KINETICS OF PEPTIDE FORMATION IN THE CHEMICAL ORIGINS OF LIFE

by

Hayley A. Boigenzahn

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemical and Biological Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON

2023

Date of final oral examination: 23 August 2023

The dissertation is approved by the following members of the Final Examination Committee:
John Yin, Professor, Chemical and Biological Engineering
David Baum, Professor, Botany
Victor M. Zavala, Professor, Chemical and Biological Engineering
David Lynn, Professor, Chemical and Biological Engineering
Regina Murphy, Professor Emeritus, Chemical and Biological Engineering

ABSTRACT

We know remarkably little about the origins of life on our planet. The presence of small organics on the early Earth is generally accepted, but our understanding of the transition from small organic molecules to life-like systems is speculative. Understanding these processes would provide insight into the early evolution of biopolymers.

Among the first organic molecules synthesized from inorganic gases were amino acids, the building blocks of peptides and proteins. Peptides have the potential functionality required to develop ‘life-like’ behaviors like autocatalysis, which in this context refers to the ability of a molecule or set of molecules to promote their own synthesis. Autocatalytic peptides exist, but have never been formed from amino acid mixtures using prebiotic reaction conditions. This begs the question of how the transition from simple peptide mixtures to autocatalytic reaction networks could occur. However, that question cannot be addressed until the more concrete questions of how to search for and identify autocatalytic peptides can be answered. Since autocatalysis is a kinetical behavior, kinetic studies are a logical way to seek it out. Peptide reaction networks can be very complex, and few studies have attempted to develop simplified approaches to characterize the kinetics of experimental systems.

We studied the kinetics of a specific reaction, the trimetaphosphate-activated polymerization of amino acids. In the first introductory chapter, we briefly discuss the chemical origins of life field, then review literature related to prebiotic peptide formation. Chapter 2 discusses two reaction mechanisms for the trimetaphosphate-activated process, and demonstrates how the products of one mechanism can promote the occurrence of the other. In Chapter 3, we define an ordinary differential equation model summarizing peptide formation and hydrolysis, then estimate rate parameters based on experimental data. The limitations and potential uses of the parameters determined from this approach, which are often low precision, are discussed. Chapter 4 applies the model developed in Chapter 3 to the reactions explored in Chapter 2, and examines how cyclic environmental conditions can interact with kinetic features to move a system away from equilibrium and increase the formation of kinetically favorable products.

DEDICATION

To my parents

ACKNOWLEDGMENTS

First, I want to thank Prof. John Yin, my advisor, for his mentorship, guidance, and faith in my abilities over the past 5 years. Through the many turns this project has taken, including the strange times with COVID, I couldn’t ask for a better boss.

Thank you to my collaborators, without whom much of this work would not exist. Thank you Jaron Thompson and Leo González for letting me use your code and being patient with my endless math questions. Thank you Graham Delafield for being an exceptionally clear communicator, and for your patience and persistence on a project that never quite worked right.

Of course, I also owe thanks to many other people who don’t appear in the author lists my papers. I want to thank the other members of the Yin lab, especially Izabela Sibilska-Kaminski for teaching me about the HPLC during my first year, Huicheng Shi, for helping me get started with lab work, Tolulope Perrin-Stowe for the interesting conversations and advice about grad school, and Nan Jiang for always being a friendly face (and giving me excellent tea).

I also want to thank the students in the Baum lab, I really enjoyed having lab meetings with you. Thank you to Stephanie Colón-Santos, Lena Vincent, Zhen Peng, Praful Gagrani, Tymofii Sokolskyi, and Pavani Gangju for consulting on lab work, modeling questions, and discussing origins of life topics with me.

Additionally, I want to thank all the amazing administrative and technical support staff in both the Chemical Engineering Department and the Wisconsin Institute of Discovery for answering questions, organizing events, and keeping the program running. I especially want to thank Kate Fanis for everything she does to help us grad students stay on track.

For the last of my professional acknowledgements, I want to thank my committee members, David Lynn, Regina Murphy, Victor Zavala, and David Baum, for their guidance and suggestions as the project has developed. I particularly want to acknowledge Prof. Baum and Prof. Zavala for supporting my collaborations with their students and their thoughtful editing.

On a more personal note, I want to thank my friends both in and out of the department. It would be too much to list everyone and what they mean to me by name here, but to my friends from WPI, BBDT, sailing, or D&D, to my fellow grad students from in and out of my department, and anyone else I might have accidentally skipped in this list – thanks for supporting me and giving me something to look forward to after work. For their assistance with grad school specifically, I want to thank Lawrence Chen, who helped me survive math during our first year, and Maya Venkataraman, who was a great ChEGS co-president with me during a time when a lot of things still felt uncertain.

This is weird, but I also want to acknowledge my cat, Hildegard. She can't read this, but she makes my life better.

Last but most certainly not least, I want to thank my family, especially my parents and my brother, for having confidence in me even when I didn't have much in myself. Thank you, Mom and Dad, for always being proud of me and supporting me through the good times and the bad. I literally couldn't have done it without you.

To everyone here, and anyone I may have missed, know that I hugely appreciate your support. Thank you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CSTR	Continuous stirred tank reactor
DKP	2, 5-diketopiperazine
DoE	Design of experiments
FMOC	Fluorenylmethoxycarbonyl protecting group
HPLC	High performance liquid chromatography
IP	Ion pairing
MLE	Maximum likelihood estimation
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MSE	Mean squared error
NMR	Nuclear magnetic resonance
ODE	Ordinary differential equation
OPA	O-pthalaladehyde
RP	Reverse phase
SIPF	Salt-induced peptide formation
SPCA	Sparse principal component analysis
TP	Trimetaphosphate

EXAMPLES OF PEPTIDE ABBREVIATIONS

GG	Diglycine
G_3	Triglycine
AG	H-Ala-Gly-OH

1. INTRODUCTION

Relatively little is known about the chemistry that led to the origins of life on earth. In 1953, the Miller-Urey experiment showed that amino acids could be synthesized from inorganic gases using an electrical spark in conditions believed at the time to simulate an early Earth environment (Miller 1953). Miller's experiments provided experimental support for the Oparin-Haldane hypothesis, a theory proposed independently by Alexander Oparin and J.B.S. Haldane in the 1920s which suggested that life originated from the spontaneous emergence and gradual organization of organic monomers in the 'primordial soup' of the early Earth (Haldane 1929; Oparin 1957; Fry 2006). However, despite all the additional research that has occurred in the years following Miller's experiments, there is still no experimental evidence of small organic molecules developing into a system with life-like properties, such as the ability to replicate and evolve, in plausible early Earth conditions. Exploring the advent of larger biomolecules and the potential emergence of life-like interactions between them is a challenging problem due to the extremely large experimental search space, which spans many unknowns and competing theories.

1.1 Theories of the origin of life

One of the challenges in investigating the origins of life-like systems is that it is not entirely clear what the minimal definition of life should be. This subject will only be briefly addressed here, as it has already been the subject of various reviews and special editions in previous literature (Tirard et al. 2010; Gayon et al. 2010). The capacity for replication and evolution are generally agreed on as features required to establish life-like behavior, with NASA endorsing the definition of life as "a self-sustaining system capable of undergoing Darwinian evolution" (Luisi 1998). In the context of chemical systems, replication often refers to autocatalysis, meaning molecules or sets of molecules that promote their own synthesis. However, autocatalysis and evolvability do not necessarily encompass all the other features that are essential to life as we know it today. Other behaviors may have arisen alongside or even preceded the emergence of replication, like metabolism and compartmentalization. The ability of a system to sustain itself using external resources (Castresana & Saraste 1995) and maintain a separate state while still exchanging materials with its environment (Monnard & Walde 2015) would have greatly increased the survivability of an early autocatalytic system. The order in which these behaviors appeared is still debated.

Various origins of life theories have been proposed, often with different classes of organic molecules being suggested as the first to appear. The best-known theory is the RNA world hypothesis, which suggests that autocatalytic oligonucleotides were the first replicating molecules to evolve (Schwartz 2010; Woese 1967). Other theories include the metabolism-first hypothesis, which suggests that autocatalytic metabolic cycles existed prior to the emergence of larger replicating biopolymers (Castresana & Saraste 1995; Shapiro 2000), and the peptide and amyloid world hypotheses, which suggest that peptides and specifically amyloidogenic peptide structures acted as informational polymers prior to the emergence of RNA (Brack 2007; Rode et al. 2007; Maury 2009; Ikehara 2014). The complexity of products formed in origins of life experiments also tends to suggest that multiple classes of organic molecules may have coexisted, and the interactions between those molecules has also been proposed as the source of life-like behaviors (Patel et al. 2015; Ruiz-Mirazo et al. 2014; Frenkel-Pinter et al. 2020).

There have also been a variety of environments proposed for the origin of life. The most popular theories involve hydrothermal fields, either on land or underwater. Organic molecules may have been able to accumulate and polymerize on the surfaces of rocks and minerals in hydrothermal pools while being periodically rehydrated by rainfall or tidal cycles (Lahav & Chang 1976; Damer & Deamer 2015). Underwater hydrothermal vents may have also been able to accumulate organic molecules in the surrounding minerals while continued polymerization was driven by the disequilibrium between the minerals around the vent and the surrounding ocean (Corliss et al. 1979; Russell & Hall 1997; Martin et al. 2008). Other theories suggest that life formed not in high-temperature hydrothermal conditions, but in cold conditions where long organic polymers would be more stable (Bada et al. 1994). These theories cover a range of pH, temperature, and reactant conditions, and are not necessarily mutually exclusive (Omran & Pasek 2020). Life-like systems may have formed more than once in different environments, or formed and adapted to multiple environments prior to the evolution of the last universal common ancestor (LUCA) of all extant life on Earth (Cantine & Fournier 2018).

Rather than advocating for any particular theory or environment, the primary aim of this dissertation is to investigate the general process of simple monomers forming more complex molecules and explore methods to seek the emergence of complex behavior within those systems. Instead of attempting to replicate a specific hypothetical environment on the early Earth, these experiments were conducted in ‘prebiotically plausible’ conditions. This refers to conditions which do not include any modern biological or synthetic molecular species that were unlikely to exist on the early Earth, but also do not use an anoxic chamber, which is required to accurately replicate the low-oxygen conditions of the early Earth (Canfield 2005). This research was not designed to evaluate the likelihood of one theory or environment, but almost exclusively discusses peptides as a model system and uses wet-dry cycles as a straightforward method of increasing peptide yield.

1.2 Peptides at the origin of life

Peptides are an interesting biopolymer to research since they almost certainly played an important role in the very early stages of the origin of life. They are relatively stable compared to nucleic acids and oligonucleotides, making them slower to degrade in adverse environmental conditions (Rode 1999; Nelson et al. 2000). Amino acids were the first organic monomers identified in the Miller-Urey experiment and have been found in many subsequent organic synthesis experiments (Miller 1953; Kitadai & Maruyama 2018) as well as in multiple meteorites (Cronin & Pizzarello 1983; Glavin et al. 2012). There is an established inorganic synthesis route for amino acids – they form from carbonyl compounds, ammonia, and hydrogen cyanide gas via Strecker synthesis – and are considered one of the most likely organic monomers to have been readily available on the early Earth (Brack 2007; Frenkel-Pinter et al. 2020). Peptides also have the potential to serve a diverse variety of molecular functions – some short peptides are active biological molecules, even in modern life (Lau & Dunn 2018; Sheehan et al. 2019). Short peptides designed to coordinate with metal ions have been shown to have esterase (Rufo et al. 2014) and hydrogenase (Timm et al. 2023) activity. Other peptides can act as structural or molecular templates by assembling into specific structures that promote the formation of a complementary peptide. Autocatalytic behavior has been found using coiled-coil structures, which involves two alpha helices coiled together (Ashkenasy et al. 2004; Dadon et al. 2015), and amyloid-like beta sheets (Rufo et al. 2014; Rout et al. 2018). These peptides were rationally designed specifically to be autocatalytic, but the success of these experiments demonstrates the potential for peptides to develop self-replicating behavior.

Various methods have been published in literature for forming peptides from amino acids in prebiotically plausible environments. Peptide bond formation is a condensation reaction and unfavorable in water at standard conditions (Fig. 1.1), so two general strategies are used – dehydration and activation (Danger et al. 2012). Dehydration often simply involves allowing samples to dry, sometimes repeatedly (Lahav & Chang 1976; Mamajanov et al. 2014; Forsythe et al. 2015; Yu et al. 2017). The drying process shifts the equilibrium toward peptide bond formation and has the added benefit of increasing the concentration of the reactants as they approach the solid phase. High concentrations of salts can also promote condensation, since many of the water molecules move to solvate the salts, which reduces the overall water activity even while the molecules are still in the liquid phase (Schwendinger & Rode 1989; Campbell et al. 2019).

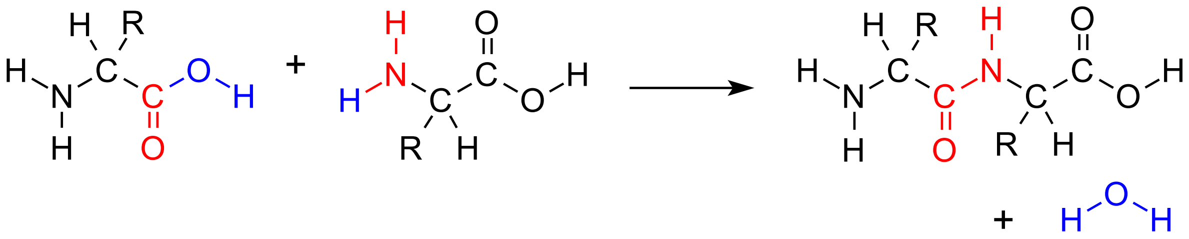


Figure 1.1 Overall mechanism of peptide bond formation. (Public domain.)

Activation is the use of additional, often high-energy, molecules or reaction conditions that form intermediates with the amino acids or peptides, which then spontaneously react to form peptide bonds. A wide variety of activating agents and conditions have been suggested in literature. An early suggestion was that the formation of peptides occurred via heterogeneous catalysis on the surfaces of minerals and clays since mineral surfaces can accumulate peptides via surface adsorption and promote peptide bond formation (Lambert 2008; Erastova et al. 2017). The role of dissolved metal ions, particularly divalent salts, on peptide bond formation has also been studied because peptides can form complexes with metal ions (Biester & Ruoff 1959; Wang et al. 2019) and there are metal ions present at the active sites of many modern enzymes (Belmonte & Mansy 2016; Kitadai et al. 2011). A specific example of metal ions promoting peptide bond formation is the salt-induced peptide formation (SIPF) method which uses Cu(II) and high NaCl concentrations (Rode & Schwendinger 1990). In addition to minerals and salts, a variety of other high-energy compounds have been used as condensing agents, such as cyanamide, dicyanamide, urea, carbonyl sulfide (COS) and carbon disulfide gas (CS_2), all of which are reviewed in Frenkel-Pinter et al. (2020). Although its stability in prebiotic conditions is uncertain, the aromatic heterocycle imidazole has also been used to promote peptide bond formation (Sawai & Orgel 1975; Serov et al. 2020). The challenge with many high-energy activating agents is that, since they are consumed during the reaction, it is unclear whether they would have been available with enough consistency and in high enough concentrations to support an emerging life-like system. However, some are very effective and produce much higher peptide yields than methods without energetic activating agents (Frenkel-Pinter et al. 2020).

The activating agent used in this work is trimetaphosphate (TP), a cyclic inorganic phosphate which may have been an inorganic precursor to adenosine triphosphate (ATP) (Rabinowitz et al. 1969). A possible volcanic synthesis route for TP has been published, and it is thought to have existed on the early Earth (Yamagata et al. 1991). Cyclophosphates are sometimes unstable because they are prone to ring-opening in certain conditions (Glonek 2021), but they were recently discovered in natural minerals, demonstrating their potential to appear and persist in the environment (Britvin et al. 2021). Although they are rare in modern minerals, the conditions of the early Earth are believed to have been more reducing than our current atmosphere (Hao et al. 2019) and were probably more conducive to forming and maintaining cyclophosphates. TP-activated peptide formation is mainly studied in basic pH conditions, although it has been shown to be an effective activator across a variety of pH and temperature settings (Chung et al. 1971; Ying et al. 2018; Sibilska et al. 2018; Serov et al. 2020).

Despite the numerous published routes for prebiotic peptide formation, a bottom-up approach using amino acids to create peptides capable of emulating the behavior seen in designed autocatalytic peptide systems has yet to be discovered. One challenge is that all systems designed so far involve peptides that are at least 8 acids long (Lee et al. 1996; Rubinov et al. 2009; Dadon et al. 2015; Rout et al 2022). The thermodynamics of amide bond formation would make the emergence of peptides of that length exceedingly rare in solution (Ross & Deamer 2016; Lambert 2008). While dehydration and activation make the formation of longer peptides feasible, these lengths are on the upper end of what has been achieved without using activated amino acid derivatives and are still challenging to reach in many conditions (Rodriguez-Garcia et al. 2015). Designed autocatalytic peptides also usually contain a variety of amino acids, some of which were more likely to exist in meaningful concentrations on the prebiotic Earth than others (Zaia et al. 2008). Some diversity of reactants is prebiotically plausible since prebiotic synthesis routes have been found for all 20 proteinogenic amino acids and many nonproteinogenic ones (Kitadai & Maruyama 2018), but the combination of length and reactant diversity quickly creates an enormous number of possible products.

Although not all peptides will form with equal likelihood, and it is probable that some species will not form at all, the number of peptides which can potentially form is combinatorial – the number of amino acid species to the power of the peptide length. Although several recent studies using mass spectrometry (MS) have made progress towards making certain kinds of experimental analyses of these combinatorically complex systems more tractable, quantitative studies of such mixtures are still daunting (Surman et al. 2019; Jain et al. 2022). Additionally, with so many potential products, even if longer peptides are formed, there is no guarantee that the species directly involved in autocatalytic behavior will form in sufficiently high concentrations or interact with enough specificity to demonstrate autocatalytic activity (Plasson et al. 2011). Thus, both forming an autocatalytic peptide using prebiotic reaction mechanisms and identifying such species if they were to form are ongoing challenges.

1.3 Kinetic studies of peptide formation

Kinetic studies can help address several of the questions of bottom-up biopolymer formation. The emergence of kinetic control, a regime in which the products formed are determined by the reaction rates and not the by the lowest energy products of the system, can allow much longer peptides than are thermodynamically favorable to emerge and persist (Eschenmoser 2007). Kinetic control is thought to be an important step in the transition to life-like behavior (Pross 2003; Pascal et al. 2013). A specific example of this is the concept of kinetic trapping in wet-dry cycles, which suggests that if bonds formed during the dry phase are slow to hydrolyze once back in the wet phase, it will result in the gradual accumulation of longer polymers (Ross & Deamer 2016). Identifying systems which appear to be following kinetic control can help surpass thermodynamic limits and determine which conditions are favorable for forming and maintaining populations of longer polymers (Astumian 2019).

Kinetic studies can also be used to identify the emergence of catalysis or autocatalysis, since the appearance of these behaviors may reasonably be expected to significantly alter the reaction rates of a system. With an unlimited reactant source, autocatalytic systems will grow exponentially, but in a system with limited reactant availability, they eventually limit their own growth, creating a distinctive sigmoidal concentration profile (Plasson et al. 2011). Identifying a sigmoidal concentration profile in an experiment is therefore suggestive of autocatalysis, though secondary environmental effects still need to be ruled out (Imai et al. 1999; Rout et al. 2022). Top-down studies, which use synthetic autocatalytic peptides, use kinetics to study the mechanisms of peptide autocatalysis and properties of autocatalytic systems, such as template dependence (Lee et al. 1997; Rubinov et al. 2009).

Bottom-up studies, which start from amino acids or short peptides, mostly use kinetics to explore the effects of various reaction conditions (Pasternack et al. 1972; Sakata et al. 2010; Mamajanov et al. 2014; Yu et al. 2017) and occasionally to verify reaction mechanisms (Yu et al. 2016; Serov et al. 2020). Many kinetic studies analyzing the hydrolysis rates of peptides in various environments have also been performed, since these rates are also significant in biochemistry (Lawrence & Moore 1951; Radzicka & Wolfenden 1996; Qian et al. 1993; Sheehan et al. 2019; Sun et al. 2019). Modern kinetic studies are generally performed using high performance liquid chromatography (HPLC), quantitative nuclear magnetic resonance (NMR), or quantitative liquid chromatography mass spectrometry (LC-MS). Two broad conclusions can be gathered from these studies as a whole: first, yield and distribution of peptide formation is highly dependent not only on the activating agent used, but on the details of the environmental conditions (Surman et al. 2019; Sibilska et al. 2018) and, second, sustained peptide formation from amino acids is quite difficult to achieve for longer than a few days, even in cyclic or activated conditions (Sakata et al. 2010; Yu et al. 2017; Serov et al. 2020).

There is still a clear gap between what has been achieved using bottom-up synthesis and the behavior found in top-down studies (van der Gulik et al. 2009). Since we do not know the conditions required for catalysis to emerge, it is not clear how large this gap actually is, which has been the subject of several theoretical studies (Kauffman 1986; Mossel & Steel 2005; Markovitch & Lancet 2012; Martin & Horvath 2013; Intoy & Halley 2017). In a system where random peptides catalyze random reactions at a given frequency, autocatalysis almost inevitably emerges as the system increases in complexity (Kauffman 1986). However, the estimated complexity of the system at which autocatalysis appears varies significantly depending on the assumptions made about the frequency of catalysis. The addition of other factors, such as external energy, catalytic specificity and efficiency, and whether the system consumes externally available ‘food’ molecules also affect the frequency of catalysis required to establish an autocatalytic system (Intoy & Halley 2017; Mossel & Steel 2005). Other modeling studies of prebiotic polymer formation, addressing topics like the effect of drying and the trade-offs of reaction network complexity have also been performed (Walker et al. 2012; Varfolomeev & Lushchekina 2014; Virgo & Ikegami 2013; Ross & Deamer 2016). The implications of these model results on the relationship between complexity and catalysis and the ease of establishing autocatalytic sets are interesting, but since theoretical studies use abstracted systems described by parameters that cannot be measured experimentally, it is difficult to directly compare them to real-world chemical systems.

Sequences and mechanisms for experimental autocatalytic peptides are based on modern proteins, which are far more developed than what likely formed on the early Earth. There may be much shorter examples of autocatalytic peptides which have yet to be identified or forms of autocatalysis with less complex mechanisms – evidence of one peptide increasing the yield of another has been found using dimers and even amino acids (Suwannachot & Rode 1999; Plankensteiner et al. 2002; Gorlero et al. 2009; Li et al. 2010). These peptides probably behave more like intermediates than traditional catalysts, forming multiple bonds between reactive species before hydrolyzing into a species that is unfavorable to form directly, but this behavior may still be significant for increasing the system complexity and potentially forming autocatalytic cycles.

Overall, the large number of uncertainties and the technical demands of quantitative studies makes searching for emergent autocatalytic peptides in plausible early Earth conditions challenging. Kinetic studies are already useful for studying reaction mechanisms and environmental effects, but exploring additional methods of performing kinetic studies can hopefully enable broader searches for autocatalysis.

1.4 Motivation

The larger challenge motivating this work is the question of how the transition from small organic molecules into larger catalytic biopolymers could occur, and how it could be detected experimentally if it did. The size of the experimental search space is already a major challenge in addressing this question, so we focused on studying the kinetics of a specific set of polymerization conditions – the formation of peptides from amino acids using drying conditions and TP. In Chapter 2, we discuss how TP-activated peptide bond formation both affects and is affected by environmental conditions. These interactions are significant because they change not only the reaction rates, but the underlying reaction mechanism of peptide bond formation. In Chapter 3, we propose a simplified kinetic model for a peptide reaction network and a method of estimating the model parameters based on experimental data. We also discuss why the approximations of this approach are limited for 'sloppy' models, which are models with characteristics that make parameter estimation very sensitive to even small amounts of noise. In Chapter 4, we apply the approximation method discussed in Chapter 3 to explore possible kinetic behaviors in simple peptide networks and explore how such behaviors may contribute to the emergence of selectivity and kinetic control. Finally, in Future Recommendations, we discuss the need for continuing analytical method development, propose further environmental conditions for kinetic studies, and suggest additional methods of searching for autocatalytic peptides.

1.5 References

Ashkenasy, G., Jagasia, R., Yadav, M., & Ghadiri, M. R. (2004). Design of a directed molecular network. *Proceedings of the National Academy of Sciences of the United States of America, 101*(30), 10872–10877. https://doi.org/10.1073/pnas.0402674101

Astumian, R. D. (2019). Kinetic asymmetry allows macromolecular catalysts to drive an information ratchet. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-11402-7

Bada, J. L., Bigham, C., & Miller, S. L. (1994). Impact melting of frozen oceans on the early Earth: Implications for the origin of life. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(4), 1248–1250. https://doi.org/10.1073/pnas.91.4.1248

Belmonte, L., & Mansy, S. S. (2016). Metal catalysts and the origin of life. *Elements*, *12*(6), 413–418. https://doi.org/10.2113/gselements.12.6.413

Biester, J. L., & Ruoff, P. M. (1959). Structural Influences on the Stability of Dipeptide-Metal Ion Complexes. *Journal of the American Chemical Society*, *81*(24), 6517–6521.

Brack, A. (2007). From interstellar amino acids to prebiotic catalytic peptides: A review. *Chemistry and Biodiversity*, *4*(4), 665–679. https://doi.org/10.1002/cbdv.200790057

Campbell, T. D., Febrian, R., McCarthy, J. T., Kleinschmidt, H. E., Forsythe, J. G., & Bracher, P. J. (2019). Prebiotic condensation through wet–dry cycling regulated by deliquescence. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-11834-1

Britvin, S. N., Murashko, M. N., Vapnik, Y., Vlasenko, N. S., Krzhizhanovskaya, M. G., Vereshchagin, O. S., Bocharov, V. N., & Lozhkin, M. S. (2021). Cyclophosphates, a new class of native phosphorus compounds, and some insights into prebiotic phosphorylation on early Earth. *Geology*, *49*(4), 382–386. https://doi.org/10.1130/G48203.1

Canfield, D. E. (2005). The early history of atmospheric oxygen: Homage to Robert M. Garrels. *Annual Review of Earth and Planetary Sciences*, *33*, 1–36. https://doi.org/10.1146/annurev.earth.33.092203.122711

Cantine, M. D., & Fournier, G. P. (2018). Environmental Adaptation from the Origin of Life to the Last Universal Common Ancestor. *Origins of Life and Evolution of Biospheres, 48*(1), 35–54. https://doi.org/10.1007/s11084-017-9542-5

Castresana, J., & Saraste, M. (1995). Evolution of energetic metabolism: the respiration-early hypothesis. *Trends in Biochemical Sciences*, *20*(11), 443–448. https://doi.org/10.1016/S0968-0004(00)89098-2

Chung, N. M., Lohrmann, R., Orgel, L. E., & Rabinowitz, J. (1971). The mechanism of the trimetaphosphate-induced peptide synthesis. *Tetrahedron*, *27*(6), 1205–1210. https://doi.org/10.1016/S0040-4020(01)90868-3

Corliss, J. B., Dymond, J., Gordon, L. I., Edmond, J. M., Herzen, R. P. von, Ballard, R. D., Green, K., Williams, D., Bainbridge, A., Crane, K., & Andel, T. H. van. (1979). Submarine Thermal Springs on the Galápagos Rift. *Science*, *203*(4385), 1–23.

Cronin, J. R., & Pizzarello, S. (1983). Amino acids in meteorites. *Advances in Space Research*, *3*(9), 5–18. https://doi.org/10.1016/0273-1177(83)90036-4

Dadon, Z., Wagner, N., Alasibi, S., Samiappan, M., Mukherjee, R., & Ashkenasy, G. (2015). Competition and cooperation in dynamic replication networks. *Chemistry - A European Journal*, *21*(2), 648–654. https://doi.org/10.1002/chem.201405195

Damer, B., & Deamer, D. (2015). Coupled phases and combinatorial selection in fluctuating hydrothermal pools: A scenario to guide experimental approaches to the origin of cellular life. *Life*, *5*(1), 872–887. https://doi.org/10.3390/life5010872

Danger, G., Plasson, R., & Pascal, R. (2012). Pathways for the formation and evolution of peptides in prebiotic environments. *Chemical Society Reviews*, *41*(16), 5416–5429. https://doi.org/10.1039/c2cs35064e

Erastova, V., Degiacomi, M. T., Fraser, D. G., & Greenwell, H. C. (2017). Mineral surface chemistry control for origin of prebiotic peptides. *Nature Communications*, *8*(1), 1–9. https://doi.org/10.1038/s41467-017-02248-y.

Eschenmoser, A. (2007). Commentary referring to the statement “the origin of life can be traced back to the origin of kinetic control” and the question "do you agree with this statement; and how would you envisage the prebiotic evolutionary bridge between thermodynamic and kinet. *Origins of Life and Evolution of Biospheres*, *37*(4–5), 309–314. https://doi.org/10.1007/s11084-007-9102-5

Forsythe, J. G., Yu, S., Mamajanov, I., Grover, M. A., Krishnamurthy, R., Fernµndez, F. M., & Hud, N. V. (2015). Ester-Mediated Amide Bond Formation Driven by Wet – Dry Cycles : A Possible Path to Polypeptides on the Prebiotic Earth. *Angewandte*, *10*, 9871–9875. https://doi.org/10.1002/anie.201503792

Frenkel-Pinter, M., Samanta, M., Ashkenasy, G., & Leman, L. J. (2020). Prebiotic Peptides: Molecular Hubs in the Origin of Life. *Chemical Reviews*, *120*(11), 4707–4765. https://doi.org/10.1021/acs.chemrev.9b00664

Fry, I. (2006). The origins of research into the origins of life. *Endeavour*, *30*(1), 24–28. https://doi.org/10.1016/j.endeavour.2005.12.002

Gayon, J., Malaterre, C., Morange, M., Raulin-Cerceau, F., & Tirard, S. (2010). Defining Life: Conference Proceedings. *Origins of Life and Evolution of Biospheres*, *40*(2), 119–120. https://doi.org/10.1007/s11084-010-9189-y

Glavin, D. P., Elsila, J. E., Burton, A. S., Callahan, M. P., Dworkin, J. P., Hilts, R. W., & Herd, C. D. K. (2012). Unusual nonterrestrial l-proteinogenic amino acid excesses in the Tagish Lake meteorite. *Meteoritics and Planetary Science*, *47*(8), 1347–1364. https://doi.org/10.1111/j.1945-5100.2012.01400.x

Glonek, T. (2021). Did Cyclic Metaphosphates Have a Role in the Origin of Life ? In *Origins of Life and Evolution of Biospheres* (Issue 0123456789). Springer Netherlands. https://doi.org/10.1007/s11084-021-09604-5

Gorlero, M., Wieczorek, R., Adamala, K., Giorgi, A., Schininà, M. E., Stano, P., & Luisi, P. L. (2009). Ser-His catalyses the formation of peptides and PNAs. *FEBS Letters*, *583*(1), 153–156. https://doi.org/10.1016/j.febslet.2008.11.052

Haldane, J.B.S. (1929) Origin of Life. The Rationalist Annual, 148, 3-10.

Hao, J., Sverjensky, D. A., & Hazen, R. M. (2019). Redox states of Archean surficial environments: The importance of H2,g instead of O2,g for weathering reactions. *Chemical Geology*, *521*(October 2018), 49–58. https://doi.org/10.1016/j.chemgeo.2019.05.022

Ikehara, K. (2014). [GADV]-Protein World Hypothesis on the Origin of Life. *Origins of Life and Evolution of Biospheres*, *44*(4), 299–302. https://doi.org/10.1007/s11084-014-9383-4

Imai, E. I. I., Honda, H., Hatori, K., & Matsuno, K. (1999). Autocatalytic synthesis of oligoglycine in a simulated submarine hydrothermal system. *Origins of Life and Evolution of the Biosphere*, *29*(3), 249–259. https://doi.org/10.1023/A:1006545711889

Intoy, B. F., & Halley, J. W. (2017). Energetics in a model of prebiotic evolution. *Physical Review E*, *96*(6), 1–15. https://doi.org/10.1103/PhysRevE.96.062402

Jain, A., McPhee, S. A., Wang, T., Nair, M. N., Kroiss, D., Jia, T. Z., & Ulijn, R. V. (2022). Tractable molecular adaptation patterns in a designed complex peptide system. *Chem*, *8*(7), 1894–1905. https://doi.org/10.1016/j.chempr.2022.03.016

Kauffman, S. A. (1986). Autocatalytic sets of proteins. *Journal of Theoretical Biology*, *119*(1), 1–24. https://doi.org/10.1016/S0022-5193(86)80047-9

Kitadai, N., & Maruyama, S. (2018). Origins of building blocks of life: A review. *Geoscience Frontiers*, *9*(4), 1117–1153. https://doi.org/10.1016/j.gsf.2017.07.007

Kitadai, N., Yokoyama, T., & Nakashima, S. (2011). Hydration-dehydration interactions between glycine and anhydrous salts: Implications for a chemical evolution of life. *Geochimica et Cosmochimica Acta*, *75*(21), 6285–6299. https://doi.org/10.1016/j.gca.2011.08.027

Lahav, N., & Chang, S. (1976). The possible role of solid surface area in condensation reactions during chemical evolution: Reevaluation. *Journal of Molecular Evolution*, *8*(4), 357–380. https://doi.org/10.1007/BF01739261

Lambert, J. F. (2008). Adsorption and polymerization of amino acids on mineral surfaces: A review. *Origins of Life and Evolution of Biospheres*, *38*(3), 211–242. https://doi.org/10.1007/s11084-008-9128-3

Lau, J. L., & Dunn, M. K. (2018). Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic and Medicinal Chemistry*, *26*(10), 2700–2707. https://doi.org/10.1016/j.bmc.2017.06.052

Lawrence, L., & Moore, W. J. (1951). Kinetics of the Hydrolysis of Simple Glycine Peptides. *Journal of the American Chemical Society*, *73*(8), 3973–3977. https://doi.org/10.1021/ja01152a123

Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K., & Ghadiri, M. R. (1996). A self-replicating peptide. *Letters to Nature,* 382, 525–528.

Lee, D. H., Severin, K., Yokobayashi, Y., & Ghadiri, M. R. (1997). Emergence of symbiosis in peptide self-replication through a hypercyclic network. *Letters to Nature*, *394*(July), 591–594.

Li, F., Fitz, D., Fraser, D. G., & Rode, B. M. (2010). Catalytic effects of histidine enantiomers and glycine on the formation of dileucine and dimethionine in the salt-induced peptide formation reaction. *Amino Acids*, *38*(1), 287–294. https://doi.org/10.1007/s00726-009-0249-4

Luisi, P. L. (1998). About Various Definitions of Life. *Origins of Life and Evolution of Biospheres*, *28*, 613–622.

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. (2014). Ester formation and hydrolysis during wet-dry cycles: Generation of far-from-equilibrium polymers in a model prebiotic reaction. *Macromolecules*, *47*(4), 1334–1343. https://doi.org/10.1021/ma402256d

Markovitch, O., & Lancet, D. (2012). Excess mutual catalysis is required for effective evolvability. *Artificial Life*, *18*(3), 243–266. https://doi.org/10.1162/artl_a_00064

Martin, W., Baross, J., Kelley, D., & Russell, M. J. (2008). Hydrothermal vents and the origin of life. *Nature Reviews Microbiology*, *6*(11), 805–814. https://doi.org/10.1038/nrmicro1991

Martin, O., & Horvath, J. E. (2013). Biological Evolution of Replicator Systems: Towards a Quantitative Approach. *Origins of Life and Evolution of Biospheres*, *43*(2), 151–160. https://doi.org/10.1007/s11084-013-9327-4

Maury, C. P. J. (2009). Self-Propagating β -Sheet Polypeptide Structures as Prebiotic Informational Molecular Entities : The Amyloid World. *Origins of Life and Evolution of Biospheres*, *39*(2), 141–150. https://doi.org/10.1007/s11084-009-9165-6

Miller, S. L. (1953). A production of amino acids under possible primitive earth conditions. *Science*, *117*(3046), 528–529. https://doi.org/10.1126/science.117.3046.528

Monnard, P. A., & Walde, P. (2015). Current ideas about prebiological compartmentalization. *Life*, *5*(2), 1239–1263. https://doi.org/10.3390/life5021239

Mossel, E., & Steel, M. (2005). Random biochemical networks: The probability of self-sustaining autocatalysis. *Journal of Theoretical Biology*, *233*(3), 327–336. https://doi.org/10.1016/j.jtbi.2004.10.011

Nelson, K. E., Levy, M., & Miller, S. L. (2000). Peptide nucleic acids rather than RNA may have been the first genetic molecule. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(8), 3868–3871. https://doi.org/10.1073/pnas.97.8.3868

Omran, A., & Pasek, M. (2020). A constructive way to think about different hydrothermal environments for the origins of life. *Life*, *10*(4), 1–8. https://doi.org/10.3390/life10040036

Oparin, A. I. (1957). *The origin of life on the earth*. Academic Press (New York, NY, USA)

Pascal, R., Pross, A., & Sutherland, J. D. (2013). Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open Biology*, *3*(NOV), 1–9. https://doi.org/10.1098/rsob.130156

Pasternack, R. F., Gipp, L., & Sigel, H. (1972). Thermodynamics and Kinetics of Complex Formation between Cobalt (II), Nickel (II), and Copper (II) with Glycyl-L-leucine and L-Leucylglycine. *Journal of the American Chemical Society*, *94*(23), 8031–8038. https://doi.org/10.1021/ja00778a017

Patel, B. H., Percivalle, C., Ritson, D. J., Duffy, C. D., & Sutherland, J. D. (2015). Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nature Chemistry*, *7*(4), 301–307. https://doi.org/10.1038/nchem.2202

Plasson, R., Brandenburg, A., Jullien, L., & Bersini, H. (2011). Autocatalysis: At the Root of Self-Replication. *Artificial Life*, *17*(3), 219–236.

Plankensteiner, K., Righi, A., & Rode, B. M. (2002). Glycine and Diglycine as Possible Catalytic Factors in the Prebiotic Evolution of Peptides. *Origins of Life and Evolution of the Biosphere*, *Ii*, 225–236.

Pross, A. (2003). The driving force for life’s emergence: Kinetic and thermodynamic considerations. *Journal of Theoretical Biology*, *220*(3), 393–406. https://doi.org/10.1006/jtbi.2003.3178

Qian, Y., Engel, M. H., Macko, S. A., Carpenter, S., & Deming, J. W. (1993). Kinetics of peptide hydrolysis and amino acid decomposition at high temperature. *Geochimica et Cosmochimica Acta*, *57*(14), 3281–3293. https://doi.org/10.1016/0016-7037(93)90540-D

Rabinowitz, J., Flores, J., Krebsbach, R., & Rogers, G. (1969). Peptide Formation in the Presence of Linear or Cyclic Polyphosphates. *Nature*, *224*(1963), 795–796.

Radzicka, A., & Wolfenden, R. (1996). Rates of uncatalyzed peptide bond hydrolysis in neutral solution and the transition state affinities of proteases. *Journal of the American Chemical Society*, *118*(26), 6105–6109. https://doi.org/10.1021/ja954077c

Rode, B. M. (1999). Peptides and the origin of life. *Peptides*, *20*(6), 773–786. https://doi.org/10.1016/S0196-9781(99)00062-5

Rode, B. M., Fitz, D., & Jakschitz, T. (2007). The First Steps of Chemical Evolution Towards the Origin of Life. *Chemistry & Biodiversity*, *4*, 2674–2702. https://doi.org/10.1002/chin.200813277

Rode, B. M., & Schwendinger, M. G. (1990). Copper-Catalyzed Amino Acid Condensation in Water - A Simple Possible Way of Prebiotic Peptide Formation. *Origins of Life and Evolution of the Biosphere*, *20*(Ii), 401–410.

