From Recursion to Functionality: Exploring the Origins of Life-Like Chemistry in Oligomer Systems

by

Pavani Ganju

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science (Botany)

at the UNIVERSITY OF WISCONSIN-MADISON

2024

Date of final oral examination: 07/17/2024

The thesis is approved by the following members of the Final Oral Committee:
David Baum, Professor, Botany
John Yin, Professor, Chemical and Biological Engineering
Marisa Otegui, Professor, Botany
Simon Gilroy, Professor, Botany
# Table of Contents

ACKNOWLEDGEMENTS .......................................................................................................................... ii

ABSTRACT .................................................................................................................................................. iii

Chapter 1: The Protometabolic Nature of Prebiotic Polymer Systems...................................................... 1

Chapter 2: A Recursive Framework for the Detection of Autocatalytic Motifs in Mineral-Associated Oligomer Systems .................................................................................................................. 10

Chapter 3: Long-Term Signals Suggest Compositional Dynamics May be Driven by Autocatalytic Feedback.................................................................................................................................. 17

REFERENCES ............................................................................................................................................. 37

SUPPLEMENTAL DATA .............................................................................................................................. 43
Acknowledgements

I would like to thank all the people who have contributed in some way to the work described in this thesis. I would like to express my deepest gratitude to my academic advisor, Professor David Baum, for giving me the opportunity to work on this research project, as well as giving me his invaluable time, patience and persistent support. His unwavering enthusiasm for this research, as well as his contributions and feedback were not only critical for the completion of this manuscript, but have also taught me, both directly and indirectly, how good science is done. I would also like to sincerely thank Professor John Yin for providing me with additional mentorship and expertise on this project and engaging me in many new ideas for research projects and side-experiments. Additionally, I would like to thank my committee members Professor Marisa Otegui and Professor Simon Gilroy for their interest in my work and their invaluable feedback on my thesis.

The members of the Baum lab have contributed immensely to my journey as a graduate student, both personally and professionally. I am grateful to Tymofii Sokolskyi for his keen enthusiasm for the research topic, willingness to help me conduct many LC-MS analyses and help in troubleshooting issues with my project. I would like to also thank Praful Gagrani and Camille Evin for their friendship and insights into my research topic, as well as Sydney Gargulak and Yichen Zhuang for helping me with many daily transfer experiments.

Lastly, I would like to thank my family, friends and pets for all their love and encouragement. This thesis would not have been possible without their constant support during my time as a graduate student, and their advice during challenging times, both personal and professional.
Abstract

Understanding how the first functional polymers of life emerged has drawn considerable focus in origins of life studies. Theoretical models have demonstrated the ability of peptides to interact in dynamic ways, widely relying on chemical autocatalysis and selection for self-propagation. Since it is highly unlikely that complex replicators emerged spontaneously on a prebiotic Earth, it is important to investigate how catalytic peptides might have been generated via non-enzymatic peptide synthesis, and whether further proliferation and an emergence in functionality could be possible in plausible prebiotic settings.

One approach to studying the emergence of such autocatalytic regimes in the laboratory involves multiple recursive dehydration cycles, which promotes amino acid condensation while imposing a selective pressure on the system through the continuous dilution of products and partial replenishment of fresh feedstock. This work tests the potential of this recursive framework to enrich for effective catalytic sequences in an equimolar mixture of ten prebiotically plausible amino acids and montmorillonite, a clay mineral that has been extensively studied for its ability to catalyze, aggregate and stabilize peptide motifs in solution.

We observe differences in relative amino acid abundances over analyzed generations that likely reflects changes in montmorillonite-associated absorbance properties over time, either through constraints posed by dehydration on system behavior, or possibly through competition with species that build over time. An analysis of global system behavior reveals a potential role for autocatalytic feedback in driving some of these trends, through the emergence of undetectable catalytic sets in the compositional space. Additional experiments will be required to robustly investigate this experimental system, as well as corroborate some of our findings.
Chapter 1: The Protometabolic Nature of Prebiotic Polymer Systems

Understanding the sequence of events that led to the emergence of life remains largely unknown. Nonetheless, the ubiquitous presence of peptides and nucleic acids in modern biochemistry has made their involvement in prebiotic processes indisputable. Numerous studies have identified many prebiotically plausible pathways for forming life’s essential monomers, especially amino acids (Kitadai and Maruyama; 2018). However, there still remains a significant leap from the monomer chemistry dominating a Hadean Earth to the complex biological machinery governing biopolymer synthesis in living organisms today.

Living systems depend on a particular biochemical organization where a network of molecules and chemical reactions are intricately linked to each other in a functionally closed and self-sustaining system. It is often assumed that the ability of such a system to evolve is synonymous with the origins of the first genetically encoded polymer systems, most likely RNA. Theory and experiments have shown that such polymer systems can be engineered to exhibit some forms of self-replicating behavior that resembling cellular life (von Kiedrowski, 1986; Doudna et. al, 1991; Li & Nicolaou, 1994; Lee et. al, 1996; Kim et. al, 2015). Moreover, experiments such as those by Haruna and Spiegelman (1965), Joyce (1987; 1991; 2002), Tuerk & Gold (1990) and Szostak (2001) became instrumental in popularizing the notion of evolution being tied to template-directed replication.

The ”RNA world” hypothesis posits that there was a period of time on hadean Earth when RNA, or a chemically similar polymer, carried out most of the information processing that life needed to emerge (Higgs & Lehman, 2015). The great appeal of this hypothesis, however, has not yet elucidated how in the absence of the enzymatic activity of ribozymes or synthetic laboratory processes, RNA could have been formed with sufficient regularity and rate to support the metabolic
activity and information processing integral to the emergence of life. Of course, it is possible that an ancestral polymer that emerged prior to RNA could have provided the necessary transitionary steps for such a process to happen, but that would yet again leave us with a formidable problem of trying to characterize the identity of such a polymer and trace our steps back to how it led to the emergence of RNA. After all, complexification is considered to be one of the essential features of functional biochemical systems, and complexification via the capacity to evolve adaptively is what truly distinguishes the living from the non-living.

