopic technique we chose for the detection of amino acids in the complex mixture was fluorescence. This was along the lines of previously established protocols,[183] where a Fluorescence detector was applied to the analysis of amino acids in the Miller-Urey broth, by coupling it to an HPLC method, see **Figure 36**. The amino-acids do not have a strong chromophore that emits in the fluorescence region and consequently, a derivatisation reaction is required to 'tag' the compounds. The HPLC-FLD analyses were performed using an Agilent 1200 HPLC system, following the standard protocol HPLC method for the analysis of amino acids. The selected derivatisation reaction was o-phthalaldehyde (OPA)/mercaptopropionic acid (MPA). This derivatisation reagent is not able to separate the amino acids by their enantiomer, but it was chosen regardless, due to its stability at room temperature when compared to the alternative (e.g. OPA/NAC). Prior to the derivatisation reactions, the samples where centrifuged and the supernatant transferred to an HPLC vial for further analysis.

For the chromatographic separation of the products, we selected a reverse phase column by Agilent (Poroshell 120 HPH C18, 3.0 x 100 mm, 2.7 µm). The samples were injected in 10 µL aliquots and eluted with a linear gradient mixture of solvents A (water w/0.1% v/v formic acid) and B (100% acetonitrile w/0.1% v/v formic acid) at 1.0 mL per minute, over 21 mins as follows: 0 min – 100% A; 3 min – 100% A; 13 min – 100% B; 15 min – 100% B; 18 min – 100% A. Also, the column over was maintained at 30 °C. The fluorescence detector was set to an excitation wavelength of 340 nm and emission detected at 450 nm. Furthermore, the instrument was controlled and data acquired using Agilent OpenLab software.

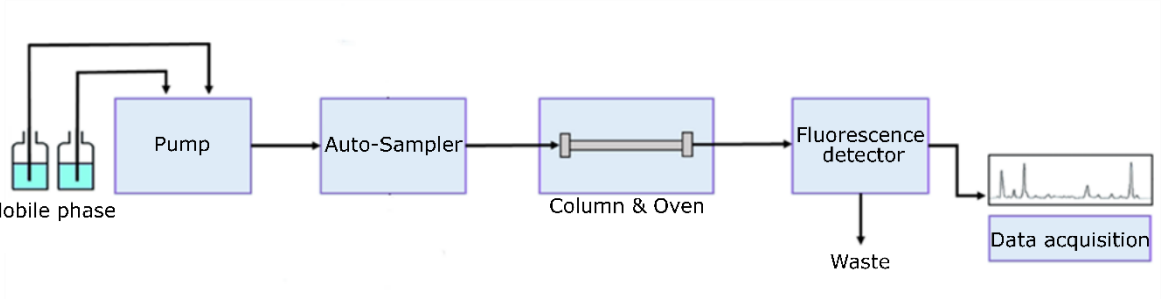


Figure 36. Diagram of the High Performance Liquid Chromatography (HPLC) system coupled to a fluorescence (FLD) detector.

Three analytical replicates were carried out for each experimental sample. Also, a series of amino acid standards of previously identified products in the spark discharge experiments (see **Section 1.6.3**) and analysed as a way to confirm the absence of significant retention time (RT) drift. The same software was used to detect and integrate all significant peaks, and extract corresponding intensities (peak height) in all runs. Only subtle variations (+/- 30 seconds) in the retention time elution window was observed for the peaks. The resulting chromatograms were plotted and over-laid with an intensity offset of 5 units, as a way to efficiently visualize the peaks, see **Figure 37**. All experimental runs have been plotted and their peaks are compared, looking for differences on the intensity which would indicate that certain amino acids where made preferentially. However, no significant differences in the peak intensity of any of the identified amino acids was identified. Therefore, we can conclude that there are no significant differences in the amino-acid product space of the Miller-Urey experiment when carried out with deuterium substitution or a ‘deuterium world’.

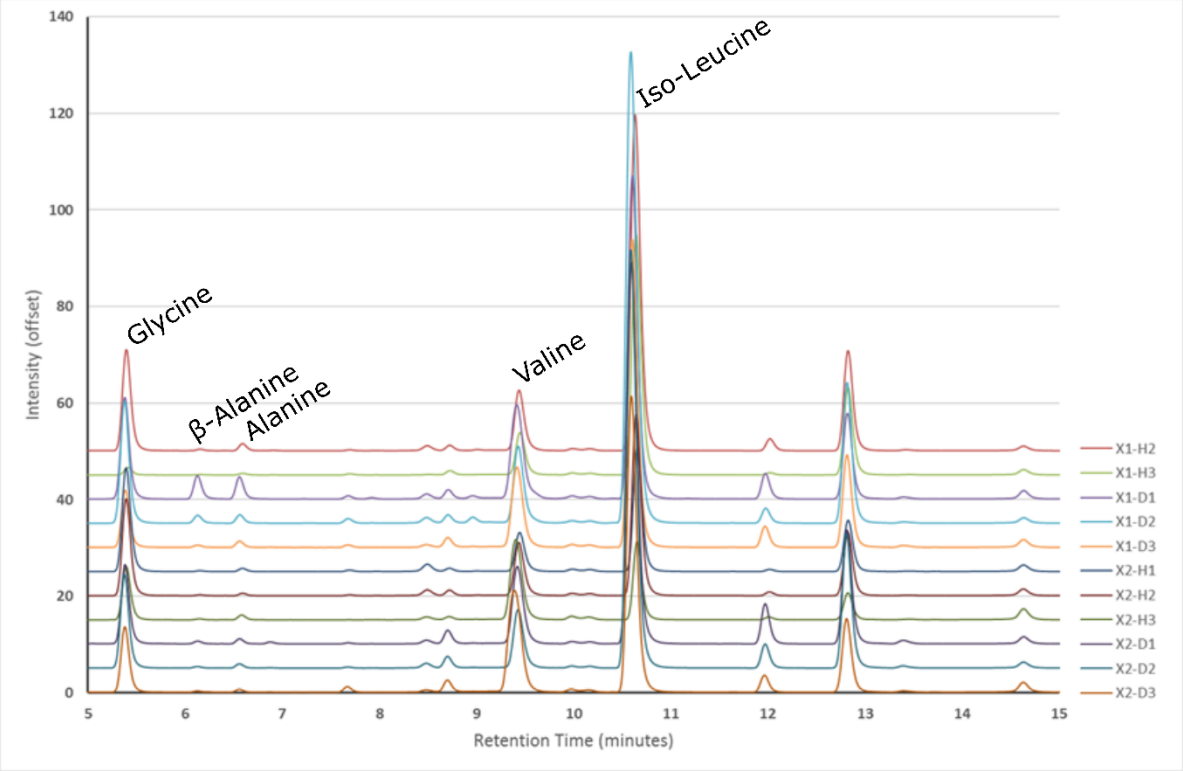


Figure 37. The HPLC-FLD plots of all the 'classic' and deuterated Miller Urey experimental runs. All experimental repeats are presented, for the deuterated (*X1-D1 to X2-D3*) and protonated (*X1-H1 to X2-H3*) version.

2.1.4 Elemental analysis and SEM-EDS

Elemental analysis is a process where a sample of some material (for example: soil, water, bodily fluids, etc.) is analysed for its elemental and sometimes isotopic composition. This allows for a general assessment of the amount of Carbon, Hydrogen and Nitrogen (C/H/N) in the products generated by the deuterated or 'classic' Miller-Urey experiments. Therefore, elemental analysis of the dried Miller-Urey samples was conducted. None of the samples were filtered prior to analysis, as we wanted to include the insoluble fraction of the samples. Also, for the second experiment of the deuterated and protonated version, we did analytical duplicates, as a way to gauge the instrumental variation across the samples. The duplicates for X1-H2 and X1-D2, did not result in large differences for the C/N/H ratios.

However, this approach into the system did not highlight any significant differences, in either the carbon, hydrogen or nitrogen content-ratio of the deuterated and 'classic' Miller-Urey samples (see **Figure 38**).

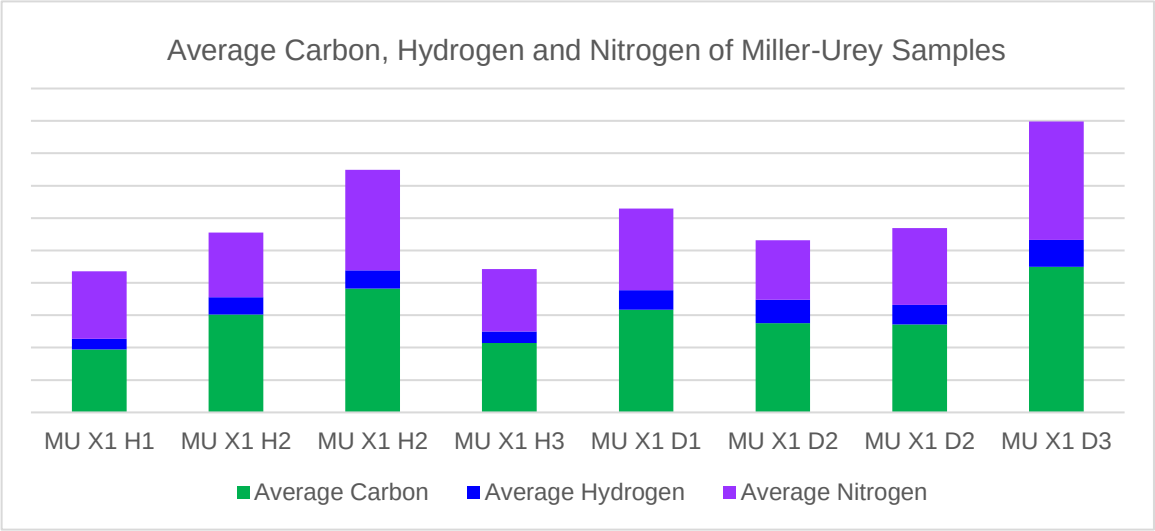


Figure 38. Results of elemental analysis, average Carbon (*green*), Hydrogen (*blue*) and Nitrogen (*purple*) of Miller-Urey samples. Instrumental repeats for X1H2 and H1D2 are also displayed.

Furthermore, as a way to assess differences across the deuterated and 'classic' Miller-Urey samples, scanning electron microscopy (SEM) coupled to energy-dispersive X-ray spectroscopy (EDS) analysis was conducted. The SEM analysis allowed us to look for visual differences in the surface of the dried material, as seen in **Figure 39**. The SEM images of 'classic' protonated Miller-Urey samples show a surface of higher ruggidity than the one we get from the deuterated (dried) material, constituting a subtle but very much observable topological difference. The EDS part of the analysis enabled us to look for qualitative differences in the elemental composition of the freeze-dried samples, by comparing which chemical elements are present and their relative abundance. However, this holds only to some extent in our case, since a limited fraction of the material (e.g. a sub-section of the surface) is taken into consideration.

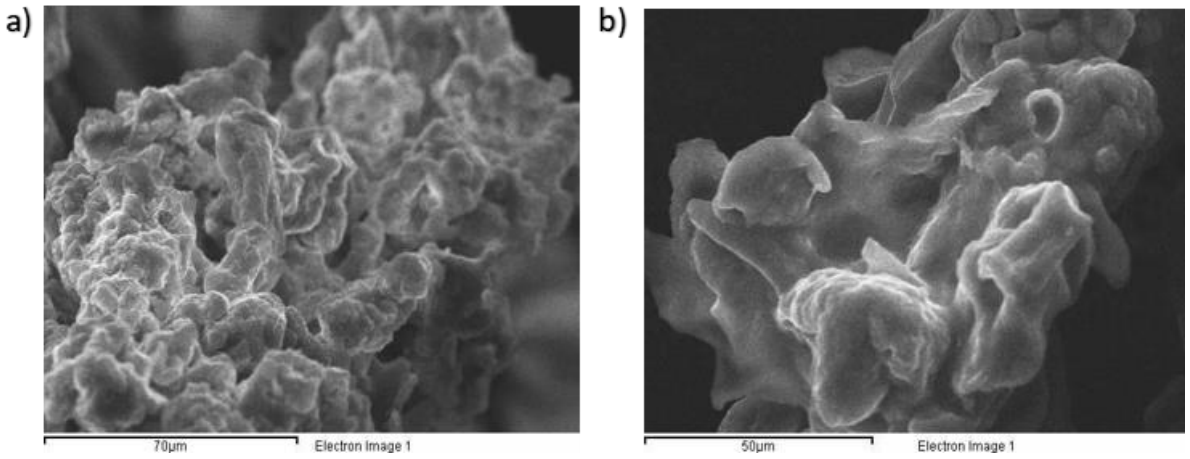


Figure 39. SEM images for a 'classic' Miller-Urey sample (**a**) and a deuterated Miller-Urey sample (**b**).

Nonetheless, we did encounter something unusual in the elemental analysis by SEM-EDS. It indicated the presence of significant amounts of silicon (Si); as well as, low amounts of aluminium (Al) and sodium (Na). These compounds are known to be present in the Pyrex glass-ware used for the experiment, with the same relative distributions (e.g. mostly silica and traces of Na and Al.[200] It must be noted that the presence of silica (or boro-silicate) in the samples was pointed out by Miller himself.[39] This is yet another great example on how advances in analythical techniques, allows to account for the compounds in trace amounts that could have an effect in the system, but have not been captured by previous analysis. (see **Figure 40** and **Figure 41**) Futhermore, that the inclusion of inorganic material into the abiotic synthesis of small organics by the Miller-Urey reaction is inevitable due to the harsh experimental conditions, a variable that deviates the system from a truly 'prebiotic' reactor and therefore, must be taken in to consideration in furter studies. Particularly, when considering the effect of long term experiments carried out under 'prebiotic' considerations.

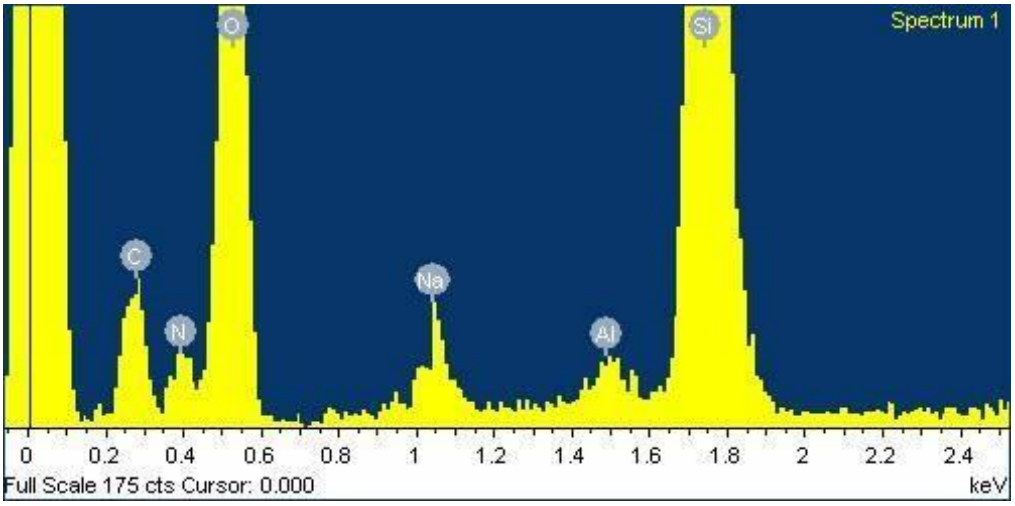


Figure 40. *SEM-EDS:* Elemental composition of a deuterated Miller-Urey sample. A significant amount of silica (Si), as well as traces of sodium (Na) and Aluminum (Al) can be observed.

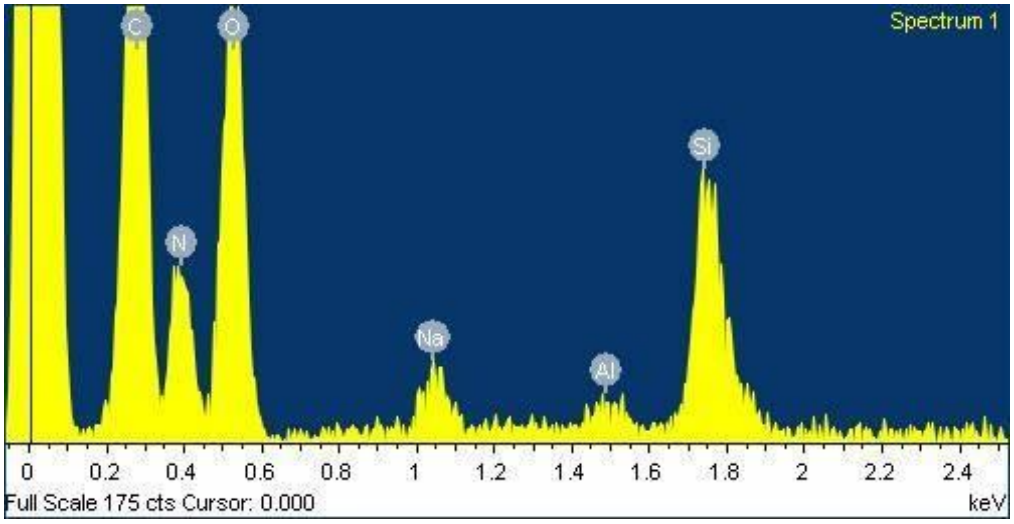


Figure 41. *SEM-EDS:* Elemental composition of a protonated ('classic') Miller-Urey sample. A significant amount of silica (Si), as well as traces of sodium (Na) and Aluminum (Al) can be observed.

2.1.5 Principal Component Analysis (PCA)

In aims to have a more 'systems' approach to the Miller-Urey GC-MS and HPLC-FLD analysis, we considered the application of statistical analysis tools to the resulting datasets. This approach is commonly used in 'systems' science when the complexity of the datasets does not allow to visually extrapolate differences across them. The statistical analysis of datasets might be a common feature across 'systems chemistry' experiments, but it had not been applied before to a 'prebiotic'-type complex mixture such as the Miller-Urey 'soup'. Therefore, we carried out Principal Component Analysis (PCA), when the product space of

the identified products (e.g. peaks) did not exhibit any clear differences between the deuterated and ‘classic’ samples. In aims that a powerful statistical tool could provide an insight into the complex data, which was not resolvable ‘by eye’ or by comparison of the products generated. Considering that this type of statistical analysis can provides a ‘fingerprint’ of the mixture, even without any peak-picking bias.

The principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used in many different applications outside chemistry, as a way to make data easy to explore and visualize. The main idea behind PCA is to reduce the dimensionality of a given data set, while retaining the variation present in multiple correlated variables. This efficiently maximizes the amount of information we retain from the variables, and therefore minimizes the loss from the original data sets. A principal component (PC) can be defined as a linear combination of optimally-weighted observed variables. In **Figure 42**, we can imagine a 2 dimensional feature space which we want to apply PCA to, the original data set is then transformed to the best representation of the variance across the points. This is achieved by making linear combinations of the original variables, which satisfy the requirement of most variance. The variation present in the PC’s decreases as we go down, therefore making the 1st one the most important.

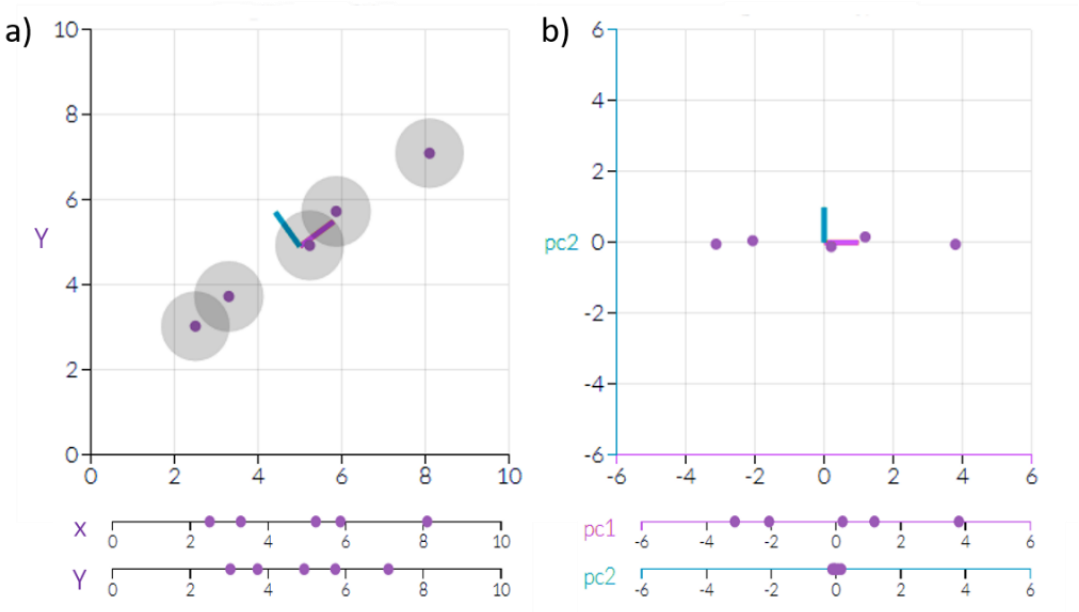


Figure 42. Example of an original dataset (a) being converted to its PC components (b) in the output of PCA.

Fortunately, a simple principal component analysis of the Miller-Urey GC-MS data did indeed reveal a systematic distinction between the product sets of H and D experiments (see **Figure 43**). The product space was clustered by into two visually clear ellipses, where the features calculated for the D experiment are spread over a larger area than in the H experiment. This can indicate that the chemical differences across the generated products within the same experiment are greater in the deuterated experiment, when compared to the classic one. This is not conclusive of any particular mechanism behind the formation of the products, but it does support the hypothesis that there is an amplified kinetic effect in the resulting product distribution of spark –discharge experiments, when an isotopic substitution is investigated (in this case, with the heavier isotope of hydrogen).

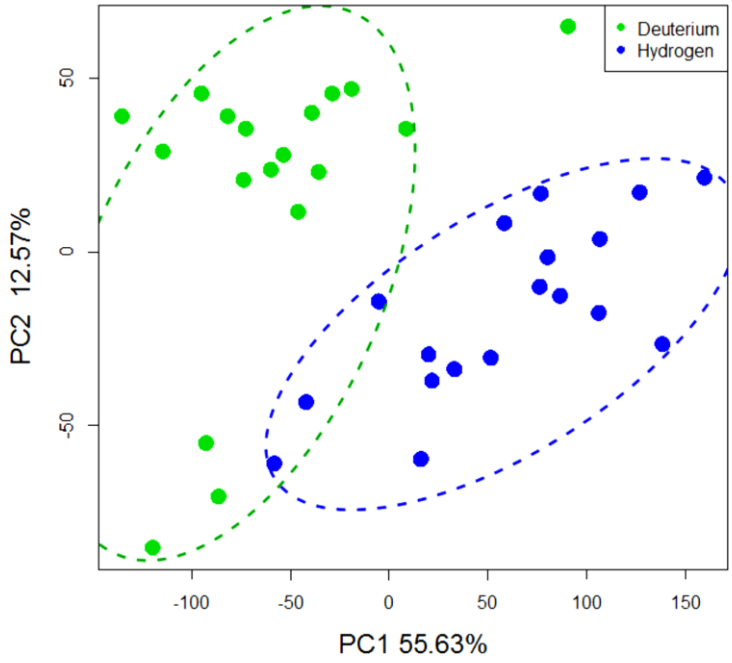


Figure 43. Simple PCA scores plot of GC-MS raw data (peak intensity vs retention time, with no peak picking and MTBSTFA derivatisation) for H (blue) and D (green). Dotted ellipses are drawn as a guide to the eye and are not calculated confidence ellipses.

Furthermore, we also conducted PCA analysis of the features generated by HPLC-FLD. While the analytical features generated for the GC-MS samples were selected without any peak-picking, the HPLC-FLD statistical analysis did require peak-picking of the chromatographic data. The peaks selected from the chromatograms can be observed in **Figure 44,** where they are highlighted by a purple triangle. Two representative samples for the H and D experiment are selected and compared to a sample blank, as a way to show that none of the selected peaks where present in the blank. This allowed us to further confirm that the peaks taken into consideration for the statistical analysis, do effectively correspond

to real amino-acid products.

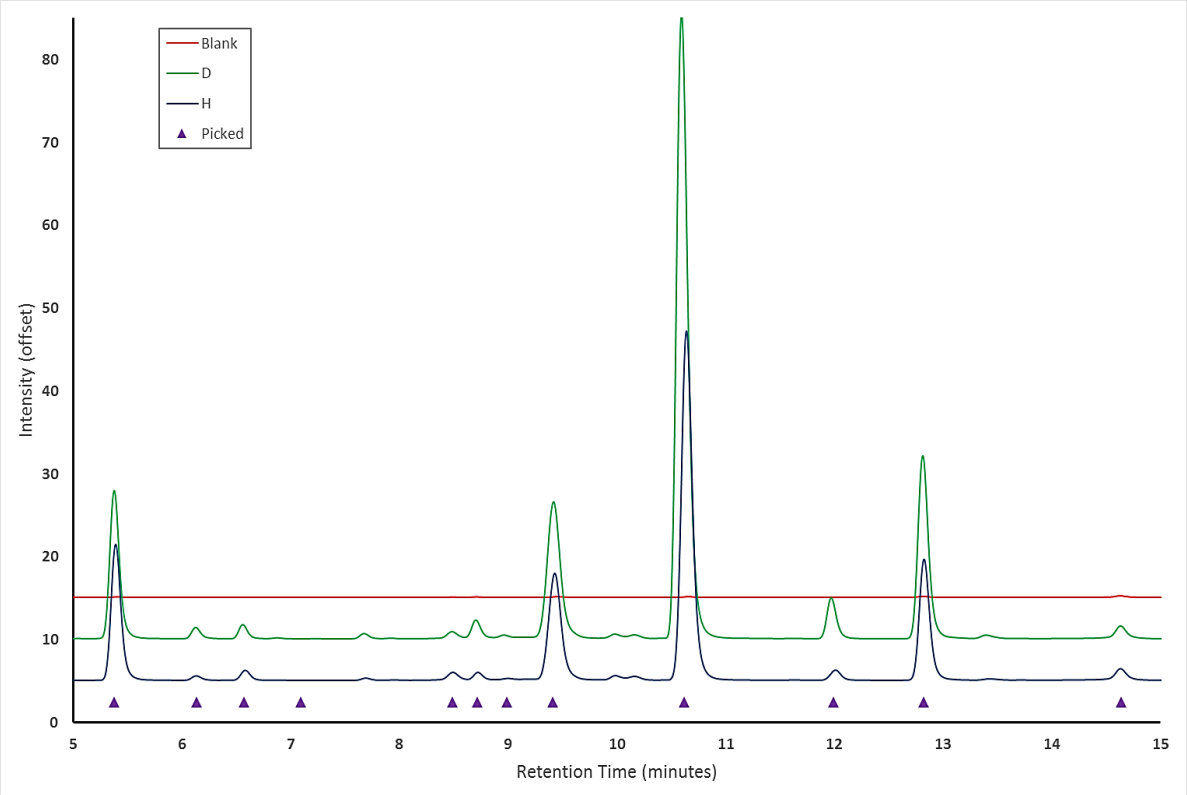


Figure 44. HPLC-FLD plots: Example of a deuterated and protonated Miller-Urey sample compared to a blank run. *Purple* triangles show the peaks that were picked in the data analysis.

The results of the PCA analysis of the HPLC-FLD data sets are comparable to those obtained by the GC-MS unpicked data. The deuterated samples separate from the 'classic' protonated samples, generating two distinct clusters, as seen in **Figure 45**. The variance associated to the PCA scores for the HPLC-FLD appears to be moderately lower than that for the GC-MS analysis. This could be due to the differences in the effectiveness of the derivatisation reaction, as well as it selectivity. In a complex matrix such as this one, which contains a broad range of chemical properties across the generated products, it would be extremely difficult not to encounter some interference of the derivatisation reaction. However, comparatively to the clusters formed by the GC-MS analysis, the deuterated products appear to spread out more than those in the protonated version.

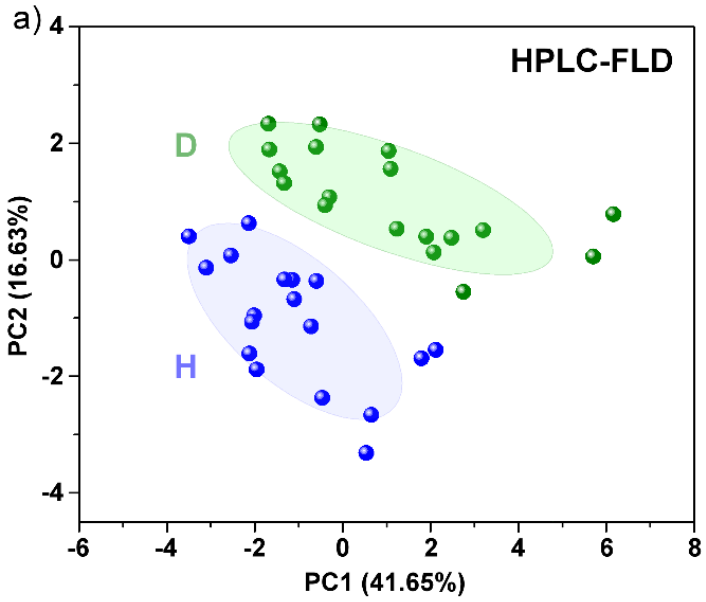


Figure 45. Simple PCA scores plot of HPLC-FLD picked peaks for H (*blue*) and D (*green*).

The statistical analysis of the Miller-Urey data sets acquired by GC-MS and HPLC-FLD produced consistent results across them. This suggest that targeted analysis of the mixtures enclose a product space that is somewhat comparable regardless of techniques used. In addition to the realization that the use of a 'more' systems approach into the data analysis, results in consistent differences across datasets. Especially, since this was not true for the conventional data processing when looking at the data generated by either technique, in a targeted manner. The aforementioned 'systems' approach enabled us to find that there is considerable variance in the data, which has perhaps not been addressed previously. Indeed, we are unaware of any work in the field where data from multiple experimental replicates, or experiments from different apparatus, are compared in this way. Furthermore, when we inspected the variance across the multiple experimental repeats, this was not significant and the resulting trends remained.

2.1.6 Section Summary

We carried out a Miller-Urey experiment in a 'deuterium world' by substituting the hydrogen to deuterium in all the materials used (e.g. gas mixture and water). Three experimental replicates were carried out to verify the reproducibility of the resulting trends. The analysis was conducted through two conventionally applied techniques in the study of Miller-type systems (e.g. prebiotic complex mixtures): GC-MS and HPLC-FLD. Sample preparation and concentration by lyophilisation was developed to remove the insoluble fraction, concentrate the organic material and allow for a water-sensitive derivatisation reaction, prior to GC-MS

analysis. The GC-MS analysis was carried out in triplicates, in order to ensure analytical consistency. The chromatograms were then manually processed and each peak was integrated to extract the corresponding mass spectral data. The appended mass-spectral pattern was searched against the NIST database, which compiles multiple libraries, extending from biological to environmental applications. The NIST search resulted in the identification of a good fraction of the peaks in the chromatograms. Also, we compared the identified products to previous results of the GC-MS analysis of Miller-Urey mixtures, finding them to be consistent with such product lists. However, no significant differences across the deuterated and 'classic' experiment are observed amongst the identified products. Additionally, HPLC-FLD analysis of the samples was carried out, in order to look for differences in the amino acid product space. The fluorescence detector allowed us to extrapolate differences in the intensity of the corresponding peaks (e.g. amino acids), since they can be representative of their relative concentration in solution. Nonetheless, as with the GC-MS results, no significant differences were seen in the amino acids formed when comparing the deuterated to the 'classic' Miller-Urey experiment. This suggests that amino acid formation in the Miller-Urey experiment is not greatly influenced by the substitution of hydrogen to its heavier analogue. The resulting intensity in a peak to peak comparison between the chromatograms, does not display any mayor variations across multiple experimental repeats, reassuring the reproducibility of the trends.

Data-sets resulting of the two (aforementioned) analytical methods are usually processed by comparing the resulting chromatograms against each other and consequently, basing the differences on the number of peaks per chromatogram, as well as their intensity. This did not result in any appreciable differences (by eye) and therefore a more 'systems' approach was carried out. Interestingly so, the statistical analysis by PCA, did display clear differences in the detected compounds for *both* GC-MS and HPLC-FLD. This makes a strong case against the conventional data acquisition and processing of these type of systems, which are usually steered towards amino-acid detection by use of derivatisation reagents to reduce analytical complexity.

The elemental analysis showed no significant differences for nitrogen, hydrogen or carbon percentages. However, through SEM analysis, physical differences in the surface of the dried material of the deuterated and protonated experiments was observed. The deuterated samples appeared to be less rugged than those of the 'classic' Miller-Urey experiment. Furthermore, the SEM was coupled to EDS analysis, which provided information on the elemental composition of the surfaces. This resulted in an interesting finding, since the amount of

inorganic material present was greater than initially thought. All samples analysed, from *both* experiments (H and D), contained silicon (Si), aluminum (Al) and sodium (Na). This indicative of glassware 'erosion' promoted by the experimental conditions used, which we can assume to be considerable if the aluminum and sodium that are only present in trace amounts within the glassware composition can be detected.

The product distribution of the Miller-Urey system changes as an effect of the isotopic substitution of hydrogen containing compounds, with that of its heavier isotope (deuterium), but could only be appreciated when an alternative processing method was used. Therefore, the analysis of prebiotic complex mixtures could benefit from integrating a less targeted approach. This would allow for a more comprehensive overview of the product distribution, regardless of its capacity to identify the resulting products. Indeed, it is the necessity for product identification that had made the aforesaid techniques, a main tool for the chemical analysis of such highly convoluted mixtures. This can be beneficial, but only when overall trends and differences have been identified. The ability to identify compounds though this techniques is also limited to amount of available stardards (HPLC-FLD / GC-MS) and mass-spectral databases (GC-MS), consistently restraining our capacity of assessing systems wide phenomena in complex mixtures.

2.2 The Formose reaction in Formamide: A model system for prebiotic complex mixtures

Several reactions have been considered in the prebiotic synthesis of life's building blocks, as implicated in **Section 1.4**. From the selection of prebiotic reactions that lead to the formation of the organic precursors of nucleic acids, two different pathways have been identified that lead to the abiotic synthesis of sugars and nucleobases: the formose reaction and the formamide condensation, respectively (see **Figure 46**). Yet, when taking into consideration recent advances on the abiotic synthesis of nucleotides by Sutherland *et al.* and others (**Section 1.4.3**), we reflected on the possibility of unifying the main two prebiotic reactions involved in the formation of the DNA/RNA monomers. In order to do this, we decided to carry out the formose reaction in a mixture of water and formamide (50:50 v/v). The resulting product distribution for each of the reactions is known to be a highly convoluted mixture of compounds. For example, the formose reaction results in a combinatorial explosion of different sugars and it would be extremely difficult (or near impossible) to completely resolve and simultaneously identify all the products generated in

the uncontrolled reaction networks. However, we can still look for overall differences in the chromatograms and explore how different environmental variations of the combined formose reaction and formamide condensation can affect the resulting product distribution.

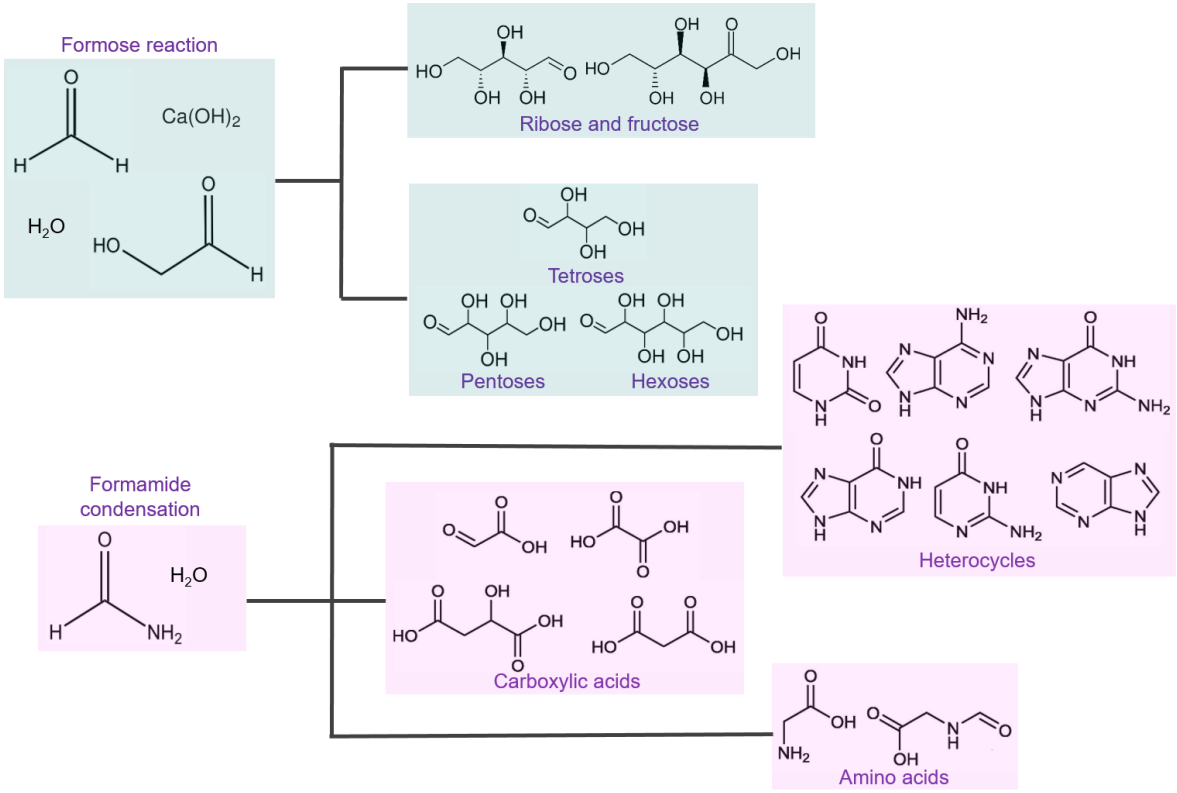


Figure 46. Two main prebiotic reactions in the synthesis of sugars and nucleobases: Reagents and previously identified products[179,239] for the Formose Reaction (*top*) and Formamide condensation (*bottom*).

Previous work in prebiotic chemistry has demonstrated that the inclusion of mineral surfaces in complex reaction networks, can effectively steer the product distribution into a particular product.[201] For example, a study by Ricardo et al, found that in the presence of borate minerals, the formose reaction yielded ribose (the sugar unit of RNA) as a preferred reaction product, over a broad distribution of sugars.[144] Furthermore, in the case of the formamide condensation, the inclusion of mineral surfaces catalyzed the formation of certain nucleobases over others, with a preferential effect seen on the type of mineral used.[202]

Another important variable is the dynamic nature of the environment itself. In a real-world scenario, the abiotically generated broths would not be in a static environment, but in a rather fluctuating one as an effect of day-night cycles. The periodicity of the day-night cycles in early earth are not yet confirmed and ongoing studies tackle this uncertainty. However, we

are certain that there was a natural atmospheric cycling process on early earth, as it is now (but with shorter or longer days).[55] The possibility of natural environmental fluctuations, which take place outside a laboratory setup, having an important role in the chemical evolution of complex mixtures has not been addressed in any prebiotic experiment to date. At least not outside the scope of the abiotic synthesis of peptides, where wet-dry cycles have been employed to achieve sequential water-removal to drive the amino-acid polymerization process.

We investigated how different environmental inputs (such as mineral surfaces) can affect the combined formose and formamide reaction, by steering them to a particular outcome. Also, the possibility of these effect being amplified by natural environmental fluctuations, giving rise to unique product distributions for each environment (mineral type). As way to assess differences in the product distribution in an objective manner, we decided on the untargeted analysis of the samples. Inspired by the metabolomics workflows designed for metabolite discovery, we conducted UPLC-MS/MS in a Data-Dependent fashion. This allows for features to be generated in a confident manner, where each one represents a product within the complex product distribution, partially mapping the resulting chemical space of the products. The features are then compared across different environments and cycle number, in order to gain an insight on how variables present in real-world scenarios could affect the product distribution of prebiotic complex mixtures.

2.2.1 Recursive cycles

In order to investigate the effect of environmental fluctuation on a complex prebiotic system, we carried out a reaction cycling process by seeding a new reaction with the products of the previous one, for several cycles, see **Figure 47**. This was done on a model-system for a prebiotic-type combinatorial explosion, which was created by joining together two (well known) analytically challenging reactions: the formose reaction and the formamide condensation. We also explored a selection of different mineral environments, to assess whether the selectivity imparted by the environment can be amplified through recursion, whilst truncating the combinatorial explosion by reducing the overall number of products.

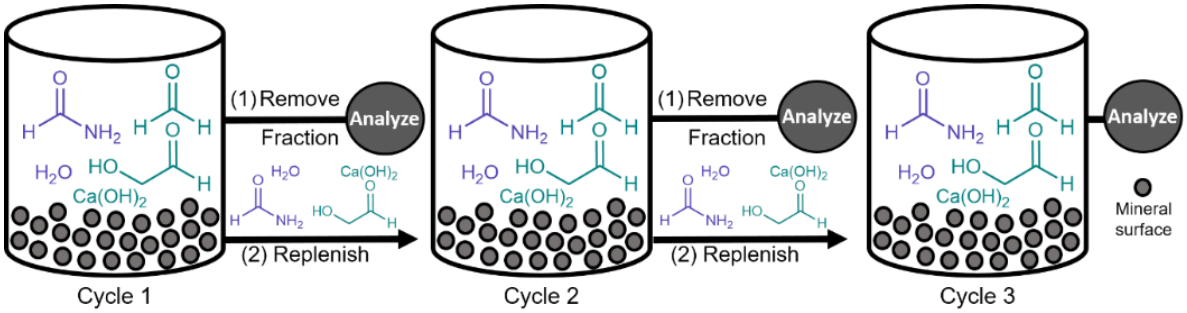


Figure 47. *Recursive cycles:* After each reaction, the supernatant is removed for analysis and a small fraction is left in the reaction vessel and used to seed a next reaction.

A general procedure for a recursive cycle entails starting a reaction by adding Formaldehyde (0.5 mL), glycolaldehyde (0.0126 g), water (2.25 mL), formamide (2.25mL) and calcium hydroxide (0.0705g) into a 22 mL borosilicate glass vial (e.g. Pyrex). The reaction is stirred at 1200rpm with a magnetic stirrer and heated at 50°C for 48 hours. After the 48 hours, the top two-thirds (1.5 mL) of the reaction volume was removed (e.g. supernatant) and placed in two 1 mL HPLC vials for analysis. The remaining fraction in the reaction vessel was used to seed the next reaction. Then, we start the next cycle by replenishing the reaction vessel with the same starting materials as the previous one and letting it react under the exact conditions as before. This process was repeated, for a final iteration, giving rise to three recursive cycles. Also, it must be noted that the reaction was allowed to cool to room temperature before the fraction was removed. This was followed by a removal of excess cations in solution (e.g. Ca^{2+}) with an Amberlite™ Ion-exchange resin, before the supernatant was diluted to 1 in a 100 with MS grade water. Finally, the solution was filtrated with a syringe filter (0.22µm cut-off) and placed in an HPLC sample vial, prior to the analysis.

We selected seven different mineral surfaces as a way to assess how different environment types could affect the product distribution of the formose-formamide reaction. The minerals were selected to ensure a good variety of mineral-types was present. For this, the selection includes phosphate, iron, borate, or alumino-silicates (e.g. clay) minerals. Furthermore, some of this minerals are hydroxides, meaning that they came into existence after the Great Oxidation event, according to Bob Hazen's theory of mineral evolution.[75] This was done purposely, as we wanted to include oxygen-containing minerals and not limit the selection to prebiotically plausible minerals. For a list of the selected mineral surfaces, please see

Table 4 below.

Name	Category	Chemical Formula
Goethite	Iron mineral	α-FeO(OH),
Montmorillonite	Clay mineral	$(Na, Ca)_{0.33}(Al, Mg)_2(Si_4O_{10})$
Hydroxyapatite	Phosphate mineral	$Ca_5(OH)(PO_4)_3$
Chalcopyrite	Iron mineral	$CuFeS_2$
Ulexite	Borate mineral	$NaCaB_5O_6(OH)_6{\cdot}5H_2O$
Zoesite	Clay mineral	$Ca_2Al_3(SiO_2)_3(OH)$
Quartz	Clay mineral	SiO_2

Table 4. Selection of mineral surfaces for the recursive formose-formamide experiments.

After the 3 recursive cycles were completed, we found that the recursive action resulted in a lower number of individual products, with or without a mineral surface. This demonstrated that reaction cycling has a significant effect on the product distribution. Furthermore, we also observed a significant increase in the yields of certain species when the minerals were present, showing that selection by the environment also plays a role in determining the product mixture, see **Section 2.2.2.3**

2.2.2 UPLC-HRMS: An untargeted method

In order to investigate and establish the nature of any differences in the product distribution without bias, untargeted analysis of the mixtures was conducted. The separation step was achieved by Hydrophilic Interaction Liquid Chromatography (HILIC), a chromatographic method that has become a recent favourite for the analysis of convoluted matrices with a broad chemical properties (such as blood serum analysis),[203] alongside Reversed-Phase chromatography. The Ultra-Performance Liquid Chromatography was coupled to tandem mass-spectrometry (UPLC-MS/MS) and carried out in a data dependent fashion, which allowed us to investigate the resulting chemical space without having to target any particular compound.

However, we found that the available data processing software is usually designed to allocate the discovered products into metabolic pathways. Consequently, most small molecule databases conventionally used in in tandem mass-spectrometry (UPLC-MS/MS) untargeted analysis would only match records corresponding to known biological systems. While this is to be expected, when we consider that the human metabolome project helped start the

process of developing chemical databases from experimental data, which resulted very useful in the mapping of our metabolic profile; a challenge remains in the analysis of complex prebiotic mixtures, since we don't want to be biased by looking for biological databases in order to match and identify the compounds. Therefore, we decided to take a more 'systems' approach into the data analysis. This was clearly influenced by the previous Miller-Urey results, where we found a different approach can elucidate non-trivial trends in the product distribution.

Classical metabolomics-type workflows use the patterns generated by mass-spectral features as a way to construct chemical biomarkers. These mass spectral features are built of a combination of retention time and mass to charge ratio (e.g. RT + m/z).[188] Therefore, we wrote a series of python scripts that would allow us to extracts the mass-spectral features, from the UPLC-MS/MS raw data, with the help of available mass-spectral libraries and deconvolution tools. This allowed us to simultaneously extract the m/z features and plot them. The resulting plots generated unique patterns for each sample and taking the assumption that each features represent a product in the complex mixture, then the features correspond to observable differences in the product distribution arising from the recursive cycles and mineral surfaces.

However, in order to validate our in-house feature generator and 'omics' based approach to complex mixture analysis, we processed the data using CompoundDiscoverer™ (Thermo Scientific),[204] a conventionally used software for processing untargeted mass-spectral data. The feature extraction process is embedded in the software's available data processing workflow and resulted in the same trend as with our scripted version, also, enabling the extraction of ion chromatograms (EIC's) in a targeted fashion. While this method generated fewer features overall, the trends were consistent throughout the experiments (see **Section 4.2.4**). Furthermore, during the data- dependent acquisition (DDA) of the mass-spectral data, the top-most intense peaks were fragmented further into MS^2 fragments. This allowed us to identify some of the products using database matching, in silico calculations and validation against pure standards to confirm chemical identities. By using the MS^2 data, we were able to perform qualitative structural analysis and identify some of the features as Ribose and Uracil, the building blocks of RNA. As well as, the traces of nucleoside formation, which had never been reported before starting from such simple precursors.

2.2.2.1 UPLC method development: UPLC-CAD

The chromatographic method development was carried out in a Thermo-Dionex UltiMate3000 UPLC system equipped with a Charged Aerosol Detector (CAD). The CAD detector was initially used to overcome the hurdles of other conventionally used detectors in liquid chromatography, such as UV/Vis or Fluorescence, which require the presence of a strong chromophore in the reaction products we want to investigate (i.e. analytes). In this sense, the CAD detector is a 'universal' detector, which makes it capable of detecting any compound without the need of an already present chromophore. This was a significant aspect behind its selection, since it permitted an assessment of the separation efficiency without the need of any extra-steps in the sample preparation procedure, prior to the chromatographic run. From this point of view, the detection system made it easier to achieve our objective, which was to develop a chromatographic method that would simultaneously separate the main products of the reactions we are trying to combine: sugars (i.e. mannose, ribose, fructose, glucose and sucrose) and nucleobases (i.e. thymine, adenine and cytosine), for the formose reaction and formamide condensation, respectively. For this, the use of external and internal standards of the products known to be present in the complex mixtures was employed, as way to evaluate the UPLC method. However, it must be noted that the UPLC-CAD system was selected to investigate the efficiency of the chromatography, but the ultimate goal is to couple the separation method to a high resolution mass spectrometer (HR-MS).

The chromatographic separation was carried out with a HILIC mode column, due to its suitability across a wide range of chemistries (i.e. good for hydrophilic, hydrophobic and neutral material). HILIC is typically used when the retention times in reverse phase mode (RP) are insufficient, as a consequence of more polar analytes being present alongside highly hydrophobic material.[205] Also, HILIC can be used as a replacement to normal-phase mode (NP), since it has a higher ionization efficiency, making it particularly useful when using electrospray mass spectrometry, as the analyte can be ionised in solution whilst still showing good retention, especially when considering that the goal is to couple the separation to HR-MS, which in this case has an ESI source. Additionally, HILIC overcomes the drawback of poor water solubility of some analytes, which is often problem with normal phase. In this context, the HILIC technology is ideal for complex mixture analysis by liquid chromatography coupled to mass spectrometry. Consequently, it comes as no surprise that since it first appeared in the 1990s, it has become a preferred column for small-molecule

analysis in complex matrices (such as, metabolite and protein analysis).[206] The mobile phases employed in HILIC chromatography contain a high degree of organic solvent (generally 70% or greater) and a typical gradient would involve altering the aqueous composition between 5 and 30%. One of the most popular solvents (if not the favourite) is acetonitrile, due to its aprotic and intermediate polarity, a characteristic which encourages the retention of polar analytes. The exact mechanism behind HILIC chromatography is not well understood to date, however most experts agree that the bulk mechanism involves the polar analyte partitioning into and out of a water layer, which is adsorbed onto the surface of the polar stationary phase.[206] Furthermore, it must be noted that the retention in HILIC is mainly affected by adjusting the eluent, the type and concentration of the buffer and the pH value; but considering the high chemical diversity within the complex mixtures in this study, the retention may depend on several additional factors.

In order to explore the efficiency of separation in HILIC mode, we compared two columns: a ZIC-HILIC and an amide-HILIC. According to the Thermo-Scientific product specifications[207] the ZIC-HILIC column offers an enhanced retention of charged and neutral polar compounds. Also, in a recent comparison study across five different HILIC columns, the ZIC-HILIC performed the best on a large set of hydrophilic metabolites for the untargeted (UPLC-MS/MS) analysis of human urine and plasma.[208] Nonetheless, we also considered a second HILIC column, the amide-HILIC. The strong hydrogen bonding interactions between the stationary phase and the analytes, provided by the amide-HILIC, results in a stronger retention of highly polar material, when to compared to other HILIC phases. This makes the amide-HILIC column better suited for separating very polar analytes and a variety of hydrophilic molecules, such as carbohydrates and peptides, as well as, having an operational pH range higher than the ZIC-HILIC. However, which one would work best in the case of the formose-formamide products, was not clear. Therefore, we decided to compare their separation efficiency using a standard solution of five different sugars. In **Figure 48** below, observable differences can be seen for the resolution and sharpness of the chromatographic peaks. The amide-HILIC resulted in better peak-shapes overall (**Figure 48a**), when compared to the ZIC-HILIC column. Correspondingly, taking into account that the formose reaction and formamide condensation produced very polar compounds (amongst them sugars and amino-sugars) the amide-HILIC was selected for the chromatographic separation of the product mixtures.

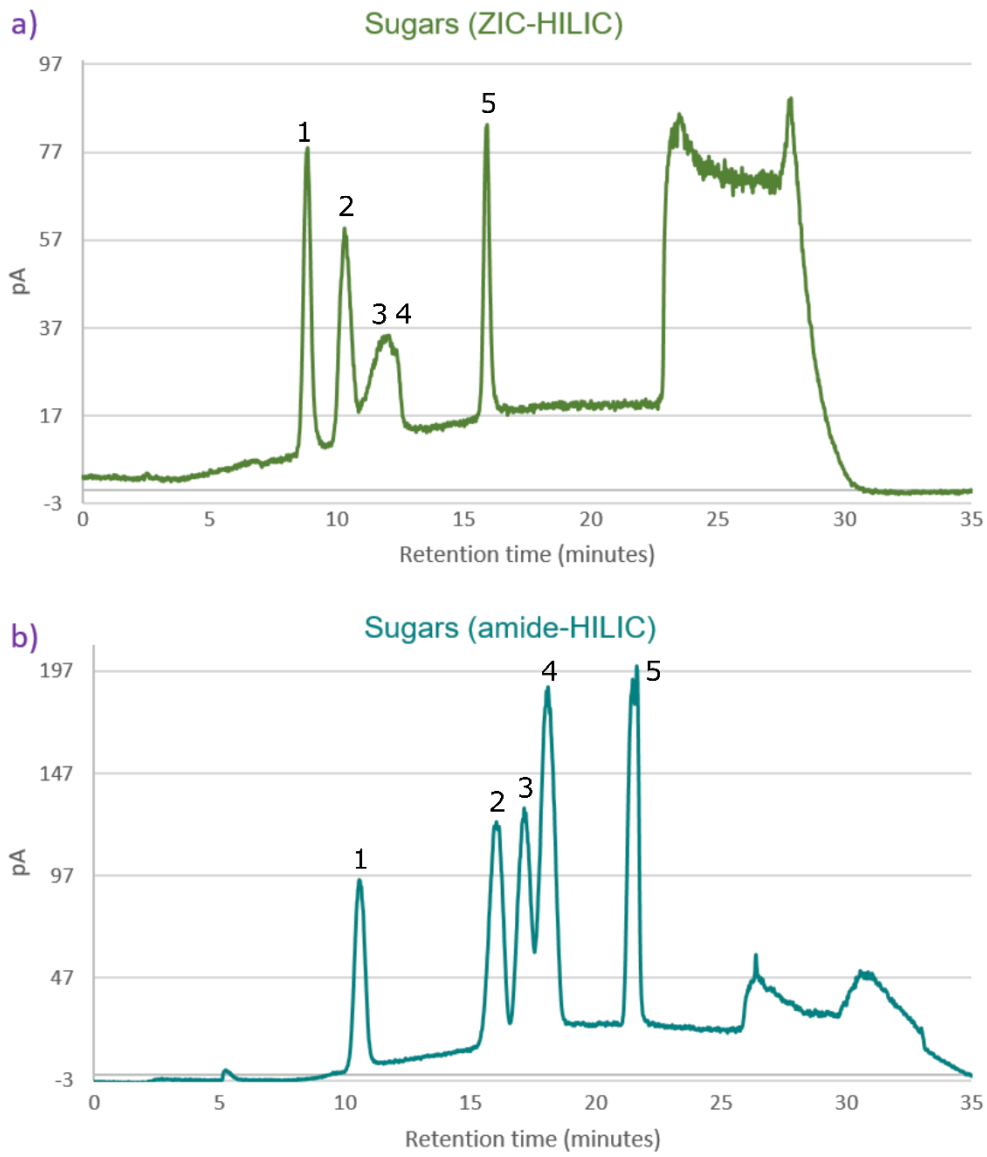


Figure 48. UPLC-CAD chromatogram of the sugar standards (**1**- mannose, **2**- ribose, **3**- fructose, **4**- glucose and **5**- sucrose) in two different HILIC mode columns: (**a**) ZIC-HILIC and (**b**) amide-HILIC. The peaks are observably better resolved with the amide-HILIC column.

After selecting the column, the method development was relatively simple. A methodology described by Idborg, *et al*. for HPLC analysis with HILIC was a good starting point.[209] The elution method was used (mostly) as described, with only minor alterations. The gradient was adjusted by making it 10 minutes longer, in order to give enough time for the peaks to resolve and elute at least within 30 seconds of each other (e.g. not co-eluting). This was done after two of the sugar standards (ribose and fructose) eluted too close to each other to be resolved properly. The aforementioned considerations resulted in the following method: A linear gradient mixture of solvents A (water w/20 mM Ammonium Acetate, pH = 5) and B (100% acetonitrile w/0.1% v/v formic acid) over 35 min as follows: 0 min, 95% B; 5% A; 15 min, 75% B - 25% A; 21 min, 5% B - 25% A; 25min, 25% B - 75% A; 35 min 95% B -

5% A; in a method adjusted from Idborg, *et al.* The column was maintained at 30 °C and the CAD detector was set to a nominal evaporator temperature of 65.0 °C (+/- 5.0 K). A selection of several standard solutions were prepared, with a 50 mM concentration (stock-solution). The sugars selected as representative of the formose reaction products where mannose, ribose, fructose, glucose and sucrose, as seen in **Figure 48**. To consider some of the main products of the formamide condensation, three nucleobases were selected: adenine, cytosine and thymine. Also, a nucleoside standard of Thymidine, in case we could detect one of the sugars (e.g. ribose) attached to the nucleobase thymine. Two standard solutions were prepared: one for the sugars and another for the nucleobases/nucleoside. The standards were then analysed one by one and as mixtures, through the same chromatographic method. These standard solutions allowed us to identify the elution profile and retention time for each chromatographic peak corresponding to the single standard compounds. See **Figure 49** for a chromatogram of the standard solution of nucleobases. Then this was used to identify the compounds within the resulting product mixtures, by comparing their chromatographic information. In other words, the match was carried out by external standard validation. In addition, we also did internal standard validation, by introducing the standard mixture in a 1:1 ratio (v/v).

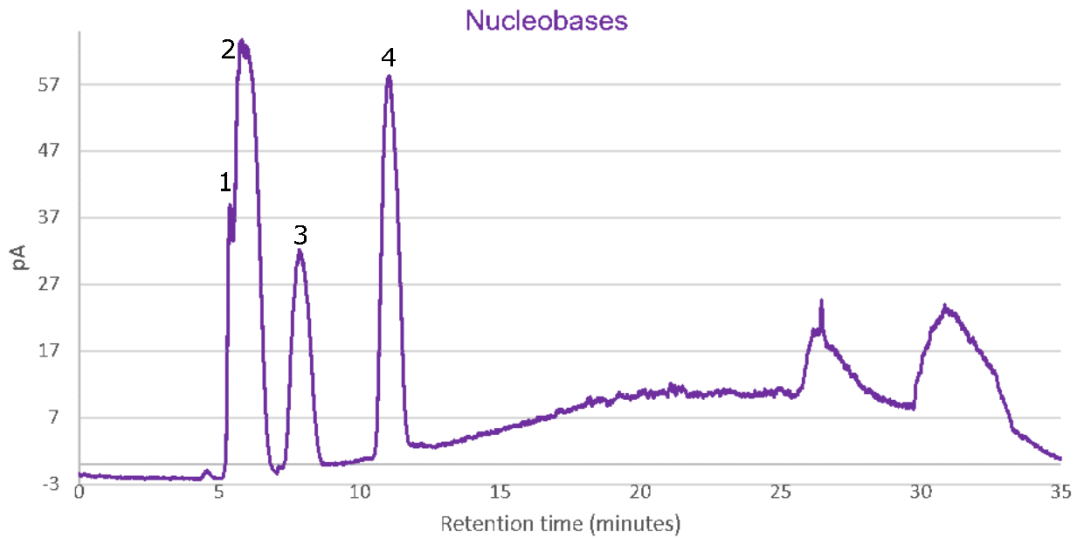


Figure 49. UPLC-CAD chromatogram of the standard mixture of Nucleobases, eluted in the following order: **1**- Thymine, **2**- Thymidine, **3**- Adenine and **4**- Cytosine.

Furthermore, the longer method enabled for the simultaneous separation of sugars and nucleobases. As observed in **Figure 48-49**, the retention time (elution range) for the nucleobases in this method does not interfere with those of the sugar standards, making it

possible to resolve both within a single chromatographic method. Also, the resulting chromatograms for the recursive cycles indicated the formation of all three nucleobases and the nucleoside thymidine, as well as several other unidentified peaks eluting at a later retention time that the nucleobase standards. The intensity of the identified peaks changes across recursive cycles, also seen in **Figure 50**. The peaks appear to reduce in intensity, over recursive action, resulting in a lower intensity for cycle 3. This would not necessary mean degradation, since there is the possibility of surface adsorption and conversion to another product/polymer.