Rodriguez-Garcia, M., Surman, A. J., Cooper, G. J. T., Suarez-Marina, I., Hosni, Z., Lee, M. P., & Cronin, L. (2015). Formation of oligopeptides in high yield under simple programmable conditions. *Nature Communications*, *6*(8385). https://doi.org/10.1038/ncomms9385

Ross, D. S., & Deamer, D. (2016). Dry/wet cycling and the thermodynamics and kinetics of prebiotic polymer synthesis. *Life*, *6*(3), 1–12. https://doi.org/10.3390/life6030028

Rout, S. K., Friedmann, M. P., Riek, R., & Greenwald, J. (2018). A prebiotic template-directed peptide synthesis based on amyloids. *Nature Communications*, *9*(234). https://doi.org/10.1038/s41467-017-02742-3

Rout, S. K., Rhyner, D., Riek, R., & Greenwald, J. (2022). Prebiotically Plausible Autocatalytic Peptide Amyloids. *Chemistry - A European Journal*, *28*(3). https://doi.org/10.1002/chem.202103841

Rubinov, B., Wagner, N., Rapaport, H., & Ashkenasy, G. (2009). Self-Replicating Amphiphilic β-Sheet Peptides. *Angewandte Chemie*, *121*(36), 6811–6814. https://doi.org/10.1002/ange.200902790

Rufo, C. M., Moroz, Y. S., Moroz, O. V., Stöhr, J., Smith, T. A., Hu, X., Degrado, W. F., & Korendovych, I. V. (2014). Short peptides self-assemble to produce catalytic amyloids. *Nature Chemistry*, *6*(4), 303–309. https://doi.org/10.1038/nchem.1894

Ruiz-Mirazo, K., Briones, C., & De La Escosura, A. (2014). Prebiotic systems chemistry: New perspectives for the origins of life. *Chemical Reviews*, *114*(1), 285–366. https://doi.org/10.1021/cr2004844

Russell, M. J., & Hall, A. J. (1997). The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. *Journal of the Geological Society*, *154*(3), 377–402. https://doi.org/10.1144/gsjgs.154.3.0377

Sakata, K., Kitadai, N., & Yokoyama, T. (2010). Effects of pH and temperature on dimerization rate of glycine: Evaluation of favorable environmental conditions for chemical evolution of life. *Geochimica et Cosmochimica Acta*, *74*(23), 6841–6851. https://doi.org/10.1016/j.gca.2010.08.032

Sawai, H., & Orgel, L. E. (1975). Prebiotic peptide-formation in the solid state - III. Condensation Reactions of Glycine in Solid State Mixtures Containing Inorganic Polyphosphates. *Journal of Molecular Evolution*, *6*(3), 185–197. https://doi.org/10.1007/BF01732355

Schwartz, A. W. (2010). Origins of the RNA world. *The Molecular Origins of Life*, 237–254. https://doi.org/10.1017/cbo9780511626180.013

Schwendinger, M. G., & Rode, B. M. (1989). Possible role of copper and sodium chloride in prebiotic evolution of peptides. *Analytical Sciences*, *5*(4), 411–414. https://doi.org/10.2116/analsci.5.411

Serov, N. Y., Shtyrlin, V. G., & Khayarov, K. R. (2020). The kinetics and mechanisms of reactions in the flow systems glycine–sodium trimetaphosphate–imidazoles: the crucial role of imidazoles in prebiotic peptide syntheses. *Amino Acids*, *52*(5), 811–821. https://doi.org/10.1007/s00726-020-02854-z

Shapiro, R. (2000). A replicator was not involved in the origin of life. *IUBMB Life*, *49*(3), 173–176. https://doi.org/10.1080/152165400306160

Sheehan, J. D., Abraham, A., & Savage, P. E. (2019). Reaction pathways and kinetics for tetra-alanine in hot, compressed liquid water. *Reaction Chemistry and Engineering*, *4*(7), 1237–1252. https://doi.org/10.1039/c9re00023b

Sibilska, I., Feng, Y., Li, L., & Yin, J. (2018). Trimetaphosphate Activates Prebiotic Peptide Synthesis across a Wide Range of Temperature and pH. *Origins of Life and Evolution of Biospheres*, *48*(3), 277–287.

Sun, Y., Frenkel-Pinter, M., Liotta, C. L., & Grover, M. A. (2019). The pH dependent mechanisms of non-enzymatic peptide bond cleavage reactions. *Physical Chemistry Chemical Physics*, *22*(1), 107–113. https://doi.org/10.1039/c9cp05240b

Surman, A. J., Rodriguez-Garcia, M., Abul-Haija, Y. M., Cooper, G. J. T., Gromski, P. S., Turk-MacLeod, R., Mullin, M., Mathis, C., Walker, S. I., & Cronin, L. (2019). Environmental control programs the emergence of distinct functional ensembles from unconstrained chemical reactions. *Proceedings of the National Academy of Sciences*, *116*(12), 5387–5392. https://doi.org/10.1073/pnas.1813987116

Suwannachot, Y., & Rode, B. M. (1999). Mutual Amino Acid catalysis in Salt-Induced Peptide Formation Supports this Mechanism's Role in Prebiotic Peptide Evolution. *Origins of Life and Evolution of the Biosphere*, *29*, 463–471.

Timm, J., Pike, D. H., Mancini, J. A., Tyryshkin, A. M., Poudel, S., Siess, J. A., Molinaro, P. M., McCann, J. J., Waldie, K. M., Koder, R. L., Falkowski, P. G., & Nanda, V. (2023). Design of a minimal di-nickel hydrogenase peptide. *Science Advances*, *9*(10), eabq1990. https://doi.org/10.1126/sciadv.abq1990

Tirard, S., Morange, M., & Lazcano, A. (2010). The definition of life: a brief history of an elusive scientific endeavor. *Astrobiology*, *10*(10), 1003–1009. https://doi.org/10.1089/ast.2010.0535

van der Gulik, P., Massar, S., Gilis, D., Buhrman, H., & Rooman, M. (2009). The first peptides: The evolutionary transition between prebiotic amino acids and early proteins. *Journal of Theoretical Biology*, *261*(4), 531–539. https://doi.org/10.1016/j.jtbi.2009.09.004

Varfolomeev, S. D., & Lushchekina, S. V. (2014). Prebiotic synthesis and selection of macromolecules: Thermal cycling as a condition for synthesis and combinatorial selection. *Geochemistry International*, *52*(13), 1197–1206. https://doi.org/10.1134/S0016702914130102

Virgo, N., & Ikegami, T. (2013). *Autocatalysis Before Enzymes: The Emergence of Prebiotic Chain Reactions*. 240–247. https://doi.org/10.7551/978-0-262-31709-2-ch036

Walker, S. I., Grover, M. A., & Hud, N. V. (2012). Universal sequence replication, reversible polymerization and early functional biopolymers: A model for the initiation of prebiotic sequence evolution. *PLoS ONE*, *7*(4), 31–37. https://doi.org/10.1371/journal.pone.0034166

Wang, M. S., Hoegler, K. J., & Hecht, M. H. (2019). Unevolved de novo proteins have innate tendencies to bind transition metals. *Life*, *9*(1), 1–15. https://doi.org/10.3390/life9010008

Woese C. 1967. The genetic code, pp. 179–195. Harper and Row, New York.

Yamagata, Y., Watanabe, H., Saitoh, M., & Namba, T. (1991). Volcanic production of polyphosphates and its relevance to prebiotic evolution. *Nature*, *352*(6335), 516–519. https://doi.org/10.1038/352516a0

Ying, J., Lin, R., Xu, P., Wu, Y., Liu, Y., & Zhao, Y. (2018). Prebiotic formation of cyclic dipeptides under potentially early Earth conditions. *Scientific Reports*, *8*(1), 1–8. https://doi.org/10.1038/s41598-018-19335-9

Yu, S. S., Krishnamurthy, R., Fernández, F. M., Hud, N. V., Schork, F. J., & Grover, M. A. (2016). Kinetics of prebiotic depsipeptide formation from the ester-amide exchange reaction. *Physical Chemistry Chemical Physics*, *18*(41), 28441–28450. https://doi.org/10.1039/c6cp05527c

Yu, S. S., Solano, M. D., Blanchard, M. K., Soper-Hopper, M. T., Krishnamurthy, R., Fernández, F. M., Hud, N. V., Schork, F. J., & Grover, M. A. (2017). Elongation of Model Prebiotic Proto-Peptides by Continuous Monomer Feeding. *Macromolecules*, *50*(23), 9286–9294. https://doi.org/10.1021/acs.macromol.7b01569

Zaia, D. A. M., Zaia, C. T. B. V., & De Santana, H. (2008). Which amino acids should be used in prebiotic chemistry studies? *Origins of Life and Evolution of Biospheres*, *38*(6), 469–488. https://doi.org/10.1007/s11084-008-9150-5

2. GLYCINE TO OLIGOGLYCINE VIA SEQUENTIAL TRIMETAPHOSPHATE ACTIVATION STEPS IN DRYING ENVIRONMENTS

Hayley Boigenzahn and John Yin

Authors' Contributions

Both authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by **HB**. The first draft of the manuscript was written by **HB** and both authors contributed to subsequent versions of the manuscript.

This chapter has been published.

Boigenzahn, H., & Yin, J. (2022). Glycine to Oligoglycine via Sequential Trimetaphosphate Activation Steps in Drying Environments. *Origins of Life and Evolution of Biospheres*, *52*(4), 249–261. https://doi.org/10.1007/s11084-022-09634-7

2.1 Abstract

Polyphosphate-mediated peptide bond formation is central to protein synthesis in modern organisms, but a simpler form of activation likely preceded the emergence of proteins and RNA. One suggested scenario involves trimetaphosphate (TP), an inorganic phosphate that promotes peptide condensation. Peptide bond formation can also be promoted by high pH and drying, but the interaction of these factors with TP has yet to be characterized kinetically. We studied the formation of glycine oligomers formed under initially alkaline conditions in the presence of TP during the process of drying. Oligopeptide products sampled over 24 hours were analyzed by functionalization and high-performance liquid chromatography with ultraviolet absorption (UV-HPLC). As they dried, two different pH-dependent mechanisms dominated during different stages of the process. The first mechanism occurs in alkaline solutions and activates monomer amino acids to form dimers while reducing the pH. Our results then become consistent with a second mechanism that proceeds at neutral pH and consumes dimers to form longer products. The possibility that a series of reactions might occur where the first reaction changes the environment to favor the second, and so on, may have broader implications for prebiotic polymerization. Studying how the environment changes during time-varying conditions, like drying, could help us understand how organic polymers formed during the origin of life.

2.2 Introduction

Early in the origin of life, short peptides probably performed essential functions analogous to the roles filled by proteins in modern life (Frenkel-Pinter et al. 2020). The wide range of possible functions that peptides can adopt, including catalysis, secondary structure organization, and template-guided polymerization, suggests that they played a significant role in the emergence and development of life (Ruiz-Mirazo et al. 2014). Although the chemistry that led to the origin of life remains a topic of much speculation, amino acids are relatively easy to form through prebiotic routes (Frenkel-Pinter et al. 2020). In contrast, peptide bond formation does not proceed favorably in water (Danger et al. 2012), which poses a key question of how amino acids polymerized into peptides, and how those peptides avoided hydrolysis prior to translation or regulated catalysis.

Various methods for forming peptide bonds in possible prebiotic conditions have been proposed, as reviewed previously (Frenkel-Pinter et al. 2020; Ruiz-Mirazo et al. 2014; Danger et al. 2012). Virtually all methods work through one or both of two mechanisms: creating a dehydrating environment and activating functional groups. Some approaches use the solvation effects of salts or minerals to create a dehydrating environment in the presence of bulk water (Lahav et al. 1978; Rode 1999), whereas others use drying to physically remove water (Ross and Deamer 2016; Campbell et al. 2019). Drying is easily justified as a prebiotic process that could occur naturally due to tidal cycles, day/night cycles, or weather variation. Nonetheless, simply drying amino acids in water produces negligible peptide yields. Rather specific environmental conditions are needed for drying to promote effective polymerization (Lahav et al. 1978; Rodriguez-Garcia et al. 2015; Kitadai and Maruyama 2018; Rode 1999; Napier and Yin 2006). One benefit of drying is that the concentration of non-volatile reactants significantly increases as the solvent evaporates, which can force dilute species to interact with each other and increases the rate of some reactions (Ross and Deamer 2016; Mamajanov et al. 2014). In addition to affecting the yield, this can also allow longer or more diverse peptides to form.

Another way to promote prebiotic peptide bond formation is to add an ‘activating agent’ – a material that interacts with amino acids or peptides to decrease the energy barrier for the condensation reaction to occur (Danger et al. 2012). One such material is trisodium trimetaphosphate (TP), a cyclic triphosphate that is known to promote peptide bond formation (Rabinowitz et al. 1969; Sibilska et al. 2017, 2018; Ying et al. 2018). Polyphosphates are key activators in many modern biological processes, including protein synthesis, which makes them interesting candidates for activating molecules in the origins of life (Lohrmann and Orgel 1973; Pasek et al. 2017). TP has relatively high solubility in water compared to other forms of phosphate (Yamagata et al. 1991), and the ring strain on the O-P-O bonds causes it to be especially reactive (Britvin et al. 2021). TP is considered prebiotically available because a pathway through which it could be formed by volcanic reactions has been proposed (Yamagata et al. 1991), and tetrametaphosphate, a closely related cyclophosphate, has been found in nature (Britvin et al. 2021). TP has been used extensively in studies of prebiotic polymerization due to the relatively high yields of peptides that it supports (Hill and Orgel 2002; Yamagata and Inomata 1997; Sibilska-Kaminski and Yin 2021). Mechanisms for TP-activated peptide bond formation have been published by several groups (Sibilska et al. 2017; Chung et al. 1971; Yamanaka et al. 1988; Inoue et al. 1993).

Although reaction mechanisms have been explored before, an account of how they act during the drying process has not yet been published. To understand how TP-activated peptide bond formation proceeds in a drying environment, we tracked the polymerization of glycine through a 24-hour drying period. We observed the samples going through two distinct phases, each consistent with a different reaction mechanism. We suggest that the shift from one mechanism to another is based on pH change, as protons are produced by the early polymerization steps. Improving our understanding of how dynamic reaction conditions such as drying produce complex molecules can give us insight into how the precursors to biological polymers may have emerged on the early Earth.

2.2 Materials and Methods

2.2.1 Materials

All chemicals were of analytical grade purity and used without further purification. Materials were obtained from suppliers as follows: glycine, diglycine, triglycine, pentaglycine, trisodium trimetaphosphate, and trifloroacetic acid from Sigma-Aldrich, tetraglycine from Bachem, sodium hydroxide from Fisher Scientific, acetone from Alfa Aesar, 9-fluorenylmethoxycarbonyl chloride (FMOC) from Creosalus, acetonitrile from VWR Chemicals, and sodium tetraborate anhydrous from Acros Organics. Reactions were carried out in 1.5 mL low-retention Eppendorf tubes.

2.2.2 Experimental Setup

Unless otherwise specified, all samples contained 0.1 M glycine, 0.1 M TP, and 0.15 M NaOH. All samples had an initial volume of 1 mL and were placed, with their caps open, in a heat block preheated to 90^{o}C. The ratio of TP to amino acids, starting pH (10.5-11), and heating temperature were determined based on what conditions were most favorable to peptide bond formation in Sibilska et al. (2018). Data was collected using at least three independent experimental replicates at each time point.

Prior to analysis, samples heated with open caps were rehydrated with milliQ water to replace what was lost during evaporation, bringing them back to their original volume (1 mL). To determine the amount of water to replace in samples, six samples were weighed to determine the mass of the 1.5 mL tube plus the sample contents. These weights only varied by ±0.01 g. After heating, each sample was individually weighed, and its weight was subtracted from the average initial mass to determine how much water was needed to reach the original volume, assuming a water density of 1 g/mL. Samples were vortexed (Pulsing Vortex Mixer, Fischer Scientific) until there were no longer any visible solids remaining in the sample, usually about

60-90 seconds on maximum speed for fully dried samples. The pH of the samples was measured using an Apera Instruments PH8500-MS Portable pH microelectrode. pH measurements were performed after the sample was replenished and vortexed to ensure there was a large enough sample volume to measure the pH.

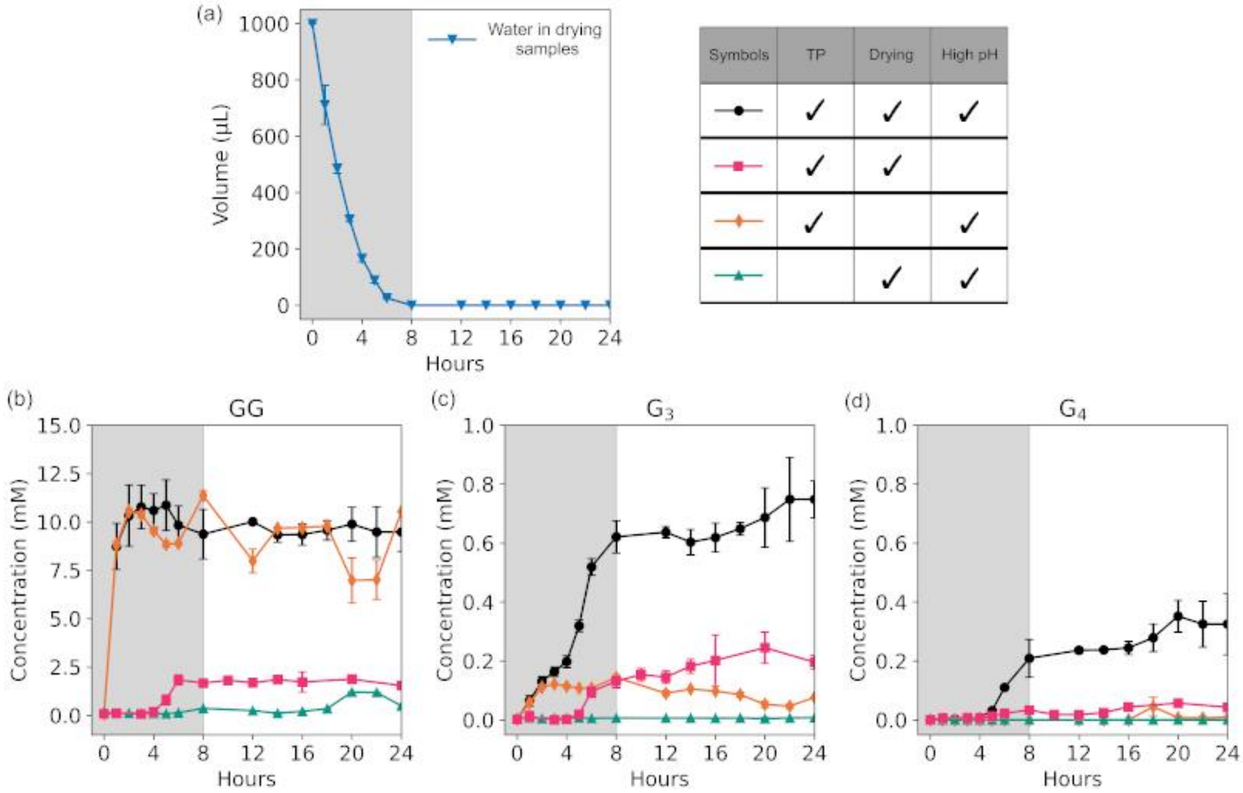
Samples were analyzed using FMOC derivatization and UV-HPLC. FMOC was used to increase the retention time and signal strength of peptide analytes. For the FMOC derivatization procedure, 25 µL of sample was diluted with 75 µL milliQ water to put the large monomer peaks in a quantifiable range. Each sample was then mixed with 100 µL 0.1 M sodium tetraborate buffer for pH control. Finally, 800 µL 3.125 mM FMOC dissolved in acetone was added to each sample. For a sample of 0.1 M amino acid, this results in an equal concentration of FMOC and amino acid, and a slight excess of FMOC in any samples where peptide bond formation had occurred. We were able to recover near-linear calibration curves for all species with this approach (Fig. 2.5), which were used to estimate concentrations from the integrated absorbance values of the HPLC peaks.

Many FMOC procedures suggest performing an extraction procedure to remove excess FMOC-OH (Jámbor and Molnár-Perl 2009), however, we found this was unnecessary as the noise peaks associated with FMOC in the UV-HPLC chromatogram were sharp and did not interfere with any of the peaks associated with our measured species. Samples were allowed to react with FMOC for at least one minute at room temperature, though most reacted longer while queued in the autosampler of the HPLC.

Samples were analyzed with a Shimadzu Nexera HPLC with a C-18 column (Phenomenex Aeris XB-C18, 150 mm x 4.6 mm, 3.6 µL). Products were measured at 254 nm. UV-HPLC analysis was performed using Solvent A: milliQ water with 0.01% v/v trifluoroacetic acid (TFA) and Solvent B: acetonitrile with 0.01% v/v TFA. The following gradient was used: 0-4 min, 30% B, 4-12 min, 30-100% B, 14-15 min, 100-30% B, 15-17 min, 30% B. The solvent flow rate was 1 mL/min. Peak integration was performed using LabSolutions with the ‘Drift’ parameter set to 10000.

2.3 Results

Amino acid condensation is promoted by TP, alkaline conditions (presence of NaOH), and drying (Sibilska et al. 2018). To clarify their roles in activating peptide bond formation, we left out each condition – drying, TP, or NaOH – one by one and measured the resulting concentrations of glycine homopolymers over 24 hours (Fig. 2.1). As expected, the samples that were treated with TP, drying, and high initial pH had the highest peptide yields. The most significant differences were in the yields of trimer (G_3) and tetramer (G_4) glycine polymers – the samples including all three conditions had notably higher yields than the other treatments (Fig. 2.1c, d). In contrast, the dimer (GG) yield of the samples treated with all three conditions was matched by the dimer yield of the samples that contained TP and started at high pH, but were not allowed to dry out (Fig. 2.1b). The similarity of the diglycine yields from samples using TP and high pH, regardless of whether or not they were dried, is explained by the observation that the vast majority of diglycine formed within the first two hours of heating. At that point, most of the bulk water was still present even in the samples being dried, so any reactions taking place had to be able to proceed in water (Fig. 2.1a). A small amount of trimer formation also occurs in the absence of drying. Collectively, these results indicate that almost all dimer (and some trimer) formation in alkaline samples containing TP occurs through a relatively fast reaction which does not require dehydration to proceed.



Symbols	TP	Drying	High pH
●	✓	✓	✓
■	✓	✓	
♦	✓		✓
▲		✓	✓

Figure 2.1 Di-, tri-, and tetraglycine yields depend on different combinations of three treatments known to promote peptide bond formation. (a) Volume remaining in drying samples. Samples that are not dried maintain a constant volume throughout the experiment. (b) Diglycine concentrations, (c) Triglycine concentrations, and (d) Tetraglycine concentrations. The shaded region from 0 to 8 hours highlights the relationship between sample drying and peptide formation. Error bars represent sample standard deviations calculated from independent experimental triplicates.

After four hours of heating, the rate of formation of trimers and tetramers increased in samples that were drying (Fig. 2.1c, d). The simplest explanation for these increases follows from the decreasing volume of water and corresponding shift towards the condensation reaction per Le Chatelier's principle, plus increasing reactant concentrations. However, if trimer and tetramers were forming through the same mechanism as dimers, then the diglycine concentration should also rise due to drying, since there is still a large amount of monomer remaining in all conditions. Instead, the dimer concentration drops as the yields of trimer and tetramer rise, presumably due to conversion into longer polymers and some quantity of 2,5-diketopiperazine

(DKP) (Table 2.1). These results suggest that in the samples being dried, trimers and tetramers were formed through a different mechanism than what formed dimers during the first two hours, and that the reactions that formed the longer peptides mostly proceeded after drying was nearly complete.

It is noteworthy that it also took about four hours for any peptide formation to occur in samples that were dried and contained TP but had no additional sodium hydroxide added, and therefore started with neutral pH conditions (pH 7) (Fig. 2.1c, d). When peptides eventually formed in these conditions, the dimer yield was low, but the yield of trimers relative to the amount of available dimer reactant was high. This suggests that the mechanism driving peptide bond formation in dry, neutral pH conditions favors trimer formation, which offered a possible explanation for the accelerated trimer formation after four hours in the samples treated with TP, drying, and NaOH.

We suspected pH might change over the course of the experiment. In the samples treated with TP, drying, and high pH, we found that the pH dropped dramatically during the first hour then continued to drop roughly linearly for another four hours. Therefore, although the pH is initially alkaline, even samples treated with NaOH have a relatively neutral pH for most of the experiment (Fig. 2.2). At the time when samples including all three conditions begin to promote trimer and tetramer formation, at about four hours, they have a similar pH to the samples that started at neutral conditions. This may suggest that the initial presence or absence of NaOH does not significantly affect the rate of formation of trimers and tetramers for the last 20 hours of the experiment. Instead, the effect of NaOH in promoting total trimer and tetramer formation is likely due to having a higher concentration of diglycine available at four hours, when drying-induced condensation begins.

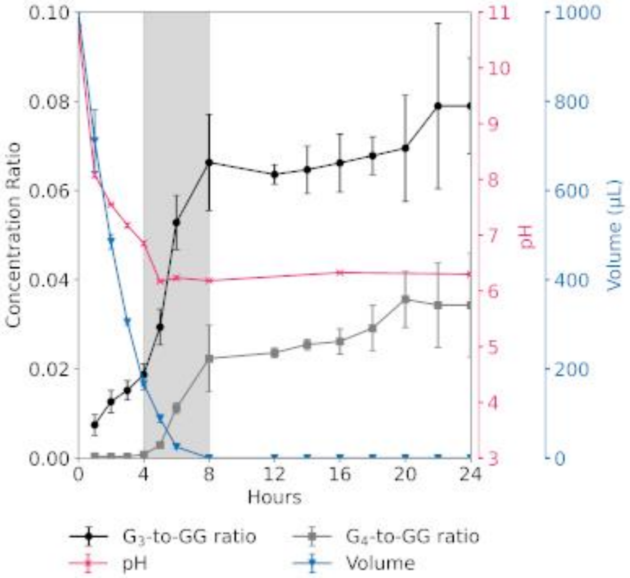


Figure 2.2 The relative concentrations of tri- and tetraglycine increase during drying at neutral pH. Relative concentrations are calculated by dividing by the diglycine concentration at each point. The remaining water volume and pH are shown by the right-hand y-axes. The shaded area highlights the period of rapid tri- and tetra-glycine formation that occurs after 4 hours. Results at 0 hours were excluded due to near-zero numbers producing high variability. Samples were treated with trimetaphosphate, drying, and started at alkaline pH. Error bars represent sample standard deviations calculated from independent experimental triplicates.

2.3.1 Effect of Solid Formation

The highest rates of trimer and tetramer formation coincide with the time when solids begin to form, but peptide formation largely stops once the samples are fully dried. This brief period of increased peptide formation could result from samples having very high reactant concentrations while still having enough solvent to avoid restricting the molecules’ mobility, a limitation that might exist in the fully solid state. We examined the relationship between the solute mass fraction, the formation of solids, and the rates of peptide formation to better understand the effects of drying.

The first consistent appearance of solids occurs at the same time as the rates of longer peptide formation begin to increase. The solids we observed were a translucent but clearly visible separate phase that did not immediately dissolve when the samples were filled back to their original volume, but would eventually dissolve when the samples were subjected to vortex mixing. Samples heated for 4 hours consistently formed solids at the bottom of the tube, despite about 20% of the original water still being present. The highest rates of trimer and tetramer formation occur just afterwards, after 5 and 6 hours of heating and corresponding to solute mass fractions of 0.4 and 0.8, respectively (Fig. 2.3). The solute mass fraction changed rapidly during this time because the sample was mostly dry, but it appeared that longer peptides formed the fastest when the solute mass fraction was neither particularly high nor particularly low. Further drying beyond 6 hours, the last point where there was still a measurable amount of solvent remaining, stops peptide bond formation almost entirely. After 8 hours, the samples were considered fully dried, and the rate of peptide bond was negligible in all the conditions tested. This suggests that further reactions are inhibited while the sample is completely dry.

We conclude that although we did not observe significant peptide bond formation after establishing the dry solid phase, the process of approaching the dry solid phase still has a significant role in promoting trimer and tetramer formation. The different ratios of dimers to trimers and tetramers forming at different times in the experiment appears to be driven by the pH shift, but completely dehydrating the sample is required to drive forward the reactions that form longer peptides.

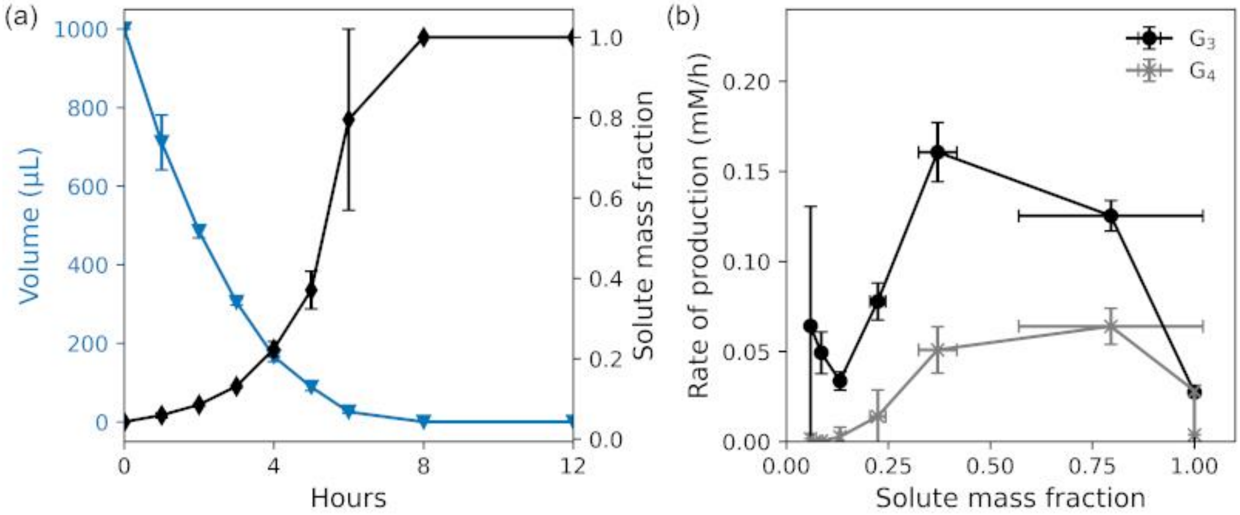


Figure 2.3 The highest rates of trimer and tetramer formation occur at intermediate solute mass fractions. (a) Volume of water remaining in samples and solute mass fraction over time. For simplicity, solute mass is assumed to be constant and equal to the theoretical mass based on concentrations and molar masses. (b) Rates of G_3 and G_4 formation. Rates were estimated using the three-point central difference formula. Error bars represent sample standard deviations calculated from independent experimental triplicates. Details on the calculation of the solute mass fraction and error propagation can be found in the Supplemental Information (Section 2.6.1).

2.4 Discussion

2.4.1 Mechanisms

A key result of our study was the identification of two distinct phases of TP-activated peptide formation that correlated with changes in the pH and hydration conditions of the samples. The two phases we observed correlate well with two different mechanisms of TP-activated peptide formation, both of which were previously described by Yamanaka et al. (1988) (Fig. 2.4). The first mechanism proceeds through activation of the N-terminus, the second proceeds through activation of the O-terminus.

Mechanism 1 likely accounts for the rapid increase in diglycine observed during the first hour in samples containing TP at alkaline conditions. First proposed by Chung et al. (1971), Mechanism 1 is generally accepted for TP-activated peptide elongation in alkaline conditions (Yamanaka et al. 1988; Inoue et al. 1993). This mechanism creates a phosphoryl-carboxyl mixed anhydride, a five-membered ring intermediate. The high reactivity of the mixed anhydride allows this reaction to occur in solution without dehydration. However, this mechanism releases hydronium ions but requires alkaline conditions to proceed, creating a negative feedback loop – as the reaction continues, it increasingly hinders itself.

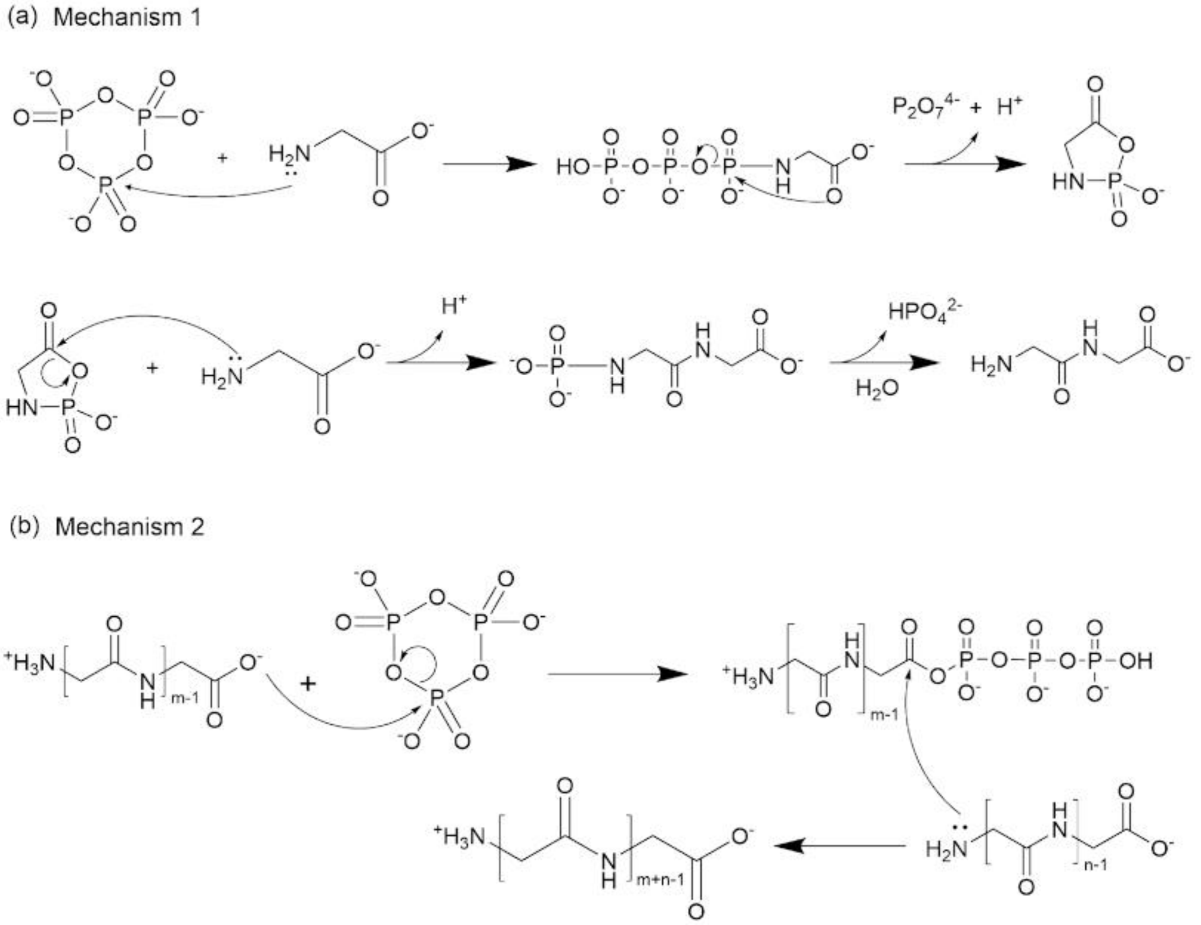


Figure 2.4 Mechanisms for TP-activated peptide bond formation. Adapted from Yamanaka et al. (1988). (a) Mechanism 1 – Dimer formation in alkaline conditions. (b) Mechanism 2 – Bond formation between peptides of arbitrary length at neutral pH.

Mechanism 1 primarily consumes monomers to produce dimers. The mixed anhydride intermediate can only form from amino acids, so at least one reactant must be a monomer. The nucleophile attacking the mixed anhydride can be a longer peptide instead of an amino acid, so it is possible for this mechanism to form peptides longer than dimers, but the excess of monomer here favors dimer formation. The formation of longer products via this mechanism is further limited by the stability of N-phosphorylated diglycine in alkaline conditions (Yamanaka et al. 1988). In N-phosphorylated diglycine, the amine group is blocked by phosphate and unable to act as a nucleophile. If diglycine reacts with TP to become N-phosphorylated instead of attacking a mixed anhydride, then it is essentially excluded from further extension while the sample is at high pH. N-phosphorylated diglycine hydrolyzes back into diglycine at neutral conditions, allowing it to potentially react again (Yamanaka et al. 1988). However, Mechanism 1 does not significantly proceed at a neutral pH because amino acids have protonated amine groups and are unable to perform the nucleophilic attack on TP. Mechanism 2, originally proposed by Yamanaka et al. (1988), proceeds in neutral conditions through an O-phosphorylated peptide that is attacked by the deprotonated amine of another peptide (Fig. 2.4b). This reaction mechanism favors the formation of trimer and tetramer in neutral pH conditions. It requires one nucleophilic attack by an amino acid or peptide with a deprotonated amine group, which at neutral pH is rare. However, it is much more common among peptides than glycine monomers due to significant differences between the basic dissociation constants for the amine groups (pK_b) of glycine and oligoglycine. The pK_b of glycine is 9.60, while the pK_b values of diglycine and triglycine are 8.13 and 7.94, respectively (Yamanaka et al. 1988; Settimo et al. 2014). The pK_b values of diglycine and triglycine are low enough that these species will have non-negligible quantities of both protonated and deprotonated amine groups at pH 7, and the deprotonated species can act as nucleophiles. Glycine is much further below its pK_b at pH 7, so virtually no glycine will be able to act as a nucleophile. Therefore, species already containing a peptide bond are proportionately more likely to participate in Mechanism 2, resulting in increased trimer and tetramer formation.

2,5-Diketopiperazine (DKP), the cyclic anhydrous form of glycine, was also found in previous studies of similar systems (Sibilska et al., 2018). We were not actively tracking it in these experiments because it is not derivatized by FMOC. DKP may interact with some of the intermediates in either mechanism, however, it is expected to be a relatively minor side product. DKP formation requires one of the amino acids to adopt a *cis* configuration, which is not the dominant configuration of linear peptides (Beaufils et al., 2016). Mechanism 1 is an unlikely source of DKP because its formation in water at basic pH conditions is unfavorable (Sakata et al. 2010), and the stability of N-phosphorylated diglycine suggests further reactions are inhibited after reacting with trimetaphosphate (Yamanaka et al. 1988). DKP formation from internal aminolysis does occur at neutral conditions and may play a role in mediating the concentrations of longer peptides (Sun et al. 2019). However, the overall rate is probably limited by the frequency of the correct configuration and correct charge state for aminolysis occurring simultaneously.

2.4.2 Significance of Drying

Although they studied conditions that favored Mechanism 2, Yamanaka et al. (1988) only observed negligible yields of triglycine and no tetraglycine from reactions starting with monomer glycine because their samples were never dried. The change in pH and resulting shift in reaction mechanism explains why longer peptides form favorably later in the experiment, but the results clearly demonstrate the important role of dehydration. Samples that were permitted to dry into a solid had distinctly higher yields of tri- and tetraglycine than those that did not. Mechanism 2 proceeds only to a limited extent in bulk water, which is further supported by samples that started at neutral conditions yielding no detectable peptides until most of the bulk water had evaporated (Fig. 2.1).

Drying increases the rate of peptide condensation by removing water and increasing amino acid and peptide concentrations. However, reaction rates depend on mobility as well as concentration (Ross and Deamer 2016). Molecules in the solid phase have a limited ability to diffuse and rotate, which can slow or stop their reactivity. As samples dry to the solid phase, it would be reasonable to expect reaction rates to increase, then abruptly slow or completely stop due to lack of mobility. In practice, this is not always the case - in some proposed prebiotic reaction conditions, peptide bond formation occurs mostly in the solid state (Napier and Yin 2006; Campbell et al. 2019), and there is some evidence suggesting peptides form slowly after drying in TP-activated samples (Sibilska et al. 2017). However, for the experiments described in this paper, peptide bond formation in the solid phase was negligible.

Condensation into longer peptides likely proceeds best when the system has very low water activity, but has not dried completely. Low water activity shifts the equilibrium towards polymerization and allows longer polymers to form without being hydrolyzed. Once the longer polymers have formed, they do not immediately hydrolyze when rehydrated. Fluctuations between the solid and dissolved states were explored in Campbell et al. (2019) using deliquescent salts, and those systems were found to produce comparable yields of peptides even in the absence of activating agents. We believe our system temporarily reaches similar levels of water activity in the period between 4 and 8 hours, when Mechanism 2 dominates. Understanding these details may be useful for finding systems that produce larger peptides with more potential for complex behavior.

We should also acknowledge the possibility that other physical properties of the dry state may contribute to the increased reaction rates as the sample approaches the solid phase. For example, glycine polymers are known to aggregate into a variety of ordered structures when dried (Yanagawa et al. 1984), and it is possible that some structures align molecules in a manner that promotes peptide bond formation. Subtle mechanistic effects like this are not possible to distinguish with our current methods but we can observe a clear connection between drying and the formation of longer oligopeptides.

2.4.3 Environmental Conditions in Prebiotic Chemistry

An interesting feature in the trimetaphosphate system is that two different mechanisms of peptide bond formation occur at different environmental conditions, and the first mechanism may contribute to creating conditions favorable for the second mechanism. The fact that these reaction mechanisms have been known for many years and there has been limited appreciation for the link between them suggests that it may be worthwhile to pay greater attention to the effects of proposed prebiotic reactions on their environment, in addition to the effects of the environment on the reactions.

Although this study is limited in scope, the idea that dynamic reaction environments can increase yields and allow more complex molecules to form is well established (Ross and Deamer 2016; Damer and Deamer 2015; Varfolomeev and Lushchekina 2014; Walker et al. 2012). Finding a path to more complex peptides would be significant since the conditions tested here may be too limiting to create peptides with more complex interactions. Glycine is the most reactive amino acid, but the longest peptides identified in this experiment were only six amino acids long, with the hexamer being present in such low abundance that it was difficult to consistently measure. This is enough polymerization occurring within 24 hours for the system to be intriguing, especially since there is some evidence that peptides as short as dimers may have catalytic activities (Gorlero et al. 2009; Plankensteiner et al. 2005). However, the nature and length of peptides that may have contributed to the origin of life is still very poorly understood (Van der Gulik et al. 2009; Raggi et al. 2016), and the peptides we observed are still far shorter than what is generally used in engineered systems used to study auto-catalytic peptides (Yao et al. 1998; Rout et al. 2018).

Our experiments use drying, one of the simplest to implement and most common dynamic environmental conditions studied in prebiotic chemistry, to demonstrate mechanistically how such conditions can allow longer peptides to form. Experiments with more diverse reactants and longer sequences of environmental conditions, such as wet-dry cycling and reactant replenishment, may be required to obtain peptides with greater length and complexity. Nevertheless, it seems reasonable to suggest that there may be other combinations of environmental conditions and reaction mechanisms that overlap in ways which facilitate the formation of larger organic molecules and build up reaction networks that occur in series, where the environmental conditions are partially controlled by the organic reactions taking place. There are many parallels between such scenarios and the cycles or cascades of reactions that constitute modern biology. A few examples of how reactions could influence the surrounding environmental conditions include pH changes, the creation of various by-products and intermediates, temperature changes caused by endothermic and exothermic reactions, and phase separation owing to the accumulation of various intermediates. Understanding these relationships would be extremely useful for discovering how chemical systems could develop enough complexity coupled with enough specificity to take on life-like behaviors.