I. Towards Prebiotic Peptide Synthesis on Mineral Surfaces

As an alternative to an RNA-dependent model, it is worth considering the possibility that, evolution was initiated by the oligomerization of amino acids into functionally active peptides. It is likely, given the results of Miller-Urey experiments and other analyses of meteorite surfaces, that amino acids existed in the prebiotic world (Miller, 1953; Bernstein et. al, 2002). However, a peptide-first model needs to confront the problem that the process of amino acid condensation, is thermodynamically and kinetically unfavorable in an aqueous environment (Martin, 1998). Some studies have investigated ways to induce condensation through salt-induced dehydration in the presence of co-precipitated iron/nickel sulfides and carbon monoxide (Schwendinger & Rode, 1992) or by using condensation agents such as cyanamide (Nooner et. al, 1977), urea (Fischer & Guichard, 2010), dicyanamide (Steinman et. al, 1966) or diaminomaleonitrile (Bizzarri et. al, 2022). However, many of these agents have varying degrees of plausibility in a prebiotic scenario (Frenkel-Pinter, 2020).
As a prebiotically plausible alternative, several studies have demonstrated that an effective way to achieve successful, and continued condensation of amino acids into higher order oligomers is through repeated processes of drying and subsequent re-wetting at high temperatures (Griffith and Vaida, 2012; Yu et. al, 2016). Such processes would have likely occurred in depositional environments, with fluctuating water activity occurring due to rainfall, tidal processes, or hydrothermal processes (Pirajno, 2020).

Repeated alternation of a water-rich period followed by a dry period, could force the system out of equilibrium and allow for polymerization to occur [Figure 1]. However, this model has some shortcomings. Although drying was plausible on an early Earth, it could likely be incompatible, or even inhibitory to other prebiotic reactions depending on the composition of the reaction mixture. Moreover, polymerization under dry conditions is partially reversible, with longer polymers being more susceptible to cleavage with every iterative wetting cycle, therefore

Figure 1: A standard model for wet-dry cycling involves iterative phases of cool temperatures and rain followed by warm sunny weather. The drying period drives condensation reactions, while the wet period supplies and distributes chemical feedstock from the environment to prepare reactions for the next cycle.
reducing reaching a threshold of polymer length after a certain number of cycles (Muchowska & Moran, 2020).

Figure 2: A comparison of the acid-base speciation of several amino acids to several oxide surfaces. Black denotes a positively charged surface, hatched is globally neutral and white is negatively charged. From Lambert, 2008.
As a complementary approach for assisting wet-dry mediated condensation, the adsorption of monomers onto mineral surfaces can provide a thermodynamically favorable environment for polymerization to occur. There is considerable evidence of the protecting, selecting, catalyzing and concentrating role of minerals in the origins of life (Cleaves et. al, 2012). Such an idea has a long history, tracing back to work by Bernal (1951), Goldschmidt (1952) and Wächtershäuser (1988) on the possible role minerals play in the emergence of life. Many minerals including silicates, zeolites, oxides, sulfides, calcite, gypsum, epsomite, halite, pyrite and fine-grained clays could have been present on an early Earth in evaporitic environments governed by frequent wet-dry cycles (Cleaves et. al, 2012). Of these, the possible role of clay minerals, which were abundant on the early Earth, has been well-studied (Jaber et. al, 2014). Clay minerals have large surface areas, which provide an effective and chemically selective means of peptide bond catalysis (Rode, 1999). However, the capacity of any mineral to promote peptide bond catalysis, and perhaps, even to modulate the set of sequences that interact with it is largely dependent on the surface chemistry of that mineral.

Some potential possibilities for different adsorption interactions arise from a combination of four major bond types; (1) non-specific, pH-dependent electrostatic bonds dictated by the overall charge of the amino acid and mineral, (2) specific covalent bonds, as have been reported for some silicate minerals, (3) specific but weak hydrogen bonds capable of forming adducts that are similar to “outer-sphere complexes” in colloidal solutions (Lambert, 2008), and (4) hydrophobic/hydrophilic interactions in layered clay minerals (Yu et. al, 2013).

To promote the condensation of adsorbed moieties, free energy of adsorption of a dimer must be more negative than twice the free energy of adsorption of two individual monomers (de Duve & Miller, 1991). This implies that even if dimerization is unfavorable in solution, it can
become more energetically favorable when adsorbed onto a surface. Furthermore, as a polymer chain grows, additional units can contribute increasingly to the stabilization of the adsorbed state, making interactions with the surface stronger. Beyond a certain length, the adsorption can become irreversible, making longer polymers less susceptible to cleavage by rehydration processes.

This argument, however, does not consider the dynamic state of water in prebiotic scenarios and assumes its activity to be almost constant. Lambert (2008) argues that even if polymerization becomes favored in a molecule’s absorbed state its kinetics would be slow. Instead, the kinetic boost provided by variable water activity, such as through dehydration approaches, would allow for polymerization reactions within adsorbed species to not only become favorable when coupled with wet-dry cycling, but also stabilize longer polymers that have been adsorbed on the mineral surface, or in the case of layered clays, within the mineral matrix.

Considering the thermodynamic and kinetic constraints of mineral adsorption, it is plausible that adsorbed chemical species could directly or indirectly modify the surface, thereby generating new adsorption phenomena and, consequently, new reaction chemistry. Such chemical variation could accumulate and persist when coupled with an iterative process such as wet-dry cycling, and systematically favor the emergence of non-random peptide sequences capable of interacting with the surface and solution chemistry in novel ways.

II. Autocatalytic Networks in Prebiotic Polymer Systems

Autocatalysis, the chemical equivalent of self-replication, has been defined in many different ways in origins-of-life studies, from Ostwald’s formalization of the phenomenon (1890; Peng et. al, 2022) to theoretical characterizations of its role in chemical reaction networks (Gagrani et. al, 2024). Despite some ambiguity as to its exact definition, it is generally agreed on that early
metabolic systems required some kind of feedback behavior that allowed them to make use of the food and energy present in their environment to make more of all system members.