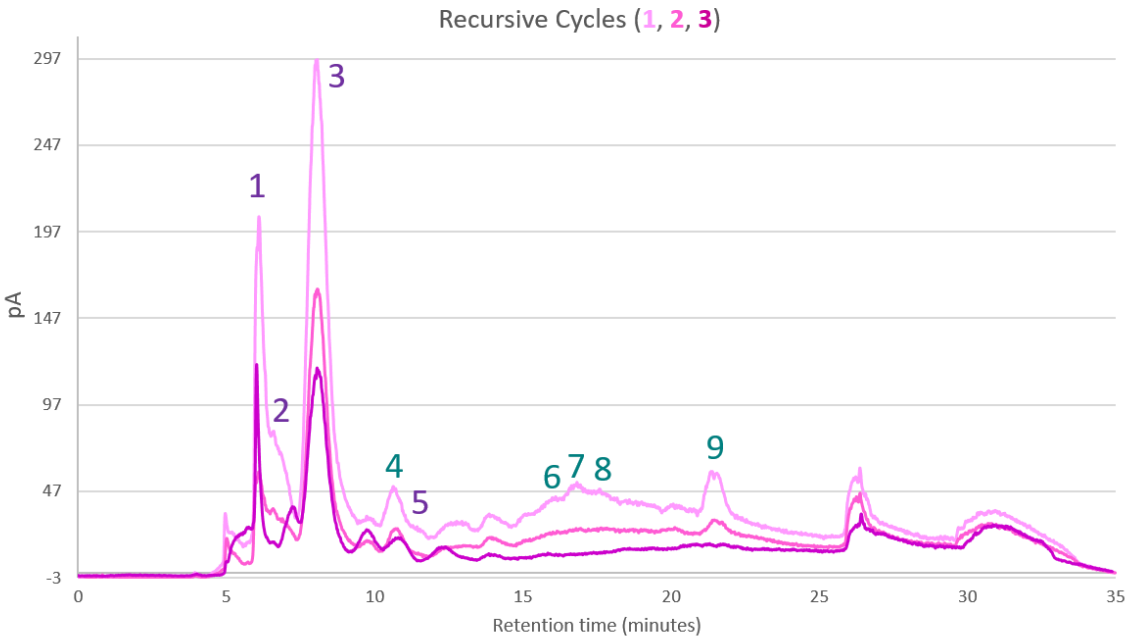


Figure 50. UPLC-CAD chromatogram of recursive cycles for the control (no-mineral) reaction. Differences in intensity can be observed for the peaks corresponding to (1) Thymine, (2) Thymidine, (3) Adenine, (4) Mannose, (5) Cytosine, (6) Ribose, (7) Fructose, (8) Glucose and (9) Sucrose. Highlighted in *blue* are the products for the formose reaction and in *purple* for the formamide condensations.

The UPLC-CAD analysis was developed as a complementary high-throughput analytical technique to compensate for the limitations/bias of other detectors, such as the diode-array detector, during the method development process. Another important aspect was the intended coupling of the chromatography to HR-MS (in this case with an UPLC-Orbitrap) detector, that would enable the plausible identification of unknown compounds. Therefore, the chromatographic conditions (i.e. column composition and mobile phases) were selected and adjusted to take into consideration the future integration of a MS detector with an ESI source.

2.2.2.2 Compound identification

Ultra-Performance Liquid chromatography coupled to tandem mass spectrometry (UPLC-MS/MS) has proven to be a powerful tool for the analysis of small molecules, by providing highly accurate and precise characterization of a broad range of analytes. However, the identification of the so-called "unknown-unknowns" (e.g. unexpected, but potentially important compounds for which there are no available spectral library data) is extremely challenging. For this reason, a powerful software control system was developed, which enables the instrument to perform data dependent acquisition (DDA). This acquisition method switches from the full-MS (MS^1) scan to MS/MS mode (MS^2) using a pre-established data-dependent criterion. The DDA method removes the need to re-analyse (re-run) the samples in MS/MS mode, once the target precursor ions have been identified by a full-MS analysis, reducing the amount of time needed to acquire and analyse the datasets. Also, this method greatly reduces the complexity of the data processing. Particularly, when compared to data-independent methods, which take into account all detected peaks in the MS^1 full-scan. However, the approach taken by the DDA method does not guarantee that all features (or analytes) of interest will be analysed in the second MS/MS step. Nonetheless, this can be mitigated by including a dynamic exclusion criteria that avoids repetitive MS/MS fragments by excluding them after a set period of time, consequently increasing the precursor feature sampling and the likelihood of detecting all the relevant features.

The chromatographic method described previously in **Section 2.2.2.1**, was performed in Thermo Vanquish Ultra-performance liquid chromatography system coupled to a Thermo Orbitrap Fusion Mass-Spectrometer. All samples were injected directly in ten microliter (10 µl) aliquots, while the chromatographic separation was achieved with an amide-HILIC C18 column. The samples were eluted in a linear gradient, with the column was maintained at 30 °C and the MS spectra was collected for 30 minutes in positive mode over a scan range of 50–500 m/z.. The Data-Dependent Acquisition (DDA) was performed by prioritizing the most intense fragments in a 3 second window with an intensity threshold of 5.0E4 (in counts) and dynamic exclusion, after selection for 15 seconds window (in order to avoid the selection of the same fragments), using the ion trap isolation with a HCD collision energy of 35 eV and a resolution 15000. For a schematic depiction of the UPLC-MS/MS DDA method, see **Figure 51**.

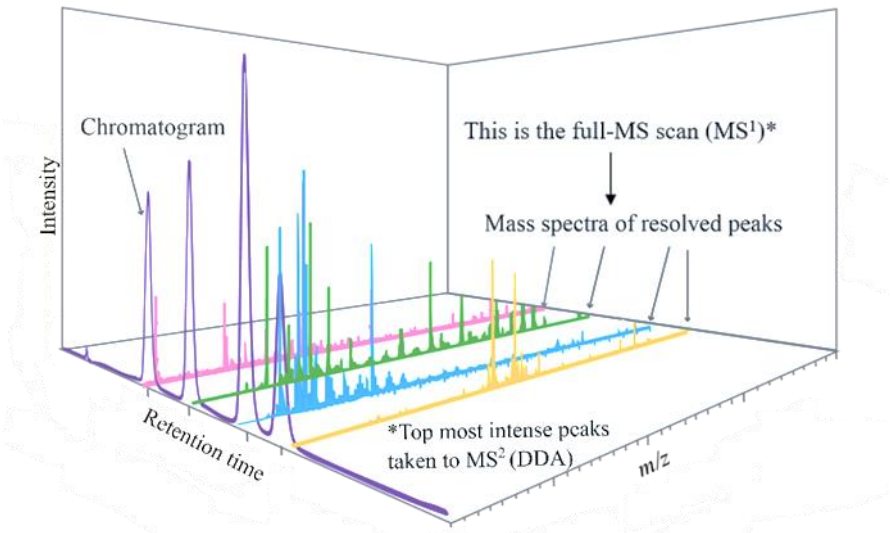


Figure 51. *UPLC-MS/MS method:* The chromatographic separation is complemented by mass-spectrometry detection in a DDA manner. A full-MS scan is done every 3 seconds, taking the top three (3) most intense peaks to MS/MS, with a dynamic exclusion after 15 seconds.

The CompoundDiscoverer™ (Thermo Scientific) software is designed to enable the identification of small-organic compounds from the UPLC-MS/MS analysis of complex chemical mixtures. It contains a selection of different workflows, in which the user is able to change many parameters (e.g. possible adducts, elements to take into consideration, intensity thresholds, signal-to-noise ratio, etc.), amongst others and adjust it to the experimental conditions of the data acquisition and chromatographic method. The workflows usually integrates a combination of methods for data extraction, including retention time alignments and mass-spectral feature analysis (ANOVA), amongst others. This is done in order to ensure chromatographic corrections of the untargeted raw data in a consistent and accurate way. Also, Extracted Ion Chromatograms (EICs) of the identified products, as well as the other (unknown) features, are easily accessible through the User-Interface. An example of the processing workflow interface is presented in **Figure A14**.

A database search was made possible by the DDA method, which allowed for the most intense MS^1 peaks to be fragmented further into their MS^2 fragments. However, to identify the detected features confidently, the resulting MS/MS (MS2) pattern was compared with those obtained by a pure standard. We identified plausible compounds for the detected features, by carrying out a database match search in MZcloud and Chemspider, made possible by the CompoundDiscoverer™ (Thermo Scientific) software suite as part of their available workflows. For details see **Section 4.2.4.5**. Nonetheless, it must be noted that the databases used in the workflow differ in how they are built, with Chemspider being made

from experimental archives and the mzCloud based through in-silico calculations, resulting in different but complementary matches.

Furthermore, once the compounds have been identified through any of the databases, the resulting mass-spectral pattern in the MS^2 spectrum is then compared with those of a pure standard to fully validate the data-base match. (see **Figure 52**). This was taken as validation strategy, through a match in the exact mass (+/- 0.5 m/z, under the same adduct) and retention time (+/- 30s-60s) of the pure standard, when subjected to the same UPLC-MS/MS method. The analysis of the MS^2 data allowed for a qualitative structural analysis and identified some of the detected mass-spectral features, indicating the presence of Ribose and Uracil, the building blocks of RNA. Also, other nucleobases such as thymine, adenine and cytosine were found, which coincides with the results published previously on the condensation of formamide.[157]

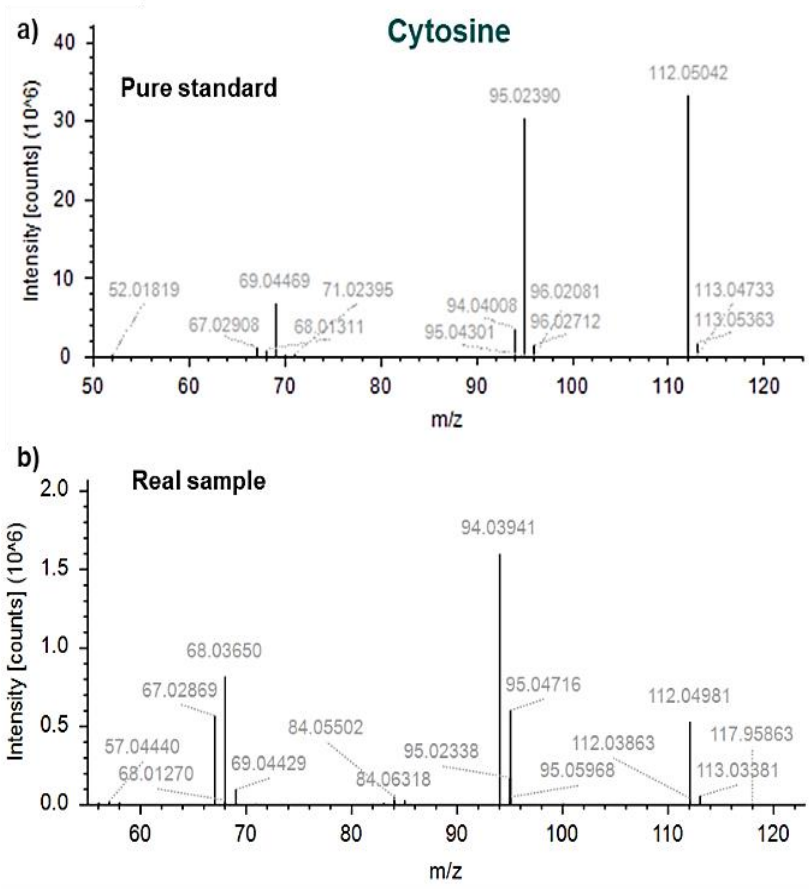


Figure 52. UPLC-MS/MS spectrum for Cytosine (*m/z*: 112.05, Adduct: [M+H]) in (**a**) a pure standard and (**b**) a real sample, (Control, Cycle 3); Characteristic MS/MS fragments are 95.03 and 69.04 m/z

However, by extracting the ion chromatograms for the features corresponding to the nucleobases, we observed that their intensity changes across recursive cycles and the type of mineral environment employed. While this cannot be taken as an absolute measure of their concentration in solution, the intensity variation across cycles does indicate that the nucleobases take part in a dynamic network of chemical reactions, since the reaction conditions were relatively mild and therefore should not have promoted the thermal degradation of the compounds. Therefore, we can assume that the change in the intensity of the EIC's corresponds to the compound being formed or consumed during the progress of a given cycle. For an example on how the EIC's change over cycles, see **Figure 53**. Where the feature that corresponds to the nucleobase thymine, in the presence of chalcopyrite, is seen to reduce over recursive cycles.

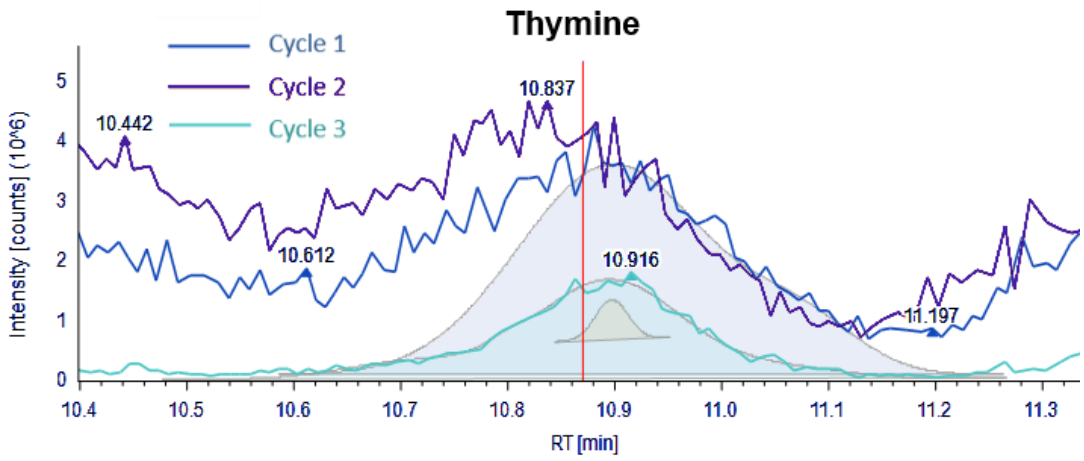


Figure 53. Intensity variation of Thymine across recursive cycles: Extracted Ion Chromatogram for thymine (*m/z:* 127.05, Adduct: [M+H]) in the presence of chalcopyrite. Differences in the intensity (in counts) can be seen from Cycle 1 to Cycle 3.

Furthermore, the nucleobases were not only present in our product mixtures but also produced preferentially on mineral surfaces, as observed in the difference between intensity scales for the selected ion in the extracted ion chromatogram (EIC) for uracil, in the presence or absence of the mineral chalcopyrite, and in the peak areas, shown in **Figure 54**. In addition, we detected hexamethylenetetramine (HMT) across all reactions. HMT was discovered by Aleksandr Butlerov in 1859 and is prepared industrially by combining formaldehyde and ammonia.[210] The significance of HMT in prebiotic chemistry has been discussed previously,[145] particularly in its role of incorporating formaldehyde (from its reaction with ammonia, which is generated *in-situ* by the decomposition of formamide) into a more stable compound, possibly allowing for it to be concentrated in a prebiotic, evaporative environment. The concentration of HMT changed across recursive cycles

(**Figure 55a-b**), with a significant drop being observed after the second cycle for all samples (including the control). We postulate that the HMT is depleted by reaction with the products of Cycle 2, but we currently have no definitive evidence for this, or a mechanism responsible.

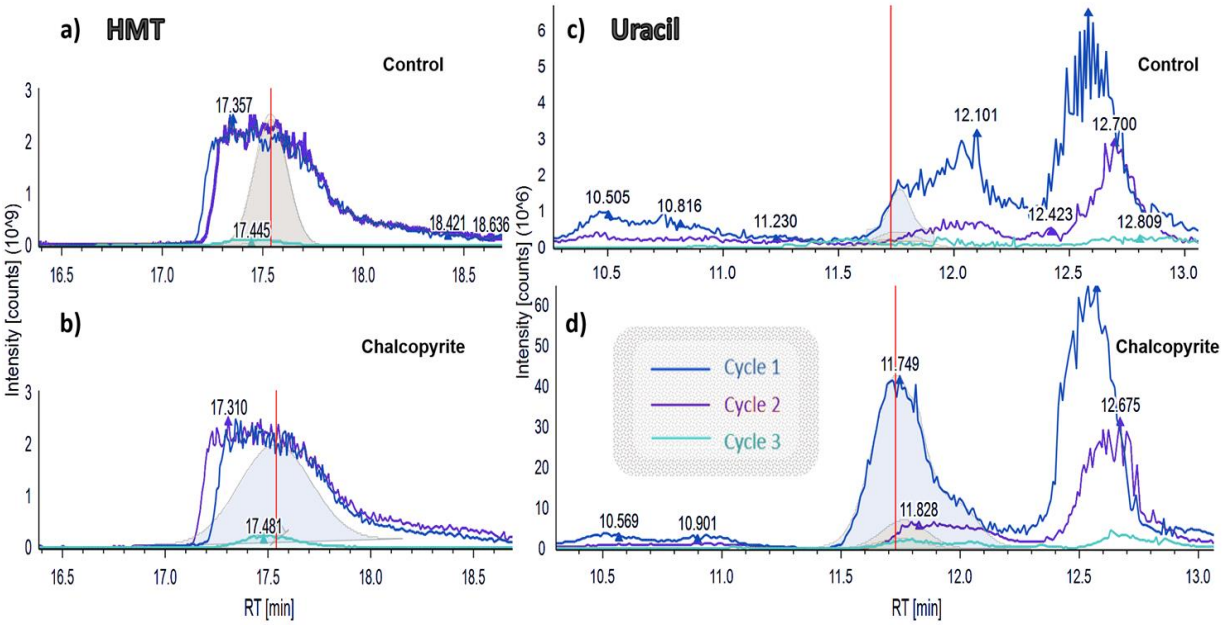


Figure 54. Identification of RNA building block (uracil) and HMT: Extracted Ion Chromatograms (EICs) of HMT (m/z: 141.11, Adduct: [M+H]) for (**a**) the control reaction and (**b**) in the presence of a mineral surface (chalcopyrite). EICs for uracil (m/z: 113.03, Adduct: [M+H]) (**c**) in the control reaction and (**d**) in the presence of a mineral surface.

In order to display how specific products change over recursive cycles, we calculated the relative abundance for the selected ions, as shown previously with **Figure 55.** The selected features had an exact mass to HMT, uracil and ribose and double validated by comparing their resulting MS^2 spectra with those of pure standards. This was done in a qualitative manner, as an accurate quantification of these compounds in different complex matrices would require a targeted analysis, which is beyond the scope of this work. We acknowledge this limitation from the untargeted acquisition method and aim to complement it with a targeted workflow in further investigations.

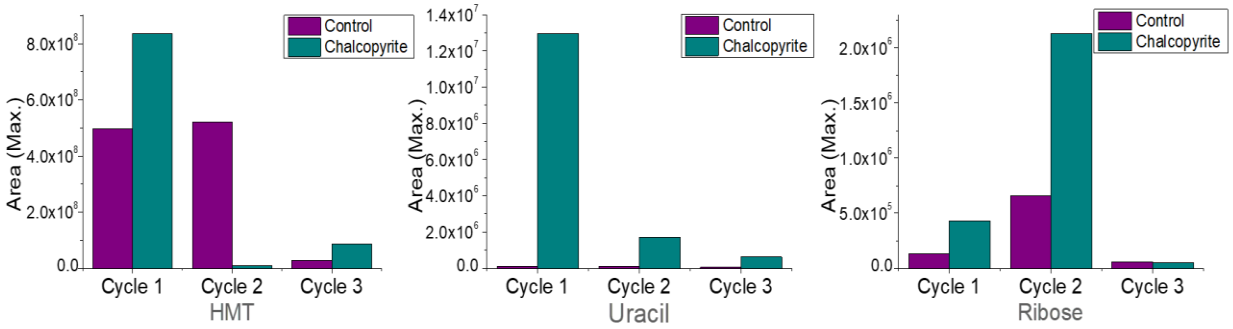


Figure 55. Relative abundance of HMT, uracil and ribose calculated by integration of EIC's for the selected ions (features).

In a targeted workflow, we would need to be able to identify the relevant features, responsible for the dynamics in a given chemical system. The relevance of certain features over others in complex product distributions is not trivial and it would benefit from a discovery-driven investigation, such as the one used in metabolomics-type workflows and this work. This approach generates a more robust overview of the highly complex product distribution generated in analytically intractable mixtures, as a mean to further our understanding of complex chemical systems and their intrinsic reproducibility. Due to the high complexity in the product distribution of combinatorial explosions, a satisfactory reproducibility assessment would need a large number of experimental replicates, which in turn requires a high-throughput experimental design. Also, a series of well formulated controls to avoid the false identification of features. This became apparent when a previous publication claimed the formation of a polymeric specie (PEG) in a prebiotic mixture,[211] which we later found to be present in the sample blank. Different to an instrument blank, the sample blank was made by having the reaction solvent (50:50 water-formamide in this case) go through exactly all the sample preparation procedure, this allowed us to narrow down that the source of PEG contamination must be coming from one of the plastics in this step and not from the instrumental equipment itself.

However, while a complementary targeted workflow is not assessed directly in this work, it has indeed enabled the possibility for such studies and therefore the possibility of quantification of the relevant features. A comprehensive overview of the resulting products would be substantial in order to investigate the relationship within the chemical features of a complex product distribution and to draw any meaningful conclusions on which features are important to the system. Justifying the need of a 'discovery' approach before being able to carry out any quantification studies through targeted workflows, which would require a

previous identification of the most-relevant features.

2.2.2.3 Feature generation

The generation of features based on exact mass (m/z) and retention time (RT), made it possible to achieve a meaningful representation of the product distribution from mass-spectral data. The features represent unique reaction products and their number corresponds to the number of individual species, providing a way to gauge the complexity of the mixture. Also, in order to validate our in-house feature generator and 'omics'-based approach to complex mixture analysis, we compared the number of resulting features with those detected by the CompoundDiscoverer™ (Thermo Scientific) software (which is usually used for the processing of untargeted mass-spectral data). The processing software includes a selection of customizable automated workflows, which can be adjusted to the experimental parameters of the acquisition method. The workflow integrates multiple nodes, which include (a) retention-time alignment, (b) compositional prediction and (c) database search through in silico (Mzcloud) and experimental (Chemspider) libraries by matching the exact mass and MS/MS fragmentation pattern of the detected features. However, we found that the overall number of features detected through this processing workflow was considerably less (roughly 30-40%) than those generated by the in-house feature generator, consistently across all samples. However, the resulting trends were conserved in all the data sets analysed, for which we could assume that the feature extraction (generation) was done correctly.

The in-house feature generation was carried out as following:

(a) The raw data was extracted from its vendor format (.raw) with MSConvert from ProteoWizard, into an .mzML format before loading it in a Python environment.[212] We selected a peak-picking algorithm that is vendor specific and performs centroiding on spectra within a selected range, in this case MS^1 and MS^2.

(b) The python package PymzML was used to extract MS^1 values, Retention Time (RT), Intensity (in counts) and MS^2 values (with their corresponding RT and Intensity).[213] The intensity threshold used to filter the m/z fragments, for both the MS^1 and MS^2 was of 1^{E04} counts.

(c) The detected MS^1 values are truncated to their fourth decimal and their intensity to the second decimal. Also, the retention time was truncated to the second decimal and all features detected after 20 minutes of the chromatographic method discarded. This was done as a way to avoid any features that elute during the re-equilibration step.

(d) The duplicate (MS^1) values where filtered further, by eliminating values that had same exact mass to the second decimal value, besides being within an acceptable retention time window (+/- 30 s) of each other. For all samples, the number of duplicates filtered out is roughly one – third (1/3) of the total number of features originally detected. All the features were truncated – intensity threshold in the extractor- and retention time to remove the features eluting during the re-equilibration step.

(e) Furthermore, the number of MS^2 fragments obtained for each feature (MS^1) was calculated, as a generic way to access the overall complexity of the molecules within the resulting product distribution.

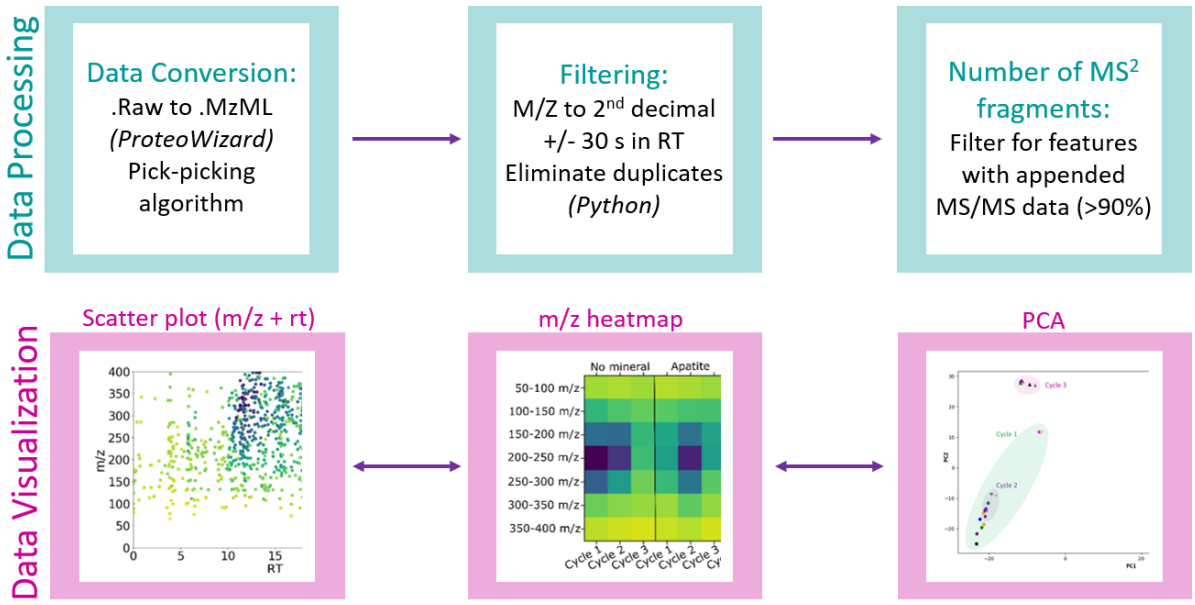


Figure 56. Simplified scheme of the feature generation and consecutive data visualization. Scatter plots in this section, heatmap in **Section 2.2.2.4** and PCA in **Section 2.2.2.9.1.**

In order to visualize the generated features, scatter plots for each sample were made in Python with the matplotlib package. This allowed us to observe and assess the differences in the detected features, by comparing their position in the resulting plots. The position of the features can be complemented with a (*very*) general assessment of their composition depending on their location, as the chromatographic method indicates that more hydrophilic

and polar material would elute at later retention times. Moreover, the number of MS/MS (MS^2) fragments would also give an insight to the compositional differences across the features; besides being indicative of their complexity, if we can roughly assume that between two features with similar m/z, a feature of higher compositional complexity will have a larger number of MS^2 fragments. This assumption is also part of ongoing research within the research group that involves a new description of chemical complexity,[214] which effectively correlates with the number of MS/MS fragments, with complex molecules producing a larger number of MS^2 fragments.

By visualization of the data sets with the scatter-plots generated, we could observe a series of trends. The number of features detected in each reaction changes as an effect of recursive action, as seen in **Figure 57**. The resulting trends show that the number of featured decreases from Cycle 1 to Cycle 3, even in the absence of any mineral surface, as seen in **Figure 57a**. However, we can also observe the number of features go down in the presence of mineral surfaces (**Figure 57b, 57c**). The feature distribution is visually different for each environment from the very first cycle, which is to be somewhat expected since the minerals will solubilize in different rates and change the overall pH of the reaction in different ways. This was not experimentally monitored and can only be a suggested from previous results in comparable studies of mineral inclusions in prebiotic reactions.[83] The initial differences where carried on through the recursive cycles, generating a unique feature pattern for each environment, as well as for each cycle in the recursive process. . This suggests that the reactions proceed along different trajectories, towards different product distributions, as a direct result of the mineral environment.

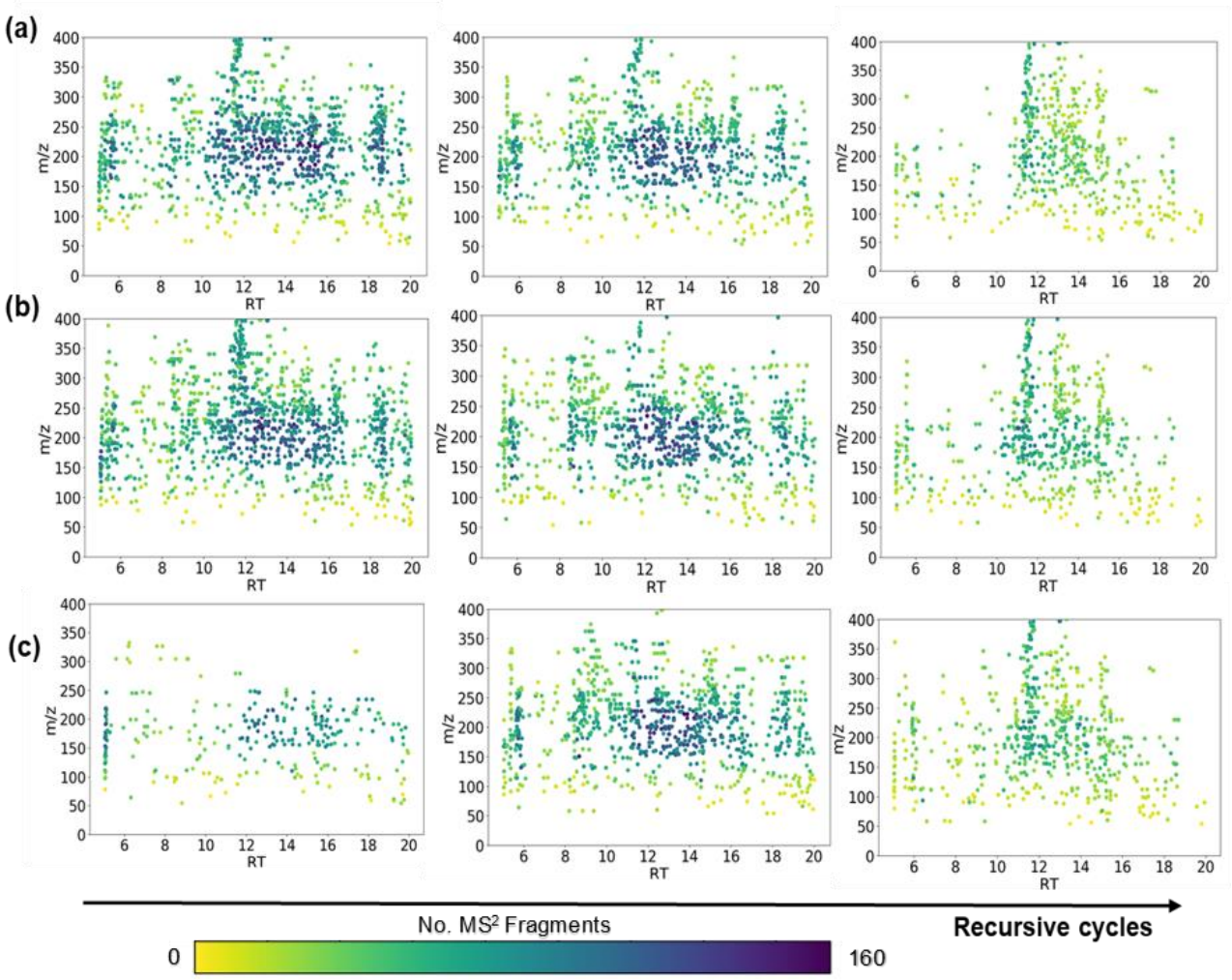


Figure 57. Product distribution (mass spectral features) for Cycle 1, Cycle 2 and Cycle 3 in the (**a**) non-mineral reaction and with (**b**) chalcopyrite and (**c**) goethite as a mineral surface. Differences across the pattern generated by the features and their corresponding number of MS^2 fragments can be observed.

Moreover, for all the reactions studied recursively, the number of detected features decreases from Cycle 2 to Cycle 3, demonstrating that the action of recursive cycles is limiting the combinatorial explosion expected from these reactions. However, this trend was observed to evolve differently across the environments. In the case of Chalcopyrite, Quartz and in the absence of any mineral surface (control), the number of features reduces linearly from Cycle 1 to Cycle 3. While in all the other mineral environments, the number of features peak in Cycle 2 and decrease again in Cycle 3. This is yet another reason to believe, that the reaction products in each environment are different and that a selective process is present, which is then sequentially amplified through the reaction cycling process, generating consistently different product distributions.

Furthermore, the number of MS2 fragments resulting from the MS/MS fragmentation of the features was calculated, see heatmap in **Figure 57**. This was done to assess the overall complexity of each feature, under the assumption that compounds with a larger molecular complexity will generate a greater number of MS^2 fragments. Far from perfect, it's a generic way to look for compositional differences across the product distribution of the reactions. This allowed us to see that the majority of the generated products with a high number of MS^2 fragments, and assumed complexity, are initially (Cycle 1) spread across a wide range of retention times. However, after two recursive cycles (Cycle 3), the features with a larger number of fragments were concentrated towards the middle of the chromatographic run (retention time between 10 to 14 min). This could be an indication (assuming a sustained elution profile for the chromatographic process) that most of the complex material generated after several recursive cycles is mostly composed of mixed or mid-polarity, when compared to the initial product distribution (Cycle 1).

2.2.2.4 Feature distribution by m/z

As a way to make the large volume of data more accessible and easier to visualize, the detected features were binned by their m/z values, resulting in yet another way to fingerprint the product distribution, **Figure 58**. The unique patterns in the feature distribution that arose as an effect of the environment are made clearer and their uniqueness can be appreciated visually. For all reactions, a larger number of features (or products) was detected in the range of 150 to 300 m/z. Also, distinct variations between the product distributions can be observed from the very beginning (in Cycle 1) for all mineral environments and the reaction with no mineral. The number of detected features in each m/z range changes differently in each cycle, as an effect of *both* the environment and the recursive cycles. As well, the observed distributions across the m/z range are different, with some having a greater number of features spread across the bins.

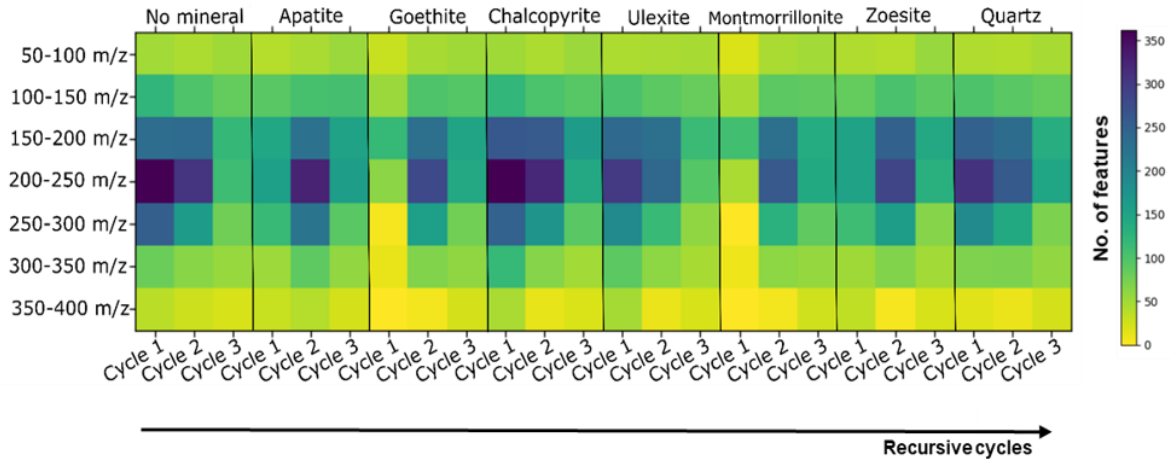


Figure 58. Feature distribution by m/z values over experiment cycle for the different mineral environments: A heatmap of the features generated by grouping the features into 50m/z bins, resulting in a unique pattern for each reaction mineral environment over the three recursive cycles.

2.2.3 The Long Cycle

The recursive cycles where carried out for a total of 3 cycles of 48 hours each and they were analysed as soon as the cycle was completed. For all mineral environments and the (non-mineral) control, the number of detected features in Cycle 3 was lower than Cycle 2. This is believed to be a consequence of the continuous cycling of material, which in turn truncated the combinatorial explosion by keeping the system out of equilibrium. However, the experimental controls for recursivity where not taken into consideration in the initial design of the experiment, since the main objective was to study the differences that arise from the presence of a mineral surface within a reaction cycling process. In this way, recursivity itself was a shared variable across all experiments. Nonetheless, we wanted to assess if the truncation was effectively caused by the process of reaction cycling and due to the absence of experimental controls, where the samples where not subjected to recursive samples, we decided to do a 'Long cycle'.

Therefore, we decided to let the samples run for a longer cycle (e.g. the long cycle), as a way to assess if the combinatorial explosion would be allowed to continue in the absence of recursive cycles. The long cycle was carried out by replenishing the reaction vessel with the same amount of starting materials after removing 2/3 of the reaction volume, exactly like with previous cycles, but letting the reaction continue for the length of three cycles (e.g. six days). The hypothesis being that in the absence of a reaction cycling process the number of

detected features would increase noticeably as the number of individual products generated continues into a thermodynamic equilibrium. This could suggest that the entropic level of the system increases in the absence of the recursive cycles, resulting in a larger product distribution. In agreement with this assumption, the number of detected features in the long cycle was significantly larger than for any of the recursive cycles, as seen in **Figure 59**. The trend for the long cycle was conserved across all mineral surfaces and the control reaction; giving a robust indication that the recursive cycles are indeed independent of the mineral environment present. However, just as observed with all the other recursive cycles, there is a unique pattern resulting in the features generated by each environment. Consequently, this supports the assumption that different mineral environments do effectively generate unique product distributions, regardless of the recursive action.

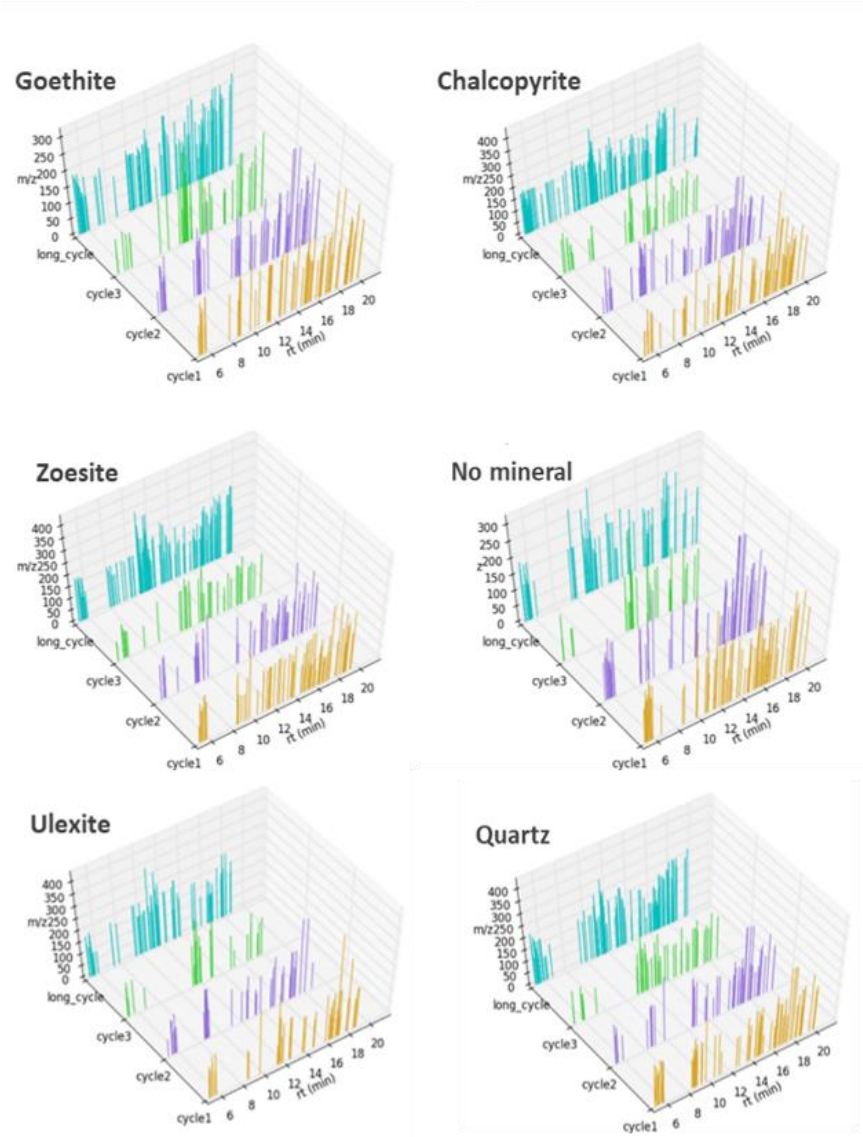


Figure 59. The Long cycle: Recursive cycles (1 to 3) result in a lower number of mass spectral features than the long cycle. Features are generated by the CompoundDiscoverer data processing software. Unique feature patterns can be seen for all mineral-containing cycles as an effect of the mineral environment.

The plots shown in **Figure 59** (above) were generated by extracting the detected features by the Compound Discoverer data processing workflow. For an example of the resulting feature list and their appended information, please see **Figure A14**. The trend seen by the Compound Discoverer features was conserved when the in-house feature generator was applied. In order to visualize the differences that arise by mineral type, while simultaneously enabling a direct comparison with previous results, the features were translated into scatter plots. This allowed us to take into consideration the number of MS^2 fragments in each feature and to observe the resulting feature patters for the products detected. In **Figure 60**, we can see how the scatter plots do show distinct patterns in the resulting feature space, as a consequence of the environment-type. As seen with the line-plots in **Figure 59**, the number of detected features is larger in the long cycle, than any other one. Also, a distinction across patterns can be seen for each mineral, reassuring a non-trivial influence of the environments in the reaction (even in the absence of recursive action).

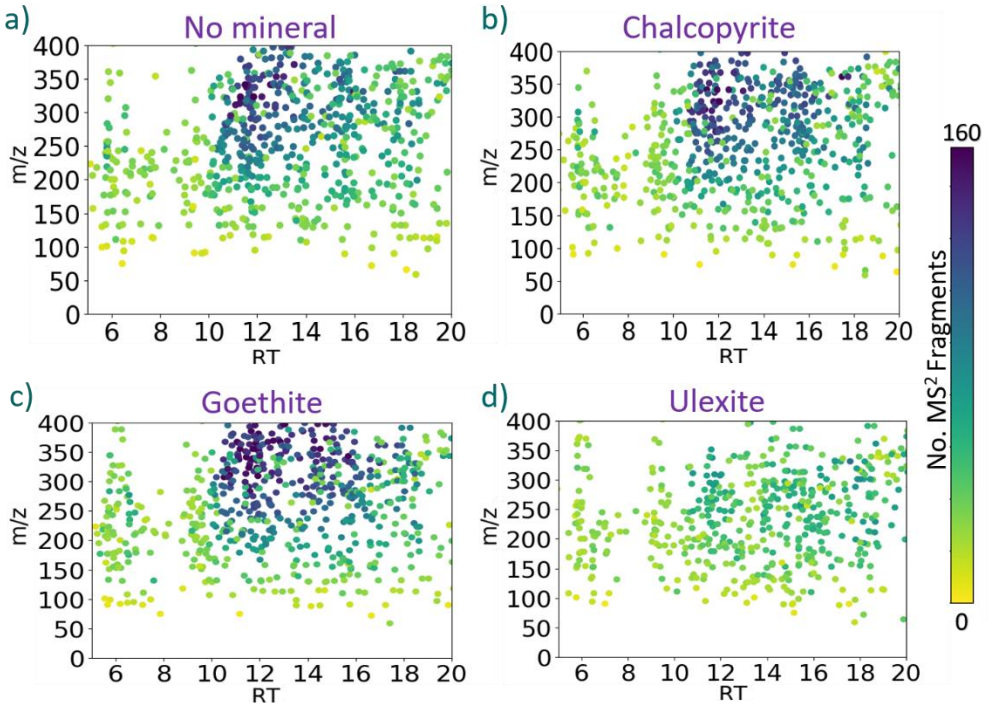


Figure 60. The Long cycle: Features are generated by the –in-house feature generator. Distinctive feature patterns can be observed for all mineral-containing cycles as an effect of the mineral environment (**b-d**).

2.2.4 Non-Recursive control

The results of the long cycle prompted us to repeat the experiment in a non-recursive manner. In order to do this, we carried out non-recursive controls by letting a reaction run for the equivalent time that it took to carry out the three recursive cycles (e.g. 3 cycles of 48 hours, for a total reaction time of 6 days). The non-recursive samples included a non-mineral control, as well as the same selection of mineral surfaces employed in the recursive experiment. The samples were then prepared and analysed following the procedure used for the previous dataset. Also, the data was also processed in the same way, as to enable a direct comparison of the experiments.

The non-recursive controls generated (in average) more features than in cycle 3 of the recursive samples, when comparing the detected features through their scatter-plots, see **Figure 61**. However, the total number of features found in the recursive control appears to be lower than those seen in the long cycle scatter plots (in **Section 2.2.3**). This is not entirely surprising, considering that the amount of material added to the reaction vessel for the recursive control was fixed. In contrast to the several additions of starting material that preceded the long-cycle in the recursive experiment.

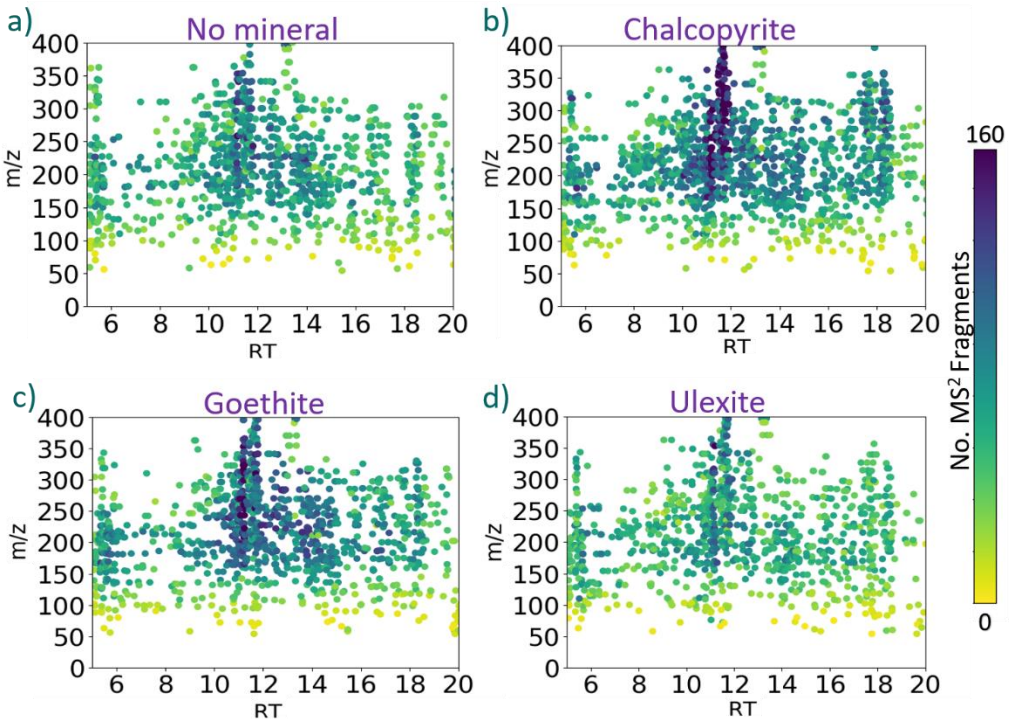


Figure 61. Non-recursive controls: The number of features detected is higher than those found for the third recursive cycle. The trend was conserved in the non-mineral reaction (**a**) and in all mineral environments, including chalcopyrite (**b**), goethite (**c**) and ulexite (**d**).

Furthermore, the patterns seen in the scatter-plots for the recursive controls, are unique for each mineral type. This correlates with previous results, where the influence of the environment effectively changes the resulting product distribution. In the **Figure 61**, we can see how the control reaction (**Figure 62a**) generates a different pattern, than when a mineral is present in the reaction vessel. Also, the results are consistent with previous patterns observed for the different environment types. For example, as seen for cycle 1 of the recursive reaction in the presence of the mineral ulexite, the number of features is lower than those detected for goethite, chalcopyrite and other minerals (with the exception of quartz, which had a similar number of feature to ulexite). The consistency across the distribution of features for each environment, in the recursive and non-recursive experiment, indicates that the effect of the environment in the reaction dynamics is not trivial and mostly reproducible. Nonetheless, experimental repeats would need to be considered to assess the magnitude of the variations in the product distribution as an effect of mineral types. Particularly, when considering the intrinsic variability within the chemical composition of a given mineral, which is embedded in their mechanisms of formation and cannot be avoided. At the very least, these results confirm that the patterns arising for each environment are relatively consistent.

2.2.5 Reproducibility Assessment

In order to address if the trend could be observed across more recursive cycles and assess the reproducibility of experimental repeats of the reactions, we have repeated the experiment with five cycles. For an experimental plan, see **Figure 54**. Given that the main objective was to address reproducibility across experimental repeats, only one of the mineral surfaces was selected for the study. The selected mineral surface was chalcopyrite and goes along the lines of the figures presented above. All the appropriate controls were conducted: (*a*) non – recursive mineral control, (*b*) non-recursive reaction, (*c*) recursive reaction and (*d*) recursive mineral reaction.

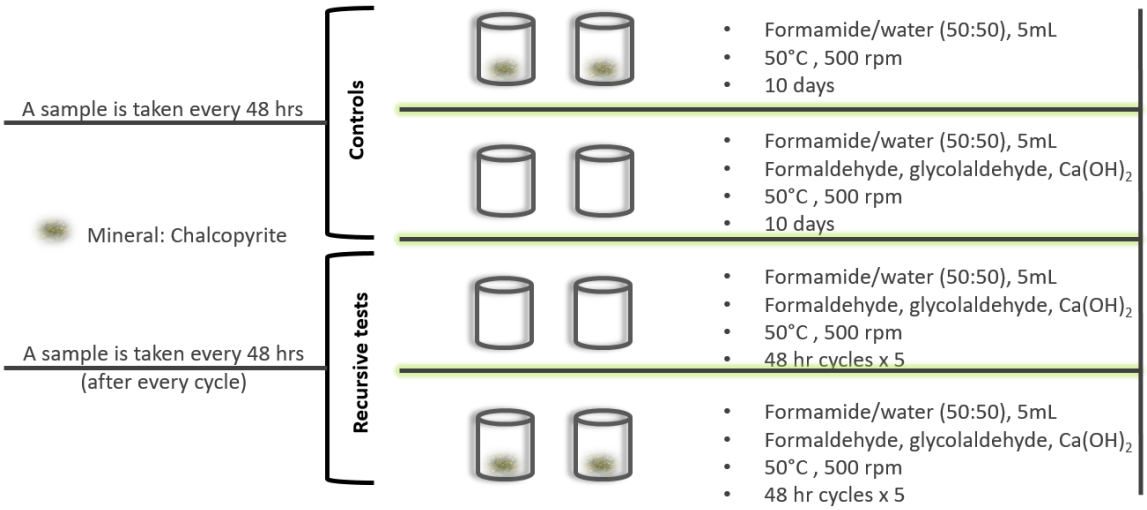


Figure 62. Experimental repeats for the Recursive Reaction as described in **Section 4.2.3**.

The feature generation was carried out in the same way as with the previous experiments (Cycle 1 to Cycle 3), but the number of overall features detected was reduced, **see Figure 55**. This is believed to be due to poor instrument maintenance, which would be expected to reduce the sensitivity, unable to detect the compounds in a lower concentration range. Unfortunately, samples where compromised and re-acquisition was not possible. However, we could assume that the instrumental error would be propagated across all the samples in the instrument run, and therefore a comparison across between the experimental replicates is possible. This would allow us to see how reproducible the experimental repeats are in themselves, even if we cannot directly compare the resulting trend with the previous experiments (e.g. assess how the features change over recursive cycles confidently).

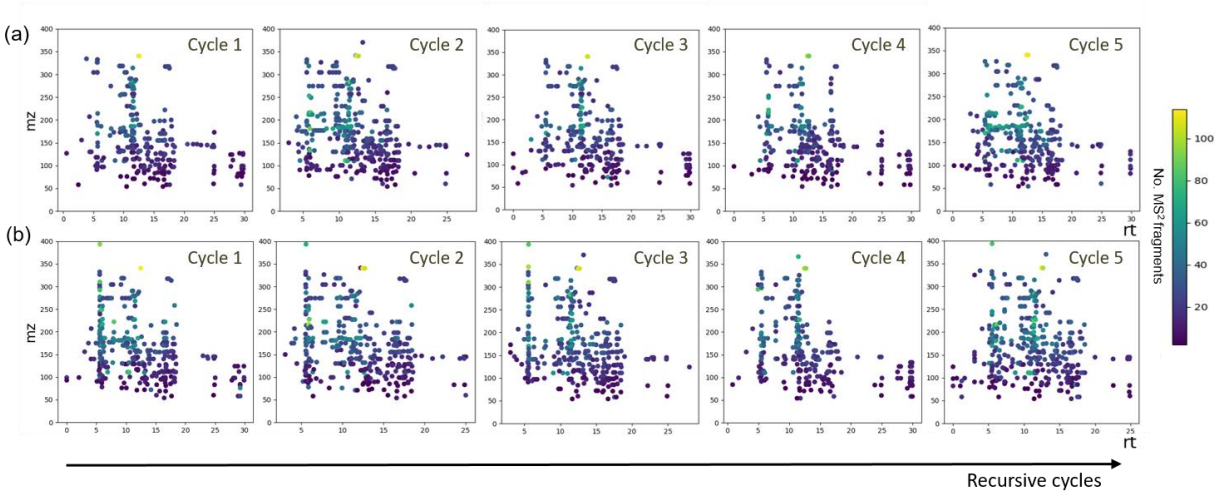


Figure 63. Experimental repeats for the Recursive Reaction: A non-mineral recursive reaction (**a**) is compared to a recursive reaction in the presence of chalcopyrite (**b**).

The features from the reproducibility experiment were also grouped into 50m/z bins, in the range of 50 to 400 m/z (as done previously for cycle 1 to cycle 3), see **Figure 64**. The resulting trend indicated variations across cycles for the non-recursive reactions, as well as, for the recursive reactions (in the presence and absence of a mineral surface). The non-recursive reaction (*Reaction Control*) has a significant increase in the number of features across the 100-200 m/z range for cycle 5, diverging from the recursive reactions quite significantly. Due to the inconsistencies in the overall number of features detected, we do not compare the resulting trend of experimental replicates (Cycle 1 to Cycle 3 vs Cycle 1 to Cycle 5) directly, but are confident that only minor variations within the system (and their product distributions) can be observed across the experimental repeats.

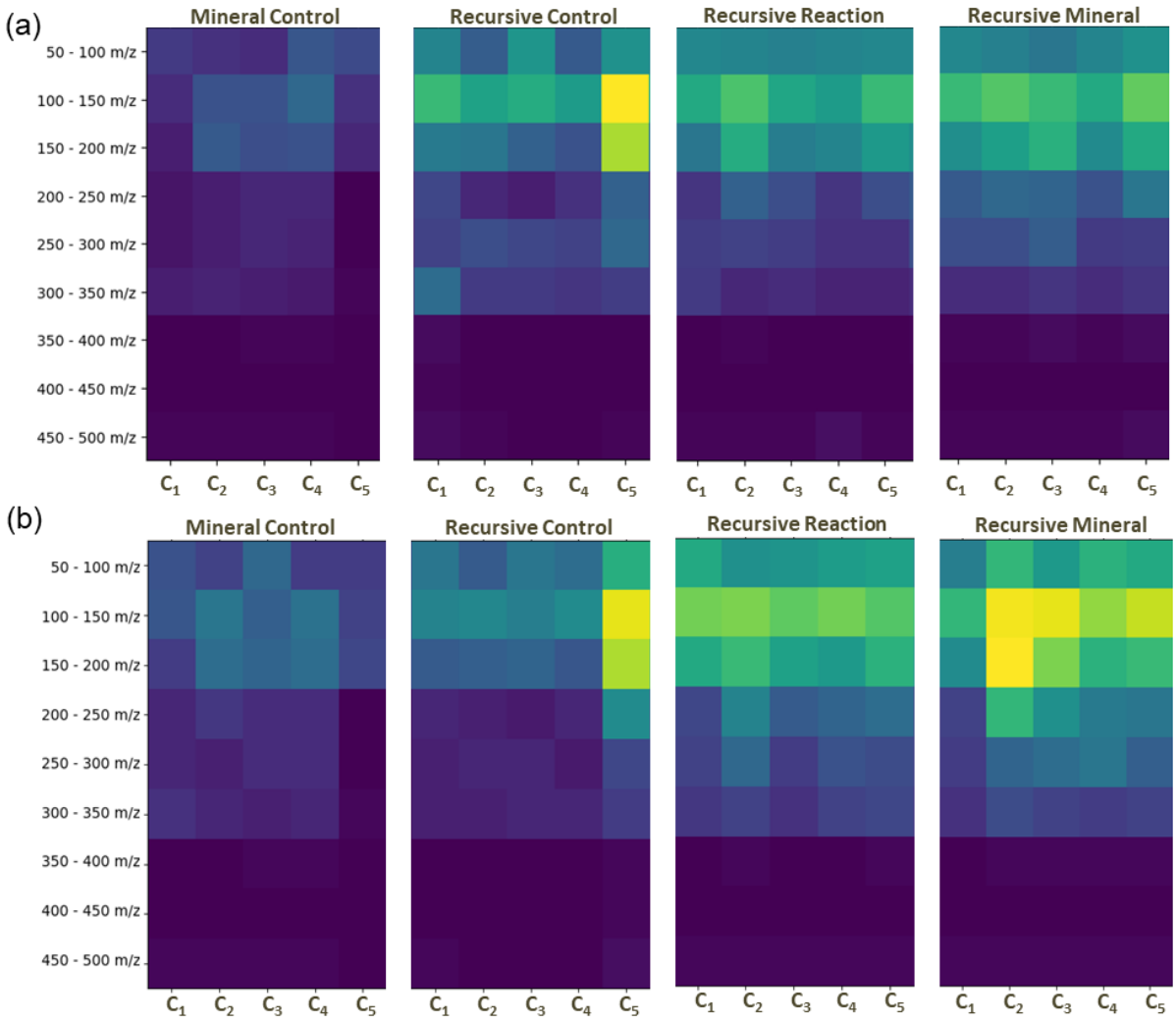


Figure 64. Molecular weight distribution of the features detected in the Cycle 1 to Cycle 5 experiment for the mineral control (non-recursive), recursive control (non-recursive reaction), recursive reaction (recursive reaction) and recursive mineral (recursive reaction in the presence of chalcopyrite). Experimental duplicates are presented as (**a**) and (**b**).

Furthermore, if we observe the variations across the detected products presented in the figure above we can appreciate a larger deviation across experimental repeats from the recursive mineral reaction, when compared to any other reaction (including the recursive reaction). This is noted, since there is the possibility that minor compositional differences in the minerals could promote this effect, which is consequently amplified over recursive cycles. The mineral surface will have differences based on the stochasticity behind the crystal formation of the mineral. As mentioned in **Section 1.2.3**, the possibility of a mineral surface being capable of selecting and therefore differentiating product distributions due to the intrinsic differences in the mineral surfaces, is believed to be one of the first mechanisms that allows the system to retain environmental information.

Regardless of the known differences and uncertainty within the datasets, we decided to compare the trends between the reaction controls and the recursive mineral reactions. To do this, we plotted the number of detected features, presented as MS^1 fragments in **Figure 57**. The recursive reaction has a significantly different number of detected fragments when compared to the recursive control. The control reaction shows the number of detected features exploding after cycle 4, resulting in a higher number of features than any of the recursive reactions.