2.5 Conclusion

We investigated TP-activated glycine homopolymer formation in drying conditions and described the results in the context of the known mechanisms for this process. There are two mechanisms for TP-activated peptide formation, which are active in different pH and concentration conditions, and favor different peptide lengths. Alkaline samples of glycine and TP naturally proceed through both mechanisms in sequence as they dry. The first mechanism forms dimers and lowers the pH, which allows the second mechanism to proceed as the sample dries. The second mechanism favors trimer and tetramer formation, further polymerizing the dimers formed during the first reaction. This particular sequence of reactions enables the formation of longer glycine polymers.

Production of longer peptides is significant because it indicates that the system can achieve a higher level of molecular complexity, which may have been useful in the development of early life. The observation that longer peptides can arise from a naturally occurring sequence of reactions suggests the possible importance of dynamic reaction conditions in developing complex molecules. Studying different prebiotic reactions, the environments they occur in, and the effect that they have on the surrounding environment may suggest routes through which longer biopolymers could have developed on the early Earth.

2.6 Supplemental Information

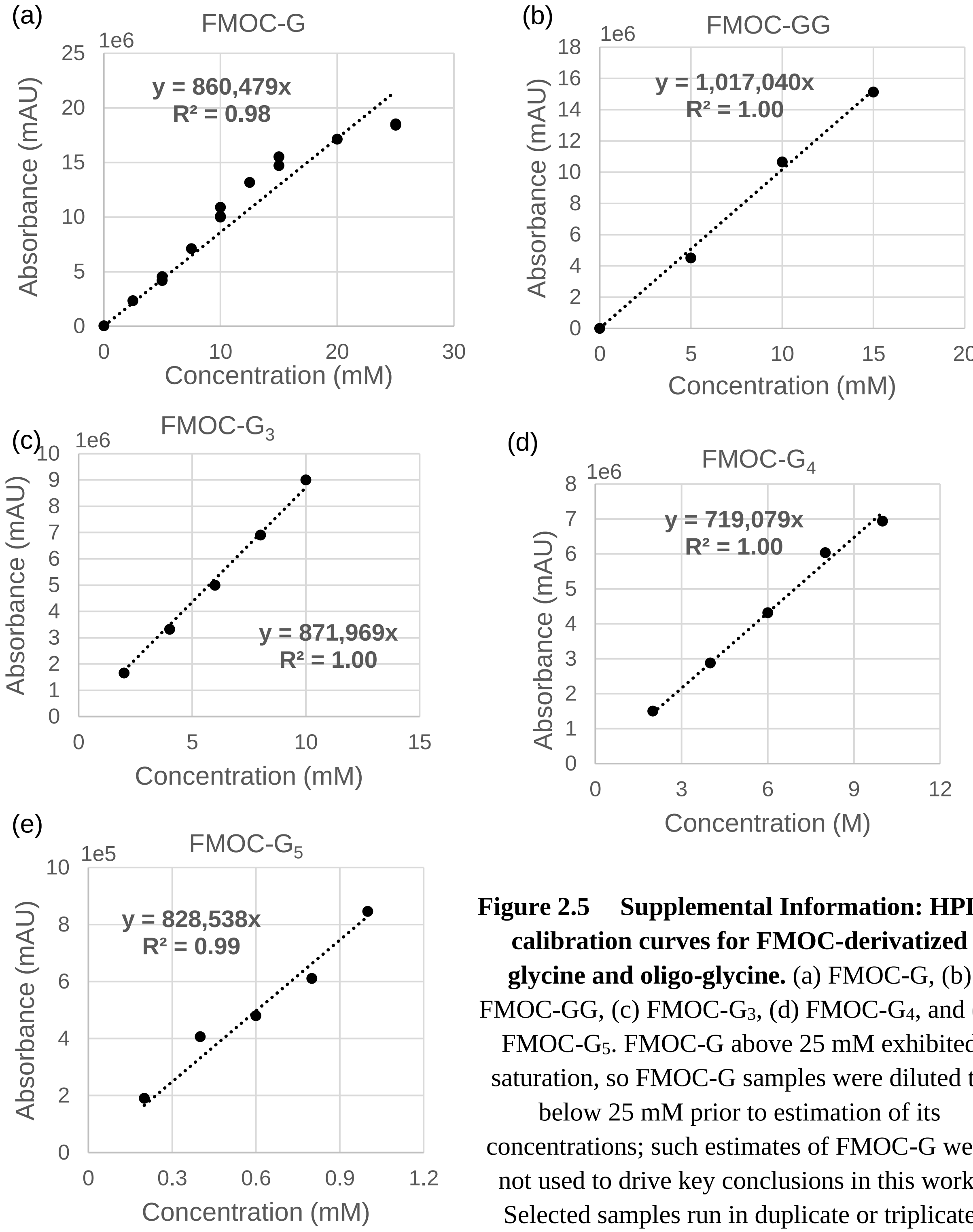


Figure 2.5 Supplemental Information: HPLC calibration curves for FMOC-derivatized glycine and oligo-glycine. (a) FMOC-G, (b) FMOC-GG, (c) FMOC-G$_3$, (d) FMOC-G$_4$, and (e) FMOC-G$_5$. FMOC-G above 25 mM exhibited saturation, so FMOC-G samples were diluted to below 25 mM prior to estimation of its concentrations; such estimates of FMOC-G were not used to drive key conclusions in this work. Selected samples run in duplicate or triplicate showed good reproducibility.

	[GG] (mM)	**[G_3] (mM)**	**[G_4] (mM)**
4 hours	10.6 ± 0.88	0.20 ± 0.021	0.0084 ± 0.0026
8 hours	9.37 ± 1.28	0.62 ± 0.055	0.21 ± 0.064
	Δ[GG] (mM)	**Δ[G_3] (mM)**	**Δ[G_4] (mM)**
4 hrs to 8 hrs	-1.23 ± 1.55	0.42 ± 0.059	0.20 ± 0.064
% of [GG] drop accounted for by rising [G_3] and [G_4] $\frac{\Delta[G_3] + 2 * \Delta[G_4]}{-\Delta[G_2]} * 100\%$			66 ± 85%

Table 2.1 Diglycine consumption to form tri- and tetraglycine.

The level of G_2 drops during 4-to-8 hours of heating, potentially feeding production of G_3 and G_4, which increase during this same period. To test this possibility, we assumed formation of each G_3 and G_4 consumes one and two G_2, respectively. However, standard deviations from triplicate measures of these species were too large for a statistically significant accounting of changes. Moreover, other reactions of G_2 may be involved: two G_2 may cyclize to form 2,5-diketopiperazine (DKP), G_2 may be elongated to form still longer oligomers, or G_2 may hydrolyze back to $2G_1$. DKP lacks an amine, so it is not detected by FMOC derivatization. Rates of hydrolysis would likely be negligible after 4 hours, since by that time the sample was mostly dry.

2.6.1 Supplemental Information: Calculations for the solute mass fraction, rate of tri- and tetramer production, and error propagation shown in Figure 2.3.

The mass of the solute (m_{solute}) was assumed to be exactly 0.044 g in all samples, which is the theoretical solute mass for 1 mL of 0.1 M glycine, 0.1 M TP, and 0.15 M NaOH. The mass of the remaining solvent was calculated based on the total weight of the initial samples minus the theoretical solute mass and the volume of water replaced at each time point ($m_{solvent}$). The density of water was assumed to be 1 g/mL.

The solute mass fraction is calculated as:

$$Solute\ mass\ fraction\ (mf) = \frac{m_{solute}}{m_{solute} + m_{solvent}} \tag{2.1}$$

The error of the solute mass fraction (σ_{mf}) was calculated using the standard deviations from the volume measurements ($\sigma_{solvent}$). For simplicity, the error of the solute mass was assumed to be zero.

$$\sigma_{mf} = \left|\frac{\sigma_{solvent}}{m_{solvent}}\right| * |mf| \tag{2.2}$$

The production rates of triglycine and tetraglycine were calculated using a three-point central difference formula. For concentration measurements (y_n, t_n) taken at time point *n* and with a standard deviation of σ_n, the derivative was approximated as:

$$\left(\frac{dy}{dt}\right)_n \cong \frac{\frac{y_n - y_{n-1}}{t_n - t_{n-1}} + \frac{y_{n+1} - y_n}{t_{n+1} - t_n}}{2} \tag{2.3}$$

For the first and the last point, when there weren’t three consecutive points to calculate from, two-point difference formulas were used instead. The error of the production rates was calculated using the measured standard deviations of the concentrations.

$$\sigma_{\left(\frac{dy}{dt}\right)_n} = \frac{\sqrt{\left(\frac{\sigma_n}{y_n}\right)^2 + \left(\frac{\sigma_{n-1}}{y_{n-1}}\right)^2}}{2(t_n - t_{n-1})} + \frac{\sqrt{\left(\frac{\sigma_{n+1}}{y_{n+1}}\right)^2 + \left(\frac{\sigma_n}{y_n}\right)^2}}{2(t_{n+1} - t_n)} * \left|\left(\frac{dy}{dt}\right)_n\right| \qquad (2.4)$$

2.7 References

Britvin, S. N., Murashko, M. N., Vapnik, Y., Vlasenko, N. S., Krzhizhanovskaya, M. G., Vereshchagin, O. S., Bocharov, V. N., & Lozhkin, M. S. (2021). Cyclophosphates, a new class of native phosphorus compounds, and some insights into prebiotic phosphorylation on early Earth. *Geology*, *49*(4), 382–386. https://doi.org/10.1130/G48203.1

Campbell, T. D., Febrian, R., McCarthy, J. T., Kleinschmidt, H. E., Forsythe, J. G., & Bracher, P. J. (2019). Prebiotic condensation through wet–dry cycling regulated by deliquescence. *Nature Communications*, 10(1). https://doi.org/10.1038/s41467-019-11834-1

Chung, N. M., Lohrmann, R., Orgel, L. E., & Rabinowitz, J. (1971). The mechanism of the trimetaphosphate-induced peptide synthesis. *Tetrahedron*, *27*(6), 1205–1210. https://doi.org/10.1016/S0040-4020(01)90868-3

Damer, B., & Deamer, D. (2015). Coupled phases and combinatorial selection in fluctuating hydrothermal pools: A scenario to guide experimental approaches to the origin of cellular life. *Life*, *5*(1), 872–887. https://doi.org/10.3390/life5010872

Danger, G., Plasson, R., & Pascal, R. (2012). Pathways for the formation and evolution of peptides in prebiotic environments. *Chemical Society Reviews*, *41*(16), 5416–5429. https://doi.org/10.1039/c2cs35064e

Forsythe, J. G., Petrov, A. S., Millar, W. C., Yu, S.-S., Krishnamurthy, R., Grover, M. A., Hud, N. V., & Fernández, F. M. (2017). Surveying the sequence diversity of model prebiotic peptides by mass spectrometry. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1711631114

Frenkel-Pinter, M., Samanta, M., Ashkenasy, G., & Leman, L. J. (2020). Prebiotic Peptides: Molecular Hubs in the Origin of Life. *Chemical Reviews*, *120*(11), 4707–4765. https://doi.org/10.1021/acs.chemrev.9b00664

Gorlero, M., Wieczorek, R., Adamala, K., Giorgi, A., Schininà, M. E., Stano, P., & Luisi, P. L. (2009). Ser-His catalyses the formation of peptides and PNAs. *FEBS Letters*, *583*(1), 153–156. https://doi.org/10.1016/j.febslet.2008.11.052

Hill, A., & Orgel, L. E. (2002). Trimetaphosphate-Induced Addition of Aspartic Acid to Oligo(glutamic acid)s. *Helvetica Chimica Acta*, *85*, 4244–4254. https://doi.org/10.1002/hlca.200290009

Inoue, H., Baba, Y., Furukawa, T., Maeda, Y., & Tsuhako, M. (1993). Formation of dipeptide in the reaction of amino acids with cyclo-triphosphate. *Chemical and pharmaceutical bulletin*, *41*(11), 1895-1899. https://doi.org/10.1248/cpb.41.1895

Jámbor, A., & Molnár-Perl, I. (2009). Amino acid analysis by high-performance liquid chromatography after derivatization with 9-fluorenylmethyloxycarbonyl chloride. *Journal of Chromatography A*, *1216*(15), 3064–3077. https://doi.org/10.1016/j.chroma.2009.01.068

Kitadai, N., & Maruyama, S. (2018). Origins of building blocks of life: A review. *Geoscience Frontiers*, *9*(4), 1117–1153. https://doi.org/10.1016/j.gsf.2017.07.007

Lahav, N., White, D., & Chang, S. (1978). Peptide formation in the prebiotic era: Thermal condensation of glycine in fluctuating clay environments. *Science, 201*(4350), 67–69. https://doi.org/10.1126/science.663639

Lohrmann, R., & Orgel, L. E. (1973). Prebiotic activation processes. *Nature*, *244*(5416), 418–420. https://doi.org/10.1038/244418a0

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. (2014). Ester formation and hydrolysis during wet-dry cycles: Generation of far-from-equilibrium polymers in a model prebiotic reaction. *Macromolecules, 47*(4), 1334–1343. https://doi.org/10.1021/ma402256d

Napier, J., & Yin, J. (2006). Formation of peptides in the dry state. *Peptides*, *27*(4), 607–610. https://doi.org/10.1016/j.peptides.2005.07.015

Pasek, M. A., Gull, M., & Herschy, B. (2017). Phosphorylation on the early earth. *Chemical Geology*, *475*, 149-170. https://doi.org/10.1016/j.chemgeo.2017.11.008

Plankensteiner, K., Reiner, H. & Rode, B.M. Catalytically Increased Prebiotic Peptide Formation: Ditryptophan, Dilysine, and Diserine. *Orig Life Evol Biosph* 35, 411–419 (2005). https://doi.org/10.1007/s11084-005-1971-x

Rabinowitz, J., Flores, J., Krebsbach, R. & Rogers, G. (1969). Peptide Formation in the Presence of Linear or Cyclic Polyphosphates. *Nature* 224, 795–796. https://doi.org/10.1038/224795a0

Raggi, L., Bada, J. L., & Lazcano, A. (2016). On the lack of evolutionary continuity between prebiotic peptides and extant enzymes. *Physical Chemistry Chemical Physics*, *18*(30), 20028–20032. https://doi.org/10.1039/c6cp00793g

Rodriguez-Garcia, M., Surman, A. J., Cooper, G. J. T., Suarez-Marina, I., Hosni, Z., Lee, M. P., & Cronin, L. (2015). Formation of oligopeptides in high yield under simple programmable conditions. *Nature Communications*, *6*(8385). https://doi.org/10.1038/ncomms9385

Rode, B. M. (1999). Peptides and the origin of life. *Peptides*, *20*(6), 773–786. https://doi.org/10.1016/S0196-9781(99)00062-5

Ross, D. S., & Deamer, D. (2016). Dry/wet cycling and the thermodynamics and kinetics of prebiotic polymer synthesis. *Life*, *6*(3), 1–12. https://doi.org/10.3390/life6030028

Rout, S. K., Friedmann, M. P., Riek, R., & Greenwald, J. (2018). A prebiotic template-directed peptide synthesis based on amyloids. *Nature Communications*, *9*(234). https://doi.org/10.1038/s41467-017-02742-3

Ruiz-Mirazo, K., Briones, C., & De La Escosura, A. (2014). Prebiotic systems chemistry: New perspectives for the origins of life. *Chemical Reviews*, *114*(1), 285–366. https://doi.org/10.1021/cr2004844

Sakata, K., Kitadai, N., & Yokoyama, T. (2010). Effects of pH and temperature on dimerization rate of glycine: Evaluation of favorable environmental conditions for chemical evolution of life. *Geochimica et Cosmochimica Acta*, *74*(23), 6841–6851. https://doi.org/10.1016/j.gca.2010.08.032Settimo, L., Bellman, K., & Knegtel, R. M. A. (2014). Comparison of the accuracy of experimental and predicted pKa values of basic and acidic compounds. *Pharmaceutical Research*, *31*(4), 1082–1095. https://doi.org/10.1007/s11095-013-1232-z

Sibilska, I., Chen, B., Li, L., & Yin, J. (2017). Effects of Trimetaphosphate on Abiotic Formation and Hydrolysis of Peptides. *Life*, *7*(4), 1–11. https://doi.org/10.3390/life7040050

Sibilska, I., Feng, Y., Li, L. et al. Trimetaphosphate Activates Prebiotic Peptide Synthesis across a Wide Range of Temperature and pH. *Orig Life Evol Biosph* 48, 277–287 (2018). https://doi.org/10.1007/s11084-018-9564-7

Sibilska-Kaminski, I.K., Yin, J. Toward Molecular Cooperation by De Novo Peptides. *Orig Life Evol Biosph* 51, 71–82 (2021). https://doi.org/10.1007/s11084-021-09603-6

Sun, Y., Frenkel-Pinter, M., Liotta, C. L., & Grover, M. A. (2019). The pH dependent mechanisms of non-enzymatic peptide bond cleavage reactions. *Physical Chemistry Chemical Physics*, *22*(1), 107–113. https://doi.org/10.1039/c9cp05240b

Surman, A. J., Rodriguez-Garcia, M., Abul-Haija, Y. M., Cooper, G. J. T., Gromski, P. S., Turk-MacLeod, R., Mullin, M., Mathis, C., Walker, S. I., & Cronin, L. (2019). Environmental control programs the emergence of distinct functional ensembles from unconstrained chemical reactions. *Proceedings of the National Academy of Sciences*, *116*(12), 5387–5392. https://doi.org/10.1073/pnas.1813987116

van der Gulik, P., Massar, S., Gilis, D., Buhrman, H., & Rooman, M. (2009). The first peptides: The evolutionary transition between prebiotic amino acids and early proteins. *Journal of Theoretical Biology*, *261*(4), 531–539. https://doi.org/10.1016/j.jtbi.2009.09.004

Varfolomeev, S. D., & Lushchekina, S. V. (2014). Prebiotic synthesis and selection of macromolecules: Thermal cycling as a condition for synthesis and combinatorial selection. *Geochemistry International*, *52*(13), 1197–1206. https://doi.org/10.1134/S0016702914130102

Walker, S. I., Grover, M. A., & Hud, N. V. (2012). Universal sequence replication, reversible polymerization and early functional biopolymers: A model for the initiation of prebiotic sequence evolution. *PLoS ONE*, *7*(4), 31–37. https://doi.org/10.1371/journal.pone.0034166

Yamagata, Y., & Inomata, K. (1997). Condensation of glycylglycine to oligoglycines with trimetaphosphate in aqueous solution. II: Catalytic effect of magnesium ion. *Orig Life Evol Biosph*, *27*(4), 339–344. https://doi.org/10.1023/A:1006529421813

Yamagata, Y., Watanabe, H., Saitoh, M., & Namba, T. (1991). Volcanic production of polyphosphates and its relevance to prebiotic evolution. *Nature*, *352*(6335), 516–519. https://doi.org/10.1038/352516a0

Yamanaka, J., Inomata, K. & Yamagata, Y. Condensation of oligoglycines with trimeta- and tetrametaphosphate in aqueous solutions. *Orig Life Evol Biosph* 18, 165–178 (1988). https://doi.org/10.1007/BF01804669

Yanagawa, H., Nishizawa, M. & Kojima, K. A possible prebiotic peptide formation from glycinamide and related compounds. *Orig Life Evol Biosph* 14, 267–272 (1984). https://doi.org/10.1007/BF00933667

Yao, S., Ghosh, I., Zutshi, R., & Chmielewski, J. (1998). Selective amplification by auto- and cross-catalysis in a replicating peptide system. *Nature*, *396*(6710), 447–450. https://doi.org/10.1038/24814

Ying, J., Lin, R., Xu, P., Wu, Y., Liu, Y., & Zhao, Y. (2018). Prebiotic formation of cyclic dipeptides under potentially early Earth conditions. *Scientific Reports*, *8*(1), 1–8. https://doi.org/10.1038/s41598-018-19335-9

3. KINETIC MODELING AND PARAMETER ESTIMATION OF A PREBIOTIC PEPTIDE REACTION NETWORK

Hayley Boigenzahn, Leonardo D. González, Jaron C. Thompson, Victor M. Zavala, John Yin

Authors’ Contributions

All experimental data was collected and analyzed by **HB**. The main parameter estimation code was developed by JCT, with minor edits and applications performed by **HB** and LDG. The manuscript was drafted by **HB** and LDG and all authors contributed to subsequent editing.

This chapter has been accepted for publication to the Journal of Molecular Evolution.

3.1 Abstract

Although our understanding of how life emerged on Earth from simple organic precursors is speculative, early precursors likely included amino acids. The polymerization of amino acids into peptides and interactions between peptides are of interest because peptides and proteins participate in complex interaction networks in extant biology. However, peptide reaction networks can be challenging to study because of the potential for multiple species and systems-level interactions between species. We developed and employed a computational network model to describe reactions between amino acids to form di-, tri-, and tetra-peptides. Our experiments were initiated with two of the simplest amino acids, glycine and alanine, mediated by trimetaphosphate-activation and drying to promote peptide bond formation. The parameter estimates for bond formation and hydrolysis reactions in the system were found to be poorly constrained due to a network property known as sloppiness. In a sloppy system, the model behavior mostly depends on only a subset of parameter combinations, but there is no straightforward way to determine which parameters should be included or excluded. Despite our inability to determine the exact values of specific kinetic parameters, we could make reasonably accurate predictions of model behavior. In short, our modeling has highlighted challenges and opportunities toward understanding the behaviors of complex prebiotic chemical experiments.

3.2 Introduction

The emergence of life on the early Earth is believed to have been preceded by the accumulation of an increasingly diverse and complex set of organic molecules (Orgel 2010). The reaction networks developed by these molecules laid the groundwork of functions critical for life, like energy and information processing. Understanding how the systems-level molecular interactions required for life-like behavior could emerge from simple precursors remains one of the key questions of prebiotic chemistry, but since this question is primarily about collective behaviors, complexity presents an ongoing challenge (Schwartz 2007; Johnson & Hung 2019). Although studying a single type of molecule or reaction to establish its properties can be useful, it limits what conclusions can be drawn about potential broader community behavior. Experiments involving a greater variety of molecules and reactions can probe more interesting interactions, but have a large search space of variables, and the complexity of the systems make them inherently more difficult to analyze.

Models are useful for understanding complex systems because they can reveal the systematic dependence of various properties on each other and allow us to describe and make predictions about the system behavior. Computational models have been used to explore hypothetical prebiotic chemical networks for many years and have produced many interesting insights (Covney et al. 2012). However, our current interest is in models that are based on experimental data. Prior experimental works mainly used basic kinetic and thermodynamic governing equations to describe individual reactions or small networks involving fewer than five reactions. For example, Arrhenius expressions have been used to determine the free energies of activation for reactions in a small network (Sakata et al. 2010; Yu et al. 2016; Lee et al. 1996). More abstractly, parameters have been fit to empirical rate equations to describe specific elements of system behavior or distinguish between candidate models (von Kiedrowski 1986; Rout et al. 2022). These methods work well for small systems, but may not apply to larger systems with multiple reactions occurring simultaneously and potentially more intricate network interactions. Serov et al. (2020) approximated the parameters for multiple reactions simultaneously in a peptide reaction network, but the parameter fitting was performed manually, and the network was small. Manual approaches are less rigorous than using a computational strategy and can be difficult to implement for even moderately sized networks. Other types of results from complex experiments have been generated using statistical methods, but these do not capture the system dynamics (Surman et al. 2019; Jain et al. 2022). There is a need for approaches to study the dynamical behavior of more complex experimental networks (Ruiz-Mirazo et al. 2014).

Complex network models are broadly applicable and have already been developed extensively for other fields (Newman 2003). One notable example is in systems biology, which has significant parallels to the origins of life. Both involve large interaction networks with potentially limited available data and may include community interactions that are critical to understanding system behavior. Bioinformatics models can be used to analyze experimental data and help understand the molecular interaction networks within living cells (Gauthier et al. 2019). Similar approaches could be useful for furthering experimental chemical origins of life research, but aside from a few reviews and computational investigations, they have generally been overlooked (Johnson & Hung 2019; Ludlow & Otto 2008; Goldman et al. 2013).

Our goal in this study was to investigate how dynamical models, described by ordinary differential equations (ODEs), might be useful for studying origins of life chemistry. These models are theoretically generalizable, but as with all modeling approaches, there are limitations that make them more difficult to apply in some situations. Presenting the benefits and limitations of a model approach in a way that is accessible to experimentalists, which we aim to do, is an important step for linking theory and experiment. Differential equation models are not always suitable for large systems, since constructing them can become difficult, but in well-defined systems they can be used to study detailed mechanistic behavior (Maria 2004). Computational methods can be used to estimate all the parameters efficiently and simultaneously in a moderately complex dynamical network, but validating the physical meaning of the results can be more challenging since these problems may not have a unique and stable solution (Transtrum et al. 2015). However, parameter fitting has still been used to describe nonlinear networks in a variety of fields, including in systems biology for biochemical pathways (Raue et al. 2013; Rodriguez-Fernandez et al. 2006).

We focus specifically on fitting parameters to a set of nonlinear ODEs describing the kinetics of short peptide formation. Peptides are interesting candidates for emergent behavior because they can engage in a variety of intermolecular interactions and their development was likely an important step during the origin of life (Frenkel-Pinter et al. 2020). We studied a simplified network describing peptide formation in a system starting with only two amino acid species, glycine and alanine. By limiting ourselves to two amino acids, we were able to obtain quantitative data on the concentrations of most peptide species as they formed through a possible prebiotic reaction mechanism involving an inorganic phosphate activating agent, trimetaphosphate (TP) (Sibilska et al. 2018).

We found that our model exhibited "sloppiness", a term originally used by the Sethna lab to describe models based on a set of highly imprecise parameters that still return reasonably accurate predictions (Gutenkunst et al. 2007a). Such models are significantly more sensitive to changes in certain parameter values while remaining largely unaffected by changes in others (Waterfall et al. 2006). We suspect sloppiness may be a common feature in networks relevant to the chemical origin of life. It is known to be extremely common in systems biology, and many of the features that contribute to it, like reversibility of reactions and limited experimental observations, are also common features of prebiotic chemistry networks (White et al. 2016).

Sloppiness occurs when parts of the parameter fitting problem are poorly constrained, resulting in highly imprecise parameter estimates. Our computational study reveals that the peptide network model is sloppy. Due to their high uncertainty, parameters fitted to a sloppy model cannot be treated as true kinetic reaction rates, limiting the hypotheses a sloppy model can be used to evaluate (Gutenkunst et al. 2007a). However, the collective model behavior predicted by fitting a sloppy system can be accurate even when fit to relatively sparse experimental data. This makes them useful for tasks such as exploring theoretical long-term model behavior and model falsification (Brown et al. 2004; Gutenkunst et al. 2007b; Hettling & van Beek 2011). For these reasons, we concluded that this system was worth investigating and disseminating despite the high variability observed in the parameter estimates.

We attempted to reduce sloppiness using model reduction and statistical design of experiments, but without improvements. As such, it is important to recognize the inherent limitations of the model structure and of the experimental setup and we conclude that fitting accurate kinetic parameters using the model we present might be difficult. However, ODE models can still be useful tools for characterizing the behavior and stability of prebiotic chemical reaction networks.

3.3 Methods

We studied the formation of peptides from amino acids using trimetaphosphate (TP) as an activating agent. For simplicity, our experiments only included two amino acids: glycine and alanine. To maximize peptide bond formation within 24 hours, samples at alkaline pH were allowed to dry completely (Sibilska et al. 2018). Various combinations of initial concentrations of glycine and alanine were used to increase the amount of relevant data for parameter fitting, and cover a larger range of potential conditions in the network, since concentrations of each species should not affect the values of the kinetic constants. The concentrations of each peptide product were determined using HPLC (see Experimental Methods for details). Each experimental data point is the average of three experimental replicates.

Parameters were fit to an ODE model describing peptide formation and decomposition in a mass-action style network, depicted in Fig. 3.1. The complete time-dependent ODEs for the system are provided in the Supplemental Information (Section 3.8.1). To keep the network a manageable size, we omitted many mechanistic details of peptide formation and only includes canonical peptides, not intermediates or possible side products. For example, no phosphate salts or intermediate products of TP activation were quantifiable in our analysis, so TP was not explicitly included anywhere in the network. To minimize any effect the concentration of TP might have on the kinetics studied, we used a constant ratio of TP to amino acids across all experiments. Isomers such as GGA, GAG, and AGG were grouped together to further reduce the number of parameters and avoid the need to resolve isomers, which tend to co-elute during HPLC analysis. A complete list of fitted parameters, organized by figure, are available on Github at https://github.com/haboigenzahn/OoL-KineticParameterEstimation.

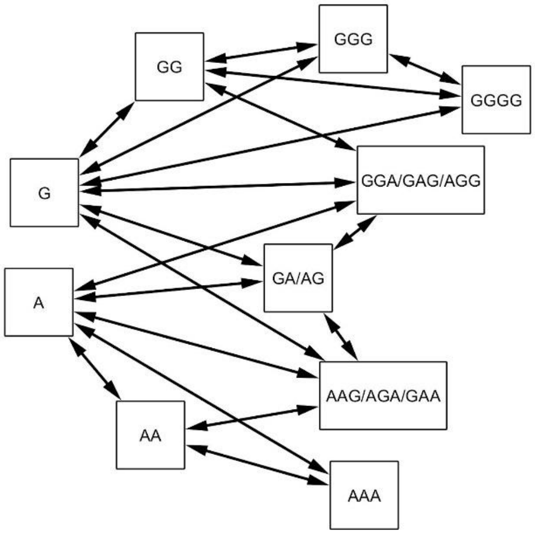


Figure 3.1 Peptide network. Double-headed arrows represent a reversible reaction connecting two species. Note that many edges share the same reaction parameter, such as the G→GGG and GG→GGG edges representing the reaction G+GG→GGG.

We expected that the network would provide a good baseline for understanding which reactions were occurring at higher rates. To improve the precision of the parameter estimates, we applied model reduction and statistical experimental design. Details about these approaches can be found in the Computational Methods section. Here we will describe the results of these tests and assess the feasibility of obtaining a predictive model and accurate parameter estimates for this system from experimental data.

3.4 Results & Discussion

3.4.1 Parameter Estimation

Parameter fitting is performed by tuning the model parameters to minimize a *cost function* ($\mathcal{L}$) that calculates the difference between the model predictions and experimental data; $\mathcal{L}$ is also called a loss function or a residual. We minimized $\mathcal{L}$ using the L-BFGS-B algorithm from Scipy's minimize function (Virtanen et al. 2020). We were also able to approximate the parameter uncertainties, which represent how well the parameters are constrained by experimental data using an asymptotic Gaussian approximation (Vanlier et al. 2013). Parameters determined using sparse or noisy experimental data are less precise than parameters fit with abundant, high precision data, but the structure of the model itself can also significantly contribute to the parameter uncertainty. Validating that the model can theoretically be solved can save time and experimental effort.

We first estimated the parameters for simulated data in the absence of noise, and we were able to accurately recover the parameters used to generate the data (Fig. 2a). When we applied the model to experimental data, it was able to capture general trends, however, the parameter uncertainties were undesirably high (Fig. 2b). For some species, the 95% confidence envelope for the model prediction was larger than the peptide concentrations themselves. Since the optimization can find a local minimum, we repeated the parameter estimation for several different initial guesses. Although the number of initial guesses was limited by the fact that the parameter estimation method takes up to a full day to finish when all of the experimental data is included, we observed that none of the different initial guesses significantly improved the precision of the parameter estimates and that there did not appear to be any positive correlation between the MSE and the number of highly uncertain parameters (Supplemental Information 2). Trying many initial guesses to find the lowest possible value for the cost function may slightly improve the model predictions, but it does not seem to cause an improvement in the precision of the parameter estimates. Despite the extremely high parameter uncertainties, the accuracy of the model predictions initially seemed promising, so we began to explore the parameter fitting process in more detail to determine how to decrease the parameter uncertainty, starting with the identifiability of the network.

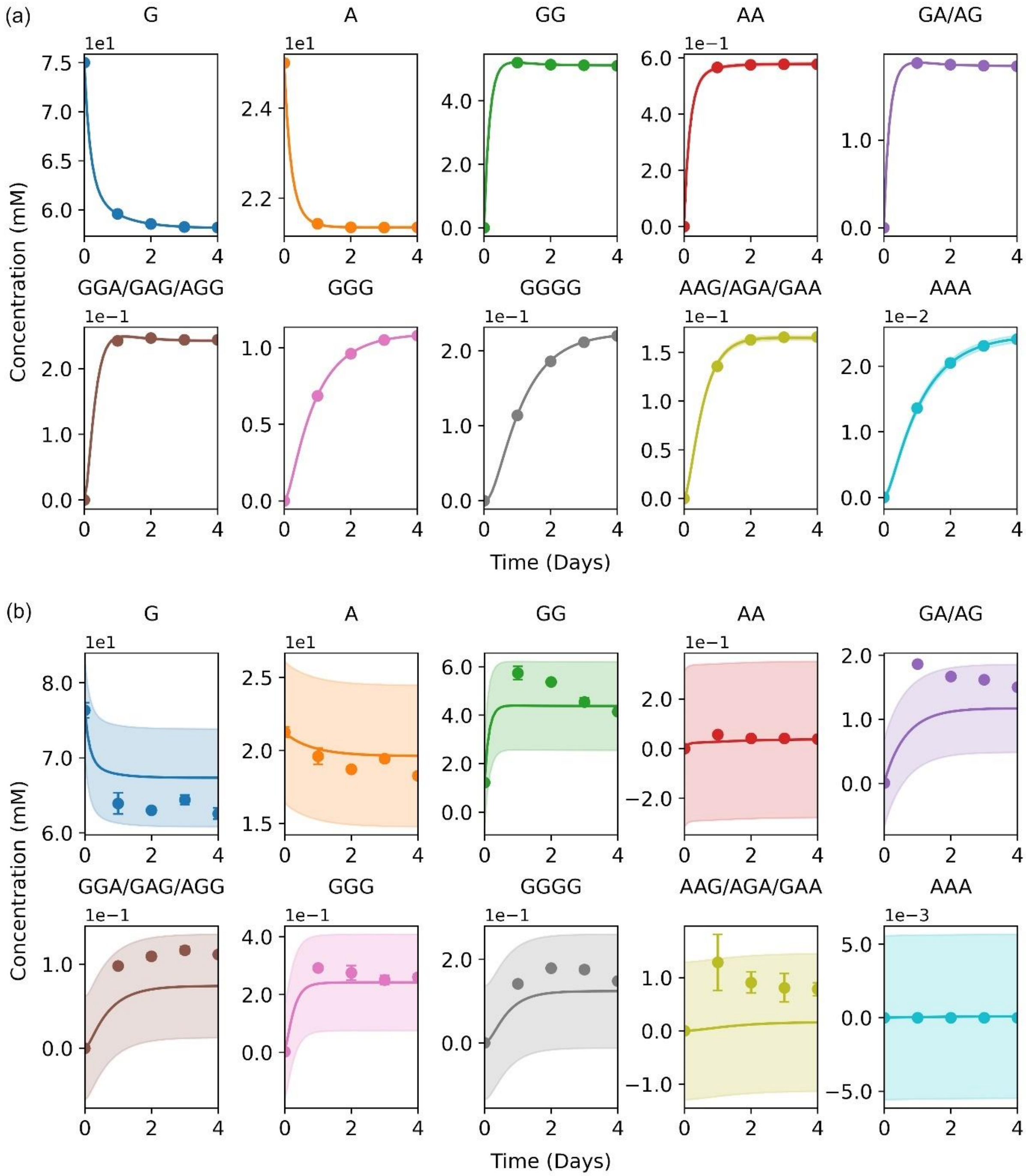


Figure 3.2 Comparison of fitting data and model predictions. Results are shown for (a) simulated data and (b) experimental data, using initial conditions of 75 mM glycine and 25 mM alanine. Both the simulated and experimental data sets included 65 data points and the simulated data had no artificially added noise.

3.4.2 Identifiability & Sloppiness

Identifiability analysis determines the possibility of a unique and precise estimate of the unknown parameters in a network (Cobelli & DiStefano, 1980; Wieland et al. 2021). If a unique solution cannot be obtained, then the model is said to be *structurally unidentifiable.* A model is *practically unidentifiable* if its parameters cannot be estimated at an acceptable level of precision. The exact definition of what is considered an acceptable level of precision varies from case to case. Practical unidentifiability indicates that regions of the objective function are relatively flat, making it difficult to find a minimum, and it typically results from overfitting (White et al. 2016). Finally, some systems exhibit a property known as *sloppiness*, which occurs when the model behavior is highly sensitive to changes in certain combinations of parameters and almost completely insensitive to changes in others (Gutenkunst et al. 2007a). Generally, sloppiness is a consequence of the model structure and its input range (White et al. 2016). Although sloppiness and practical unidentifiability are not synonymous, in practice they often coincide (Chis et al. 2014).

Sloppiness can be recognized by examining the spectrum eigenvalues of the Hessian matrix, sometimes called the sensitivity eigenvalues (see Computational Methods section for further detail) (Gutenkunst et al. 2007a). The sensitivity eigenvalues are an indirect estimate of the sensitivity of the cost function to changes in the parameter values and represent the confidence in the estimate of the parameter combination in the direction of the corresponding eigenvector. Small eigenvalues represent high uncertainties and large confidence intervals. Sloppy systems have sensitivity eigenvalues that are roughly evenly spaced across three or more orders of magnitude. When the eigenvalue spectrum is this large, the smallest sensitivity eigenvalues tend to correspond to parameter combinations that have minimal effect on the system behavior – these combinations are ‘sloppy’ eigenvectors. The eigenvectors of the largest eigenvalues are referred to as ‘stiff’ and control most of the model behavior. In some systems, there is a clear division between the large and small eigenvalues, usually corresponding to a clear separation in length or time scales that renders some of the physical details of the system irrelevant – for example, the kinetic models of many chemical reactions can be simplified when there is a known rate-limiting step (White et al. 2016). In sloppy systems, no clear division exists, and the small eigenvalues are rarely united by a single physical phenomenon.

Since rigorously checking for structural identifiability in nonlinear systems can be challenging, we tested the identifiability of our model by determining if it could recover the parameters used to generate a set of noiseless, simulated data. We found that all parameters could be recovered with acceptably high accuracy, suggesting that the model was identifiable. Here, we define acceptable accuracy to be when a parameter’s standard deviation is at least one order of magnitude smaller than the value of the associated parameter. However, when we examined the effect of noise on model performance, we observed that the parameter standard deviations rise rapidly when even a small amount of noise is introduced (Fig. 3.3a). The error of the model predictions, on the other hand, rose relatively slowly as noise increased. This suggests that despite the high parameter uncertainties, the general behavior predicted by the model can be accurate even when it is fit using noisy data (Fig. 3.3b).

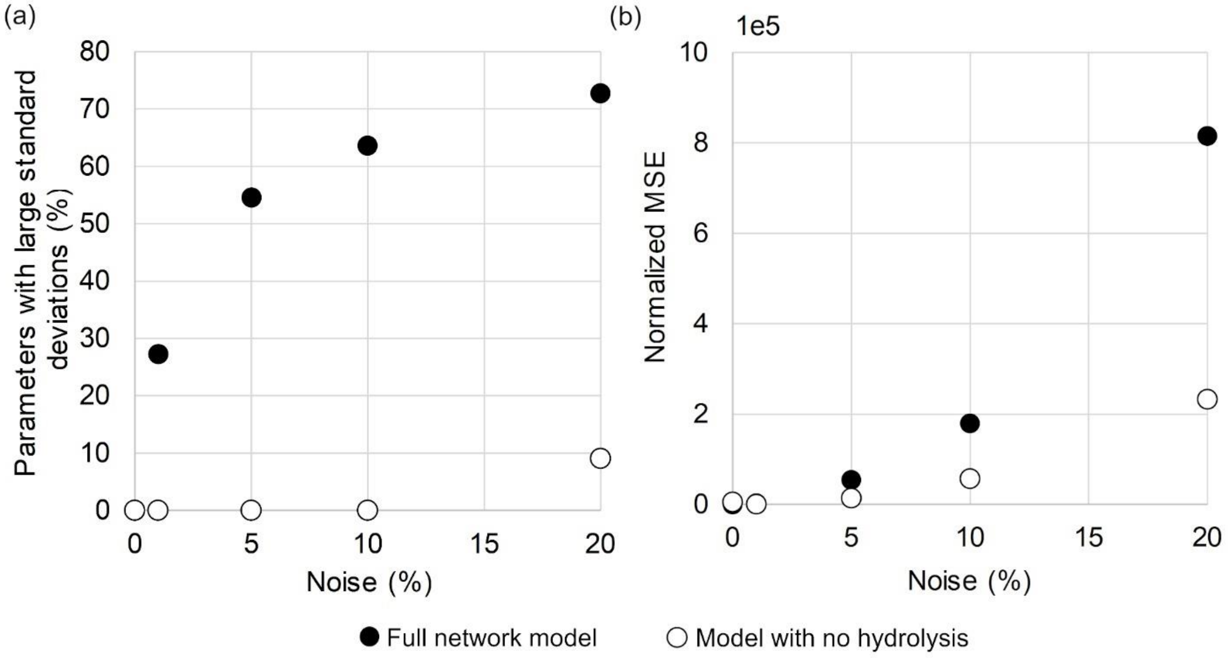


Figure 3.3 Comparison of parameter accuracy and mean squared error (MSE) for two different network structures at various noise levels. For the full reaction network, as the noise in the input data is increased, (a) the number of parameters with standard deviations within one order of magnitude of the parameter value rises rapidly compared to (b) the error of the model predictions. When the hydrolysis reactions are removed from the full network, the parameter estimates remain precise as noise is introduced. The MSE of the model predictions are normalized to the MSE of the full network with no artificial noise (2.85e-11). All data sets used simulated experiments created from 25 different initial conditions and 125 data points. The added noise was normally distributed with a constant signal-to-noise ratio with all negative values were set to zero to prevent negative concentrations.