Of particular relevance to peptide associated autocatalysis, Kauffman (1971; 1986; 1993) proposed that life started as collectively self-reproducing sets of catalytic polymers (proteins), where *autocatalytic sets* (ASs) - sets of chemical reactions that are all catalyzed by at least one molecule from the set – arise spontaneously. Inspired by Kauffman’s model, a rich body of theory has been developed to describe and analyze such Reflexively Autocatalytic, Food-generated (RAF) sets (e.g., Hordijk & Steel 2004, 2012; Mossel and Steel 2005; Hordijk et al. 2011, 2017). RAF sets are defined as sets of chemical species where the synthesis of each species from food is catalyzed by chemicals in the set (Hordijk et. al, 2010). In its essence, this theory suggests that the production of RAF networks is inevitable given a large enough chemical space and a moderate level of catalysis (Hordijk & Steel, 2012). It is difficult to determine the number of species and the frequency of catalysis needed for RAF sets to emerge, although theory suggests that this
probability will go up as the number of chemical reactions increases and as the probability that any species has the ability to catalyze reactions increases.

In the case of autcatalytic “cores” involving one or more autocatalytic species, Vasas et al (2012) demonstrated that multiple viable and evolvable cores can emerge from RAF sets [Figure 3]. Since cores remain inactive until triggered by the presence of one of their members, they can act as a sort of information system; mutations caused by rare reactions can seed new cores over time, resulting in changing system-level behavior, which could become a target of selection.

Without committing to a particular kind of chemistry, the evolvability of autocatalytic networks can also be approached as a problem of chemical ecology. Hunding et. al (2006) hypothesized initially that life started as a prebiotic ecology of interacting chemical systems. Over the years, this theory has been further refined to explain the origins of evolution in the absence of prior adaptive processes (Baum, 2015; 2018; Baum et al. 2023). The emerging view is that chemical ecosystems composed of interacting autocatalytic cycles are ubiquitous and provide a basis for continual change that can resemble ecological succession (Peng et. al, 2019) and gradually acquire the characteristics needed for conventional Darwinian evolution (Baum et. al, 2023).

III. Conclusions

Four main variables govern the efficiency of polymerization chemistry in the context of the origins of life, and consequently this thesis; (1) the number of monomer types and, thus the number of potential sequences at a given length; (2) the length of the longest polymers; (3) the fraction of polymers that display catalytic activity; and (4) the nature and efficiency of that catalysis. The last two are perhaps the most important to the emergence of autocatalytic sets in a
chemical system. At the same time, longer peptides have also been associated with higher rates of catalytic behavior than smaller oligopeptides (Carvalho et al., 2022), although shorter oligomers can be capable of catalyzing some reactions. Catalytic sets are also more likely to emerge when there are more diverse interactions between different monomer types, as an increase in compositional space is also increasingly favorable for the propagation of collectively autocatalytic sets (Hordjick et al., 2010).

There have been many theoretical models describing how autocatalytic chemical networks could emerge in the absence of complex replicating polymers. However, despite the large number of theories, it is still unclear how such models might behave under a realistic kinetic framework. To reveal systems suitable for the emergence of catalytic sets, experimental strategies must provide favorable conditions for maximizing polymerization efficiency in prebiotic scenarios. One of these approaches, wet-dry cycling, has already been discussed in an earlier section. Another powerful strategy that can potentially enrich for motifs of catalytic sequence sets in solution is through a recursive dilution approach (Vincent et al., 2019), where a subset of the system is transferred into the next cycle and diluted with fresh ingredients. This provides a means of imparting a “chemical history” onto generations of chemical systems, allowing the outcome of one reaction cycle to be partly influenced by all other reactions that have preceded it. The remainder of this thesis describes experiments that utilize a combination of both approaches within a mineral environment, to evaluate their capability to drive the emergence of complexity.
Chapter 2: A Recursive Framework for the Detection of Autocatalytic Motifs in Mineral-Associated Oligomer Systems

I. Materials and Methods

Stock Preparation

The reaction mixture was composed of a library of 10 amino acids that have been commonly observed in non-biological scenarios and were likely available for utilization in a prebiotic context (Higgs and Pudritz, 2009; Trifonov, 2000; Cleaves, 2010) [Table 1]. Trisodium trimetaphosphate was also incorporated into the reaction mixture as a prebiotically plausible activating agent to boost peptide synthesis in addition to wetting-drying cycles (Sibilska et. al, 2018). All chosen reagents were of analytical grade purity (>99.0%) and were used without any further purification. 2L nanopore water was also autoclaved to minimize biological contamination and stored at room temperature (22-25 °C) for the duration of the experiment. Solubility constraints for each amino acid were taken into consideration by preparing each stock solution at different concentrations to allow for maximal dissolution of each reagent during preparation. In addition to their adjusted starting concentrations, 100 μL of sodium hydroxide was also added for every 3 mL of aspartic acid and glutamic acid stock prepared due to their increased dissolution at higher pH values (Bowden et. al, 2018). Each prepared stock was subsequently vortexed for 30 seconds and stored at room temperature. Prior to each transfer, a mixed amino acid master stock was assembled every generation. All individual stock solutions were replaced by freshly prepared solutions every fourth generation (beginning of every week) during the course of the experiment.
<table>
<thead>
<tr>
<th>Compound</th>
<th>Manufacturer</th>
<th>CAS #</th>
<th>Stock Concentration (mM)</th>
<th>Final Concentration (mM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>Sigma-Aldrich</td>
<td>-</td>
<td>1000</td>
<td>10</td>
</tr>
<tr>
<td>L-Alanine</td>
<td>Sigma-Aldrich</td>
<td>56-41-7</td>
<td>1000</td>
<td>10</td>
</tr>
<tr>
<td>L-Aspartic Acid</td>
<td>Alfa Aesar</td>
<td>56-84-8</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>L-Valine</td>
<td>Alfa Aesar</td>
<td>72-18-4</td>
<td>500</td>
<td>10</td>
</tr>
<tr>
<td>L-Glutamic Acid</td>
<td>Acros Organics</td>
<td>56-86-0</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>L-Serine</td>
<td>Alfa Aesar</td>
<td>56-45-1</td>
<td>1000</td>
<td>10</td>
</tr>
<tr>
<td>L-Isoleucine</td>
<td>Dot Scientific inc.</td>
<td>73-32-5</td>
<td>250</td>
<td>10</td>
</tr>
<tr>
<td>L-Leucine</td>
<td>Dot Scientific inc.</td>
<td>61-90-5</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>L-Proline</td>
<td>Dot Scientific inc.</td>
<td>147-85-3</td>
<td>1000</td>
<td>10</td>
</tr>
<tr>
<td>L-Threonine</td>
<td>Dot Scientific inc.</td>
<td>72-19-5</td>
<td>600</td>
<td>10</td>
</tr>
<tr>
<td>Trisodium Trimetaphosphate</td>
<td>Sigma-Aldrich</td>
<td>-</td>
<td>500</td>
<td>100</td>
</tr>
<tr>
<td>Sodium Hydroxide</td>
<td>Dot Scientific inc.</td>
<td>1310-73-2</td>
<td>3000</td>
<td>150</td>
</tr>
</tbody>
</table>