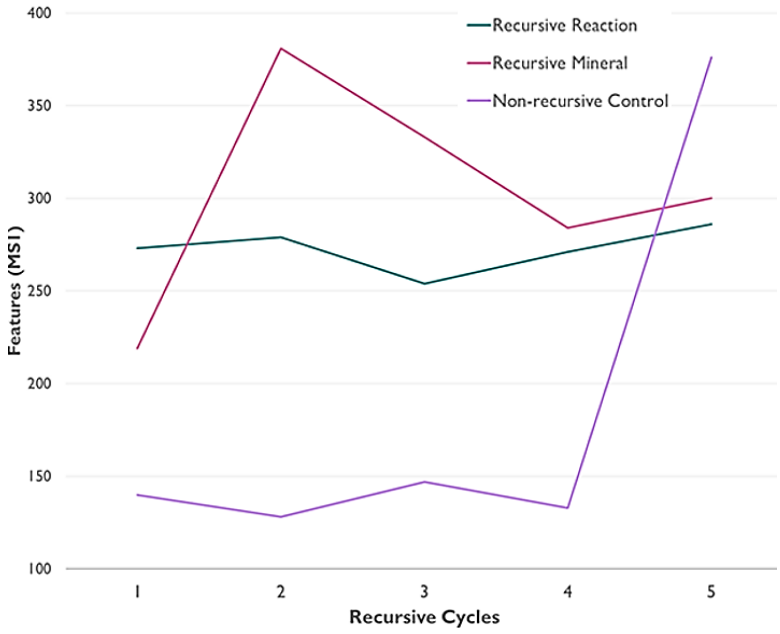


Figure 65. *Detected features for the experimental replicates:* Number of (MS^1) features from cycle 5 cycles (Non-recursive reaction, *purple* - Recursive reaction, *green* - Recursive mineral, *pink*), in a range of 100 to 400 features detected across cycles.

2.2.6 Comparison test: Formose reaction and Formamide condensation on minerals

The model-system for complex mixtures is based on two well-known combinatorial explosions: The formose reaction and formamide condensation (**Figure 46**). Each of these reactions result in a highly convoluted chemical mixture and a complete characterization of their reaction products has not been accomplished to date. The analytical intractability of these two reactions are what made them strong candidates for the model system, alongside with their 'prebiotic' significance. The reactions proceed via two main reagents; formaldehyde and formamide, which have been identified as main intermediates in the formation of organic compounds from spark discharge experiments, such as the Miller-Urey experiment. The broad product distribution arising from these reactions has been a matter of discussion in many prebiotic studies, where the need for selective mechanism is highlighted. For this reason, a variety of mineral inclusions as a mean to promote environmental selection have been tested. As mentioned in **Section 1.2.3**, the inclusion of mineral surfaces in either reaction resulted in a selective truncation of the combinatorial explosion, by means of many different mechanisms besides catalysis, including the generation mineral-organic complexes that stabilize certain products over others. However, how do any of the selective processes came to be remains a largely a mystery.

The effect of mineral inclusions in these reactions has been studied extensively, but how does the product distribution of these systems look after the inclusion of a mineral surface compared to the reaction on its own, as not been assessed in an untargeted fashion. The analysis for these type of reactions systems, as with all other prebiotic mixtures, has been tailored to see changes in the 'prebiotically' relevant products. For example, the formose reaction which was found to stabilize ribose over any other sugar in the presence of a borate mineral, was analysed by GC-MS analysis following a targeted derivatisation reaction. Therefore, in order to effectively compare the resulting product distribution for the single reactions with that of the formose-formamide system, untargeted analysis of each reaction must be carried out.

The synthesis of the formose reaction and the formose condensation was executed, alongside with the model system reactions, as a way to evaluate the differences that arise from combining the two systems. This allowed us to address two main questions: (a) is the resulting product distribution of the formose-in-formamide different than the one that arises from the reactions in isolation? and (b) are there any observable differences arising from the untargeted analysis of the single reaction in the presence of different mineral surfaces?

The untargeted analysis of all reactions was done with the exact same UPLC-MS/MS methods. This allowed us to extract and generate mass-spectral features, while appending their MS/MS information, also with the same processing scripts as used in **Section 2.2.2.2**. The processed data-sets were then plotted in the same fashion as in **Section 2.2.2.3**, to allow for a direct comparison of the results. As expected, we found that resulting product distribution for the formose rection and the formamide consendation are visually different to each other, and to the model system (see **Figure 66**). This was true for the control reaction (e.g. no mineral) as well as, with all mineral surfaces employed. The MS/MS information allowed us to identify where the features with the larger number of MS^2 fragments arise and map out, to some extent, differences in the molecular complexity of the unidentified products. Assuming that the higher number of MS^2 fragments, the more complex is the MS/MS resulting spectra, and therefore the molecular complexity corresponding to a given feature. The position of the features with a higher number of MS^2 fragments also changes for each reaction type. In the case of the formose reaction, we can observe how most of the features with more than 80 MS^2 fragments have an m/z value higher that 200 m/z. On the other hand, the formamide condensation appears to produce the features with the highest number of MS^2 fragments in the mz range of 150- 250 m/z.

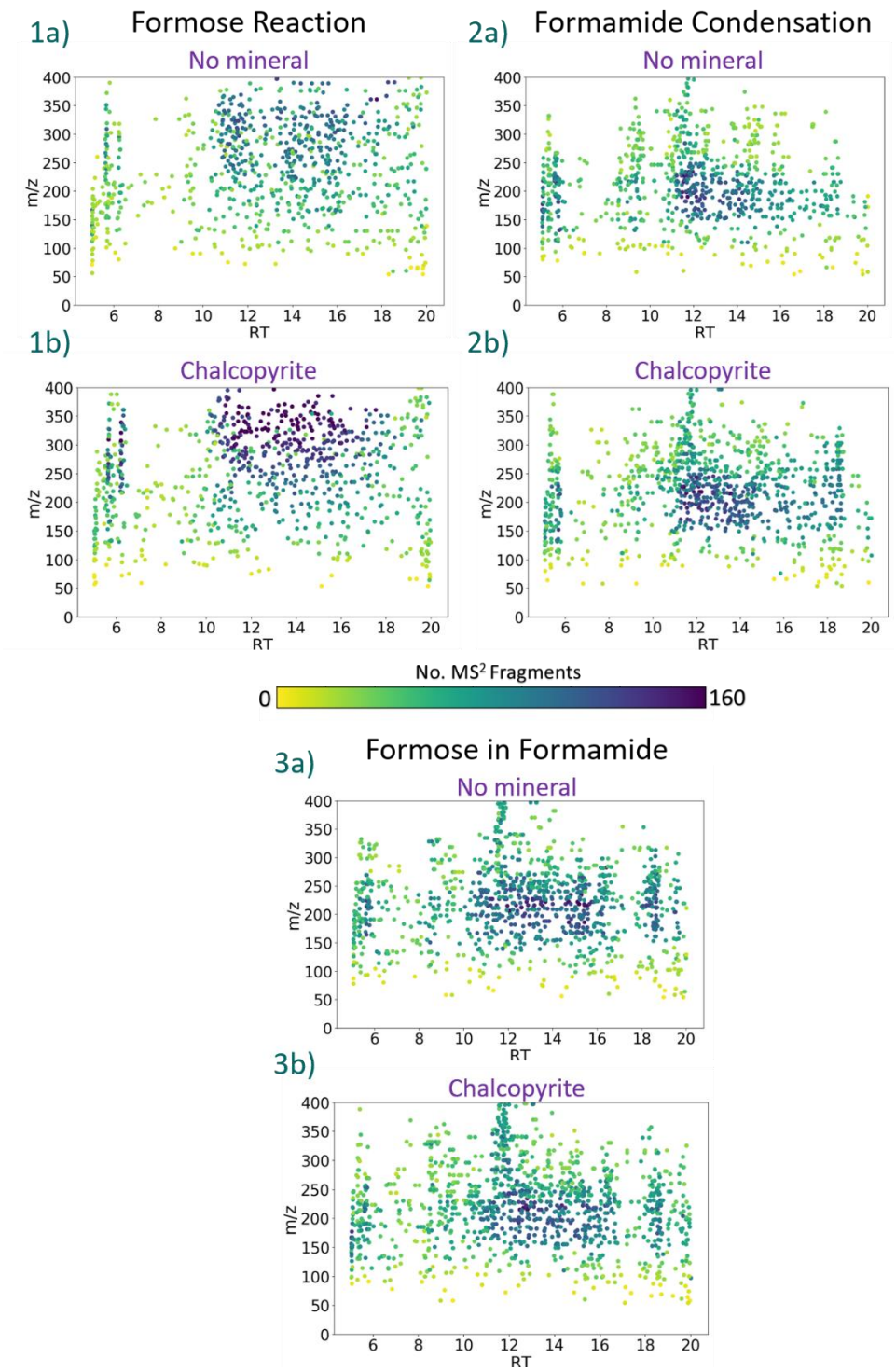


Figure 66. Distribution of the features detected for the formose Reaction (**1a**), formamide condensation (**2a**) and the formose-formamide (**3a**), in the non-mineral reaction. Compared to the reactions in the presence of a mineral surface, chalcopyrite (**1b- 3b**).

The formose reaction is conventionally carried out for shorter periods of time than the reaction time we selected for the dual-system. This was done in order to allow the formamide condensation to have a reaction time of at least 24 hours, as with most of the previously reported experiments. However, it is known that if the formose reaction is carried out for longer than 6 hours at temperatures comparable to ones used in the model-reaction, it will

produce tar.[215] The plotted features in **Figure 67**, clearly display how most of the material with a large number of MS^2 fragments corresponds to the higher m/z region. This can be suggestive of the tarry material being present, bearing in mind that polymerized sugar moieties would produce a large number of MS/MS fragments, compared to monomeric sugars. Furthermore, considering that the elution profile of the UPLC method is conserved, we should see oligosaccharides towards the end of the chromatographic run. Furthermore, it must be noted that in the formose reaction with ulexite, the selected mineral surface to represent borate minerals, appears to produce less features (e.g. reaction products) than any other surface or the non-mineral (control) reaction. Taking into consideration the previously reported result on the stabilization of ribose by this mineral surface, we could assume some type of stabilization effect that transcends through the overall product distribution, as we can observe by the untargeted analysis of the mixture.

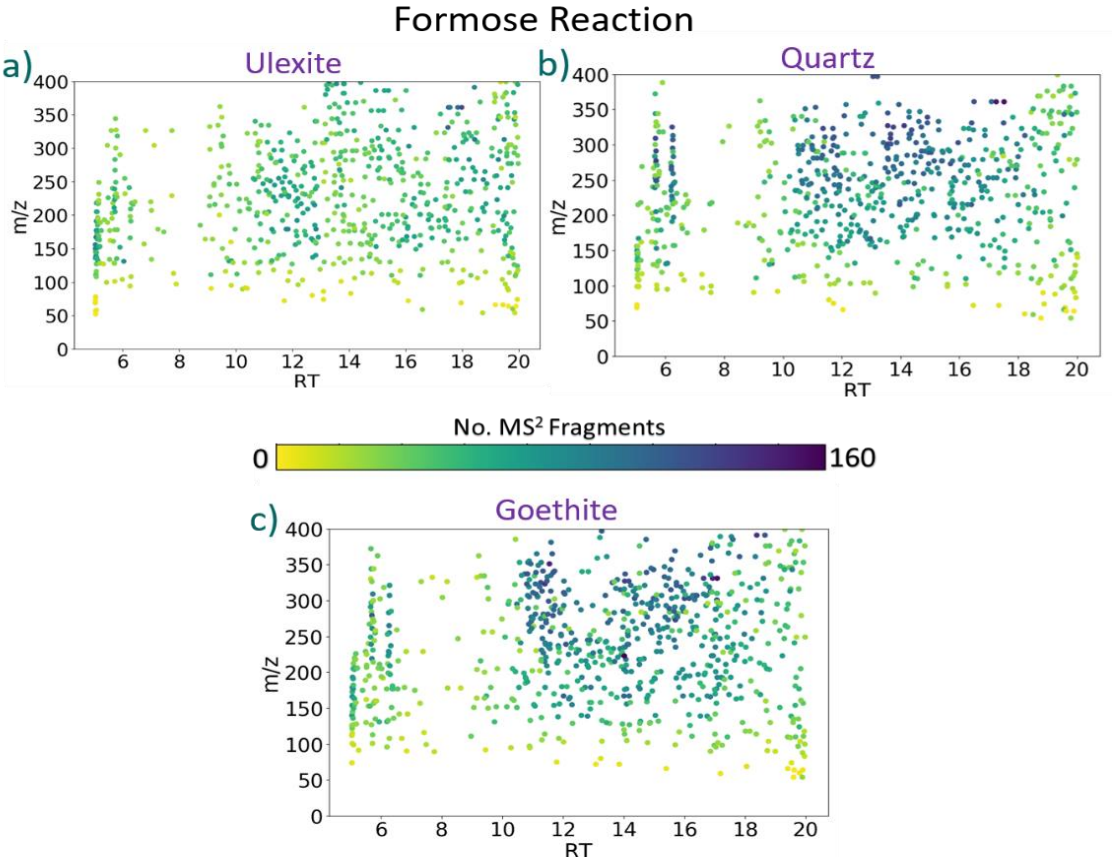


Figure 67. Formose reaction in a selection of environments: Mass-spectral features for the resulting product distribution in the presence of ulexite, quartz or goethite mineral surfaces.

The formamide condensation is usually carried out at much higher temperatures than the conditions selected for the combined system. This was done in order to prevent the degradation of the sugar products arising from the formose condensation. This was a compromise we took assuming that over time, the overall nucleobase formation would be achieved. To our surprise, the nucleobases where detected from the first cycle, meaning that the conventionally targeted products can be formed under milder conditions than originally thought, even in the absence of any mineral catalyst (surface). The feature pattern that arises from the presence of a mineral surface is visually different for each environment employed, see **Figure 68** for a comparison between ulexite, quartz and goethite. We can also observe how the number of features detected is different for each environment (and sample type), with the region corresponding to the 7 to 10 minutes (Retention Time) being of particular interest, since it displays the most variability in the amount features.

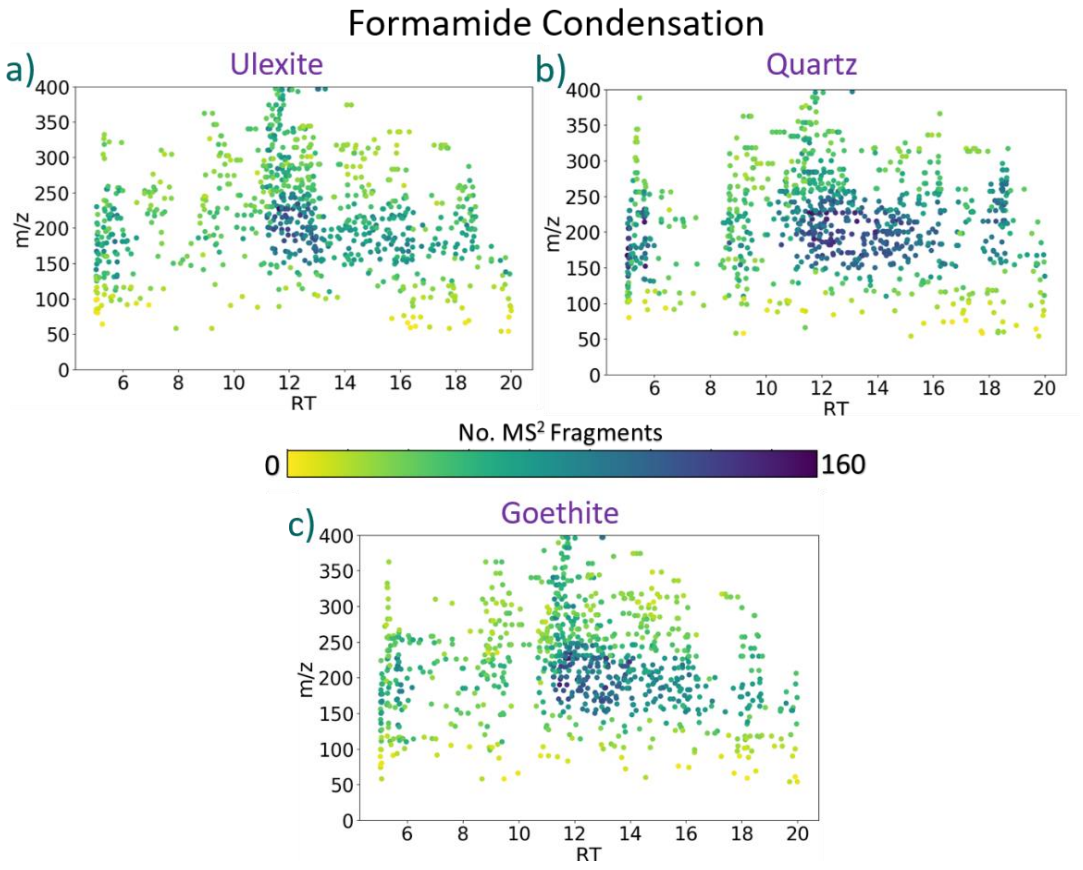


Figure 68. Formamide condensation in a selection of environments: Mass-spectral features for the resulting product distribution in the presence of ulexite, quartz or goethite mineral surfaces.

2.2.7 Feature Analysis

The features generated can visually be appreciated to be different, but in order to validate these differences, they have been analysed through PCA. This type of systems approach to data analysis of the experimental data was employed in the Miller-Urey 'deuterium world' experiment (as seen in **Section 2.1.5**), resulting in an efficient distinction across data-sets. Therefore, we aimed to carry out this type of statistical analysis on the detected features by UPLC-MS/MS. However, there are differences on the way the PCA was carried out for the untargeted data. For example, the dimensionality of the data-sets was reduced in a different way than in **Section 2.1.5**, as we thought it was more representative of the 'feature' approach taken for the data processing described in **Section 2.2.2**.

Furthermore, we generated van Krevelen diagrams on the identified fraction by the CompoundDiscoverer data processing workflow. This was done as a way to assess compositional differences in the chemical speciation of the product distribution, which can arise as an effect of the recursive cycling and the presence of a mineral environment. Also, acting as a secondary validation of different compositions within the resulting product space, as observed when the MS/MS fragments were counted and resulted in different numbers of MS^2 fragments for each feature.

2.2.7.1 PCA

The mass spectral features where subjected to Principal Component Analysis (PCA) as a way to validate the differences in the feature distribution. Initially, the online software for metabolomics untargeted analysis, XCMS online by Scripps institute, was used to analyse the samples and generate the PCA plots.[216] The data processing within the software, encloses retention time alignment and mass-spectral deconvolution methods, which are tuned by specifying the type of instrument in which the data was acquired. In our case, we chose the settings for UPLC-Orbitrap, which fits perfectly for the selected analytical method. The samples are then analysed within the XCMS software, by generating multiple features and appending their mass-spectral information, such as intensity, m/z and retention time (calculated as their average across the full set). For an example, the features detected by XCMS, please see **Table A2**. This feature list is integrated to an interactive PCA, which allows for multiple types of data scaling and centering. By applying unit-variance scaling of the datasets, we could generate a PCA plot that effectively spaces out the samples by cycle.

Therefore, these methods could indicate differences across the data sets based on the statistical analysis of the features. See **Figure 69** below.

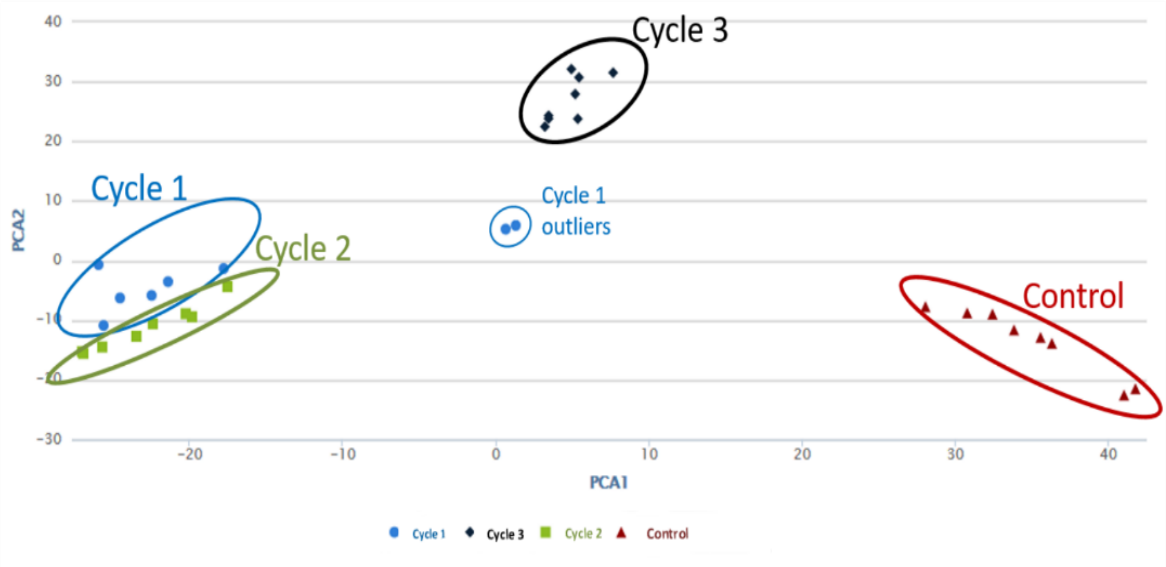


Figure 69. Simple PCA scores plot of UPLC-MS/MS mass spectral features of the recursive reactions by XCMS online data processing software.

However, there are limitations of using a software designed for the analysis of biological samples, such as the extraction of information from the calculated features. Therefore, we decided to construct our own data processing scripts to generate mass-spectral features from the raw data, as described in **Section 2.2.2.3**. By manually going through the data and looking at the resulting scatter plots for the features, we could visually appreciate a trend across the recursive cycles and sharp differences across the patterns of each mineral environment. Nonetheless, if we wanted to assess differences without having to manually look at each resulting plot, we would need to another way to observe multiple data sets from a single figure. This prompted us to assess the statistical differences of the data sets by applying PCA directly on the generated mass-spectral features.

In order to do this in a way that is representative on how we visualized the differences across the samples, the generated features were distributed into two types of bins: 400 bins for the m/z values (1 m/z bins, for a range of 0 - 400 m/z values) and 20 bins for the retention time (1 minute bins, for a range 0-20 minutes in retention time). In **Figure 70**, we can observe how these bins construct an array of ‘cells’ (or grids), each containing a particular number of features. The number of features in each cell is counted to fill a 400 x 20 matrix. This is one of many ways to describe the features and somewhat closer to the approach taken into image processing techniques.

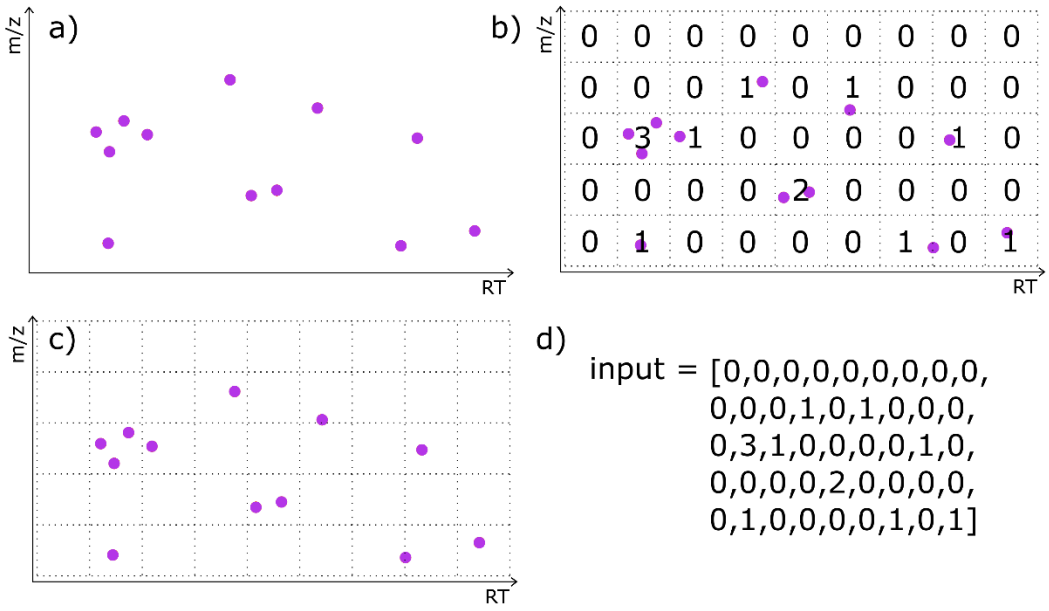


Figure 70. Example on how the data was introduced to the PCA: (**a**) the features are generated by a combination of m/z and retention time (RT), (**b**) then the number of features in each resulting cell is counted, (**c**) which are produced by the m/z and RT bins producing a grid (of cells). Finally, (**d**) the number of features in each cell is converted into a 400 X 20 matrix.

The PCA of the recursive samples resulted in unique groups of the datasets by cycle number, as shown in **Figure 71**. The results are comparable to those obtained by the XCMS software, particularly in the way the 'grouping' is distributed. Therefore, we could assume that the feature representation approach taken to generate the PCA's was successful. The distance between the points in each cycle 'group', varies across the cycle number and control reaction. For example, in cycle 3 the samples are closer to each other than in any other cycle or the controls. The grouping section is expanded in Cycle 1 and the control samples. This indicates larger differences in the frequency of the features detected for the control and cycle 1 samples. Also, the outliers in cycle 1 can be explained by the initial difference in the detected compounds, which was observed to be significantly lower for those minerals by visualization through the scatter plots. This is believed to be a result of the interaction of the minerals and highlights the influence of the environment in the initial product distribution of the formose-formamide reaction. Furthermore, in **Figure 72** the PCA displays grouping of the reproducibility assessment samples (i.e. **Section 2.2.5**) based on the recursivity variable. The non-recursive samples separate from the recursive samples, resulting in two (observable) groups.

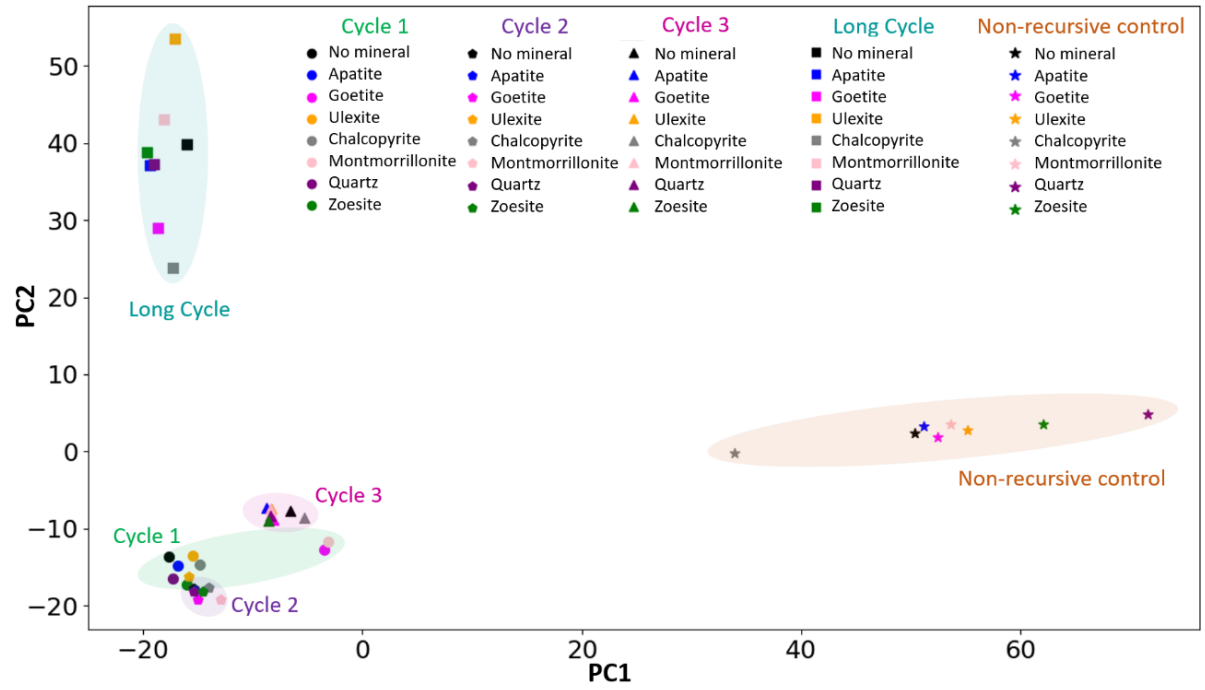


Figure 71. Simple PCA scores plot of UPLC-MS/MS mass spectral features of the recursive reactions: Cycle 1 (*green*), Cycle 2 (*purple*), Cycle 3 (*pink*), the long cycle (*blue*) and the non-recursive controls (*orange*). Generated by the scktlearn package in python.

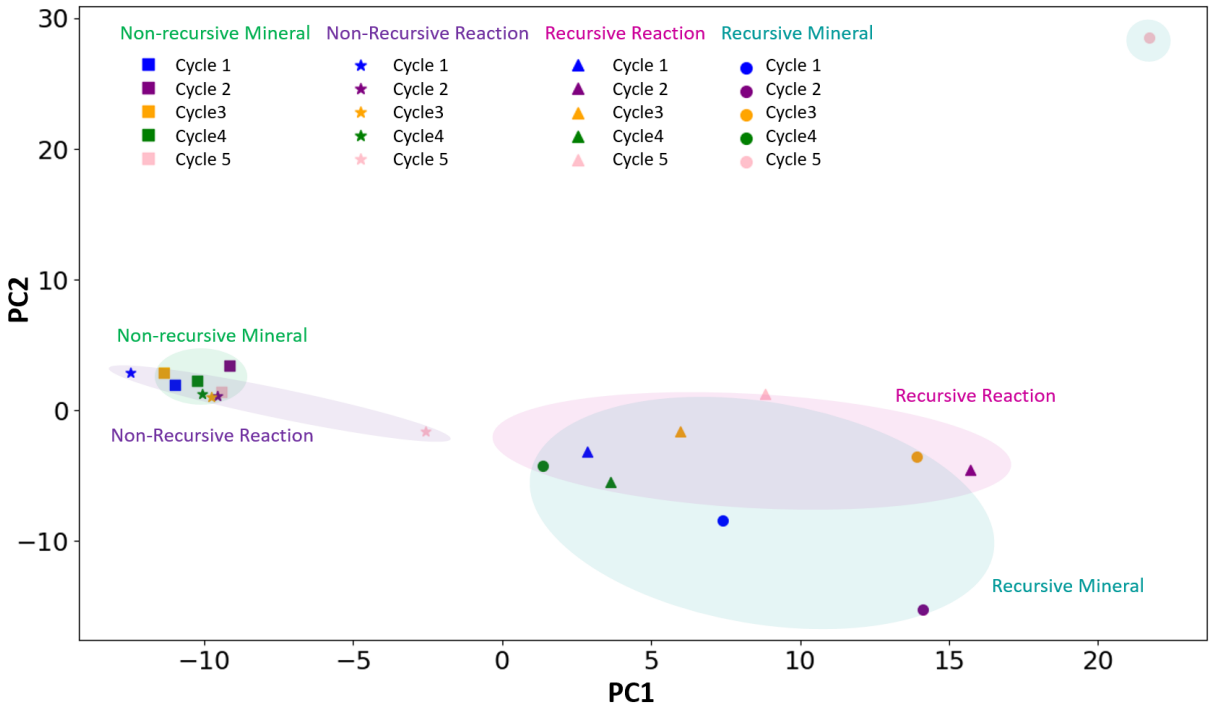


Figure 72. Simple PCA scores plot of UPLC-MS/MS mass spectral features of the recursive reactions in the reproducibility assessment (e.g. experimental repeat with 5 cycles): Non-recursive mineral reaction (*green*), Non-recursive reaction (*purple*), Recursive reaction (*pink*) and the Recursive mineral reaction (*blue*). Generated by the scktlearn package in python.

The PCA results suggest that this approach takes into consideration the number of features detected primordially. This is desirable, since the selected reaction is expected to be a combinatorial explosion and a measure for complexity would be a truncation of the number of individual products, which can be translated to some selection mechanism taking place. The approach can also be applied to the comparison test within the formose reaction, formamide condensation, and the selected hybrid system. The same method for feature representation was used to generate the PCA's shown in **Figure 73**. The results for these datasets display a strong differentiation across the reaction types. The formose reaction displays the largest separation within the 'group', arising from the differences across the mineral surface. On the other hand, the formamide condensation resulted in the smallest 'grouping' or 'cluster', suggesting that the variation of the resulting product distribution as an effect of the mineral environment is less significant in this type of system. The formose-formamide system, being a combination of the two reactions, results in a cluster of medium size.

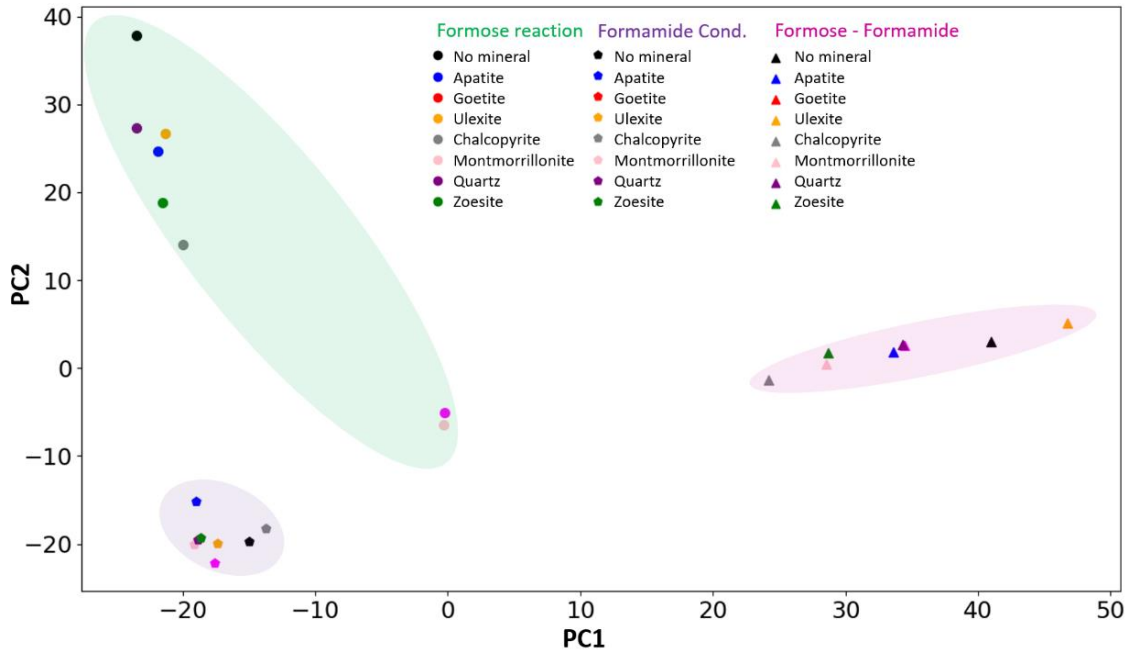


Figure 73. Simple PCA scores plot of UPLC-MS mass spectral data of the single reactions (i.e. Formamide consensation, *purple* and formose reaction, *green*) and the dual 'Formose in Formamide' system, *pink*.

2.2.7.2 van Krevelen diagrams

The van Krevelen diagrams are graphical plots developed by Dirk Willem van Krevelen (chemist and professor of fuel technology) and typically used to assess the origin and maturity of kerogen and petroleum.[217] The diagram cross-plots the hydrogen:carbon atomic ratio as a function of the oxygen:carbon atomic ratio. This visualization method for complex mixtures as been used before by the analysis of known 'prebiotic-type' mixtures, as well as, the Murchinson meteorite. A convensional use of this plots results from the analysis by FT-ICR data. However, through the application of high resolution mass specteometry and data processing methods, it is possible to get accurate chemical formula from the MS data. This allowed us to use this technique as a way to assess the differences in molecular diversity of the resulting product distribution, as an effect of the recursive cycles and mineral environment. The data selected for the van Krevelen plots corresponds to only a subset (~20%) of the features detected by the Compound Discoverer® workflow described in **Section 4.2.4.5**, but it can still highlight differences in the chemical distribution of the data-sets. Therefore, we used the chemical formula calculated in the automated (Compound Discoverer) workflow by the composition prediction node. The plots were generated in Python with the matplotlib library, after manual extraction of the features with annotated chemical formula. This resulted in a series of van Krevelen plots, seen in **Figure 66**, for Cycle 1, Cycle 2, Cycle 3, Long Cycle and Control samples. The number of points (e.g. features) present in the plots are not representative of the total number of features detected for the samples, since the fraction of the features that had an automated assignment of their chemical formula was not equal for all the cases. Nonetheless, this approach displayed compositional differences within the features, across *both* the recursive cycles and mineral environments.

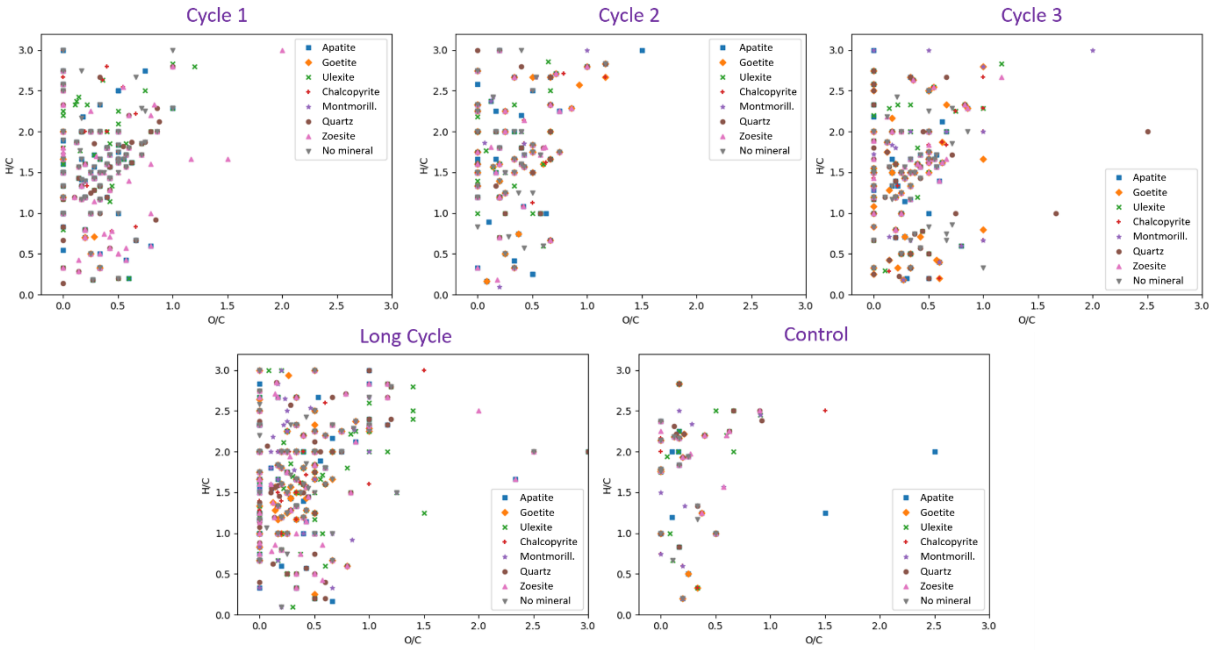


Figure 74. van Krevelen diagrams for the recursive cycles, the long cycle and the control samples. They represent a subset of the detected features which had an annotated chemical formula from the Compound Discoverer® processing workflow. Differences in the composition of the features can be observed for the mineral environments and the recursive action.

2.2.8 Section Summary

In summary, we carried out the formose reaction and formamide condensation in a one-pot fashion, under milder conditions than previously reported,[215] while a recursive environment was applied to the resulting mixture in a series of cycles. We found that recursive cycles not only truncated the combinatorial explosion by reducing the number of individual products, but also successfully generated sugars and nucleobases from potentially prebiotic routes, in an integrated fashion. Traces of nucleoside formation were also detected after two recursive cycles, for the first time in this simple-precursor systems (e. g. Formose reaction/ Formamide condensation). Also, we found a molecule with a strong connection to prebiotically-relevant compounds, hexamethylenetetramine (HMT), which might have a non-trivial relationship with the formation of these building blocks.

The untargeted analysis of the mixtures allowed for an unprecedented exploration of the chemical space generated in analytically intractable (prebiotic) combinatorial explosions. Furthermore, we were able to observe significant differences across the data-sets by the development of an in-house feature generator, which allowed us to represent the compounds in solution without having to identify them, an approach only seen before in ‘discovery’-

type workflows within the 'omics' field.[218] The resulting features were plotted in order to observe their distribution and how they changed as an effect of the recursive cycles *and* the mineral environments. Through this approach, we could see distinctive patterns arising from the different mineral environments. As well as, how these distributions changed as a consequence of the recursive action. We found that the number of features detected reduces as the recursive cycles take place, in all the mineral environments tested *and* the non-mineral reaction. Therefore, concluding that the recursive action does have a notable influence in the product distribution of the formose-in-formamide reaction.

Moreover, we conducted a series of control tests: a long-cycle, a non-recursive control, a reproducibility assessment, and a comparison with the 'classic' formose reaction or formamide condensation. The number of detected features for the 6 day "long cycle" is higher than those for cycle 3, indicating that in the absence of 2 day recursive cycles, the number of resulting products increases. This can be seen as a validation of the influence of the recursive action in the experimental outcome. Also, the same trend was observed for the non-recursive controls. In the absence of the recursive cycles, all the non-recursive controls (even in the absence of a mineral environment) resulted in a larger number of features than when the recursive action was employed. On the other hand, the reproducibility assessment was accomplished by executing experimental replicates of the recursive reaction. As well as, carrying out all the appropriate controls (all at once), which included a non-recursive mineral reaction and a non-recursive mineral reaction. The experimental duplicates displayed no significant variability amongst themselves, with the largest discrepancies seen for the recursive reaction in the presence of a mineral surface. However, due to problems with the instrumentation, a trend for the recursive variable could not be assessed.

Furthermore, a comparison test with the formose reaction and formamide condensation, resulted in clear distinctions across the detected features for each reaction-type. Likewise, significant differences can be seen when comparing them to the formose in formamide reaction, displaying a divergent pattern when the two-have been combined. In the presence of a mineral environment, all three reactions studied exhibited unique feature distributions for each mineral-type employed. These results indicate that the formose-in formamide reaction has a different product distribution than the reactions in isolation, along with distinct variations when done in the presence of a particular environment.

In addition, the generated features were analysed further (e.g. not by eye) with Principal Component Analysis. This provided us with a way to visualize the differences in one plot, opposed to multiple scatter-plots. The results from the PCA indicate that the recursive cycles are different than the non-recursive experiments, as well as, due to the presence of a mineral environment. Furthermore, we generated van Krevelen diagrams based on the features detected by the Compound Discover processing workflow that had an annotated chemical formula. These features correspond to only a subset of the total features, but were enough to assess compositional differences across the recursive cycles. The diagrams also display different chemical compositions for the different mineral environments, reassuring that the distinct variations observed for the distribution of the features, also translate to compositional differences in the product distribution of the reactions.

Through the untargeted approach, we developed a way to look for distinguishing aspects across experimental variables, without targeting any particular compound. In this section, we have discussed how the relevance of certain features over others, particularly in complex product distributions, is not inconsequential and it can benefit from a discovery-driven investigation; such as the one used in metabolomics-type workflows and this work. This approach generates a more robust overview of the highly complex product distribution generated in analytically intractable mixtures, as a mean to further our understanding of complex chemical systems and their intrinsic reproducibility. Due to the convoluted nature of combinatorial explosions, a satisfactory reproducibility assessment would need a large number of experimental replicates, where a high-throughput experimental design is required. While this is not assessed directly in this work, it has indeed enabled the possibility for such studies, in which a comprehensive overview of the resulting products would be substantial in order to draw any meaningful conclusions.

We believe that recursive experiments bring us one step closer to a reasonable ‘real-life’ scenario and combined with this analytical approach, it provides an improved experimental regime for looking at the evolution of complex mixtures from simple precursors under non-equilibrium conditions.

2.3 The Recursive Miller-Urey experiment

The Miller-Urey experiment is known as the landmark experiment in the evolution of the Origins of life – Prebiotic Chemistry field. Its significance remains to this day, as many researchers across the globe continue to replicate the experiment under different atmospheric conditions (i.e. other gas compositions than those used to replicate a reducing atmosphere, as a consequence of the ongoing dispute on early Earth's conditions and whether they were strongly reducing, neutral or oxidizing, by the time the first living entity came to be). However, not many different versions of the experiment have been carried out so far. This is to say, besides variations within the energy source used in the experiment or the gas mixture employed, there has been limited investigation on other parameters. Consequently, this prompted us to investigate the effect of reaction cycling in the Miller-Urey reaction. If we take into consideration current prebiotic earth models, then some of the abiotically generated material would have been deposited on the surface and subsequently replenished, as the recycling process of the water cycle develops. The effect of natural processes such as atmospheric cycling, is an important but not yet addressed variable within the prebiotic broths framework.[219] Therefore, we decided to investigate what effect could this have in the overall product distribution of the famous experiment.

2.3.1 Recursive cycles

The Miller-Urey experiment was carried out in a recursive manner as a way to assess the effect of reaction cycling in the products generated by the spark-discharge experiment. The recursive cycles are executed in a similar fashion to the method described in **Section 2.2.1**. After an experiment is completed (see **Section 4.3.2** for a complete description of the experimental procedure), about 70% of the resulting solution was removed from the reaction vessel (e.g. round bottom flask) and replenished with fresh water. The system was then (as per usual) degassed and filled with a gas mixture of methane, hydrogen and ammonia, before igniting the spark discharge through two tungsten electrodes. This process was repeated five times, for the same duration (7days) through all the recursive experiments (**Figure 75**).

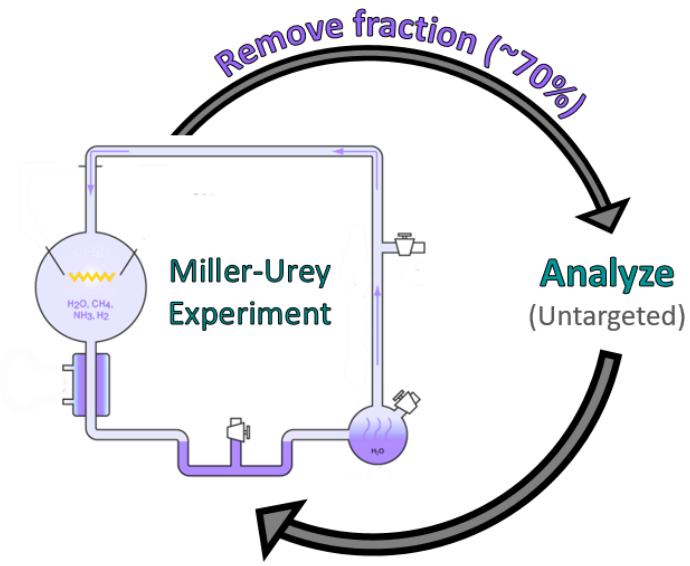


Figure 75. Recursive cycles: After each reaction, the supernatant is removed for analysis and a small fraction is left in the reaction vessel and used to seed a next reaction.

After each cycle was completed, the removed fraction was collected in a 500 mL Duran® bottle and stored at room temperature. Prior to analysis, 10 mL of the Miller Urey 'broth' was transferred to 25 mL plastic Eppendorf® tubes, centrifuged at 5,000 rpm for 10 minutes and the supernatant transferred to a clean tube, before freeze-drying. All samples were lyophilized as a way to concentrate the resulting mixture by water removal. Then, the dried material was re-dissolved in 1mL of 50:50 (v/v) acetonitrile/water and sonicated (in an ultra-sonic bath at 38 Hz) for 15 minutes. The samples were then diluted further (1:10 in acetonitrile/water), filtrated by a syringe filter with a 0.22 µm cut-off, and placed in HPLC vials to be analysed. Also, we prepared a sample blank by subjecting the (50:50 v/v) mixture of acetonitrile/water (used to dilute the samples) to the same process, in order to address the possible inclusion of contaminants from the sample preparation process.

2.3.2 UPLC-MS/MS

An untargeted analysis of the recursive cycles was conducted in a comparable manner as in the method described in **Section 2.2.2**. The samples were analysed with an UPLC-HRMS system, allowing automated MS/MS fragmentation by a DDA method. The detector settings were conserved in this analysis, but the chromatography method was adjusted for Miller-Urey samples. In this case, we decided to select for a chromatographic method that is known for its flexibility and robustness, as well as, efficiency over a wide range of analytes (e.g. polar and non-polar compounds): Reverse-Phase (RP) mode chromatography. The separation efficiency of the reverse-phase column can be comparable with that of the HILIC

column, but is superior to a normal-phase column for samples with a high- degree of chemical diversity. The elution gradient is reversed, when compared to a normal phase column and hence the name. Besides, it is frequently used in HPLC –MS analysis and has become a staple for such applications, making it the most popular column for the last 40 years.

The reverse phase chromatography was performed with an Agilent Poroshell 120 EC-C18 (4.6 x 50mm, 2.7μm) column. Also, the samples were injected in 10 μL aliquots and eluted with a linear gradient mixture of solvents A (water w/0.1% v/v formic acid) and B (100% acetonitrile w/0.1% v/v formic acid) over 35 minutes, while the column compartment was maintained at 30°C. The separation was carried out in a Thermo Vanquish Ultra-performance liquid chromatography system (UPLC), which was coupled to a Thermo Orbitrap Fusion Mass-Spectrometer (MS). The mass-spectral method was set to the same parameters as before, allowing for an automated fragmentation of the three most intense fragments in each full scan (MS^1) and excluding them after 15 seconds of detection. The UPLC-MS method permitted the simultaneous separation and MS/MS fragmentation of the majority of the detected compounds.

However, two new considerations have been taken this time, which were not present in the method described in **Section 4.2.4**, but have been added to the analytical procedure and feature generation process. First, all samples were analysed three times, to ensure that the resulting trends are conserved through instrumental repeats, a necessary step when taking into consideration the highly convoluted matrix of the analytes. Secondly, we have included three instrumental blanks and the sample blank (i.e. subjected to the sample preparation procedure). They were analysed and the detected features subtracted from the features found in the real samples. This was done as a way to remove any features that are not representative of the products within the Miller-Urey broth, but introduced by any other process (eg. by the sample preparation method or mobile phases used in the chromatography, for example). In this case, all features found in the blanks and also in the samples, if not subtracted, would constitute false positives in the feature generation process.

The analytical method allowed us to generate mass-spectral features to represent the detected products, in the same fashion as presented in previous sections. The generated features were also plotted into scatter-plots, as a way to visualize the results. In **Figure 76**, all three instrumental repeats have been plotted, from cycle 1 to cycle 5. The highest variability across instrumental repeats can be observed for cycle 1. From cycle 2 to cycle 5, the repeats are

relatively consistent, with no significant discrepancies across them. Over recursive cycles, the resulting product distributions do not exhibit any major variations (by eye) and no absolute trend can be determined by observation of the scatter-plots. However, it is noticed that as the recursive action progresses, different features (i.e. dots) appear and disappear, even if no pattern can be determined (as seen in **Figure 76,** below).

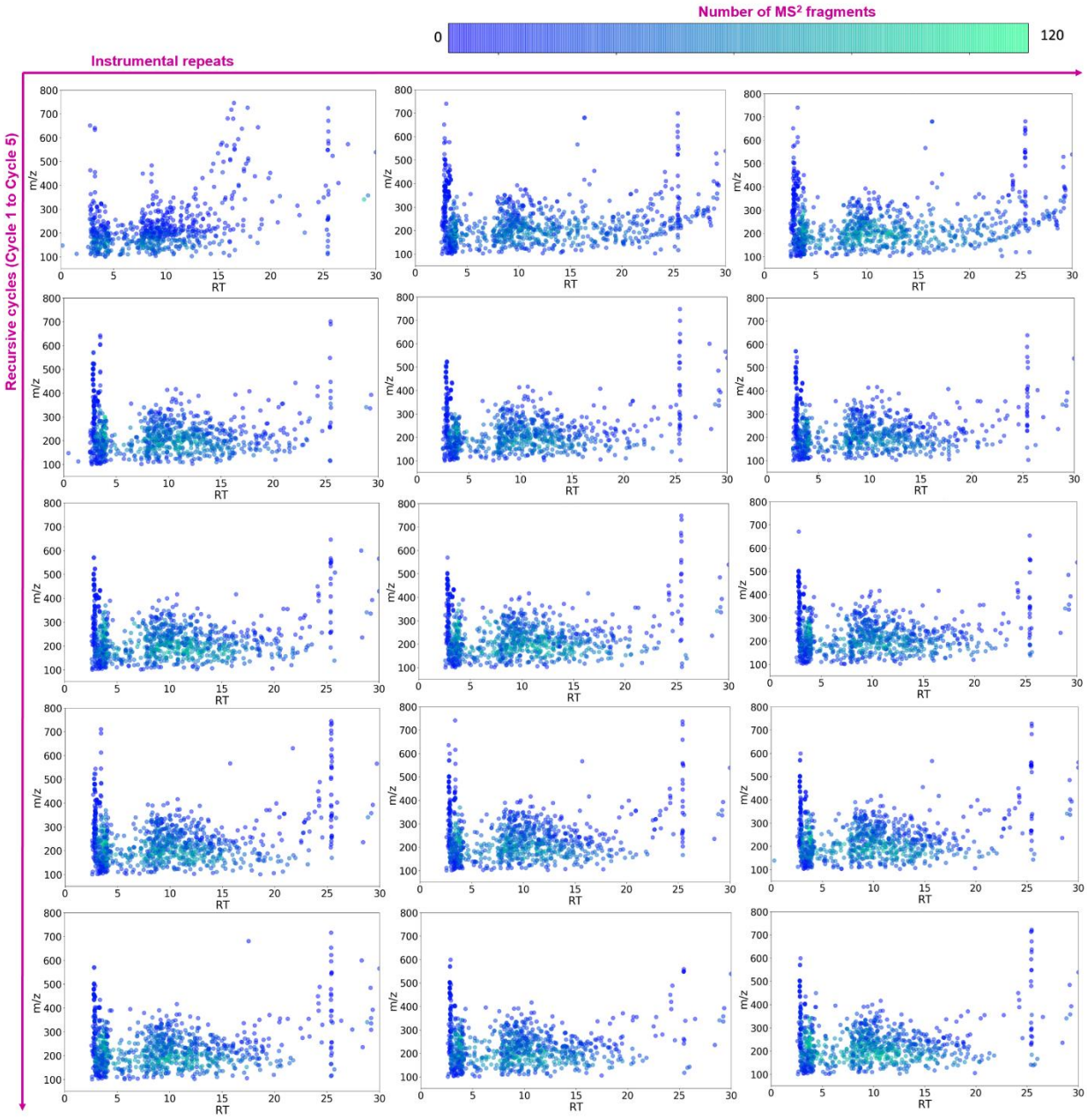


Figure 76. Generated features for the Recursive Miller-Urey samples: All analytical triplicates have been plotted from Cycle 1 to Cycle 5. The number of MS^2 fragments detected for each feature can be seen as a heatmap, colorbar presented at the *top*.

2.3.3 Feature Analysis

Since no clear patterns or trends could be assessed by observing the resulting scatter-plots, we have conducted an alternative analysis of the feature. First, we carried out a Principal Component Analysis of the detected features with the same method described in **Section 2.2.7.1**. This allowed us to compare differences across instrumental replicates, as well as, through the recursive cycles. Furthermore, we looked for features that were present in all samples, but not in any blanks, and considered them as 'Generic features' to the system. These are features that are consistently part of all samples and don't change as an effect of recursive cycles. Also, we wanted to assess if any of the detected features where only present in a particular cycle and in no other one, as the recursive cycles progressed. Therefore, we filtered the features by cycle, looking for those only seen in a given cycle. The features present exclusively in a cycle number are defined as the 'unique features'. Finally, we also wanted to determine if any of the features that where neither 'generic' nor 'unique', had appeared as a consequence of the recursive action. For this, we filtered the detected features looking for those that appeared after cycle one and persisted through the recursive cycles, which we then referred to as the 'New features'. See **Figure 77** in the following page for an overview of the feature analysis workflow.

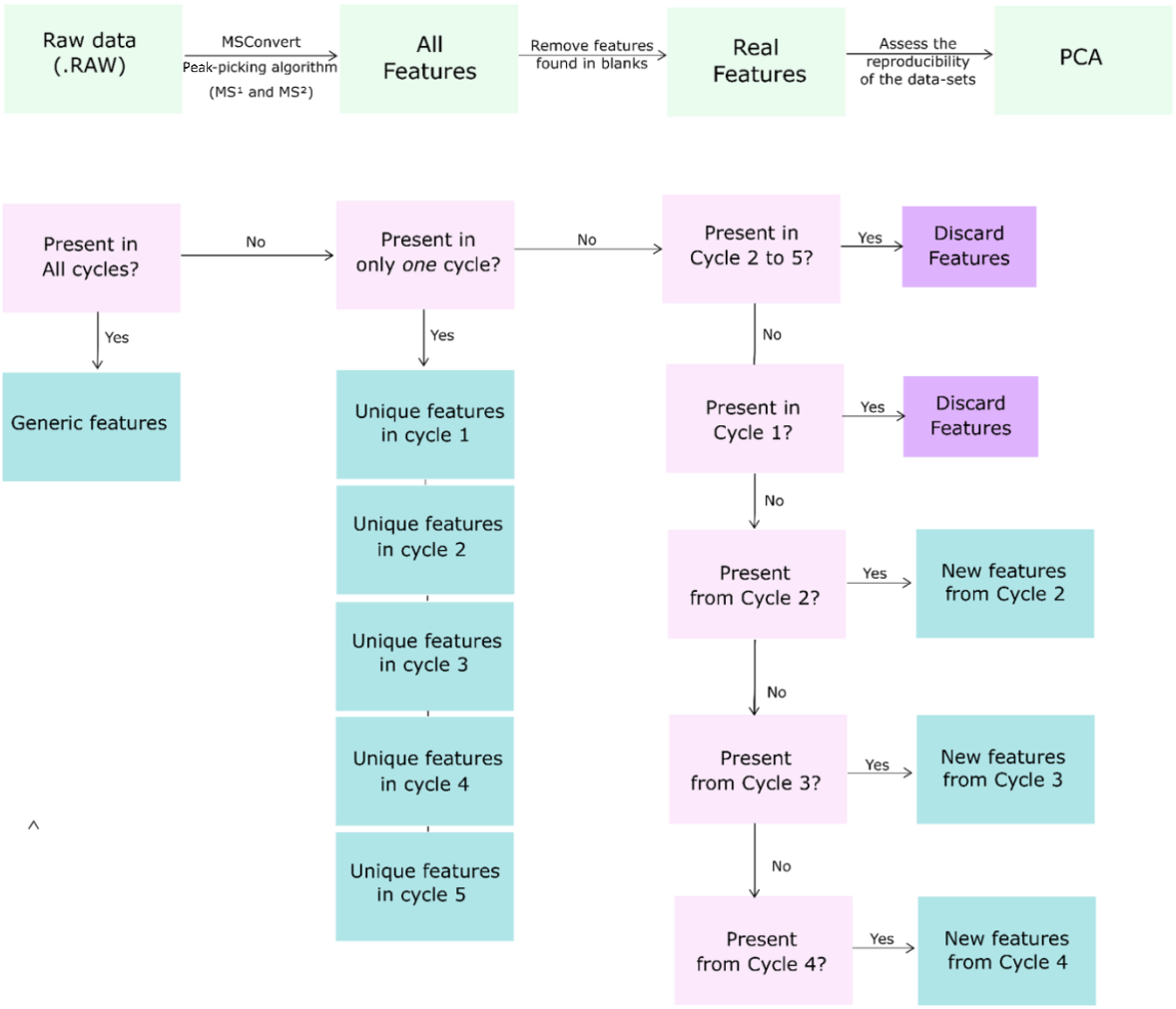


Figure 77. *Feature Generation and analysis:* Data conversion and correction (*green*) is carried out with MSConvert, followed by background subtraction and PCA assessment of the reproducibility across instrumental repeats. The feature analysis (*pink*) executes a filtering protocol, which finds the samples that are Generic, Unique or New to the recursive cycles (*blue*).

2.3.3.1 PCA

Principal component analysis of the features was conducted, as discussed in previous sections. However, in this case the bins were adjusted to coincide with the m/z range of the detected features and the retention time of the chromatographic method. Therefore, the size of the input matrix has been changed to 800 x 30, since we generated 800 bins for the m/z values and 30 bins for the retention time. This approach enabled us to determine variations across the multiple instrumental repeats, which reflected the observable differences within the replicates seen in the scatter plots presented on the previous section. The variability between instrumental repeats of cycle 1, can be seen in **Figure 78** below, where the biggest difference comes from the first triplicate when compared to the second and third repeat. This

variation was the only one that was observed by eye from the scatter-plots and is consistent with the resulting PCA plot. Likewise, the consistency across triplicates seen from cycle 2 to cycle 5, is also reflected in the PCA, where we can see the repeats clustering very closely to each other.