Given that this behavior is typical in sloppy systems, we checked the sensitivity eigenvalues for both our simulated data and experimental data (Fig. 3.4a, b). We found that the peptide reaction network is unambiguously sloppy, because the sensitivity eigenvalues of the simulated and experimental data span nearly nine and seven orders of magnitude respectively. To compare the behavior of the peptide reaction network with a similar system that was not sloppy, we modified the network to exclude all hydrolysis reactions (Fig. 3.8). Removing reversible pathways from the network eliminates many combinations of parameters that can compensate for one another, which significantly reduced sloppiness (Fig. 3.4c). To demonstrate that it was the modification to the structure of the model, rather than its smaller size, that was responsible for the reduction in sloppiness, we also compared it to an even smaller network describing reversible homopolymer reactions (Fig. 3.9); this model was determined to have a much larger eigenvalue span (Fig. 3.4d). To investigate whether the grouping of some species in the peptide network was responsible for the sloppiness of the system, we also checked the sensitivity eigenvalues for a network with the trimer species separated using simulated data (Fig. 3.10), and found it made the eigenvalue spread larger (Fig. 3.4e).

The parameter standard deviations were far more sensitive to noise in the full, sloppy network than in the network with no hydrolysis reactions (Fig. 3.3a). Despite the difference in the confidence of the parameter fits, the prediction accuracy was not significantly different between the two models until 20% noise had been introduced to the system (Fig. 3.3b). This demonstrates a previously mentioned key consequence of sloppy systems – although they can make reasonably accurate predictions of system behavior, they should not be used to calculate the values of individual system parameters, since the precision required for accurate parameter estimations cannot be experimentally realized.

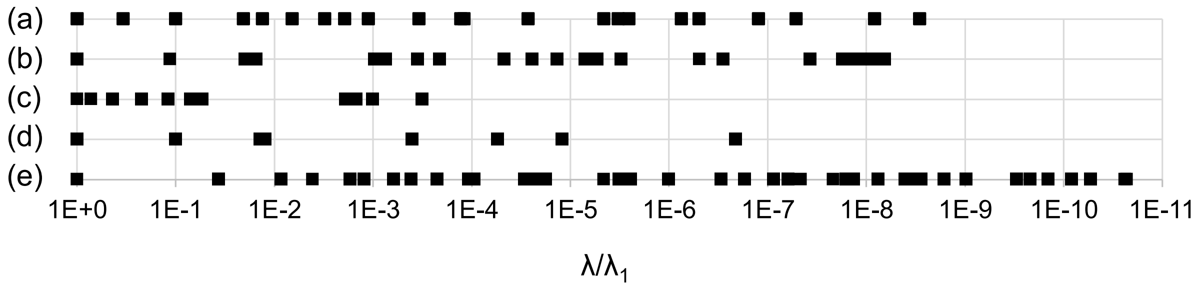


Figure 3.4 Sensitivity eigenvalues for different systems. (a) Simulated data for the full network (22 parameters, 35 data points), (b) experimental data generated from a mixture of glycine and alanine (22 parameters, 65 data points), (c) simulated data for a variation of the main network that excludes all hydrolysis reactions (11 parameters, 35 data points), (d) simulated data for network including only one amino acid forming peptides up to tetramer length with hydrolysis reactions included (8 parameters, 35 data points) and (e) simulated data for a network with separated trimers (40 parameters, 80 data points). Each system is normalized to its largest eigenvalue (λ_1). All simulated data has no additional noise included.

Sloppiness is a common property in systems biology models, and some of the characteristics that result in sloppiness are likely shared by prebiotic chemistry models. Reversible reactions and cyclic behaviors can increase the likelihood of sloppiness because they create situations where a particular combination of parameters (for example, the ratio between forward and reverse rates defining an equilibrium constant) is more important for describing the system behavior than the individual parameters themselves. The parameters may become 'sloppy' because their individual values can essentially vary freely without affecting the overall model behavior, as long changes in other parameters can compensate to produce a similar model prediction. Reaction networks that are mostly or entirely reversible, like the peptide reaction network, can therefore become significantly more difficult to fit with high precision than models with comparable sizes, but fewer reversible reactions (Maity et al. 2020). The emergence of cycles and reversible reactions are expected to be important features in the emergence of life-like chemistry (Varfolomeev & Lushchekina 2014; Mamajanov et al. 2014). Therefore, we anticipate that sloppiness may be a common and potentially unavoidable feature of ODE models found in prebiotic chemistry, and its implications should be examined.

3.4.3 Consequences of Collective Fitting

Sloppy systems can provide surprisingly accurate model predictions despite having low confidence parameter estimates. The collective fit of all the parameters tends to be more accurate and require less data than the individual parameter uncertainties might suggest, since only the stiff parameter combinations must be constrained to achieve accurate predictions. One of the consequences of collective fitting is that the numerical values of parameters estimated for sloppy systems cannot be treated as independent kinetic parameters whose quantitative values have physical meaning. Situations where a reaction occurs faster in the presence of one molecule than another are of interest to the chemical origins of life because of their semblance to catalysis. Unfortunately, in sloppy systems, the numerical values of the parameters fit in each case are often not comparable. For example, even if the rate constant of one reaction in the peptide network was significantly higher than another, that is not necessarily good evidence that one reaction proceeds faster than the other. The parameters are only meaningful when the entire system is used to describe the specific environment to which they were fit. Fixing individual parameter values to reflect direct measurements or literature values can potentially break the collective fit and significantly increase the error of the prediction, often to the point that it is no longer useful. The lack of physical meaning of the individual parameter values is a significant drawback of modeling sloppy systems. However, sloppy models can still be useful for certain tasks. For example, a sloppy model can still be used if the goal of the model is to generate predictions about the behavior of a similar system with slightly different initial conditions, or to predict responses at longer time spans. Moreover, we highlight that sloppiness might simply be a fundamental property of the actual reaction network, that arises from inherent redundancies in the system.

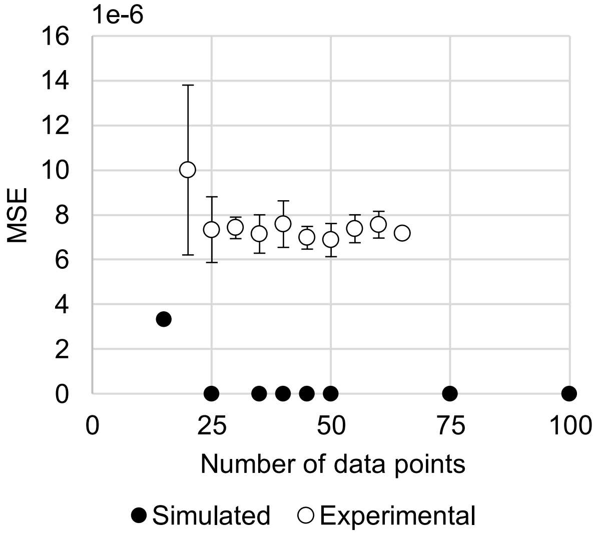


Figure 3.5 MSE of model predictions depend on quantity of experimental and simulated data. Except for the final points, which include all applicable data, parameters were estimated for three arbitrarily selected data subsets of varying sizes, then the average MSE of those models was determined. Noise was neglected. Error bars show the standard deviation of the three subsets, but are too small to be visible for the simulated data.

To estimate the minimal data required to get relatively accurate predictions, we created at least three different subsets of the data, trained the model individually with each subset, and compared their MSEs (Fig. 3.5). The simulated data was sampled at time intervals similar to the experimental results, since those were the points that were physically relevant. When training the model using simulated data, increasing the amount of data used improved the model predictions up to about 40 data points, but with even 25 data points, the error was negligible compared to the experimental results. Similarly, when we repeated the process with experimental data, the average error did not decrease as more data was added beyond 25 data points.

We also investigated the effect of using more frequent measurements, as opposed to using a greater number of simulated experiments with different initial conditions. We compared the results of simulated data with a similar number of total data points, but double the usual sampling frequency to the simulated results in Fig. 3.5. Increasing the sampling frequency was comparable or slightly worse than including data from additional simulated initial conditions, except possibly when there is little data available overall (Fig. 3.11). It did not improve the system's sensitivity to noise (Fig. 3.12).

Different subsets of the data with the same number of data points could have fairly different MSEs, suggesting that some combinations of experiments may be better for parameter fitting than others. This subject will be discussed further in the section on the design of experiments (DoE). Overall, these results suggest that as few as 25 to 30 data points are required to fit the system as accurately as the model constraints allow; therefore, reasonably accurate predictive model fits can be achieved with a realistically obtainable amount of data. The ability to extrapolate accurate model predictions from short-term experiments has some uses for studying prebiotic chemical reactions, since long time spans are potentially relevant. Models like the one we present here could be used to predict the expected equilibrium outcome of slow reactions based on data from a shorter time span and compare candidate model structures. They may also be a useful way to predict the outcomes of sequential or cyclic processes, provided that the parameters are fit in compatible experimental conditions. Sensitivity analysis can be used to validate the predictions from sloppy models independently from the parameter uncertainties (Gutenkunst et al. 2007a). Model selection, which involves comparing two or more different model structures to determine which one reflects the experimental data most accurately, can also still be performed with sloppy models (Brown & Sethna 2003). However, if finding physically meaningful terms for the parameter values is an important goal, then the aim should be to reduce the sloppiness of the model.

3.4.4 Model Reduction

To address high parameter uncertainty, one may seek to simplify the structure of the model, ideally without compromising the accuracy of the model predictions. This task is referred to as *model reduction* or *network reduction*, and it can be an effective way to improve overparameterized models (Apri et al. 2012; Transtrum et al. 2015). However, model reduction methods are generally based on statistical principles and not physical knowledge, and the results should be interpreted within an experimental context. The user must ensure that parameters that might be statistically problematic but are known to be physically significant are not removed from the model.

Since one of the main features of sloppy models is that they contain parameter combinations that are insensitive to changes, model reduction may initially appear to be a straightforward task for sloppy models. However, the fact that the sensitivity eigenvalues are evenly distributed over multiple orders of magnitude poses a challenge for accurate model reduction, as there is no clear cut-off between the parameter combinations that are important and those that are not. Additionally, in practice some parameters are so poorly constrained that they are randomly distributed throughout the sensitivity eigenvectors, so the components of the sensitivity eigenvectors are not entirely reliable indicators of what parameters are influencing them (Gutenkunst et al. 2007a).

We attempted model reduction with the peptide reaction network to determine if it was over-parameterized and if it might be possible to reduce the reactions considered. For example, we expected that some of the hydrolysis reactions could be ignored. Since we wanted to use a model reduction technique that is accessible and easily interpretable for experimentalists, we used sparse principal component analysis (SPCA). SPCA is an extension of principal component analysis (PCA), a popular dimensionality reduction method for linear models (Zou et al. 2006). Using SPCA, we can identify the inputs that capture most of the information in the data. It has been used successfully in control theory and gene network analysis, and there are existing implementations of it in MATLAB and Python (Ma & Dai 2011).

When SPCA was applied to the peptide reaction network, the results were highly variable and unable to adequately represent the data. SPCA frequently suggested removing reactions known to be physically significant, such as the formation of dimers from monomers (Supplemental Information 3.8.5). Not only does this not make physical sense, but because these are the initial reactions that occur in the system, removing them severely limits the pathways for longer species to form. Other methods of network reduction may be more effective for sloppy systems, but are less commonly used and may be more difficult to implement (Transtrum & Qiu 2014; Maiwald et al. 2016). If we choose to pursue additional model reduction efforts, one logical next step may be to inspect the inverse of the covariance matrix to identify which parameters are the most correlated and least constrained by the data (Wasserman, L. 2004). This information may be useful for determining which parameters are best to remove or to combine into a single term.

3.4.5 Design of Experiments

If the model structure cannot be altered, another method for reducing sloppiness is to determine if experimental data can be gathered strategically to explore the variable space more thoroughly (Apgar et al. 2010). However, to reduce parameter uncertainty, the selected experiments must provide new information not already captured in the model. *Design of* *experiments* (DoE), or experimental design, seeks to identify the experiments that would provide the most useful information for improving prediction accuracies. DoE methods such as factorial design (Fisher 1935), response surface methodology (Box & Wilson 1951), and screening (Shevlin 2017), have been widely adopted across various fields. However, there are several notable caveats in relation to sloppy systems (Jagadeesan et al. 2022). First, the precision of parameter fitting for sloppy systems is limited by the least accurately determined eigenvectors, so more data measured with the same uncertainty may not help. Second, there is some debate over whether DoE can be used with approximate models without risking the collective fit, as it can inadvertently place too much importance on details not included in the model (White et al. 2016).

In this work, we use a Bayesian experimental design (BED) method that selects experimental designs based on the expected reduction in parameter uncertainty as quantified by the determinant of the Fisher information matrix (FIM) (Transtrum et al. 2012; Thompson et al. 2022). To determine if there was any significant benefit obtained using DoE, we compared the reduction in parameter uncertainty from performing experiments suggested by the BED method to the reduction achieved from performing arbitrarily chosen experiments (Fig. 3.6). We evaluated the results using a couple of metrics – by the percentage of parameters with standard deviations that were large (within an order of magnitude of the relevant parameter) to indicate the overall precision of the parameter estimates, and by the MSE to indicate the accuracy of the model's predictions.

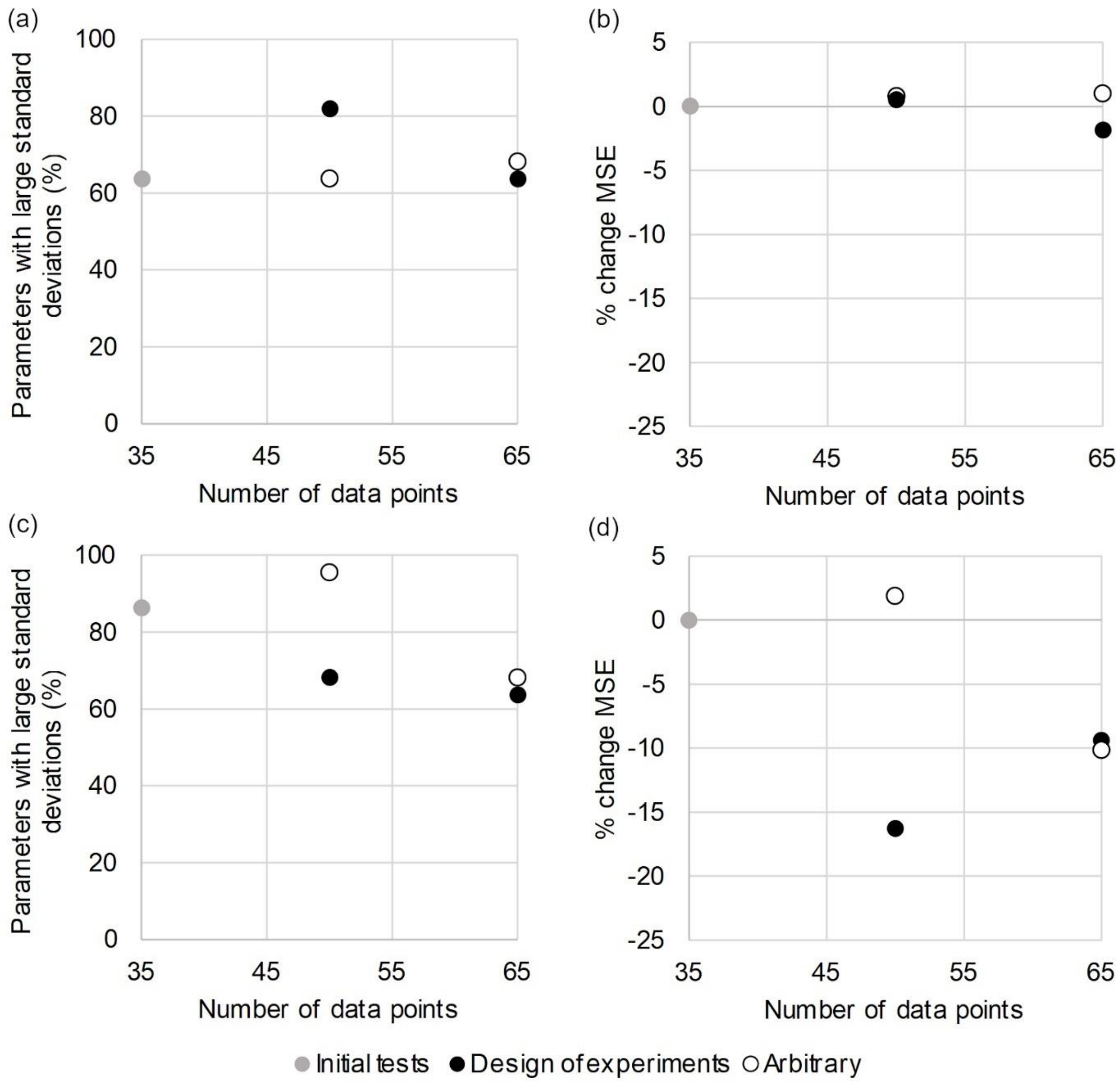


Figure 3.6 DoE slightly improved the precision of the parameter estimates and the model prediction accuracy. (a) Using simulated data with 15% noise, the percentage of large parameter uncertainties (standard deviation within one order of magnitude of the parameter value) remained consistent and (b) the MSE did not change significantly compared to the initial tests. (c) Using experimental data, the percentage of large parameter uncertainties decreased slightly and (d) the model predictions improved relative to the initial tests, but did not continue to improve as more data was added. Each round added 3 additional experiments, consisting of 5 time points measured for each experiment. For the DoE rounds, 3 experiments chosen from the top 20 experiments suggested by the DoE algorithm were added. For the control rounds, data from 3 initial conditions not included in the DoE suggestions were added (50 mM Gly, 25 mM Gly and 25 mM Ala, and 50 mM Ala).

In our preliminary tests using simulated data with artificial noise, adding results from experiments suggested by the DoE method did not reduce the number of parameters with large standard deviations or improve the accuracy of the model predictions. This suggests that the poor precision of the parameter estimates may not be caused by poor data coverage, and is instead a consequence of the model structure. When applied to our experimental data, the addition of results suggested by the algorithm did decrease the number of parameters with large standard deviations and improved the model predictions relative to the initial tests, however, there was significantly less improvement from the second round of additional experiments than there was in the first. The simulated results suggest a limit to how much additional data can improve the parameter estimates and highlight that the model structure is responsible for sloppiness. Even after nearly doubling the amount of data included in our original tests, neither the experimental nor the simulated system ever had fewer than 60% of parameters with large standard deviations and the model predictions were essentially unchanged. Overall, it seems unlikely that continued cycles would significantly improve the parameter estimates to the extent that it would allow us to attach any physical significance to their numerical values.

Data suggested by the DoE algorithm typically had similar or better performance than the data that was added arbitrarily. However, we cannot conclude there is a significant improvement from using the DoE algorithm, because during the second round of experiments using arbitrary data produced very similar results in all cases. Concerning the experimental results, conclusively determining whether the selections of the DoE algorithm are an improvement over randomly selected conditions would require performing many additional experiments. Within the existing results, we noted that model prediction errors occasionally increased when more data was added, which can be a consequence of overfitting, however, there was no consistent trend of samples outside of the training data set having significantly higher prediction errors, suggesting overfitting is not likely (Supplemental Information 3.8.6). Because the increases in prediction error are small, they are probably an incidental consequence of the noise in the data and the limited sample size.

There are several possible reasons why DoE did not consistently improve the precision of the parameter estimates this system. The precision of a sloppy model is limited by the most variable parts, so experimental noise may be preventing key features from being determined more precisely (Gutenkunst et al. 2007a). The prescribed range of initial conditions may have also been too restrictive. We only included initial conditions with various concentrations of monomers because amino acids and peptides can participate in different reaction mechanisms with TP. Since these mechanisms were not being explicitly separated in the model, initial conditions with large concentrations of peptides could have inadvertently led to measuring the parameters for a different reaction mechanism. Rather than risk measuring the kinetics of a different mechanism, which would undermine the assumption that each experiment had the same kinetic parameter value, we chose to use a more limited system definition. However, this also may have limited our ability to constrain some parts of the network. Finally, as DoE methods are statistically based approaches that rely on existing results, they can be sensitive to noise in the data. As a result, it may be difficult to predict how parameter uncertainties will change as additional data is added. Therefore, because sloppy networks tend to be better at producing accurate predictions than accurate parameter estimates, approaches that aim to improve predictions rather than parameter uncertainties may be more useful.

3.4.6 Model Limitations

The mass-action style model used here is a significant simplification of the reactions occurring in the actual experimental system. TP-activated peptide bond formation involves not only multiple intermediates but likely multiple reaction mechanisms, which were not fully described in this model (Boigenzahn & Yin 2022). Certain products, like the cyclic dimer 2,5-diketopiperazine were not detectable or quantifiable in our analysis. Merging the isomeric peptide species also may have increased the experimental error slightly, since not all isomers have the same absorbance. However, on average, the species balances of glycine and alanine were about 90% accurate, suggesting that any products missed by our analysis were probably not dominant products in the system. While we acknowledge the simplifications and sources of noise in our experiments, it is important to note that the model generated high parameter standard deviations when extremely small amounts of noise were added to simulated data. It may not be possible to fit the current version of the peptide network with high precision from experimental data.

It might be possible to alleviate sloppiness by replacing the generic reversible reactions in this model with more detailed descriptions and measurements of intermediates. However, this would significantly increase the resources needed for experimental and statistical analysis. Additionally, this model does not account for increasing concentration of all species as the sample dries. The volume could be included as a dynamic term in the network model, but it complicates parameter estimation because of the infinite limits that occur as the volume approaches zero. There are also potential reactions that occur almost exclusively in the solid phase (Napier & Yin, 2006). We chose to neglect any concentration effects or details of the TP reaction mechanism and instead explored the feasibility of creating a model that predicted overall peptide production.

3.5 Conclusion

Although we were able to fit kinetic parameters to the peptide reaction network in our simulated tests, in practice the parameter estimations were poorly constrained due to sloppiness. Neither network reduction nor statistical design of experiments were particularly successful for reducing sloppiness or improving the precision of the parameter estimates for this example. Sloppiness precludes us from drawing any physical conclusions based on the individual values of the parameters estimated in these models, but this approach is still an effective way to make model predictions based on relatively few time points. The predictive capacity of the model may be useful for forming hypotheses about the behavior of systems that pass through multiple conditions sequentially, or simply estimating equilibrium conditions based on short-term experiments.

Our goal was not only to explore the kinetics of these specific reactions, but to evaluate the potential challenges and opportunities of applying mathematical tools, which were originally developed for biological networks to prebiotic chemical systems. Sloppiness is a challenge when studying the kinetics of complex nonlinear systems but may be an interesting property in the broader context of the chemical origins of life; sloppiness has been suggested as a possible non-adaptive explanation for the robustness of many multiparameter biological systems (Daniels et al. 2008). This idea suggests that many complex networks, ranging from those found in biology to those that are randomly generated, have similar behavior across large areas of the parameter space. This implies that robustness, in this case a reaction network's ability to achieve similar outcomes despite variation in its parameter values, can emerge from complexity even when it is not specifically selected for. The feature of intrinsic robustness in sufficiently large multiparameter networks observed in deep neural networks, which can be dramatically complex but highly accurate, and is an open area of investigation in the machine learning community (Belkin, et al. 2019). As a result, there is a significant incentive to work towards studying more complex experimental origins of life systems.

Adapting systems biology tools to study complex origins of life experiments lends itself to an interdisciplinary approach, since many methods can be difficult to implement or even approach without expert assistance. Demonstrative studies like this one can improve experimentalists' understanding of what data analysis approaches are available, what their limitations are, and what results they can provide. We hope that using computational networks to analyze experiments will become more commonplace and enable the study of more complex origins of life reaction networks.

3.6 Computational Methods

The usefulness of a parametric model is limited by our ability to accurately determine the values of the corresponding parameters. A large body of work has detailed various parameter fitting or regression techniques that can be used to build these models (Bard 1974). The most popular parameter estimation method is maximum likelihood estimation (MLE). In MLE, the noise from experimental measurements (ϵ) is treated as a random variable that captures the error between the model predictions and the observed output values:

$$\boldsymbol{y} = m(X; \boldsymbol{\theta}) + \boldsymbol{\epsilon} \tag{3.1}$$

where $\epsilon \in \mathbb{R}^S$, S is the number of observations (measurements) available, m is the model and $\boldsymbol{\theta} \in \mathbb{R}^n$ are its n parameters. The set of output observations is stored in the vector $y \in \mathbb{R}^S$, and $X \in \mathbb{R}^{S \times K}$, known as the design or feature matrix, is structured so that the s^{th} row corresponds to the s^{th} observation, $\mathbf{x}_s$, and the k^{th} column corresponds to the k^{th} input variable x_k. Combining MLE's assumption that $\boldsymbol{\theta}$ and X are deterministic variables with the most common noise model, the Gaussian or normal distribution ($\left(\epsilon \sim \mathcal{N}(0, \Sigma)\right)$, where Σ is the covariance of the noise) allows us to exploit the fact that the sum of normal distributions is also a normal distribution. We can use this to calculate the distribution for the observations vector, $y \sim \mathcal{N}(m(X, \boldsymbol{\theta}), \Sigma)$. The goal of MLE is then to find the values of $\boldsymbol{\theta}$ that best account for the experimental observations, or the values for $\boldsymbol{\theta}$ that best parameterize this output distribution. This is done by determining the values that maximize the log-likelihood function, $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, log\, f(\boldsymbol{y}|X, \boldsymbol{\theta}, \Sigma) \tag{3.2}$$

where $f\left(\mathbf{y} \,\middle|\, X, \boldsymbol{\theta}, \Sigma\right)$ is the likelihood (or conditional probability) that the outputs in $\mathbf{y}$ would be observed given values for X, $\boldsymbol{\theta}$, and Σ. For the given distribution of $\mathbf{y}$:

$$f(\mathbf{y}|X, \boldsymbol{\theta}, \Sigma) = \left((2\pi)^{-\frac{S}{2}}|\Sigma|^{-\frac{1}{2}}\right)\exp\left(-\frac{1}{2}\left(\boldsymbol{y} - m(X; \boldsymbol{\theta})\right)^T \Sigma^{-1}\left(\boldsymbol{y} - m(X; \boldsymbol{\theta})\right)\right). \tag{3.3}$$

The well-known ordinary least squares regression problem is a special case of MLE where the model is linear and Σ is a diagonal matrix composed of identical values (σ^2).

A common issue with MLE is that (2) can have multiple solutions ($L(\boldsymbol{\theta})$ is nonconvex), as is often the case with nonlinear models. However, some of these solutions may contain parameter values that are not physically sensible, making the solution invalid. One way to overcome this limitation is to shift the goal of (2) from maximizing the probability of measuring the observed outputs given a set of parameters to maximizing the probability of a set of parameters being correct given a set of observations. Mathematically, this is done using Bayes' theorem, $f\left(\boldsymbol{\theta} \,\middle|\, \mathbf{y}\right) \propto f\left(\mathbf{y} \,\middle|\, \boldsymbol{\theta}\right) f(\boldsymbol{\theta})$, and changes the likelihood function to:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \log f(\mathbf{y}|X, \boldsymbol{\theta}, \Sigma) + \log f(\boldsymbol{\theta}) \tag{3.4}$$

where now we no longer assume that $\boldsymbol{\theta}$ is deterministic but instead has some distribution (e.g., $\boldsymbol{\theta} \sim \mathcal{N}(\overline{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$) that is captured by the prior $f(\boldsymbol{\theta})$. This term can be used to input any prior knowledge or expectation one might have over the values of the model parameters (e.g., must have a certain sign, lay within a specified range, etc.) and thereby constrain the search to values of $\boldsymbol{\theta}$ that satisfy the desired criteria. If $\mathbf{y}$ and $\boldsymbol{\theta}$ are normally distributed, then (4) can be expressed as:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2}\big(\mathbf{y} - m(X; \boldsymbol{\theta})\big)^T \Sigma^{-1} \big(\mathbf{y} - m(X; \boldsymbol{\theta})\big) + \frac{1}{2}(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})^T \Sigma_{\theta}^{-1} (\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}) \tag{3.5}$$

Note that the first term will be minimized when the model predictions exactly match the output observations, while the second term will be minimized when $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$. To perform the optimization of model parameters, we use the L-BFGS-B algorithm from SciPy's minimize function with a tolerance for termination of 1e-3. As a result, Bayes' estimation seeks to balance the fit of the model with the prior knowledge over the parameters that is available. We use an Expectation-Maximization (EM) algorithm to determine the covariance matrix of the measurement noise and the parameter prior that maximizes the model evidence (Thompson et al. 2022).

Due to the randomness in $\mathbf{y}$, the selected parameters $\boldsymbol{\theta}^*$ will exhibit an inherent uncertainty that is determined by how well the estimates are constrained by experimental data. The parameter uncertainty is largely controlled by the model structure as well as the quality and quantity of the available data. If a model is selected where certain inputs are not strong predictors of the outputs or are dependent on other inputs, or if the dataset is too small or contains redundant samples, then $\boldsymbol{\theta}^*$ will be imprecise. This is a major issue as it can lead to overfitting, where m is not able to make accurate predictions at values of x that are outside of the dataset.

An estimate of the parameter uncertainty can be obtained from the eigenvalues of the Hessian matrix, $\mathcal{H}(\mathbf{y};\boldsymbol{\theta})$, also known as the Fisher information matrix (FIM) in the context of parameter estimation, which is defined as:

$$\mathcal{H}_{i,j} = \frac{\partial^2 L}{\partial\theta_i \partial\theta_j} \tag{3.6}$$

The eigenvalues of the Hessian serve as an estimate of data sufficiency. From calculus we know that the second derivative of a function, f'', determines if a critical point $(f' = 0)$ is a maximum $(f'' < 0)$, a minimum $(f'' > 0)$, or an inflection point $(f'' = 0)$, which could be either a minimum, a maximum, or neither. Additionally, we can also estimate how sharp or defined an extremum is from the value of $f"$. As a result, we can use $\mathcal{H}(\mathbf{y};\boldsymbol{\theta})$ to gauge the quality of the obtained solution. For example, if all the eigenvalues of $\mathcal{H}(\mathbf{y};\boldsymbol{\theta})$ are large and positive $(\gg 0)$, this implies that $\boldsymbol{\theta}^*$ sits in a well-defined minimum and provides a precise estimate of the parameters. If all the eigenvalues are positive and one or more are small $(\ll 1)$, then the minimum is not sharp, and the parameter estimates will be ill-defined and exhibit high variability. Finally, if $\mathcal{H}(\mathbf{y};\boldsymbol{\theta})$ has any eigenvalues equal to zero, then $\boldsymbol{\theta}^*$ lays on a flat surface and cannot be uniquely estimated from the data; in other words, $\boldsymbol{\theta}^*$ has infinite variability.

If the precision of $\boldsymbol{\theta}^*$ is deemed to be too low, there are two methods that can be used to improve the quality of the estimates. The first, known as system identification, involves the structure of the model and the selection of the input variables. We can determine the relative importance of the input variables using a feature importance technique such as automatic relevance determination (ARD), or model class reliance (MCR), or as used in this paper, sparse principal component analysis (SPCA) (Zou et al. 2006). This information can then be used to restructure *m* to eliminate any redundant inputs.

If system identification is not able to reduce the uncertainty of the parameter estimates to a desired level, a second approach is to collect additional data. However, the data must provide additional information beyond what is already contained in the current dataset to have any chance of improving the parameter estimates. One way to achieve this is by using a design of experiments (DoE) algorithm to select experiments that have a maximal value. Depending on the goal of the experiments (optimization, discovery, or both), their value can be measured by the information content they provide or by their predicted proximity to a desired set of properties. There is a rich variety of DoE algorithms to select from such as response surface methodology (RSM), screening, factorial design, etc. (Fisher 1937; Box & Wilson 1951; Shevlin 2017). A common metric to evaluate the optimality of candidate experimental designs is the determinant of the FIM. For any candidate experimental design, X, the FIM is computed as

$$\mathcal{H}_{i,j} = \frac{\partial^2 L}{\partial\theta_i \partial\theta_j} = \Sigma^{-1}_{\theta_{i,j}} + \frac{\partial m(X,\boldsymbol{\theta})}{\partial\theta_i}\Sigma^{-1}\frac{\partial m(X,\boldsymbol{\theta})}{\partial\theta_j} \tag{3.7}$$

where evaluations of the gradient with respect to model parameters is computed using the forward sensitivity equations (Ma et al. 2021).

While DoE can be very useful for improving parameter uncertainties, there are several challenges. Calculating the expected information gain (EIG) can be time consuming due to the number of operations that need to be performed for larger systems. As a result, obtaining a new batch of experiments can easily take on the order of hours depending on the size of the dataset and the number of parameters involved. Even for moderately sized models, the quantity or precision of an experimental system may not be sufficient for accurate predictions of the information generated by each experiment to be made in the first place, or the experiments that would provide the information may not be feasible in reality. Both cases seriously hinder the effectiveness of DoE methods.

Selection of experiments for the DoE method was performed as in Thompson et al. (2022). Experimental data was normalized using linear scaling to ensure that the concentration values for each species spanned [0,1]. Scaling the data ensures that low abundance species still affect the parameter fits, which was necessary since the experimental results span several orders of magnitude. Parameters values were limited to [0,10] for simplicity, though we found that raising the upper bound had no effect if the initial guesses were single digit. Negative values had no physical meaning since both directions of the reversible reactions were already included. All computational methods were performed using Python 3.2.2. We used automatic differentiation in PyTorch to calculate the gradients of the loss function and SciPy to solve the initial value problems. Relevant code is available at https://github.com/haboigenzahn/OoL-KineticParameterEstimation.

Simulated data for testing was generated in Python 3.2.2 using SciPy 1.7.1 solve_ivp. The parameters for the simulated data were loosely based on the parameter fits of the experimental data, but were rounded to integers (Table 3.3). Network figures were generated using Cytoscape 3.7.2 (Shannon et al. 1971).

3.7 Experimental Methods

All chemicals were of analytical grade purity and used without further purification. Materials were obtained from suppliers as follows: trisodium trimetaphosphate (TP) and trifloroacetic acid (TFA) from Sigma-Aldrich, sodium hydroxide from Fisher Scientific, acetone from Alfa Aesar, 9-fluorenylmethoxycarbonyl chloride (FMOC) from Creosalus, acetonitrile from VWR Chemicals, and sodium tetraborate anhydrous from Acros Organics. Reactions were carried out in 1.5 mL low-retention Eppendorf tubes. Peptide standards came from various sources: glycine, diglycine, triglycine, pentaglycine, dialanine and Ala-Gly from Sigma-Aldrich, tetraglycine from Bachem, Gly-Gly-Ala from Chem-Impex International, Ala-Gly-Gly from ChemCruz, Ala-Ala-Gly from Pepmic, and Gly-Ala-Gly, Gly-Ala-Ala, Ala-Gly-Ala and trialanine from Biomatik.

Samples were prepared with 0.15 M NaOH, various concentrations of glycine and alanine, and TP in equimolar concentration to the total amount of amino acid. Details of the initial conditions chosen are included in Supplemental Information 3.8.8. Samples were placed on a heat block preheated to 90°C with the caps open and allowed to dry for 24 hours. At the end of each day of drying, samples were rehydrated with 1000 μL milliQ water preheated to about 65°C, capped and vortexed (Pulsing Vortex Mixer, Fisher Scientific) 3000 rpm until everything was dissolved, which took 1-3 minutes per sample.

To analyze the samples with UV-HPLC, they were first derivatized using FMOC, which increases the retention time and signal strength of peptide analytes. For the FMOC derivatization, 25 μL of sample was diluted with 75 μL milliQ water to put the large monomer peaks in a quantifiable range. Each sample was then mixed with 100 μL 0.1 M sodium tetraborate buffer for pH control. Finally, 800 μL 3.125 mM FMOC dissolved in acetone was added to each sample. For a sample of 0.1 M amino acid, this results in an equal concentration of FMOC and amino acid, and a slight excess of FMOC in any samples where peptide bond formation had occurred. Linear calibration curves were determined for all species using this approach (Fig. 3.15), which were used to estimate peptide concentration based on the integrated absorbance values of the HPLC peaks of the samples.

Samples were analyzed with a Shimadzu Nexera HPLC with a C-18 column (Phenomenex Aeris XB-C18, 150 mm x 4.6 mm, 3.6 μL). Products were measured at 254 nm. UV-HPLC analysis was performed using Solvent A: milliQ water with 0.01% v/v trifluoroacetic acid (TFA) and Solvent B: acetonitrile with 0.01% v/v TFA. The following gradient was used: 0-4 min, 30% B, 4-12 min, 30-100% B, 14-15 min, 100-30% B, 15-17 min, 30% B. The solvent flow rate was 1 mL/min. Peak integration was performed using LabSolutions with the ‘Drift’ parameter set to 10000.

3.8 Supplemental Information

3.8.1 Supplemental Information: Complete ODE equations for the network in Fig. 3.1.

$$\begin{aligned}\partial[G]/\partial t = {}& -2k_1\,[G]^2 + 2k_2\,[GG] - k_5\,[G][A] + k_6\,[GA/AG] - k_7\,[GA/AG][G] \\ & + k_8\,[GGA/GAG/AGG] - k_9\,[G][GG] + k_{10}\,[GGG] - k_{11}\,[G][GGG] \\ & + k_{12}\,[GGGG] - k_{17}[AA][G] + k_{18}[AAG/AGA/GAA]\end{aligned}$$

$$\begin{aligned}\frac{\partial[GG]}{\partial t} = {}& k_1[G]^2 - k_2[GG] - k_9[G][GG] + k_{10}[GGG] - 2k_{13}[GG]^2 + 2k_{14}[GGGG] \\ & - k_{15}[GG][A] + k_{16}[GGA/GAG/AGG]\end{aligned}$$

$$\begin{aligned}\frac{\partial[A]}{\partial t} = {}& -2k_3[A]^2 + 2k_4[AA] - k_5[G][A] + k_6[GA/AG] - k_{15}[GG][A] \\ & + k_{16}[GGA/GAG/AGG] - k_{19}[GA/AG][A] + k_{20}[AAG/AGA/GAA]\end{aligned}$$

$$\begin{aligned}\frac{\partial[AA]}{\partial t} = {}& k_3[A]^2 - k_4[AA] - k_{17}[AA][G] + k_{18}[AAG/AGA/GAA] - k_{21}[A][AA] \\ & + k_{22}[AAA]\end{aligned}$$

$$\begin{aligned}\frac{\partial[GA/AG]}{\partial t} = {}& k_5[G][A] - k_6[GA/AG] - k_7[GA/AG][G] + k_8[GGA/GAG/AGG] \\ & - k_{19}[GA/AG][A] + k_{20}[AAG/AGA/GAA]\end{aligned}$$

$$\begin{aligned}&\partial[GGA/GAG/AGG]/\partial t \\ &\quad = k_7\,[GA/AG][G] - k_8\,[GGA/GAG/AGG] + k_{15}\,[GG][A] \\ &\quad + k_{16}\,[GGA/GAG/AGG]\end{aligned}$$

$$\frac{\partial[GGG]}{\partial t} = k_9[G][GG] - k_{10}[GGG] - k_{11}[G][GGG] + k_{12}[GGGG]$$

$$\frac{\partial[GGGG]}{\partial t} = k_{11}[G][GGG] - k_{12}[GGGG] + k_{13}[GG]^2 - k_{14}[GGGG]$$

$$\partial[AAG/AGA/GAA]/\partial t = k_{17}\,[AA][G] - k_{18}\,[AAG/AGA/GAA] + k_{19}\,[GA/AG][A] - k_{20}\,[AAG/AGA/GAA]$$

$$\frac{\partial[AAA]}{\partial t} = k_{21}[A][AA] - k_{22}[AAA]$$

3.8.2 Supplemental Information: Results of multiple initial guesses

To evaluate the effect of the initial condition on the quality of the parameter estimation, we fit the experimental results at 7 different initial guesses. Two were chosen to be the same integer for all parameters and 5 were randomly generated lists of integers between 0 and 9. We found that there was no correlation between initial guesses that generated a lower MSE and initial guesses that improved the precision of parameter estimation. These results suggest that using a multi-start approach can help minimize the MSE of the parameter estimation, does not seem to significantly improve the uncertainty of the parameter estimates.

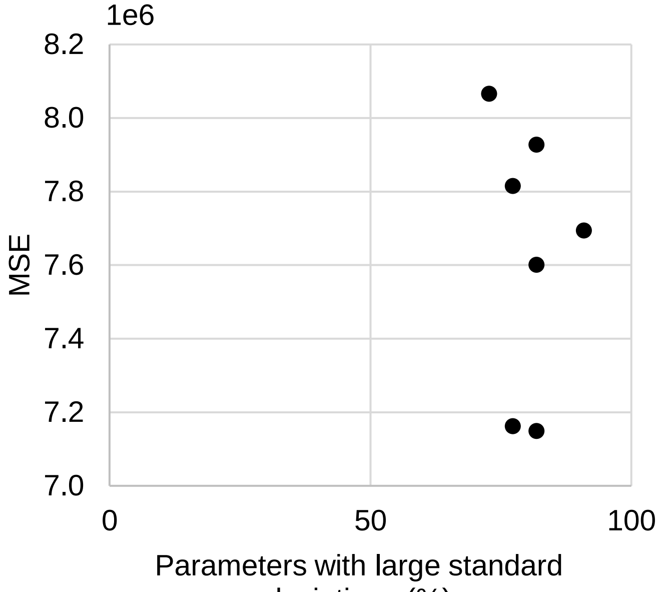


Figure 3.7 Supplemental Information: Fraction of parameters with standard deviations within one order of magnitude of the parameter value versus MSE for various initial guesses. All points were generated using the complete experimental data.