Table 1: Composition of Reaction Mixture

Montmorillonite was chosen as a candidate mineral environment for this study due to its prebiotic relevance (Cleaves et. al, 2012) and its well-documented role in peptide catalysis, aggregation and stabilization (Hsu, 1977; Bujdák et. al, 1996; Block et. al, 2015). K10 montmorillonite, an acid-activated form of the mineral that can have enhanced catalytic activities due to its increased surface area and pore volume (Kumar et. al, 2014), was obtained commercially (Alfa Aesar; CAS: 1318-93-0) and utilized in the experiment without further processing.
Experimental Protocol

A recursion-with-dilution strategy was adopted for the course of this study in conjunction with wet-dry cycling to investigate whether the reaction environment might be well-suited to yielding autocatalytic feedback through the emergence of effective catalytic sequences [Figure 4]. This framework was implemented in two parallel, exploratory experiments; two sets of ten replicates where either 25% of the supernatant or the mineral were continually transferred into the next experimental cycle and replenished with fresh feedstock. In the event that catalytic sequences emerge during the course of the experiment, either suddenly or over time as a result of rare polymerization reactions gradually accumulating, it might be possible to not only enrich for any catalytic sequences present in solution, but also sequences that largely propagate on the montmorillonite surface/interlayer.

To initiate the experiment (generation 0), eight scoops (~0.0562g) of montmorillonite powder were scooped into 20 2.0 mL sterile microtubes each, which were then capped at room temperature until a mixed amino acid stock solution was assembled [Table 1]. 550 µL of this
solution was dispensed into each vial, along with 200 µL of trimetaphosphate, 50 µL of sodium hydroxide and 200 µL of autoclaved nanopure water, for a total volume of 1 mL of the reaction mixture in each microtube. Each replicate was then vortexed for 15 seconds to ensure adequate mixing of both components in the reaction environment, and then dehydrated while uncapped at 90 °C for 24 hours.

To perform each 25% transfer, 20 new microtubes were divided and designated either as replicates for the upcoming solid transfer or as replicates for the supernatant transfer. Newly dehydrated replicates from the previous generation (generation 0) were also divided into two groups of 10. For the first 10 replicates as part of the solid transfer, a reaction mixture was prepared in the same way as described above, except for two scoops (~0.01405g) of the dehydrated material from the previous generation, representing the 25% transfer, and 6 scoops (~0.04215g) of fresh montmorillonite as the replenishing ‘feedstock’.

Since replicates for both transfer variants were analyzed at the same experimental stage, i.e right after dehydration is complete, 25% (2 scoops) of the dehydrated material in the supernatant transfer replicates was removed and discarded to make the mineral environments in both transfer variants more comparable, in case this discrepancy might introduce unwanted transformations between the transfer and data acquisition process. The remaining 75% of the mineral fraction was then rehydrated, vortexed for 15 seconds and then centrifuged at ~6000 RPM for 30 seconds. 250 µL of the supernatant was transferred into the next generation for each of the ten replicates, with a 75% replenishment of the supernatant involving the incorporation of 345 µL of the mixed amino acid stock solution, 180 µL of trimetaphosphate and water and 45 µL of sodium hydroxide. Eight scoops of fresh montmorillonite were also added. All 20 replicates were then vortexed for 15 seconds and then cycled under the same conditions for a total of 30 generations. Every four
generations (end of the week), samples were capped and frozen over the weekend at -80 °C instead of being prepped for a subsequent transfer. Each Monday, the frozen sample trays were removed from the freezer and thawed, during which new stock solutions were prepared to resume the transfer process. At the end of each transfer, all remaining samples were capped and placed in the -80 °C freezer for long-term storage prior to analysis.

Data Acquisition

In both transfer variants, the bulk chemical composition of all replicate lineages was characterized using an untargeted metabolomics approach with reverse-phase Ultra-Performance Liquid Chromatography coupled with tandem mass-spectrometry (UPLC-MS/MS). Stored samples from every third generation were run together in groups of 2 (ex. Generations 1 and 4, 7 and 10 etc.) until generation 22. Each sample went through multiple processing steps for subsequent data acquisition to prevent any mineral particulates from interfering with the acquisition process; they were first thawed and centrifuged for 30 seconds, until there were no visible mineral grains suspended in solution. The resulting supernatant was then removed and filtered using 96-well 0.20 µM hydrophilic polypropylene filter plates (AcroPrep™) and a vacuum manifold to separate the solution from any suspended mineral grains. After another centrifugation step, 110-180µL of the filtered mixture was transferred to LC vials for subsequent analysis. For every analyzed generation, 10 control replicates were also created an hour before the vials were loaded onto the LC-MS instrument using a freshly prepared mixed amino acid stock solution with fresh montmorillonite as described by Table 1. This way, it was possible to have a rough proxy of early system behavior in the reaction environment before any dehydration cycles or transfer histories were initiated.
The UPLC (Thermo Fisher Scientific, Waltham, USA) with a C18 column (Agilent, Santa Clara, USA) was coupled to a Thermo Q-Exactive Plus Orbitrap MS (Thermo Fisher Scientific, Waltham, USA). The sample vials were loaded onto the instrument by replicate; i.e replicate 1 was loaded first followed by replicate 2 and so on, within which each transfer variant, control replicate, and generation were organized in a random pattern to minimize order effects. 5 µL of samples were injected and then eluted in a linear gradient mixture from 0.1% v/v formic acid in water (47146-M6, Thermo Fisher Scientific, Waltham, USA) to 0.1% v/v formic acid in acetonitrile (47138-K7, Thermo Fisher Scientific, Waltham, USA) over a 14-minute period. Full MS spectra was collected for every sample for 10 minutes in positive ion mode over a scan range of 50-750 m/z, with a resolving power of 70000. Acquisition was conducted using fragmentation data collection with full MS for pooled samples in both transfer variants followed by a data-dependent MS2 analysis.