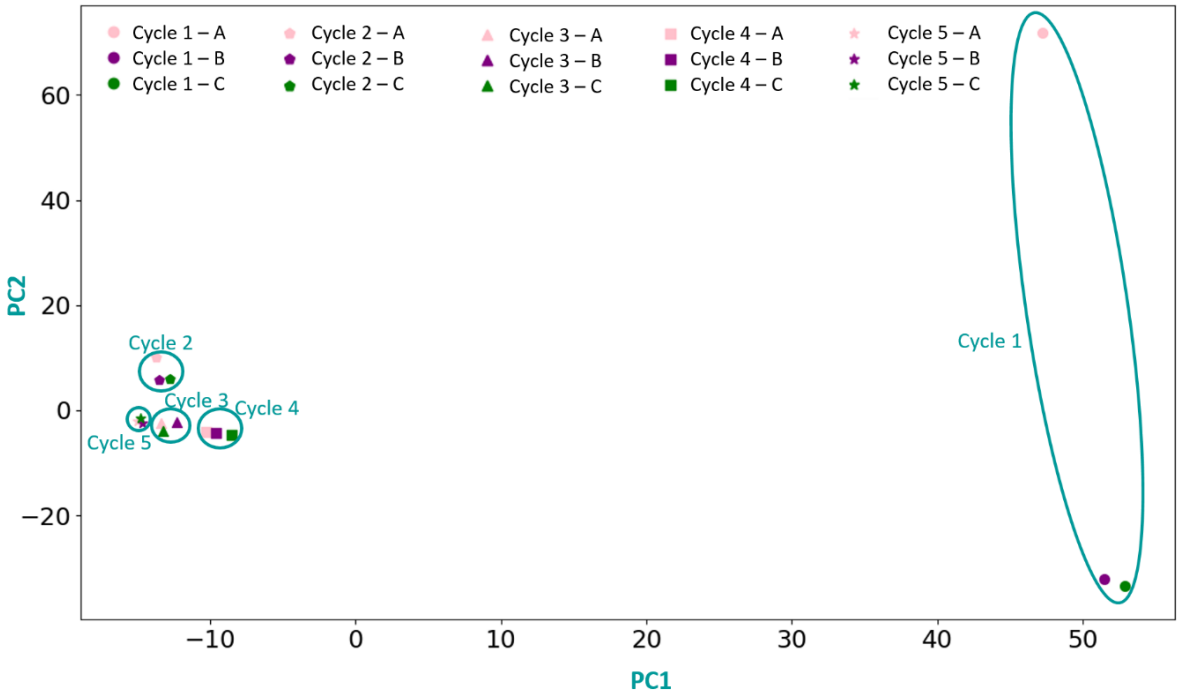


Figure 78. Principal component analysis of the Recursive Miller-Urey: Cycle 1 exhibits a larger variability across repeats than any other cycle. Also, recursive cycles have been highlighted (*blue*), to facilitate the visualization of their clustering.

On the other hand, the PCA also displayed 'grouping' of the analytical repeats for each cycle, making for a seamless distinction across the cycle numbers, for which we can conclude that the resulting product distribution for each cycle, even when not appreciable by eye, is effectively different as a result of the recursive action. The recursive cycles are not as spaced out as in the case of previous studies in the presence of a mineral surfaces, but are distinctly generating unique groups with their instrumental repeats, resulting in a strong indication that the differences across the analytical triplicates is lower than that of the recursive cycles. Also, confirming that the recursive cycles, even if not visually different in their scatter-plots, are effectively different and therefore the product distribution that they represent also is.

2.3.3.2 Filtering for relevant features

Another way to assess the differences across the cycles was to filter for the relevant features by employing different sets of criteria. To do this, we first filtered the detected features into multiple categories that indicated their level of persistence within the system, as seen in **Figure 79**. The number of features detected through different stages of the recursive action can tell us how many of them are rare or how persistent they are. The number of features present in all cycles refers to those products that remained unchanged from cycle 1 or where produced consistently in each cycle; for which we would see it present across all cycles. Also, we calculated the number of features that were present in only one cycle, which describes how many products are short-lived and not capable of survival through the recursive action. As well as, counting the number of features that are present in 2, 3 or 4 cycles; indicating the number of products that are in the 'middle ground', meaning that they find a way to survive or regenerate themselves through the recursive cycles.

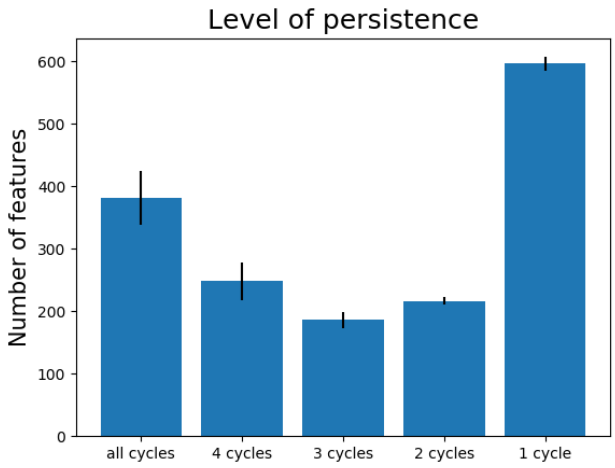


Figure 79. Levels of persistence within the detected features in the recursive cycles: The recursive cycles. The error bars represent the variations across instrumental triplicates

This approach allowed us to get a general measure of how many products are present on different levels of persistence. The number of features that are rare or unique to a particular cycle are more (~600) than in any other case. Followed by the number of features that are present across all cycles (~400), implicating that the amount of unique, and therefore short-lived products is larger than those that are persistent through all cycles. Also, the number of detected features present in 2, 3 or 4 cycle is relatively similar (~250), when compared to the other levels of persistence. In addition, we have calculated the number of features detected for each of the three instrumental replicates to make the observations more robust,

this can be seen in the errors bars of **Figure 79**.

Once the initial classification was done, we referred to the different levels of persistence as follows:

(a) Generic Features (*all cycles*) – Those features that are present across all recursive samples, but not in any blanks. They represent the unchanged or constant part of the product distribution, as they don't change due to the effect of reaction cycling.

(b) Unique Features (*one cycle*) – These features are the ones that are only present in a particular cycle. Their exclusivity towards a particular cycle, indicated that they are short-lived and have effectively been created as an effect of reaction cycling.

(c) New features – All features that are not present in cycle 1 and only appear after consecutive cycles, but don't disappear after that, are what we call the 'new' features. This features needed recursive action to be formed and managed to persist through more recursive cycles.

In this way, we are able to classify the features depending on their persistence within the system and also visualize their distribution. The features which were present across all cycles (generic features) were filtered and plotted, as a way effectively compare them with the previous scatter-plots, see **Figure 80**. The generic features spread all the way across the chromatographic run, indicating that the products they represent entail a broad chemical diversity and only appear to be subtly enriched in the polar region of the elution profile, also the largest density of these features is in the m/z range of 100 to 400 m/z. All instrumental repeats where plotted, with no significant differences across their distribution, see **Figure A30**.

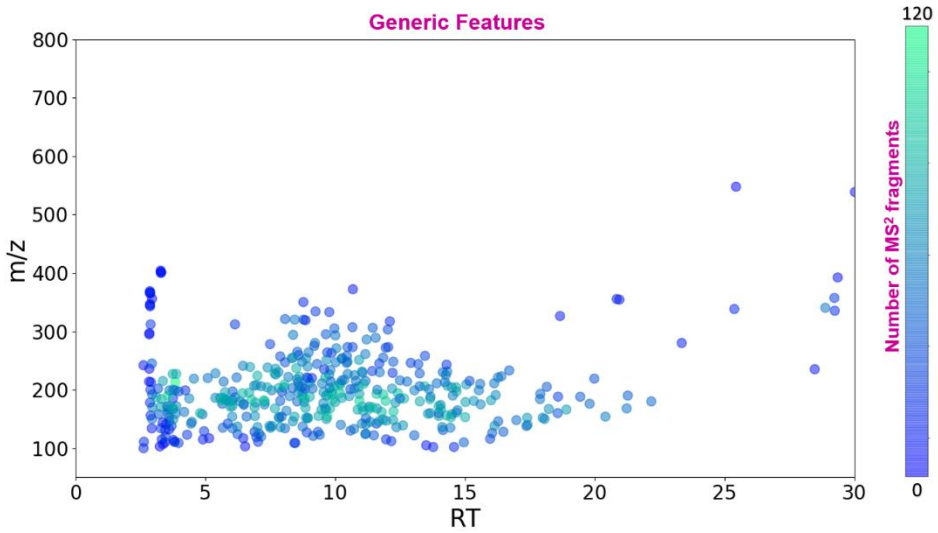


Figure 80. Generic features in the Recursive Miller-Urey samples and their distribution across the chromatographic profile. The number of MS^2 fragments for each of the features are presented as a heatmap.

The features that were present in only one cycle are referred to as the 'Unique' features. These features are representative of those products that are generated in a given cycle number through the recursive cycles, but did not persist through any other cycle. In order to address the distribution of the unique features though the recursive cycles, we have filtered them by cycle, see **Figure 81**. The number of unique features in cycle 1 is larger than in any other cycle (~300), indicating that there is a selective process arising from seeding the experiment with the outcome of a previous one. Moreover, the number of unique features detected for cycle 2 to cycle 5 is roughly one-third (~100), than those found for cycle one. This results suggest that the recursive cycles do produce unique features in each cycle, but it far less than in the initial cycle. In terms of how the recursive action can modify the product distribution of the Miller-Urey experiment, this observation indicated that the experimental variable has a non-trivial effect on the distribution of the products generated. Also, the presence of unique features in the consecutive cycles can be seen as confirmation that when a recursive cycle takes place, a new product distribution is achieved.

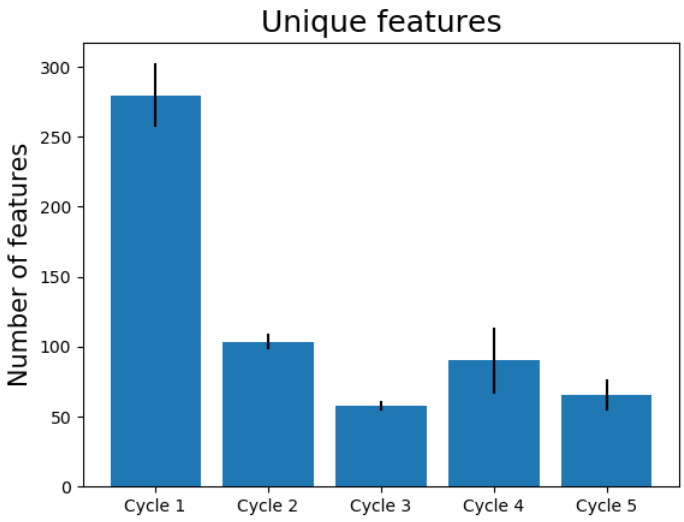


Figure 81. Unique features within the detected features in the recursive cycles: The number of unique features detected in cycle 1 to cycle 5. The error bars represent the variations across instrumental triplicates.

Furthermore, we also wanted to address what was the distribution of the unique features present in each recursive cycle. Therefore, we have plotted the filtered features for each cycle into their scatter-plots, see **Figure 82**. The distribution of these features is observably different for each cycle. For example, in cycle 1 the features are distributed across all of the chromatographic run, indicating these unique features represent compounds of both polar and non-polar character. However, from cycle 2 to cycle 5 we observe that the distribution of the features is relatively scarce in the middle of the elution profile, which infers that most of the unique features have either highly polar or non-polar affinities. Also, in cycle 5 we can see how the unique features seem to be enriched in polar compounds. The unique features in each cycle where generated for all instrumental replicates and as in the case of the generic features, there are no significant variations across the distribution of the repeats, see **Figure A31**.

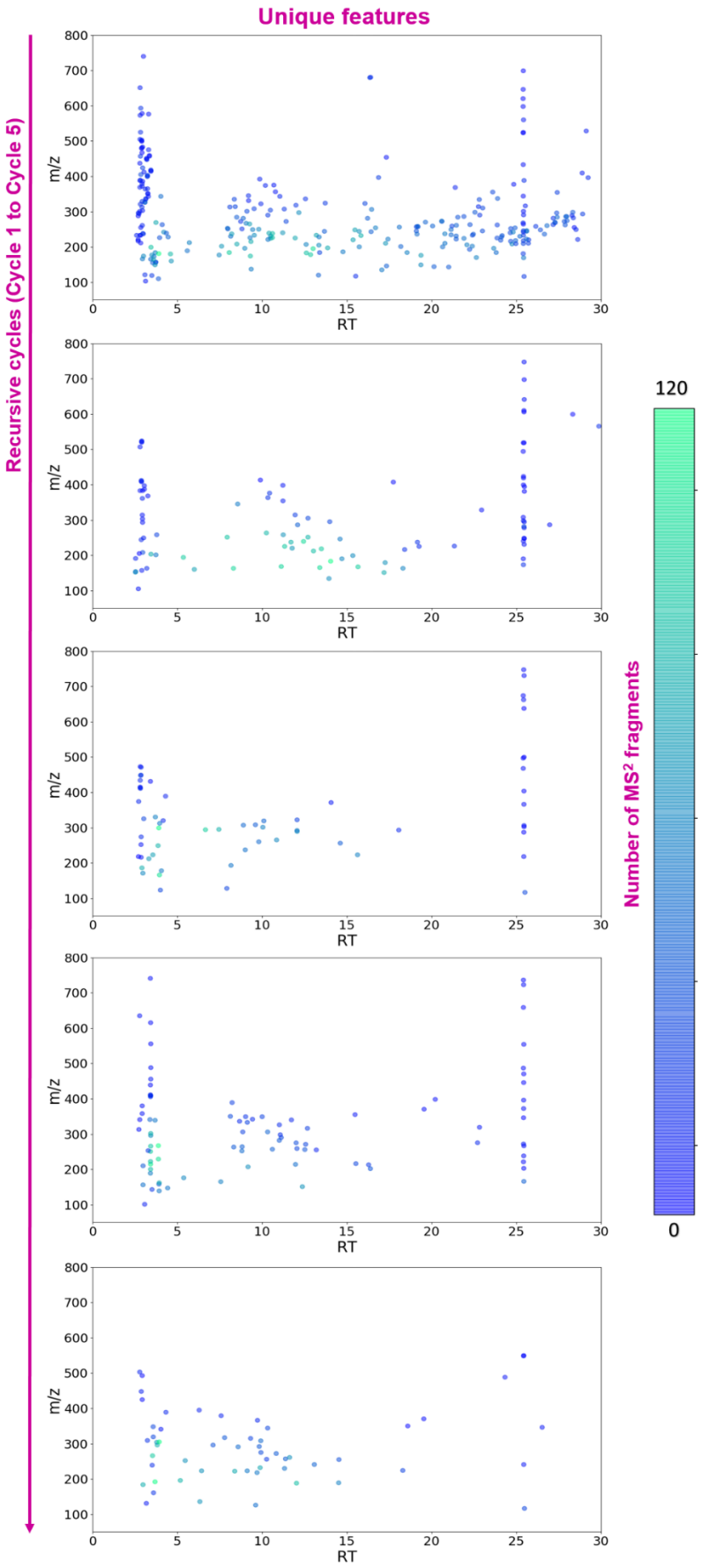


Figure 82. Distribution of the Unique features across recursive cycles. The number of MS^2 fragments for each feature has also been included (heatmap).

Additionally, we wanted to investigate if any new features appeared as an effect of the recursive cycles managed to survive through the recursive action. This is to say, that they were only present as a consequence of recursivity, but are not consumed or destroyed in the consecutive cycles. In this sense, the new features are different from the generic or unique, since they are capable of survival through environmental stressors. To achieve this, we have looked for features that only appeared from cycle 2 (e.g. not present in cycle 1) and where present in all the following cycles after that (e.g. Cycle 3 to Cycle 5). In the same manner, we looked for which features where new in cycle 3 (e.g. not present in Cycle 1 or 2), but remained through Cycles 4 and 5. And finally, we looked for which features where new to cycle 4, not present in cycle 1 to cycle 3, and persisted in cycle 5. Also, in the case of the new features that appeared in cycle 5, they would be equal to the unique features for cycle 5 since it is the last cycle of the experiment. The filtering strategy allowed us to look for the number of features that persist across the recursive action (**Figure 83**), as well as, their distribution (**Figure 84**).

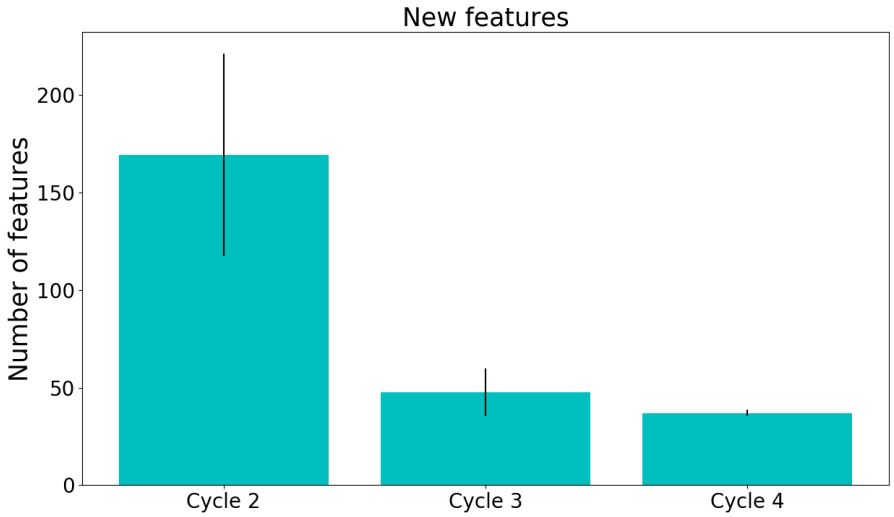


Figure 83. New features arising from the recursive cycles: The number of new features detected from Cycle 2, Cycle 3 and Cycle 4. The error bars represent the variations across instrumental triplicates

The number of new features in cycle 2 that persisted all the way to cycle 3, is larger than for those detected from cycle 3 and cycle 4. To some extent, this can be expected. Particularly, when considering that as the recursive action progresses, the number of new features will be increasingly limited by the resulting distribution in the previous cycle. This effect can also be seen for cycle 3 and cycle 4, where the number of new features reduces linearly. Also, the variation across instrumental triplicates is larger (see error bars in **Figure 83**) for cycle 2, since it has the largest amount of new features. As well, we can expect that the more

features are present (i.e. data points), so will be the associated error for their instrumental repeats.

The distribution of the new features for the recursive cycles 2, 3 and 4 was also addressed. In **Figure 84** (below), we can observe distinct variations across the resulting distribution for the new products arising on each cycle. The new features do not appear to have any compositional selectivity (e.g. polar or non-polar characteristics) in any of the cycles, as they are distributed across the whole chromatographic run. Also, the m/z range of which all new features (cycle 2 to cycle 4) appear, 100 to 450 m/z, is comparable between the cycles. Moreover, the distribution of the features is conserved across the instrumental repeats, reassuring that the error associated with the triplicates does not result in different features being detected. See **Figure A32**, for the distribution of the other two instrumental replicates. Likewise, the number of MS^2 fragments for the detected features has been plotted as a heatmap, giving an extra dimension (or insight) into the differences within the detected features. In this way, we are able to confirm that the distribution of the features in multiple instrumental repeats is not only the same, but the number of MS^2 fragments detected for each of them is also consistent. Besides, another piece of information provided by the number of MS^2 is the relative complexity of the feature, if we can assume that the higher the number (*in green*), the more intrinsically complex is the product that the feature represents. If this can be assumed, then the new features contain all sorts of molecular complexities and are not necessarily more complex than the generic or unique ones.

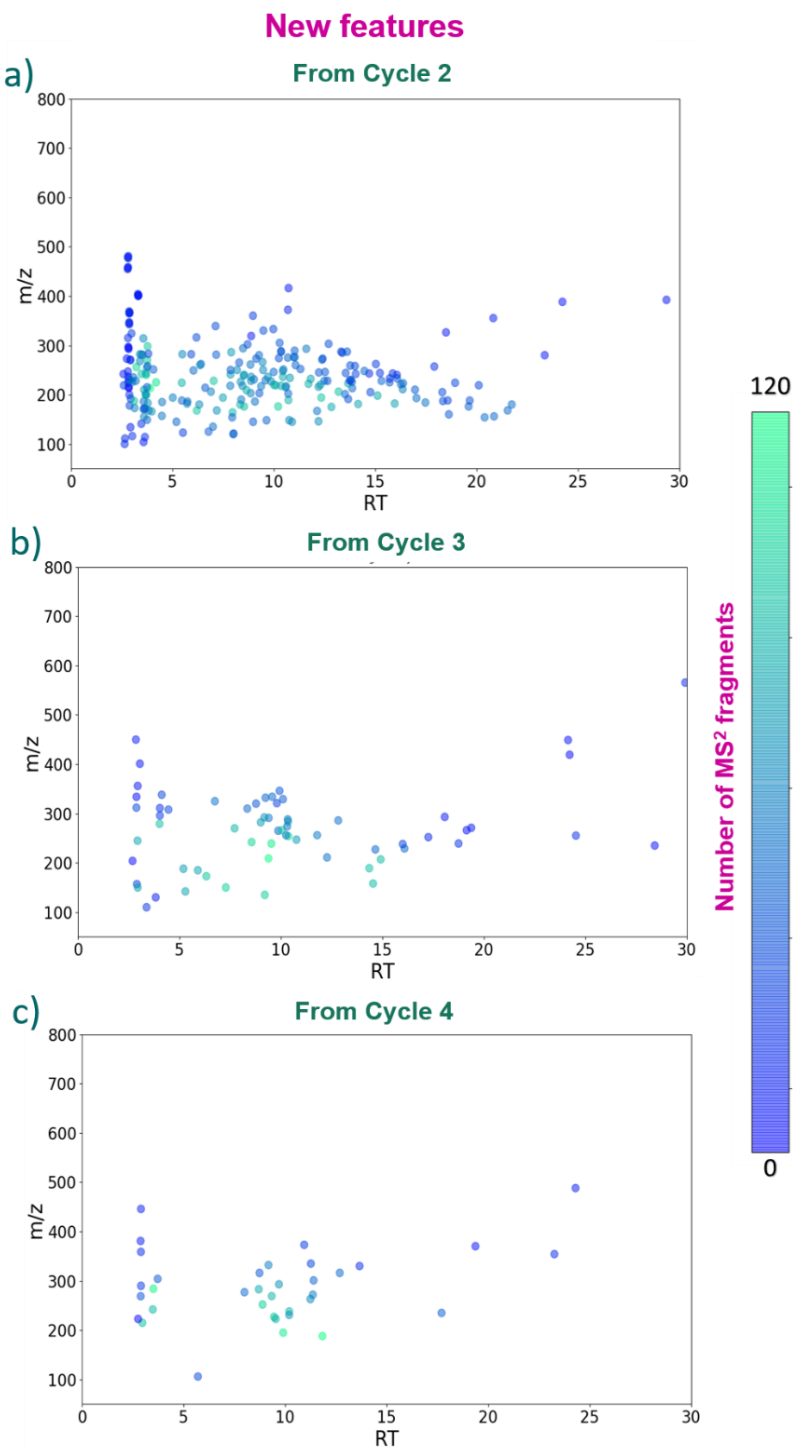


Figure 84. Distribution of the New features from Cycle 2, 3 and 4. The number of MS^2 fragments for each feature has also been included as a heatmap.

2.3.4 ^{1}H –Nuclear Magnetic Resonance (^{1}H-NMR) of the insoluble fraction

During the sample preparation process, the insoluble fraction is removed prior to the lyophilisation step, for the UPLC-MS/MS analysis. Therefore, this fraction is not addressed in the sections presented above. In light of this, we wanted to assess if any differences could also be seen for the insoluble part of the Miller-Urey reaction, as an effect of the recursive

cycles. However, in order to do this, we had to enlist yet another analytical technique. The insoluble fraction was dissolved in deuterated DMSO-d6, after removing the excess water (also by freeze-drying) and analysed by ^{1}H- NMR. The resulting ^{1}H-NMR spectra was compared across the resulting cycles, as seen in **Figure 77**. This allowed us to assess general differences across the recursive cycles, by giving rise to variations in the intensity of some peaks, indicating that the relative abundance of the material changes as a consequence of the recursive action. For instance, at 1.9ppm in the ^{1}H spectrum of Cycle 2, we noticed significant increase in peak intensity which returned to its original intensity level by Cycle 5. Also, at the 7.9 ppm there is a peak that is increasing linearly from Cycle 1. Moreover, some peaks appear and disappear across the recursive samples. For example, at the 3.5 ppm there are two peaks appearing in cycle 2, but by cycle 5 only 1 of the two remain. Similarly, there is a new peak (at low intensity) appearing around 11.5 ppm, possibly to a carboxylic acid or aldehyde group in the mixture. All of these observation are quite general, however they do constitute notable differences amongst the insoluble fraction of the Miller-Urey recursive samples.

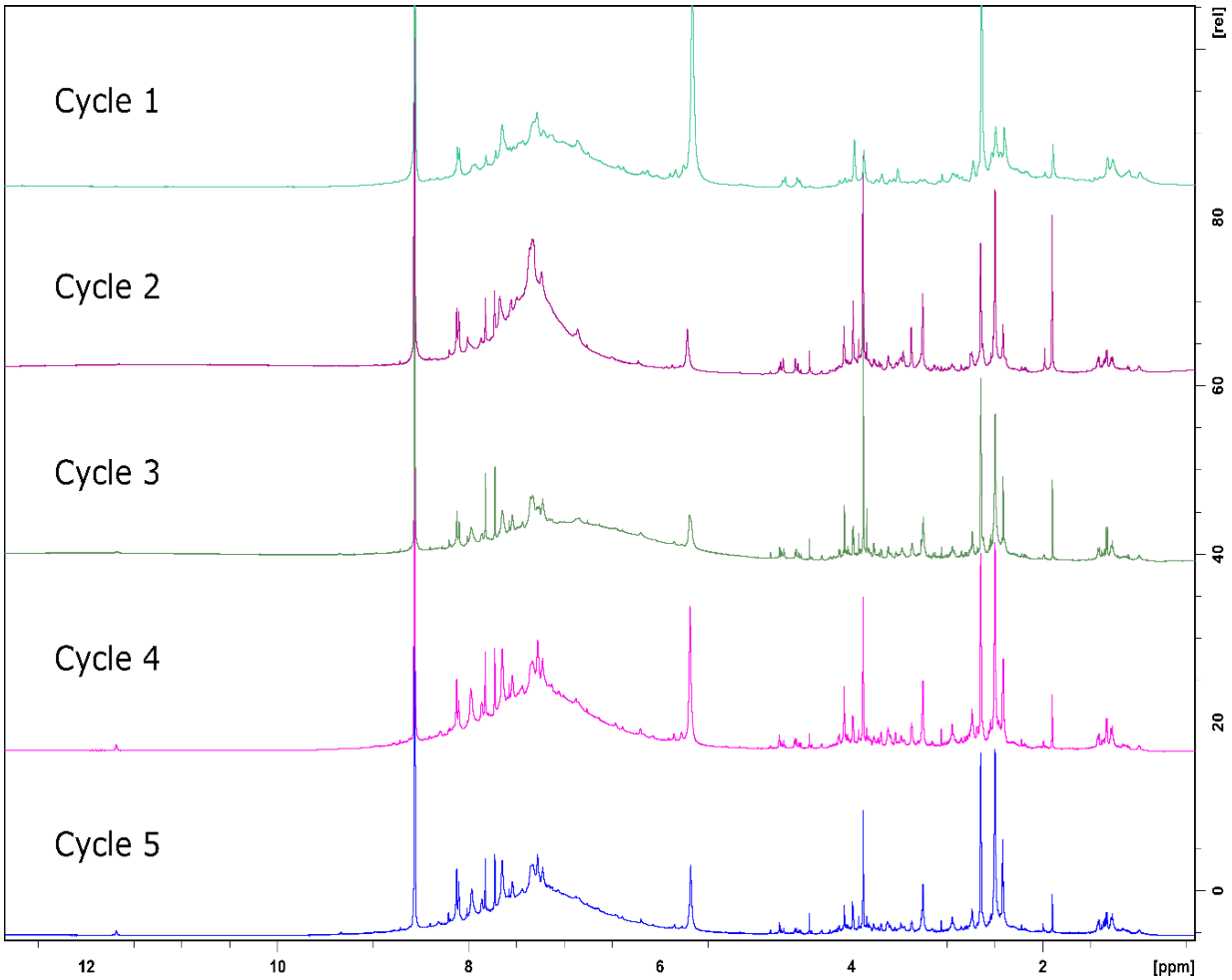


Figure 85. ^{1}H-NMR spectra of the insoluble material from the Miller-Urey recursive cycles: Observable differences in the intensity of multiple peaks can be seen for Cycle 1 to Cycle 5.

2.3.5 Section summary

The Miller-Urey experiment has been carried out in a recursive fashion. The approach aims to recreate the material cycling that might have taken place through natural processes on early earth. In aims to pursue a more realistic insight on the synthesis of prebiotic mixtures, when performing them in a laboratory setting. The effect of recursive action on the Miller-Urey mixture resulted in distinct variations on the resulting product distribution of the cycles. In order to execute this assessment, we have generated mass-spectral features to represent the products generated. Through the untargeted UPLC-MS/MS analysis of the mixtures, we have screened for general differences across the resulting features, as an effect of the process variable. After a general evaluation of the generated features, we realized that a comparison between their distributions by eye, was not a satisfactory way to assess their differences. Therefore, we carried out PCA on the detected features, finding that the analysed samples

did generate unique groups according to their cycle number.

Furthermore, we carried out a series of classifications on the detected features by filtering them over multiple criteria. This enables us to assess the levels of persistence of the detected features across the recursive cycles. First, we looked for those that were present in all cycles, as they would represent the unchanged fraction within the product distribution and called them ‘generic features’. Secondly, we filtered for the features that were exclusive to a particular cycle, which we deemed as ‘unique’ features. The number of unique features is over two-times higher in cycle 1 than in any other cycle. The distribution of the generic features and the unique features was also plotted as a way to assess if they were enriched in any particular region of the chromatographic run (in hopes to roughly assess their relative polarity). However, this did not show any particular preference or pattern on the distribution of either. Nonetheless, we also filtered the features for those that only appeared as an effect of the recursive cycles but managed to persist through the recursive action. These features we called, the ‘new’ features. As the recursive cycles progress, the number of new features reduces linearly, from cycle 2 to cycle 4. The distribution of these features was also seen to vary from cycle to cycle.

Moreover, the insoluble fraction of the recursive Miller-Urey samples was addressed by ^{1}H-NMR. This technique is far less sensitive than the one used to generate the mass-spectral features (i.e. about one-hundred-to ten-thousand-times[169]),but can provide an overview of general differences across the recursive cycles. After comparing their spectra, we can see differences in the intensity of multiple peaks, encompassing different chemical shift regions. Also, some peaks appear and disappear through the recursive cycles, indicating notable differences arising from the recursive process (i.e. changes in functional groups in the molecules of the mixture).

The results presented in this section suggest that the effect of reaction cycling in prebiotic mixtures, such as the ones produced by the Miller-Urey experiment, is non-trivial and should be taken into considerations when investigating such systems, especially if we want to investigate the evolution of complex chemical mixtures and provide a fair analogy of the natural processes of our early earth. An approach that should bring experiments aiming to a prebiotic context, a little closer to genuine ‘prebiotic plausibility’.

3. Conclusions and Future Work

The work reflected in this thesis has mainly been focused on the analysis of complex systems, which in one way or another, are related to the Origins of Life research. Different analytical techniques have been explored, each having its strengths and weaknesses. However, assessing which analytical technique is best, is closely tied to the aim of the projects.

3.1 Miller-Urey: In a deuterium world

A series of deuterated and classic Miller-Urey experiments were carried out with parallel experimental setups as a way to generate experimental replicates. This was done in order to gauge differences in the product distribution arising from the substitution of hydrogen with its heavier isotope, deuterium. The samples were analysed with two different approaches: GC-MS and HPLC-FLD. Initially, neither the GC-MS nor HPLC-FLD analysis yielded significant differences within the identified products of the deuterated and 'classic' Miller-Urey experiment. However, when unbiased statistical analysis (i.e. PCA) of the datasets was conducted, differences arising from the experimental variable were observed and confirmed through the experimental repeats, demonstrating that there are variations in the product distribution of a deuterated experiment that cannot be appreciated by eye, since they were not necessarily exposed through a direct comparison of the resulting chromatograms of the analytical techniques employed. Nonetheless, minor differences were also present between experimental repeats, which lead us to the conclusion that even in the best case of scenarios, intrinsic variations due to the reaction setup are inherently present in the Miller-Urey experiments, something we should take into consideration, particularly when comparing results of multiple experimental setups across the globe.

In the GC-MS analysis, the identification of all substances in solution was challenging due to the high dynamic range of the compounds concentration. Also, confirmation by an external standard was not always possible due to the complexity of the sample matrix and the inevitable retention time shifts. Therefore, the use of internal standards is still needed in order to fully confirm the unexpected compounds detected in the samples. This consideration will be taken on in future methodologies, alongside with Single Ion Monitoring of the identified species, as a way to optimize their quantification. A targeted analysis does lead to an increase in bias, but it can be beneficial as complementary strategy, since the

quantification of any products in a complex mixture requires it. For this reason, the HPLC-FLD analysis was conducted and tailored towards the identification of amino-acids in the product mixture. The relative intensities of the detected amino-acids did not exhibit any significant changes or identifiable trends as an effect of the deuterated substitution. Consequently, we did not find any evidence of the experimental variable imparting selectivity on the distribution of the amino acid product space.

Furthermore, as a way to assess what kind of material was 'lost' by the filtration step, SEM-EDS analysis of the unfiltered (dried) samples was conducted. This led us to believe that a significant amount of the insoluble particulate retained in the filter came from the glassware of the apparatus itself as a result of the degradation through the extended exposure to harsh (reducing) conditions. A feature of the experimental setup that was mentioned by Stanley Miller himself.[194]

For the first time in the Origins of Life research, we looked at the differences between the product distributions in an untargeted fashion alongside with conventional analytical regimes, and tried to assess the relevance of an isotopic substitution within the atmospheric composition of the Miller-Urey experiment. We imagine that the next series of Miller-Urey experiments will have an inclusion of various minerals where the influence of the environment will be explored. Future generations of the Miller-Urey experiment should also include an experimental variable for wet/drying cycles, in which the classic setup would be altered by adding a valve to enable the enclosure or the reaction vessel, allowing the promotion of peptide formation within the system.[220] The targeted analysis of peptides through HPLC-MS of the Miller-Urey samples will have to be included. This should be complemented with the untargeted acquisition and analysis of the mixtures since, as this section demonstrated, it can provide non-trivial insights into the dynamics of the complex system.

3.2 Recursive cycles of the Formose – Formamide reaction

The one-pot reactions of simple precursors, such as those found in the formose reaction or formamide condensation, continuously lead to combinatorial explosions in which simple building blocks capable of function exist, but are in insufficient concentration to self-organize, adapt, and thus generate complexity. In light of this, we explored the effect of recursion on such complex mixtures by 'seeding' the product mixture into a fresh version of the reaction, with the inclusion of different mineral environments, over a number of reaction

cycles. Complemented with the untargeted analysis (UPLC-HRMS) of the mixtures, we found that as a result of the recursive action, the overall number of products detected reduces as the number of cycle's increases, thus limiting the combinatorial explosion. Also, different product distributions where observed for all mineral environments studied, which developed into unique patterns through the recursively enhanced environmental selectivity. This discovery demonstrates how the involvement of mineral surfaces with simple reactions could lead to the emergence of some building blocks found in RNA, ribose and uracil, under much simpler conditions than originally thought.

Our results conclude that an untargeted approach generates a more robust overview of the highly complex product distribution obtained in analytically intractable mixtures. However, an array of limitations were identified during the process of generating the untargeted workflow. For instance, the simultaneous exploration of multiple parameters (such as different mineral environments) increases the variability of the sample matrix, challenging our capacity of analyte comparison across samples. When multiple experimental variables want to be addressed, a need for experimental controls becomes a necessity and a careful design of experiments is required to carry out complex prebiotic reactions in a confident manner. Alongside with a minimum of two experimental replicates and three instrumental replicates, to ensure reproducibility on any trend arising as an effect of the experimental variable(s). This is also extended to the analysis of samples generated in the systematic studies of prebiotic reactions, where a normalization step must be addressed, as to account for as many analytical features as possible without an increase in false-positives. The inclusion of internal standards in the analysis of the samples can circumvent some of the associated errors, such as the loss of mass-accuracy over analytical repeats and sample matrix effects. As well, background subtraction in the datasets must be executed beyond an instrumental blank, to account for any contaminant that might be present in the experimental setup or coming from the sample preparation procedure.

The data-processing would benefit from careful selection of the algorithms for raw data conversion, retention-time correction and signal pre-processing. In this sense, modern advances in the analytical workflow for complex mixtures (initially developed for the 'omics' field), will aid in choosing the right combination of parameters as to minimize the loss of information from the system. Yet, this will not be enough and the workflows must be adapted to include non-biological pathways in network analysis. To date, -denovo- network analysis in complex prebiotic ensembles has not been addressed experimentally. We believe that the work presented in this chapter, paves the way to jump-start the integration from

computer models that deal with an in-silico chemical space in search for patterns arising from the networks of molecules with real experimental data

Furthermore, studying combinatorial explosions requires a large number of experiments and consequently a high-throughput experimental design, which can be aided by the construction of an automated liquid-handling platform that can carry out the recursive process in a reliable way. While the levels of reproducibility in combinatorial explosions are not assessed directly in this work, it has indeed enabled the possibility for such studies, in which a comprehensive overview of the resulting products would be substantial in order to draw any meaningful conclusions. Future work will take on this approach and an automated platform will be integrated with real-time analytics as a way to tackle the compositional complexity of the mixtures. Under the assumption that an increased sampling rate of the system will result in more information being retained, in a time-resolved manner. Ultimately, we believe that automated recursive experiments will bring us one step closer to a reasonable 'real-life' scenario detached from human intervention. The combination of such system with the analytical approach presented will provide an improved experimental regime for looking at the evolution of complex mixtures from simple precursors under non-equilibrium conditions.

3.3 The Recursive Miller-Urey Experiment

The effect of reaction cycling on the landmark Miller-Urey experiment was explored. To achieve this, the experimental setup was not modified but rather replenished and restarted, leaving a portion of the resulting mixture behind each time. This process aims to recreate natural environmental conditions on early earth, since current consensus agrees on the presence of atmospheric cycles by the time the first organism came to existence.[55] We believe that this work is an extension of Miller's original vision of simulating the prebiotic environment in a laboratory setting. To this date, no other attempts to include the influence atmospheric processes into a bottoms-up prebiotic (spark-discharge) experiments had been carried out. Moreover, the analysis of continuous evolution of abiotically generated material has been largely limited, due to the highly convoluted nature of the resulting product mixtures. In other words, investigations of the variations in the product distribution as an effect of experimental variables continues to be focussed on biologically-relevant material in prebiotic broths, as 'a defence mechanism' against their perceived analytical intractability. For this reason, we have extended the analysis of the Miller-Urey recursive cycles to an

untargeted workflow which gives a secondary importance to the identity of the detected compounds. We aimed to represent the products, as confidently as possible, by constructing mass-spectral features that can encompass a larger fraction of the product space than a targeted approach. Therefore, instead of looking for particular products, we focussed on addressing how the features change over the recursive action (i.e. process variables).

The samples were analysed in an extension of the untargeted UPLC-HRMS method presented in the previous section. In order to measure the differences within the system as the process variable unravels, we constructed a filtering algorithm that simultaneously processes and filters the features according to their level of persistence. This allowed us to generate a series of classifications on the detected features by filtering them over multiple criteria. Initially, we looked for features present in all cycles, as they would denote the unchanged fraction within the product distribution and referred to them as ‘generic ’. These features must have found a way to be generated in each experiment, or to remain unreacted through multiple cycles. We can think of them as the noise-level of the chemical system. Then we filtered for short-lived features, which only appeared in one particular cycle and called them ‘unique’. These features could eventually lead to a mechanistic insight into the recursively-induced changes in the product distribution of the prebiotic broth. Finally, a third category was constructed and deemed to be the most relevant for this study: the ‘new’ features. These features only arose as an effect of the recursive action (i.e. not present in cycle 1) and managed to survive the recursive process over generations. Each cycle (e.g. cycle 2-4) has both unique and new features arising, which continuously modify the product distribution. The identification of different classes of features within the product mixture is one step forward into elucidating the complex chemical networks that give rise to the combinatorial explosion that are bottoms-up prebiotic studies, such as the Miller-Urey experiment. The results presented in this section conclude that the effect of reaction cycling in prebiotic mixtures, such as the ones produced by the Miller-Urey experiment, is non-trivial and should be taken into considerations when investigating such systems. Especially, if we want to investigate the evolution of complex chemical mixtures and provide a fair analogy of the natural processes of our early earth. We believe that such approach would also eventually push the experiments designs towards a prebiotic context a little closer to genuine ‘prebiotic plausibility’.

Future work will focus on developing an automated way to execute recursive Miller-Urey experiments, as well as extend our chemical knowledge of the features. The classified features can be analysed by a complementary targeted approach, which expands our scope

into the quantitative territory. Also, we can extrapolate chemical formulae from the mass-spectral features and use this to generate van Krevelen plots, which can serve as a guide of the chemical identity of the features. Data-dependent acquisition of tandem spectra will also be beneficial as to correct for chemical formula calculations and aid in the identification of the compounds represented by the features. As presented in the previous section, database search enables the high-throughput identification by matching the exact mass and MS^2 mass-spectral pattern with the ever-growing array of existing libraries. However, it would be beneficial for the entire prebiotic chemistry community to start developing small molecule databases for the elucidation of abiotically generated compounds in a prebiotic context. Currently, the identification process for complex mixtures is largely built out of biological experimental data and therefore limits our searches into 'biologically-relevant' compounds. Libraries developed for environmental analysis of complex mixtures (i.e. pesticides, petroleum and others) do help to include other small-molecules, but are not as widely-distributed or enriched as the biological ones used in the 'omics' field. Therefore, there is a gap to be filled in the documentation of analytical protocols and collection of experimental data from the identified compounds in prebiotic chemistry experiments. Future efforts should be made in constructing an Origins of Life equivalent of the Human Genome Project, where multiple experimental setups across the globe work together to construct a database of the (potential) chemical space on early earth. Once there is a path for the elucidation of compounds within the prebiotic mixtures, network analysis of the compounds can be conducted. Nonetheless, since molecular networking models used to re-construct biological pathways are based on our understanding of living systems and there is no equivalent for this in the process of chemical evolution, new challenges will be inevitable. For this, the many models developed in the past 100 years of the Origins of Life field [221,222,223,224,113,225,226] can act as a brute-force initial approach into finding out how to elucidate (or hopefully, optimize) a model of chemical evolution, constructed with assumptions derived from experimental data.

4. Materials and Methods

4.1 The Miller-Urey experiment a 'Deuterium world'

4.1.1 Reagents and Gases

All amino-acid standards and the HPLC (mobile phase) additive were obtained from Sigma-Aldrich and used without further purification. HPLC grade water and deuterium oxide used for the Miller-Urey Experiment were supplied by Goss Scientific. The gas mixtures were supplied pre-mixed by the British Oxygen Company (BOC) and CK Special Gases Ltd. For the HPLC-FLD analysis, the mobile phases (e.g. acetonitrile and water solvents, 99.9%) were obtained from VWR. Also, the ophthaldialdehyde / 3-mercaptopropanoic acid (OPA/MPA) mixture was purchased from Agilent Technologies.

4.1.2 Experimental procedure

In order to obtain experimental replicates, experiments where conducted in parallel, using a set-up was built of two rigs sharing the same gas supplier (see **Image 3**). The parallel rigs were supplied with exactly the same gas mixture, as a way to ensure experimental reproducibility (see **Table A1**). Three experimental runs were carried out for both for the deuterated and 'classic' experiment. This resulted in twelve samples, when taking into consideration the experimental replicates (i.e. X1 and X2).

Image 3. Image of the Miller-Urey experimental set-up, the two rigs in parallel are displayed.

A typical Miller-Urey experiment was carried out in the following fashion:

(a) 400mL of HPLC water (or deuterium oxide) was placed in the reaction flask.

(b) The rig was then pumped down three times to degas the water. After the third evacuation, the system was pressurized (1 atm) with gas mixture of roughly: 40% methane, 40% ammonia and 20% hydrogen.

(c) The round bottom flask (i.e. reaction vessel, 500 mL) was heated with a heating mantle, tuntil the water started boiling.

(d) Once the water was boiling and recirculation established, the 24 kV spark discharge was turned on, in a 10 seconds alternating ("on" - "off") duty-cycle.

(e) All experiments were run for seven days. During this time the water in the flask changed from clear to different shades of brown.

(f) After seven days, the spark discharge and the heating mantel were turned off and the system was allowed to cool down.

(g) Once cooled, the sample was collected by placing it in 500mL Duran® bottles and stored at room temperature.

All Miller-Urey experiments were carried out with the help of Dr. Geoff Cooper.

4.1.3 Sample preparation and derivatisation reactions

4.1.3.1 GC-MS

The Miller-Urey sample preparation procedure for GC-MS analysis was conducted as follows:

(a) 10 mL of each sample was transferred into 45 mL falcon tubes.

(b) The tubes where centrifuged for 10 minutes at 4,400 rpm.

(c) The supernatant was filtrated with a syringe filter (0.22 µm cut-off) and transferred (~8mL) to 15 mL falcon tubes.

(d) Samples were lyophilized by freeze-drying after placing them at -20 °C for 8-12 hours.

Image 4. Image of the Miller-Urey samples in 15 mL (plastic) tubes after freeze-drying. An array of different shades of brown can be observed.

For the derivatisation reaction, a methodology was adapted from Molnár-Perl *et al.*[177]:

(e) 450 µL of MTBSTFA and 50 µL of acetonitrile (HPLC grade) were added to the dried sample (in a glass vial), then placed in a ultra-sonic bath at a frequency of 37 Hz, at a temperature range of 50 C° to 65 °C for 60 minutes.

(f) After derivatisation, the samples are diluted 1:10 v/v in acetonitrile (HPLC grade) and filtered using a syringe filter with a 0.22 µm cut-off.

(g) Finally, 1mL of the supernatant is filtered with a syringe filter (0.22 µm cut-off). This was done twice and the filtered fraction was transferred into sample vials (total volume per sample: approx. 0.5 mL to 0.8 mL).

Images of the full step-by-step derivatisation procedure is presented in **Figure A5**.

4.1.3.2 HPLC- FLD

The final product of the Miller-Urey reaction was prepared for HPLC-FLD analysis in the following manner:

(a) 1 mL of the reaction mixture was placed in an Eppendorf tube and centrifuged at 10,000 rpm for 5 minutes, in order to precipitate any possible suspended particles.

(b) The supernatant was then transferred to a HPLC vial.

This was followed by a derivatisation reaction, which imparts a strong chromophore to the analytes (i.e. amino-acids) with an OPA/MPA procedure (see **Figure 86**) and was carried out as follows:

(c) The boric acid buffer (pH=9.5) for the OPA/MPA method was prepared in a 50:50 v/v solution of 0.2 M H_3BO_3 in KCl and 0.2 M NaOH, from already prepared solutions. Also, the pH was adjusted when necessary with H_3BO_3.

(d) Into an HPLC (glass vial), 50 µL HPLC water was added to 20 µL 0.2M boric acid buffer (pH=9.5), 75 µL sample and 20 µL OPA/MPA mixture, before adding an extra 20 µL HPLC water.

(e) The HPLC vials where then vortexed two times.

(f) Finally, the samples were filtered by a spin-filter (0.22 µm cut-off), transferred to a fresh HPLC vial (equipped with a 250 µL insert) and placed in the auto-sampler.

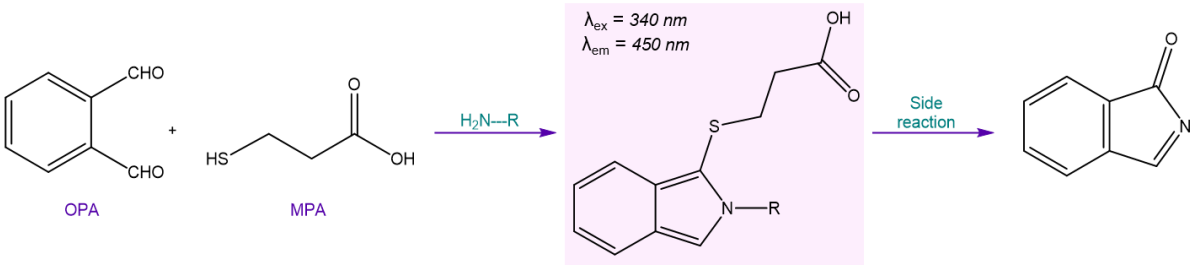


Figure 86. A general reaction scheme of the OPA/MPA derivatisation.

4.1.4 GC-MS method

The GC-MS was performed using an Agilent 7890 GC / 5975 MSD. 2 µL per sample injected into the GC in split-mode 1:20. The injector temperature was set at 250 °C, MS transfer line to 230 °C and the detector at 230 °C. Helium was used as the carrier gas at a constant flow of 1.0 mL/min. A HP-5MS capillary column (95% dimethylpolysiloxane, 5% diphenyl; 30m x 0.25m x 0.25 mm) from Agilent J&W was programmed at 75 °C, held for 3 minutes, then increased at a rate of 3 °C/per minute to 140 °C, held for 3 minutes, then increased at 3 °C/per minute to 200 °C, held for a 1 minute and increased 5 °C/per minute to 230 °C. The ion polarity for all MS scans was positive, with an EMV mode of 1.0 (or 2448 V) gain factor; acquisition mode was set to scan at a normal speed.

Analytical triplicates were carried out for all the samples analysed. All recorded instrumental blanks were made with acetonitrile (MS grade). A sample blank was taken by subjecting the solvent (i.e. MeCN) to the sample preparation procedure post-derivatisation. The compounds were identified by a match in the MS fragmentation pattern to a pre-existing compound within the NIST database (i.e. includes all known amino acids, and other small organic compounds known to be present in the Miller-Urey product mixture) or an amino acid-standard. The amino acid standards were prepared from 50 mM stock-solutions, see **Figure A6**. Validation of the compounds identified in the samples by means of an amino acid standard was done through external validation, were a match in retention time and the resulting mass-spectral pattern (with the pure standard) was required for confirmation.

The NIST 14 database search included (within its libraries) compounds that have been derivatized with a silylation reaction, which made the mass-spectral matching even easier. The database search was enabled through the Chemstation® Data Analysis software, available from the Agilent GC-MSD vendor, see **Figure 87**. The commercial mass spectral libraries can be searched using a probability-based matching algorithm (PBM) included with

the software, compiling a maximum of 3 libraries to be searched at once. The PBM algorithm uses a reverse search to verify that peaks in the reference spectrum are present in the unknown spectrum. Consequently, the extra peaks in the unknown are also ignored, thus allowing the analysis to be carried out in short-periods of time (i.e. in a condensed manner) and the capacity of tackling spectrum resulting from a mixture of compounds (due to poor separation of the analytes). This aspect can roughly explain why some of the confirmed peaks retrieved such low quality percentages (meaning a high level of uncertainty within the match) or why unreasonable matches were confirmed. Nonetheless, since not all mass-to-charge (m/z) values of a mass spectrum are equally likely to occur, the PBM algorithm uses both the mass and abundance values to identify the most significant peaks in a reference spectrum. This is also combined with a pre-filter within the search routine, which assigns a significance to each of the peaks in the unknown spectrum and uses these to find the most probable matches in the 'condensed' reference library. The selected (condensed) spectra are then compared using the reverse search described above, with the complete unknown spectrum. In addition, all the chromatograms were analysed in a qualitative manner and no quantitative method was developed (or appended).

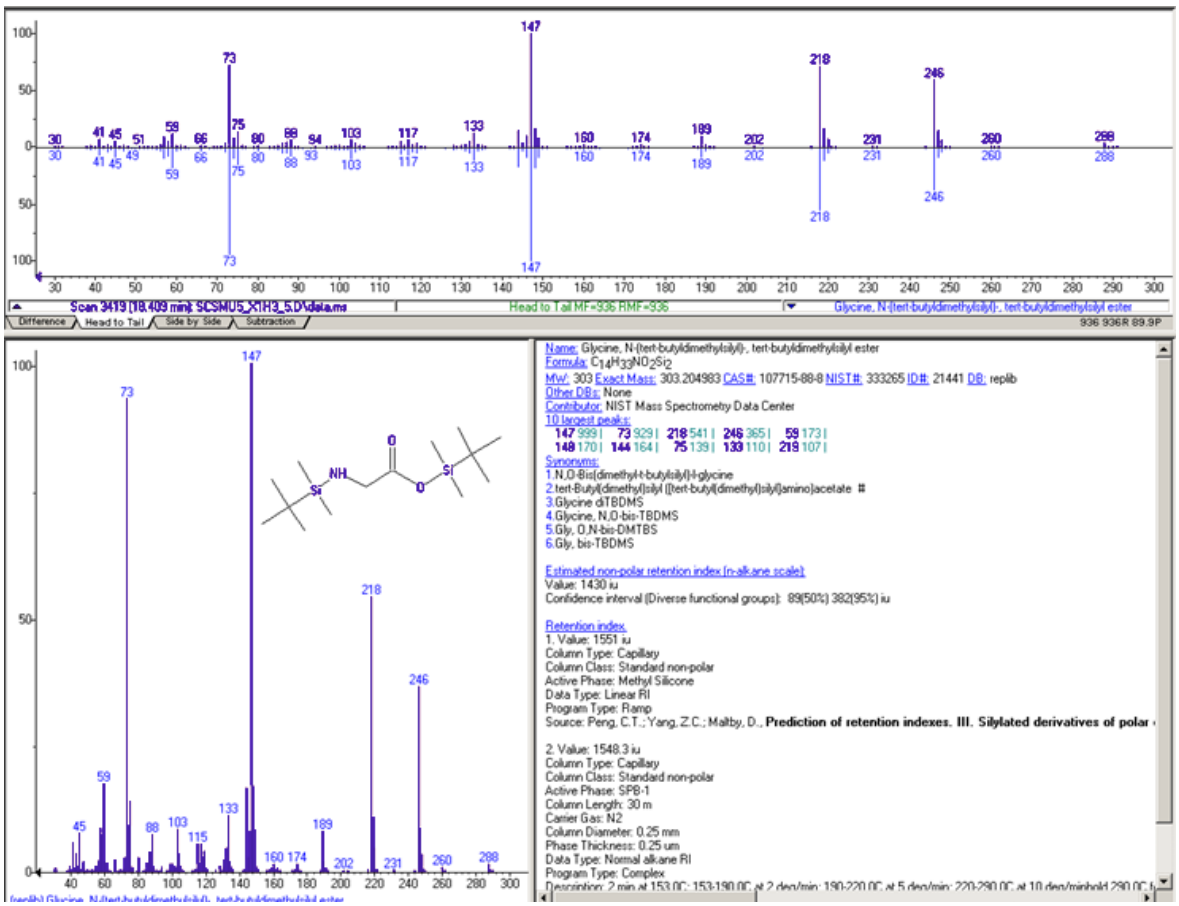


Figure 87. NIST 14 software for integrated data-base search of EI mass-spectral pattern, a comparison is made from the experimental sample (purple) and the reference data file (blue). It represents a match for the amino-acid Glycine (89% quality).

4.1.5 HPLC-FLD method

The HPLC-FLD analysis was performed using an Agilent 1200 HPLC system / FLD detector, based on a standard protocol method for the analysis of amino acids, involving derivatisation of amine groups with o-phthalaldehyde (OPA)/mercaptopropionic acid (MPA) to allow retention of the products on a reverse phase column (Agilent Poroshell 120 HPH C18, 3.0 x 100 mm, 2.7 µm) and detection using a fluorescence detector (excitation at 340 nm, emission detected at 450 nm).[227] For the chromatographic separation of the products, a reverse phase column by Agilent (Poroshell 120 HPH C18, 3.0 x 100 mm, 2.7 µm) was selected. The samples were injected in 10 µL aliquots and eluted with a linear gradient mixture of solvents A (water w/0.1% v/v formic acid) and B (100% acetonitrile w/0.1% v/v formic acid) at 1.0 mL per minute, over 21 mins as follows: 0 min – 100% A; 3 min – 100% A; 13 min – 100% B; 15 min – 100% B; 18 min – 100% A. The column compartment was maintained at 30 °C. The fluorescence detector was set to an excitation wavelength of 340 nm and emission detected at 450 nm.

The instrument was controlled and data acquired using Agilent OpenLab software. Three identical analyses were recorded for each experimentally-produced sample; in addition, a series of standards of products identified in previous spark discharge experiments were analysed for comparison as well as a means to confirm the absence of significant retention time (*rt*) drift; see **Figure 88**. The same software was used to detect and integrate all significant peaks, and extract corresponding intensities in all runs (+/- 2% retention time 'window'; also checked manually).

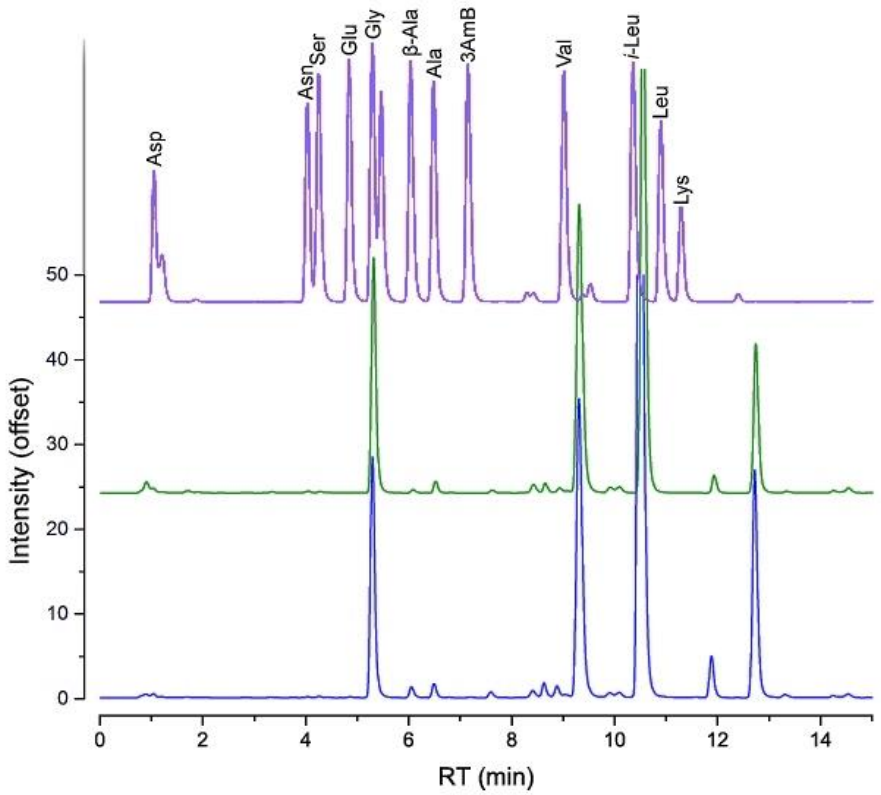


Figure 88. HPLC-FLD plots of amino acid standards (*purple*, top), in a representative deuterated run (*green*, middle) and a representative non-deuterated run *(blue*, bottom), showing a positive identification of glycine and tentative identification of alanine and β-alanine. Standards of aspartic acid, asparagine, serine, glutamine, glycine, β-alanine, alanine, 3-amino butyric acid, valine, iso-leucine, leucine and lysine were made up at 2.5mM in HPLC grade water and carried out with the help of Dr. Andrew Surman.

4.1.6 Elemental Analysis and SEM-EDS

The samples were lyophilized without any prior centrifugation or filtration. Then, ~0.5.grams of dried material was transferred into glass vials (8mL) and handed in for analysis. Elemental analysis was carried out by Mr. Jim Tweedie in the University of Glasgow (analytical service). The SEM-EDS analysis was conducted by Dr. Stefan Glatzel in the Summer of 2016.

4.1.7 Principal Component Analysis

The resulting data sets for the GC-MS and HPLC-FLD analysis of the (deuterated or 'classic') Miller-Urey experiment, implemented as outlined in the sections above, were subjected to simple PCA with a scaling of the raw data, using the FactoMineR library in R.[228] For the GC-MS data, multiplative signal correction was applied before the PCA analysis.[229] Also, all ellipses were drawn to highlight the grouping of the samples (i.e. done by eye) and do not represent 'confidence ellipses'. The PCA analysis was carried out by Dr. Piotr Gromski, after I provided the data-sets in a CSV format.