3.8.3 Supplemental Information: Other model networks

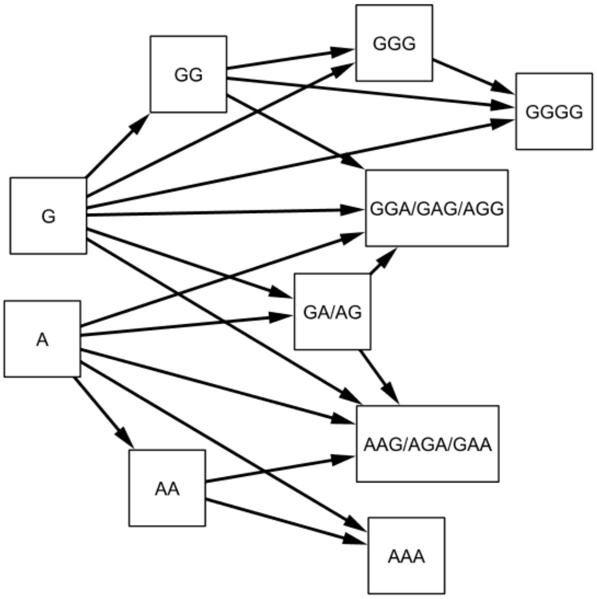


Figure 3.8 Supplemental Information: Model of peptide network with no hydrolysis. Used for the irreversible network in Fig. 3.3.

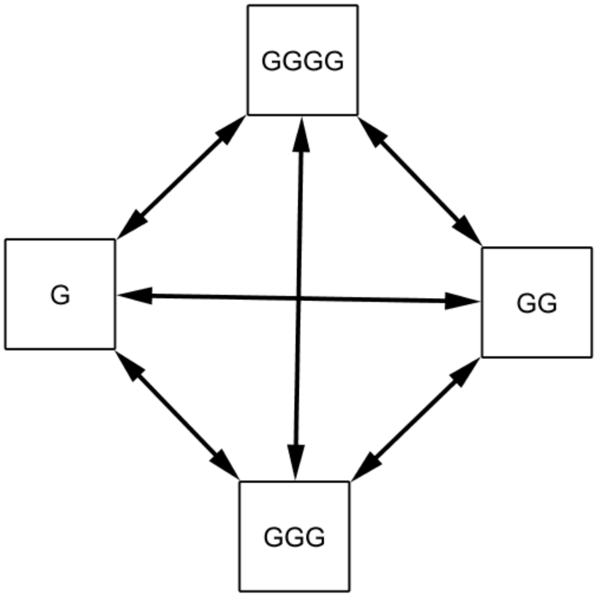


Figure 3.9 Supplemental Information: Glycine homopolymer model used for Fig. 3.4d.

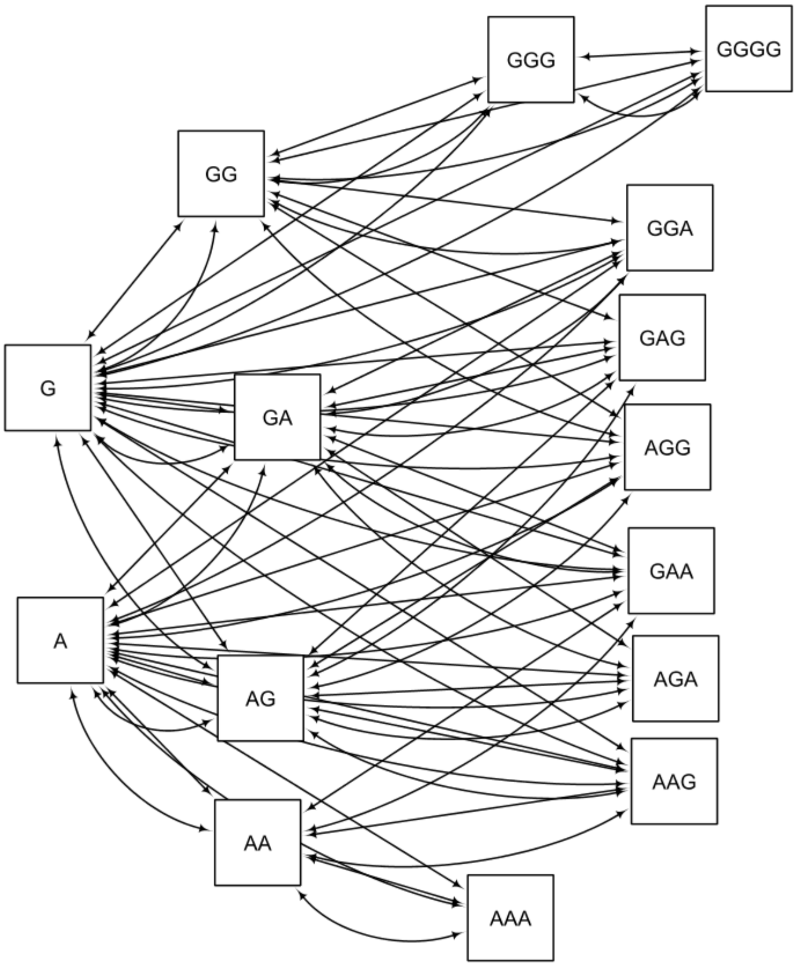


Figure 3.10 Supplemental Information: Network with separated trimers used for Figure 3.4e.

3.8.4 Supplemental Information: High frequency sampling results

Since the peptide concentrations change most rapidly during the first 1-2 days of reactions, we checked to see if increasing the frequency of the time points where data was gathered would improve the model predictions. We found that for simulated data, doubling the sampling frequency was only a small improvement relative to the same number of data points generated using a greater number of simulated initial conditions.

When comparing an equal number of total data points, doubling the sampling frequency is only helpful when the number of total data points was low. Above about 40 total data points, the MSE was lower when we included data from additional simulated initial conditions (Fig. 3.11). Using fewer than 40 data points risks reducing the accuracy of the model predictions due to insufficient data, so we concluded that it is generally better to use more initial conditions rather than to try to increase the sampling rate. These results may vary slightly depending on the initial conditions of the data used, but this analysis generally suggests that increasing the sampling frequency does not significantly improve the accuracy of the model predictions. It should also be noted that due to the complexity of the underlying mechanism, experimental data gathered at two different sampling rates may not be directly comparable, since one sampling rate may include a detail of the system behavior which the other does not. This issue was one of the motivations for consistently using integer sampling.

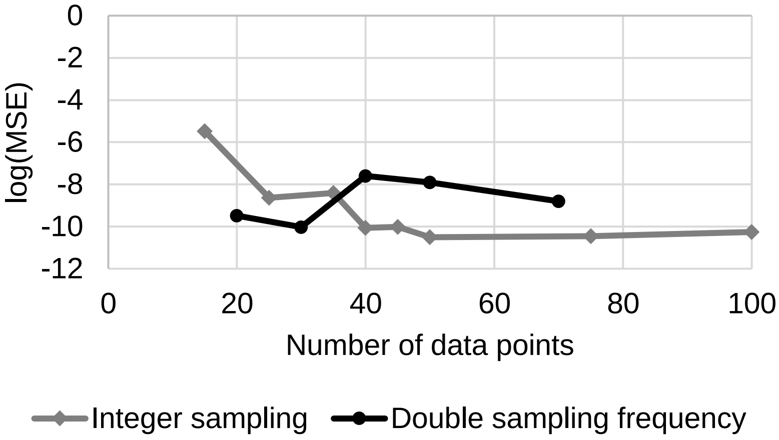


Figure 3.11 Supplemental Information: Model prediction accuracy of noiseless, simulated data created with different sampling frequencies. The data in this manuscript uses integer sampling to match the experimental conditions. Data generated with the doubled sampling frequency included a data point at every integer and half-integer.

We also examined whether using more frequent sampling decreased the sensitivity of the parameter estimations to artificial noise. These values are not directly comparable to those in Fig. 3.3, since they use fewer total data points for expediency, however, the qualitative behavior, particularly the rapid initial rise when even 1% noise is added, is recognizably similar to the parameters estimated using integer sampling.

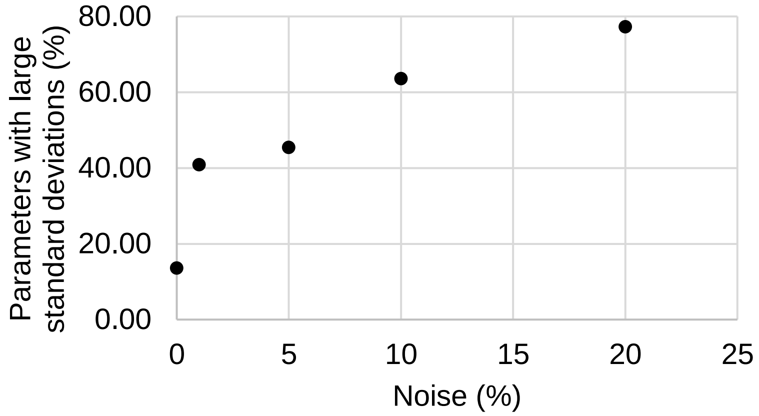


Figure 3.12 Supplemental Information: Parameter accuracy of parameters estimated from data with increased sampling frequency at various noise levels. Parameter standard deviations are considered large if they are within one order of magnitude of the parameter value. All data was simulated and included a total of 70 data points.

3.8.5 Supplemental Information: SPCA Results

First we will discuss the SPCA results for noiseless, simulated data including 35 data points. We chose to split at the 9th eigenvalue, which meant removing variables with entries > 0 in the last 13 columns of the W_{sparse} matrix. The parameters removed were k_1, k_2, k_4, k_6, k_7, k_{12}, k_{14}, k_{15}, k_{16}, and k_{18}.

The network reduction introduces large errors in the fitting that were not present when the complete network was used. These results were typical among the variations of this process we attempted using different cutoffs for λ, different data sets (both simulated and experimental), and SPCA seeds. Reduced networks calculated using this method increased the parameter uncertainty rather than decreasing it, and in many cases also decreased the accuracy of the model predictions. The fact that model reduction was not able to lower parameter uncertainties when the fits are highly accurate might be expected, but the reduced models also consistently failed to improve the parameter accuracy of the experimental data.

We also tried repeating this test with a very conservative estimate for the eigenvalue cutoff by splitting at the 19th eigenvalue. Although the results were an improvement compared to those shown here, the prediction accuracy was still lower than when the full network was included, without any noticeable improvement in the parameter standard deviations.

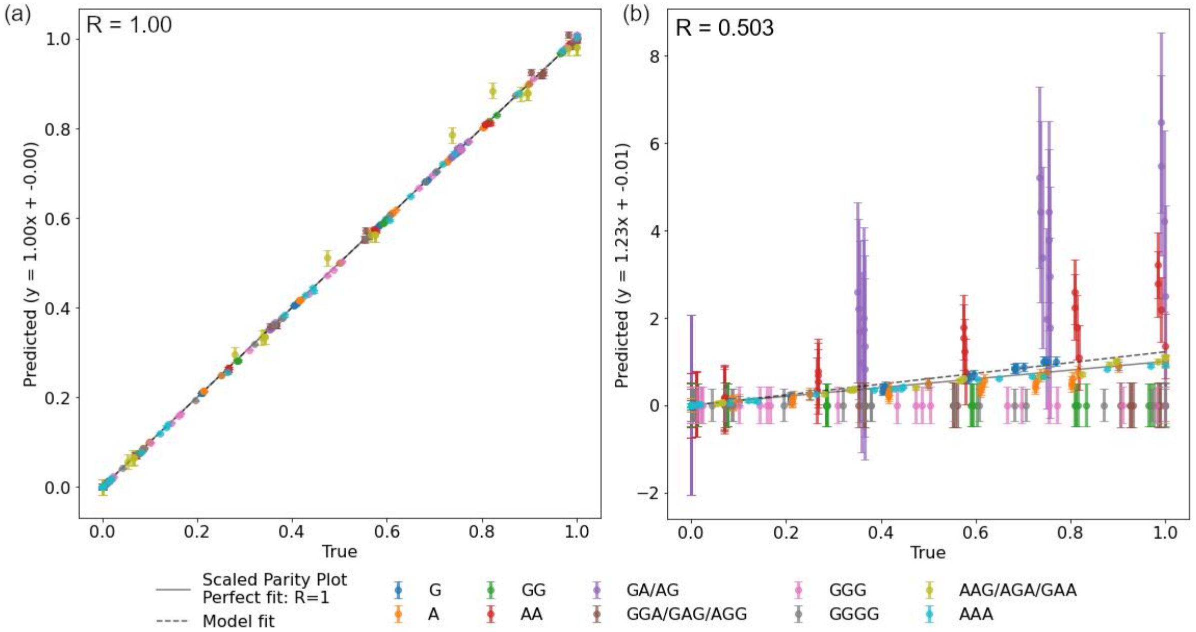


Figure 3.13 Supplemental Information: Parity plots showing model predictions vs. simulated values (a) before and (b) after model reduction by SPCA. Data is scaled so that each species spans the interval [0,1] to ensure the lower abundance species are visible.

W_{sparse}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
k1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
k2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
k3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
k5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k6	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
k7	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
k8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k9	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k11	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
k12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
k13	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
k15	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
k16	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
k17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k18	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
k19	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k20	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k22	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3.1 Supplemental Information: SPCA matrix for simulated data. Shaded values indicate parameters selected for removal based on the exclusion of everything after the 9th eigenvalue.

Next we will discuss the SPCA results for the experimental data, which included 65 data points. Repeating the process above with the experimental data, we split at the 13th eigenvalue, which meant removing variables with entries > 0 in the last 9 columns of the W_{sparse} matrix. The parameters removed were k_3, k_6, k_7, k_9, k_{10}, k_{13}, k_{21}, and k_{22}.

The suggested network reduction included one of the reactions that forms dimers from monomers ($k_3 = A + A \rightarrow A_2$), which does not agree with our physical understanding of the behavior of the network. This network significantly over-predicted the formation of GA/AG, possibly due to it being one of the few remaining pathways connecting alanine to the rest of the network in the absence of the formation of dialanine.

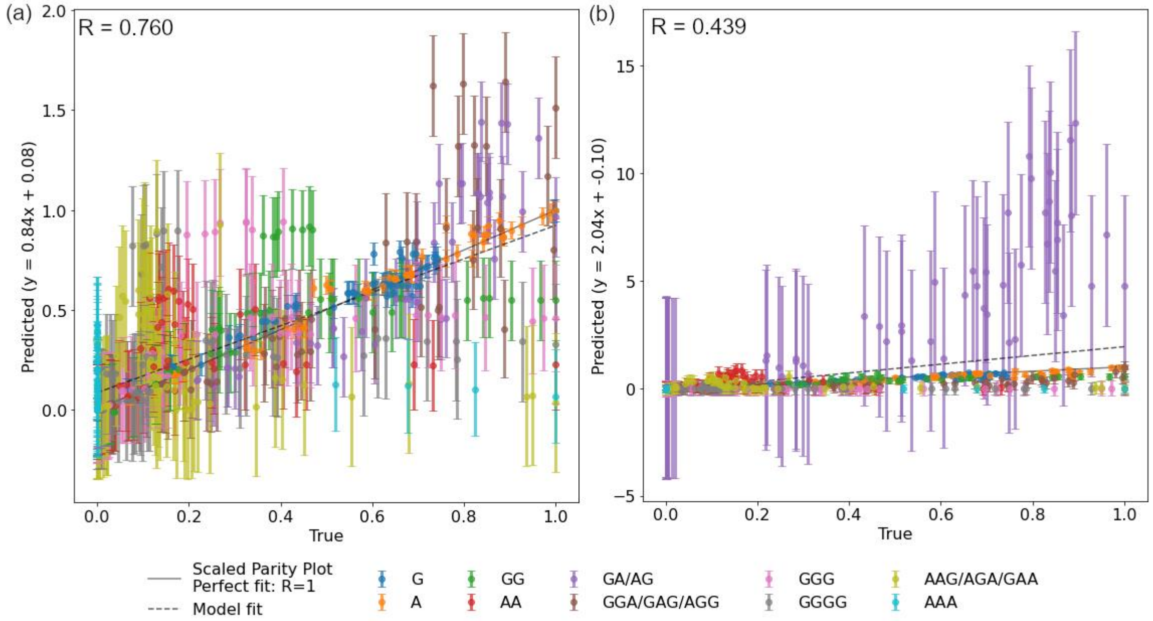


Figure 3.14 Supplemental Information: Parity plots showing model predictions vs. experimental values (a) before and (b) after model reduction by SPCA. Data is scaled so that each species spans the interval [0,1] to ensure the lower abundance species are visible.

Wsparse	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
k1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
k4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
k5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
k6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
k7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
k8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
k10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
k11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
k12	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
k13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
k14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k15	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
k16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k17	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k18	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k19	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
k20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
k22	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 3.2 **Supplemental Information: SPCA matrix for experimental data.** Shaded values indicate parameters selected for removal based on the exclusion of everything after the 13th eigenvalue.

3.8.6 Supplemental Information: Comments on overfitting

One method of checking for overfitting to check if the model predictions are significantly better for the data the model was trained on than similar data that it was not trained on. In this case, the training data refers to the data that was used to estimate the model parameters. Data not included in the training data will be referred to as the test data.

We checked for overfitting in our experimental design tests since the total model MSE increased when additional data was added. Using the first seven experiments as the training data and the other six experiments as the test data, the test results ($\mu = 2.04^{-4}$, $\sigma = 9.92^{-3}$) did have a higher average MSE than the training data ($\mu =1.29^{-4}$, $\sigma = 0.0195$), but the result was not statistically significant ($p = 0.214$). When we repeated the test using the data from the first round of DoE testing as the training data and only the three experiments that were later performed for the second round of DoE testing as the test data, the test data ($\mu = 1.115^{-4}$, $\sigma = 1.46^{-4}$) actually had a slightly lower average MSE than the training data ($\mu = 1.387^{-4}$, $\sigma = 2.186$), though again the results were not statistically significant ($p = 0.387$). Collectively, these results suggest that overfitting was not the cause of the increasing MSE as more data was added during the DoE tests.

During these tests, we observed that specific experiments had much higher MSE values than others. Experiments starting from 0.1 M glycine were notable outliers in all tested cases. This may be the result of longer glycine products forming that were not included in the model or another source of experimental error involving the glycine oligomers.

3.8.7 Supplemental Information: Parameters for simulated data

Table 3.3 Supplemental Information: Parameter values used to generate simulated data for Chapter 3.

k1	6
k2	4
k3	5
k4	4
k5	6
k6	4
k7	5
k8	2
k9	4
k10	1
k11	5
k12	1
k13	5
k14	1
k15	4
k16	2
k17	4
k18	1
k19	5
k20	1
k21	2
k22	1

3.8.8 Supplemental Information: Initial conditions for simulated and experimental data

Data was gathered or generated using various initial conditions were used to improve the coverage of the reaction space. Initial experiments were originally done for seven pairs of glycine and alanine adding to 0.1 M total amino acid.

Glycine	0.1 M	0.09 M	0.075 M	0.05 M	0.025 M	0.01 M	0 M
Alanine	0 M	0.01 M	0.025 M	0.05 M	0.075 M	0.09 M	0.1 M

Table 3.4 Supplemental Information: The first 7 initial conditions used to generate the basic data set for design of experiments (DoE). Each column represents one initial condition.

Additional experimental starting points were selected by the DoE algorithm. Three experimental conditions were chosen from the top 20 experiments suggested by the algorithm. The selected experiments were as follows: for experimental data, the first round added 0.1 M G and 0.1 M A, 0.1 M G and 0.07 M A, and 0.05 M G and 0.1 M A. For the second round using experimental data, the tests added started from 0.09 M G and 0.09 M A, 0.07 M G and 0.1 M A, and 0.1 M G and 0.05 M A. For the first round of simulated data with 15% noise, the tests added were 0.1 M G and 0.1 M A, 0.07 M G and 0.1 M A, and 0.1 M G and 0.07 M A. For the second round, 0.1 M G and 0.09 M A, 0.09 M G and 0.1 M A, and 0.09 M G and 0.09 M A were added. The DoE algorithm consistently suggested tests near the upper limit of the amino acid concentrations we included as possible tests, so for the arbitrary experiments to compare the DoE selections to were 0.05 M G and 0 M Ala, 0.025 M G and 0.025 M A, and 0 M G and 0.05 M A.

For the simulated data, an additional set of five standard experiment were added – 0.06 M G and 0.06 M A, 0.07 M G and 0.07 M A, 0.08 M G and 0.08 M A, 0.09 M G and 0.09 M A, and 0.1 M G and 0.1 M A, These were included based loosely on the suggestions from the DoE model, which seemed to imply the need for better coverage in the region with high amino acid concentrations. Data files that included more data points than the original seven plus, these additional five, and the suggestions selected from the DoE process included randomly generated initial points where the glycine and alanine values each had to be within [0, 0.1].

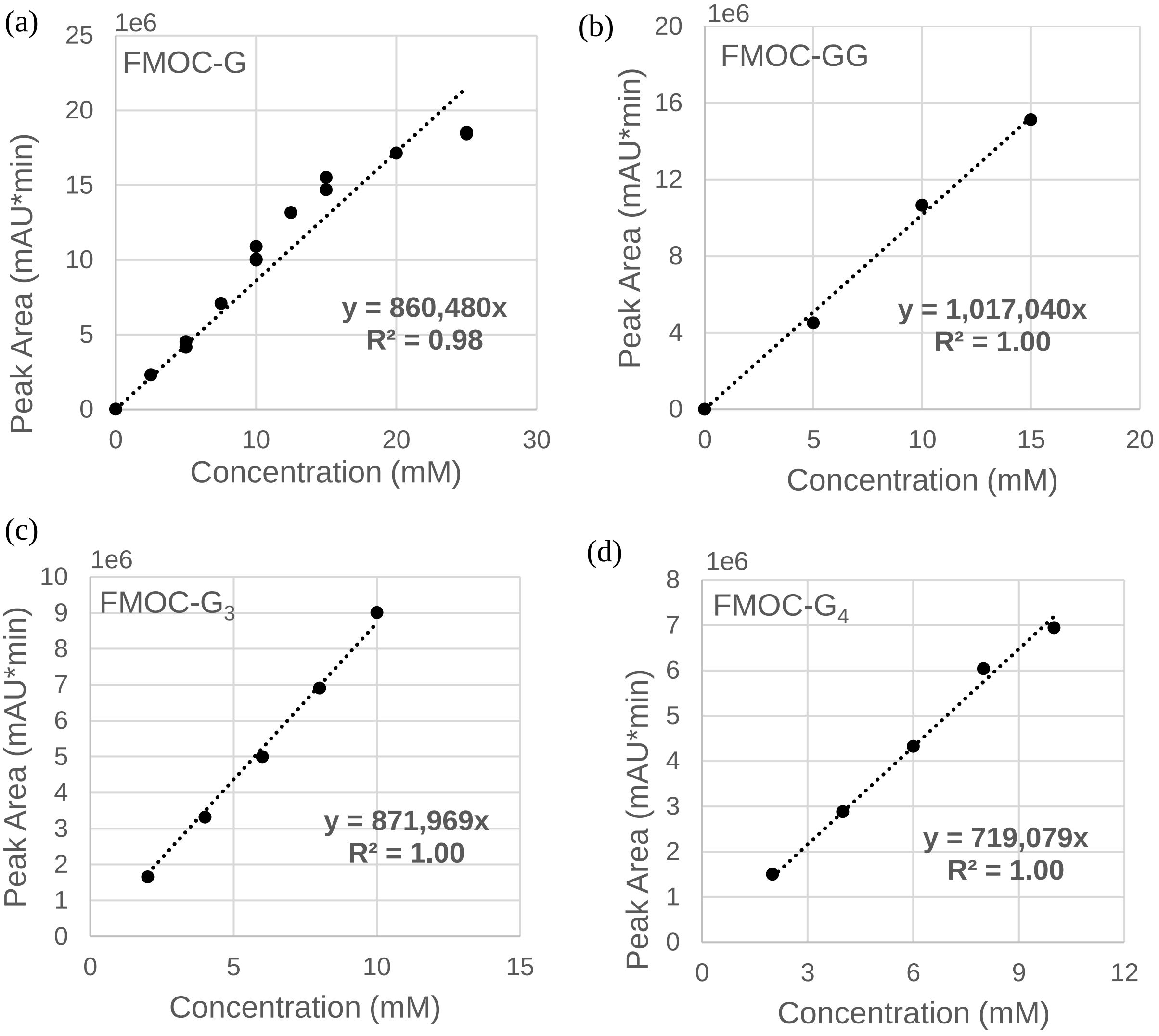
3.8.9 Supplemental Information: HPLC Calibration Curves

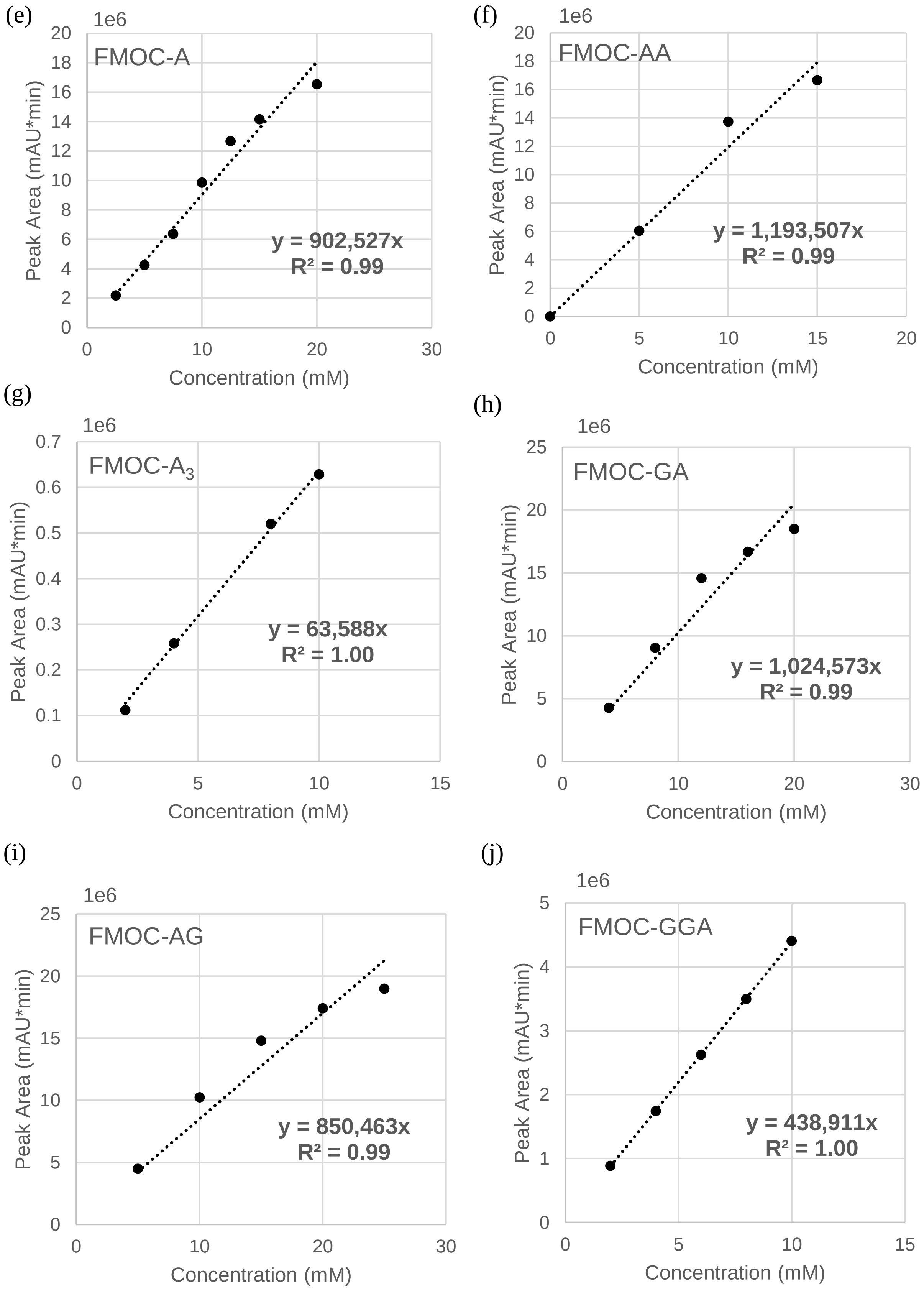
Calibration curves based on laboratory standards for the various species measured in this paper. Samples were diluted to ¼ their original concentration prior to analysis to bring them within range of the calibration curves. For curves that appeared to saturate at higher concentrations, only the relatively linear ($R^2 \geq 0.98$) region of each set of measurements was used to determine the calibration curve. The concentrations found experimentally fall within these regions.

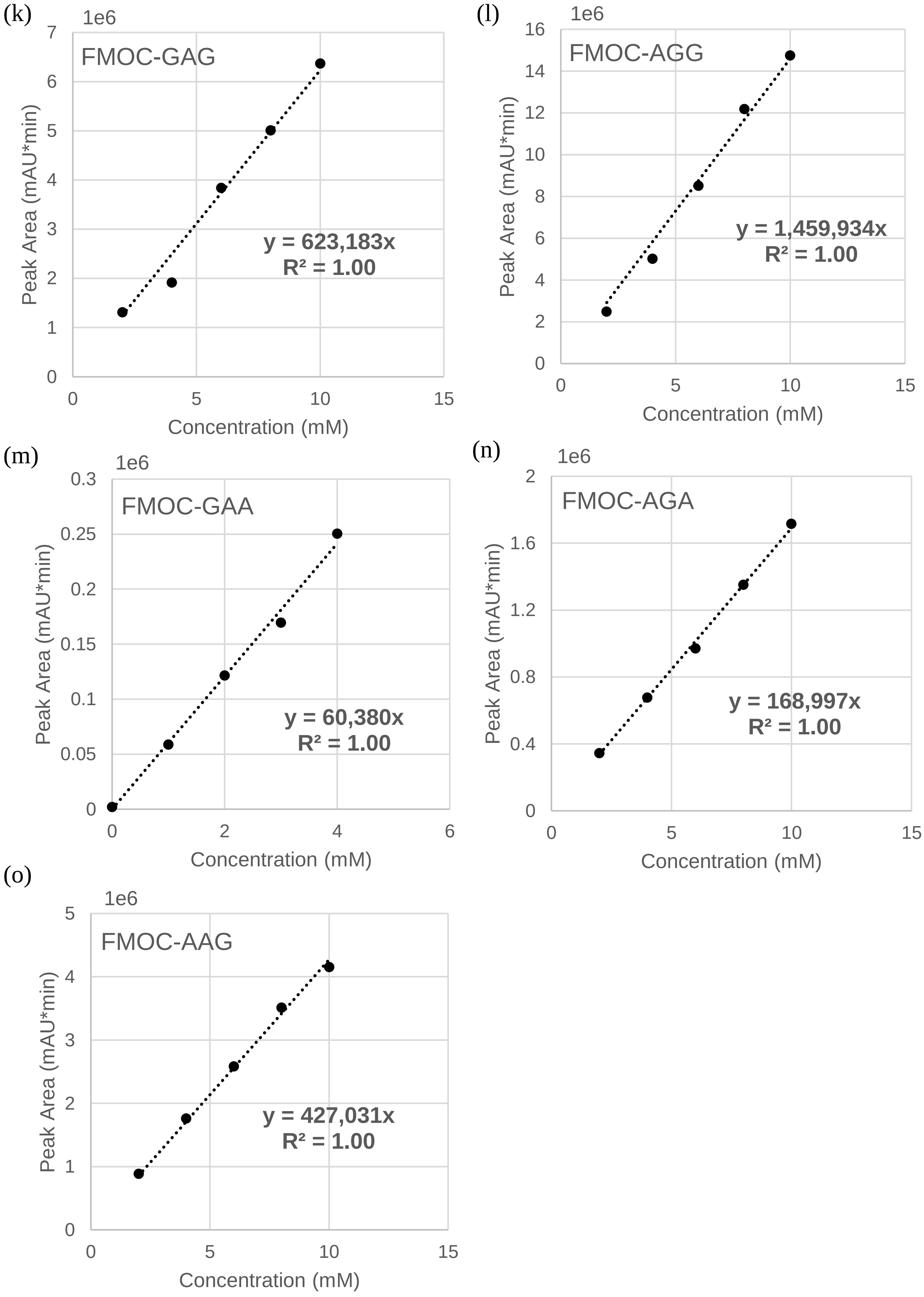
We found that most isomers had nearly identical retention times, which is a known challenge of using HPLC to identify short peptides. To calculate the concentrations of overlapping isomer species, we used the average of the calibration curves of each species, essentially assuming that all isomers were present in equal concentrations. This is a probable source of experimental error since some isomers may be more abundant than others and they can have significantly different calibration factors. However, given the similarity of the experimental data and the simulated data, which had no concept of underlying separate isomers, we believe it is unlikely that a different method of calculating the merged isomer concentrations would substantially change any of the results discussed.

The only case where there was significant peak overlap outside of isomers was between AAA and GA/AG. The qualitative behavior of this peak was what would be expected from GA/AG, maximizing when the initial condition was half glycine and half alanine and decreasing the more the initial condition was biased towards one amino acid or the other. Since AAA was a low abundance peak even in samples containing only alanine, we chose to ignore its contribution to this peak. The concentration of AAA was set to zero for every experiment except those starting with only alanine. The alanine-only experiment was sufficient for the model to predict non-zero concentrations of AAA, but we expect these predictions to have high uncertainty since they are based on incomplete data.

Figure 3.15 Supplemental Information: Calibration curves for glycine and alanine peptide species. Calibration curves are shown for (a) G, (b) GG, (c) G_3, (d) G_4, (e) A, (f) AA, (g) A_3, (h), GA, (i) AG, (j) GGA, (k) GAG, (l) AGG, (m) GAA, (n) AGA, (o) AAG




(e)
1e6
FMOC-A
Peak Area (mAU*min)
Concentration (mM)
y = 902,527x
R² = 0.99
0
2
4
6
8
10
12
14
16
18
20
0
10
20
30
(f)
1e6
FMOC-AA
Peak Area (mAU*min)
Concentration (mM)
y = 1,193,507x
R² = 0.99
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
(g)
1e6
FMOC-A3
Peak Area (mAU*min)
Concentration (mM)
y = 63,588x
R² = 1.00
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
5
10
15
(h)
1e6
FMOC-GA
Peak Area (mAU*min)
Concentration (mM)
y = 1,024,573x
R² = 0.99
0
5
10
15
20
25
0
10
20
30
(i)
1e6
FMOC-AG
Peak Area (mAU*min)
Concentration (mM)
y = 850,463x
R² = 0.99
0
5
10
15
20
25
0
10
20
30
(j)
1e6
FMOC-GGA
Peak Area (mAU*min)
Concentration (mM)
y = 438,911x
R² = 1.00
0
1
2
3
4
5
0
5
10
15


(k)
1e6
FMOC-GAG
y = 623,183x
R² = 1.00
Peak Area (mAU*min)
Concentration (mM)
0
1
2
3
4
5
6
7
0
5
10
15
(l)
1e6
FMOC-AGG
y = 1,459,934x
R² = 1.00
Peak Area (mAU*min)
Concentration (mM)
0
2
4
6
8
10
12
14
16
0
5
10
15
(m)
1e6
FMOC-GAA
y = 60,380x
R² = 1.00
Peak Area (mAU*min)
Concentration (mM)
0
0.05
0.1
0.15
0.2
0.25
0.3
0
2
4
6
(n)
1e6
FMOC-AGA
y = 168,997x
R² = 1.00
Peak Area (mAU*min)
Concentration (mM)
0
0.4
0.8
1.2
1.6
2
0
5
10
15
(o)
1e6
FMOC-AAG
y = 427,031x
R² = 1.00
Peak Area (mAU*min)
Concentration (mM)
0
1
2
3
4
5
0
5
10
15

3.9 References

Apgar, J. F., Witmer, D. K., White, F. M., & Tidor, B. (2010). Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular BioSystems*, *6*(10), 1890–1900. https://doi.org/10.1039/b918098b

Apri, M., de Gee, M., & Molenaar, J. (2012). Complexity reduction preserving dynamical behavior of biochemical networks. *Journal of Theoretical Biology*, *304*, 16–26. https://doi.org/10.1016/j.jtbi.2012.03.019

Bard, Y. (1974). *Nonlinear parameter estimation* (No. 04; QA276. 8, B3.).

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849-15854. https://doi.org/10.1073/pnas.1903070116

Box, G. E. P. and Wilson K. B. (1951) On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society*, 13, 1-45. https://doi.org/10.1111/j.2517-6161.1951.tb00067.x

Brown, K. S., Hill, C. C., Calero, G. A., Myers, C. R., Lee, K. H., Sethna, J. P., & Cerione, R. A. (2004). The statistical mechanics of complex signaling networks: Nerve growth factor signaling. *Physical Biology*, *1*(3), 184–195. https://doi.org/10.1088/1478-3967/1/3/006

Brown, K. S., & Sethna, J. P. (2003). Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *68*(2), 9. https://doi.org/10.1103/PhysRevE.68.021904

Chaloner, K., & Verdinelli, I. (1995). Bayesian Experimental Design: A Review. *Statistical Science*, *10*(3), 273–304.

Chis, O.-T., Banga, J. R., & Balsa-Canto, E. (2014). *Sloppy models can be identifiable*. 1–35. http://arxiv.org/abs/1403.1417

Cobelli, C., & DiStefano, J. J. (1980). Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *The American Journal of Physiology*, *239*(1), 7–24. https://doi.org/10.1152/ajpregu.1980.239.1.R7

Coveney, P. V., Swadling, J. B., Wattis, J. A. D., & Greenwell, H. C. (2012). Theory, modelling and simulation in origins of life studies. *Chemical Society Reviews*, *41*(16), 5430–5446. https://doi.org/10.1039/c2cs35018a

Daniels, B. C., Chen, Y. J., Sethna, J. P., Gutenkunst, R. N., & Myers, C. R. (2008). Sloppiness, robustness, and evolvability in systems biology. *Current Opinion in Biotechnology*, *19*(4), 389–395. https://doi.org/10.1016/j.copbio.2008.06.008

Fisher, R. A. (1937). *Design of experiments.* Edinburgh: Oliver and Boyd, 1935.

Frenkel-Pinter, M., Samanta, M., Ashkenasy, G., & Leman, L. J. (2020). Prebiotic Peptides: Molecular Hubs in the Origin of Life. *Chemical Reviews*, *120*(11), 4707–4765. https://doi.org/10.1021/acs.chemrev.9b00664

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, *20*(6), 1981–1996. https://doi.org/10.1093/bib/bby063

Goldman, A. D., Bernhard, T. M., Dolzhenko, E., & Landweber, L. F. (2013). LUCApedia: A database for the study of ancient life. *Nucleic Acids Research*, *41*(D1), 1079–1082. https://doi.org/10.1093/nar/gks1217

Gutenkunst, R. N., Casey, F. P., Waterfall, J. J., Myers, C. R., & Sethna, J. P. (2007). Extracting falsifiable predictions from sloppy models. *Annals of the New York Academy of Sciences*, *1115*, 203–211. https://doi.org/10.1196/annals.1407.003

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., & Sethna, J. P. (2007a). Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, *3*(10), 1871–1878. https://doi.org/10.1371/journal.pcbi.0030189

Hettling, H., & van Beek, J. H. G. M. (2011). Analyzing the functional properties of the creatine kinase system with multiscale “sloppy” modeling. *PLoS Computational Biology*, *7*(8), 11–16. https://doi.org/10.1371/journal.pcbi.1002130

Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K., & Ghadiri, M. R. (1996). A self-replicating peptide. *Nature*, *382*, 525–528.

Ludlow, R. F., & Otto, S. (2008). Systems chemistry. *Chemical Society Reviews*, *37*(1), 101–108. https://doi.org/10.1039/b611921m

Jagadeesan, P., Raman, K., & Tangirala, A. K. (2022). Bayesian Optimal Experiment Design for Sloppy Systems. *IFAC-PapersOnLine*, *55*(23), 121-126. https://doi.org/10.1016/j.ifacol.2023.01.026

Jain, A., McPhee, S. A., Wang, T., Nair, M. N., Kroiss, D., Jia, T. Z., & Ulijn, R. V. (2022). Tractable molecular adaptation patterns in a designed complex peptide system. *Chem*, *8*(7), 1894–1905. https://doi.org/10.1016/j.chempr.2022.03.016

Johnson, E. O., & Hung, D. T. (2019). A Point of Inflection and Reflection on Systems Chemical Biology. *ACS Chemical Biology*, *14*(12), 2497–2511. https://doi.org/10.1021/acschembio.9b00714

Ma, S., & Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics*, *12*(6), 714–722. https://doi.org/10.1093/bib/bbq090

Ma, Y., Dixit, V., Innes, M. J., Guo, X., & Rackauckas, C. (2021). A Comparison of Automatic Differentiation and Continuous Sensitivity Analysis for Derivatives of Differential Equation Solutions. *2021 IEEE High Performance Extreme Computing Conference, HPEC 2021*, *2*, 1–9. https://doi.org/10.1109/HPEC49654.2021.9622796

Maity, S., Ottelé, J., Santiago, G. M., Frederix, P. W. J. M., Kroon, P., Markovitch, O., … Roos, W. H. (2020). Caught in the Act: Mechanistic Insight into Supramolecular Polymerization-Driven Self-Replication from Real-Time Visualization. *Journal of the American Chemical Society*, *142*(32), 13709–13717. https://doi.org/10.1021/jacs.0c02635

Maiwald, T., Hass, H., Steiert, B., Vanlier, J., Engesser, R., Raue, A., Kipkeew, F., Bock, H. H., Kaschek, D., Kreutz, C., & Timmer, J. (2016). Driving the model to its limit: Profile likelihood based model reduction. *PLoS ONE*, *11*(9), 1–18. https://doi.org/10.1371/journal.pone.0162366

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. (2014). Ester formation and hydrolysis during wet-dry cycles: Generation of far-from-equilibrium polymers in a model prebiotic reaction. *Macromolecules*, *47*(4), 1334–1343. https://doi.org/10.1021/ma402256d

Maria, G. (2004). A review of algorithms and trends in kinetic model identification for chemical and biochemical systems. *Chemical and Biochemical Engineering Quarterly*, *18*(3), 195-222.