Data Processing

All raw data files acquired during the course of the experiment were processed together in Compound Discoverer™ using an untargeted metabolomics workflow that is optimized to process data acquired on Thermo Fisher Orbitraps. In addition to the default workflow, 18 other databases were added (Across Organics, Alfa Aesar, Alfa Chemistry, BioCyc, Cambridge Structural Database, CAS Common Chemistry, ChemBank, DrugBank, FDA, Human Metabolome Database, KEGG, MassBank, Merck, Millipore, MeSH, NIST, NPAtlas, Peptides) and pooled samples were set to ‘Identification only’. The resultant list of identified compounds and their relative abundances were extracted as .CSV files, and differences between both transfer variants and the control were evaluated using heteroscedastic t-tests with a Bonferroni alpha correction.
The compound dataset was then processed in python and filtered to remove any compounds that were unidentified by both name and formula, as well as compounds that were likely contaminants (i.e. formulas containing elements other than those present in the starting mixture etc.). Partial compound ID matches were also used to identify amino acid residues on identified compounds and extract them into another dataset for further analysis that might be targeted specifically towards oligomers and other amino acid-associated compounds.
Chapter 3: Long-Term Signals Suggest Compositional Dynamics May be Driven by Autocatalytic Feedback

I. Results

Here the results of a recursive mineral-dependent amino acid experiment are reported, which was designed to seek evidence of the production of functional peptides via non-enzymatic peptide synthesis. Two separate transfer experiments were carried out on a library of 10 amino acids one where the supernatant fraction of the system was transferred, and another where the solid fraction of the system was transferred, which will be referred to as “Liquid” or “Mineral” respectively. Each experimental variant contained 10 replicate lineages that underwent one wet-dry cycle before a 25% recursive transfer into the next generation of the experiment. These were denoted by an “L” (Liquid) or an “M” (Mineral) in front of their replicate number.

Figure 5: Abundances of the first 100 significant compounds in generation 1, sorted by control. The statistical significance of differences between the control and experimental samples were determined using a two-tailed t-test.
Processing the raw files with an untargeted metabolomics workflow produced a list of 15,860 mass spectral features for the Liquid variant, and 15,088 features for the Mineral variant. The extracted area values were then filtered and yielded a list of 251 and 254 peptides/amino acid associated compounds for the two experiments respectively. The unfiltered top 100 compounds for generation 1 with a significant p-value with an α set at 5% comparing each experimental treatment with the control are ordered by decreasing abundance values in the control and shown in Figure 5. Linear oligomers up to 11mers were obtained for the Liquid transfer variant and up to 9mers for the Mineral transfer variant. Some cyclic peptides were detected, likely synthesized through dehydration steps during the drying process, containing up to 4 and 6 monomers for the Liquid and Mineral transfers, respectively.

![Graphs showing Euclidean distances](image)

*Figure 6: A Euclidean Distance analysis for the Liquid and Mineral experimental variants comparing experimental lineages for a filtered dataset and an unfiltered dataset. Distances G1-22 are compared relative to generation 1.*

To compare the filtered and the raw datasets, the Euclidean distance of each generation from generation 1 was calculated for both the experimental variants [Figure 6]. The distances calculated in the filtered dataset are consistently higher for the filtered dataset for the Liquid transfer, but higher only after generation 10 in the Mineral transfer. The patterns reveal relatively
similar trends in both datasets, especially for the filtered data. It is possible that due to the larger diversity of compounds in the unfiltered dataset, the Euclidean distance between any two points becomes more similar, effectively reducing overall distance at every measured generation.

Areas for all compounds identified in the filtered dataset were also aggregated for both the Liquid and Mineral experiment variants [Figure 7A]. Overall compound abundance in both the variants shows a similar decreasing trend, but aggregated amino acid areas do not reveal any obvious trends [Figure 7B]

Principle Component Analysis (PCA) of the filtered data produced 10 principal components explaining a combined total of 80% of the variance (35.45% and 11.46% for PC1 and PC2 in the Liquid transfer, and 35.36% and 14.96% for PC1 and PC2 in the Mineral transfer) [Figure 8S]. The PCA plots, standardized with z-scores to ensure equal weighting between compound abundances and account for outliers, were plotted for the Liquid and Mineral variants in Figures 8 and 9 respectively. The PCA analysis for both transfer variants reveals an observable segregation in generations 1, 4 and 22 across all principal components, meanwhile generations 7-19 seem to converge to one area on the PCA space close to the control. However, it is apparent with a closer look at generations 7-19 from the Liquid variant that all generations show separation

![Figure 7: Aggregated compound areas and amino acid areas for the Liquid and Mineral transfer variants](image-url)
from the control [Figure 10; 10S]. Moreover, while it seems like the generations largely stabilize in the same space at around generation 10 and don’t show any distinct separation, generations 16 and 19 start separating again. It is important to note that the generation 22 lineages seem to have similar overall distance from both generations 1 and 19, suggesting that generation 22 has a similar degree of variance from both generations, and is better separated from the other generations along an entirely different axis.

Figure 8: PCA plots comparing control and experimental lineages in the Liquid transfer variant. Each color represents a different generation in the 9 plots representing PCs 1-10.
Figure 9: PCA plots comparing control and experimental lineages in the Mineral transfer variant. Each color represents a different generation in the 9 plots representing PCs 1-10.