4.2 Formose reaction in Formamide

4.2.1 Reagents

Formaldehyde (ASG reagent, 37% wt. in H_2O), glycoaldehyde (97%), formamide (Reagent Plus®, >99.0 (GC)), calcium carbonate (purity >96%) and formic acid (reagent grade, >95%) were purchased from Sigma Aldrich. Calcium hydroxide (purity > 96.0%) was purchased from Fluka Analytical. Analytical solvents (water and acetonitrile, HPLC-MS grade) and ammonium acetate (Ambion® Molecular Biological Grade (5M), >98%) were purchased from ThermoFisher Scientific UK. Analytical standards of hexamethylenetetramine (HMT), ribose, adenine, guanine, thymine, cytosine, uracil. Adenosine and thymidine were purchased from Tokyo Chemical Industry UK. Amberlite™ (cation) Ion-exchange resin was purchased from Sigma-Aldrich.

4.2.2 Minerals

Goethite, α-FeO(OH), montmorillonite, $(Na, Ca)_{0.33}(Al, Mg)_2(Si_4O_{10})$ and hydroxyapatite, $Ca_5(OH)(PO_4)_3$ were purchased from Sigma-Aldrich. Chalcopyrite, $CuFeS_2$ was purchased from Alpha-Aesar by Thermo Fisher Scientific. Ulexite, $NaCaB_5O_6(OH)_6{\cdot}5H_2O$, zoesite, $Ca_2Al_3(SiO_2)_3(OH)$ and quartz, SiO_2 were purchased from Richard Taylor Minerals, a private collection from the United Kingdom.

The minerals acquired in the private collection, quartz and zoesite where not pre-treated as to remove organics. However, only 2 of the 7 minerals are natural, the rest are synthetically made and purchased from SigmaAldrich and AlphaAesar.

4.2.3 Experimental procedure and Sample preparation

(a) *Recursive Cycles:* A mixture of formaldehyde (0.5 mL), glycolaldehyde (0.0126 g), water (2.25 mL), formamide (2.25 mL) and calcium hydroxide (0.0705 g) was prepared on seven different mineral surfaces (1 mg each) in 22 mL borosilicate glass vials. It was stirred at 1200rpm with a magnetic stirrer and heated at 50°C, for 48 hours. Then, about 70% of the reaction volume (supernatant) was removed for analysis. The remaining fraction was used to seed the next reaction. The reaction was then topped up to the initial volume with a solution of starting material at the same concentration. This process was repeated three times.

(b) *Long cycles:* At the end of the 3rd recursive cycle, the reaction vessel was replenished with a mixture of formaldehyde (0.5 mL), glycolaldehyde (0.0126 g), water (2.25 mL), formamide (2.25 mL) and calcium hydroxide (0.0705 g). This was done after removing 70% of the reaction volume, exactly like with previous cycles. The reaction was then stirred at 1200rpm with a magnetic stirrer and heated at 50°C, for the total length of three cycles: 144 hours = 6 days.

(c) *Non-recursive controls:* Formaldehyde (0.5 mL), glycolaldehyde (0.0126 g), water (2.25 mL), formamide (2.25 mL) and calcium hydroxide (0.0705 g) was carried out in 22mL borosilicate glass vials. Reactions were stirred at 1200rpm with a magnetic stirrer and heated at 50°C and allowed to react for the equivalent time that it took to carry out the three recursive cycles (6 days). The non-recursive samples included a non-mineral control, as well

as the same selection of mineral surfaces (7, 1 mg each) employed in the recursive experiment.

(d) *Reproducibility assessment:* (**1d**) Non –recursive mineral control - water (2.50 mL) and formamide (2.50 mL) mixture (1:1 v/v) in the presence of the mineral chalcopyrite (1 mg) was carried out in 22 mL borosilicate glass vials. Reactions were stirred at 1200rpm with a magnetic stirrer and heated at 50°C and allowed to react for the equivalent time of 5 recursive cycles (i.e. 10 days) (**2d**) Non-recursive reaction - formaldehyde (0.5 mL), glycolaldehyde (0.0126 g), water (2.25 mL), formamide (2.25 mL), calcium hydroxide (0.0705g) was carried out in 22 mL borosilicate glass vials. Reactions were stirred at 1200rpm with a magnetic stirrer and heated at 50°C and allowed to react for the equivalent time of 5 recursive cycles (i.e. 10 days). (**3d**) Recursive reaction - formaldehyde (0.5 mL), glycolaldehyde (0.0126 g), water (2.25 mL), formamide (2.25 mL) and calcium hydroxide (0.0705 g) were added to a 22mL borosilicate glass vials, in the presence of the mineral chalcopyrite. It was stirred at 1200rpm with a magnetic stirrer and heated at 50°C, for 48 hours (2 days). Then, about 70% of the reaction volume (supernatant) was removed for analysis. The remaining fraction was used to seed the next reaction. Topping up with the same concentration of starting materials, but conserving the total reaction volume; we repeated the process five times. (**4d**) Recursive mineral reaction - formaldehyde (0.5 mL), glycolaldehyde (0.0126 g), water (2.25 mL), formamide (2.25mL) and calcium hydroxide (0.0705g) was carried out in the presence of chalcopyrite (1 mg) in 22mL borosilicate glass vials. It was stirred at 1200rpm with a magnetic stirrer and heated at 50°C, for 48 hours (2 days). Then, about 70% of the reaction volume (supernatant) was removed for analysis. The remaining fraction was used to seed the next reaction. Topping up with the same concentration of starting materials, but conserving the total reaction volume; we repeated the process five times. It must be noted that the mineral surface selected, chalcopyrite, was discontinued from Alpha-Aesar by the time we carried out this assessment. Therefore, we have used a natural source of chalcopyrite, acquired from Richard Taylor Minerals. In order to use it, we have ground the mineral with a ball-mill and then used sieved fractions [300 μm-2 mm] to constrain the size of the resulting fragments. We addressed the possible interference of contaminants by carrying out the mineral control reaction (**1d**), see **Figure A11**.

(e) *Formose reaction:* Formaldehyde (0.5 mL), glycolaldehyde (0.0126 g), water (5 mL) and calcium hydroxide (0.0705g) was carried out on seven different mineral surfaces (1 mg) plus a control in 22mL borosilicate glass vials. It was stirred at 1200rpm with a magnetic stirrer

and heated at 50°C, for 48 hours.

(f) *Formamide condensation:* Formaldehyde (0.5 mL), glycolaldehyde (0.0126 g), formamide (5 mL) and calcium hydroxide (0.0705g) was carried out on seven (7) different mineral surfaces (1 mg) plus control in 22mL borosilicate glass vials. It was stirred at 1200rpm with a magnetic stirrer and heated at 50°C, for 48 hours.

4.2.4 UPLC-MS/MS analysis

Ultra-Performance liquid chromatography and tandem mass spectrometry (UPLC-MS/MS) analysis was performed with a Thermo Vanquish Ultra-performance liquid chromatography system coupled to a Thermo Orbitrap Fusion Mass-Spectrometer. Samples were injected directly (no splitting) in 10 µl aliquots and chromatographic separation was achieved with a amide-HILIC C18 (4.6 × 150 mm, 2.7 µm) column, eluted in a linear gradient mixture of solvents A (water w/20 mM ammonium acetate, pH = 5) and B (100% acetonitrile w/0.1% v/v formic acid) over 25 min as follows: 0 min, 100% A; 4min, 100% A; 19 min, 100% B; 23 min, 100% A; 25min, 100% A; in a method adjusted from Idborg, *et al.*[209] The column was maintained at 30 °C and the MS spectra was collected for 30 minutes in positive mode over a scan range of 50–500 m/z. Ion transfer tube was set to 275 °C, RF lens 60%, and acquisition was performed in a data-dependent (DDA) manner.

The Data-Dependent Acquisition (DDA) was performed by prioritizing the top most intense fragments in a 3 second window with an intensity threshold of 5.0E4 and dynamic exclusion, after one time for 15 seconds (in order to avoid the selection of the same fragments), using the ion trap isolation with a HCD collision energy of 35 eV and a resolution 15000. In order to minimize the risk of saturating the detector and avoid false positives during the re-equilibration step the eluted material was set to waste for the first and last five minutes of the chromatographic run.

4.2.4.1 Sample preparation

For all cycles, the removed fraction was allowed to cool to room temperature, then a 1000µL aliquot was taken for analysis. Followed by removal of excess cations in solution (i.e. Ca^{2+}) with Amberlite™ Ion-exchange resin, before the supernatant was diluted 1 in a 100 with MS grade water. Finally, the solution was filtrated with a syringe filter (0.22µm cut-off) and

placed in an HPLC sample vial, before analysis.

4.2.4.2 Method development: UPLC-CAD

A Thermo-Dionex UltiMate3000 UPLC system equipped with a Charged Aerosol Detector (CAD) was used for method development, as a complementary analytical technique to compensate for the limitations/bias of other detectors, such as UV/Vis –DAD- and Fluorescence. Primarily due to the limited amount of instrument time available in the HR-MS. However, it's the UPLC-HR-MS (Orbitrap-Lumos) analysis that allowed for the plausible identification of the unknown compounds. A methodology described by Idborg *et al.* for HPLC analysis with HILIC column chromatography was adapted for the Formose-in-Formamide samples.[209] The elution method was executed in the linear gradient described, but elongated for 10 more minutes. The column was maintained at 30 °C and the CAD detector was set to a nominal evaporator temperature of 65.0 °C (+/- 5.0 °C). A selection of sugar (mannose, ribose, fructose, glucose and sucrose, see **Figure 89**) and nucleobase (adenine, cytosine and thymine) standard were prepared, as a way to identify the peaks. Also, a nucleoside standard of Thymidine was also made and analysed through the chromatographic method. A set of HILIC columns were tested: a ZIC – HILIC and amide-HILIC column (*ThermoScientific*). A comparison between the columns was performed by making a standard solution (mixture) composed of all five (5) sugar standards and assessing the differences in their resulting chromatograms, when eluted with the same (gradient) method. The separation and sharpness of the peaks was superior with the amide-HILIC column and therefore it was selected for the UPLC-MS/MS analysis.

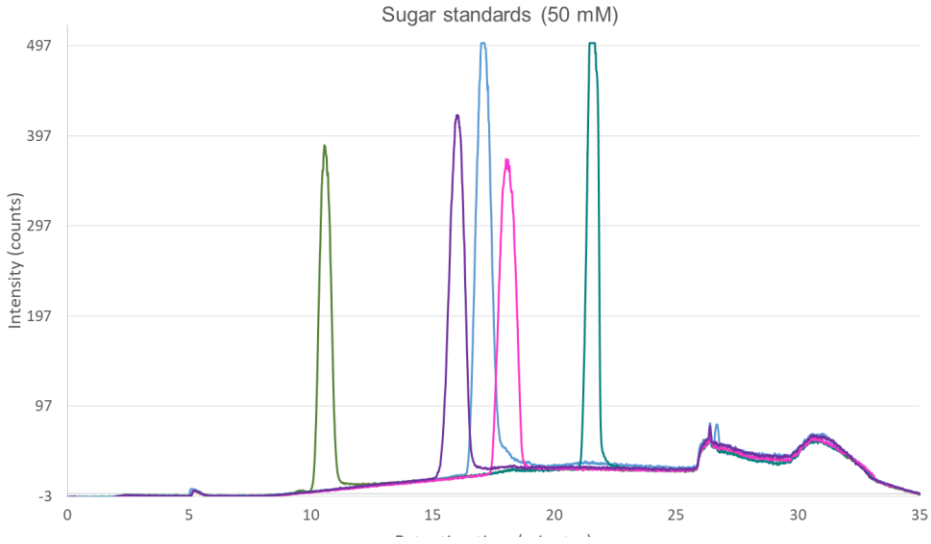


Figure 89. UPLC-CAD chromatogram of the standard mixture of sugars, elution profile proceeds in the following fashion: Mannose-Ribose-Fructose-Glucose-Sucrose

4.2.4.3 Compound identification and validation

Identification of the compounds in the reaction mixture was performed by ThermoScientific™ Compound Discoverer 2.0 [230] by matching the exact mass and the resulting MS^2 spectra with the all the available libraries, through MZcloud® or ChemSpider database search (**Figure 90**). The identity of the compounds was further validated with pure standards, where a match in retention time, exact mass and a robust correlation with the MS/MS mass-spectral pattern was observed. Additionally, further validation was performed manually through ThermoScientific™ Mass Frontier™ spectral interpretation software.

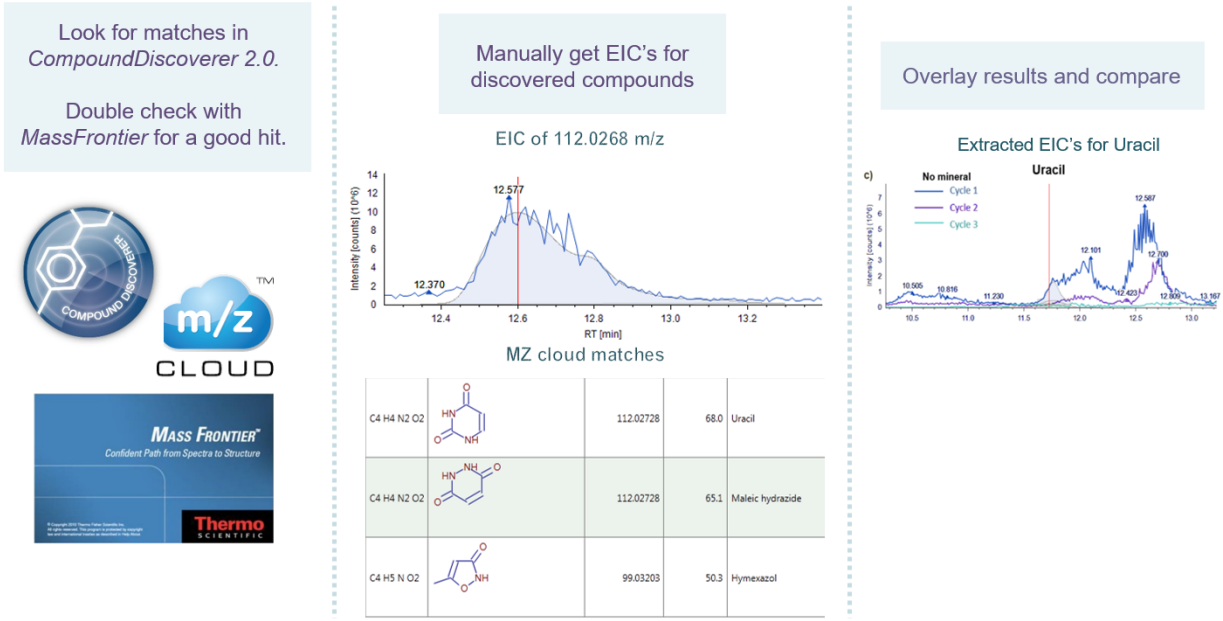


Figure 90. Identification of unknown compounds through database search within the Compound Discoverer and Mass Frontier data processing software. Extracted Ion Chromatograms (EICs) are also generated within the CompoundDiscoverer workflow.

The CompoundDiscoverer™ (Thermo Scientific) software is designed to enable the identification of small-organic compounds from the UPLC-MS/MS analysis of complex chemical mixtures. It contains a selection of workflows, which integrate different methods for data extraction, including retention time alignments and statistical analysis of the mass spectral features. The workflow selected for this analysis was 'Untargeted metabolomics workflow with statistics and ID using mzCloud and Chemspider', see **Figure 91**. The step-by-step procedure goes as follows: (*1*) Alignment of retention times, (*2*) Unknown compounds detection, (*3*) Grouping of unknown compounds, (*4*) MzCloud search, (*5*) Composition prediction, (*6*) ChemSpider search, (*7*) Gap filling and (*8*) Marking of

background compounds. This workflow also allows the user to set a range of accepted amount for each element, as well as, which adducts to take into consideration. In addition to establishing the thresholds for peak intensity (min. 2.0 E06), signal to noise ratio (3) and mass tolerance (1 – 20 ppm).

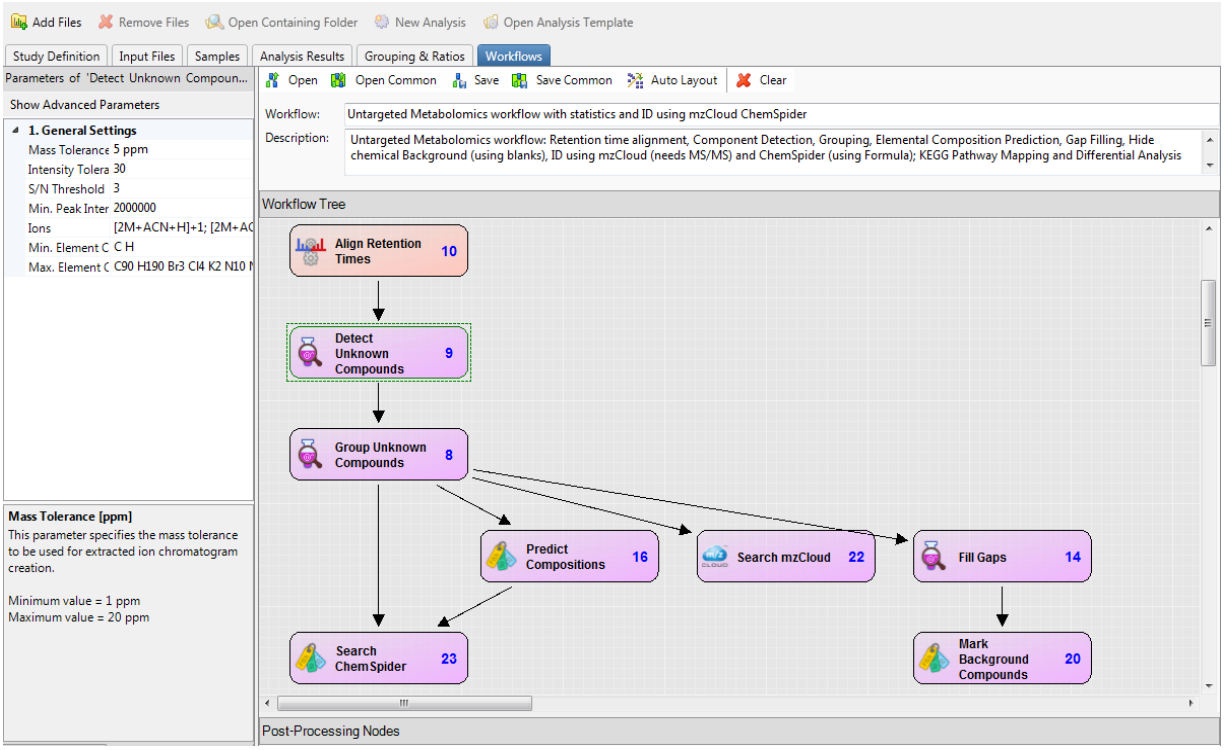


Figure 91. *General overview of the processing workflow in CompoundDiscoverer 2.0™:* The integrated data analysis workflow does adduct calculations, predicts compositions and conducts a search through the MZcloud® and ChemSpider databases.

The aforementioned workflow enabled us to detect features confidently, since it takes into consideration multiple adducts, besides chromatographic alignments and corrections. Identified compounds for the detected features were done by doing a comparison of the resulting MS/MS pattern with a reference spectrum of any compound present in the searched libraries. The databases used in the workflow differed in how they are built, with Chemspider constructed from experimental data and the mzCloud based through in-silico calculations. Consequently, the matches were not equal but complementary in most cases. Nonetheless, to validate the identified features confidently, the resulting MS/MS (i.e. MS^2) pattern was also compared with the one obtained by a pure standard (see **Figure 92**).

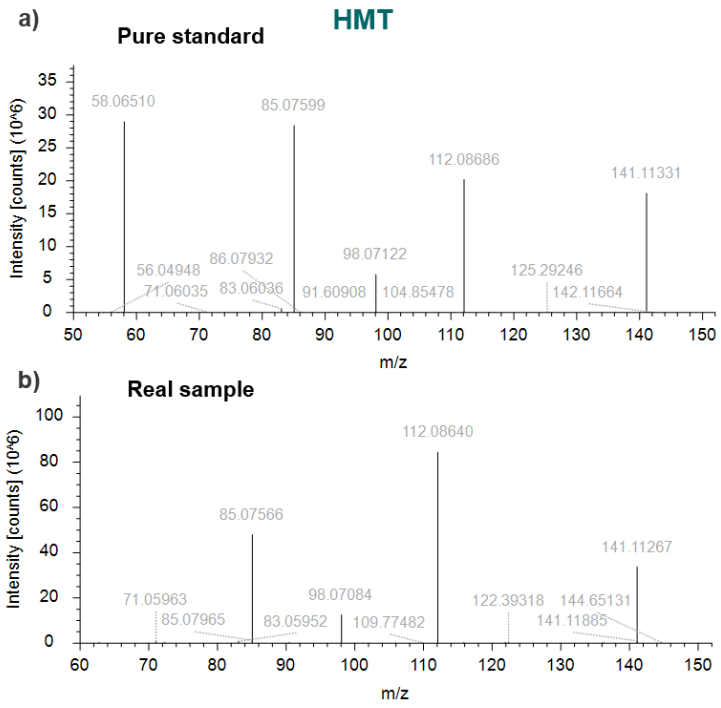


Figure 92. UPLC-MS/MS spectrum for hexamethylenetetramine, HMT (m/z: 141.11, Adduct: [M+H]) in (**a**) a pure standard and (**b**) a real sample, (chalcopyrite, Cycle 3)

The ThermoScientific™ Mass Frontier™ allowed for a complementary validation of the identified features. Through this software, a selected precursor ion and its resulting MS/MS pattern can be searched against multiple other databases (e.g. HighChem ESI pos and NIST, amongst others) through its own server manager. Also, integrated substructure annotation for each fragment is enabled within the software (FiSH), giving an insight to the chemical compositions behind every fragment that constructs the resulting MS/MS spectra, see **Figure** A15. In order to do this, it automatically generates all possible fragments following a FragmentationLibrary™ or general fragmentation rules, for each precursor ion (feature).

Figure 94 and **Figure 93**, show two examples of compounds (e.g. uracil and ribose) validated by external standard, whose resulting MS/MS spectra was compared using the MassFrontier software. This was necessary, since the corresponding features were not picked up by the Compound Discoverer workflow and would have been missed otherwise. For this reasons, we found it to be a useful tool to complement the CompoundDiscoverer® results and identification process.

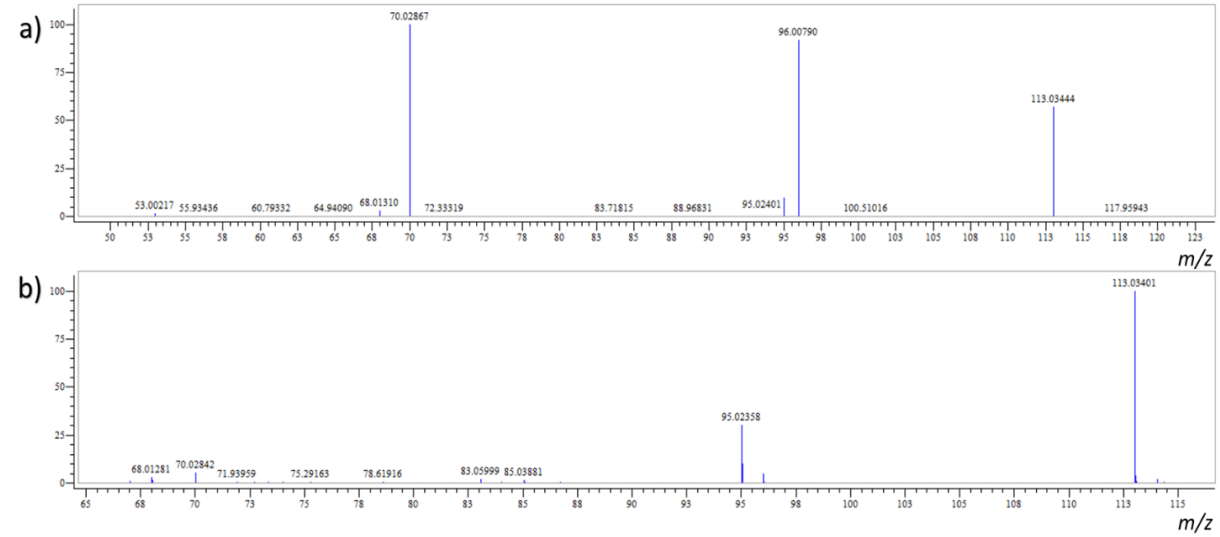


Figure 94. UPLC-MS/MS spectrum for uracil (*m/z:* 113.03, Adduct: [M+H]) in (**a**) a pure standard and (**b**) a real sample, (chalcopyrite, Cycle 3).

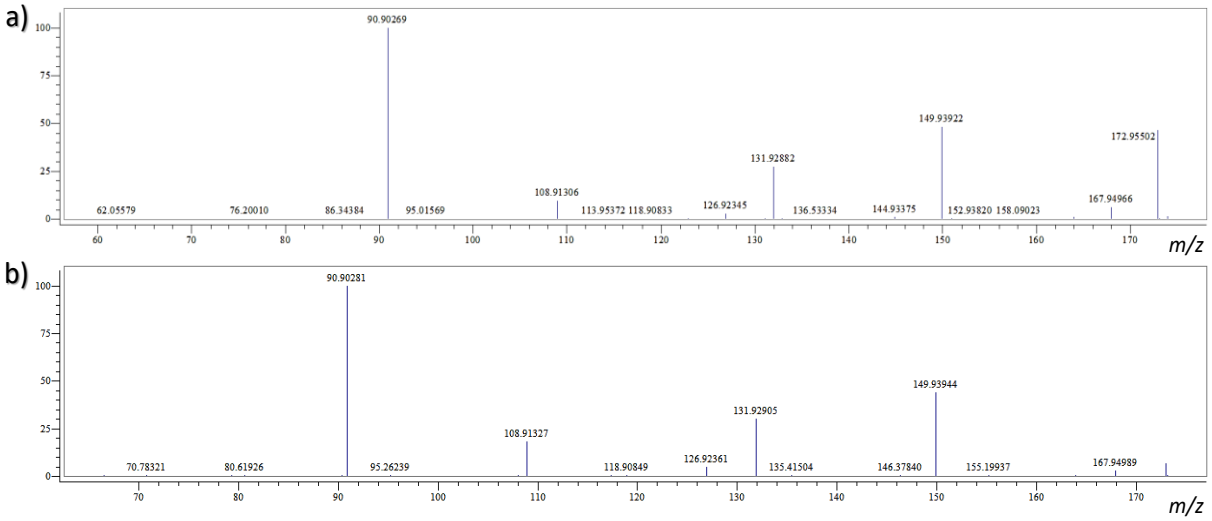


Figure 93. UPLC-MS/MS spectrum for ribose (*m/z*: 172.96, Adduct: [M+Na]) in (**a**) a pure standard and (**b**) real sample (chalcopyrite, Cycle 3).

Regardless the limitations of the Compound Discoverer workflow, Extracted Ion Chromatograms (EICs) of the identified products (as well as, other unknown features) are easily accessible through the User-Interface. An example of the processing workflow interface can be seen in **Figure A15**. Moreover, the software also appends the information of each feature across multiple samples, allowing for a direct comparison of the resulting intensities, see **Figure 95**. These intensities are not representative of the absolute concentration in solution, but can give rough indication of the relative concentration for a particular feature (assuming that the matrix complexity is comparable) . Through this assumption, we have done a qualitative analysis of the differences across the intensity of specific features.

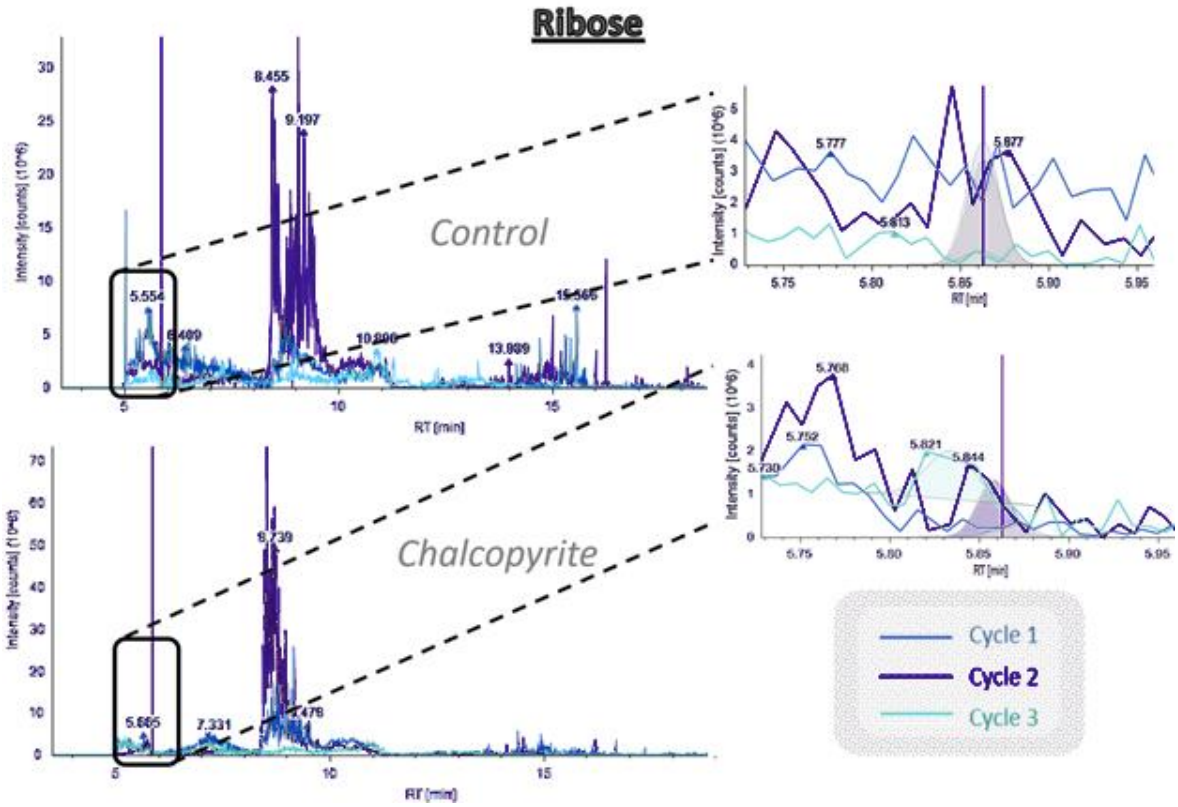


Figure 95. . UPLC-MS/MS spectrum for ribose (*m/z:* 172.92, Adduct: [M+Na]) in (**a**) a pure standard and (**b**) a real sample, (chalcopyrite, Cycle 3).

4.2.4.4 Detection of nucleosides (traces)

Traces of adenosine and thymidine nucleosides where found in the last cycle (3) of most samples (including the control reaction). The presence of ribose and several nucleobases in Cycle 1 and 2, led us to believe there was a possibility for nucleoside formation in Cycle 3. To address this, we looked for the corresponding nucleosides of adenine and thymine, in Cycle 3. **Figure 96** illustrates the cross validation with an external standard of Thymidine, by having an acceptable chromatographic match in retention time (+/-40s), the same exact mass and a matching MS/MS (MS^2) fragmentation pattern. The same criteria was extended for the (external) standard validation of adenosine, as seen in **Figure A17**.

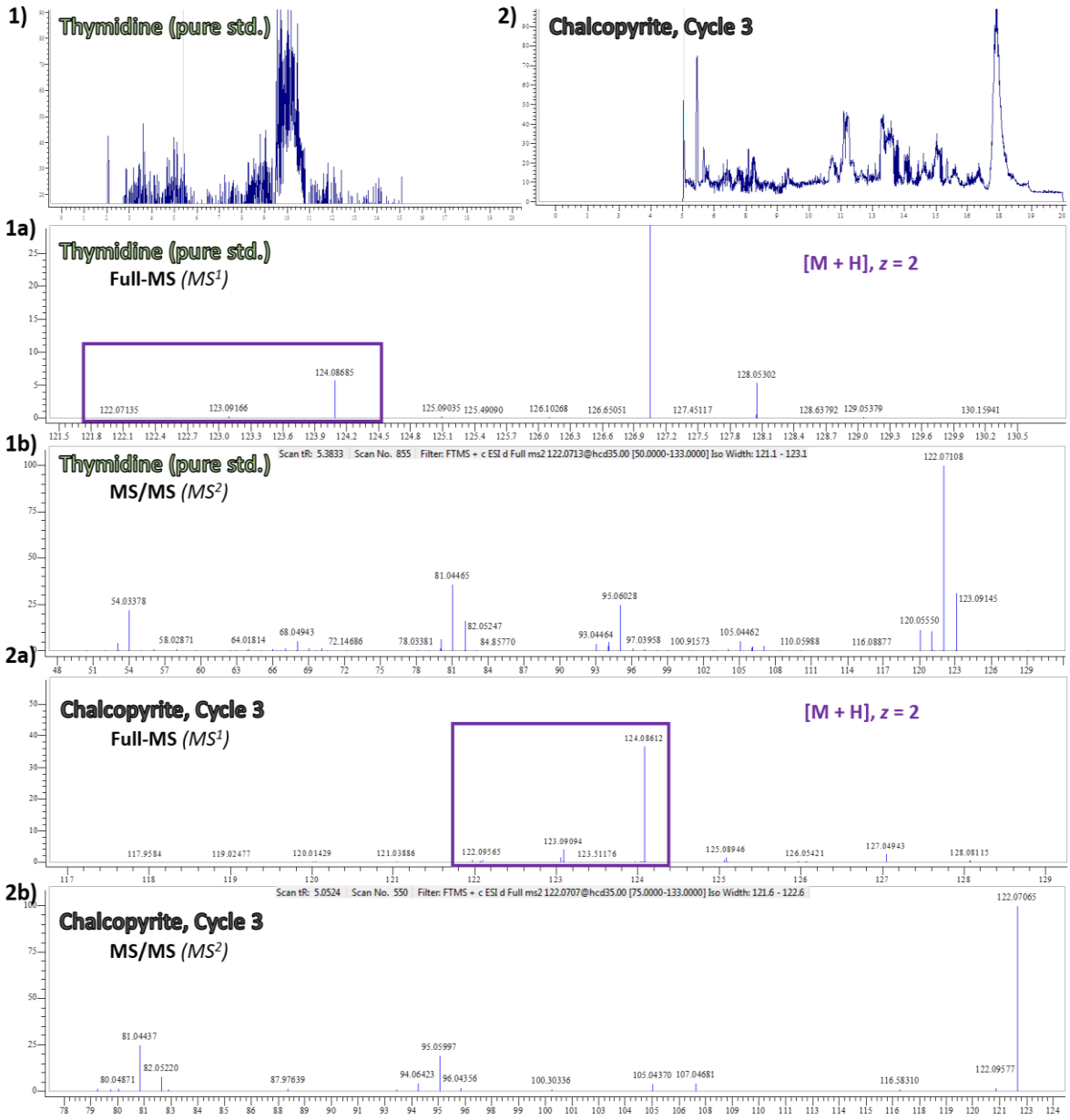


Figure 96. UPLC-MS/MS analysis and comparison of (**1a-1b**) a pure standard of thymidine (m/z: 122.07), Adduct: [M+H] z = 2, with (**2a-2b**) a real sample (chalcopyrite, Cycle 3).

Furthermore, by running the pure standards through the same chromatographic method, we found that preferred adducts for Thymidine were not necessarily the $[M+H]^+$, but rather the charged (z = 2, for m/z) and sodium adducts predominated (see **Figure 97**). The exact isomer of the nucleosides was not assessed, since the pure standards used for validation were not differentiated by their isomeric position. However, we believe this to be satisfactory evidence towards the presence of nucleosides in the product distribution ensemble; *particularly*, within a mixture of this complexity.

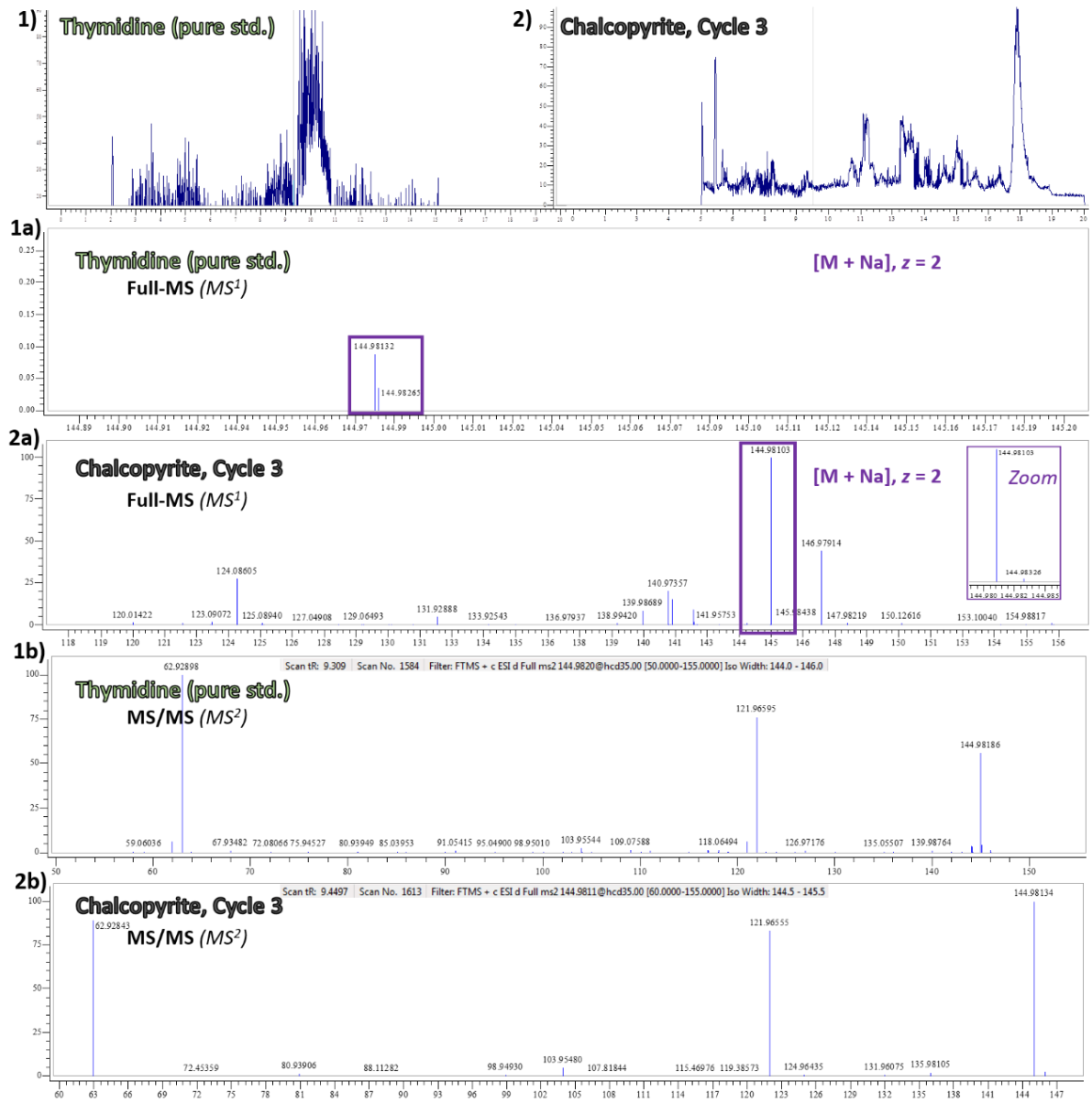


Figure 97. UPLC-MS/MS analysis and comparison of (**1a-1b**) a pure standard of thymidine (m/z: 144.98), Adduct: [M+Na] z = 2, with (**2a-2b**) a real sample (chalcopyrite, Cycle 3).

4.2.4.5 Feature Generation and m/z distribution

The raw data was extracted from its vendor format (.raw) with MSConvert from ProteoWizard [2], into an .mzML format before introducing it to Python. The converter allowed for a vendor-specific algorithm to perform peak-picking of the raw data and centroidization, as a way to extract all the relevant information. The package PymzML [3] was used to extract MS^1 values, Retention Time (RT), Intensity (in counts) and MS^2 values (with their corresponding RT and Intensity). Also, the extractor was set to have a measured precision of 1E-04 for both levels: MS^1 and MS^2.

All detected MS^1 values where then filtered by selecting those which were taken for MS^2 fragmentation by the DDA method, which we found were representative of the most important features within the complex mixture. In addition, all m/z values were truncated to their second decimal, as a way to account for the instrument uncertainty associated to its calibration status. As well, all retention time values were truncated to the second decimal. Finally, the number of MS^2 fragments was counted and appended to their corresponding feature. This was done as a way to (roughly) gauge the molecular complexity of the features, if we can assume that branched and multi-element compounds will have a higher number of fragments. Similarly, it can be used as a way to differentiate between the unknown features, giving them an extra-dimension based on their chemical composition (e.g. MS/MS patterns are unique for each chemical compound, even if it does not account for isotopologues).

The features generated are based on unique Retention Time (RT), Exact Mass (m/z) and MS^2 fragments were plotted using python 'matplotlib' library as scatter-plots, in order to enable their visualization. In such plots (see **Figure 98**) each point ideally represents a compound within the product mixture, giving a representation of the reaction products, even if we cannot identify each one of the compounds.

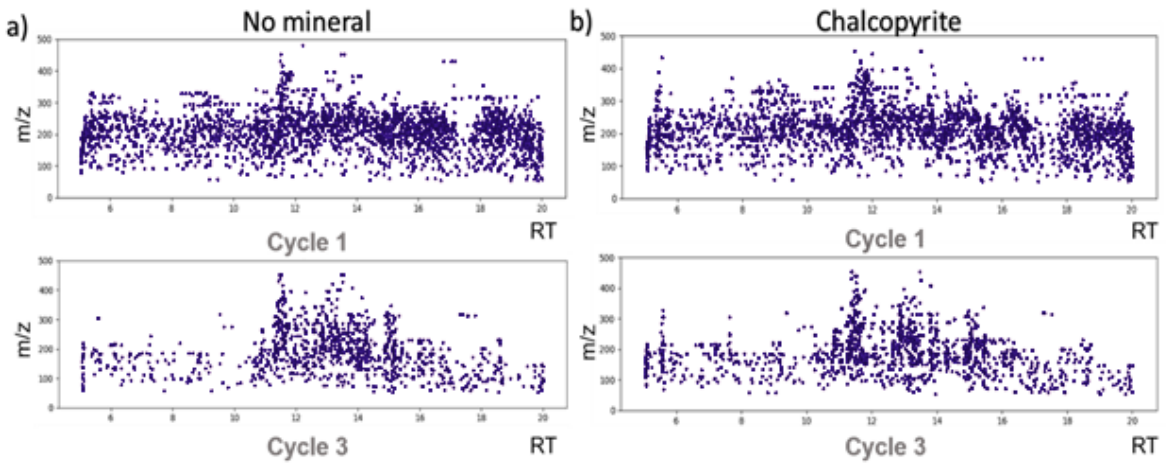


Figure 98. The feature generation based on unique retention time (RT) and *m/z* combinations for Cycle 1 and Cycle 3, in the reaction control (**a**) and in the presence of the mineral chalcopyrite (**b**). The plots are generated for all the detected features, *prior* to background subtraction (e.g. instrument blanks) and removal of duplicates features.

Duplicate (MS^1) values where further filtered, by eliminating values that had same exact mass (to the second decimal value) and eluted within an acceptable retention time window (+/- 30 s) of each other. Furthermore, the number of MS^2 fragments obtained for each feature

(MS^1) was calculated, as a generic way to access the overall complexity of the molecules within the complex mixture. Isotopic peaks and noise peaks might have been included in this process, but the validation of 'real' MS^2 fragments was beyond the scope of this study.

As seen in **Figure 99**, the changes in the product distribution were consistent with the features generated in CompoundDiscoverer™ (Thermo Scientific). The number of features generated by the software's processing method where far less than the ones generated by our in-house feature extraction method, but preserved the same trend. This comparison provided a good validation of the bespoke feature generator and our 'omics' based approach to complex mixture analysis.

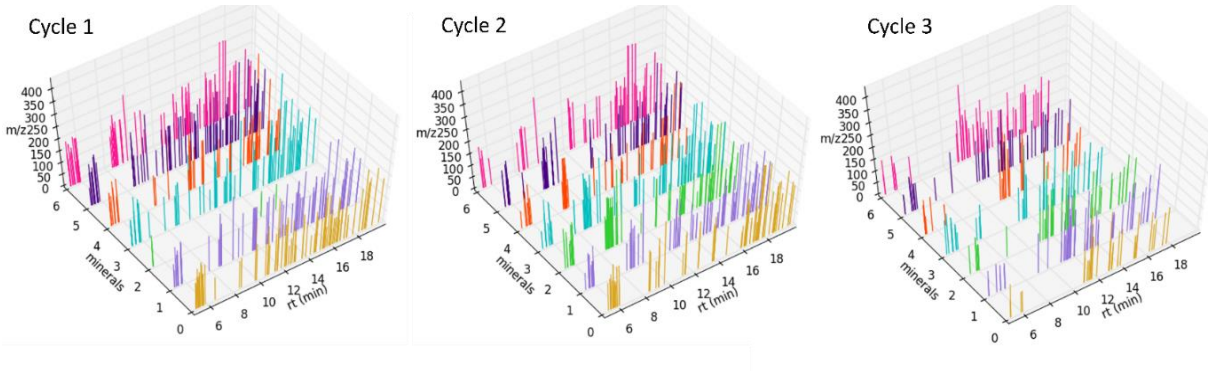


Figure 99. *Multi-colour plots:* Differences in the feature distribution are displayed across mineral surfaces (***1-6***) and the reaction with no mineral (***0***) and the number of detected features reduces over the recursive cycles, in features extracted by the CompoundDiscoverer™ software suite.

Additionally, we have calculated the difference between the number of MS^2 fragments across consecutive cycles for the recursive reaction with chalcopyrite and in the absence of a mineral surface (see **Table 5**). This worked as an empirical validation of the trend, since the number of MS^2 fragments (across all detected features in each cycle) does effectively reduce over recursive cycles.

Recursive cycles (no mineral)	Difference of MS^2 fragments	Recursive cycles (with chalcopyrite)	Difference of MS^2 fragments
C1 – C2	44627	**C1 – C2**	16542
C2 – C3	101321	**C2 – C3**	86579

Table 5. Difference between the number of MS^2 fragments across recursive cycles (1 to 3), for a recursive reaction in the absence of any mineral and in the presence of chalcopyrite.

Furthermore, the filtered features were manually grouped (in Excel) into 50m/z bins, for a range of 50 to 400 m/z, as seen in **Figure 64**. The number of features detected for each bin were conventionally displayed in a heatmap, generated in matplotlib through Python. The heatmap produces a unique pattern for each reaction, in which the number of features decreases over recursive cycles consistently, in the absence *and* presence of a mineral surface. An additional display of m/z distribution of the detected features can be seen below in **Figure 100.**

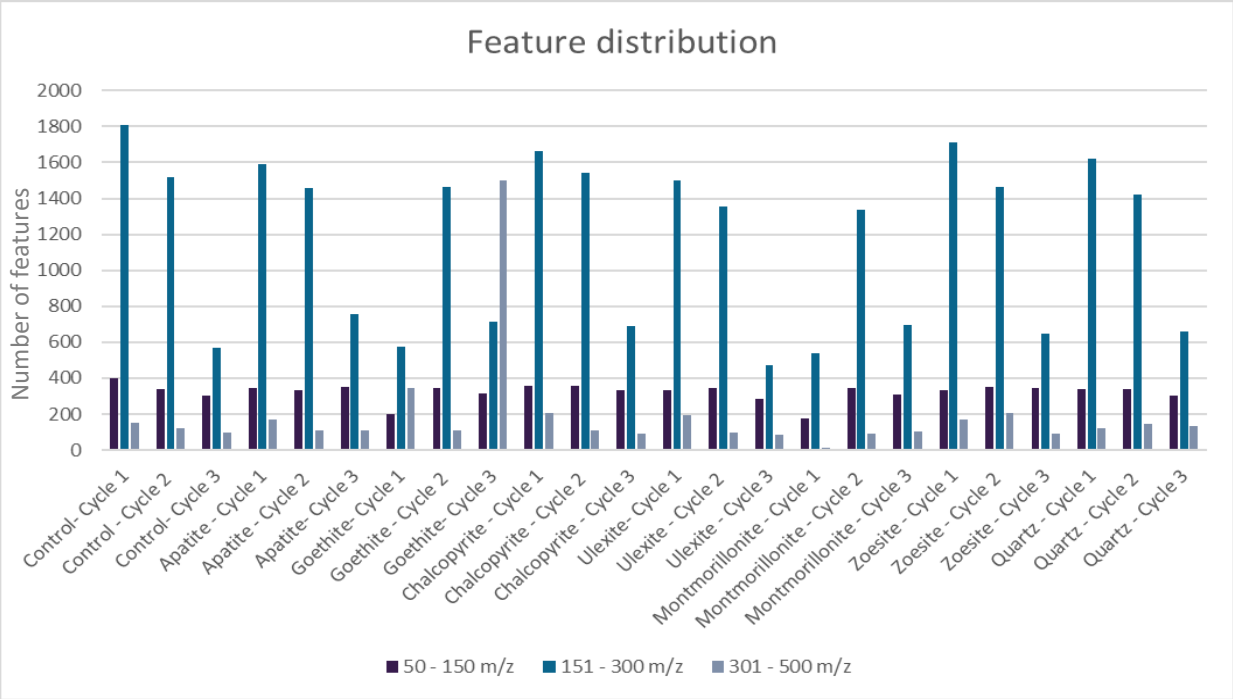


Figure 100. *Feature distribution by m/z:* A histogram of the features detected for each reaction mineral environment over the three recursive cycles, generated by grouping the features into m/z bins.

4.2.5 Feature Analysis

4.2.5.1 PCA

The PCA analysis was initially carried out with XCMS Online, a web-based platform developed for the batch processing of untargeted metabolomics data. The XCMS Online was created as a user-friendly-interface for the XCMS software, which is widely used for UPLC-MS/MS data processing but requires an entry-level training by the providers or a good level of programming knowledge (i.e. R, Python or C). The data processing workflow was constructed aiming to automate the multiple-steps needed for data deconvolution and therefore, generate a 'processing' standard across different laboratories executing similar experiments (instrument-wise). All data files are uploaded in their .RAW format and

analysed in batch by grouping the samples according to their cycle number or mineral type. The data is then processed through the following steps: (1) *peak detection algorithm* - uses 0.1 m/z bins to generate Extracted Ion Base Peak Chromatograms (EIBPC) and are consequently filtered by a Second Derivative Gaussian that generates zero-crossing points (in the X-axis) which define the area of peak integration (2) *peak matching*- identified peaks in each sample are compared across the whole data-set by setting overlapping 0.25 m/z bins and removing any duplicates (3) *retention time alignment algorithm* – calculates the median retention time and the deviation median for the peak-groups, complemented with a loess fitting method for correction. This process automatically generates a feature list for all the samples analyses, see **Table A2** , for an example. In addition, the PCA had a total of a 1000 loadings, the data was scaled with unit-variance and centered, see **Figure A26**.

On the other hand, the PCA analysis performed on the generated features by our in-house scripts, was carried out in Python with the Scikit library,[231] see Appendix for a copy of the script. All parameters were used in their default settings. In order to input the data, the features were distributed into two bins: 400 bins for the m/z values (1 m/z bins, for a range of 0 - 400 m/z values) and 20 bins for the retention time (1 minute bins, for a range 0-20 minutes in retention time). Then, the number of features in each cell is counted to fill a 400 x 20 matrix. See **Figure 70**. The resulting matrix was scaled onto unit scale (e.g. mean = 0 and variance = 1), as a way to standardize the features, before reducing the dimensionality of the data.

4.2.5.2 van Krevelen

The data selected for the van Krevelen plots corresponds to only a subset (~20%) of the total amount of features detected by the Compound Discoverer® workflow descriped in **Section 3.2.4.5**, since only a fraction had an annotated chemical formula, but it can still highlight differences in the chemical distribution of the data-sets. The chemical formula calculated in the automated (Compound Discoverer) workflow by the composition prediction node (see **Figure A18)** was used in the calculation. The formula assignment was limited to the elements carbon, hydrogen, nitrogen and oxygen in the following elemental ratios: H/C min=0.1 max=6, N/C min=0 max=4 & O/C min=0 max=3. Regardless of the instrumental accuracy, we did not extract the chemical formulas from the m/z value, in order to avoid the poor assumption that all m/z values represent [M+H] adducts. Considering that the identification of species is outside the remit of this work, more in-depth analysis to determine formulae was not pursued.

All plots were generated in Python with the matplotlib library, after manual extraction of the features with annotated chemical formula. This resulted in a series of van Krevelen plots, seen in **Section 2.2.7.2**, for Cycle 1, Cycle 2, Cycle 3, Long Cycle and Control samples. The number of points (e.g. features) are not representative of the total number of features detected, since the fraction of the features that had an automated assignment of their chemical formula was not equal for all the cases.

4.3 Recursive Miller-Urey Samples

4.3.1 Reagents and Gases

All Analytical solvents (water and acetonitrile, HPLC-MS grade) and mobile phase additive (e.g. formic acid) were purchased from ThermoFisher Scientific UK. For the Miller-Urey experiment, HPLC grade water and deuterium oxide were supplied by Goss Scientific. The gas mixtures were supplied pre-mixed by the British Oxygen Company (BOC) and CK Special Gases Ltd.

4.3.2 Experimental Procedure

Similar to the procedure described in **Section 2.1.1**, the Miller-Urey Recursive experiment was executed in the following fashion (without parallel rigs):

(a) 400mL of HPLC water (or deuterium oxide) was placed in the reaction vessel.

(b) The rig was then pumped down three times to degas the water. After the third evacuation, the system was pressurized (1 atm) with a gas mixture of roughly: 40% methane, 40% ammonia and 20% hydrogen .

(c) The round bottom flask (e.g. reaction vessel, 500 mL) was heated with a heating mantle, until the water started to boil.

(d) As soon as the water started boiling and recirculation was established, the 24 kV spark discharge was turned on, in a 10 seconds alternating (“on” - “off”) duty-cycle.

(e) The experiment was run for seven days, during which time the water in the flask progressively changed from clear to different shades of brown.

(f) The spark discharge and the heating mantel were turned off and the system was allowed to cool down.

(g) When cooled, a fraction of the solution was removed (e.g. 300 mL), collected in 500mL Duran® bottles and stored at room temperature.

(h) The reaction vessel was then replenished with fresh water (e.g. 300mL) and the process (a-g) was repeated.

(i) This was repeated a total of 4 times to generate 5 recursive samples.

The recursive Miller-Urey experiment was carried out with the help of Dr. Geoff Cooper.

4.3.3 UPLC-MS/MS analysis

4.3.3.1 Sample preparation

10 mL of each sample was transferred into 45 mL falcon tubes and centrifuged for 10 minutes at 4,400 rpm. The supernatant was filtrated with a syringe filter (0.22 µm cut-off) and transferred (~8mL) to 15 mL falcon tubes. Samples were then frozen by placing them at -20 °C for 8-12 hours and put in the freeze-drier (to be lyophilized).

The dried material was dissolved in 200 µL of acetonitrile (MS grade) and placed in the ultrasonic bath (35 Hz, room temperature) for 10 minutes. After, the samples were diluted further, 1 in a 100 with 50:50 (v/v) MS grade acetonitrile/water with 0.1% Formic Acid. Finally, the solution was filtrated with a spin-filter (0.22µm cut-off)] and placed in an HPLC sample vial for analysis.

4.3.3.2 Method

The reverse phase chromatography was performed in a Thermo Vanquish Ultra-performance liquid chromatography system (UPLC) coupled to a Thermo Orbitrap Fusion Mass-

Spectrometer (MS), fitted with an Agilent Poroshell 120 HPH-C18 (4.6 x 50mm, 2.7μm) column. All samples were injected in 10 μL aliquots and eluted with a linear gradient mixture of solvents A (water w/0.1% v/v formic acid) and B (100% acetonitrile w/0.1% v/v formic acid) over 35 minutes, while the column compartment was maintained at 30°C. The mass-spectral method was set to the same parameters as before, allowing for an automated fragmentation of the three most intense fragments in each full scan (MS1) and excluding them after 15 seconds of detection. The UPLC-MS method permitted the simultaneous separation and MS/MS fragmentation of the majority of the detected compounds.

All the MS spectra was collected for 30 minutes in positive mode over a scan range of 50–500 m/z. Ion transfer tube was set to 275 °C, RF lens 60%, and acquisition was performed in a data-dependent (DDA) manner. The Data-Dependent Acquisition (DDA) was performed by prioritizing the top most intense fragments in a 3 second window with an intensity threshold of 5.0E4 and dynamic exclusion, after one time for 15 seconds (in order to avoid the selection of the same fragments), using the ion trap isolation with a HCD collision energy of 35 eV and a resolution 15000. The analytical blanks were made in a 50:50 v/v mixture of HPLC grade water and acetonitrile. Experimental blanks were made by subjecting the 50:50 v/v water/acetonitrile mixture to the sample preparation procedure, in order to incorporate any contaminants arising from this process.

4.3.4 Feature Analysis

The features were generated with the method described in **Section 3.2.4.5**. No mayor alterations were made to the conversion, extraction or plotting process. The only distinction, is that this time we included a sample blank to be taken into consideration along-side with the instrumental blanks, as a way to avoid any false positives arising from the sample preparation process. Moreover, the PCA analysis of the features was conducted as described in **Section 5.2.5.1.** (except for the aforesaid modification on the inputted features).

Nonetheless, the feature analysis was taken further by creating a filtering algorithm, which allowed for the classification of the features according to their prevalence. In order to assess the different levels of persistence of the detected features across the recursive cycles, we generated a series of classes through a bespoke Python script. First, we identified which features were present across all the recursive cycles, labelling them as ‘generic’ to the system. Then, we looked for which features were only present in a particular cycle, which

we called ‘unique’. The number of unique features on each cycle was counted and compared across all three instrumental replicates, both to make the observations more robust, and to assign an error bar in the produced figures. Additionally, we looked for the features that were new in each cycle but only produced by the cursive action, for which we: (1) gathered all the features present from cycle 2 to cycle 5, and made sure they are not in cycle 1, (2) searched for features that are present from cycle 2 onwards, (3) and did the same for cycle 3 and cycle 4. We did not applied this concept to cycle 5, since being the last cycle, they would be equal to the unique features for cycle 5. Once we had identified and classified the features, we plotted them into scatter-plots as a way to visualize the differences across them. The plots were generated with the matplotlib library in python and had the number of MS^2 fragments (for each feature) appended represented as the colour of the point.

4.3.5 H-NMR

The samples were prepared by taking the insoluble part (precipitated pellet) at the end of the initial centrifugation (of the Miller-Urey solution) and freezing it with liquid nitrogen, before lyophilisation, to remove any remaining water. The dried (insoluble) material is dissolved in deuterated-DMSO (~600 µL) and placed in NMR tubes to be analysed. All data was acquired with the help of Dr. Jessica Bame.

All NMR data was recorded on a Bruker Advance 600 MHz. ^{1}H NMR at 600 MHz, in deuterated solvent, at T = 298 K, using TMS as the scale reference. Chemical shifts are reported using the δ-scale, referenced to the residual solvent protons in the deuterated solvent for ^{1}H (i.e. ^{1}H: δ (d-DMSO) = 7.26). All chemical shifts are given in ppm.