Monsalve-Bravo, G. M., Lawson, B. A. J., Drovandi, C., Burrage, K., Brown, K. S., Baker, C. M., Vollert, S. A., Mengersen, K., McDonald-Madden, E., & Adams, M. P. (2022). Analysis of sloppiness in model simulations: Unveiling parameter uncertainty when mathematical models are fitted to data. *Science Advances*, *8*(38). https://doi.org/10.1126/sciadv.abm5952

Napier, J., & Yin, J. (2006). Formation of peptides in the dry state. *Peptides*, *27*(4), 607–610. https://doi.org/10.1016/j.peptides.2005.07.015

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167-256. https://doi.org/10.1137/S003614450342480

Nghe, P., Hordijk, W., Kauffman, S. A., Walker, S. I., Schmidt, F. J., Kemble, H., Yeates, J. A. M., & Lehman, N. (2015). Prebiotic network evolution: Six key parameters. *Molecular BioSystems*, *11*(12), 3206–3217. https://doi.org/10.1039/c5mb00593k

Orgel, L. E. (2010). The origin of life: A review of facts and speculation. *The Nature of Life: Classical and Contemporary Perspectives from Philosophy and Science*, *0004*(December), 121–128. https://doi.org/10.1017/CBO9780511730191.012

Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B. D., Theis, F. J., Klingmüller, U., & Timmer, J. (2013). Lessons Learned from Quantitative Dynamical Modeling in Systems Biology. *PLoS ONE*, *8*(9). https://doi.org/10.1371/journal.pone.0074335

Rodriguez-Fernandez, M., Mendes, P., & Banga, J. R. (2006). A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *BioSystems*, *83*(2-3 SPEC. ISS.), 248–265. https://doi.org/10.1016/j.biosystems.2005.06.016

Rout, S. K., Rhyner, D., Riek, R., & Greenwald, J. (2022). Prebiotically Plausible Autocatalytic Peptide Amyloids. *Chemistry - A European Journal*, *28*(3). https://doi.org/10.1002/chem.202103841

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv Preprint arXiv:1609.04747*. http://arxiv.org/abs/1609.04747

Ruiz-Mirazo, K., Briones, C., & De La Escosura, A. (2014). Prebiotic systems chemistry: New perspectives for the origins of life. *Chemical Reviews*, *114*(1), 285–366. https://doi.org/10.1021/cr2004844

Sakata, K., Kitadai, N., & Yokoyama, T. (2010). Effects of pH and temperature on dimerization rate of glycine: Evaluation of favorable environmental conditions for chemical evolution of life. *Geochimica et Cosmochimica Acta*, *74*(23), 6841–6851. https://doi.org/10.1016/j.gca.2010.08.032

Schwartz, A. W. (2007). Intractable mixtures and the origin of life. *Chemistry and Biodiversity*, *4*(4), 656–664. https://doi.org/10.1002/cbdv.200790056

Serov, N. Y., Shtyrlin, V. G., & Khayarov, K. R. (2020). The kinetics and mechanisms of reactions in the flow systems glycine–sodium trimetaphosphate–imidazoles: the crucial role of imidazoles in prebiotic peptide syntheses. *Amino Acids*, *52*(5), 811–821. https://doi.org/10.1007/s00726-020-02854-z

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (1971). Cytoscape: A Software Environment for Integrated Models. *Genome Research*, *13*(22), 426. https://doi.org/10.1101/gr.1239303.metabolite

Shevlin, M. (2017). Practical high-throughput experimentation for chemists. *ACS medicinal chemistry letters*, *8*(6), 601-607. https://doi.org/10.1021/acsmedchemlett.7b00165

Sibilska, I., Feng, Y., Li, L., & Yin, J. (2018). Trimetaphosphate Activates Prebiotic Peptide Synthesis across a Wide Range of Temperature and pH. *Origins of Life and Evolution of Biospheres*, *48*(3), 277–287. https://doi.org/10.1007/s11084-018-9564-7

Surman, A. J., Rodriguez-Garcia, M., Abul-Haija, Y. M., Cooper, G. J. T., Gromski, P. S., Turk-MacLeod, R., Mullin, M., Mathis, C., Walker, S. I., & Cronin, L. (2019). Environmental control programs the emergence of distinct functional ensembles from unconstrained chemical reactions. *Proceedings of the National Academy of Sciences*, *116*(12), 5387–5392. https://doi.org/10.1073/pnas.1813987116

Thompson, J. C., Zavala, V. M., & Venturelli, O. S. (2022). Integrating a tailored recurrent neural network with Bayesian experimental design to optimize microbial community functions. *BioRxiv*, 1–24.

Transtrum, M. K., Machta, B. B., Brown, K. S., Daniels, B. C., Myers, C. R., & Sethna, J. P. (2015). Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *Journal of Chemical Physics*, *143*(1). https://doi.org/10.1063/1.4923066

Transtrum, M. K., & Qiu, P. (2012). Optimal experiment selection for parameter estimation in biological differential equation models. *BMC bioinformatics*, *13*(1), 1-12. https://doi.org/10.1186/1471-2105-13-181

Vanlier, J., Tiemann, C. A., Hilbers, P. A. J., & van Riel, N. A. W. (2013). Parameter uncertainty in biochemical models described by ordinary differential equations. *Mathematical Biosciences*, *246*(2), 305–314. https://doi.org/10.1016/j.mbs.2013.03.006

Varfolomeev, S. D., & Lushchekina, S. V. (2014). Prebiotic synthesis and selection of macromolecules: Thermal cycling as a condition for synthesis and combinatorial selection. *Geochemistry International*, *52*(13), 1197–1206. https://doi.org/10.1134/S0016702914130102

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

von Kiedrowski, G. (1986). A Self-Replicating Hexdeoxynucleotide. *Angewandte Chemie - International Edition*, *25*(10), 932–935. https://doi.org/10.1002/anie.198609322

Wasserman, L. (2004). *All of statistics: a concise course in statistical inference* (Vol. 26, p. 86). New York: Springer.

Waterfall, J. J., Casey, F. P., Gutenkunst, R. N., Brown, K. S., Myers, C. R., Brouwer, P. W., Elser, V., & Sethna, J. P. (2006). Sloppy-model universality class and the vandermonde matrix. *Physical Review Letters*, *97*(15), 1–4. https://doi.org/10.1103/PhysRevLett.97.150601

White, A., Tolman, M., Thames, H. D., Withers, H. R., Mason, K. A., & Transtrum, M. K. (2016). The Limitations of Model-Based Experimental Design and Parameter Estimation in Sloppy Systems. *PLoS Computational Biology*, *12*(12), 1–26. https://doi.org/10.1371/journal.pcbi.1005227

Wieland, F. G., Hauber, A. L., Rosenblatt, M., Tönsing, C., & Timmer, J. (2021). On structural and practical identifiability. *Current Opinion in Systems Biology*, *25*, 60–69. https://doi.org/10.1016/j.coisb.2021.03.005

Yu, S. S., Krishnamurthy, R., Fernández, F. M., Hud, N. V., Schork, F. J., & Grover, M. A. (2016). Kinetics of prebiotic depsipeptide formation from the ester-amide exchange reaction. *Physical Chemistry Chemical Physics*, *18*(41), 28441–28450. https://doi.org/10.1039/c6cp05527c

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286. https://doi.org/10.1198/106186006X113430

4. ENHANCEMENT OF PREBIOTIC PEPTIDE FORMATION IN CYCLIC ENVRIONMENTS

Hayley Boigenzahn, Praful Gagrani, and John Yin

Authors' Contributions

All experimental data and model figures were prepared and analyzed by **HB**. The mathematical formulation of overshoot was written by PG. The manuscript was drafted by **HB** with input from PG and JY, and all authors contributed to the subsequent editing.

This paper has been submitted to Origins of Life and Evolution of Biospheres.

4.1 Abstract

The dynamic behaviors of prebiotic reaction networks may be critically important to understanding how larger biopolymers could emerge, despite being unfavorable to form in water. We focus on understanding the dynamics of simple systems, prior to the emergence of replication mechanisms, and what role they may have played in biopolymer formation. We specifically consider the dynamics in cyclic environments using both model and experimental data.

Cyclic environmental conditions prevent a system from reaching thermodynamic equilibrium, improving the chance of observing interesting kinetic behaviors. We used an approximate kinetic model to simulate the dynamics of trimetaphosphate (TP) activated peptide formation from glycine in cyclic wet-dry conditions. The model predicts that environmental cycling allows trimer and tetramer peptides to sustain concentrations above the predicted fixed points of the model due to overshoot, a dynamic phenomenon. Our experiments demonstrate that oscillatory environments can shift product distributions in favor of longer peptides. However, experimental validation of certain behaviors in the kinetic model is challenging, considering that open systems with cyclic environmental conditions break many of the common assumptions in classical chemical kinetics. Overall, our results suggest that the dynamics of simple peptide reaction networks in cyclic environments may have been important for the formation of longer polymers on the early Earth. Similar phenomena may have also contributed to the emergence of reaction networks with product distributions determined not by thermodynamics, but rather by kinetics.

4.2 Introduction

How organic monomers developed into polymers capable of replication, metabolism, and other key behaviors of life remains unknown. Repeated reactions, such as those produced from cycling a sample between wet and dry conditions, can promote the formation of more complex molecules (Lahav & Chang 1976; Ross & Deamer 2016), but also tend to produce intractable tars that would be unlikely to support continuous replication (Shapiro 2000). Life-like behavior probably emerged in at least partially open systems that were able to exchange materials and energy with their environment (Wagner et al. 2019; Baum 2018). Open systems allow fresh reactants to be supplied while removing side products from previous reactions. Removing products from a system can also decrease the overall system complexity by favoring products that form quickly, potentially avoiding the formation of tars and supporting the proliferation of catalytic or autocatalytic reactions (Martin & Horvath 2013; Colón-Santos et al. 2019). Experiments with open systems have been performed using various mixtures of organic molecules (Lahav et al. 1978; Maio et al. 2021; Bartolucci et al. 2022), and some have suggested the possible emergence of function, though the details of those functions remain ambiguous (Doran et al. 2019; Vincent et al. 2019). Some dynamic phenomena, such as sustained oscillations, are only thermodynamically possible in open systems (Wagner et al. 2019).

Another significant feature of open systems in the origins of life is that they can remain away from thermodynamic equilibrium indefinitely. One of the hallmarks of a life-like system is that it should remain out of equilibrium with its environment, meaning that life almost certainly originated in far-from-equilibrium conditions (Eigen & Schuster 1977; Pross 2003; Pascal et al. 2013; Mamajanov et al. 2014). The kinetic behavior of systems which are far from equilibrium may have provided the driving force necessary for the emergence of organization from an unordered system (Prigogine 1978; Astumian 2019). Chemical reaction networks in far-from-equilibrium conditions may exhibit overshoot, a dynamic phenomenon in which a species passes through its equilibrium point one or more times before actually reaching it (Jia et al. 2014). Overshoot is a kinetically driven phenomena and has been correlated with the ability of biochemical systems to recover their original state after a perturbation (Jia & Qian 2016; Ma et al. 2009).

Nonlinear dynamics can emerge in relatively simple chemical reaction networks and lead to complex behaviors in open or partially open systems (Epstein & Showalter 1996), but these behaviors have not been extensively explored to evaluate their significance to the chemical origins of life. Computational models of replenished systems have been investigated, but many of the existing models either assume the presence of an autocatalytic network, are deliberately vague about the identities of the molecules and mechanisms involved, or both (Kindermann 2005; Walker et al. 2012; Wynveen et al. 2014; Peng et al. 2020). The inclusion of chemicals in an autocatalytic set, sometimes referred to as chemical replicators, increases the diversity of potential dynamics in a model system, but such models do not help explain how chemical replicators arose in the first place.

Here we will discuss the possible significance of open-system dynamics that can arise in chemical networks prior to the emergence of catalysis or autocatalysis. We examined the kinetic behavior in fed batch systems of glycine polymerizing into oligoglycine through a combination of wet-dry cycles and activation by trimetaphosphate (TP), an inorganic phosphate that significantly enhances peptide bond formation across a wide range of environmental conditions (Sibilska et al. 2018). Using parameters fitted from experimental data, a simplified mass-action based ordinary differential equation (ODE) model predicts the emergence of overshoot in this open system. We explored the effect that cycling between two different reaction mechanisms has on the system dynamics and compared model results to experiments. The experimental results support the idea that oscillations can favor certain products over others. However, our work highlights the difficulty of evaluating kinetics in open systems, specifically where reactions are driven by drying into the solid state.

4.3 Methods

4.3.1 Experimental methods

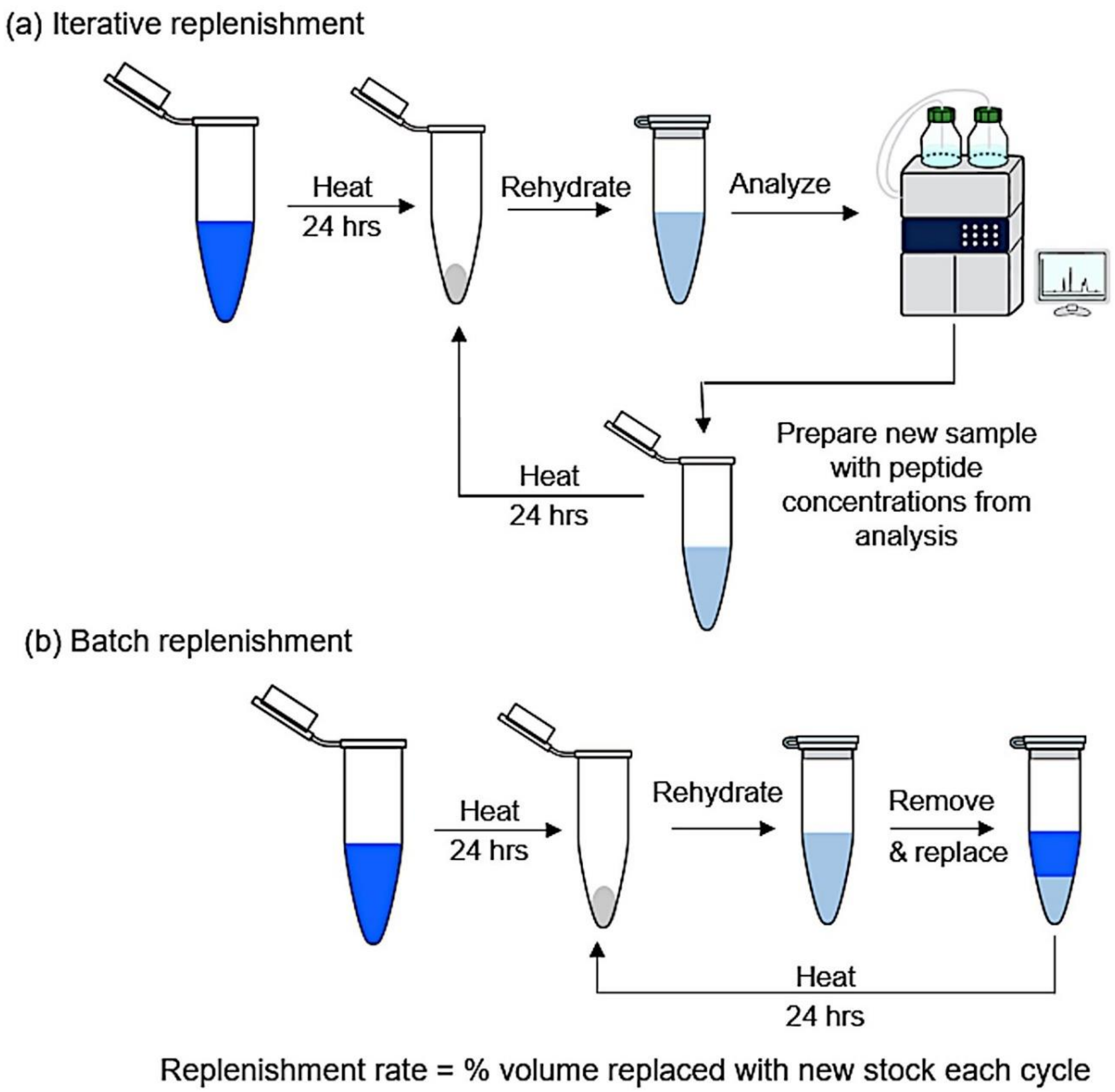

(a) Iterative replenishment
Heat
24 hrs
Rehydrate
Analyze
Heat
24 hrs
Prepare new sample with peptide concentrations from analysis
(b) Batch replenishment
Heat
24 hrs
Rehydrate
Remove & replace
Heat
24 hrs
Replenishment rate = % volume replaced with new stock each cycle

Figure 4.1 Methods of experimental replenishment.

We performed several cyclic, multi-day experiments to compare with model predictions. Samples were prepared as 1 mL samples containing 0.1 M glycine, 0.1 M TP, and 0.15 M NaOH, and heated with the caps open at 90^{o}C for 24-hour intervals. Since reaction rate parameters depend on the environment, we kept the initial conditions of each cycle as close to those conditions as possible; this required keeping the trimetaphosphate and base conditions close to their initial conditions, since they are not explicitly included in the model. To maintain these conditions without causing continuous accumulation of TP, base, or their byproducts, we used an iterative strategy, recreating samples using peptide standards to match the concentrations measured after each cycle (Fig. 4.1a). This ensured complete replacement of TP and base. To keep the amino acid to TP ratio constant, we also adjusted the total glycine species balance with each cycle by adding glycine monomers to compensate for any glycine lost to side products, such as 2,5-diketopiperazine or glycine oligomers with a length of seven or above, which we did not quantify. Although this may underestimate the significance of side-products, this experimental setup adheres most rigorously to the assumptions of the model.

Compartmentalization or adsorption to solid surfaces can allow some molecules to be diluted less rapidly than others in an open system, but complete removal of waste and replacement of activating molecules, as occurs in iterative replenishment, is probably too idealized to be physically realistic in an origins of life context. Therefore, we also performed a set of experiments using batch replenishment, in which the reaction products from one cycle were directly transferred into the subsequent cycle (Fig. 4.1b). This process is a discontinuous analog of what would occur in a continuously stirred tank reactor (CSTR), where side-products from reactions can build up over time. This approach to replenishment is a more realistic representation of flow in open systems insofar as we might expect all the molecules to be equally affected by the flow rate.

For the batch replenishment experiments, all samples were initially 1 mL samples and contained 0.1 M glycine, 0.1 M TP, and 0.15 M NaOH. Samples were heated at 90^{o}C with the caps open for 24 hours to allow them to dry, then rehydrated with 1 mL water and vortexed until the solid was fully dissolved. The term ‘replenishment rate’ refers to the percentage of fresh material that is replaced in each subsequent cycle; for example, a 75% replenishment rate indicates that 250 µL of the dissolved reaction product is transferred to the next generation, and 750 µL of a solution of 0.1 M glycine, 0.1 M TP, and 0.15 M NaOH solution is added. In practice, the glycine, TP, and NaOH solutions were stored separately and mixed only during the preparation of the subsequent cycles to prevent them from reacting in storage. We compared the results of multiple replenishment rates – 50%, 75%, and 90%.

All samples were analyzed using fluorenylmethyloxycarbonyl chloride (FMOC) derivatization and high-performance liquid chromatography (UV-HPLC) for improved retention and quantitation, as in our previous work (Boigenzahn & Yin 2022; Boigenzahn et al. 2023). FMOC derivatization was used to improve the retention time and signal strength of the peptide analytes. For the FMOC derivatization procedure, 25 µL of sample was diluted with 75 µL milliQ water and mixed with 100 µL 0.1 M sodium tetraborate, which acted as a buffer. Finally, 800 µL 0.0391 M FMOC dissolved in acetone was added to each sample, equating to 25% excess FMOC to possible amino acid.

Between cyclic transfers and prior to derivatization, samples were vortexed at maximum speed until there were no visible solids remaining (Pulsing Vortex Mixer, Fischer Scientific), usually approximately two minutes. pH was measured using an Apera Instruments PH8500-MS

Portable pH microelectrode. For the batch replenishment samples, the pH at the start of each heating cycle was measured using duplicate samples prepared from the material remaining after the volume needed for HPLC analysis and transfer to subsequent generations was removed. Since a small amount of sample tends to stay on the pH probe after each measurement due to surface tension, the use of duplicate samples prevented effects due to volume changes.

Derivatized samples were analyzed using a Shimadzu Nexera HPLC with a C-18 column (Phenomenex Aeris XB-C18, 250mm x 4.6mm, 3.6µL) and quantified using calibration curves generated from laboratory standards (Section 4.6.1). All analysis was carried out using Solvent A: milliQ water with 0.01 v/v trifluoroacetic acid (TFA) and Solvent B: acetonitrile with 0.01% v/v TFA. Replenished samples with serial transfer were analyzed using the following gradient: 0-3 min, 30% B; 3-16 min, 30-100% B, 16-19 min, 100% B; 19-21 min, 100-30% B; 21-24 min, 30% B. Iteratively recreated samples were analyzed using the following gradient for improved resolution of G_6: 0-3 min, 30% B; 3-16 min, 30-70% B, 16-19 min, 70% B; 19-21 min, 70-30% B; 21-24 min, 30% B. The solvent flow rate was 1 mL/min. Peak integration was performed in LabSolutions with the ‘Drift’ parameter set to 1000.

4.3.2 Materials

All materials were of analytical grade purity and used without additional purification. Materials were obtained from the following suppliers: Glycine and triglycine (G_3) from Alfa Aesar (Heysham, LA3 2XY, England), diglycine, TP and TFA from Sigma-Aldrich (St. Louis, MO, USA), tetraglycine (G_4), pentaglycine (G_5), and hexaglycine (G_6) from Bachem (Torrance, CA, USA), acetone and sodium hydroxide from Fisher Scientific (Fair Lawn, NJ, USA), acetonitrile from VWR International (Radnor, PA, USA), and FMOC from Creosalus (Louisville, KY, USA).

4.3.3 Model formulation

Peptide concentrations were modeled using mass-action ordinary differential equations (ODEs) for four reversible reactions that describe how a single amino acid (glycine) forms peptides of up to four amino acids in length (Eqn. 4.1-4.4). Although we observed and quantified G_5 and G_6 in our experiments, their empirical concentrations are low during the time frame that used to estimate parameters for the network, so we chose to exclude them and instead focus on species that could be clearly quantified (Boigenzahn & Yin 2022). Parameters were estimated from experimental data according to the procedure outlined in (Boigenzahn et al. 2023). As in Boigenzahn et al. (2023), the effects of volume change due to drying and the presence of any reaction intermediates were excluded for simplicity.

$$G + G \rightleftharpoons G_2 \tag{4.1}$$

$$G + G_2 \rightleftharpoons G_3 \tag{4.2}$$

$$G + G_3 \rightleftharpoons G_4 \tag{4.3}$$

$$G_2 + G_2 \rightleftharpoons G_4 \tag{4.4}$$

In Boigenzahn & Yin (2022), we described two mechanisms of trimetaphosphate (TP) activated peptide formation, which are differentially dependent on pH and water activity. We fit two sets of parameters to the network in Scheme 1, representing the two different reaction mechanisms. Briefly, Mechanism 1 occurs in alkaline conditions and proceeds readily in water (Chung et al. 1971), and Mechanism 2 occurs in neutral conditions and proceeds as the sample approaches the solid state due to drying (Yamanaka et al. 1988). Mechanism 1 lowers the pH of the samples when it occurs, so samples containing amino acids and TP dried in alkaline conditions naturally transition from Mechanism 1 to Mechanism 2. This transition can be used to explore how environmental oscillations can alter kinetic behavior within a reaction network. The code and data for this manuscript can be found at https://github.com/haboigenzahn/Cyclic-Environments.

4.3.4 Model formulation

To estimate parameters for each mechanism, we selected experimental data in which one mechanism strongly dominated over the other. The parameters for Mechanism 1 were generated based on 8-hour time courses obtained from 0.1 M glycine or 0.05 M diglycine (GG) heated at 90°C with 0.1 M trimetaphosphate and 0.15 M base without drying (closed caps). Additional data came from the first four hours of equivalent experiments that were allowed to dry (open caps), since during the early stages of drying there is still bulk water present such that Mechanism 1 dominates (Boigenzahn & Yin 2022). Finally, to predict behavior in second and subsequent drying cycles while minimizing the number of experiments required, we included peptide concentrations from the first four hours of only the first day of iterative experiments. We deliberately kept the time span of the training data relatively short so that later experimental cycles could be used to assess the predictive accuracy of the model.

Mechanism 2 was estimated from 0.1 M glycine and 0.1 M TP with and without 0.15 M base, starting after the samples had been drying for 4 hours and including the next 20 hours. Additional results from samples of 0.1 M glycine dried with 0.1 M TP and 0.15 M base were prepared and measured after 4 hours of heating in 24-hour intervals for up to 72 hours to improve the long-term estimates of Mechanism 2. Finally, several measurements from the first day of the iterative experiments were also taken after the first four hours of heating and included in the training data for Mechanism 2. The specific fitted parameter values for Mechanism 1 and Mechanism 2 are detailed in the supplemental information (Section 4.6.2).

4.4 Results and discussion

To evaluate the accuracy of the model predictions using separate parameters for Mechanism 1 and Mechanism 2, we compared the predictions to peptide concentration profiles measured in Boigenzahn & Yin (2022) (Fig. 4.2). We used the assumption that Mechanism 1 occurred during the first 4 hours and Mechanism 2 accounted for the following 20 hours after verifying the timing of this transition by testing parameters fit to alternative timings (Section 4.6.3). Although the model tended to slightly underestimate peptide formation, we determined that this could be interpreted as a conservative estimate of the rates of polymer formation and was acceptable for the behaviors we were aiming to explore.

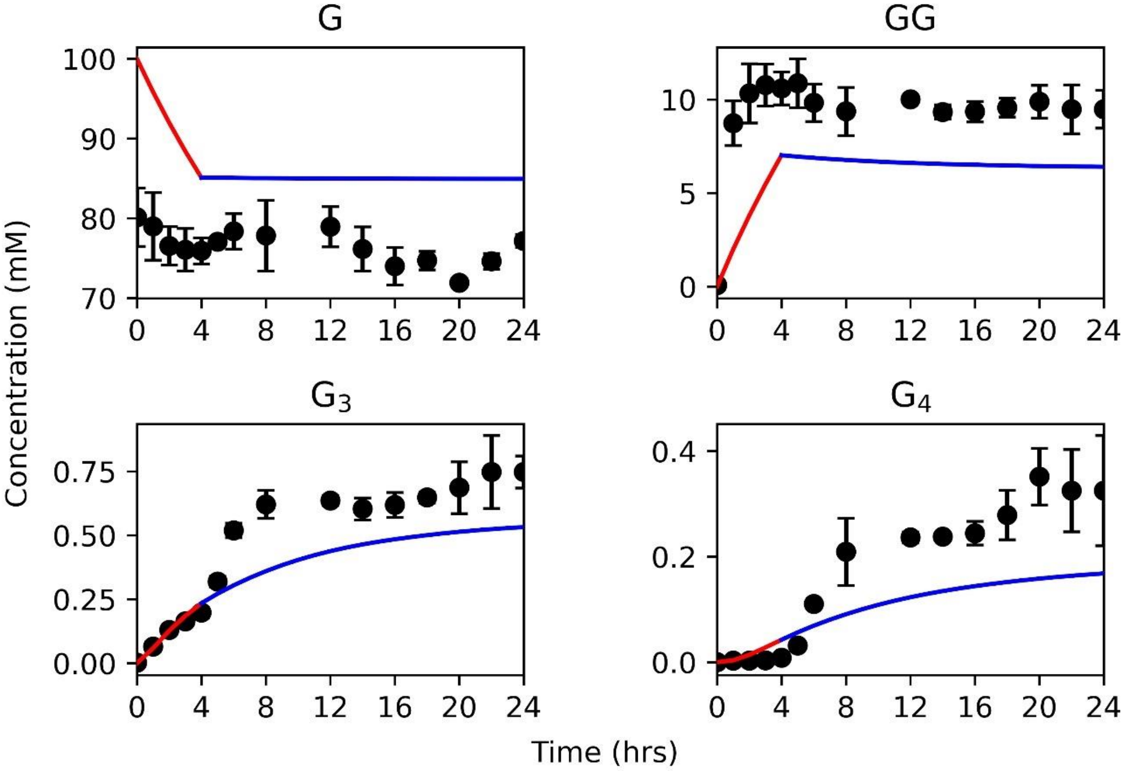


Figure 4.2 Comparison of two-step model to analogous experimental data. Experimental data is from Boigenzahn & Yin (2022).

We used these parameters to model peptide formation in a cyclic environment and observed that for certain combinations of parameters and initial conditions, the average yield of the longer glycine polymers exceeded the 'thermodynamic equilibrium' predicted by the fitted parameters of either mechanism (Fig. 4.3). Each parameterized model inherently has a fixed-point attractor, or a state which the system approaches as time approaches infinity. In an ideal mass action kinetics system, the attractor is equivalent to the concentrations of the species at thermodynamic equilibrium. However, since the reference model is not a true mass-action representation of the system, we will henceforth use the term 'attractor' when referring to the steady state solutions to the model system; this allows us to distinguish the mathematical predictions from the simplified model from concepts of true thermodynamic equilibrium.

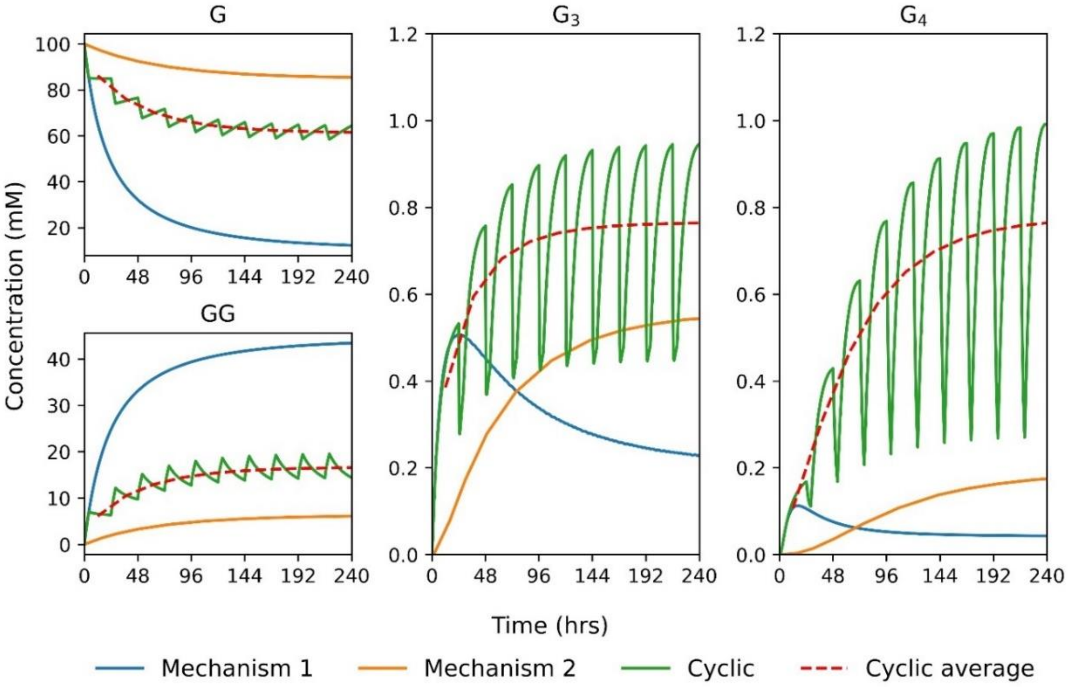


Figure 4.3 Cycling conditions where the average yield of G_3 and G_4 at steady state exceeds the value of the attractor due to overshoot. The model is initialized with 0.1 M G. The approach of each mechanism to its attractor in a non-cyclic system is shown for comparison. The cyclic trajectory alternates between 4 hours of Mechanism 1 and 20 hours of Mechanism 2. Parameters are estimated from experimental data.

Many simple chemical reactions exhibit monotonic kinetics in all species. In such systems, the maximum yield of each species is limited by the level of the attractor. In the system predicted by our model, G_3 and G_4 surpass the attractor of both Mechanism 1 and Mechanism 2 due to a dynamic phenomenon called overshoot. Overshoot is a non-monotonic behavior which occurs in a wide variety of systems, including biological and man-made control networks (Ogata 1995; Jia et al. 2014; Chen et al. 2016). Systems with overshooting kinetics will eventually return to the attractor, however, in this example the environmental cycling between the two mechanisms causes overshoot to occur repeatedly.

Overshoot occurs when a system passes through the coordinates of its stable point in any dimension before the stable point is reached. For a mathematical definition of overshoot, consider a dynamical system whose state at time t is given by a vector of real numbers $x(t)$ such that

$$\frac{dx}{dt} = F(x) \tag{4.5}$$

where $F(x)$ is the rate vector for the system. Assume the system has at least one stable fixed-point $\underline{x}$, such that

$$F(x) = 0. \tag{4.6}$$

Starting from an initial condition in the vicinity of the fixed point $x(0)$, the system overshoots in dimension j if there exists a time $t \in (0, \infty)$ for which the state at t is further away from but lies in the same direction of the fixed point as the initial condition,

$$|x_j(0) - \underline{x}_j| < |x_j(t) - \underline{x}_j| \tag{4.7}$$

$$sign\left(x_j(0) - \underline{x}_j\right) = sign\left(x_j(t) - \underline{x}_j\right)$$

This characterization provides a straightforward algorithm for determining if a dynamical system is capable of overshooting in any direction around a stable fixed point. First, discretely sample a small sphere centered around the stable fixed point. Then from each point on the sphere, integrate the dynamical system backwards in time and track the distance of the state from the fixed point. If the maximum distance is not monotonic in any dimension, then the system overshoots in that direction.

It should also be noted that whether overshoot occurs in an experimental system can be restricted by thermodynamics. Reactions occurring very close to equilibrium will never overshoot, since doing so would violate the Second Law of Thermodynamics. Overshoot requires a system to be under kinetic control, which is more likely in systems which are further from equilibrium (Epstein & Showalter 1996).

Overshoot depends on the kinetic reaction network and initial conditions of a system. In the example with alternating mechanisms shown in Figure 4.3, Mechanism 2 approaches its attractor monotonically for a system initialized with pure glycine monomer, but the formation of diglycine by Mechanism 1 creates initial conditions where Mechanism 2 tends to overshoot G_3 and G_4. One factor is that Mechanism 1 hydrolyzes some of the longer polymers produced, but also generates more GG. Repeated cycles of Mechanism 1 and Mechanism 2 in a 24-hour pattern created a dynamic steady state for which the average yields of G_3 and G_4 exceed the predicted attractor values for either mechanism. We also found that for some initial conditions, this behavior was possible using a single reaction mechanism with batch replenishment. This demonstrates that these dynamics can occur within a single reaction mechanism, though there still needs to be oscillations (Fig. 4.10).

To understand why it is significant that the longer species exceed the system attractors, it is useful to revisit the perspective of thermodynamic equilibrium. The formation of longer peptides in water is thermodynamically limited (Ross & Deamer 2016). Allowing samples to dry favors polymerization, but the lack of molecular mobility in the dry state limits the potential for long-term reactions. Overshoot provides an explanation for how systems subjected to wet-dry cycling drying could not only form and maintain populations of longer polymers, but could also exhibit selectivity since the distribution of polymers formed depends on the reaction kinetics and the timing of the environmental cycles. Species that form quickly can more easily recover from dilution or hydrolysis. In many cases there may be an ‘optimal’ cycle timing, or resonance time, for each species, which maximizes the yield of a species of interest (Haugerud et al. 2023).

Repeated overshoot can allow a system to remain away from its predicted attractor for long periods of time, and may help provide the dynamic, non-equilibrium conditions that have been proposed as being critical for the origin of life (Pross 2011; Pascal et al. 2013). Therefore, overshoot and far-from-equilibrium conditions may be mutually reinforcing: overshoot pushes systems away from their attractor and alters system compositions, creating the potential for the system to experience overshoot again. These paired behaviors could begin to move the system towards a regime of kinetic control, in which product compositions are influenced more heavily by reaction rates than by product energy levels. Autocatalysis, one of the essential features of a life-like system, is an example of a behavior which is strongly under kinetic control (Pross 2003).

Overshoot has been observed in multiple experimental systems containing designed peptide replicators (Dadon et al. 2015; Miao et al. 2021). Similar dynamics have also been linked to the emergence of biochemical adaptation, which refers to the ability of a system to return to its original state after an environmental perturbation (Jia et al. 2014; Ma et al. 2009; François & Siggia 2008). It is interesting that our simple system of just four reversible reactions shows qualitatively similar behavior to these much more complex cases.

4.4.1 Experimental comparison

Since the kinetic parameters estimated for Mechanism 1 and Mechanism 2 were primarily based on short term data, we wanted to determine how accurately the model captured the long-term behavior of the system. Therefore, we performed experiments to evaluate the capacity of the kinetic model to make predictions about the yields in multi-day experiments. We sought to identify potential signs of nonlinear dynamics, which may include features like non-monotonic concentration profiles, sigmoidal growth, or evidence of multiple steady states.

Due to the various simplifications included in the model formulation, we used the iterative replenishment approach to replicate the model assumptions as closely as possible. While the method is somewhat contrived, it has the advantage of eliminating as many known confounding variables as possible to facilitate the comparison of the model and the experimental data. Variables not included in the model that might impact the reactivity of the system included the concentrations of TP, base, and orthophosphate side products.

There was reasonably good agreement between the model predictions and the experimental results, especially the final yields of the peptides (Fig. 4.4). Significantly more G_4 forms during the first few cycles than the model predicts (Fig. 4.4d), but the model underestimating the yield of G_4 is consistent with the results of the training data (Fig. 4.2). However, there are some discrepancies between the first few days of the model trajectories and the experimental results. These discrepancies could be the result of inaccuracies in parameter estimation, experimental variance, or a consequence of confounding variables such as those caused by drying or the formation of longer polymers in the system.

One noteworthy discrepancy is that the experimentally measured maximum yield of GG and G_3 occurs after two days, then the yield drops in subsequent days, which is not predicted by the model (Fig. 4.4b, c). These experimental trajectories are notable since a non-monotonic result would not be expected from systems with monotonic underlying kinetics. Non-monotonic trajectories like those observed in GG and G_3 can occur in species that are initially overshooting, but the overshoot is damped and decreases as the reaction cycles progress. However, damped overshoot behavior was not predicted by the model. Additionally, although GG and G_3 appear to slightly exceed their eventual final yields then gradually drop in concentration, this is not necessarily experimental evidence of overshoot, since their change relative to an attractor is not defined.

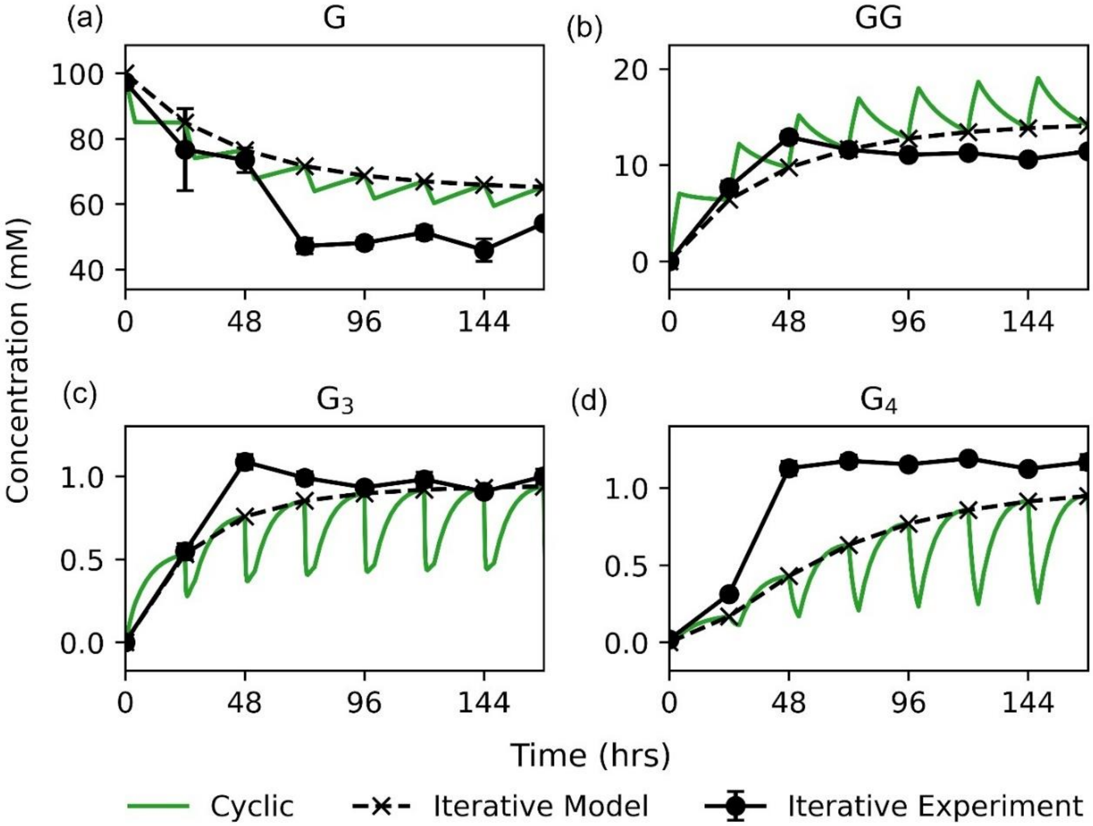


Figure 4.4 Comparison of experimental and model results for 7 days. Plots show concentrations for (a) glycine, (b) diglycine, (c) triglycine, and (d) tetraglycine. Experimental data was generated using the iterative approach. Models use an initial condition of 0.1 M Gly. Error bars represent the sample standard deviation of experimental triplicates.