Figure 10: PCA plot for PCs 1 and 2 comparing control and experimental lineages in the Liquid transfer variant for generations 7-19.
Bray-Curtis dissimilarities were also calculated for individual lineages to summarize compositional dynamics of the experimental lineages in the filtered dataset [Figure 11]. There is an increase in dissimilarity observed over time in both transfer variants relative to generation 1. The Mineral transfer lineages appeared to show a slower divergence from initial conditions, whereas the Liquid transfer seems to have attained a stable distance by generation 10. The average pairwise dissimilarity between each lineage was calculated to assess inter-lineage separation and revealed an increase in separation between lineages until generation 10 but no consistent increase in divergence between lineages over time [Figure 12].

Figure 11: Bray-Curtis dissimilarity measures of peptide and amino acid derivative species compositions in liquid and mineral replicate lineages relative to g01. The x-axis corresponds to the analyzed generations, while the y-axis corresponds to calculated Bray-Curtis dissimilarity. This measure is bounded between 0 and 1, where 0 implies no dissimilarities between chemical composition and 1 means there are no shared species present relative to generation 1.

Figure 12: Average pairwise Bray-Curtis distance between lineages for generations 1-22. These distances were calculated by averaging the individual distances between successive lineages at each generation, allowing for the visualization of lineage divergence over time.
In order to account for any global lineage effects that could be driven by the compositional profiles of all species present in the chemical space rather than just peptide oligomers, Bray-Curtis dissimilarities of all compound types from the raw dataset (after a processing step to remove obvious contaminants and unidentified compounds) were determined as well and presented in Figure 13. There is an overall trend of increasing lineage divergence and the emergence of unique lineage-specific patterns over time.

As an additional metric, Canberra distances were also quantified for generations 1-22 in both the Liquid and Mineral transfer variants [Figure 14]. This metric was primarily chosen to complement the Bray-Curtis dissimilarity metric due to its increased sensitivity for species with lower abundance, making it suitable to track small, early changes in peptide abundances that may precede larger, more noticeable shifts in catalytic activities. The patterns do not reveal any variations from the Bray-Curtis metric and are largely similar.

![Figure 13: Bray-Curtis dissimilarity measures of all compound classes in liquid and mineral replicate lineages relative to g01. The x-axis corresponds to the analyzed generations, while the y-axis corresponds to calculated Bray-Curtis Dissimilarity.](image-url)
The area values of free amino acids from the starting mixture were also tracked [Figures 15 and 16]. Due to limitations in the resolution of the untargeted metabolomics workflow to confidently detect glycine, its contribution to any observable trends was not considered. Free amino acid area show three broad observable patterns. In both variants, Alanine, valine, isoleucine, and threonine are largely stable and show no significant differences between generations; this is also true for aspartic acid in the Mineral variant. Proline decreases in area over time in both transfer variants, however, shows an increase in area after generation 13 for the Liquid transfer, and in generation 22 for the Mineral transfer. Serine and Leucine’s area decreases significantly over time compared to aspartic acid (Liquid variant only) and glutamic acid, which show an opposing trend. Other amino acids not from the starting library were also detected in high abundances as both free monomers and as part of peptide oligomers, including beta-alanine, histidine, lysine, phenylalanine, tryptophan, but these do not show any significant patterns of change over generations.

Figure 14: Canberra distance calculated from a mean G01 in liquid and mineral replicate lineages using the filtered dataset. The x-axis corresponds to the analyzed generations, while the y-axis corresponds to calculated Canberra distances.
Amino Acid Areas for G01-22: Mineral (By Lineage)

Figure 15: Free amino acid areas for replicate lineages M1-10 in generations 1-22. The x-axis corresponds to the analyzed generations while the y-axis corresponds to identified areas for the mineral transfer variant.
Amino Acid Areas for G01-22: Liquid (By Lineage)

Figure 16: Free amino acid areas for replicate lineages L1-10 in generations 1-22. The x-axis corresponds to the analyzed generations while the y-axis corresponds to identified areas for the liquid transfer variant.
In order to standardize amino acid abundances, area values for the starting 10 free amino acids in each generation of the Liquid transfer experiment were compared to their respective values in the control replicates [Figure 17]. Isoleucine, serine, leucine, threonine and proline have significantly higher abundances in generation 1 relative to the control, whereas aspartic acid and glutamic acid show the opposite with starting ratios at 0.69 and 0.35 respectively. Trends across the 22 generations also vary, with decreasing abundances found in serine (1.77 to 0.65) and threonine (1.25 to 0.68) and increasing abundances in aspartic acid (0.69 to 2.69) and, less clearly, in glutamic acid, whose abundance increases between generations 1-16, and then decreases (0.35 to 0.89, dropping to 0.70).

Figure 17: Free amino acid areas of the liquid transfer variant relative to the control. Left: The x-axis represents the analyzed generations while the y-axis depicts a relative ratio of amino acid areas in each generation relative to its associated value in the control. Right: Free amino acid areas in generation 22 relative to the control. Ratios < 1 imply lower amino acid abundance in the experimental relative to the control.
Finally, to assess the incorporation of free amino acids into peptides or amino acid associated compounds over the course of all analyzed generations, a relative scaled value was determined for each detected compound based on the number of amino acid residues it contained [Table 2; Figure 18]. This yields a scaled value for each of the 9 amino acids that were present in the starting library. These values were then aggregated across all compound areas for each generation and graphed as a relative measure. Broadly, this approach provides a rough proxy for the amount of incorporated amino acids present in the supernatant. While additional experiments need to be conducted to precisely quantify direct relationships between amino acid and their

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Number of Incorporated Residues</th>
<th>Longest Oligomer Formed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine (A)</td>
<td>62</td>
<td>11</td>
</tr>
<tr>
<td>Aspartic Acid (D)</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>Valine (V)</td>
<td>31</td>
<td>11</td>
</tr>
<tr>
<td>Glutamic Acid (E)</td>
<td>43</td>
<td>10</td>
</tr>
<tr>
<td>Serine (S)</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>Isoleucine (I)</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>Leucine (L)</td>
<td>46</td>
<td>11</td>
</tr>
<tr>
<td>Proline (P)</td>
<td>101</td>
<td>10</td>
</tr>
<tr>
<td>Threonine (T)</td>
<td>17</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2: Total number of incorporated amino acid residues identified (left) and the longest polymer they contributed to (right) for each of the 9 stock amino acids across all generations in the Liquid transfer
associated polymerization products, this can still be used to gain some preliminary insights into the global dynamics governing amino acid utilization and participation in this chemical space.