References

1. Knell, S. J. & Lewis, C. L. E. Celebrating the age of the Earth. *Geol. Soc. London, Spec. Publ.* **190,** 1–14 (2001).
2. Sleep, N. H., Zahnle, K. J., Kasting, J. F. & Morowitz, H. J. Annihilation of ecosystems by large asteroid impacts on the early Earth. *Nature* **342,** 139–142 (1989).
3. Mojzsis S.J., Arrhenius G., McKeegan K.D., Harrison T.M., Nutman A.P. & Friend CR. Evidence for life on Earth before 3 , 800 million years ago. **384,** 55–59 (1996).
4. Catling, D. C. & Claire, M. W. How Earth's atmosphere evolved to an oxic state: A status report. *Earth Planet. Sci. Lett.* **237,** 1–20 (2005).
5. Vankranendonk, M. Volcanic degassing, hydrothermal circulation and the flourishing of early life on Earth: A review of the evidence from c. 3490-3240 Ma rocks of the Pilbara Supergroup, Pilbara Craton, Western Australia. *Earth-Science Rev.* **74,** 197–240 (2006).
6. Schopf, J. W. The Early Archean Microfossils Apex Chert : New Evidence of the Life Antiquity. *Science.* **260,** 640–646 (1993).
7. Nisbet, E. G. & Sleep, N. H. The habitat and nature of early life. *Nature* **409,** 1083–1091 (2001).
8. Maruyama, S., Ikoma, M., Genda, H., Hirose, K., Yokoyama, T., & Santosh, M. The naked planet Earth: Most essential pre-requisite for the origin and evolution of life. *Geosci. Front.* **4,** 141–165 (2013).
9. Hansen, L. D., Criddle, R. S., & Battley, E. H. Biological calorimetry and the thermodynamics of the origination and evolution of life. *Pure Appl. Chem.* **81,** 1843–1855 (2009).
10. Matsuno, K. & Swenson, R. Thermodynamics in the present progressive mode and its role in the context of the origin of life. *BioSystems* **51,** 53–61 (1999).
11. Deamer, D. & Weber, A. L. Bioenergetics and Life's Origins. *Cold Spring Harb. Perspect. Biol.* **2,** a004929–a004929 (2010).
12. de Marcellus, P., Meinert, C., Myrgorodska, I., Nahon, L., Buhse, T., d'Hendecourt, L. L. S. & Meierhenrich, U. J. Aldehydes and sugars from evolved precometary ice analogs: importance of ices in astrochemical and prebiotic evolution. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 965–70 (2015).
13. Tuck, A. The role of atmospheric aerosols in the origin of life. *Surv. Geophys.* **23,** 379–409 (2002).
14. Trainer, M. G., Pavlov, A. A., Curtis, D. B., Mckay, C. P., Worsnop, D. R., Delia, A. E. & Tolbert, M. A.. Haze Aerosols in the Atmosphere of Early Earth: Manna from Heaven. *Astrobiology* **4,** 409–419 (2004).
15. He, C. & Smith, M. A. Identification of nitrogenous organic species in Titan aerosols analogs: Implication for prebiotic chemistry on Titan and early Earth. *Icarus* **238,** 86–92 (2014).
16. Brack, A. From interstellar amino acids to prebiotic catalytic peptides: A review. *Chem. Biodivers.* **4,** 665–679 (2007).
17. Pizzarello, S. Chemical Evolution and Meteorites: An Update. *Orig. Life Evol. Biosph.* **34,** 25–34 (2004).
18. Schwartz, A. W. Phosphorus in prebiotic chemistry. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **361,** 1743–1749 (2006).
19. Darwin, C. On the Origin of the Species. *Repr. New Sci.* **5,** 386 (1859).
20. Henderson, L. J. The Fitness of the Environment. *Science* **38,** 337–342 (1913).
21. Henderson, L.J. The Fitness of the Environment, an Inquiry into the Biological Significance of the Properties of Matter *The American Naturalist* **47,** 105–115 (2016).
22. Luisi, P. L. About Various Definition of Life. **28,** 613–622 (1998).

23. Schrödinger, E. What is Life? Mind and Matter. *Cambridge university press* (1944).
24. Boiteau, L. & Pascal, R. Energy Sources, Self-organization, and the Origin of Life. *Orig. Life Evol. Biosph.* **41,** 23–33 (2011).
25. Long, A. Soul and Body in Stoicism. *Phronesis* **27,** 34–57 (1982).
26. Preus, A. Galen ' s Criticism of Aristotle ' s Conception Theory. *J. Hist. Biol.* **10,** 65–85 (1977).
27. Needham, J. T. A summary of some late observations upon the generation, composition, and decomposition of animal and vegetable substances. *Philos. Trans. R. Soc. London* **45,** 615–666 (1748).
28. Owen, R. The Spontaneous Generation Controversy (1700-l860) : The Origin of Parasitic Worms *J. Hist. Biol.* **5,** 95–125 (1972).
29. Lipman, T. O. Wohler's Preparation of Urea and the Fute of Vitalism. *J. Chem. Educ.* **41,** 452-458 (1828).
30. Schwartz, M. The life and works of Louis Pasteur. *J. Appl. Microbiol.* **91,** 597–601 (2001).
31. Miller, S. L., Schopf, J. W. & Lazcano, A. Oparin's "Origin of Life": Sixty Years Later. *J. Mol. Evol.* **44,** 351–353 (1997).
32. Monnard, P. A. & Deamer, D. W. Membrane self-Assembly processes: Steps toward the first cellular life. *Minimal Cell Biophys. Cell Compart. Orig. Cell Funct.* **207,** 123–151 (2011).
33. Ikehara, K. Evolutionary Steps in the Emergence of Life Deduced from the Bottom-Up Approach and GADV Hypothesis (Top-Down Approach). *Life* **6,** 1-15 (2016).
34. Vaneechoutte, M. & Fani, R. From the primordial soup to the latest universal common ancestor. *Res. Microbiol.* **160,** 437–440 (2009).
35. Peretó, J., Bada, J. L. & Lazcano, A. Charles Darwin and the Origin of Life. *Orig Life Evol Biosph* **39,** 395–406 (2009).
36. Huxley, T. H. Biogenesis and Abiogenesis. *Collected Essays VIII*, 229–271 (1870).
37. Bailey, C. H. The Origin of Life (Oparin, A. I.). *J. Chem. Educ.* **15,** 399 (1938).
38. Lazcano, A. & Miller, S. L. The origin and early evolution of life: Prebiotic chemistry, the pre-RNA world, and time. *Cell* **85,** 793–796 (1996).
39. Miller, S. L. A production of amino acids under possible primitive earth conditions. *Science* **117,** 528–529 (1953).
40. Burton, A. S., Stern, J. C., Elsila, J. E., Glavin, D. P. & Dworkin, J. P. Prebiotic chemistry themed issue Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteoritesw. *Chem. Soc. Rev. Chem. Soc. Rev* **41,** 5459–5472 (2012).
41. J.D & Watson, C. F. H. The structure of DNA. *Cold Spring Harb. Symp. Quant. Biol.* **18,** 123–131 (1953).
42. Orgel, L. E. Origin of Life on Earth: A Review *Trends Biochem. Sci.* **23,** 491–495 (1998).
43. Ferus, M. *et al.* Formation of nucleobases in a Miller–Urey reducing atmosphere. *Proc. Natl. Acad. Sci.* **114,** 4306–4311 (2017).
44. Damer, B. A Field Trip to the Archaean in Search of Darwin's Warm Little Pond. *Life* **6,** 21 (2016).
45. Ostrovskii, V. E. & Kadyshevich, E. A. Life Origination Hydrate Theory (LOH-Theory) and the Explanation of the Biological Diversification. *J. Mol. Evol.* **79,** 155–178 (2014).
46. Ross, D. & Deamer, D. Dry/Wet Cycling and the Thermodynamics and Kinetics of Prebiotic Polymer Synthesis. *Life* **6,** 1-12 (2016).
47. Rodriguez-Garcia, M., Surman, A. J., Cooper, G. J., Suárez-Marina, I., Hosni, Z., Lee, M. P., & Cronin, L. Formation of oligopeptides in high yield under simple programmable conditions. *Nat. Commun.* **6,** 8385 (2015).
48. Higgs, P. The Effect of Limited Diffusion and Wet–Dry Cycling on Reversible

Polymerization Reactions: Implications for Prebiotic Synthesis of Nucleic Acids. *Life* **6,** 24 (2016).
49. Pearce, B. K. D., Pudritz, R. E., Semenov, D. A. & Henning, T. K. Origin of the RNA world: The fate of nucleobases in warm little ponds. *Proc. Natl. Acad. Sci.* **114,** 11327–11332 (2017).
50. Levin, S. A. Ecosystems and the Biosphere as Complex Adaptive Systems. *Ecosystems* **1,** 431–436 (1998).
51. Felnagle, E. A., Chaubey, A., Noey, E. L., Houk, K. N. & Liao, J. C. Engineering synthetic recursive pathways to generate non-natural small molecules. *Nat. Chem. Biol.* **8,** 518–526 (2012).
52. Ignat, I. & Victorovich, M. O. Origin of Life and Living Matter in Hot Mineral Water. *Интернет-журнал Науковедение* **7,** 1–19 (2013).
53. Cafferty, B. J. *et al.* Freeze-thaw cycles as drivers of complex ribozyme assembly. *Processes* **7,** 471–475 (2015).
54. Schrödinger, E. What is life? The physical aspect of the living cell. Erwin Schrödinger. *Am. J. Phys. Anthropol.* 1–32 (1944).
55. Lathe, R. Fast tidal cycling and the origin of life. *Icarus* **168,** 18–22 (2004).
56. Kaneko, K. Recursiveness, switching, and fluctuations in a replicating catalytic network. *Phys. Rev. E* **68,** 031909 (2003).
57. Baum, D. A. & Vetsigian, K. An Experimental Framework for Generating Evolvable Chemical Systems in the Laboratory. *Orig. Life Evol. Biosph.* 1–17 (2016).
58. Roy, L. & Case, M. A. Recursively enriched dynamic combinatorial libraries for the self-selection of optimally stable proteins. *J. Phys. Chem. B* **115,** 2454–2464 (2011).
59. Brush, S. G. Early history of selenogony. *Orig. Moon* **18,** 3–15 (1986).
60. Martin, W. & Russell, M. J. On the origin of biochemistry at an alkaline hydrothermal vent. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **362,** 1887–925 (2007).
61. Martin, W., Baross, J., Kelley, D. & Russell, M. J. Hydrothermal vents and the origin of life. *Nat. Rev.* **6,** 805–814 (2008).
62. Wächtershäuser, G. Before enzymes and templates: theory of surface metabolism. *Microbiol. Rev.* **52,** 452–484 (1988).
63. Szostak, J. W., Bartel, D. P. & Luisi, P. L. Synthesizing life. *Nature* **409,** 387–390 (2001).
64. Ignatov, I. Which Water is Optimal for the Origin (Generation) of Life? *Euromedica, Hanover* 34–37 (2010).
65. Herschy, B. *et al.* An Origin-of-Life Reactor to Simulate Alkaline Hydrothermal Vents. *J. Mol. Evol.* **79,** 213–227 (2014).
66. Cleaves, H. J., Aubrey, a D. & Bada, J. L. An Evaluation of the Critical Parameters for Abiotic Peptide Synthesis in Submarine Hydrothermal Systems. *Orig. Life Evol. Biosph.* **39,** 109–126 (2009).
67. Koufaki, M., Fotopoulou, T. & Heropoulos, G. A. Synergistic effect of dual-frequency ultrasound irradiation in the one-pot synthesis of 3,5-disubstituted isoxazoles. *Ultrason. Sonochem.* **21**, *35-39* (2014).
68. Bujdák, J. & Rode, B. M. Activated alumina as an energy source for peptide bond formation: Consequences for mineral-mediated prebiotic processes. *Amino Acids* **2**, 281-291 (2001).
69. Murillo-Sánchez, S., Beaufils, D., González Mañas, J. M., Pascal, R. & Ruiz-Mirazo, K. Fatty acids double role in the prebiotic formation of a hydrophobic dipeptide. *Chem. Sci.* **7,** 3406–3413 (2016).
70. Cairns-Smith, A. G. & Hartman, H. *C*lay Minerals and the Origin of Life. *Cambridge University Press* (1986).
71. Cairns-Smith, A. G., Ingram, P. & Walker, G. L. Formose production by minerals: Possible relevance to the origin of life. *J. Theor. Biol.* **35,** 601–604 (1972).
72. Rimola, A., Sodupe, M. & Ugliengo, P. Affinity scale for the interaction of amino

acids with silica surfaces. *J. Phys. Chem. C* **113,** 5741–5750 (2009).
73. Saladino, R., Neri, V. & Crestini, C. Role of clays in the prebiotic synthesis of sugar derivatives from formamide. *Philos. Mag.* **90,** 2329–2337 (2010).
74. 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).
75. Hazen, R. M. & Sverjensky, D. A. Mineral Surfaces, Geochemical Complexities, and the Origins of Life. *Cold Spring Harb. Perspect. Biol.* **2,** a002162–a002162 (2010).
76. Hazen, R.M., Papineau, D., Bleeker, W., Downs, R.T., Ferry, J.M., McCoy, T.J., Sverjensky, D.A. & Yang, H. Mineral evolution. *Am. Mineral.* **93,** 1693–1720 (2008).
77. Hazen, R. M. Paleomineralogy of the Hadean Eon: A preliminary species list. *Am. J. Sci.* **313,** 807–843 (2013).
78. Rimola, A., Sodupe, M. & Ugliengo, P. Role of Mineral Surfaces in Prebiotic Chemical Evolution. In Silico Quantum Mechanical Studies. *Life* **9,** 10 (2019).
79. Schoonen, M., Smirnov, A. & Cohn, C. A Perspective on the Role of Minerals in Prebiotic Synthesis. *AMBIO A J. Hum. Environ. Publ.* **33,** 539–551 (2004).
80. Liu, R. & Orgel, L. E. Polymerization on the rocks: beta-amino acids and arginine. *Orig. Life Evol. Biosph.* **28,** 245–257 (1998).
81. Lambert, J.-F. Adsorption and polymerization of amino acids on mineral surfaces: a review. *Orig. Life Evol. Biosph.* **38,** 211–42 (2008).
82. Zaia, D. A. M. A review of adsorption of amino acids on minerals: Was it important for origin of life? *Amino Acids* **27,** (2004).
83. de Souza-Barros, F. & Vieyra, A. Mineral interface in extreme habitats: A niche for primitive molecular evolution for the appearance of different forms of life on Earth. *Comp. Biochem. Physiol. - C Toxicol. Pharmacol.* **146,** 10–21 (2007).
84. Radzicka, A. & Wolfenden, R. Rates of uncatalyzed peptide bond hydrolysis in neutral solution and the transition state affinities of proteases. *J. Am. Chem. Soc.* **118,** 6105–6109 (1996).
85. Surman, A. J., Rodriguez-Garcia, M., Abul-haija, Y. M., Cooper, G. J. T. & Gromski, P. S. Environmental control programs the emergence of distinct functional ensembles from unconstrained chemical reactions. *PNAS* **116,** 5387–5392 (2019).
86. Ferris, J. P., Hill, a R., Liu, R. & Orgel, L. E. Synthesis of long prebiotic oligomers on mineral surfaces. *Nature* **381,** 59–61 (1996).
87. Cossetti, C., Crestini, C., Saladino, R. & di Mauro, E. Borate minerals and RNA stability. *Polymers (Basel).* **2,** 211–228 (2010).
88. Ferus, M., Knížek, A. & Civiš, S. Meteorite-catalyzed synthesis of nucleosides and other prebiotic compounds. *Proc. Natl. Acad. Sci.* **112,** 7109–7110 (2015).
89. Pasek, M., Herschy, B. & Kee, T. P. Phosphorus: a Case for Mineral-Organic Reactions in Prebiotic Chemistry. *Orig. Life Evol. Biosph.* **45,** 207–218 (2015).
90. Gull, M. Prebiotic Phosphorylation Reactions on the Early Earth. *Challenges* **5,** 193–212 (2014).
91. Furukawa, Y., Kim, H.-J., Hutter, D. & Benner, S. A. Abiotic regioselective phosphorylation of adenosine with borate in formamide. *Astrobiology* **15,** 259–267 (2015).
92. Powner, M. W., Gerland, B. & Sutherland, J. D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459,** 239–242 (2009).
93. Hansma, H. G. Possible origin of life between mica sheets. *J. Theor. Biol.* **266,** 175–188 (2010).
94. Kumar, A. & Kamaluddin. Possible Role of Metal(II) Octacyanomolybdate(IV) in Chemical Evolution: Interaction with Ribose Nucleotides. *Orig. Life Evol. Biosph.* **43,** 1–17 (2013).
95. Kim, H. J. *et al.* Synthesis of carbohydrates in mineral-guided prebiotic cycles. *J. Am. Chem. Soc.* **133,** 9457–9468 (2011).

96. Leman, L. Carbonyl Sulfide-Mediated Prebiotic Formation of Peptides. *Science* **306,** 283–286 (2004).
97. Goldford, J. E., Hartman, H., Smith, T. F. & Segrè, D. Remnants of an Ancient Metabolism without Phosphate. *Cell* **168,** 1126–1134.e9 (2017).
98. Grew, E. S., Bada, J. L. & Hazen, R. M. Borate Minerals and Origin of the RNA World. *Orig. Life Evol. Biosph.* **41,** 307–316 (2011).
99. Benner, S. a., Kim, H. J. & Carrigan, M. A. Asphalt, Water, and the Prebiotic Synthesis of Ribose, Ribonucleosides, and RNA. *Acc. Chem. Res.* **45,** 2025–2034 (2012).
100. Goldenfeld, N. & Woese, C. Life is Physics: Evolution as a Collective Phenomenon Far From Equilibrium. *Annu. Rev. Condens. Matter Phys.* **2,** 375–399 (2011).
101. Orgel, L. E. The origin of life – a review of facts and speculations. *TIBS* **4,** 491–495 (1998).
102. Kawamura, K. Drawbacks of the ancient RNA-based life-like system under primitive earth conditions. *Biochimie* **94,** 1441–1450 (2012).
103. Hernández, A. R. & Piccirilli, J. A. Prebiotic RNA unstuck. *Nat. Chem.* **5,** 360–362 (2013).
104. Hud, N. V. Our Odyssey to Find a Plausible Prebiotic Path to RNA: The First Twenty Years. *Synlett* **28,** 36–55 (2017).
105. Mast, C. B., Schink, S., Gerland, U. & Braun, D. Escalation of polymerization in a thermal gradient. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 8030–8035 (2013).
106. Šponer, J. E., Šponer, J., Novakova, O., Brabec, V., Šedo, O., Zdráhal, Z. & Di Mauro, E. Emergence of the First Catalytic Oligonucleotides in a Formamide-Based Origin Scenario. *Chem. - A Eur. J.* **22,** 3572–3586 (2016).
107. 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).
108. Lee Phillips, M. & Phillips, M. L. The Origins Divide: Reconciling Views on How Life Began. *Bioscience* **60,** 675–680 (2010).
109. Pross, A. Causation and The Origin of Life. Metabolism or Replication first? *Orig. Life Evol. Biosph.* **34,** 307–321 (2004).
110. Huber, C. & Wachtershauser, G. Alpha-Hydroxy and alpha-Amino Acids Under Possible Hadean, Volcanic Origin-of-Life Conditions. *Science (80-.).* **314,** 630–632 (2006).
111. Huber, C. & Wächtershäuser, G. Peptides by activation of amino acids with CO on (Ni,Fe)S surfaces: implications for the origin of life. *Science* **281,** 670–672 (1998).
112. Martin, W. & Russell, M. J. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **358,** 59–83 (2003).
113. Smith, E. & Morowitz, H. J. Universality in intermediary metabolism. *Proc Natl Acad Sci U S A* **101,** 13168–13173 (2004).
114. Bonfio, C., Valer, L., Scintilla, S., Shah, S., Evans, D.J., Jin, L., Szostak, J.W., Sasselov, D.D., Sutherland, J.D. & Mansy, S.S.UV-light-driven prebiotic synthesis of iron-sulfur clusters. *Nat. Chem.* **9,** (2017).
115. Zhang, X. V. & Martin, S. T. Driving Parts of Krebs Cycle in Reverse through Mineral Photochemistry. *J. Am. Chem. Soc.* **128,** 16032–16033 (2006).
116. Muchowska, K. B., Varma, S. J. & Moran, J. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569,** 104–107 (2019).
117. Keller, M. A., Kampjut, D., Harrison, S. A. & Ralser, M. Sulfate radicals enable a non-enzymatic Krebs cycle precursor. *Nat. Ecol. Evol.* **1,** 1–9 (2017).
118. Ruiz-Bermejo, M., Menor-Salván, C., Mateo-Martí, E., Osuna-Esteban, S., Martín-Gago, J. Á. & Veintemillas-Verdaguer, S.. $CH_4/N_2/H_2$ spark hydrophilic tholins: A

systematic approach to the characterization of tholins. *Icarus* **198,** 232–241 (2008).
119. Johnson, A. P., Cleaves, H. J., Dworkin, J. P., Glavin, D. P., Lazcano, A. & Bada, J. L. The Miller Volcanic Spark Discharge Experiment. *Science* **322,** 404–404 (2008).
120. Cleaves, H. J., Chalmers, J. H., Lazcano, A., Miller, S. L. & Bada, J. L. Prebiotic organic synthesis in neutral planetary atmospheres. *ACS Symposium Series* **981,** 282–292 (2008).
121. Parker, E.T., Zhou, M., Burton, A.S., Glavin, D.P., Dworkin, J.P., Krishnamurthy, R., Fernández, F.M. & Bada, J.L. A plausible simultaneous synthesis of amino acids and simple peptides on the primordial earth. *Angew. Chemie - Int. Ed.* **53,** 8132–8136 (2014).
122. Miller, S. L. The mechanism of synthesis of amino acids by electric discharges. *Biochim. Biophys. Acta* **23,** 480–489 (1957).
123. Parker, E.T., Cleaves, H.J., Dworkin, J.P., Glavin, D.P., Callahan, M., Aubrey, A., Lazcano, A. & Bada, J.L. Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark discharge experiment. *Proc. Natl. Acad. Sci.* **108,** 5526–5531 (2011).
124. Rodriguez, L. E., House, C. H., Smith, K. E., Roberts, M. R. & Callahan, M. P. Nitrogen heterocycles form peptide nucleic acid precursors in complex prebiotic mixtures. *Sci. Rep.* **9,** 1–12 (2019).
125. Wollrab, E. & Ott, A. A Miller–Urey broth mirrors the mass density distribution of all Beilstein indexed organic molecules. *New J. Phys.* **20,** 105003 (2018).
126. Leman, L. J., Huang, Z.-Z. & Ghadiri, M. R. Peptide Bond Formation in Water Mediated by Carbon Disulfide. *Astrobiology* **15,** 709–716 (2015).
127. Danger, G., Plasson, R. & Pascal, R. Pathways for the formation and evolution of peptides in prebiotic environments. *Chem. Soc. Rev.* **41,** 5416–5429 (2012).
128. Jakschitz, T. E. & Rode, B. M. Chemical evolution from simple inorganic compounds to chiral peptides. *Chem. Soc. Rev.* **41,** 5484–5489 (2012).
129. Rodriguez-Garcia, M., Surman, A.J., Cooper, G.J., Suárez-Marina, I., Hosni, Z., Lee, M.P. & Cronin, L. Formation of oligopeptides in high yield under simple programmable conditions. *Nat. Commun.* **6,** 8385-8392 (2015).
130. Mamajanov, I., MacDonald, P.J., Ying, J., Duncanson, D.M., Dowdy, G.R., Walker, C.A., Engelhart, A.E., Fernández, F.M., Grover, M.A., Hud, N.V. & Schork, F.J. Ester formation and hydrolysis during wet-dry cycles: Generation of far-from-equilibrium polymers in a model prebiotic reaction. *Macromolecules* **47,** 1334–1343 (2014).
131. Forsythe, J.G., Yu, S.S., Mamajanov, I., Grover, M.A., Krishnamurthy, R., Fernández, F.M. & Hud, N.V. Ester-Mediated Amide Bond Formation Driven by Wet-Dry Cycles: A Possible Path to Polypeptides on the Prebiotic Earth. *Angew. Chemie Int. Ed.* **54,** 9871–9875 (2015).
132. Faculty, T. A., Yu, S. & Fulfillment, I. P. Ester-Mediated Amide Bond Formation : a Possible Path To Proto-Peptides on the Prebiotic Earth. *Doctoral dissertation, Georgia Institute of Technology* (2017).
133. Jia, T. Z., Chandru, K., Hongo, Y., Afrin, R., Usui, T., Myojo, K. & Cleaves, H. J. Membraneless polyester microdroplets as primordial compartments at the origins of life. *Proc. Natl. Acad. Sci.* **116,** 201902336 (2019).
134. Schmitt-Kopplin, P., Gabelica, Z., Gougeon, R.D., Fekete, A., Kanawati, B., Harir, M., Gebefuegi, I., Eckel, G. & Hertkorn, N. High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall. *Proc. Natl. Acad. Sci.* **107,** 2763–2768 (2010).
135. Engel, M. H. & Macko, S. A. Isotopic evidence for extraterrestrial non-racemic amino acids in the Murchison meteorite. *Nature* **389,** 265–268 (1997).
136. Krishnamurthy, R. V., Epstein, S., Cronin, J. R., Pizzarello, S. & Yuen, G. U. Isotopic and molecular analyses of hydrocarbons and monocarboxylic acids of the Murchison meteorite. *Geochim. Cosmochim. Acta* **56,** 4045–4058 (1992).

137. Martins, Z., Botta, O., Fogel, M.L., Sephton, M.A., Glavin, D.P., Watson, J.S., Dworkin, J.P., Schwartz, A.W. & Ehrenfreund, P. Extraterrestrial nucleobases in the Murchison meteorite. *Earth Planet. Sci. Lett.* **270,** 130–136 (2008).
138. Meinert, C., Myrgorodska, I., De Marcellus, P., Buhse, T., Nahon, L., Hoffmann, S.V., d'Hendecourt, L.L.S. & Meierhenrich, U.J. Ribose and related sugars from ultraviolet irradiation of interstellar ice analogs. *Science* **352,** 208–212 (2016).
139. Caro G.M., Meierhenrich U.J., Schutte W.A., Barbier B., Segovia A.A., Rosenbauer H., Thiemann W.P. & Brack A, G. J. Amino acids from ultraviolet irradiation of interstellar ice analogues. *Nature* **416,** 403–406 (2002).
140. Crick, F. H. C. & Orgel, L. E. Directed panspermia. *Icarus* **19,** 341–346 (1973).
141. Butlerow, A. Bildung einer zuckerartigen Substanz durch Synthese. *Justus Liebigs Ann. Chem.* **120,** 295–298 (1861).
142. 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).
143. 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).
144. Ricardo, A. Borate Minerals Stabilize Ribose. *Science* **303,** 196–196 (2004).
145. Cleaves, H. J. The prebiotic geochemistry of formaldehyde. *Precambrian Res.* **164,** 111–118 (2008).
146. Schwartz, A. W. & Goverde, M. Acceleration of HCN oligomerization by formaldehyde and related compounds: Implications for prebiotic syntheses. *J. Mol. Evol.* **18,** 351–353 (1982).
147. Weber, A. L. The sugar model: Autocatalytic activity of the triose-ammonia reaction. *Orig. Life Evol. Biosph.* **37,** 105–111 (2007).
148. Oro, J. Mechanism of synthesis of adenine from hydrogen cyanide under possible primitive Earth conditions. *Nature* **191,** 1193–1194 (1961).
149. Saladino, R., Botta, G., Pino, S., Costanzo, G. & Di Mauro, E. From the one-carbon amide formamide to RNA all the steps are prebiotically possible. *Biochimie* **94,** 1451–1456 (2012).
150. Pino, S., Sponer, J., Costanzo, G., Saladino, R. & Mauro, E. From Formamide to RNA, the Path Is Tenuous but Continuous. *Life* **5,** 372–384 (2015).
151. Andersen, J., Andersen, T., Flamm, C., Hanczyc, M., Merkle, D. & Stadler, P. Navigating the chemical space of hcn polymerization and hydrolysis: Guiding graph grammars by mass spectrometry data. *Entropy* **15,** 4066–4083 (2013).
152. Bada, J. L., Chalmers, J. H. & Cleaves, H. J. Is formamide a geochemically plausible prebiotic solvent? *Phys.Chem.Chem. Phys.* **18,** 20085–20090 (2016).
153. Saitta, A. M. & Saija, F. Miller experiments in atomistic computer simulations. *Proc. Natl. Acad. Sci.* **111,** 13768–13773 (2014).
154. Barks, H. L., Buckley, R., Grieves, G. A., Di Mauro, E., Hud, N. V. & Orlando, T. M. Guanine, adenine, and hypoxanthine production in UV-irradiated formamide solutions: Relaxation of the requirements for prebiotic purine nucleobase formation. *ChemBioChem* **11,** 1240–1243 (2010).
155. Saladino, R., Barontini, M., Cossetti, C., Di Mauro, E. & Crestini, C. The effects of borate minerals on the synthesis of nucleic acid bases, amino acids and biogenic carboxylic acids from formamide. *Orig. Life Evol. Biosph.* **41,** 317–30 (2011).
156. Nguyen, H. T. & Nguyen, M. T. Decomposition pathways of formamide in the presence of vanadium and titanium monoxides. *Phys. Chem. Chem. Phys.* **17,** 16927–16936 (2015).
157. Šponer, J. E., Mládek, A., Šponer, J. & Fuentes-Cabrera, M. Formamide-Based Prebiotic Synthesis of Nucleobases: A Kinetically Accessible Reaction Route. *J. Phys. Chem. A* **116,** 720–726 (2012).
158. Saladino, R., Carota, E., Botta, G., Kapralov, M., Timoshenko, G.N., Rozanov, A.Y., Krasavin, E. & Di Mauro, E. Meteorite-catalyzed syntheses of nucleosides and of

other prebiotic compounds from formamide under proton irradiation. *Proc. Natl. Acad. Sci.* **112,** 2746–2755 (2015).
159. Schoffstall, A. Prebiotic phosphorylation of nucleosides in formamide. *Orig. Life* **7,** 399–412 (1976).
160. Sutherland, J. D. Opinion: Studies on the origin of life — the end of the beginning. *Nat. Rev. Chem.* **1,** 1–12 (2017).
161. Nitschke, J. R. Systems chemistry: Molecular networks come of age. *Nature* **462,** 736–738 (2009).
162. 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).
163. Sutherland, J. D. The Origin of Life-Out of the Blue. *Angew. Chemie Int. Ed.* **55,** 104–121 (2016).
164. Kua, J. & Thrush, K. L. HCN, Formamidic Acid, and Formamide in Aqueous Solution: A Free-Energy Map. *J. Phys. Chem. B* **120,** 8175–8185 (2016).
165. Tsukahara, H., Imai, E. I., Honda, H., Hatori, K. & Matsuno, K. Prebiotic oligomerization on or inside lipid vesicles in hydrothermal environments. *Orig. Life Evol. Biosph.* **32,** 13–21 (2002).
166. Islam, S. & Powner, M. W. Prebiotic Systems Chemistry: Complexity Overcoming Clutter. *Chem* **2,** 470–501 (2017).
167. Mamajanov, I. & Herzfeld, J. HCN polymers characterized by solid state NMR: Chains and sheets formed in the neat liquid. *J. Chem. Phys.* **130,** 134503 (2009).
168. Mamajanov, I., Callahan, M. P., Dworkin, J. P. & Cody, G. D. Prebiotic Alternatives to Proteins: Structure and Function of Hyperbranched Polyesters. *Orig. Life Evol. Biosph.* **45,** 123–137 (2015).
169. Keller, M. A., Kampjut, D., Harrison, S. A., Driscoll, P. C. & Ralser, M. Reply to 'Do sulfate radicals really enable a non-enzymatic Krebs cycle precursor?' *Nat. Ecol. Evol.* **3,** 139–140 (2019).
170. Geisberger, T., Diederich, P., Steiner, T., Eisenreich, W., Schmitt-Kopplin, P. & Huber, C. Evolutionary Steps in the Analytics of Primordial Metabolic Evolution. *Life* **9,** 50 (2019).
171. Buch, A., Glavin, D.P., Sternberg, R., Szopa, C., Rodier, C., Navarro-González, R., Raulin, F., Cabane, M. & Mahaffy, P.R. A new extraction technique for in situ analyses of amino and carboxylic acids on Mars by gas chromatography mass spectrometry. *Planet. Space Sci.* (2006).
172. Ruiz-Bermejo, M., de la Fuente, J. L., Rogero, C., Menor-Salván, C., Osuna-Esteban, S. & Martín-Gago, J. A. New insights into the characterization of 'insoluble black HCN polymers'. *Chemistry and Biodiversity* (2012).
173. Mccollom, T. M., Ritter, G. & Simoneit, B. R. T. Lipid synthesis under hydrothermal conditions by Fischer-Tropsch-type reactions. *Orig. Life Evol. Biosph.* **29,** 153–166 (1999).
174. Medeiros, P. M. & Simoneit, B. R. T. Analysis of sugars in environmental samples by gas chromatography-mass spectrometry. *J. Chromatogr. A* **1141,** 271–278 (2007).
175. Cordell, R. L., Pandya, H., Hubbard, M., Turner, M. A. & Monks, P. S. GC-MS analysis of ethanol and other volatile compounds in micro-volume blood samples-quantifying neonatal exposure. *Anal. Bioanal. Chem.* **405,** 4139–4147 (2013).
176. Scherer, S., Wollrab, E., Codutti, L., Carlomagno, T., da Costa, S.G., Volkmer, A., Bronja, A., Schmitz, O.J. & Ott, A. Chemical Analysis of a "Miller-Type" Complex Prebiotic Broth Part 2: Gas, Oil, Water and the Oil/Water interface. *Orig. Life Evol. Biosph.* **47**, 381–403 (2017).
177. Molnár-Perl, I. & Katona, Z. F. GC-MS of amino acids as their trimethylsilyl/t-butyldimethylsilyl Derivatives: In model solutions III. *Chromatographia* **51,** 228236 (2000).
178. Mompeán, C., Marín-Yaseli, M. R., Espigares, P., González-Toril, E., Zorzano, M. P.

& Ruiz-Bermejo, M. Prebiotic chemistry in neutral/reduced-alkaline gas-liquid interfaces. *Sci. Rep.* **9,** 1916–1928 (2019).
179. Zweckmair, T., Böhmdorfer, S., Bogolitsyna, A., Rosenau, T., Potthast, A. & Novalin, S. Accurate analysis of formose reaction products by LC-UV: An analytical challenge. *J. Chromatogr. Sci.* **52,** 169–175 (2014).
180. Iqbal, M. Z. & Novalin, S. Analysis of formose sugar and formaldehyde by high-performance liquid chromatography. *J. Chromatogr. A* **1216,** 5116–5121 (2009).
181. Ulusoy, S., Halil, A., Ulusoy, I., Pleissner, D. & Eriksen, N. T. Nitrosation and analysis of amino acid derivatives by isocratic HPLC. *RSC Advances* **6**, 13120-13128 (2016).
182. Nemkov, T., D'Alessandro, A. & Hansen, K. C. Three-minute method for amino acid analysis by UHPLC and high-resolution quadrupole orbitrap mass spectrometry. *Amino Acids* **47,** 2345–2357 (2015).
183. Parker, E.T., Cleaves, J.H., Burton, A.S., Glavin, D.P., Dworkin, J.P., Zhou, M., Bada, J.L. & Fernández, F.M. Conducting Miller-Urey Experiments. *J. Vis. Exp.* **83**, e51039 (2014).
184. Heeren, R. M. A., Kleinnijenhuis, A. J., McDonnell, L. A. & Mize, T. H. A mini-review of mass spectrometry using high-performance FTICR-MS methods. *Anal. Bioanal. Chem.* **378,** 1048–1058 (2004).
185. Kim, S., Rodgers, R. P., Blakney, G. T., Hendrickson, C. L. & Marshall, A. G. Automated Electrospray Ionization FT-ICR Mass Spectrometry for Petroleum Analysis. *J. Am. Soc. Mass Spectrom.* **20,** 263–268 (2009).
186. Hertkorn, N., Harir, M. & Schmitt-Kopplin, P. Nontarget analysis of Murchison soluble organic matter by high-field NMR spectroscopy and FTICR mass spectrometry. *Magn. Reson. Chem.* **53,** 754–768 (2015).
187. Guttenberg, N., Virgo, N., Chandru, K., Scharf, C. & Mamajanov, I. Bulk measurements of messy chemistries are needed for a theory of the origins of life. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **375,** 20160347 (2017).
188. Lu, W., Bennett, B. D. & Rabinowitz, J. D. Analytical strategies for LC-MS-based targeted metabolomics. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **871,** 236–242 (2008).
189. Schiffman, C., Petrick, L., Perttula, K., Yano, Y., Carlsson, H., Whitehead, T., Metayer, C., Hayes, J., Rappaport, S. & Dudoit, S Filtering procedures for untargeted LC-MS metabolomics data. *BMC Bioinformatics* **20,** 334-344 (2019).
190. Patti, G. J., Yanes, O. & Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13,** 263–269 (2012).
191. Guttenberg, N., Virgo, N., Chandru, K., Scharf, C. & Mamajanov, I. Bulk measurements of messy chemistries are needed for a theory of the origins of life. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **375,** 20160347 (2017).
192. Ruf, A., Poinot, P., Geffroy, C., Le Sergeant d'Hendecourt, L. & Danger, G. Data-Driven UPLC-Orbitrap MS Analysis in Astrochemistry. *Life* **9,** 35 (2019).
193. Kitadai, N. & Maruyama, S. Origins of building blocks of life: A review. *Geosci. Front.* **9,** 1117–1153 (2018).
194. Miller, S. L. A Production of Amino Acids Under Possible Primitive Earth Conditions. *Science* **117,** 528–529 (1953).
195. Brickwedde, F. G. Harold Urey and the discovery of deuterium. *Phys. Today* **35,** 34–39 (1982).
196. Shampo, M. A., Kyle, R. A. & Steensma, D. P. Frederick Soddy--pioneer in radioactivity. *Mayo Clin. Proc.* **86,** 4065 (2011).
197. Elliott, C., Vijayakumar, V., Zink, W. & Hansen, R. National Instruments LabVIEW: A Programming Environment for Laboratory Automation and Measurement. *J. Lab. Autom.* **12,** 17–24 (2007).
198. Vasanits, A., Kutlán, D., Sass, P. & Molnár-Perl, I. Retention/quantitation properties

of the o-phthaldialdehyde-3-mercaptopropionic acid and the o-phthaldialdehyde-N-acetyl-L-cysteine amino acid derivatives in reversed-phase high-performance liquid chromatography. *J. Chromatogr. A* **870,** 271–287 (2000).

199. McCollom, T. M. Miller-Urey and Beyond: What Have We Learned About Prebiotic Organic Synthesis Reactions in the Past 60 Years? *Annu. Rev. Earth Planet. Sci.* **41,** 207–229 (2013).
200. Corning. Low Actinic PYREX Code 7740 Glasses. (2015).
201. Stüeken, E.E., Anderson, R.E., Bowman, J.S., Brazelton, W.J., Colangelo-Lillis, J., Goldman, A.D., Som, S.M. & Baross, J.A. Did life originate from a global chemical reactor? *Geobiology* **11,** 101–126 (2013).
202. Saladino, R., Crestini, C., Pino, S., Costanzo, G. & Di Mauro, E. Formamide and the origin of life. *Phys. Life Rev.* **9,** 84–104 (2012).
203. Karlsson, G., Winge, S. & Sandberg, H. Separation of monosaccharides by hydrophilic interaction chromatography with evaporative light scattering detection. *J. Chromatogr. A* **1092,** 246–249 (2005).
204. Stratton, T. & Scientific, T. F. Compounding insights for small molecule research Thermo Scientific Compound Discoverer software. *Thermo Fisher Scientific, MA, USA* (2017).
205. Ikegami, T., Horie, K., Saad, N., Hosoya, K., Fiehn, O. & Tanaka, N. Highly efficient analysis of underivatized carbohydrates using monolithic-silica-based capillary hydrophilic interaction (HILIC) HPLC. *Anal. Bioanal. Chem.* **391,** 2533–2542 (2008).
206. Jandera, P. Stationary and mobile phases in hydrophilic interaction chromatography: A review. *Anal. Chim. Acta* **692,** 1–25 (2011).
207. Bailey, B., Ullucci, P., Bauder, R., Plante, M., Crafts, C. & Acworth, I. Carbohydrate Analysis using HPLC with PAD , FLD , Charged Aerosol Detection , and MS Detectors. *Thermo Fisher Scientific, MA, USA.*1–6 (2013).
208. Ge, Y., Tang, Y., Guo, S., Liu, X., Zhu, Z., Liu, P. & Duan, J. A. Comparative Analysis of Amino Acids, Nucleosides, and Nucleobases in Thais clavigera from Different Distribution Regions by Using Hydrophilic Interaction Ultra-Performance Liquid Chromatography Coupled with Triple Quadrupole Tandem Mass Spectrometry. *Int. J. Anal. Chem.* **2015,** 1–10 (2015).
209. Idborg, H., Zamani, L., Edlund, P. O., Schuppe-Koistinen, I. & Jacobsson, S. P. Metabolic fingerprinting of rat urine by LC/MS: Part 1. Analysis by hydrophilic interaction liquid chromatography-electrospray ionization mass spectrometry. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **828,** 9–13 (2005).
210. Zeffiro, A., Lazzaroni, S., Merli, D., Profumo, A., Buttafava, A., Serpone, N. & Dondi, D. Formation of Hexamethylenetetramine (HMT) from HCHO and NH 3 – Relevance to Prebiotic Chemistry and B3LYP Consideration. *Origins of Life and Evolution of Biospheres.* **46**,223–231 (2016).
211. Wollrab, E., Scherer, S., Aubriet, F., Carré, V., Carlomagno, T., Codutti, L. & Ott, A. Chemical Analysis of a 'Miller-Type' Complex Prebiotic Broth Part I: Chemical Diversity, Oxygen and Nitrogen Based Polymers. *Orig. Life Evol. Biosph.* **46,** 149–169 (2016).
212. Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J. & Hoff, K. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30,** 918–920 (2012).
213. Kösters, M., Leufken, J., Schulze, S., Sugimoto, K., Klein, J., Zahedi, R.P., Hippler, M., Leidel, S.A. & Fufezan, C. PymzML v2.0: introducing a highly compressed and seekable gzip format. *Bioinformatics* **34,** 2513–2514 (2018).
214. Marshall, S. M., Murray, A. R. G. & Cronin, L. A probabilistic framework for identifying biosignatures using Pathway Complexity. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **375,** (2017).

215. Benner, S. A., Kim, H. & Biondi, E. *Prebiotic Chemistry and Chemical Evolution of Nucleic Acids*. **35,** 31-83 (2018).
216. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS online: A web-based platform to process untargeted metabolomic data. *Anal. Chem.* **84,** 5035–5039 (2012).
217. Mann, B.F., Chen, H., Herndon, E.M., Chu, R.K., Tolic, N., Portier, E.F., Chowdhury, T.R., Robinson, E.W., Callister, S.J., Wullschleger, S.D. & Graham, D.E. Indexing permafrost soil organic matter degradation using high-resolution mass spectrometry. *PLoS One* **10,** 1–16 (2015).
218. Verrastro, I., Pasha, S., Jensen, K. T., Pitt, A. R. & Spickett, C. M. Mass spectrometry-based methods for identifying oxidized proteins in disease: Advances and challenges. *Biomolecules* **5,** 378–411 (2015).
219. Mamajanov, I. Wet-Dry Cycling Delays the Gelation of Hyperbranched Polyesters: Implications to the Origin of Life. *Life* **9,** 56 (2019).
220. Rode, B. M. Peptides and the origin of life. *Peptides* **20,** 773–786 (1999).
221. Nghe, P. *et al.* Prebiotic network evolution: six key parameters. *Mol. Biosyst.* **11,** 3206–3217 (2015).
222. Cronin, L. & Walker, S. I. Beyond prebiotic chemistry. *Science (80-.).* **352,** 1174-1175 (2016).
223. de Duve, C. & Miller, S. L. Two-dimensional life? *Proc. Natl. Acad. Sci.* **88,** 10014–10017 (2006).
224. Luisi, Pier Luigi; Varela, F. Self-replicating micelles — A chemical version of a minimal autopoietic system. *Orig. Life Evol. Biosph.* **19,** 633–643 (1989).
225. Kauffman, S. A. The Origins of Order. Self-Organization and Selection in Evolution. *J. Evol. Biol.* 731 (1993).
226. Dyson, F. *Origins of Life*. *Cambridge University Press* (1999).
227. Long, W. & Agilent Technologies, I. Automated Amino Acid Analysis Using an Agilent Poroshell HPH-C18. *Agil. Technol.* 1–9 (2015).
228. Lê, S., Josse, J. & Husson, F. FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.* **25,** 1–18 (2008).
229. Jong, S. Book review: Multivariate calibration. *J. Protozool.* **37,** 605–606 (1990).
230. Hung, E. Integrated software solutions for comprehensive metabolomics analysis Compound Discoverer 2.1 Software. 1–49 (2017).
231. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Vanderplas, J. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).
232. Hazen, R. M. Paleomineralogy of the Hadean Eon: A preliminary species list. *Am. J. Sci.* **313,** 807–843 (2013).
233. Joyce, G. F. The antiquity of RNA-based evolution. *Nature* **418,** 214–221 (2002).
234. Bada, J. L. & Lazcano, A. Prebiotic Soup — Revisiting the Miller Experiment. **300,** 745–747 (2003).
235. Cleaves, H. J. Prebiotic Chemistry: What We Know, What We Don't. *Evol. Educ. Outreach* **5,** 342–360 (2012).
236. Ashe, K., Fernández-garcía, C., Corpinot, M. K., Powner, M. W. & Coggins, A. J. Selective prebiotic synthesis of phosphoroaminonitriles and aminothioamides in. *Commun. Chem.* **2**, 23-30 (2019). doi:10.1038/s42004-019-0124-5
237. Saladino, Raffaele, Mauro, Di Ernesto, Garcia-Ruiz, J. M. A Universal Geochemical Scenario for Formamide Condensation and Prebiotic Chemistry. *Chem. - A Eur. J.* **25,** 3181–3189 (2012).
238. Chandru, K., Guttenberg, N., Giri, C., Hongo, Y., Butch, C., Mamajanov, I. & Cleaves, H. J. Simple prebiotic synthesis of high diversity dynamic combinatorial polyester libraries. *Commun. Chem.* **1,** 1–8 (2018).
239. Saladino, R., Di Mauro, E. & García-Ruiz, J. M. A Universal Geochemical Scenario for Formamide Condensation and Prebiotic Chemistry. *Chem.-A Eur. J.* **25,** 3181–

3189 (2019).

Appendix

Gas	D Mix	H Mix
Ammonia	40.6%	41.8%
Methane	39.1%	39.4%
Hydrogen (or Deuterium)	BALANCE	BALANCE

Table A1. The gas mixture used for non-deutarated (‘classic’) and deuterated Miller-Urey experiments.

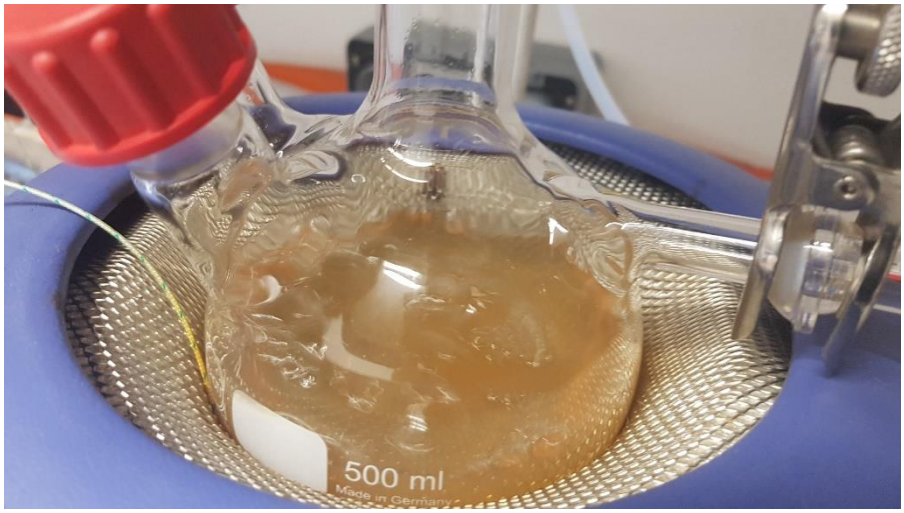


Image A1. *Miller-Urey experiment:* The 500 mL round bottom flask after 7 days of the experiment. A brown solution can be observed.

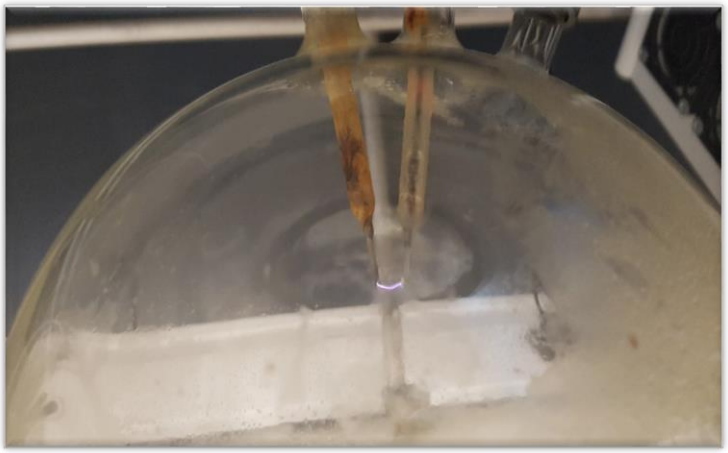


Image A2. Miller-Urey electrodes in mid-spark after 5 days. The degradation of the electrodes is clearly visible.

Image A3. *SEM-EDS:* Colour-based differences in the dried material of Miller-Urey experimental replicates.

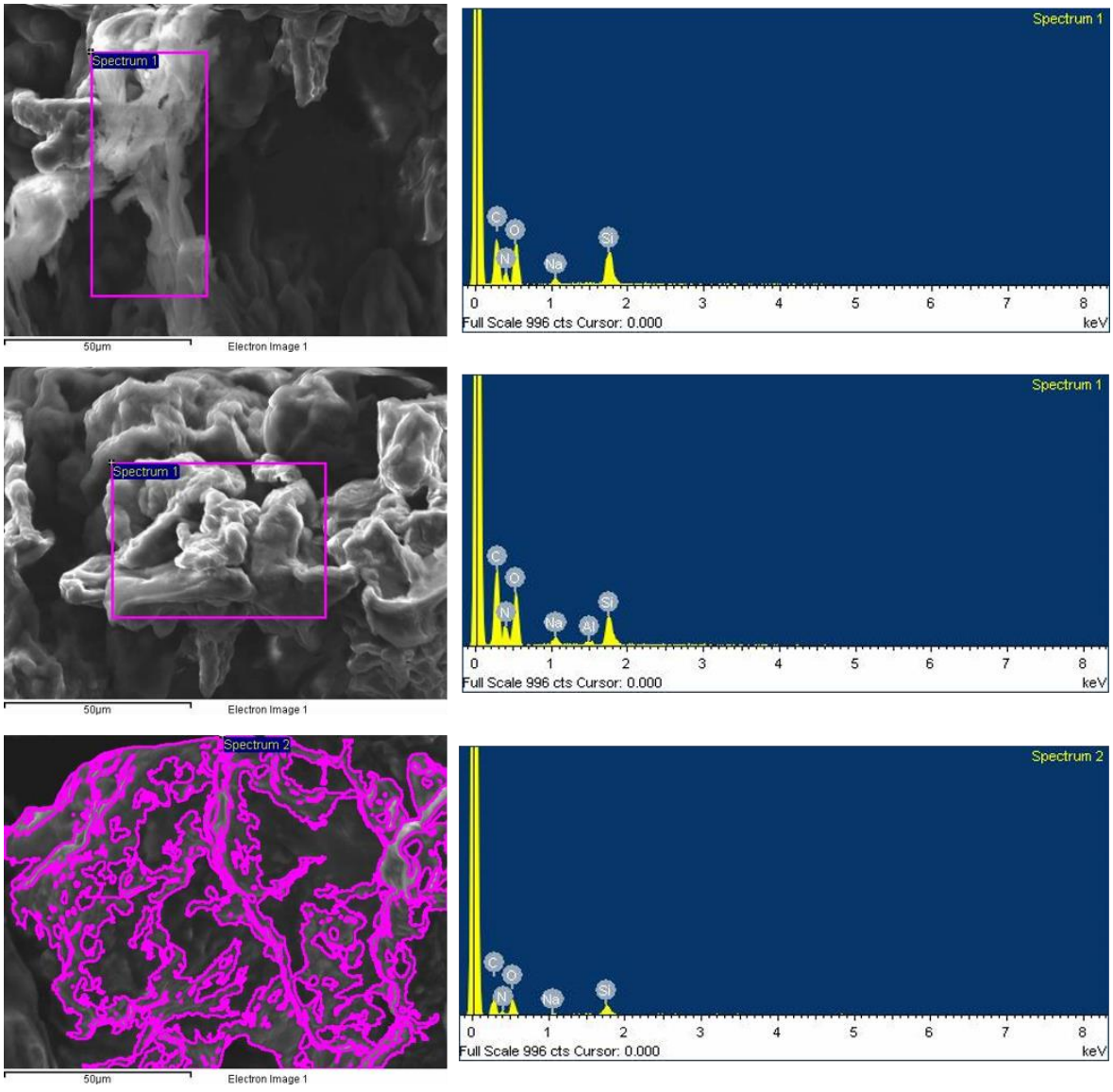


Figure A1. *SEM-EDS of the deuterated Miller-Urey samples:* Fraction taken for the EDS is highlighted in pink in the SEM image.

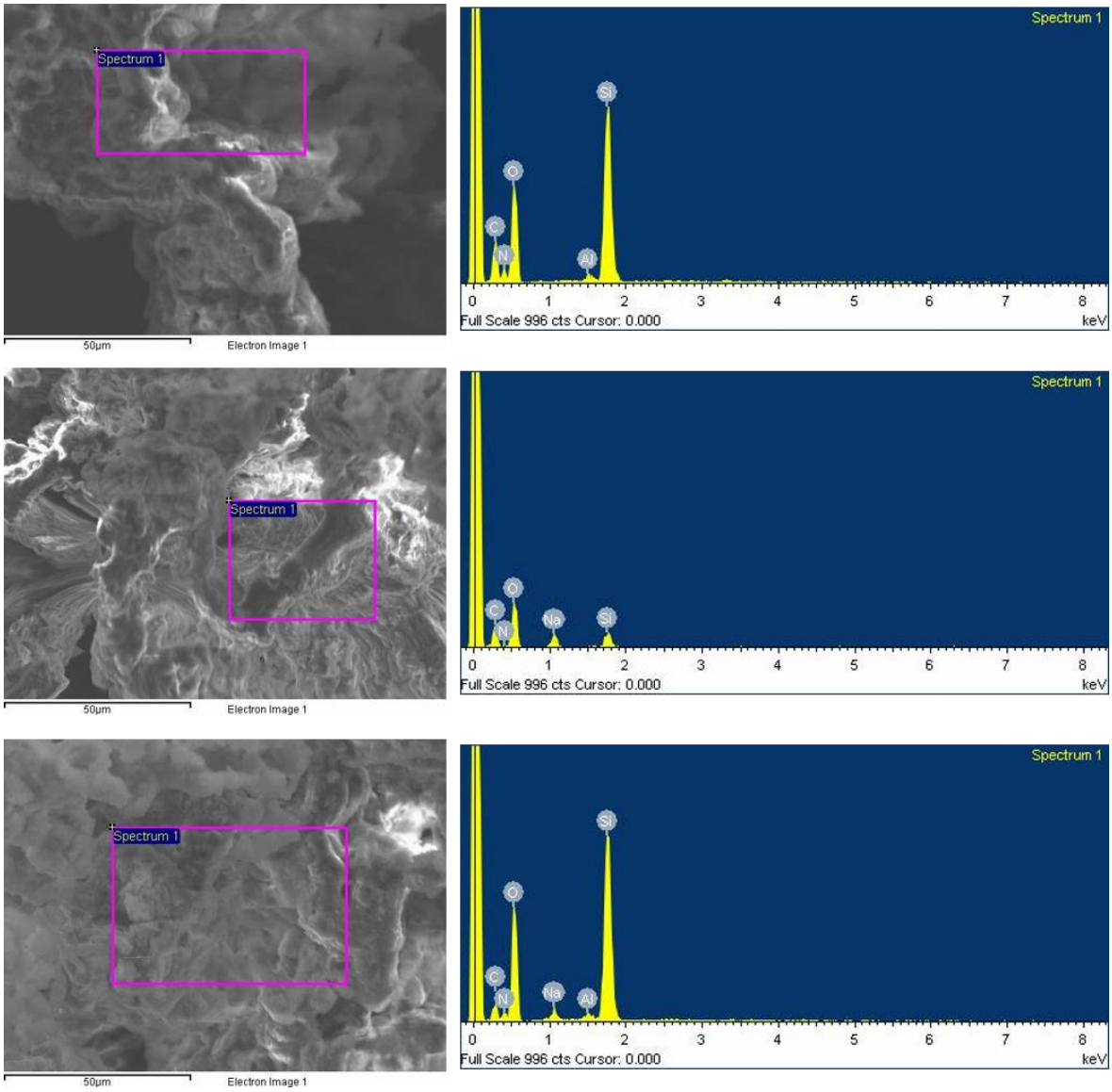


Figure A2. *SEM-EDS of the protonated ('classic') Miller-Urey samples:* Fraction taken for the EDS is highlighted in pink in the SEM image.

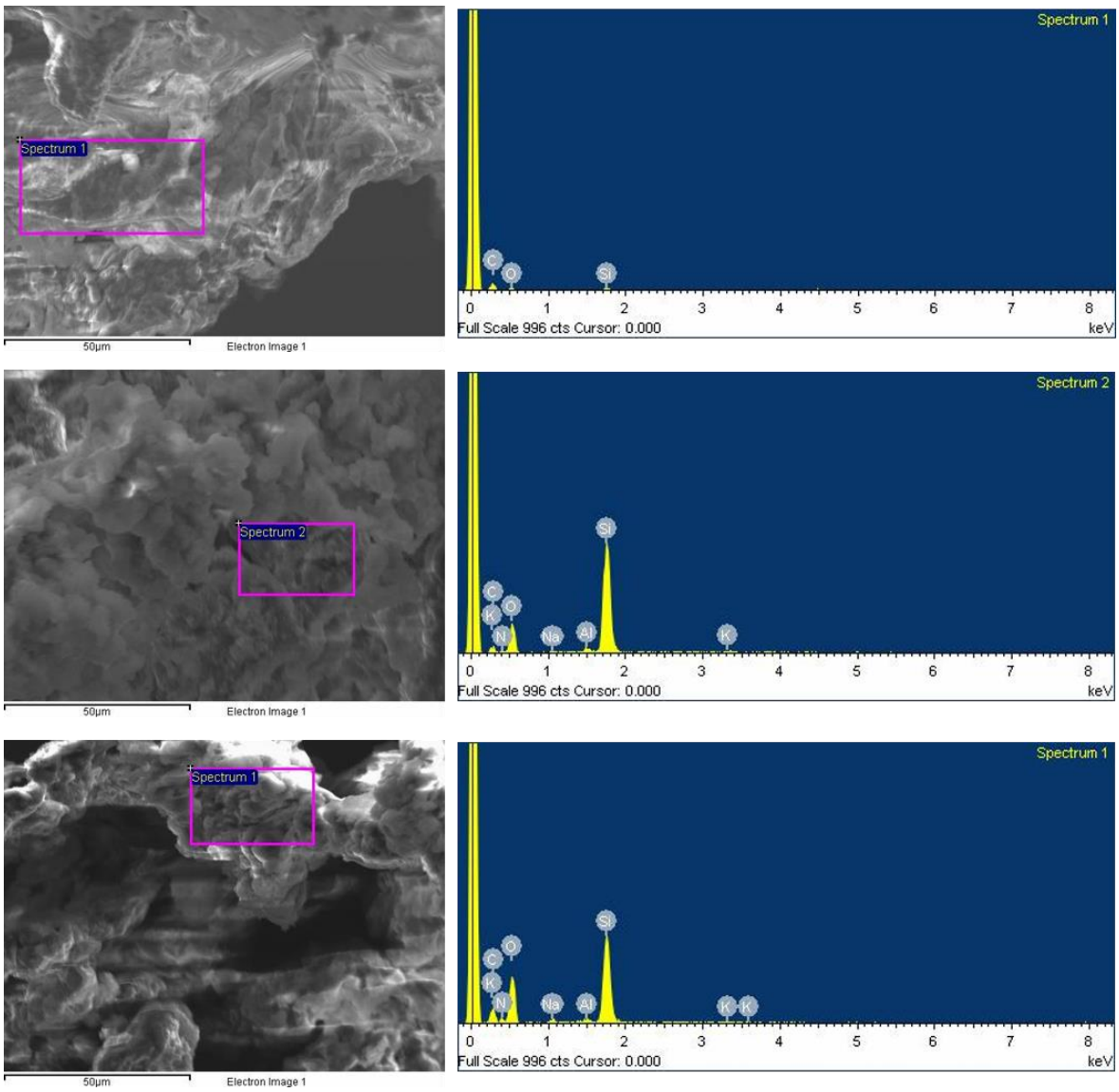
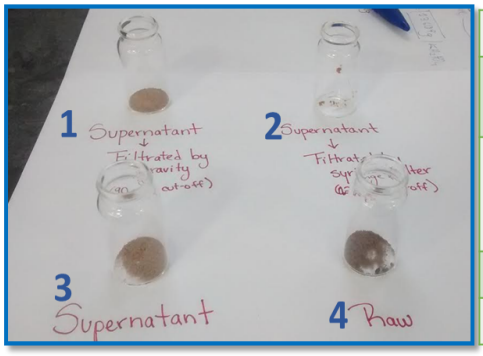


Figure A3. *SEM-EDS of the protonated ('classic') Miller-Urey samples:* Fraction taken for the EDS is highlighted in pink in the SEM image.