We can evaluate the accuracy of the model and look for evidence of nonlinear dynamics, but we cannot confirm the existence or absence of overshoot in the experimental system. Maintaining repeated overshoot relative to a mathematically predicted attractor requires an open system, but in practice overshoot is not well defined for open systems, since there is no experimentally measurable point which is equivalent to the attractor. In closed systems, the attractor represents the thermodynamic equilibrium of the system, but open systems violate a key principle of equilibrium, namely that there cannot be any energy or mass exchanged with the environment.

Instead, the real systems reach a dynamic steady state, in which the overall species concentrations remain constant from cycle to cycle, but there are bonds which form and break during each cycle. This state may have been a precursor to the emergence of dynamic kinetic stability, a concept introduced by Pross to describe the behavior of systems of replicators which maintain their population while experiencing a continuous flux of energy (Pross 2009).

Ideally, evidence of overshoot might be found experimentally by allowing Mechanism 2 to continue and looking for evidence of non-monotonic behavior. However, simply allowing the system to continue heating indefinitely would not result in the continuation of Mechanism 2 with its estimated rate constants, since those parameters were primarily estimated using data from samples probably still contained some residual water. It is therefore unsurprising that experiments which were continuously heated did not show non-monotonic behavior (Fig. 4.5), since once the samples were completely dry they were unlikely to undergo any significant hydrolysis.

When comparing the steady states reached by the iteratively cycled experiments to those of non-cycled experiments, we found the iterative experiments had higher yields of G_4, but lower yields of G_3 (Fig. 4.5c, d). This supports the idea that cyclic environmental conditions can select for certain species over others. This type of selectivity is interesting because it is driven primarily by kinetics, with the actual thermodynamic stability of each species playing a secondary role. This type of behavior may contribute to the ability of some chemical reaction networks to move away from thermodynamic equilibrium within their local environment, even in the absence of chemical replicators. Further experiments using different replenishment procedures, such as varying the cycle timing to study how it affects selectivity, may produce valuable insights. It may also be possible to observe additional kinetic effects by diluting a species relative to the others with each cycle, creating artificial pressure against it.

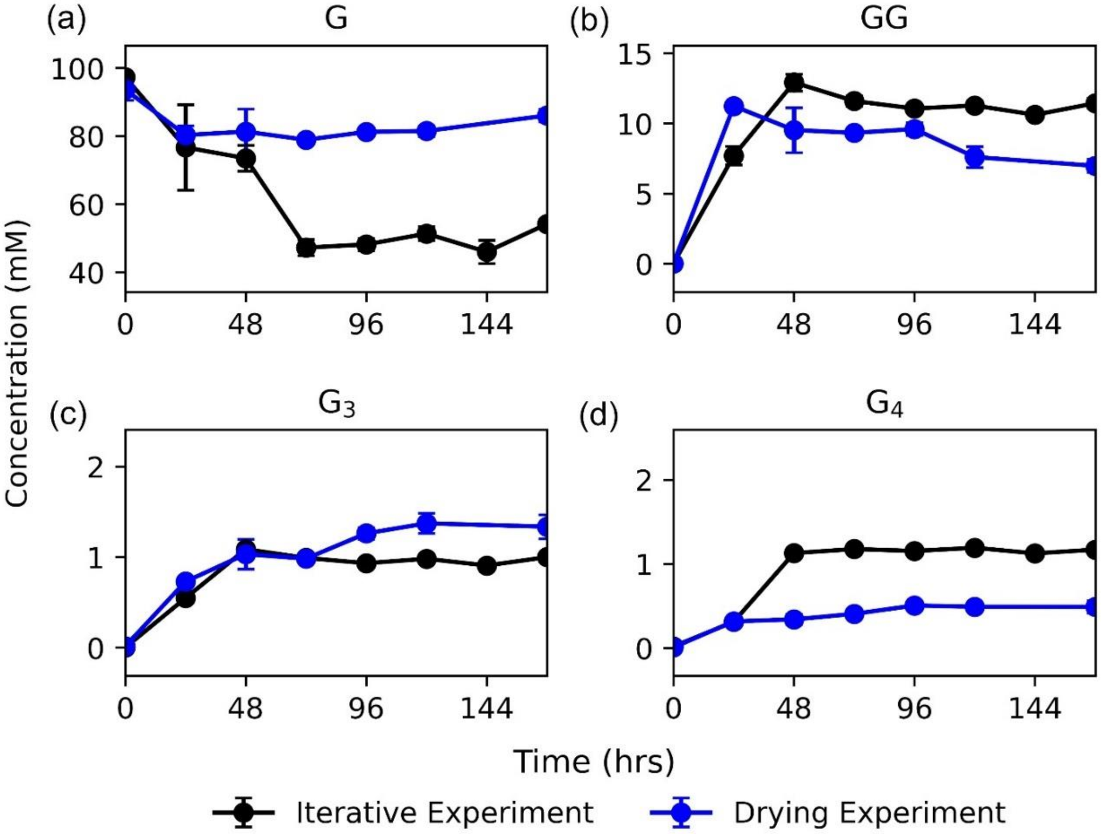


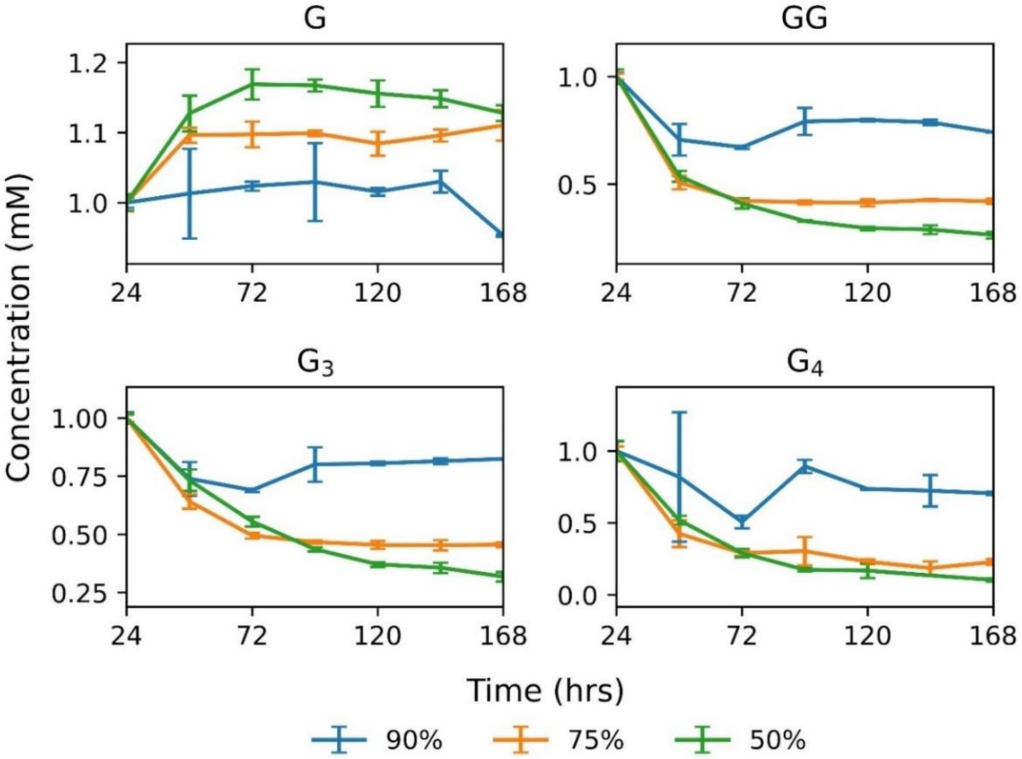
Figure 4.5 Replenishment affects species selectivity. Plots show concentrations for (a) glycine, (b) diglycine, (c) triglycine, and (d) tetraglycine. Samples carried out using iterative replenishment have higher yields of G_4, while samples that were continuously dried have higher yields of G_3. Samples were analyzed at the end of each 24-hour cycle.

4.4.2 Complete replenishment

Since the iteratively recreated experiments are not entirely physically realistic owing to the complete removal of waste and replacement of food without any dilution of the desired products, we also investigated the behavior of the system using replenishment conditions in which a portion of the dissolved products from the previous cycle are transferred and replaced with fresh reactants, as shown in Figure 4.1b. We examined three different replenishment rates, or three different fractions of products to be transferred from sample to sample and compared them to one another. This method creates a trade-off between the costs and benefits of dilution and replenishment. High replenishment rates have high dilution rates, so less of the product formed in previous cycles is transferred into subsequent generations. Low replenishment rates transfer more previously formed products, but have limited access to fresh reactants, which is detrimental when the reactants include activating agents, like TP, that are needed for reactions to occur. Since the model does not include the effects of reduced TP or base, it predicts higher peptide yields for lower replenishment rates (Figure 4.11).

Figure 4.6 Higher replenishment rates have higher peptide yields in replenished systems with serial transfer. Replenishment percentages indicate how much of the sample was replaced with fresh reactant mixture each day. Each data set is normalized relative to the results from the first day.

However, when we tried these tests experimentally, we found that the opposite was true; higher replenishment rates had higher peptide yields (Fig. 4.6). This implies that the yield per drying cycle is not limited by the availability of oligopeptides, but by the pH and concentration of TP. The difference in initial pH is particularly significant because it determines whether both reaction mechanisms occur during drying. There are two factors which change the pH of the samples between the initial setup and the subsequent replenishment steps. First, less base is added during the replenishment cycles than was added during the initial setup, since the samples are only replenished with a fraction of their original reactants. Second, some phosphate



byproducts from the TP reactions are transferred between cycles, which changes the buffering capacity of the samples. The initial pH of the samples was 10.78 ± 0.06, and the pH of the samples at the start of the first cycle was 9.72 ± 0.06 for 90% replenishment, 9.24 ± 0.15 for 75% replenishment and 7.01 ± 0.01 for 50% replenishment (Fig. 4.12). Mechanism 1 requires the deprotonated amine groups, and the pH at which most of the amine groups are deprotonated for glycine is 9.60. Therefore, we expect to see reduced GG formation when the initial pH of the solution is significantly below that point. This likely explains why samples with 75% replenishment have disproportionately lower GG formation than samples with 90% replenishment relative to the difference in their initial pH.

Although it has a less profound impact on the underlying reaction mechanism, peptide yields tend to decrease when the amino acid-to-TP ratio is less than equimolar (Section 4.6.7), so the availability of TP also contributes lowering peptide yields at lower replenishment rates. This is likely to pose a challenge in any systems which include activating agents that are consumed. It is still theoretically possible for these systems to accumulate higher peptide concentrations even when the activating agent is only partially replaced; for example, a system with slow reverse kinetics whose forward rate constants were minimally impacted by the presence of side products could potentially accumulate polymers despite having less available activating agent after the first cycle. However, in many cases the combination of decreased reactivity and sample dilution is difficult to overcome, which causes the yields of longer products at steady state to be lower than their yields after one reaction cycle.

Activating materials that act as catalysts, like solid surfaces or metal salts, may help maintain a more constant reactivity in replenished systems since they are not depleted after the first cycle (Bujdák et al. 1995; Erastova et al. 2017). However, even with reusable activating agents, recursive reactions can be inhibited by waste products that are never fully removed. Thus, there is significant interest in methods that selectively retain certain products while removing most other materials. Realistic scenarios that have been suggested to produce this behavior include absorption onto mineral surfaces or containment in coacervates (Bedoin et al. 2020; Fares et al. 2020). These scenarios allow cycling of energy sources and waste products without equal dilution of biopolymers.

Given the number of possible environmental and temporal conditions, thorough design of replenished experiments can be a difficult task. Moreover, the fact that kinetic reaction networks may be highly nonlinear means that relatively minor differences in experimental design may have significant consequences for the product distribution. Fully understanding the details of how members of the system interact with one another and with the environment is often challenging and may not currently be possible for some complex systems. Approximations of system behavior can be useful but need to be applied carefully. Despite these difficulties, it is worthwhile to develop an understanding of the dynamics of short monomers and their biopolymers undergoing hydrolysis and condensation reactions, as these interactions must have been involved in the origins of long, functional biopolymers such as enzymes.

4.5 Conclusion

The dynamics of chemical reaction networks prior to the emergence of replicating chemical systems have not been extensively explored but may provide valuable insights into how biopolymer systems begin to develop complex behaviors. The behaviors of open systems, which are consistently held away from equilibrium and therefore are heavily influenced by kinetics, are of particular interest. However, studying chemical reaction networks in an open system can be challenging because environmental changes can make the results more difficult to interpret and experimental variables such as replenishment method, replenishment rate, and cycle timing greatly expand the experimental parameter space. As we have shown, despite these challenges, interesting behaviors can be captured using simple experimental systems paired with kinetic models.

We found that a simple ODE model of glycine polymerization using parameters fit from experimental data displayed overshoot, a dynamic phenomenon in which a species passes through its attractor at least once before reaching it. Simulated environmental cycles that alternated between two reaction mechanisms resulted in yields of trimers and tetramers that were well above the calculated attractors of the two reaction mechanisms involved. Since equilibrium as represented by the attractors is not defined for an open system, we were unable to look for overshoot directly, but iterative experiments found relatively good agreement between the model and experimental results. Different experimental cycling conditions were able to select for different species, which is significant since it suggests the possibility of species selectivity. These findings illustrate that even simple reaction networks, when pushed by repeated environmental changes, can start showing kinetic control, which is an important feature of reactions in life-like systems. Future modeling and experimental studies of non-linear reaction systems in open environments, thus, seem critical to help us understand the origins of complex, life-like dynamics in chemical reaction networks.

4.6 Supplemental Information

4.6.1 Supplemental Information: Calibration curves

Figure 4.7 Supplemental Information: Calibration curves for the 30% B to 70% B gradient. (a) Glycine, (b) diglycine, (c) triglycine, (d) tetraglycine

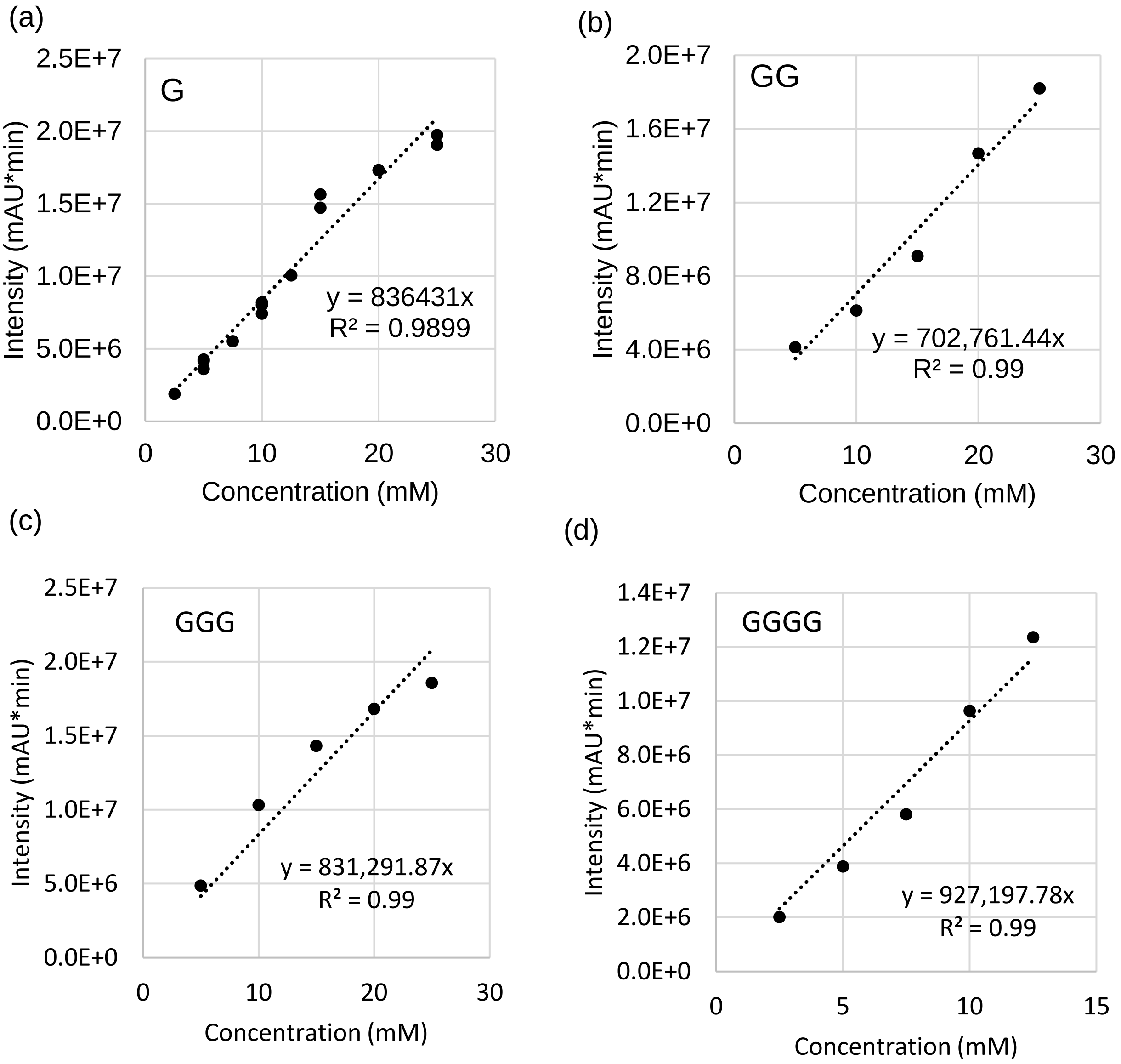
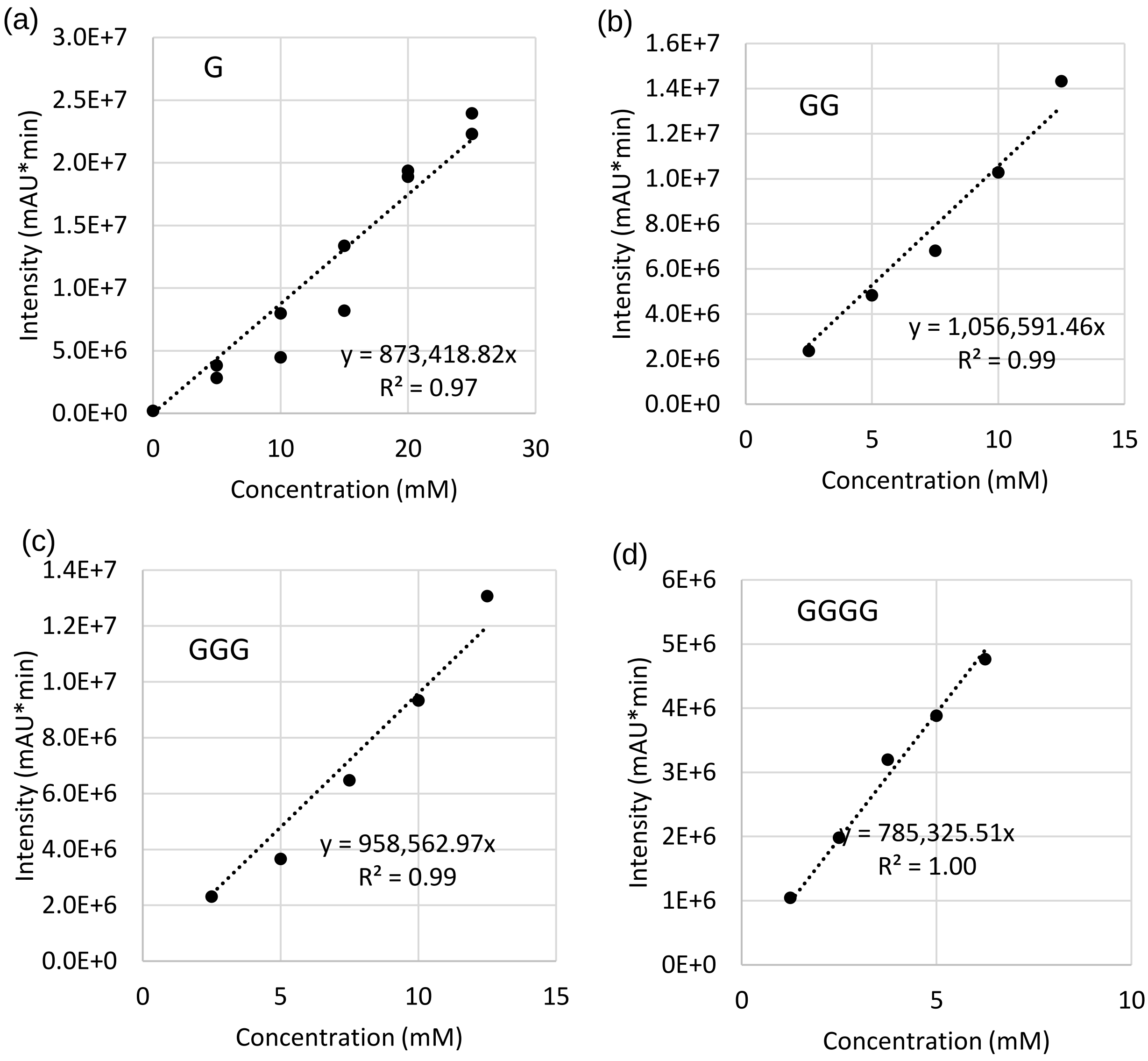


Figure 4.8 Supplemental Information: Calibration curves for the 30% B to 70% B gradient. (a) Glycine, (b) diglycine, (c) triglycine, (d) tetraglycine



4.6.2 Supplemental Information: Model equations and parameter details

The following are the ordinary differential equations for the oligoglycine network.

$$\partial[G]/\partial t = -2k_1\,[G]^2 + 2k_2\,[GG] - k_3\,[G][GG] + k_4\,[GGG] - k_5\,[G][GGG] + k_6\,[GGGG]$$

$$\frac{\partial[GG]}{\partial t} = k_1[G]^2 - k_2[GG] - k_3[G][GG] + k_4[GGG] - 2k_7[GG]^2 + 2k_8[GGGG]$$

$$\frac{\partial[GGG]}{\partial t} = k_3[G][GG] - k_4[GGG] - k_5[G][GGG] + k_6[GGGG]$$

$$\frac{\partial[GGGG]}{\partial t} = k_5[G][GGG] - k_6[GGGG] + k_7[GG]^2 - k_8[GGGG]$$

	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8
Mechanism 1	0.214	6.063e-4	1.478	3.423	1.602	0.0427	0.00734	0.386
Mechanism 2	0.0108	0.124	0.106	0.101	0	0	0.367	0.0776

Table 4.1 Supplemental Information: Parameters estimated for Mechanism 1 and Mechanism 2. The concentrations of the training data are molar (M) units, and the time scale is in hours.

Although the $G + G_3 \leftrightarrow G_4$ reactions happen to be eliminated in Mechanism 2, a smaller reaction network is not required for any of the behavior discussed in this paper. The system is still capable of overshoot when those parameters are manually changed to values comparable to the other parameters in the mechanism. Note that the existence of zero-valued parameters do not necessarily indicate that these reactions are not physically present, or even that these reactions are particularly slow in the real system. As discussed in Boigenzahn et al. (2023), the numerical values of these parameters do not have independent physical significance due to uncertainty that naturally arises during nonlinear parameter fitting.

4.6.3 Supplemental Information: Cycle timing of Mechanism 1 and Mechanism 2

New data files were prepared for each alternating timing by editing the existing training data for Mechanism 1 and Mechanism 2 to shift the training data to the appropriate time span. Parameters were estimated for each dataset, then the scaled mean squared error (MSE) of the model estimates was compared to the results of the glycine reactions in Boigenzahn & Yin (2022). The scaled MSE is calculated by normalizing all species to a range between 0 and 1 to ensure that the accuracy of the lower abundance species is significant. Splitting the data as 4 hours of Mechanism 1 and 20 hours of Mechanism 2 provides the lowest scaled MSE and corresponds to the most obvious visible shift in Gly_3 and Gly_4 formation.

Cycle Timing (hr+hr)	3+21	4+20	5+19	6+18
Scaled MSE	1.309	0.556	0.711	0.593

Table 4.2 Supplemental Information: Model fitting accuracy for various possible transitions between Mechanism 1 and Mechanism 2.

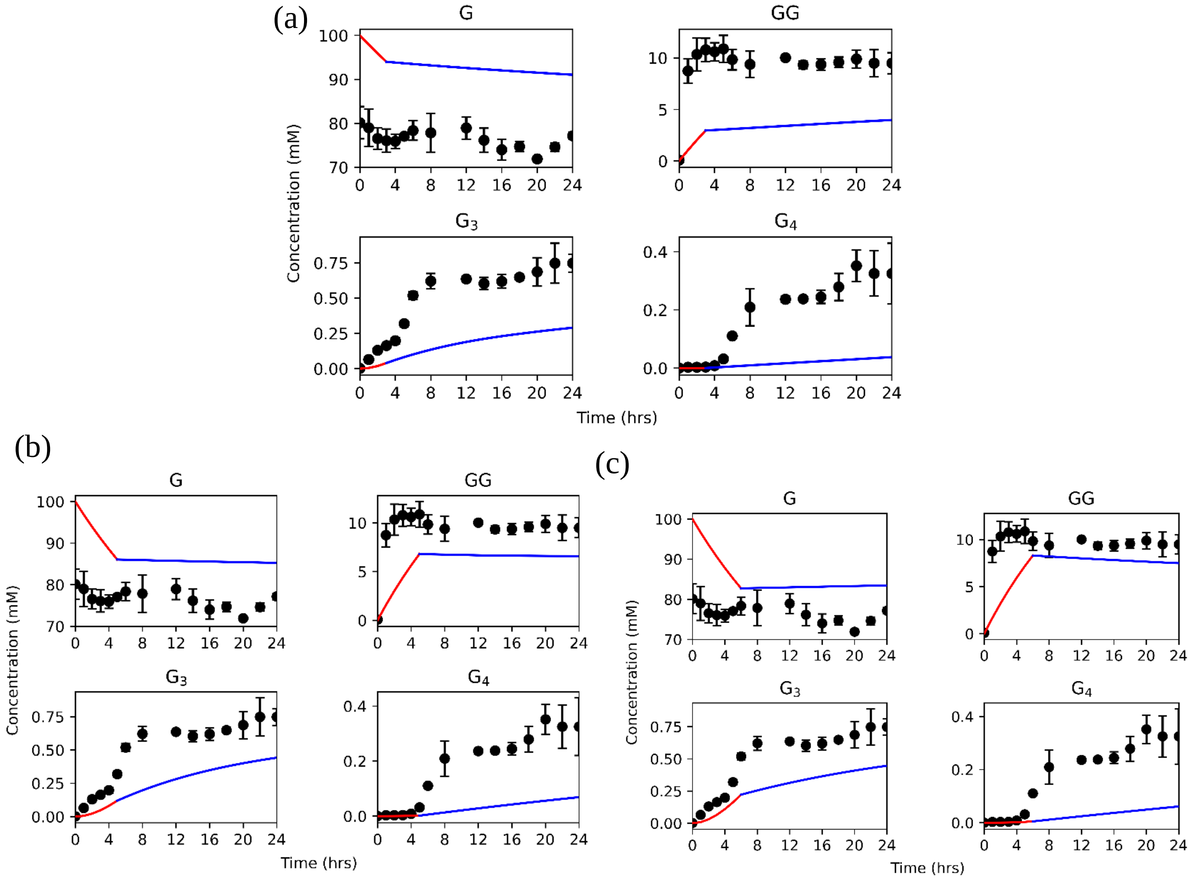


Figure 4.9 Supplemental Information: Model estimates for alternative timings for the transition between Mechanism 1 and Mechanism 2. Results shown for (a) 3 hrs Mechanism 1 plus 21 hours Mechanism 2 (b) 5 hours Mechanism 1 plus 19 hours Mechanism 2 (c) 6 hours Mechanism 1 plus 18 hours Mechanism 2.

4.6.4 Supplemental Information: Overshoot with a single reaction mechanism

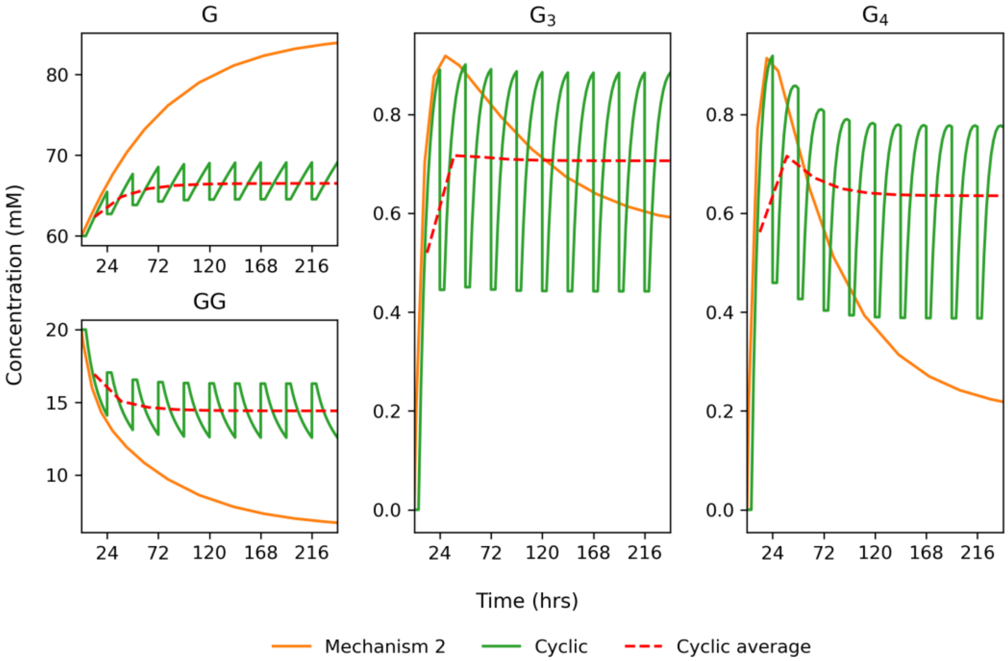


Figure 4.10 Supplemental Information: Overshoot can be driven by replenishment in a system with a single reaction mechanism. The model is initialized with 0.06 M G and 0.02 M GG, since overshoot does not occur starting from the pure monomer condition. Mechanism 2 proceeds during the cyclic trajectory for 24 hours between replenishment steps. The replenishment method is analogous to Fig. 1b occurs and occurs at a replenishment rate of 50%.

4.6.5 Supplemental Information: Model predictions of replenishment with serial dilution

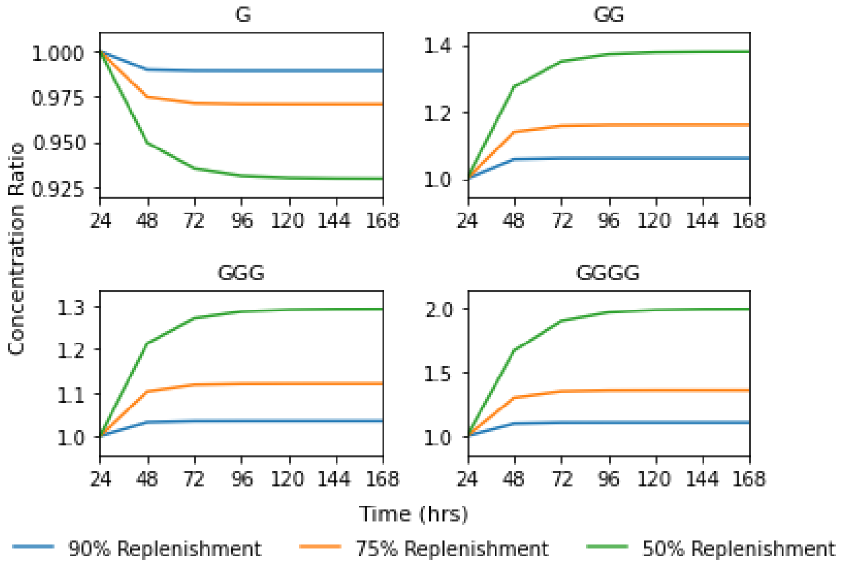


Figure 4.11 Supplemental Information: The model predicts higher peptide yield in systems with lower replenishment rates. Replenishment percentages indicate how much of the sample was replaced with fresh reactant mixture each day. Each dataset is normalized relative to the results from the first day.

4.6.6 Supplemental Information: Model predictions of replenishment with serial dilution

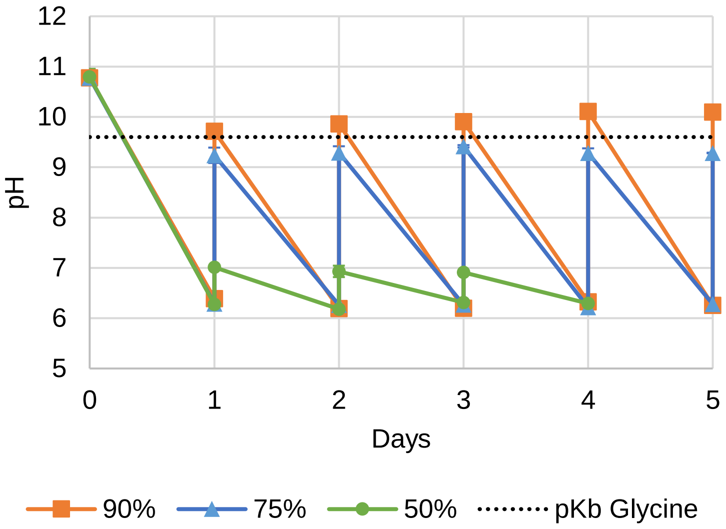


Figure 4.12 Supplemental Information: pH of batch replenishment samples relative to the pK$_b$ of glycine. pH was measured before and after replenishment. For samples plotted on the same day, the lower pHs were measured prior to replenishment, higher pHs were measured after replenishment. Error bars show the standard deviations of experimental triplicates.

4.6.7 Supplemental Information: Effect of TP availability on peptide bond formation

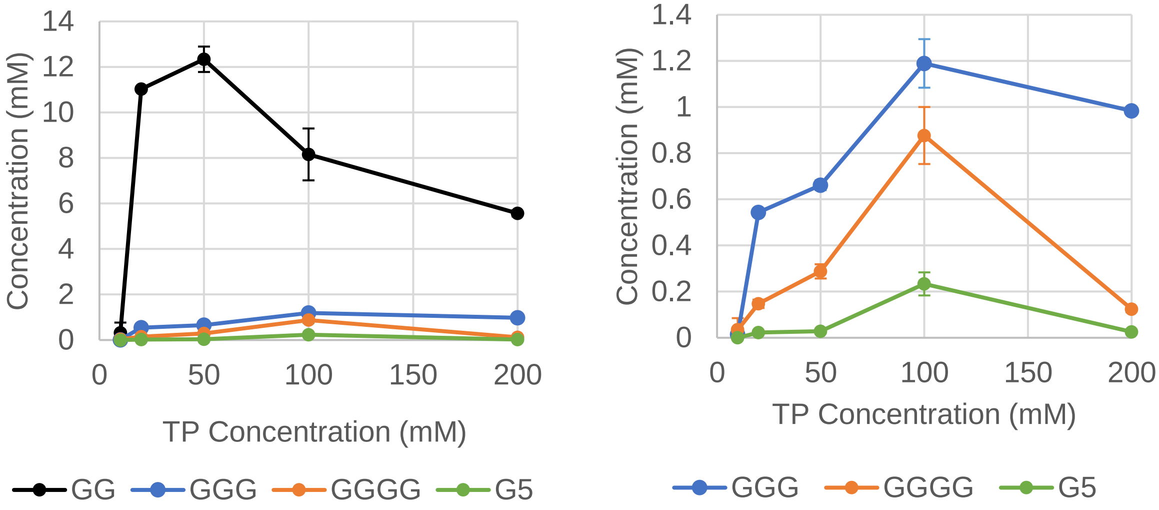


Figure 4.13 Supplemental Information: Lower TP concentrations result in lower peptide yields. Results are measured after 24 hours of heating samples with an open cap at 90°C. The initial glycine concentration was 100 mM (0.1 M). Error bars show the standard deviation of experimental triplicates.

Dimer yields were highest when there was a 2:1 ratio of amino acid to TP, but the other peptides have the highest yields when there is a 1:1 ratio of amino acid to TP. The decreased concentration of TP in the samples shown in Figure 4.6 probably contributed to the decreasing peptide concentration after the first day.

4.7 References

Astumian, R. D. (2019). Kinetic asymmetry allows macromolecular catalysts to drive an information ratchet. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-11402-7

Bartolucci, G., Serrão, A. C., Schwintek, P., Kühnlein, A., Rana, Y., Janto, P., Hofer, D., Mast, C. B., Braun, D., & Weber, C. A. (2022). *Selection of prebiotic oligonucleotides by cyclic phase separation*. http://arxiv.org/abs/2209.10672

Baum, D. A. (2018). The origin and early evolution of life in chemical composition space. *Journal of Theoretical Biology, 456,* 295–304. https://doi.org/10.1016/j.jtbi.2018.08.016

Bedoin, L., Alves, S., & Lambert, J. (2020). Origins of life and molecular information : selectivity in mineral surface induced prebiotic amino acids polymerization. *ACS Earth and Space Chemistry*. https://doi.org/10.1021/acsearthspacechem.0c00183

Boigenzahn, H., & Yin, J. (2022). Glycine to Oligoglycine via Sequential Trimetaphosphate Activation Steps in Drying Environments. *Origins of Life and Evolution of Biospheres, 52*(4), 249–261. https://doi.org/10.1007/s11084-022-09634-7

Boigenzahn, H., González, L., Thompson, J., Zavala, V. & Yin, J. (2023). Kinetic modeling and parameter estimation of a prebiotic peptide reaction network. *Journal of Molecular Evolution.* Manuscript in revision.

Bujdák, J., Faybíková, K., Eder, A., Yongyai, Y., & Rode, B. M. (1995). Peptide chain elongation: A possible role of montmorillonite in prebiotic synthesis of protein precursors. *Origins of Life and Evolution of the Biosphere, 25*(5), 431–441. https://doi.org/10.1007/BF01581994

Chen, X., Wang, Y., Feng, T., Yi, M., Zhang, X., & Zhou, D. (2016). The overshoot and phenotypic equilibrium in characterizing cancer dynamics of reversible phenotypic plasticity. *Journal of Theoretical Biology, 390,* 40–49. https://doi.org/10.1016/j.jtbi.2015.11.008

Chung, N. M., Lohrmann, R., Orgel, L. E., & Rabinowitz, J. (1971). The mechanism of the trimetaphosphate-induced peptide synthesis. *Tetrahedron, 27*(6), 1205–1210. https://doi.org/10.1016/S0040-4020(01)90868-3

Colón-Santos, S., Cooper, G. J. T., & Cronin, L. (2019). Taming the Combinatorial Explosion of the Formose Reaction via Recursion within Mineral Environments. *ChemSystemsChem, 1*(3), 1–5. https://doi.org/10.1002/syst.201900014

Dadon, Z., Wagner, N., Alasibi, S., Samiappan, M., Mukherjee, R., & Ashkenasy, G. (2015). Competition and cooperation in dynamic replication networks. *Chemistry - A European Journal, 21*(2), 648–654. https://doi.org/10.1002/chem.201405195

Doran, D., Abul-Haija, Y. M., & Cronin, L. (2019). Emergence of Function and Selection from Recursively Programmed Polymerisation Reactions in Mineral Environments. *Angewandte Chemie - International Edition, 58*(33), 11253–11256. https://doi.org/10.1002/anie.201902287

Eigen, M., & Schuster, P. (1977). The Hypercyde. *Naturwissenschaften, 64,* 541–565.