In both transfers, amino acids largely reveal decreasing trends in their associated compound areas over time, however it is possible that this occurs due to an overall trend towards the production of lower abundance oligomers and other compounds that are each individually below the limit of detection. Despite each amino acid having a different number of incorporated residues that were detected, leucine (46 residues), serine (21 residues) and proline (101 residues) are similarly scaled on the Liquid transfer [Figure 18B]. This is either reflective of serine’s incorporation into peptides with relatively large areas, or leucine’s, and more prominently proline’s, preference for species diversity over abundance. In the other transfer variant [Figure 18A], serine does not scale similarly to leucine and proline, instead grouping with alanine (62 residues) and valine (31 residues).

Notably, aspartic acid and isoleucine diverge from this expected trend; after an initial decrease in area until generation 7, aspartic acid’s scaled area exponentially increases until generation 22, while isoleucine increases until generation 13, and then again from generation 16.

![Figure 18: Aggregated compound area scaled by amino acid count for the Liquid and Mineral transfer. The x-axis represents the analyzed generations, while the y-axis corresponds to an aggregated value of peptide abundances scaled by the number of their constituent amino acids. Each line represents a relative estimate of the cumulative contributions of an amino acid to its associated compounds over time.](image)
In aspartic acid’s case, three main oligomers largely drive this trend, aspartyl-proline, aspartyl-glutamyl-valine and aspartyl-glutamyl-isoleucine [Figure 19].

![Figure 19: Most abundant peptides with aspartic acid residues detected after generation 7.](image)

As suggested by the diversity of trends associated with amino acid utilization by the compositional space, these results highlight the dynamic nature of our experimental system, where initial drying conditions and partial feedstock replenishment heavily influence early monomer acquisition for peptide synthesis and proliferation.

**II. Discussion**

*Limitations and considerations for future experiments*

It is first necessary to acknowledge some limitations to the experimental framework and data acquisition process. The procedure used in the mineral transfer variant was originally designed to target any reaction products that were adsorbed onto the mineral surface. However, because the protocol did not include a washing step, it would also have transferred all the soluble components. The presence of the soluble fraction may explain the similar trends observed between the mineral
and liquid experimental variants. It is possible to incorporate a washing step in a future iteration of this experiment which could help elucidate whether the observed similarities in patterns are indeed due to an influence from the soluble fraction, or if both transfer approaches truly yield similar experimental outcomes.

Another potential problem with this experiment is the low number of peptides identified in our filtered dataset. We used an untargeted LC-MS data acquisition strategy designed for metabolomic applications. While such untargeted acquisition can be useful for facilitating the detection and identification of a large number of compounds, especially when using a UPLC coupled to an Orbitrap MS, which results in very high resolving power with dynamic area ranges starting from $10^3$ for complex prebiotic mixtures. Nonetheless, existing metabolomics workflows are biased towards compounds of biological and pharmaceutical significance, which may explain why only a small number of oligomers were conclusively identified by CompoundDiscoverer™. Furthermore, any products of rare reactions occurring in the transfer replicates, such as peptides beyond a certain length, would not be present in sufficiently high abundances, and may fall to be above the detection threshold of the LC-MS. Due to this, these analyses are based on only a small subset of the actual chemical space.

Whether it can be ultimately shown that catalytic sequences have emerged will require the characterization of later generations and would benefit from the experiment being replicated in its entirety. This would allow for any findings to be corroborated and would help evaluate which results repeat exactly, and which, if any, vary and suggest historical contingency. In addition to the incorporation of a washing step into the mineral variant, as mentioned above, it might be worth considering not just analyzing the dissolved products but working to separate mineral grains and then extracting products from the mineral surface, such as with a 0.5M KCl solution (Naidja &
Huang, 1994). Furthermore, an additional set of controls could be assembled and dehydrated under the same conditions as the transfers, but without undergoing recursion-with-dilution as to account for any variation between stock solutions or positional effects during the wet-dry cycle.

Despite these caveats, investigations into compound-specific and global trends suggest a dynamic non-linear interplay between the compositional space and the montmorillonite surface that highlights the simple, yet powerful nature of this recursive framework in enriching for the emergence of autocatalytic feedback.

*Key changes in montmorillonite-associated adsorbance properties drive amino acid dynamics*

In the comparisons of free amino acid abundances between the transfers and the control, glutamic acid and aspartic acid reveal much lower experimental/control ratios compared to other amino acids from the starting mixture, particularly in generation 1. Since these controls were designed to represent the system’s initial state, i.e. before wet-dry cycling or transfers, this low ratio likely represents either an incorporation of the amino acid into other products or adsorption onto the montmorillonite surface during the dehydration process, barring environmentally induced degradation. Conversely, a high experimental/control ratio could become possible if either the amino acid species were regenerated through various chemical transformations, such as peptide bond cleavage or transamination reactions, or through desorption from the montmorillonite surface during the wetting-drying cycle before analysis.

As with other smectites, adsorption on montmorillonite tends to favor charged groups over those with neutral side chains through electrostatic interactions. With its net negative charge resulting from various isomorphous substitutions within the clay structure (Avena & Pauli, 1998), montmorillonite preferentially adsorbs positively charged, basic amino acids (lysine, arginine,
histidine) relative to amino acids with acidic or neutral side chains (aspartic acid and glutamic acid). It is worth noting that basic pH of the reaction environment (at least before drying) would ensure most of the aspartic acid and glutamic acid would be deprotonated and repelled by the clay’s surface. In contrast, other amino acids are neutral and would be expected to have a higher binding affinity in these conditions. If it is likely that these associations establish quickly upon first contact, it means that the concentrations of the dissolved amino acids seen in control replicates is likely skewed towards aspartic acid and glutamic acid relative to the other amino acids.