#	Description	Recovery
1	Centrifuged (Supernatant) and filtered by gravity (0.90 mm)	0.129%
2	Centrifuged (Supernatant) and filtered with syringe filter (0.22 µm)	0.071%
3	Centrifuged (Supernatant)	0.137%
4	Raw- nothing done	0.225%

Figure A4. *Sample preparation of the Miller-Urey samples:* Filtration prior to lyophilisation with different methods (e.g. supernatant filtered by gravity (**1**), supernatant after filtration by syringe filter (**2**), supernatant (**3**), no filtration (**4**). Recovery percentage is calculated from the dried material, assuming the density of pure water for conversion).

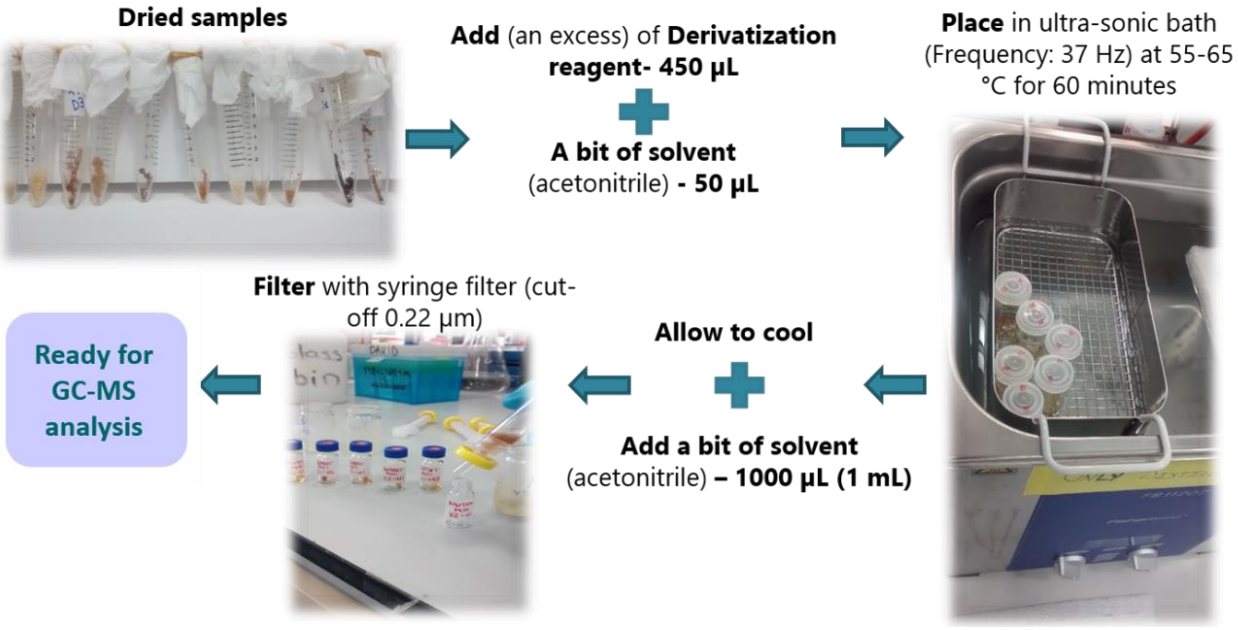


Figure A5. *Sample preparation of the Miller-Urey samples:* A Step by Step procedure of the derivatisation reaction with MSTBFA.

Image A4. Agilent 7890 GC / 5975 MSD GC-MS instrument.

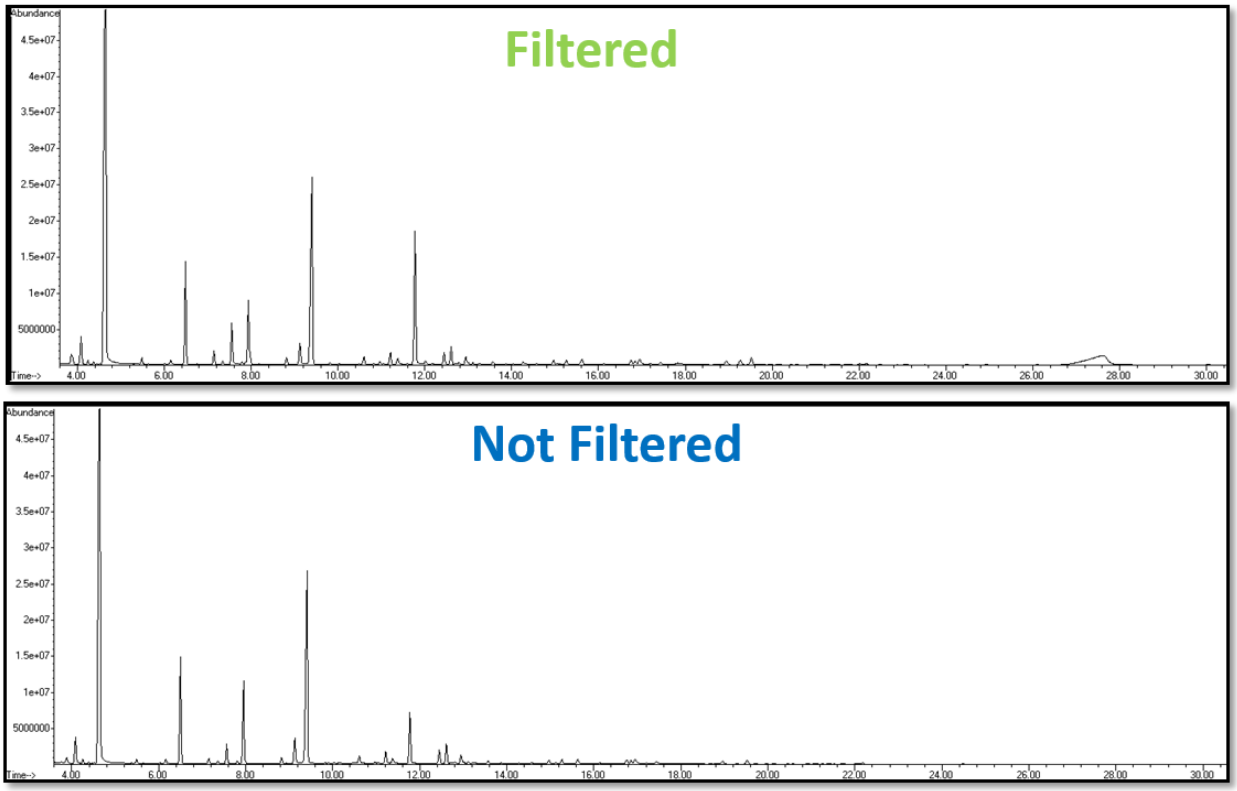


Figure A7. *Sample preparation of the Miller-Urey samples:* Two GC-MS Total- Ion Chromatograms (TIC's) for the filtered and non-filtered samples, prior to analysis and post derivatisation.

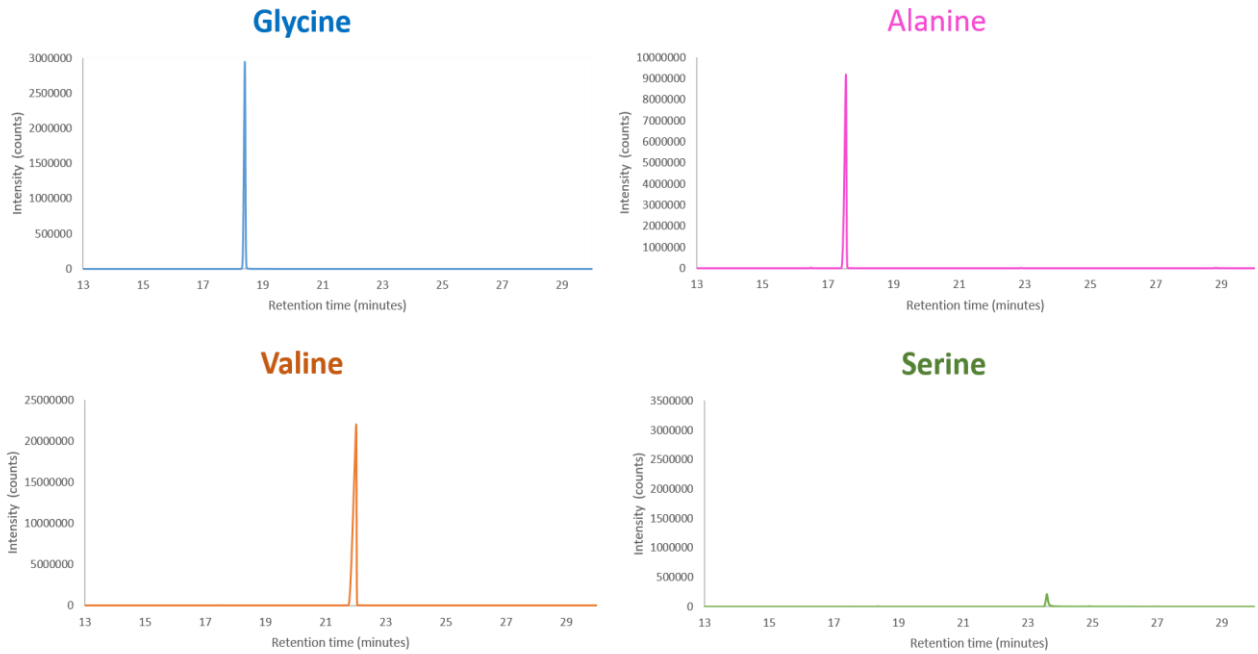


Figure A6. Gas Chromatograms for four amino acid standards solutions (50 mM, in acetonitrile) known to be present in the Miller-Urey experiment: *glycine*, *alanine*, *valine* and *serine*.

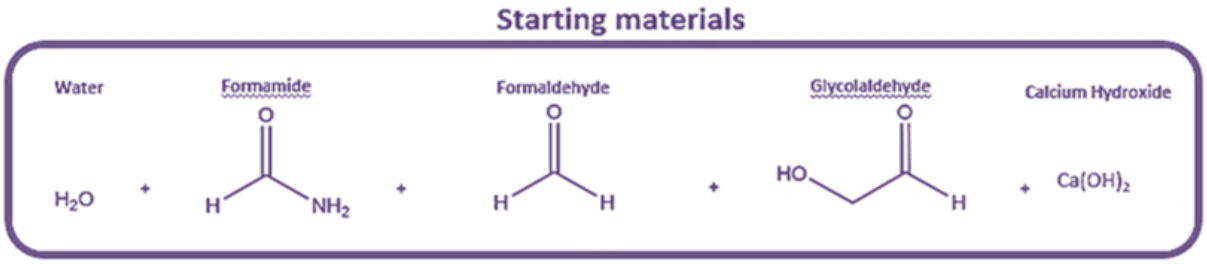
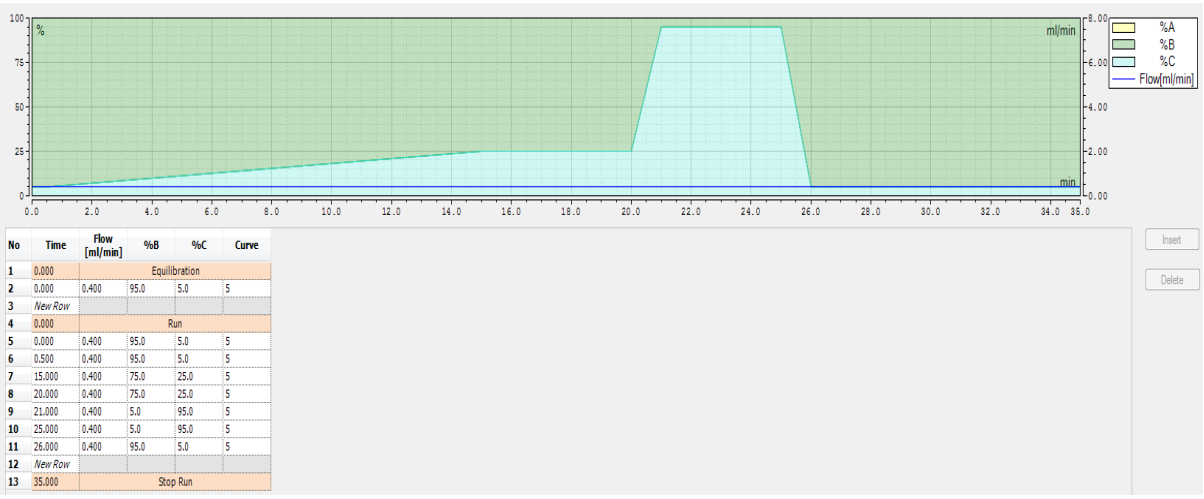


Figure A9. Starting Materials for the formose in formamide model system for complex prebiotic mixtures.

Image A5. Thermo-Dionex UltiMate3000 UPLC system equipped with a Charged Aerosol Detector (UPLC- CAD)



No	Time	Flow [ml/min]	%B	%C	Curve
1	0.000	Equilibration			
2	0.000	0.400	95.0	5.0	5
3	New Row				
4	0.000	Run			
5	0.000	0.400	95.0	5.0	5
6	0.500	0.400	95.0	5.0	5
7	15.000	0.400	75.0	25.0	5
8	20.000	0.400	75.0	25.0	5
9	21.000	0.400	5.0	95.0	5
10	25.000	0.400	5.0	95.0	5
11	26.000	0.400	95.0	5.0	5
12	New Row				
13	35.000	Stop Run			

Figure A8. *UPLC-CAD analysis:* Elution gradient for the HILIC method. Also employed in the UPLC-MS/MS analysis.

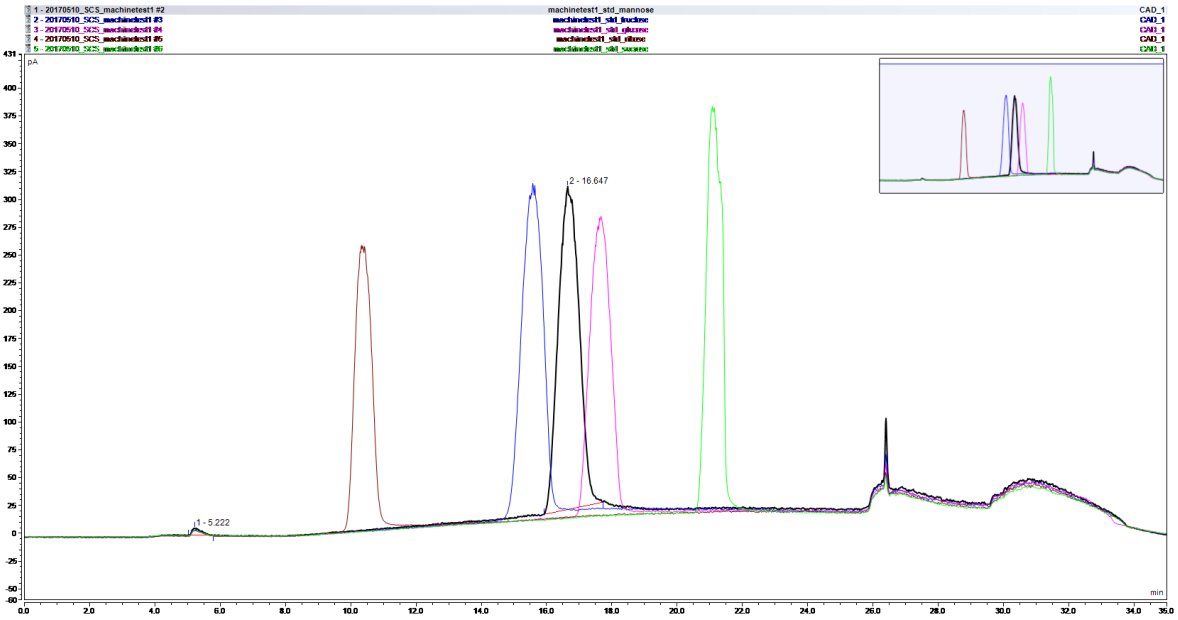


Figure A10. *UPLC – CAD chromatogram for the standard mixture of sugars*: ribose shown in *brown*, fructose shown in *blue*, mannose shown in *black*, glucose shown in *pink* and sucrose shown in *green*.

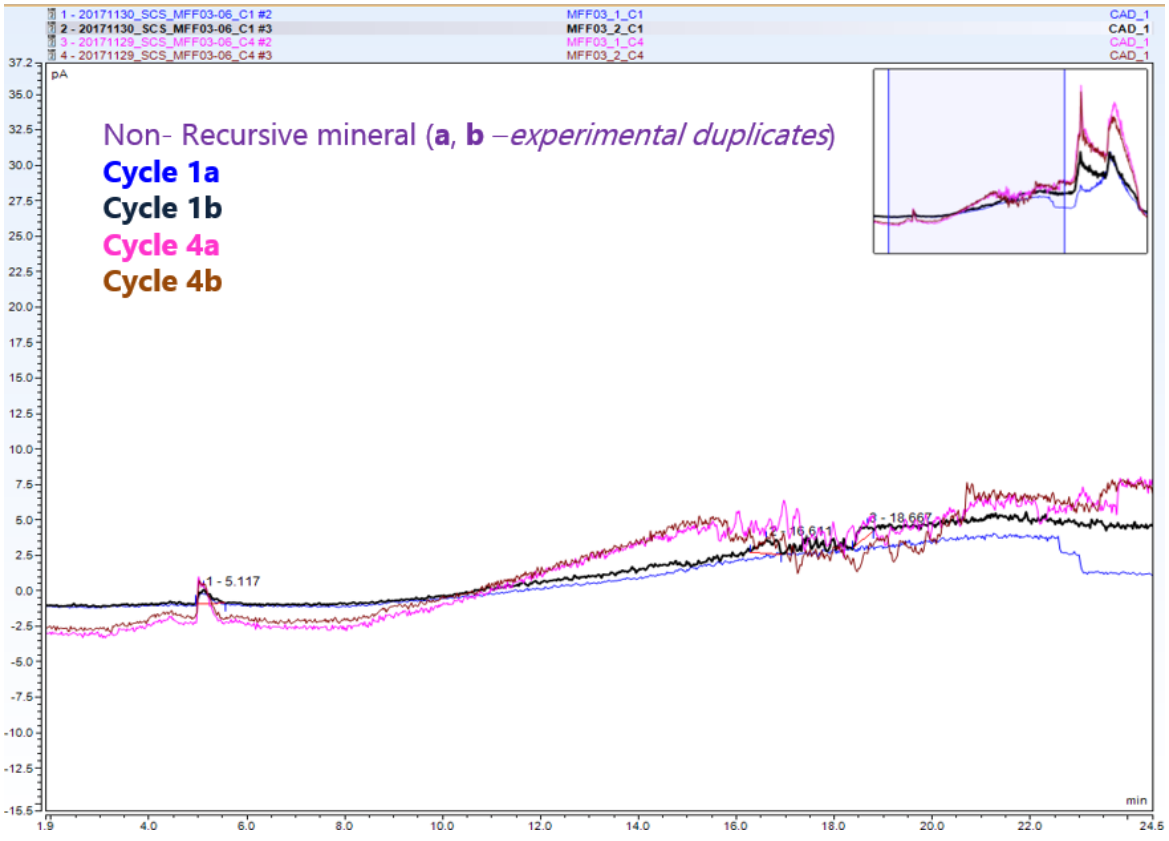


Figure A11. *UPLC – CAD chromatogram for mineral control (non-recursive):* Cycle 1 and Cycle 4 time-points. Experimental duplicates are presented as (**a**) and (**b**).

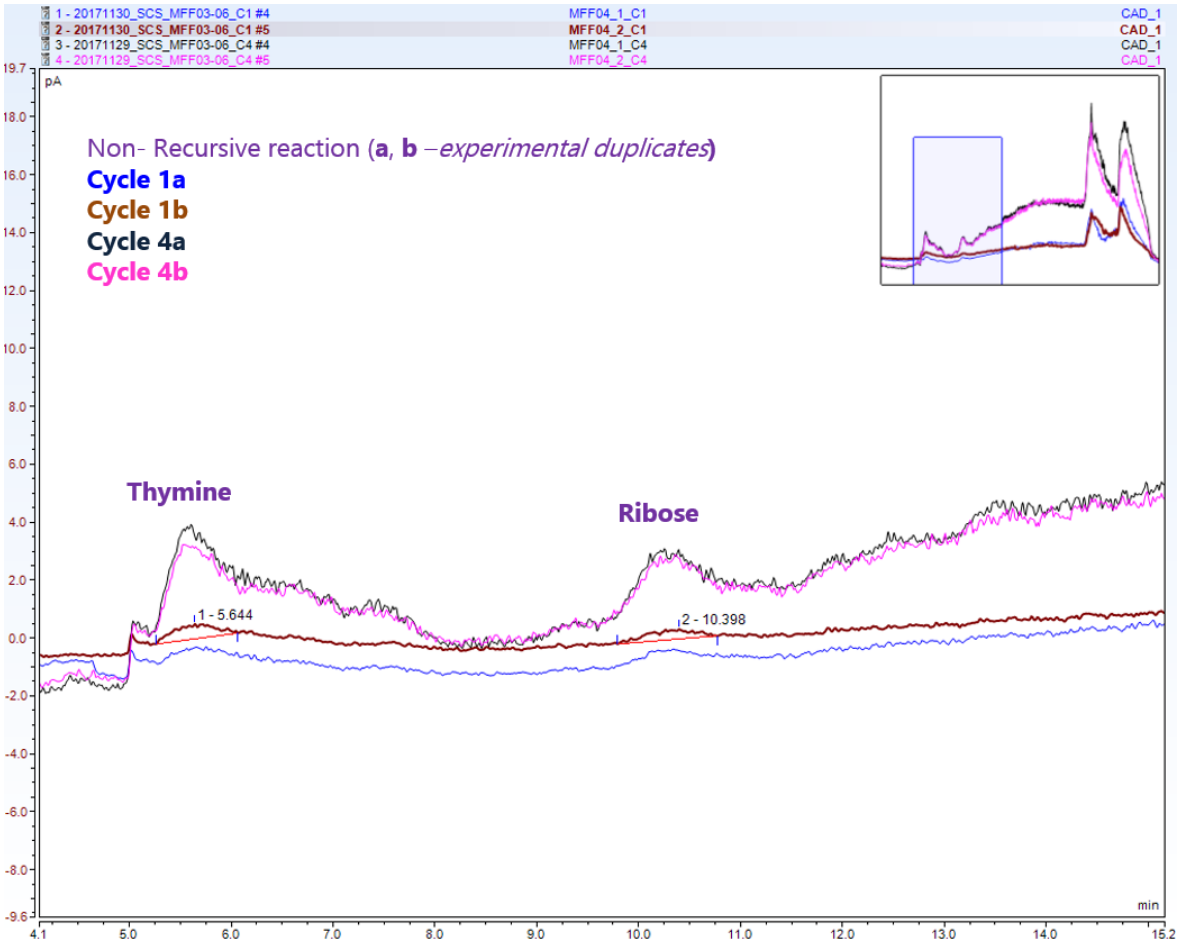


Figure A12. *UPLC – CAD chromatogram for the recursive reaction (no mineral):* Cycle 1 and Cycle 4 time-points. Experimental duplicates are presented as (**a**) and (**b**).

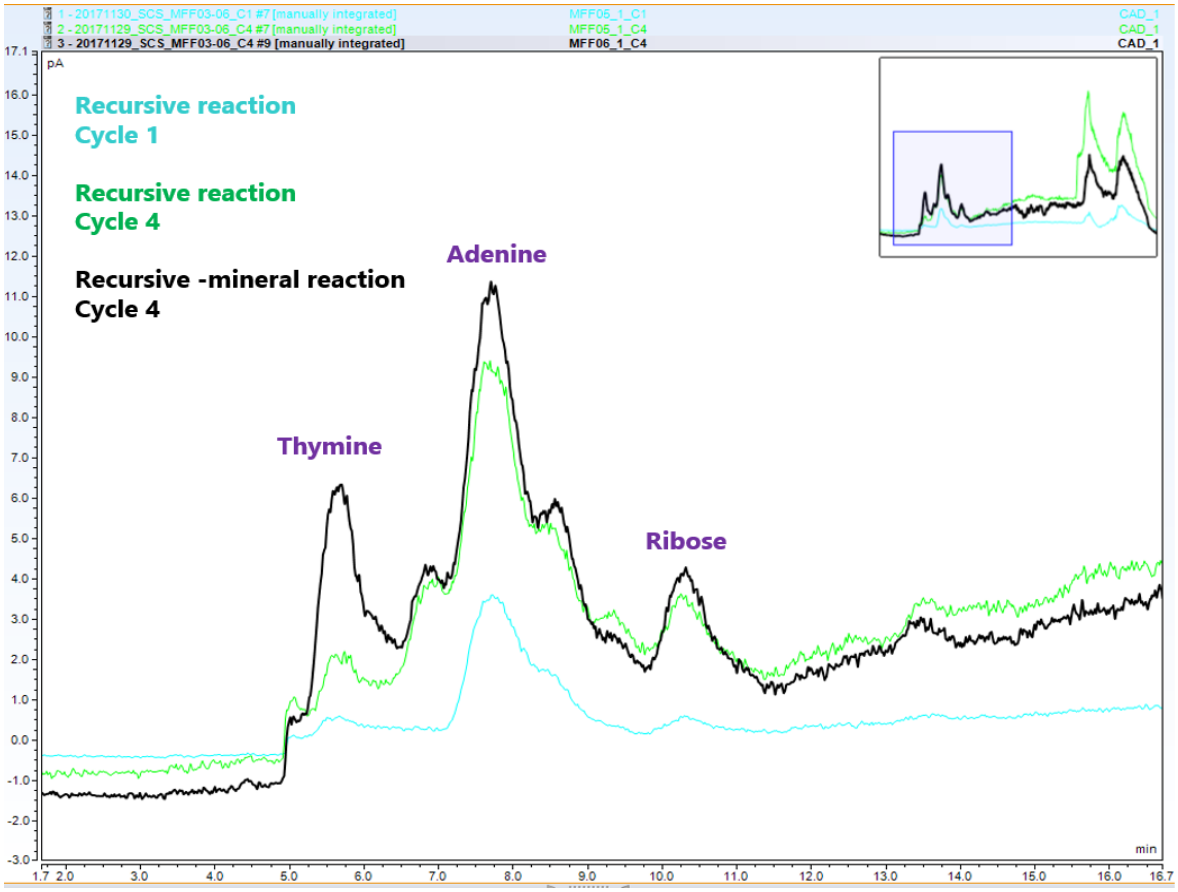


Figure A13. *UPLC – CAD chromatogram for the recursive mineral reaction (with chalcopyrite):* Cycle 1 and Cycle 4 time-points. Experimental duplicates are presented as (**a**) and (**b**).

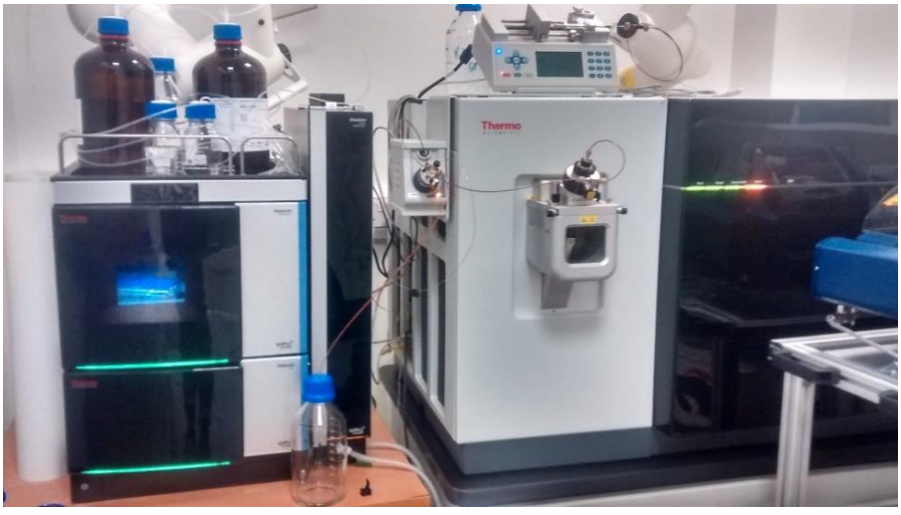


Image A6. *UPLC-Orbitrap:* Thermo Vanquish Ultra-performance liquid chromatography system coupled to a Thermo Orbitrap Fusion Mass-Spectrometer.

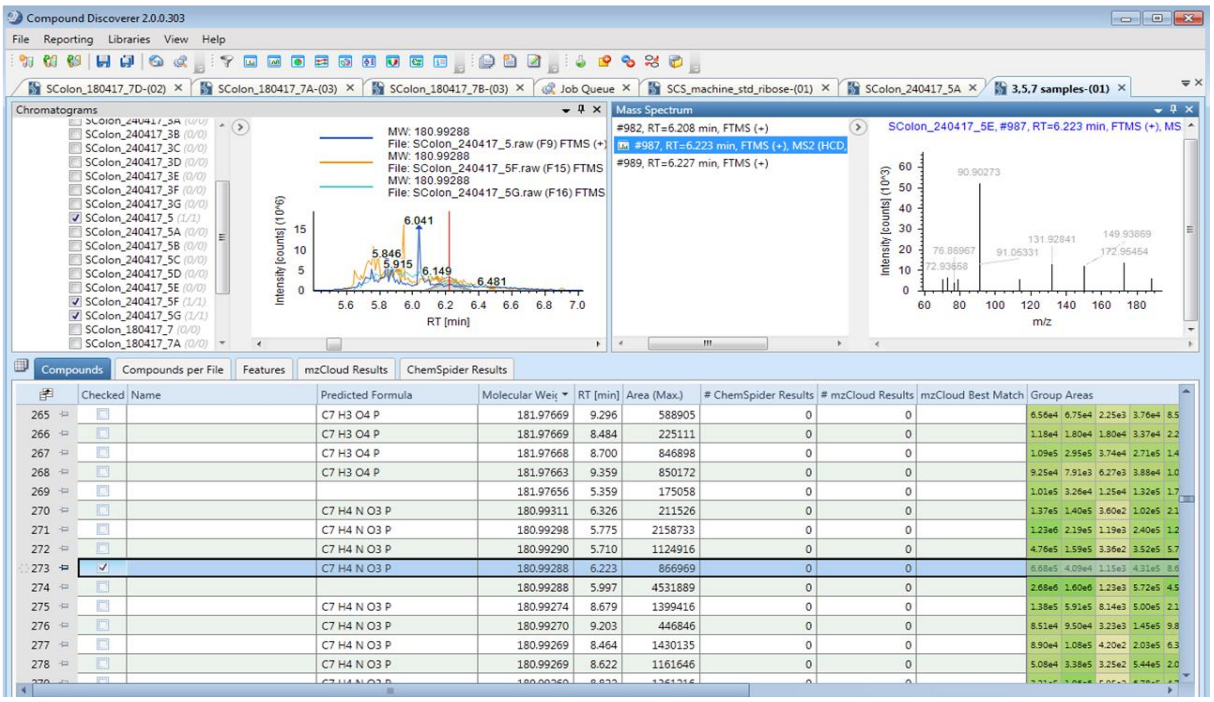


Figure A14. *General overview of the analytical workflow in CompoundDiscoverer 2.0™:* Automated extraction of features and their MS/MS fragmentation (*top-right*). Extracted Ion Chromatograms (EIC's) are generated for all features in each sample (*top-left*).

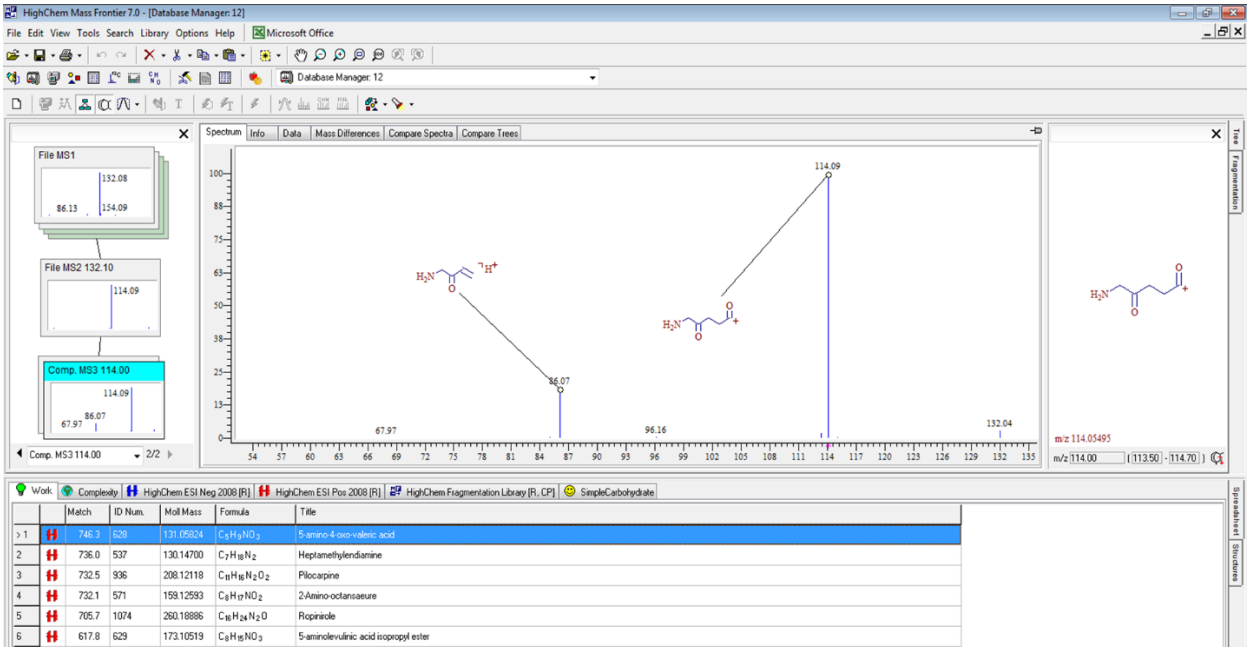


Figure A15. *Mass-Frontier® data-base search:* Integrated substructure annotation for each fragment is enabled within the software (i.e. FiSH), giving an insight to the chemical compositions behind every fragment that constructs the resulting MS/MS spectra.

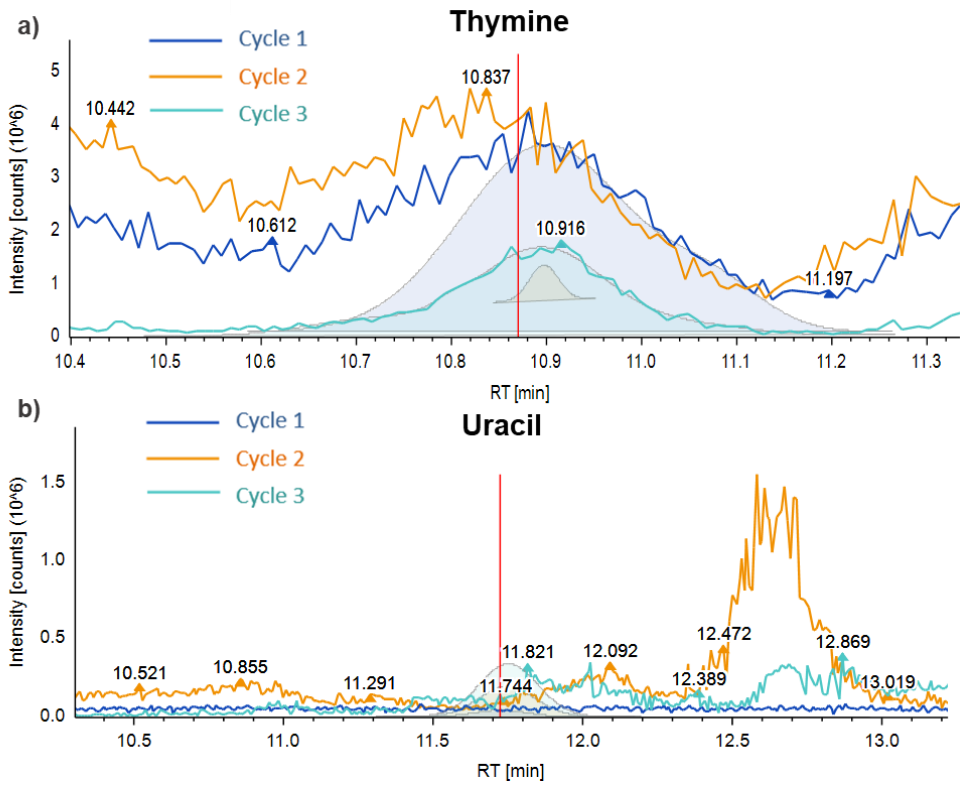


Figure A16. *UPLC-MS/MS of the formose in formamide samples:* Extracted Ion Chromatogram for **(a)** thymine (Molecular Weight: 127.05, Adduct: [M+H]), in Cycle 1 to Cycle 3, recursive reaction with chalcopyrite; and uracil (*m/z:* 113.03, Adduct: [M+H]), in Cycle 1 to Cycle 3, recursive reaction with goethite.

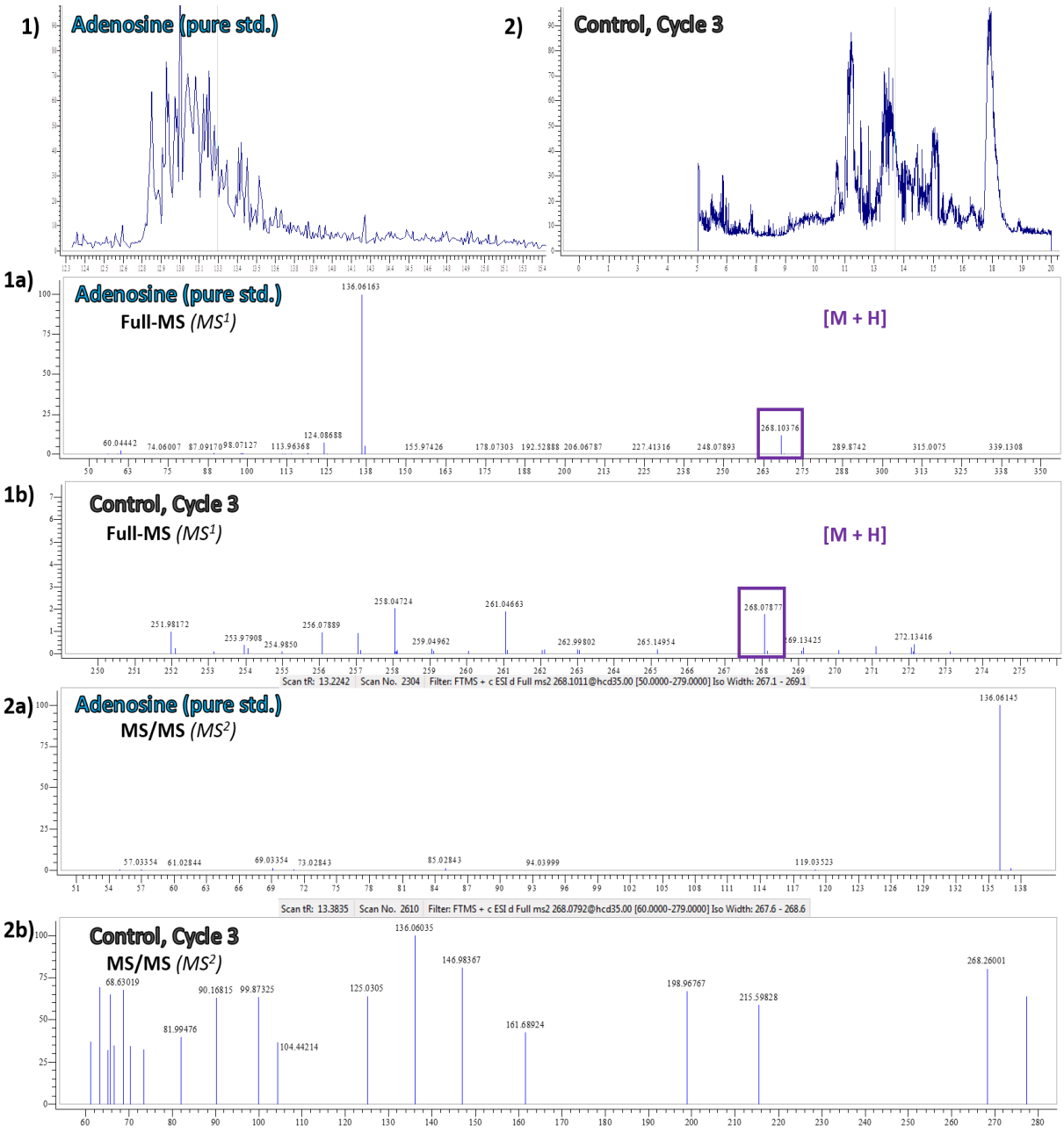


Figure A17. UPLC-MS/MS analysis and comparison of (**1a-1b**) a pure standard of adenosine (m/z: 268.1), Adduct: [M+H]) with (**2a-2b**) a real sample (chalcopyrite, Cycle 3).

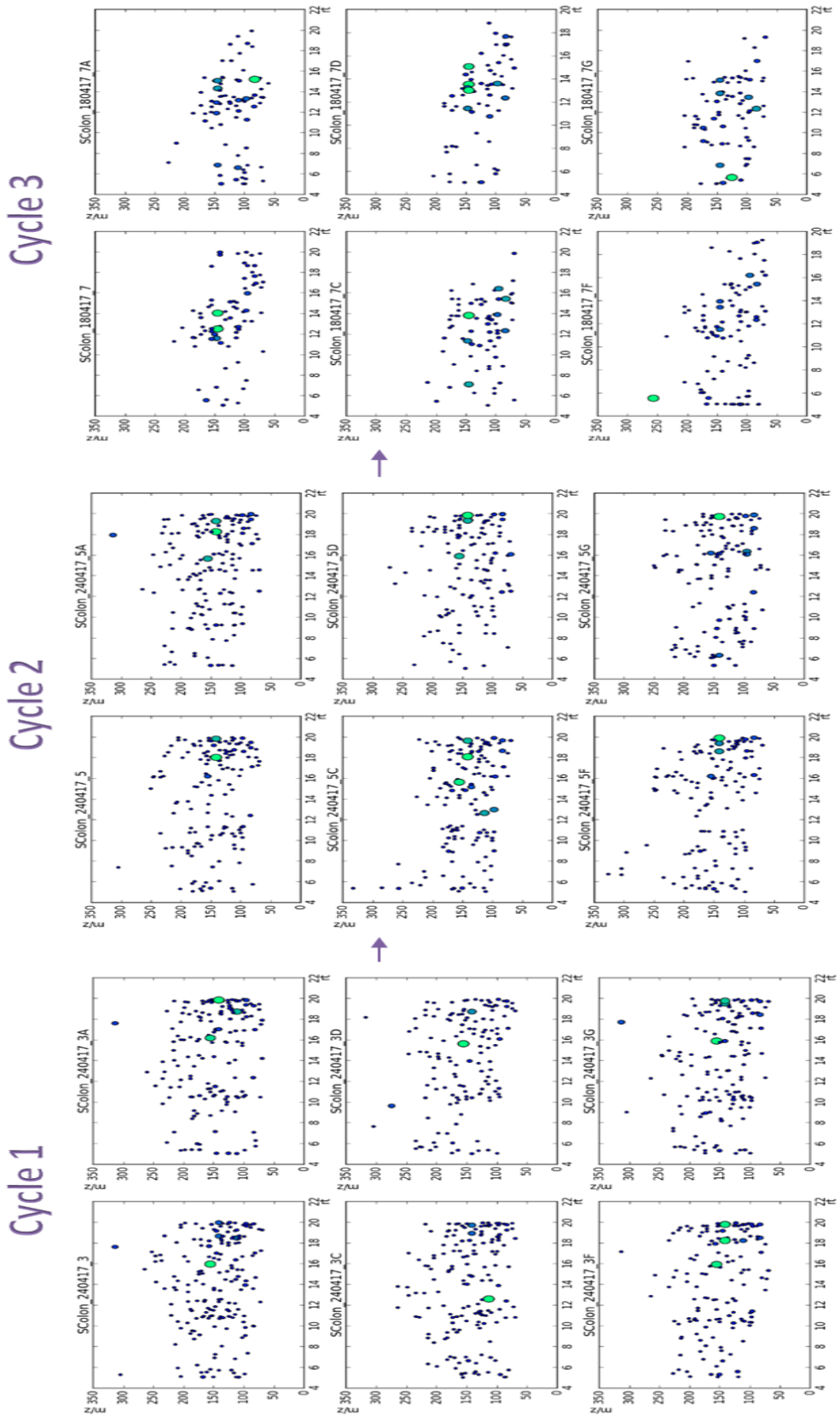


Figure A18. *Feature Generation of the Formose in Formamide mass-spectral features: From* Cycle 1 to Cycle 3, under multiple mineral environments, **A-G**).

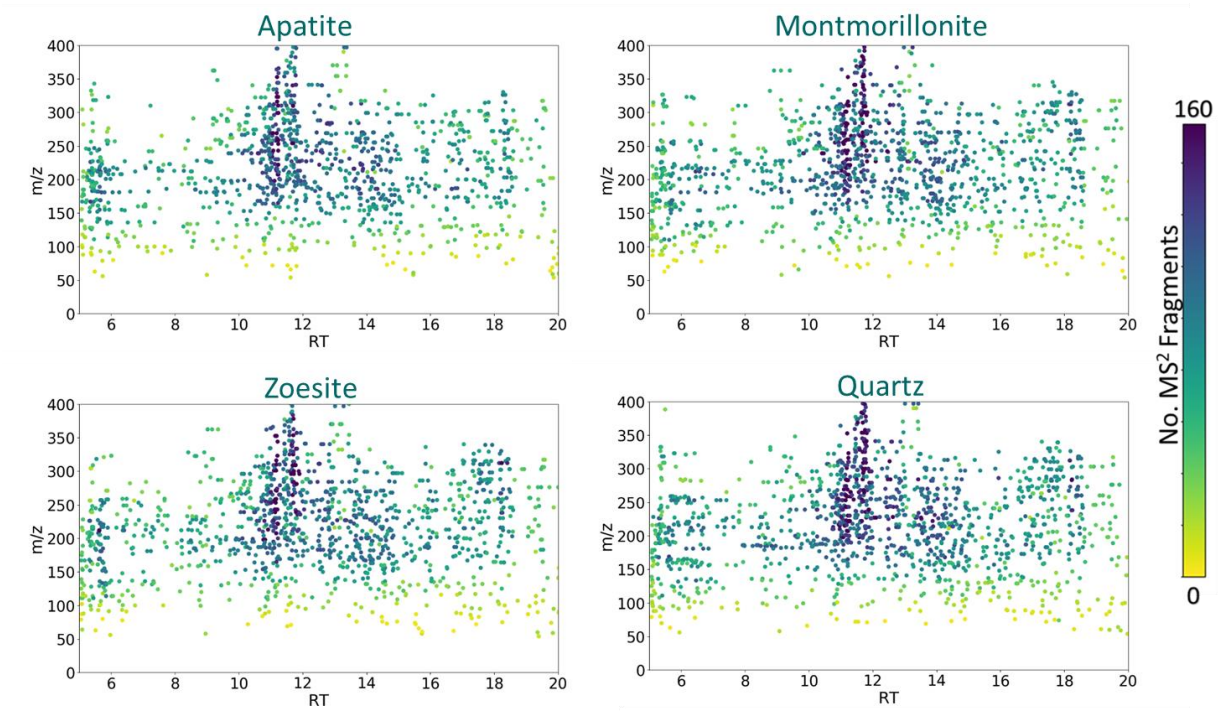


Figure A19. *The Long cycle:* Features are generated by the –in-house feature generator. Distinctive feature patterns can be observed for all mineral-containing cycles as an effect of the mineral environment.

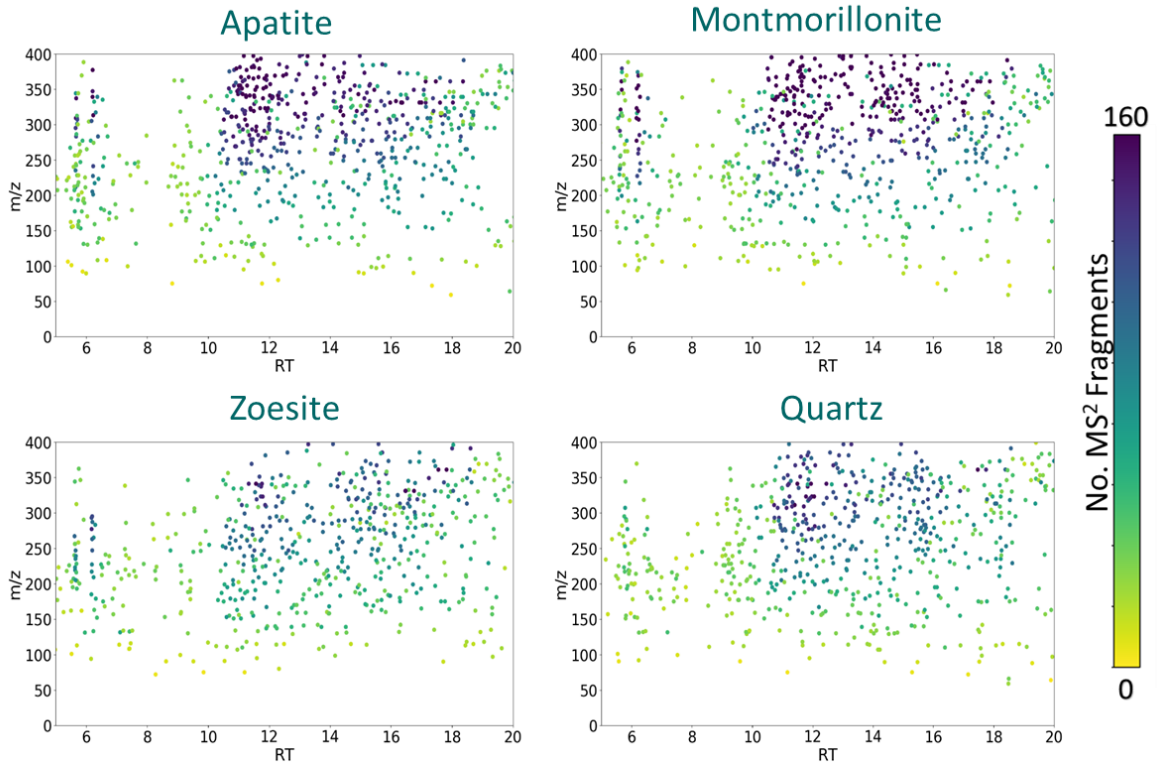


Figure A20. *Non-recursive controls:* The number of features detected is higher than those found for the third (3) Recursive Cycle. The trend was conserved in the presence of all mineral environments, including apatite, montmorillonite, zoesite and quartz.

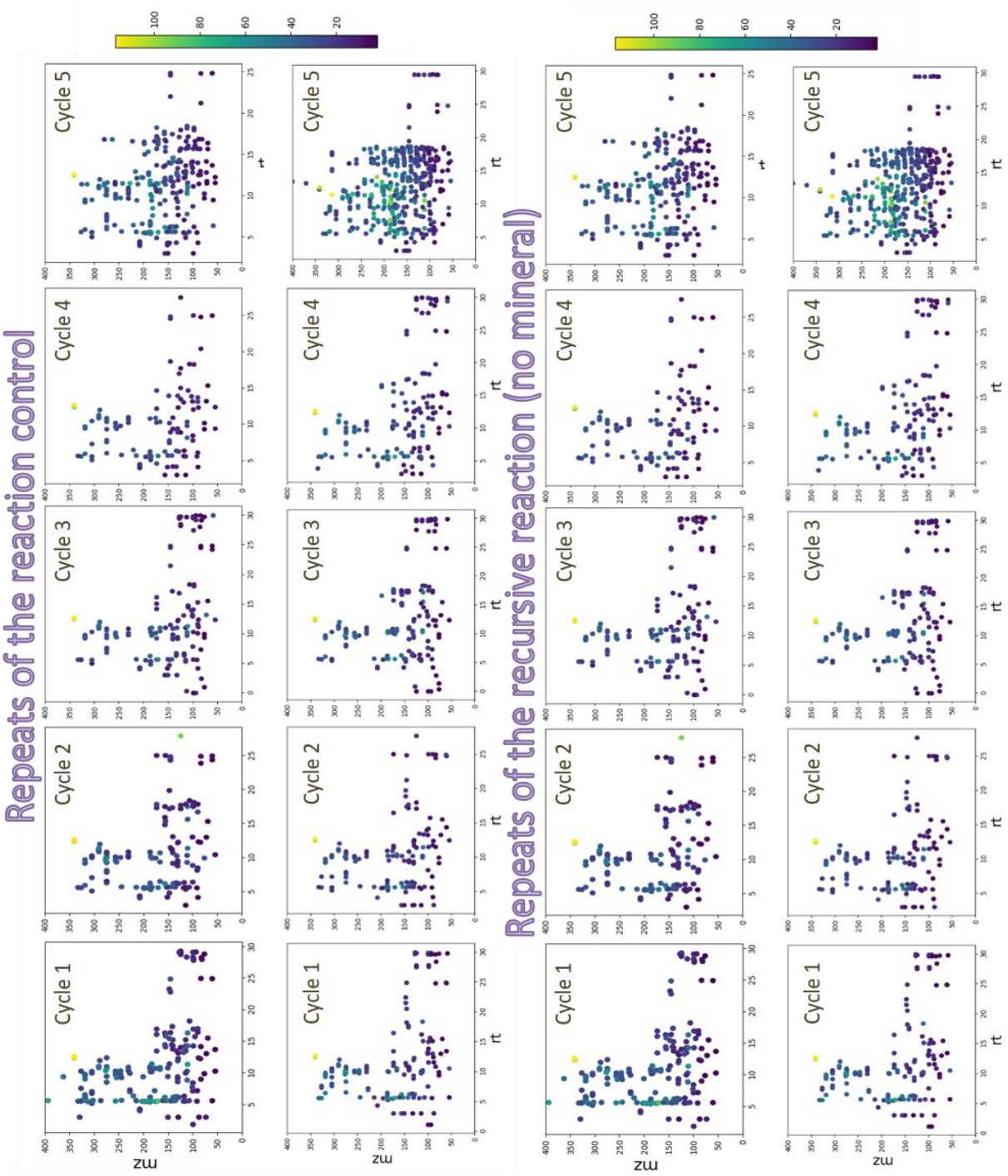


Figure A21. *Reproducibility Assessment:* Feature Generation of the formose in formamide mass-spectral features for the non-recursive reaction (reaction control) and the recursive reaction, Cycle 1 to Cycle 5.

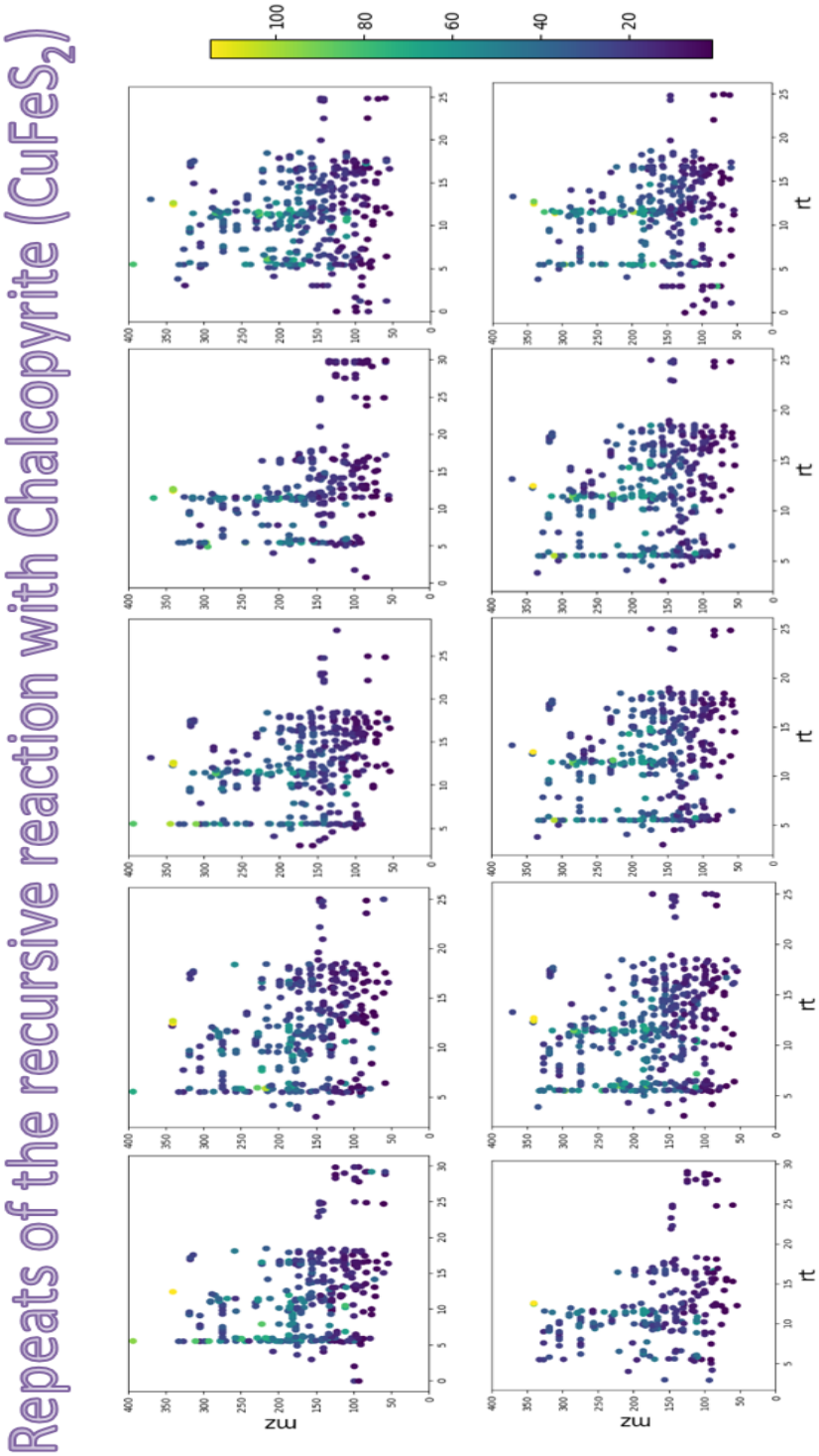


Figure A22. *Reproducibility Assessment:* Feature generation of the formose in formamide mass-spectral features in the presence of chalcopyrite, Cycle 1 to Cycle 5.

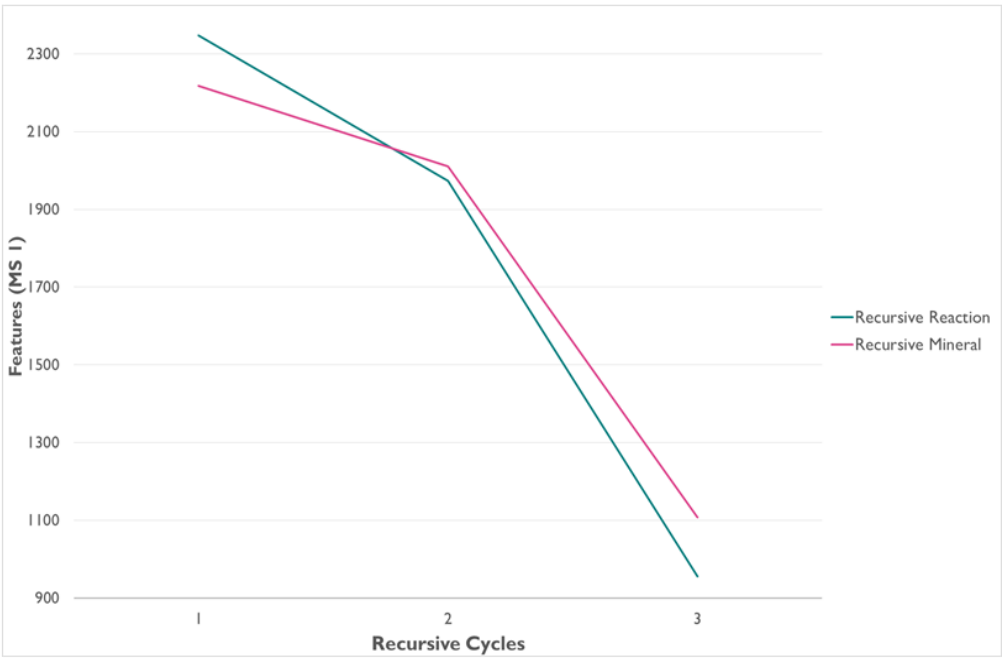


Figure A23. *Reproducibility Assessment:* Detected features for the experimental replicates. Number of (MS^1) features in the experiment with 3 cycles (Recursive reaction, *green* - Recursive mineral, *pink*), in a range of 900 to 2400 features detected across cycles.