It is thermodynamically and kinetically more complicated to estimate how the configuration of the experimental system might change during the dehydration process, where increased molecular crowding could introduce additional associations between free and adsorbed amino acids, reaction products and the clay edges or interlayers of the montmorillonite. While research documenting the behavior of a complex interacting system such as this is scant, it is difficult to believe that the adsorption properties of montmorillonite would remain unchanged after a wet-dry cycle, especially given evidence that adsorbed species can “condition” surface properties to make the surface more or less suitable for other species in the system (Gao et.al, 2017). Therefore, it seems quite likely that after one or more dehydration cycles, montmorillonite’s adsorption selectivity would not simply boil down to comparing adsorption constants for individual amino acids in the starting library, but rather would be a function of all interacting species present in the compositional space.

It would then not be unreasonable to assume that one or more wetting-drying cycles could result in a system configuration that is able to preferentially adsorb negatively charged amino acid species such as aspartic acid or glutamic acid at surface or interlayer sites after a dehydration cycle. This could explain the low experimental/control ratios at generation 1 for these amino acids
relative to other amino acids. Additionally, given that there is no clay transfer in the Liquid transfer variant, and therefore no propagation of mineral-associated history, it is quite interesting that the experimental/control ratio appears to increase for free aspartic acid and glutamic acid whereas other amino acids tend to decrease over time. If it is assumed that conditions for negatively charged amino acids become more favorable after a dehydration cycle, and the process is ‘reset’ after every transfer, then it possible that some process propagated through the supernatant is either allowing free aspartic acid and glutamic acid to react more efficiently within solution over time or is able to ‘re-direct’ the way in which each dehydration cycle manipulates the montmorillonite surface.

Comparing changes in free amino acid abundance over time for the Liquid transfer reveal directional trends for many amino acids in the starting mixture (Figure 16), particularly aspartic acid and glutamic acid, which increased over subsequent generations, serine, leucine and perhaps proline, which decreased. In a system where the rate of amino acid utilization is constant, free amino acid abundance would be expected to converge to a stable value after a few initial generations. Many amino acids do not deviate from this expected pattern and do no display any remarkable changes from the pattern.

Free aspartic acid’s abundance in solution shows an increase over time, even when the abundance of peptides containing its residues shows an increasing trend after generation 7. It is difficult to come to firm conclusions on whether this discrepancy reflects an actual increase in peptide incorporation, or through a potential hydrolysis of longer peptides. If it is the latter interpretation, these results are insufficient to determine whether this might occur due to bond cleavage events due to the effects of frequent rehydration, or due to some underlying catalytic processes since many of the reaction products likely driving such behavior might not even be
present under detectable ranges of the LC-MS. The presence of an underlying catalytic motif, however, could be suggestive of enrichment by some type of positive feedback in the system, likely enriched in later generations.

With serine and leucine, measures of free amino acid abundance reveal a consistently declining trend over time. A sustained decrease such as this suggests that these amino acids are being utilized by some process that becomes more favorable over time. Even without a detailed understanding of the participating species that might be driving these trends, it is evident that free serine and leucine are consistently, and efficiently being utilized over time despite the constraints posed by the recursive framework of this experiment. This implies that these trends might likely be driven by some kind of positive feedback process present in the system. For example, it is possible that serine or leucine are incorporated into catalytic sequences that then change the rate or efficiency at which they are incorporated into peptides, or interact with the mineral surface after dehydration.

*Preliminary evidence for global variation driven by potentially autocatalytic feedback*

The PCA plots for both transfer variants reveal a similar pattern in the overall composition of the solutions in generations 1 and 4, which are quite distinct, after which the system seems to stabilize in a relatively narrow region of composition space through generation 19, with generation 22 being distinct again. The initial period of change can be seen as a phase when diverse and abundant compounds are generated during early dehydration cycles, which converge towards a steady state by generation 7. This is the expected patterns in the case of linear chemistry. However, the fact that the system seems to transition into another state starting from generation 16 that begins diverging slowly at first and then accelerating by generation 22 implies non-linearity of some kind.
The Bray-Curtis metric reveals similar findings in its quantification of compositional dissimilarity over time. In both measures, i.e using only oligomers or including all compound classes, the dissimilarity index ranges from 0 to 1, where 0 indicates an identical composition to generation 1. There is a steady increase in dissimilarity that plateaus by generation 10, indicating that the compounds detected in the compositional space are likely maintaining relatively stable concentrations by this time. However, similar to the PCA analysis, a second, more subtle change in compositional dissimilarity is observed starting from generation 16, suggesting the possibility of a shift to another system state.

An obvious hypothesis for this pattern in overall composition is that certain sequences emerged gradually in concentrations individually below the detection limit of the LC-MS over the first 16 generations. This implies autocatalytic feedback, where longer, potentially catalytic sequences might need to accumulate to a certain threshold before they can display significant levels of catalytic activity that can influence compound abundances in the compositional space. This could also explain why the transition to a new system state starts slowly in generation 16 and accelerates for generation 19 and then generation 22. However to conclusively confirm the presence of either autocatalytic feedback, or catalytic peptide sequences that might be driving this behavior, much more work. In particular, additional data from later generations will be required to confirm the presence of such a process and elucidate its behavior. If these assumptions are largely valid, future generations might be expected to either stabilize close to generation 22 on the PCA space or continue changing as the compositional space continually explores new states.
References


Figure 8S: Scree plots for the Liquid and Mineral treatment showing variance explained by principal components (PCs) standardized by z-score. The x-axis represents the PCs while the y-axis denotes variance explained (%). Approximately 80% of the variance is explained by the first 10 PCs.

Figure 10S: Scree plot for the Liquid showing variance explained by principal components (PCs) standardized by z-score for generations 7 to 19. The x-axis represents the PCs while the y-axis denotes variance explained (%). Approximately 80% of the variance is explained by the first 14 PCs.