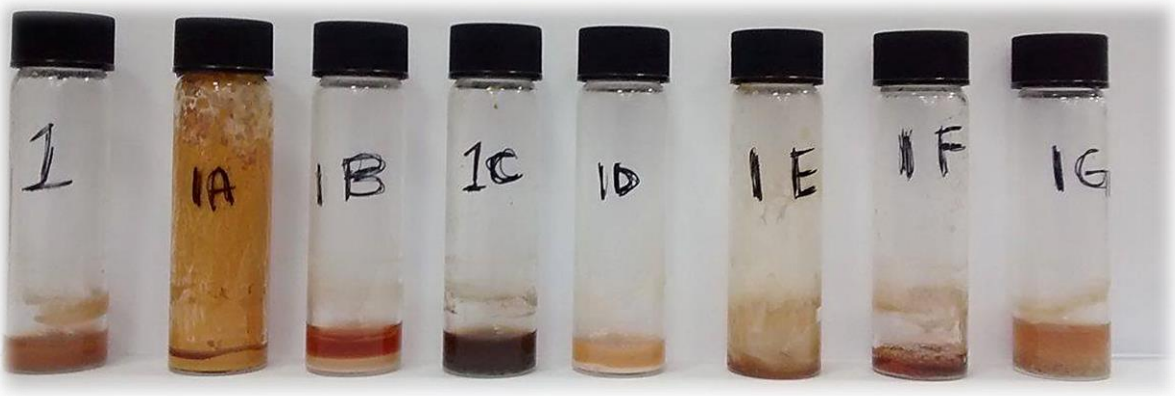


Image A7. *Comparison test:* Formose reaction vessels after the recursive cycles were completed, differences across the mineral surfaces employed can be observed.

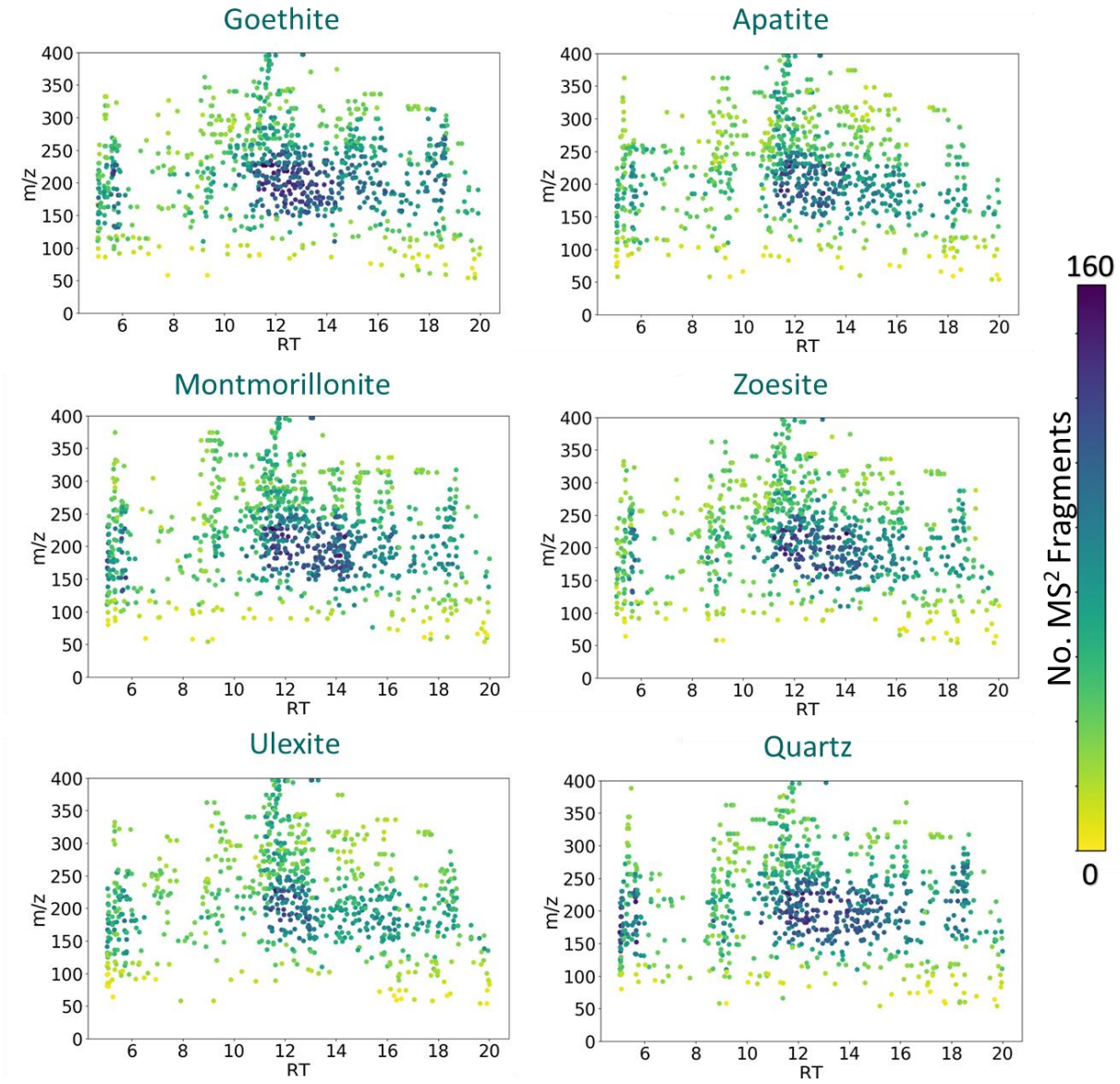


Figure A24. *Formose Reaction in a selection of environments:* Mass-spectral features for the resulting product distribution in the presence of goethite, apatite, montmorillonite, zoesite, ulexite or quartz mineral surfaces.

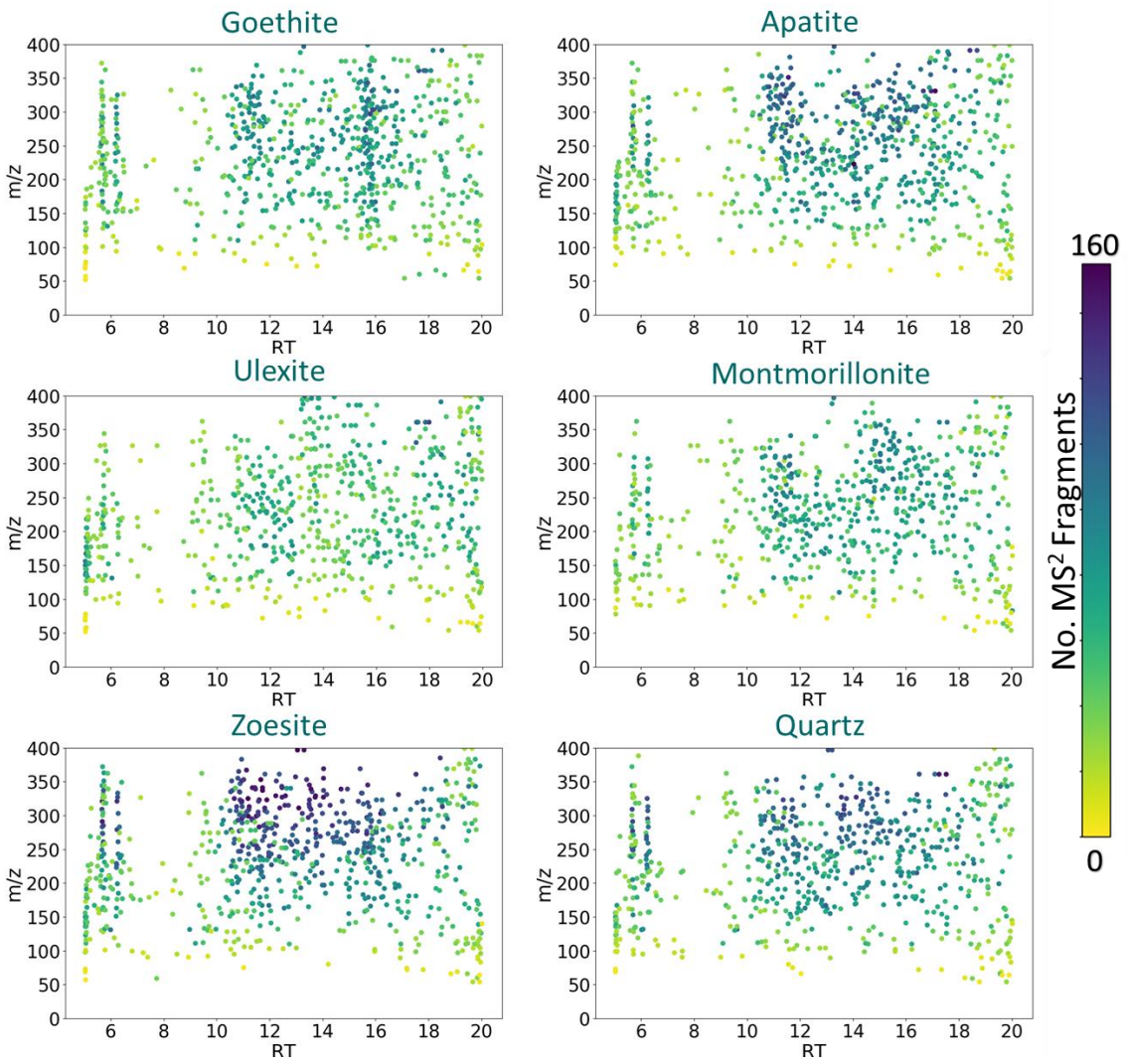


Figure A25. *Formamide condensation in a selection of environments:* Mass-spectral features for the resulting product distribution in the presence of goethite, apatite, montmorillonite, zoesite, ulexite or quartz mineral surfaces.

Job#1207768 : cycl 1-3

Columns | Hide isotopic peaks | Page 1 of 16 | 100 | View 1 - 100 of 1,586

featureidx	pvalue	qvalue	CV	mzmed	rtmed	maxint	isotopes	adducts	peakgroup	usernotes
1	4.29407e-15	8.01344e-13	0.000	83.0600	17.43	46,887,892	[8][M]+		1	
2	3.14385e-14	2.93347e-12	0.000	81.0443	14.38	1,861,346			9	
3	4.31773e-13	2.68587e-11	0.000	155.0917	17.11	10,265,568		[M+Na]+ 132.098	112	
4	5.60229e-12	2.61370e-10	0.000	141.1834	17.50	879,114		[2M+Na]+ 59.1041	1	
5	6.80627e-11	2.19796e-9	0.000	71.0600	18.14	1,809,604			48	
6	9.67569e-11	2.73748e-9	0.000	128.0700	11.65	4,009,268		[M+H]+ 127.062	145	
7	1.21647e-10	3.10770e-9	0.000	109.0753	18.69	1,154,962		[M+H-CO]+ 136.066 [M+H-	2	
8	1.48086e-10	3.42962e-9	0.000	100.0387	5.38	207,671			173	
9	1.81169e-10	3.75657e-9	0.000	58.0647	14.10	2,153,119			196	
10	3.25263e-10	5.94576e-9	0.000	90.0545	5.43	637,228			194	
11	4.40582e-10	7.35661e-9	0.000	95.0598	10.83	836,864			142	
12	6.07611e-10	9.01525e-9	0.000	74.0596	19.87	3,333,415			91	
13	6.57961e-10	9.44510e-9	0.000	147.0746	18.73	148,164		[M+K]+ 108.111	257	
14	9.71952e-10	1.20283e-8	0.000	145.0599	10.86	2,466,714		[M+K]+ 106.095	142	
15	1.01988e-9	1.23613e-8	0.000	91.0537	5.40	116,324			55	
16	1.11093e-9	1.29573e-8	0.000	58.0647	17.84	19,826,414		[M+H]+ 57.058	12	
17	1.49683e-9	1.56247e-8	0.000	190.1059	11.50	142,145		[M+H-C6H8O6]+ 365.132	121	
18	1.72686e-9	1.69639e-8	0.000	97.0755	5.46	2,859,975		[2M+Na]+ 37.0509	171	
19	1.75816e-9	1.71340e-8	0.000	130.0491	12.33	4,585,704		[M+H-CH2]+ 143.066	143	
20	1.93304e-9	1.80369e-8	0.000	118.0686	15.16	366,767	[30][M+1]+		136	
21	2.29197e-9	1.99967e-8	0.000	58.0648	18.25	7,162,488		[M+H-H20]+ 75.0792	44	
22	3.08174e-9	2.35246e-8	0.000	125.0701	18.03	4,873,882			154	
23	3.23557e-9	2.41115e-8	0.000	126.1270	12.88	2,683,673		[3M+H+Na]2+ 76.0863	147	
24	3.27873e-9	2.42712e-8	0.000	99.0628	5.38	308,562		[M+H]+ 98.052	173	
25	3.67735e-9	2.56518e-8	0.000	186.0862	11.44	9,459,129	[65][M]+	[M+H]+ 185.076	77	
26	3.87537e-9	2.62801e-8	0.000	182.1889	5.43	93,585			194	
27	3.88600e-9	2.63128e-8	0.000	144.9812	16.95	10,683,487			107	
28	3.99957e-9	2.66566e-8	0.000	95.0599	18.67	7,054,872		[M+H-COCH2]+ 136.066	2	
29	5.43884e-9	3.34086e-8	0.000	97.0755	9.65	1,050,173			222	
30	5.69692e-9	3.45090e-8	0.000	170.1390	17.48	2,819,194			1	
31	5.81321e-9	3.49949e-8	0.000	130.0855	5.33	365,030		[M+H+HCOOH]+ 83.0744	123	

Columns | Export | Page 1 of 16 | 100 | View 1 - 100 of 1,586

Table A2. *XCMS online:* Automatically extracted mass-spectral features from the XCMS workflow. Average (*med*) values are taken for the m/z, retention time and intensity of each feature. Multiple adducts have also been considered.

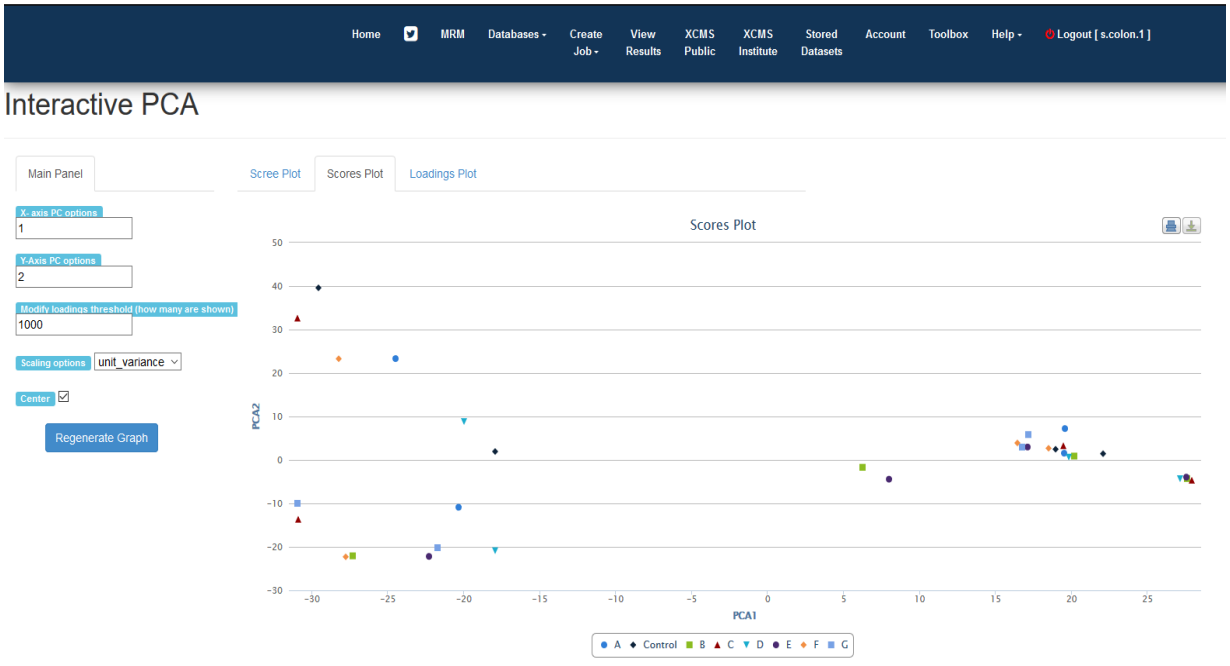


Figure A26. *XCMS online:* Generation and parameters for the PCA analysis of the recursive formose-in-formamide reaction in the presence of different mineral surfaces (**A-G** and -no mineral- **control**).

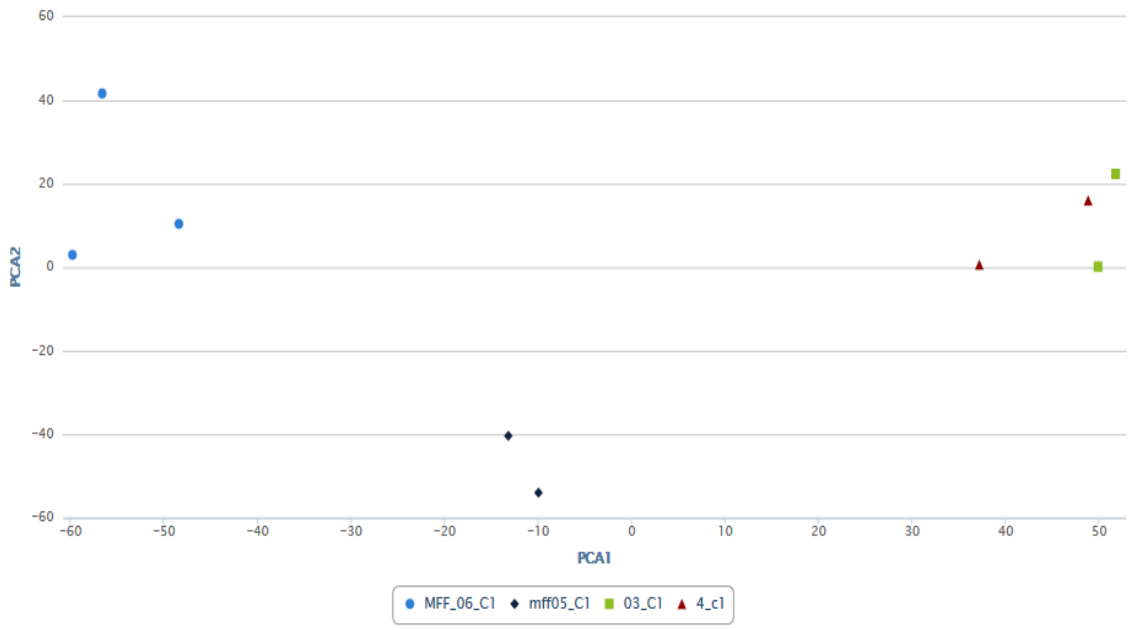


Figure A27. *XCMS online:* PCA analysis of the Reproducibility Assessment. Cycle 1 of the experimental repeats for the non-recursive mineral control (*green*), non-recursive reaction (*red*), recursive reaction (*black*) and recursive reaction in the precense of chalcopyrite (*blue*).

Compounds | Compounds per File | Features | mzCloud Results | ChemSpider Results

	Checked	Name	Predicted Formula	Molecular Weight	RT [min]	Area (Max.)	# ChemSpider Results	# mzCloud Results	mzCloud Best Match	Group Areas
16			C6 H9 N3 O	139.07364	14.514	1358076	4	0		1.36e6
17			C5 H14 O5	154.08432	16.683	1332174	0	0		1.33e6
18			C7 H19 N O5	197.12613	16.690	1215949	1	0		1.22e6
19			C7 H11 N3 O3	185.07903	11.525	1150692	2	0		1.15e6
20			C4 H4 N2 O	96.03193	7.707	1112816	0	0		1.11e6
21			C5 H3 N3 O3 S	184.98778	11.378	1102540	1	0		1.10e6
22				127.07363	16.646	1075919	0	0		1.08e6
23			C4 H5 N O	83.03666	12.427	950038	0	0		9.50e5
24			C5 H6 N2 O	110.04749	18.138	927739	4	0		9.28e5
25			C3 H6 N2 O	86.04766	11.365	863938	1	0		8.64e5
26			C3 H7 N9	169.08403	16.304	861269	6	0		8.61e5
27			C5 H8 N2 O2	128.05802	12.921	808405	6	0		8.08e5
28			C4 H7 N3	97.06342	17.407	749552	2	0		7.50e5
29		Hexamethylenetetramine	C6 H12 N4	140.10550	6.697	734100	1	1	85.0	7.34e5
30			C7 H7 N7 O4	253.05634	12.929	731448	2	0		7.31e5
31				99.03162	14.991	727500	3	0		7.27e5
32				57.05746	17.279	586555	2	0		5.87e5
33			C4 H4 N2 O2	112.02672	12.630	580286	3	0		5.80e5
34			C4 H4 N2	80.03697	16.900	559043	2	0		5.59e5
35			C3 H7 N3 O	101.05844	13.816	547623	0	0		5.48e5
36			C6 H11 N3 O3	173.07912	11.531	531440	1	0		5.31e5
37			C7 H11 N3 O	153.08939	13.650	519616	2	0		5.20e5
38			C6 H10 N4 O2	170.07945	12.975	512214	2	0		5.12e5
39			C10 H7 N2 O2 P	218.02469	5.848	491330	2	0		4.91e5
40			C7 H14 N4	154.12100	15.447	478648	2	0		4.79e5
41			C14 H40 N10 O9 P2 S2	618.18882	11.371	477768	0	0		4.78e5
42			C5 H10 N4 O2	158.07941	13.833	474681	0	0		4.75e5
43			C4 H6 N2	82.05256	5.553	470020	5	0		4.70e5

Table A3: Features detected by the CompoundDiscoverer® workflow. Predicted chemical formulas are presented in the second column and are used to generate the van Krevelen diagrams.

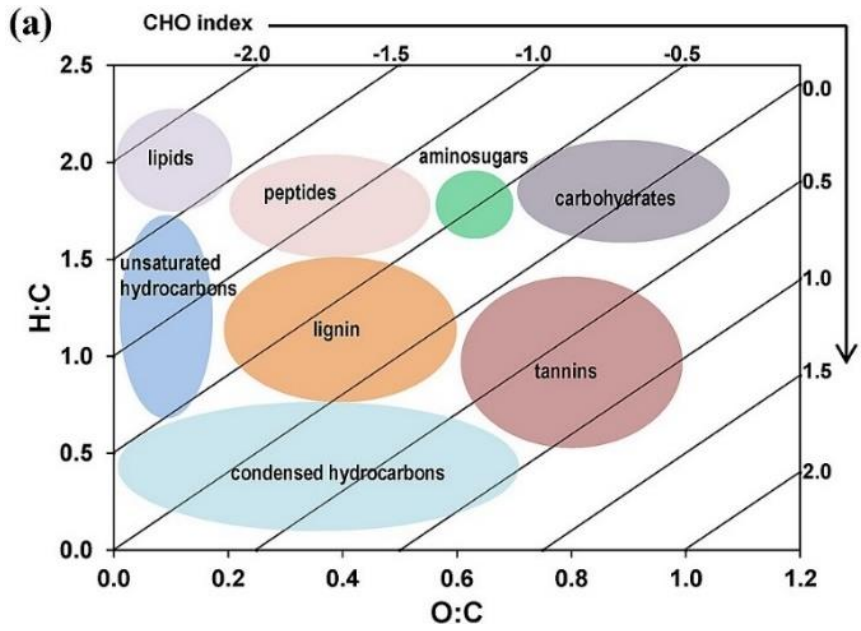


Figure A28. Example of a van Krevelen diagram distribution of organic compounds. Reproduced without permission from PLOS One, 2014.[217]

Figure A29. *XCMS online:* Generation and parameters for the PCA analysis of the recursive Miller-Urey reaction –*Cycle 1 to Cycle 5*- and the sample *blank*.

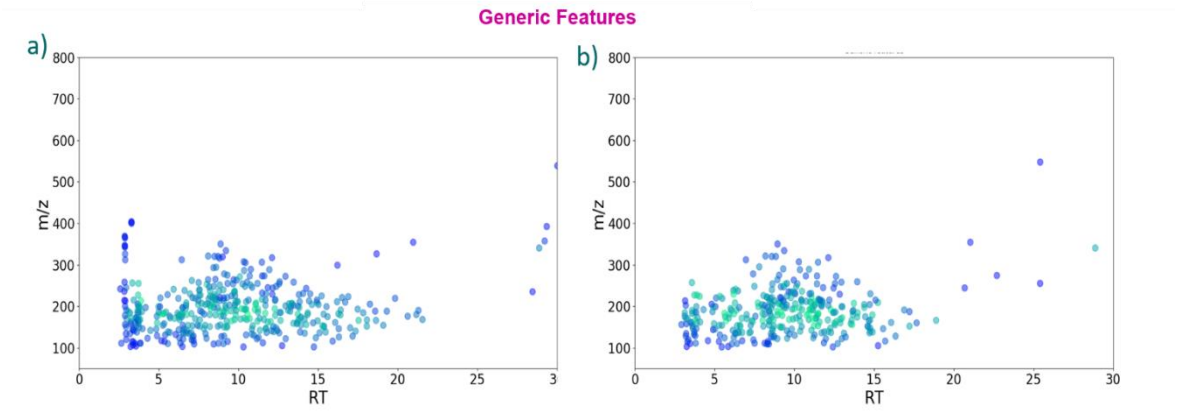


Figure A30. Generic features in the Recursive Miller-Urey samples and their distribution across the chromatographic profile. Repeats 2 and 3 are presented.

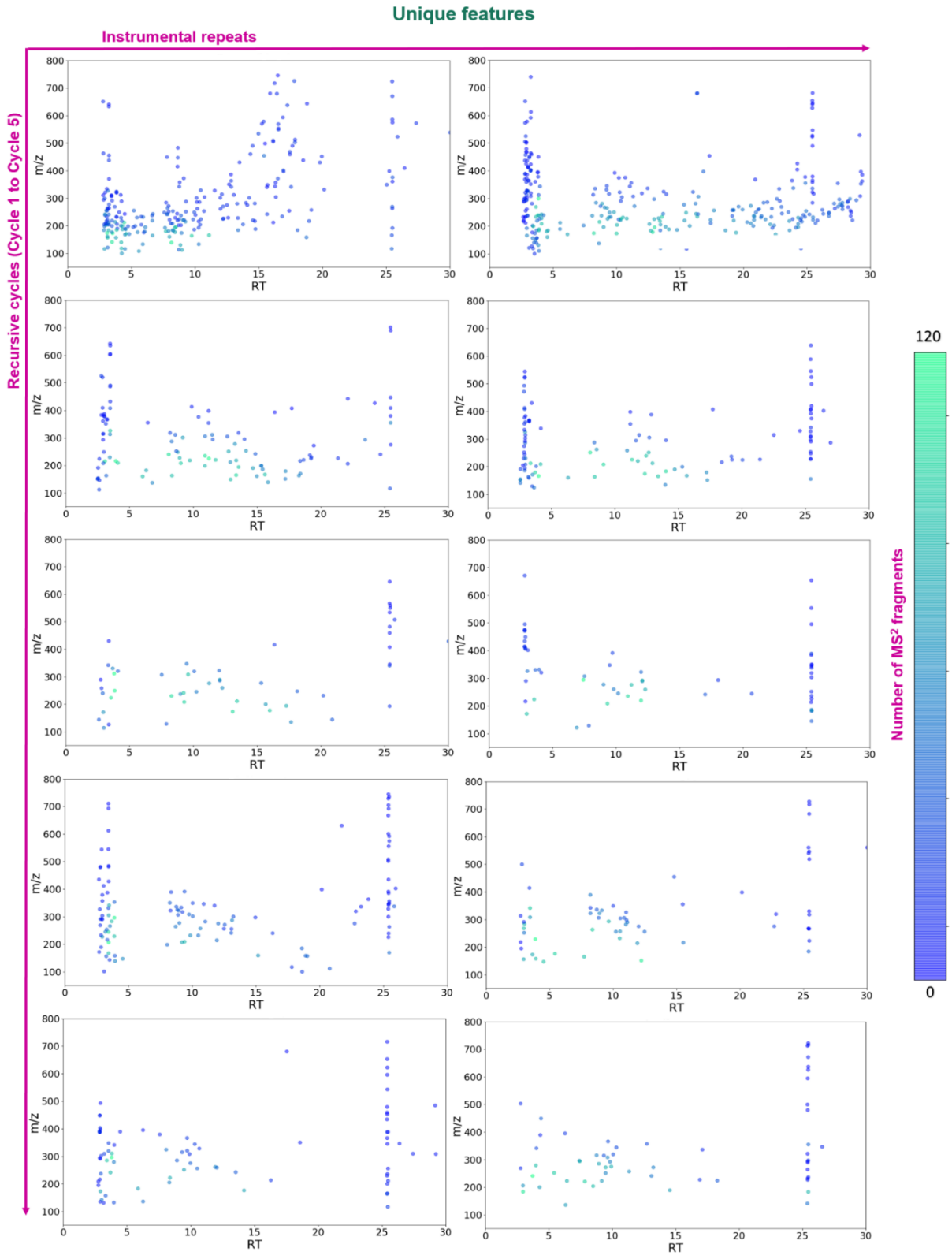


Figure A31. Unique features in the Recursive Miller-Urey samples for the instrumental repeats, *2 - 3*, and their distribution across the chromatographic profile. The number of MS^2 fragments for each of the features are presented as a heatmap.

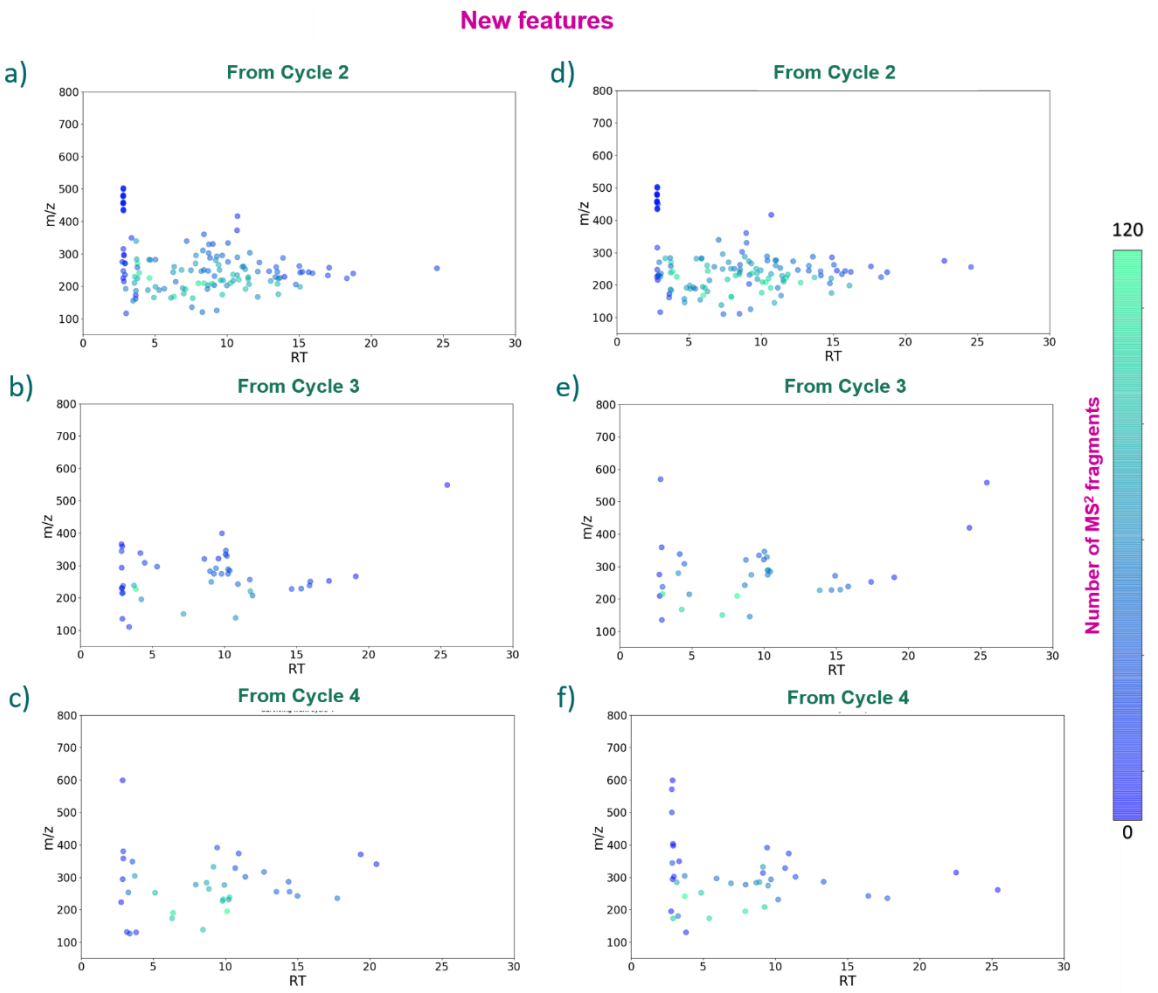


Figure A32. New features in the Recursive Miller-Urey samples for the instrumental repeats, *2 **(a-c)** – 3 **(d-f)***, and their distribution across the chromatographic profile. The number of MS^2 fragments for each of the features are presented as a heatmap.

Main Python Scripts:

Orbi_DDA_Data_Extractor_and_Complexity_Index_Calculator.py

```
"""
Script for calculating the M/Z index from MS2 data directly from mzml files
(Files must have been converted from RAW to MZML first: with MSConvert)
M/Z Index = Total MS2 peaks
"""

import os
import sys
import json
import time
import pymzml
import inspect
```

```
HERE =
os.path.dirname(os.path.abspath(inspect.getfile(inspect.currentframe())))
TEST = os.path.join(HERE, "test", "Cytidine.mzML")

LEVEL = 2


def extract_mz_index(mzml):

    msrun = pymzml.run.Reader(mzml, extraAccessions=[("MS:1000512",
["name"]), ("MS:1000504", ["name"])])
    msrun.ms_precisions[3] = 1e-04
    msrun.ms_precisions[4] = 1e-04

    full_data = {}
    for spectrum in msrun:
        complexity_data = {}
        if spectrum["ms level"] == LEVEL:
            # Filter all the MS2 spectra (Gets the parent MS1 peak)
            split = spectrum["filter string"].split("ms2 ")[1].split("
")[0]
            parent_peak = "%.4f" % (float(split.split("@hcd")[0]))
            if parent_peak not in complexity_data.keys():
                complexity_data["parent"] = parent_peak
            else:
                print('yeah')
            # Add this spectrum's data to the dictionary
            spectrum_data = []
            for mz, i in spectrum.peaks("raw"):
                mz_i = (float("%.2f" % (mz)), int(i))
                if mz_i not in spectrum_data:
                    spectrum_data.append(mz_i[0])
            # complexity_data["mz_intensity"] = spectrum_data
            complexity_data['rt'] = float(spectrum["scan start time"])
            # adding intensity
            complexity_data['intensity'] = float(spectrum["base peak
intensity"])


            # Sort the peaks and calculate complexity
            mz_peaks = sorted(spectrum_data)
            complexity = len(mz_peaks)
            complexity_data["complexity_value"] = complexity
            full_data[parent_peak] = complexity_data

    output_json_file(full_data, mzml)

def output_json_file(data, mzml_filename):
    """
6       Outputs the obtained data to JSON file
7
8       Args:
9           data (Dict): Data to be written
10          mzml_filename (str): Path of the mzML file to build the JSON
path
11      """
    filename = "{}.json".format(mzml_filename.split(".")[0])
    filename = os.path.join(HERE, filename)
    with open(filename, "w") as f:
        json.dump(data, f)


if __name__ == "__main__":
```

```
    folder =
'C:\\Users\\group\\Documents\\git_repositories\\LCMSprocessing\\Json\\new
_mzml\\reproducibility'

    for root, dirs, filenames in os.walk(folder):

        for file in filenames:
            if 'mzML' in file:
                print('rolling over {}'.format(file))
                extract_mz_index(folder + '\\' + file)
```

van_Krevelen_plot_from_CSV.py

```
from openpyxl import load_workbook
import pandas as pd
import itertools as it
from matplotlib import pyplot as plt

# file = 'Z:\\group\\Stephanie Colon\\Data\\Chemical
composition\\Complete csv set\\cycle 1\\A.xlsx'

experiments = ['Cycle 1.xlsx', 'Cycle 2.xlsx', 'Cycle 3.xlsx']

Chalcopyrite =
'C:\\Users\\group\\Documents\\git_repositories\\LCMSprocessing\\Goethite\
\'
# Control = 'Z:\\group\\Chemical composition\\Complete csv set\\cycle
2\\'
# cycle3 = 'Z:\\group\\Chemical composition\\Complete csv set\\cycle 3\\'
# # long_cycle = 'Z:\\group\\Chemical composition\\Complete csv set\\long
cycle\\'
# control = 'Z:\\group\\Chemical composition\\Complete csv
set\\control\\'

# folders = [cycle1,cycle2,cycle3,control]
folders = [Chalcopyrite]

def get_data(folder, file):
    wb = load_workbook(folder+file)
    ws = wb.active
    data = ws.values
    cols = next(data)[1:]
    data = list(data)
    idx = [r[0] for r in data]
    data = (it.islice(r, 1, None) for r in data)
    df = pd.DataFrame(data, index=idx, columns=cols)
    return df

def get_ratio(folder, file):
    # fig = plt.figure()
    data = get_data(folder, file)
    mols = [i for i in data['Predicted Formula'].tolist() if i != '']
    mol_dics = []

    for mol in mols:
        atoms_count = {'C':0, 'O':0, 'N':0, 'H':0, 'P':0}
        atoms = mol.split(' ')
        for a in atoms:
            alist = list(a)
            atom = alist.pop(0)
            try:
```

```
                atoms_count[atom] = int(''.join(alist))
            except ValueError:
                atoms_count[atom] = 1
        mol_dics.append(atoms_count)

    CH = []
    CO = []

    for a in mol_dics:
        CH.append(a['H']/a['C'])
        CO.append(a['O']/a['C'])
        CO_rat = a['O'] / a['C']
        # if CO_rat >2.5:
        #     print(a)
    # print(CO)
    return CH,CO

def print_all():
    for fold in folders:
        fig = plt.figure()
        # shapes = ["v", "^", "o", "*", "+", "x", "D", "s"]
        for exp in experiments:
            CH,CO = get_ratio(fold, exp)
            plt.scatter(CO, CH, s=20, label=exp)
            # plt.scatter(CO, CH, marker=shapes.pop(), s=20, label=exp)
        plt.legend()
        plt.title(fold.split('\\')[-2])
        plt.xlabel('O/C')
        plt.ylabel('H/C')
        plt.ylim(0, 3.05)
        plt.xlim(0.02, 1.8)
    plt.show()

# plt.legend(experiments)
```

Plot_heatmap_fromcsv.py

```
1    "Generates a heatmap from MS data / features from a CSV data file -
Steph 2019"
from matplotlib import pyplot as plt
import pandas as pd

# Folder to output image
folder =
'C:\\Users\\group\\Documents\\git_repositories\\LCMSprocessing\\Json\\New
_CSV\\long_cycle\\'

#Defines values to be plotted (note that mz_index can be changed to
intensity)
def get_data(name, fold):
    global best_index
    exp = pd.read_csv(fold + name+'.csv')
    exp_sorted = exp.sort_values('rt')
    # print(exp_sorted)

    rt = exp['rt'].tolist()
    mz = exp['Parent'].tolist()
    mz_index = exp['complexity_value'].tolist()

    return rt, mz, mz_index
```

```
#CSV data to be plotted (converted from JSON (e.g.
Convert_JSON_into_CSV.py)/ already extracted (or filtered) from mzml)
rt, mz_long, mz_index = get_data('SCS_machine_8G', folder)


mz = [round(i,3) for i in mz_long]
points = []
filtered = []


for i in range(len(rt)):
    p_rt = rt[i]
    p_mz = mz[i]
    p_mz_index = mz_index[i]
    points.append([p_rt,p_mz,p_mz_index])

df = pd.DataFrame(points, columns=['rt', 'mz', 'mz_index'])

#I think this bit selects a value from potential duplicates?
for i in mz:
    filt = df.loc[df['mz'] == i].values.tolist()
    best = [0,0,0]
    for point in filt:
        if point[2] > best[2]:
            best = point
    # print(best_point)
    if best not in filtered:
        filtered.append(best)

filt_df = pd.DataFrame(filtered, columns=['rt', 'mz', 'mz_index'])

#Calls out values to be plotted
rt_filt = filt_df['rt'].tolist()
mz_filt = filt_df['mz'].tolist()
mz_index_filt = filt_df['mz_index'].tolist()

#Sanity check - are we losing too many values by removing duplicates?
print (len(df))
print (len(filt_df))

# print(df)
# get_data('MU_t0_mu', folder)   = is this to plot single files

#Making the plot lee-readable and pretty
plt.scatter(rt_filt,mz_filt,c=mz_index_filt,
cmap=plt.get_cmap('viridis_r'), vmin=0, vmax=175)
plt.ylim([0, 400])
plt.xlim([5, 20])
plt.yticks(fontsize = 25)
plt.xticks(fontsize = 25)
plt.xlabel('RT', fontsize = 27)
plt.ylabel('m/z', fontsize = 27)
# plt.colorbar()

plt.show()
```

Orbi_DDA_feature_filtering_v2.py

```
1    "Filters out blanks / Removes duplicates / Gets REAL features- Steph
2019"
import json
import filetools
from matplotlib import pyplot as plt
import pickle
import os
import sys
import numpy as np

#All data files to process within an instrument run (JSON)
folder =
"C:\\Users\\group\\Documents\\git_repositories\\LCMSprocessing\\filtering
\\compare_to\\MU_A\\"
folder2 =
"C:\\Users\\group\\Documents\\git_repositories\\LCMSprocessing\\filtering
\\compare_to\\MU_B\\"
folder3 =
"C:\\Users\\group\\Documents\\git_repositories\\LCMSprocessing\\filtering
\\compare_to\\MU_C\\"
#Blanks in the instrument run (JSON)
folder_b =
"C:\\Users\\group\\Documents\\git_repositories\\LCMSprocessing\\filtering
\\blank_to\\"

folders = [folder, folder2, folder3]

#Folder to put the list of all features found across the samples
drop = folder + "filtered_parents.json"
drop_unique = folder + "parents_info_unique.json"
drop_generic = folder + "parents_info_generic.json"

output_names =
['filtered_parents.json','parents_info_unique.json','parents_info_generic
.json']

#Read parent JSON
def read_json(filename):
    with open(filename) as f:
        return json.load(f)

#Write an updated Parent JSON file
def write_json(mario, filename):
    with open(filename, "w") as f:
        json.dump(mario, f)


def make_parent_list_n_filesdick(folder):
    #Create an all parents list = combine all parents from multiple JSON
    parent_list = [] #this is a list of all parent peaks together
    files = {}

    for filename in filetools.list_files(folder):
        if filename not in output_names: #if you run this shit already
and the output is in the folder, ignore it!
            data = read_json(filename)
            filtered_data = {}
            for key, value in data.items():
                round_peak = float("%.2f" % float(key))
                if round_peak not in filtered_data.keys():
                    filtered_data[round_peak] = value
```

```
                elif filtered_data[round_peak]['intensity'] <
value['intensity']:
                    filtered_data[round_peak] = value


            files[filename.split('\\')[-1]] = filtered_data
            for par in data.keys():
                # print(par)
                parent_list.append(float("%.2f" % float(par)))

    #Sanity check -list must not be empty
    print('parent list:' + str(len(parent_list)))
    return parent_list, files
# parent_list, files = make_parent_list_n_filesdick(folder)
# sys.exit()

def filter_blank(parent_list):
    # parent_list, files = make_parent_list_n_filesdick(folder)
    blank_list = []
    for filename in filetools.list_files(folder_b):
        data_b = read_json(filename)
        for par_b in data_b.keys():
          blank_list.append(float("%.2f" % float(par_b)))
    #2nd Sanity check - this list must be smaller than previous
    print('blank list:' + str(len(blank_list)))
    # filters out the background (substracts blank_list from parent_list)
    parent_elite = []

    for true_parent in parent_list:
        counter = False
        for check_b in blank_list:
            if true_parent == check_b:
                counter = True
        if not counter:
            parent_elite.append(true_parent)

    # Sanity check 3 - Should be smaller than all parents (substraction
of blanks
    print('parent_elite:' + str(len(parent_elite)))

    # Removes for duplicate features
    parent_blankremoved = set(parent_elite)

    # Sanity check - should be less than previous list (parent_elite)
    print('parent_ultra_elite:' + str(len(parent_blankremoved)))

    return parent_blankremoved

def make_filtered_dictionary(parent_list, files):
    # parent_list, files = make_parent_list_n_filesdick(folder)
    parent_blankremoved = filter_blank(parent_list)

    counted_parents = {}

    for parent in parent_blankremoved:
        counted_parents[parent] = {}
        for name, file in files.items():
            if parent in file.keys():
                info = file[parent]
                counted_parents[parent][name] = info

    # for key, value in counted_parents.items():
    #     print(key, value)
    return counted_parents
```

```
#For testing if the data is loaded correctly:
# print(files)
# file = files['RMU2_low_1b.json']

def plot_individual(folder):
    parent_list, files = make_parent_list_n_filesdick(folder)

    for title, file in files.items():
        parents = file.keys()
        parents_filtered = filter_blank(parents)

        intensity = []
        rt = []
        complexity_value = []
        mz = []

        for par in parents_filtered:
            intensity.append(file[par]['intensity'])
            rt.append(file[par]['rt'])
            complexity_value.append(file[par]['complexity_value'])
            mz.append(file[par]['parent'])

        max_int = max(complexity_value)
        int_norm = [((i/max_int)+0.1)*100 for i in complexity_value]
        # plt.scatter(rt, [float(i) for i in mz], c=int_norm,
cmap=plt.get_cmap('winter'), alpha=0.5)
        plt.scatter(rt, [float(i) for i in mz], s=100, c=int_norm,
cmap=plt.get_cmap('winter'), alpha=0.5)
        plt.title(title)
        plt.ylim([50, 800])
        plt.xlim([0, 30])
        plt.yticks(fontsize=25)
        plt.xticks(fontsize=25)
        plt.xlabel('RT', fontsize=28)
        plt.ylabel('m/z', fontsize=28)
        # plt.colorbar(orientation="horizontal")

        plt.show()

# plot_individual(folder) #if you want to plot for a different folder
change this one!!

def plot_generic_features(folder):
    parent_list, files = make_parent_list_n_filesdick(folder)
    dick = make_filtered_dictionary(parent_list, files)
    generic_features = {}

    for feature, info in dick.items():
        if len(info) == len(files):
            generic_features[feature] = info

    intensity = []
    rt = []
    complexity_value = []
    mz = []

    for peak, info in generic_features.items():
        v_intensity = []
        v_rt = []
        v_complexity_value = []

        for version in info.values():
            v_intensity.append(version['intensity'])
            v_rt.append(version['rt'])
            v_complexity_value.append(version['complexity_value'])
```

```
            intensity.append(np.mean(v_intensity))
            rt.append(np.mean(v_rt))
            complexity_value.append(np.mean(v_complexity_value))
            mz.append(peak)
        plt.title('Generic features')
        plt.scatter(rt, [float(i) for i in mz], s=90, c=complexity_value,
cmap=plt.get_cmap('winter'), alpha=0.5)
        plt.yticks(fontsize=20)
        plt.xticks(fontsize=20)
        plt.ylim([50, 800])
        plt.xlim([0, 30])
        plt.xlabel('RT', fontsize=25)
        plt.ylabel('m/z', fontsize=25)
        plt.show()

# plot_generic_features(folder) #if you want to plot for a different
folder change this one!!

def plot_level_of_persistence():

    lops = []

    for folder in folders:

        parent_list, files = make_parent_list_n_filesdick(folder)

        dick = make_filtered_dictionary(parent_list, files)

        levels_of_persistence = [0 for i in range(len(files))]

        for feature, info in dick.items():
            levels_of_persistence[len(info)-1] += 1
        lops.append(levels_of_persistence)

    repetitions = []

    for cycle in range(len(lops[0])):
        reps = []
        for rep in range(len(lops)):
            reps.append(lops[rep][cycle])
        repetitions.append(reps)
    means = [np.mean(i) for i in repetitions]
    errors = [np.std(i) for i in repetitions]

    print(repetitions)
    print(means)
    print(errors)

    plt.bar(range(len(means)), means, yerr = errors)
    plt.xticks(range(len(means)), ['1 cycle','2 cycles','3 cycles','4
cycles','all cycles'])
    plt.gca().invert_xaxis()
    plt.ylabel('Number of features', fontsize = 15)
    plt.title('Level of persistence', fontsize = 18)
    plt.show()

# plot_level_of_persistence()

def plot_unique_features_barplot():

    laps = []

    for folder in folders:
```

```
        parent_list, files = make_parent_list_n_filesdick(folder)

        dic = make_filtered_dictionary(parent_list, files)

        unique_features = {}
        for feature, info in dic.items():
            if len(info) == 1:
                unique_features[feature] = info

        in_which_unique = [0 for i in range(len(files))]
        sort_filenames = sorted(files.keys())


        for feature, info in unique_features.items():
            for file in info.keys():
                in_which_unique[sort_filenames.index(file)] += 1
        laps.append(in_which_unique)

    repetitions = []

    for cycle in range(len(laps[0])):
        reps = []
        for rep in range(len(laps)):
            reps.append(laps[rep][cycle])
        repetitions.append(reps)
    means = [np.mean(i) for i in repetitions]
    errors = [np.std(i) for i in repetitions]

    print(repetitions)
    print(means)
    print(errors)

    plt.bar(range(len(means)), means, yerr=errors)
    plt.xticks(range(len(means)), ['Cycle 1', 'Cycle 2', 'Cycle 3',
'Cycle 4', 'Cycle 5'])
    # plt.gca().invert_xaxis()
    plt.ylabel('Number of features', fontsize = 15)
    plt.title('Unique features', fontsize = 18)
    plt.show()

    parent_list, files = make_parent_list_n_filesdick(folder)
    plt.figure()
    dick = make_filtered_dictionary(parent_list, files)

    unique_features = {}
    for feature, info in dick.items():
        if len(info) == 1:
            unique_features[feature] = info

    in_which_unique = [0 for i in range(len(files))]
    sort_filenames = sorted(files.keys())


    for feature, info in unique_features.items():
        for file in info.keys():
            in_which_unique[sort_filenames.index(file)] += 1

    plt.bar(range(len(in_which_unique)), in_which_unique)
    plt.xticks(range(len(sort_filenames)), sort_filenames)
    plt.ylabel('number of unique features')
    plt.title('unique features')
    plt.show()

# plot_unique_features_barplot()  #if you want to plot for a different
folder change this one!!
```

```
def unique_plot(folder):
    parent_list, files = make_parent_list_n_filesdick(folder)
    dic = make_filtered_dictionary(parent_list, files)
    unique_features = {}

    for feature, info in dic.items():
        if len(info) == 1:
            unique_features[feature] = info

    sort_filenames = sorted(files.keys())

    for cycle in sort_filenames:
        plt.figure()

        intensity = []
        rt = []
        complexity_value = []
        mz = []

        for peak, where in unique_features.items():
            for file in where.keys(): #this is retarded, no real loop
                if file == cycle:
                    info = where[file]
                    intensity.append(info['intensity'])
                    rt.append(info['rt'])
                    complexity_value.append(info['complexity_value'])
                    mz.append(peak)
        plt.title('unique in {}'.format(cycle))
        plt.scatter(rt, [float(i) for i in mz], s=50, c=complexity_value,
cmap=plt.get_cmap('winter'), alpha=0.5)
        plt.yticks(fontsize=20)
        plt.xticks(fontsize=20)
        plt.ylim([50, 800])
        plt.xlim([0, 30])
        plt.xlabel('RT', fontsize=25)
        plt.ylabel('m/z', fontsize=25)
    plt.show()

# unique_plot(folder3)  #if you want to plot for a different folder
change this one!!

def extents(f):
  delta = f[1] - f[0]
  return [f[0] - delta/2, f[-1] + delta/2]

def plot_middle(folder):
    parent_list, files = make_parent_list_n_filesdick(folder)
    for ind in [2,3,4]:
    # for ind in [3]:
    # for ind in [2]:

        dic = make_filtered_dictionary(parent_list, files)
        filtered_features = {}
        for feature, info in dic.items():
            if len(info) == ind:
                filtered_features[feature] = info

        full_matrix = []

        sort_filenames = sorted(files.keys())

        peaks = [i for i in filtered_features.keys()]
        sort_peaks = sorted(peaks)
        for peak in sort_peaks:
```

```
            peak_vector = [0 for i in range(len
                                                (files))]
            where = filtered_features[peak]
            for file, info in where.items():
                peak_vector[sort_filenames.index(file)] = 1
            full_matrix.append(peak_vector)
        x = range(5)
        y = range(len(full_matrix))
        plt.figure()
        plt.xticks(range(len(sort_filenames)), sort_filenames)
        # plt.matshow(full_matrix)
        plt.imshow(full_matrix, cmap='cool', aspect='auto',
interpolation='none', extent=extents(x) + extents(y), origin='lower')
    plt.show()

# plot_middle(folder2)  #if you want to plot for a different folder
change this one

def remove_cycles(dic, to_be_removed): #to_be_removed is a list of which
ones you want to remove, example [1,2] to plot the graph of surviving
from 3
    dic_nocycle1 = {}
    print('lenght of the long dic: {}'.format(len(dic)))

    for peak in dic.keys():
        flag = False
        for cycle in dic[peak].keys():
            if int(cycle.split('_')[-1][0]) in to_be_removed:
                flag = True
                break
        if not flag:
            dic_nocycle1[peak] = dic[peak]

    print('lenght of the dic after removing cycles:
{}'.format(len(dic_nocycle1)))
    return dic_nocycle1

def plot_surviving_from_X(folder, from_what): #folder is which folder are
you plotting, from what is an integer of from which plot you plotting
    parent_list, files = make_parent_list_n_filesdick(folder)
    long_dic = make_filtered_dictionary(parent_list, files)  #this is
with all cycles

    cycles_to_remove = list(range(from_what))

    dic = remove_cycles(long_dic, cycles_to_remove)
    surviving_from_X = {}

    for feature, info in dic.items():
        if len(info) == len(files)-from_what+1:
            surviving_from_X[feature] = info
    print('lenght of peaks surviving from {}: {}'.format(from_what,
len(surviving_from_X)))
    intensity = []
    rt = []
    complexity_value = []
    mz = []

    for peak, info in surviving_from_X.items(): #this part is because
there are multiple values for duplicates
        v_intensity = []
        v_rt = []
        v_complexity_value = []
```

```
        for version in info.values():
            v_intensity.append(version['intensity'])
            v_rt.append(version['rt'])
            v_complexity_value.append(version['complexity_value'])

        intensity.append(np.mean(v_intensity))
        rt.append(np.mean(v_rt))
        complexity_value.append(np.mean(v_complexity_value))
        mz.append(peak)

    plt.title('Surviving from cycle {}'.format(from_what))
    plt.scatter(rt, [float(i) for i in mz], s=90, c=complexity_value,
cmap=plt.get_cmap('winter'), alpha=0.5)
    plt.yticks(fontsize=20)
    plt.xticks(fontsize=20)
    plt.ylim([50, 800])
    plt.xlim([0, 30])
    plt.xlabel('RT', fontsize=25)
    plt.ylabel('m/z', fontsize=25)
    plt.show()

    return len(surviving_from_X)
#
plot_surviving_from_X(folder3, 4)

repetitions = []

for cycle in [2,3,4]:
    reps = []
    for folder in folders:
        surviving = plot_surviving_from_X(folder, cycle)
        reps.append(surviving)
    repetitions.append(reps)
means = [np.mean(i) for i in repetitions]
errors = [np.std(i) for i in repetitions]
plt.bar(range(len(means)), means, yerr=errors, color = 'c')
plt.xticks(range(len(means)), ['Cycle 2', 'Cycle 3', 'Cycle 4'])
plt.ylabel('Number of features', fontsize=25)
plt.yticks(fontsize=20)
plt.xticks(fontsize=20)
plt.title('New features', fontsize=25)
plt.show()
```

PCA.py

```
1    import pandas as pd
import numpy as np
import random as rd

from sklearn.decomposition import PCA
from sklearn import preprocessing
import matplotlib.pyplot as plt
import sys
import json
import os

def get_data(name, fold):
    global best_index
    exp = pd.read_csv(fold + name)
    exp_sorted = exp.sort_values('rt')
    # print(exp_sorted)

    rt = exp['rt'].tolist()
    mz = exp['Parent'].tolist()
    mz_index = exp['complexity_value'].tolist()

    return rt, mz, mz_index
#
# folder =
'C:\\Users\\group\\Documents\\git_repositories\\LCMSprocessing\\Json\\New
JSON\\'
folder =
'C:\\Users\\group\\Documents\\git_repositories\\LCMSprocessing\\Json\\MZM
L_test\\'
file = 'SColon_180417_7.csv'
# file = 'MFF_05_1_C1.csv'

def get_vector(file, folder):
    rt, mz, _ = get_data(file, folder)
    bins = np.histogram2d(mz, rt, bins=[500,20], range=[[0,500],[0,25]])
    vector = np.matrix(bins[0]).flatten().tolist()
    return vector

files = os.listdir(folder)
# files.remove('rmu')
print(files)
all_exp = []

for file in files:
    tag = file.split('_')[-1]
    # if '7' in tag:
    vector = get_vector(file, folder)
    all_exp.append(vector[0])
    print('done {}'.format(file))
    print(len(vector[0]))

# print(len(all_exp[0]))

# print(all_exp)

scaler=preprocessing.StandardScaler()
scaler.fit(all_exp)
X_scaled=scaler.transform(all_exp)
print('scaled')
pca = PCA(n_components=2)
pca.fit(X_scaled)
output = pca.transform(X_scaled)
```

```
markers =  ['o', 'v', '^', '<', '>', '8', 's', 'p', '*', 'h', 'H', 'D',
'd', 'P', 'X']

# mark_dic = {'3': 'o', '5':'p', '7': '^', '8':'s', 'R': '*'}
# color_dic = {'A':'blue', 'B':'magenta', 'C': 'orange', 'D':'grey',
'E':'pink', 'F':'purple', 'G':'green', 'none':'black'}

# mark_dic = {'3': 'o', '2':'p', '1': '^'}
# color_dic = {'A':'blue', 'B':'magenta', 'C': 'orange', 'D':'grey',
'E':'pink', 'F':'purple', 'G':'green', 'none':'black'}

# mark_dic = {'1': 'o', '2':'p', '3': '^', '4':'s', '5': '*'}
# color_dic = {'a':'pink', 'b':'purple', 'c':'green'}

# mark_dic = {'7': 'o'}
# color_dic = {'A':'blue', 'B':'red', 'C': 'orange', 'D':'grey',
'E':'pink', 'F':'purple', 'G':'green', 'none':'black'}

# color_dic = {'A': 'purple', 'B': 'pink'}
# mark_dic = {'1':'o', '2':'*', '3': '^', '4':'s', '5':'p'}

mark_dic = {'N': 'o', 'R': '*', 'M': 's', 'P': '^'}
color_dic = {'1':'blue', '2':'purple', '3': 'orange', '4':'green',
'5':'pink'}


#
# # print(output)
# for i in range(len(output)):
#     name = files[i]
#     tag = name.split('_')[-1]
#     cycle = tag[0]
#     if len(tag) == 5:
#         mineral = 'none'
#     else:
#         mineral = tag[1]
#     plt.scatter(output[i][0], output[i][1], marker=mark_dic[cycle],
c=color_dic[mineral], label = files[i])
# # plt.legend()
# plt.tick_params(labelsize=10)
# # plt.scatter([i[0] for i in output], [i[1] for i in output], label= )
# plt.show()


# print(output)
for i in range(len(output)):
    name = files[i]
    tag = name.split('_')[-1]
    cycle = tag[0]
    if len(tag) == 5:
        mineral = 'none'
    else:
        mineral = tag[1]
    plt.scatter(output[i][0], output[i][1], marker=mark_dic[cycle], s=
60,  c=color_dic[mineral], label = files[i])
# plt.legend()
plt.tick_params(labelsize=18)
# plt.scatter([i[0] for i in output], [i[1] for i in output], label= )
plt.show()
```