Epstein, I. R., & Showalter, K. (1996). Nonlinear chemical dynamics: Oscillations, patterns, and chaos. *Journal of Physical Chemistry,* 100(31), 13132–13147. https://doi.org/10.1021/jp953547m

Erastova, V., Degiacomi, M. T., Fraser, D. G., & Greenwell, H. C. (2017). Mineral surface chemistry control for origin of prebiotic peptides. *Nature Communications*, *8*(1), 1–9. https://doi.org/10.1038/s41467-017-02248-y

Fares, H. M., Marras, A. E., Ting, J. M., Tirrell, M. V., & Keating, C. D. (2020). Impact of wet-dry cycling on the phase behavior and compartmentalization properties of complex coacervates. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19184-z

François, P., & Siggia, E. D. (2008). A case study of evolutionary computation of biochemical adaptation. *Physical Biology*, *5*(2). https://doi.org/10.1088/1478-3975/5/2/026009

Haugerud, I. S., Jaiswal, P., & Weber, C. A. (2023). Wet-dry cycles away from equilibrium catalyse chemical reactions. *arXiv e-prints, arXiv-2304*. https://doi.org/10.48550/arXiv.2304.14442

Jia, C., Qian, M., & Jiang, D. (2014). Overshoot in biological systems modelled by Markov chains: a non-equilibrium dynamic phenomenon. *IET systems biology*, *8*(4), 138-145. https://doi.org/10.1049/iet-syb.2013.0050

Jia, C., & Qian, M. (2016). Nonequilibrium Enhances Adaptation Efficiency of Stochastic Biochemical Systems. *PLoS ONE*, 11(5), 1–19. https://doi.org/10.1371/journal.pone.0155838

Kindermann, M., Stahl, I., Reimold, M., Pankau, W. M., & von Kiedrowski, G. (2005). Systems chemistry: kinetic and computational analysis of a nearly exponential organic replicator. *Angewandte Chemie*, *117*(41), 6908-6913. https://doi.org/10.1002/ange.200501527

Lahav, N., & Chang, S. (1976). The possible role of solid surface area in condensation reactions during chemical evolution: Reevaluation. *Journal of Molecular Evolution*, *8*(4), 357–380. https://doi.org/10.1007/bf01739261

Lahav, N., White, D., & Chang, S. (1978). Peptide formation in the prebiotic era: Thermal condensation of glycine in fluctuating clay environments. *Science*, *201*(4350), 67–69. https://doi.org/10.1126/science.663639

Ma, W., Trusina, A., El-Samad, H., Lim, W. A., & Tang, C. (2009). Defining Network Topologies that Can Achieve Biochemical Adaptation. *Cell*, *138*(4), 760–773. https://doi.org/10.1016/j.cell.2009.06.013

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. (2014). Ester formation and hydrolysis during wet-dry cycles: Generation of far-from-equilibrium polymers in a model prebiotic reaction. *Macromolecules*, *47*(4), 1334–1343. https://doi.org/10.1021/ma402256d

Martin, O., & Horvath, J. E. (2013). Biological Evolution of Replicator Systems: Towards a Quantitative Approach. *Origins of Life and Evolution of Biospheres*, 43(2), 151–160. https://doi.org/10.1007/s11084-013-9327-4

Miao, X., Paikar, A., Lerner, B., Diskin-Posner, Y., Shmul, G., & Semenov, S. N. (2021). Kinetic Selection in the Out-of-Equilibrium Autocatalytic Reaction Networks that Produce Macrocyclic Peptides. *Angewandte Chemie - International Edition*, *60*(37), 20366–20375. https://doi.org/10.1002/anie.202105790

Ogata, K. (1995). *Discrete-time control systems.* Prentice-Hall, Inc.

Pascal, R., Pross, A., & Sutherland, J. D. (2013). Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open Biology*, *3*(11), 130156. https://doi.org/10.1098/rsob.130156

Peng, Z., Plum, A. M., Gagrani, P., & Baum, D. A. (2020). An ecological framework for the analysis of prebiotic chemical reaction networks. *Journal of Theoretical Biology*, *507*, 110451. https://doi.org/10.1016/j.jtbi.2020.110451

Prigogine, I. (1978). Time, structure, and fluctuations. *Science*, *201*(4358), 777-785.

Pross, A. (2003). The driving force for life's emergence: Kinetic and thermodynamic considerations. *Journal of Theoretical Biology*, *220*(3), 393–406. https://doi.org/10.1006/jtbi.2003.3178

Pross, A. (2009). Seeking the chemical roots of Darwinism: Bridging between chemistry and biology. *Chemistry - A European Journal*, *15*(34), 8374–8381. https://doi.org/10.1002/chem.200900805

Pross, A. (2011). Toward a general theory of evolution: Extending Darwinian theory to inanimate matter. *Journal of Systems Chemistry*, 2(1), 1–14. https://doi.org/10.1186/1759-2208-2-1

Ross, D. S., & Deamer, D. (2016). Dry/wet cycling and the thermodynamics and kinetics of prebiotic polymer synthesis. *Life*, *6*(3), 1–12. https://doi.org/10.3390/life6030028

Shapiro, R. (2000). A replicator was not involved in the origin of life. *IUBMB Life*, *49*(3), 173–176. https://doi.org/10.1080/152165400306160

Sibilska, I., Feng, Y., Li, L., & Yin, J. (2018). Trimetaphosphate activates prebiotic peptide synthesis across a wide range of temperature and pH. *Origins of Life and Evolution of Biospheres*, *48*, 277-287. https://doi.org/10.1007/s11084-018-9564-7

Vincent, L., Berg, M., Krismer, M., Saghafi, S. S., Cosby, J., Sankari, T., Vetsigian, K., Cleaves, H. J., & Baum, D. A. (2019). Chemical ecosystem selection on mineral surfaces reveals long-term dynamics consistent with the spontaneous emergence of mutual catalysis. *Life*, *9*(4). https://doi.org/10.3390/life9040080

Wagner, N., Hochberg, D., Peacock-Lopez, E., Maity, I., & Ashkenasy, G. (2019). Open prebiotic environments drive emergent phenomena and complex behavior. *Life*, *9*(2). https://doi.org/10.3390/life9020045

Walker, S. I., Grover, M. A., & Hud, N. V. (2012). Universal sequence replication, reversible polymerization and early functional biopolymers: A model for the initiation of prebiotic sequence evolution. *PLoS ONE*, *7*(4), 31–37. https://doi.org/10.1371/journal.pone.0034166

Wynveen, A., Fedorov, I., & Halley, J. W. (2014). Nonequilibrium steady states in a model for prebiotic evolution. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *89*(2), 1–10. https://doi.org/10.1103/PhysRevE.89.022725

Yamanaka, J., Inomata, K., & Yamagata, Y. (1988). Condensation of Oligoglycines with Trimeta- and tetrametaphosphate in aqueous solution. *Origins of Life and Evolution of the Biosphere*, *18*(31), 165–179. https://doi.org/10.1007/bf01804669

5. CONCLUSION

In Chapter 2, we observed that peptide bond formation appears to take place in two distinct steps and explained this feature of the system based on the presence of two reaction mechanisms separated by water activity and pH conditions. When the reaction is performed starting with alkaline conditions, the products of the first reaction mechanism create the environment and reactants for the second. This behavior increases the yield of longer peptides, and reactions that naturally occur in sequence like this may have helped to promote the formation of longer molecules prior to the emergence of fully developed biological cycles.

In Chapter 3, we described peptide formation using an ODE network which summarizes the rates of peptide formation and hydrolysis. The goal of the model was to determine whether kinetic parameters could be determined from experimental results and what those parameters indicated about the underlying reaction rates. For the network we defined, the rate parameters could not be identified with high precision using experimental data due to sloppiness, a property which makes parameter estimation extremely sensitive to noise in the data. However, the parameters collectively estimated for the system can still accurately predict peptide yields, so the model yields are in good agreement with the experimental results.

Chapter 4 applies the parameter estimation method to study the kinetics of the two mechanisms for TP-activated peptide formation in cyclic environmental conditions. The model suggests that longer peptides may be able to exceed their mathematically predicted equilibrium concentrations in oscillating environments due to kinetic effects. Although this effect was difficult to capture experimentally, we did find experimental evidence that cyclic replenishment conditions influence reaction selectivity. We suggest that the relationship between reaction network dynamics and environmental cycles may have contributed to the development of important features of life-like reaction networks, such as kinetic control and sustained out-of-equilibrium behavior.

Overall, we established that multiple reaction mechanisms were occurring in series during TP-activated peptide formation, then built and applied a model to describe peptide formation based on simplified reaction kinetics. Though further work is needed to efficiently search for and identify autocatalytic peptides, we made progress in characterizing and abstracting the kinetics of peptide bond formation in prebiotic reaction networks.

6. FUTURE DIRECTIONS

There are various ways in which both the analytical development and the experimental design of this project could be extended for future work. Both aspects of the project should be developed in parallel since the experiments need to be planned in the context of what questions can realistically be addressed with the analytical methods available.

6.1 Analytical Development

This work progresses the analytical methods of our lab, particularly through the use FMOC derivatization to make the analysis of peptide mixtures, particularly those containing small hydrophilic species, more robust and reliable. FMOC derivatized products allowed us to use reverse phase liquid chromatography (RP-HPLC) for hydrophilic amino acids instead of relying on ion pairing chromatography (IP-HPLC) (Sibilska et al. 2018), which can be more difficult to optimize due to the need to find the correct pH and buffer concentrations in addition to managing gradient timing and solvent composition (Jandera 2020). IP-HPLC methods may also be sensitive to contamination, and the buffer salts can precipitate and cause blockages in the HPLC or the column if they are not properly rinsed out (MilliporeSigma n.d.). Derivatization also allowed us to analyze data at 254 nm instead of 195 nm, which reduces variation from the UV source and noise from contaminants. Many derivatization procedures use a large amount of excess FMOC, but we found that this excess, and the subsequent extraction procedures used to remove FMOC side products (Molnár-Perl 2011), were not required for these experiments. These results made analysis faster and reduced the amount of material required relative to the extraction procedure (Appendix A). However, there are still significant improvements to the analytical methods of this project which would be useful to pursue.

One of the goals of further analytical development should be to be able to quantitatively analyze more complex peptide mixtures. Although absolute quantitation is ideal, relative quantitation can also be a useful measurement. Analyses of similar systems are usually performed using HPLC, quantitative NMR, or LC-MS/MS. HPLC is quantitative, flexible, and readily available, but is unreliable for determining peptide sequences and may require extensive standards. Quantitative NMR is primarily used to study cases where the structure of the product, rather than just the composition of the peptide, is significant. LC-MS/MS is extremely sensitive, can determine peptide sequences, and requires far fewer standards HPLC. However, MS is sensitive to high salt concentrations and not inherently quantitative, so establishing reliable quantitative MS methods can be difficult.

We collaborated with Prof. Lingjun Li from the School of Pharmacy at UW Madison and her graduate student, Graham Delafield, to explore using quantitative MS/MS methods to study the samples discussed in this work. A summary of those tests and their results is included in Appendix B. We also began investigating the possibility of using unlabeled HPLC profiles to study peptide distributions without having quantitative information for most individual species. The goal of this approach was to consider methods of looking for change over time, or other indications of intermolecular interactions, which do not require us to know the exact concentration of every peptide species in the mixture. Further details on these methods can be found in Appendix C. While direct rate parameter estimates cannot be gathered from unlabeled data, it can still be used to evaluate various hypotheses about selectivity and peptide diversity. Several other groups have published works studying complex short peptide mixtures, mainly using MS, and their work should provide useful guidelines for approaching analytical method development (Le Maux et al. 2015; Forsythe et al. 2017; Surman et al. 2019; Doran et al. 2021; Jain et al. 2022).

An additional challenge of studying complex mixtures is the data analysis, particularly if there are aspects which are noisy, incomplete, or unknown. The exact approach required will depend on what information is available and what hypotheses are being investigated, but fortunately, the need to draw conclusions from data which may be sparse or incomplete is not unique to origins of life research. Various network analysis methods have been used to study both chemical (Searson et al. 2007; Burham et al. 2008) and biochemical reactions (Crampin et al. 2004; Schnoerr et al. 2017) and may provide useful insights. Similar tasks have also been extensively investigated in bioinformatics research (Cohen 2004; Gauthier 2019), and although there are differences between biological and chemical networks, it may still be possible to draw useful parallels between them. A preliminary project exploring the connectivity of the peptide reaction network using network inference methods is included in Appendix D. Developing these analyses would allow us to extract the most important features from experimental results and summarize important features of complex prebiotic reaction networks.

6.2 Experimental Design

There are also many directions in which these experiments of this project could potentially be developed. As mentioned above, studying more complex amino acid mixtures is a promising extension and is important for finding more generalizable results. Complex mixtures may be more likely to develop emergent intermolecular interactions between peptides both because they can form a greater variety of species and because there are more opportunities for secondary interactions between the side chains of diverse amino acid species. Preliminary experiments using mixtures of 5 amino acids and 10 amino acids were explored but did not lead to clear conclusions; the results of these tests are included in Appendix C.

Similar experiments to the ones discussed in this work could easily be conducted with various environmental changes to determine how those conditions affect the reaction network and peptide yields. For example, the timing of the replenishment cycles performed in Chapter 4 could be varied to explore how these changes affect the distribution of peptides formed. The inclusion of additional species, such metal ions, may help promote more significant changes in the system. Metal ions are known to promote peptide bond formation (Yamagata & Inomata 1997; Rode 1999; Belmonte & Mansy 2016) and catalytic activity (Rufo et al. 2014; Timm et al. 2023). At minimum, the effect of metal ions and their counterions on the reaction rates of various amino acids and the corresponding peptide selectivity could be assessed. Various preliminary studies using divalent cations were performed, and the basic methods and results are included in Appendix E. Additionally, coacervate formation, or liquid-liquid phase separation driven by the properties of polymers in solution, has also recently gained interest as a mechanism for promoting the formation of organic polymers, including peptides (Matsuo & Kurihara 2021; Hansma 2023). Developing experiments involving coacervate materials may be another way to promote peptide formation and encourage interactions leading to catalytic effects.

Although these variations could provide interesting results about peptide formation in bottom-up experiments and the effects of different environmental factors, they are not necessarily an efficient way of approaching the issue of finding autocatalysis in general. A more targeted approach using designed peptides is still currently the most efficient method of studying autocatalysis. A variety of autocatalytic peptides of various lengths have been published, and have tended to get shorter over time: Lee et al. (1996) used a 32 amino acid long template, Rubinov et al. (2009) used a 12 amino acid long template, and Rout et al. (2022) used a peptide template of that was only 8 amino acids long. The knowledge of these peptides could be used to guide various experiments. For example, we could link top-down and bottom-up studies by attempting to form one part of an autocatalytic template using minimally guided reactions. It is also possible to continue to shorten autocatalytic peptides to try to find the minimal template that still exhibits autocatalysis or look for evidence of simple catalysis in peptides which are only slightly too small to develop autocatalytic behavior. Additional details on the suggested studies involving autocatalytic peptides are included in Appendix F. These experiments are a larger departure from those explored in this work, but they pursue concretely understood mechanisms of autocatalysis and may allow for more targeted method development, since the identity of the peptide and mechanism of autocatalysis are already known.

6.3 References

Belmonte, L., & Mansy, S. S. (2016). Metal catalysts and the origin of life. *Elements*, *12*(6), 413–418. https://doi.org/10.2113/gselements.12.6.413

Burnham, S. C., Searson, D. P., Willis, M. J., & Wright, A. R. (2008). Inference of chemical reaction networks. *Chemical Engineering Science*, *63*(4), 862–873. https://doi.org/10.1016/j.ces.2007.10.010

Cohen, J. (2004). Bioinformatics - An introduction for computer scientists. *ACM Computing Surveys*, *36*(2), 122–158. https://doi.org/10.1145/1031120.1031122

Crampin, E. J., Schnell, S., & McSharry, P. E. (2004). Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Progress in Biophysics and Molecular Biology*, *86*(1), 77–112. https://doi.org/10.1016/j.pbiomolbio.2004.04.002

Doran, D., Clarke, E., Keenan, G., Carrick, E., Mathis, C., & Cronin, L. (2021). Exploring the sequence space of unknown oligomers and polymers. *Cell Reports Physical Science*, *2*(12), 100685. https://doi.org/10.1016/j.xcrp.2021.100685

Forsythe, J. G., Petrov, A. S., Millar, W. C., Yu, S.-S., Krishnamurthy, R., Grover, M. A., Hud, N. V., & Fernández, F. M. (2017). Surveying the sequence diversity of model prebiotic peptides by mass spectrometry. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1711631114

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, *20*(6), 1981–1996. https://doi.org/10.1093/bib/bby063

Hansma, H. G. (2023). Liquid–liquid phase separation at the origins of life. In *Droplets of Life*. Elsevier Inc. https://doi.org/10.1016/b978-0-12-823967-4.00006-3

Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K., & Ghadiri, M. R. (1996). A self-replicating peptide. *Letters to Nature*, *382*, 525–528.

Le Maux, S., Nongonierma, A. B., & Fitzgerald, R. J. (2015). Improved short peptide identification using HILIC-MS/MS: Retention time prediction model based on the impact of amino acid position in the peptide sequence. *Food Chemistry*, *173*, 847–854. https://doi.org/10.1016/j.foodchem.2014.10.104

Jain, A., McPhee, S. A., Wang, T., Nair, M. N., Kroiss, D., Jia, T. Z., & Ulijn, R. V. (2022). Tractable molecular adaptation patterns in a designed complex peptide system. *Chem*, *8*(7), 1894–1905. https://doi.org/10.1016/j.chempr.2022.03.016

Jandera, P. (2020). Comparison of various modes and phase systems for analytical HPLC. In *Handbook of Analytical Separations* (Vol. 8, pp. 1-91). Elsevier Science BV.

Matsuo, M., & Kurihara, K. (2021). Proliferating coacervate droplets as the missing link between chemistry and biology in the origins of life. *Nature Communications*, *12*(1), 1–13. https://doi.org/10.1038/s41467-021-25530-6

MilliporeSigma. (n.d.). *How to Identify, Isolate, and Correct the Most Common HPLC Problems.* HPLC troubleshooting guide. https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/analytical-chemistry/small-molecule-hplc/hplc-troubleshooting-guide

Molnár-Perl, I. (2011). Advancement in the derivatizations of the amino groups with the o-phthaldehyde-thiol and with the 9-fluorenylmethyloxycarbonyl chloride reagents. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, *879*(17–18), 1241–1269. https://doi.org/10.1016/j.jchromb.2011.01.027

Rode, B. M. (1999). Peptides and the origin of life. *Peptides*, *20*(6), 773–786. https://doi.org/10.1016/S0196-9781(99)00062-5

Rout, S. K., Rhyner, D., Riek, R., & Greenwald, J. (2022). Prebiotically Plausible Autocatalytic Peptide Amyloids. *Chemistry - A European Journal*, *28*(3). https://doi.org/10.1002/chem.202103841

Rubinov, B., Wagner, N., Rapaport, H., & Ashkenasy, G. (2009). Self-Replicating Amphiphilic β-Sheet Peptides. *Angewandte Chemie*, *121*(36), 6811–6814. https://doi.org/10.1002/ange.200902790

Rufo, C. M., Moroz, Y. S., Moroz, O. V., Stöhr, J., Smith, T. A., Hu, X., Degrado, W. F., & Korendovych, I. V. (2014). Short peptides self-assemble to produce catalytic amyloids. *Nature Chemistry*, *6*(4), 303–309. https://doi.org/10.1038/nchem.1894

Schnoerr, D., Sanguinetti, G., & Grima, R. (2017). Approximation and inference methods for stochastic biochemical kinetics - A tutorial review. *Journal of Physics A: Mathematical and Theoretical*, *50*(9). https://doi.org/10.1088/1751-8121/aa54d9

Searson, D. P., Willis, M. J., Horne, S. J., Wright, A. R., Searson, D. P., Willis, M. J., Horne, S. J., & Wright, A. R. (2007). Inference of chemical reaction networks using hybrid s-system models. *Chemical Product and Process Modeling*, *2*(1). https://doi.org/10.2202/1934-2659.1029

Surman, A. J., Rodriguez-Garcia, M., Abul-Haija, Y. M., Cooper, G. J. T., Gromski, P. S., Turk-MacLeod, R., Mullin, M., Mathis, C., Walker, S. I., & Cronin, L. (2019). Environmental control programs the emergence of distinct functional ensembles from unconstrained chemical reactions. *Proceedings of the National Academy of Sciences*, *116*(12), 5387–5392. https://doi.org/10.1073/pnas.1813987116

Timm, J., Pike, D. H., Mancini, J. A., Tyryshkin, A. M., Poudel, S., Siess, J. A., Molinaro, P. M., McCann, J. J., Waldie, K. M., Koder, R. L., Falkowski, P. G., & Nanda, V. (2023). Design of a minimal di-nickel hydrogenase peptide. *Science Advances*, *9*(10), eabq1990. https://doi.org/10.1126/sciadv.abq1990

Yamagata, Y., & Inomata, K. (1997). Condensation of glycylglycine to oligoglycines with trimetaphosphate in aqueous solution. II: Catalytic effect of magnesium ion. *Origins of Life and Evolution of the Biosphere*, *27*(4), 339–344. https://doi.org/10.1023/A:1006529421813

APPENDICES

APPENDIX A: FMOC DERIVATIZATION METHODS AND VALIDATION

Appendix A.1: Motivation for using FMOC

Fluorenylmethoxycarbonyl chloride (FMOC) is an organic synthesis reactant that is commonly used for derivatizing amino acids (Fig. A.1). It reacts with amines has been used to study complex amino acid mixtures in a variety of fields, including food science, medicine, and peptide synthesis (Roturier et al. 1995; Gartenmann & Kochhar 1999; Jámbor & Molnár-Perl 2009a; Jámbor & Molnár-Perl 2009b).

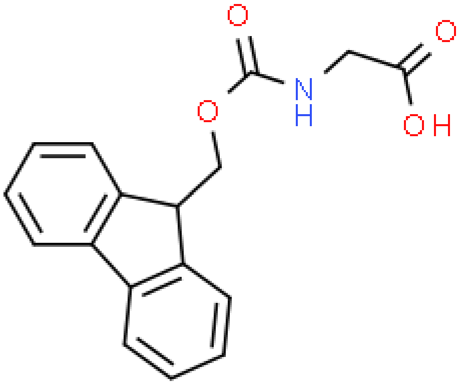


Figure A.1 Structure of FMOC-Glycine. Image from ChemSpider. Accessed July 17, 2023. (http://www.chemspider.com/Chemical-Structure.84070.html).

Prior analyses of hydrophilic amino acids in our lab were performed using ion pairing chromatography (IP-HPLC) (Sibilska et al. 2017), however, these methods can be sensitive to very small changes in solvent conditions and it can therefore be challenging to ensure consistency (Yamamoto et al. 1999). Although these methods were effective, they required significant method development time for relatively minor changes in procedure, particularly when different amino acids were used or adjustments were made that affected the pH of the sample. FMOC derivatization increases the hydrophobicity of amino acids and peptides, improving their retention time for reverse phase chromatography (RP-HPLC). FMOC is also visible at higher UV wavelengths than peptide bonds, which allowed us to analyze our samples at 254 nm, compared to 195 nm in the underivatized samples. The higher wavelength showed significantly less background noise and had better long-term stability, since the intensity of the UV lamp may start to vary in its extreme lower range as the bulb ages (COoL Boigenzahn/HPLC Hardware Maintenance Notes). Derivatization can also increase the sensitivity of a method, allowing for improved identification and quantitation of low abundance products.

Other derivatization agents have also commonly been used in conjunction with amino acids and HPLC. O-phthalaldehyde (OPA) is frequently referenced, but we were advised against it by coworkers who had previously used it and had difficulty with the stability of the derivatization, which seems to be a common problem (Fürst et al. 1990; Bütikofer et al. 1991). Other derivatization agents should still be considered for cases where the FMOC derivatization cannot provide the expected results. For more hydrophobic amino acids, derivatization may not even be necessary for separation, although it may still be useful to reduce noise and increase sensitivity.

Appendix A.2: Preliminary FMOC derivatization method and stability studies

Two versions of FMOC derivation methods were developed and characterized. The original method was significantly more labor-intensive, requires more material, and has much higher variance, so it is not generally recommended; it is included here mostly for context. The details of the original method are recorded in COoL Boigenzahn/HPLC Backups (Standards)/Fall2020_AnalysisProtocol.docx file. This method used a much higher ratio of FMOC to amino acid (FMOC:Gly = 5.4:1) and included a liquid-liquid extraction step performed with pentane to remove some of the excess FMOC and side products. The pentane was then evaporated off, and the remaining material dissolved in water and isopropyl alcohol, which was added to prevent clumping. This method was developed based on the discussion and Table 9 in Molnár-Perl (2011), in which most experiments used significant excess FMOC and performed an extraction method to remove it.

The stability of FMOC-Gly was evaluated using this method to determine if the derivatization product was degrading at a significant rate. Using this approach, the same sample of FMOC-Gly was analyzed three times. The sample was capped and kept in the autosampler during the day and refrigerated at night between analyses. For these conditions, the derivatization proved to be extremely stable. There was no significant decrease in the FMOC-Gly peak or increase in noise peaks over this period (Fig. A.2). In general, we avoided analyzing samples more than two days after they were derivatized, but these studies indicate that the derivatization remained reliably stable within that time. These results can be found in COoL Boigenzahn/Experiments/2020-07-23_G_MassBalance.

Appendix A.3: Updated FMOC derivatization method and validation

We eventually developed and validated a simpler FMOC derivatization protocol which requires fewer intermediate steps. A basic version of this procedure is included below, and details can also be found in COoL Boigenzahn/HPLC Backup (Standards)/2022_FMOCQuarterDilutionProcedure.docx and in the folders of relevant experiments. Because it did not include an extraction step, this approach had larger peaks for excess FMOC and its side products. However, these peaks did not overlap with or affect the peaks of interest. Note that this may not always be true, depending on the retention time of the peptides of interest and the gradient conditions. It is usually easier to adjust the gradient to improve retention time when possible, but if the peak of interest cannot be separated from the noise, it may be necessary to use an extraction procedure to remove as much of the side products as possible.

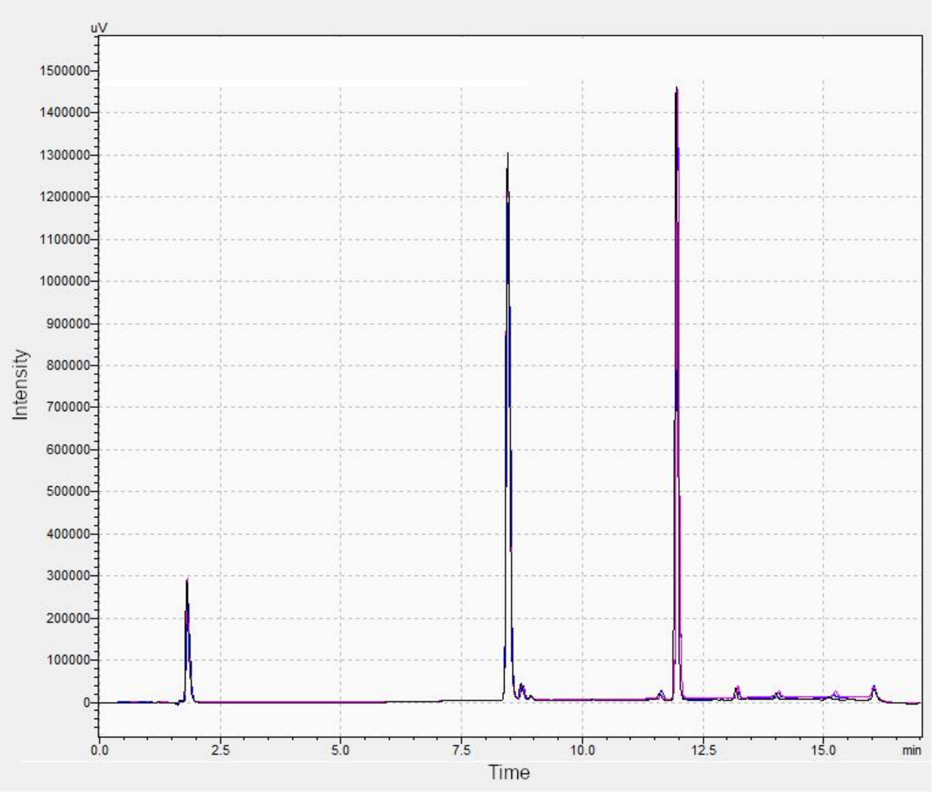


Figure A.2 Chromatograms of FMOC-Gly on various days after derivatization. Data was collected on the day of preparation (black), day after preparation (pink), and week after preparation (blue). The initial concentration of glycine in the sample prior to derivatization was 0.02 M.

Appendix A.3.1: Derivatization method

Quantities are provided for a sample containing a maximum of 0.1 M amine groups to derivatize, such as an experiment starting with 0.1 M amino acids.

1. To an 1.5 mL Eppendorf tube, add:
    a. 25 μL of sample to an Eppendorf tube
    b. 75 μL of milliQ water to the tube

    c. 100 µL 0.1 M sodium tetraborate buffer

2. Prepare 0.003906 M FMOC in acetone
    a. NOTE: I often did this by creating a 0.03 M solution of FMOC in acetone and diluting it in a ratio of 1:6.7 with pure acetone before adding it to the sample, because the higher concentration solution was easier to prepare and store.
3. Add 800 uL FMOC solution to the Eppendorf tube
4. Vortex the sample briefly to ensure mixing
5. Transfer at least 500 µL derivatized sample into HPLC vials, checking for precipitation
    a. May need to adjust sample concentration if there are precipitates beyond some very fine, suspended particles
6. Analyzing with the HPLC using a gradient appropriate for the column and amino acids in question.
    a. Unless otherwise specified, Solvent A: 100% milliQ water + 0.01% TFA v/v, Solvent B: 100% ACN + 0.01% TFA v/v for experiments using this derivatization method.
    b. The gradient will vary by experiment. Examples can be found in our published work, or the methods or HPLC files from my various experiments.

Appendix A.3.2: Quantitation and validation

This method uses 25% excess FMOC relative to the possible number of amine groups, which is much lower than what is commonly suggested. We performed experiments using variations of the method above to determine how much excess FMOC was required for quantitative results and found that above a 1:1 FMOC to amino acid ratio, there was no change in the intensity of the glycine peak, but the intensity of the primary FMOC noise peak continued to increase as expected (Fig. A.3). These results suggest that glycine is derivatizes readily, but once all glycine has been derivatized, the peak for excess FMOC or FMOC noise products begin to rise. Since we wanted to minimize the significance of the noise peaks, but including no excess FMOC provides very little margin of error, we used 25% excess FMOC relative to the maximum possible amino acids in future experiments.

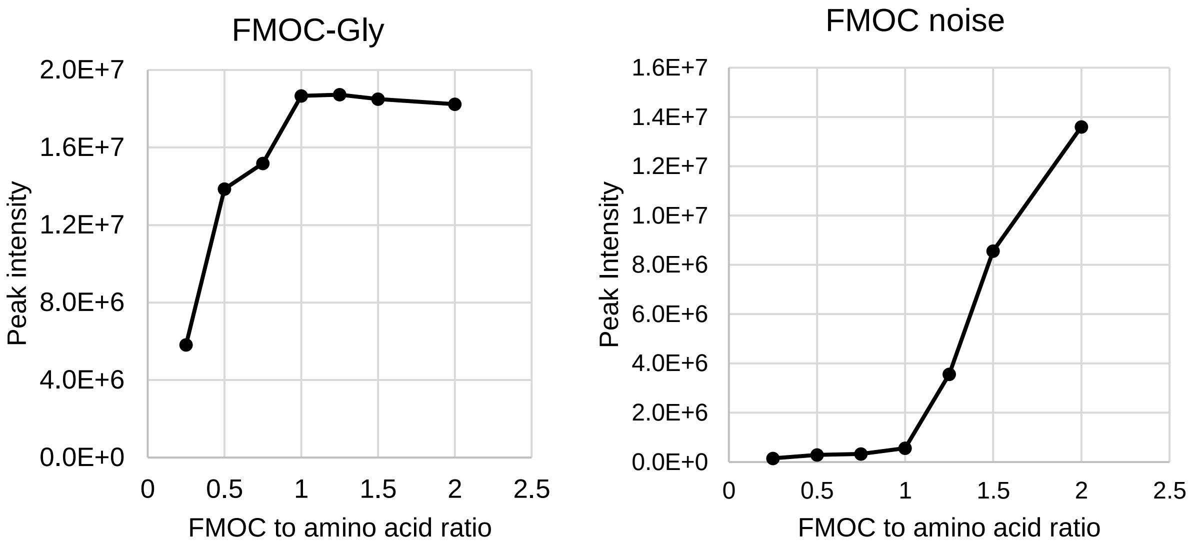


Figure A.3 Peak intensity of FMOC-Gly and FMOC noise as a function of increasing FMOC to amino acid ratio. 0.1 M glycine standard was used in the place of the sample in the protocol and FMOC concentration was varied by adjusting the concentration of FMOC solution added to the sample, base, and water.

We also assessed the quantitative accuracy of this approach by comparing multiple measurements from samples that were theoretically identical and compared it to a similar quantitative study performed using the original FMOC method. The new approach tended to slightly underestimate the quantitation of glycine; however, the results were much more precise than in the original method (Fig. A.4). This was the method we eventually used for our published results due to its significantly improved precision and speed. The materials for these experiments can be found in COoL Boigenzahn/Experiments/2021-06-04_FMOC-Scavenge-Experiments.

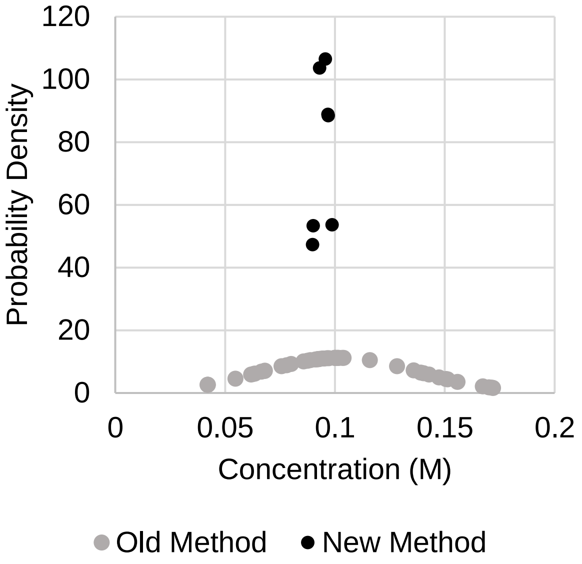


Figure A.4 Distribution of results from analysis of a 0.1 M Gly for old and new FMOC derivatization methods.

APPENDIX B: MASS SPECTROMETRY STUDIES

These studies were performed in collaboration with Prof. Lingjun Li and Graham Delafield from the School of Pharmacy at the University of Wisconsin-Madison. We would like to extend our thanks for their help and persistence in these efforts.

Appendix B.1: Benefits and limits of mass spectrometry

There are several benefits to using LC-MS instead of HPLC to characterize complex peptide mixtures. First, LC-MS does not always require extensive standards to identify the peptides in the mixture. HPLC retention times are often difficult to predict, so laboratory standards are required to determine the identity of each peak (Haddad et al. 2021). This limits the complexity of the mixture that can be studied if full labeling of the peptide species is desired, since ordering or synthesizing peptide standards for a combinatorial number of potential peptides eventually becomes intractably expensive and laborious. Even when all standards are available, some peptides may coelute and be difficult to separate.

With LC-MS, the composition of a peptide can be deduced based on its molar mass. While there are some peptides with overlapping molar masses, these could ideally be distinguished using tandem mass spectrometry (LC-MS/MS). MS/MS can be used to determine peptide sequences, which is often difficult with HPLC, since short peptides with the same amino acid composition are prone to coeluting. MS also generally has higher sensitivity than HPLC and can be used to study very low abundance species.

However, there are several challenges to using LC-MS for the types of studies conducted in this work. First, MS is not an inherently quantitative method. Although quantitative approaches have been developed and are relatively common in established fields like proteomics, designing and validating new quantitative methods is still a nontrivial process since the quantitation can be affected by many factors in the analytes, sample matrix (solvent), and instrument parameters (Li et al. 2011). Next, LC-MS is challenging to use with samples that contain high concentrations of salt (above roughly 100 mM, though this may change significantly depending on the method and the instrument) (Constantopoulos et al. 1999; Choi et al. 2000; King et al. 2000). Salts can strongly affect the behavior of electrospray ionization (ESI), which can complicate quantitation, and particular care needs to be taken to keep the instrument clean since they can be prone to leaving residue. Next, the fact that MS is highly sensitive to low abundance products can also be a source of noise which may make the results more difficult to interpret. Although there are some early explorations by groups in the origins of life field, many of the methods and tools developed to analyze peptides using MS are targeted towards proteomics research (Forsythe et al. 2017; Doran et al. 2021). This includes preparation protocols, product databases, and various data analysis tools, and while some of these approaches may be possible to adapt to study prebiotic peptide mixtures, doing so is not a straightforward or trivial exercise.

In this section, we will describe several methods of MS analysis we pursued and discuss what we achieved and what challenges we encountered. We did not develop a reliable, complete quantitative method of MS for peptide mixtures, but we were able to use MS for the detection and sequencing of longer, low abundance peptides generated from TP-activated mixtures that were not distinguishable via HPLC.

Appendix Section B.2: Exploration of quantitative MS methods

Appendix Section B.2.1: Sample preparation

The first challenge in analyzing the peptide samples using LC-MS was the salt concentration. Our typical samples included 0.1 M sodium trimetaphosphate (TP) and 0.15 M

NaOH, which gives an initial sodium ion concentration of 0.45 M Na^+. Eventually, we were able to run our samples using capillary electrophoresis (CE) (ZipChip), which separated the salts out before sample is analyzed by MS (Ramos-Payán et al. 2018). A more detailed description of the ZipChip and an example procedure can be found in COoL Boigenzahn/Misc MS Tests/2020-08-03_GlyPheHisAnalysis. Using the capillary electrophoresis method, we were able to search for peptides in a variety of sample conditions, including in samples containing magnesium chloride in addition to trimetaphosphate (Fig. B.1). These experiments are discussed further in Appendix 5 and further details can be found in COoL Boigenzahn/Experiments/2019-10-28_D+G_MgEffect.

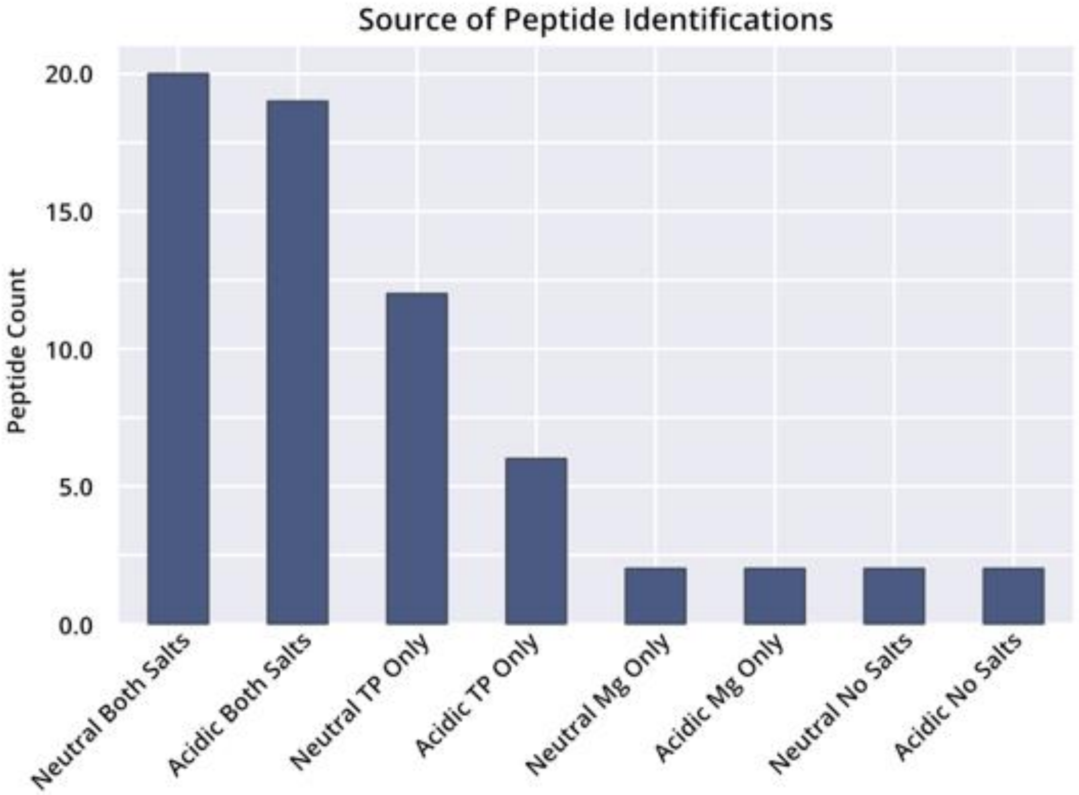


Figure B.1 Number of peptides identified via database search using MS results. These results were obtained from samples containing aspartic acid and glycine at various pHs and with various salt concentrations. Peptide counts may be inflated by low confidence identifications or identifications not properly constrained to the amino acids added to the sample.

As part of our later quantitation efforts (see Section 2.2.4), we also explored several sample preparation methods aimed at cleaning up the salt prior to any HPLC or MS analysis. We explored several methods of doing this using pipette tips containing chromatography media, which are used for sample preparation. We tried MilliporeSigma ZipTips, Agilent OMIX tips, and TopTip HILIC tips, but found that these methods often had poor separation of short peptides and at times had extremely inconsistent results. Details of these methods can be found in COoL Boigenzahn/Misc MS Tests/2021-09-24_DesaltingForMSTests. This approach may have been more effective with more hydrophobic amino acids or longer peptides, but for glycine oligomers we were not able to establish a reliable method of reducing the concentration of salt and excess monomer using this approach.

Although determining which peptides formed is interesting and a useful first step, it is not the only information required for kinetic studies, which involve at least measuring relative quantitation between samples, or ideally taking absolute measurements of species concentration. We investigated various possible methods of achieving quantitation.

Appendix B.2.2: Quantitation by internal standards

The first approach we tested to generate quantitative MS results used histidine as an internal standard to calibrate the relative peak heights of amino acids and peptides from experiments. Internal standards are used in MS to correct for errors caused by sample preparation, differences in ionization efficiency caused by matrix effects (how the solvent and other molecules in the environment effect ionization), or signal variability (Jeanne Dit Fouque et al. 2018). Histidine was added to each sample or standard such that it had a final concentration of 10 μM. However, internal standards are still subject to other forms of signal interference and differences in ionization efficiency, which we found made this approach impossible to use for reliable quantitation.

We tested samples prepared using only glycine or phenylalanine, as well as standards of those amino acids and their relevant polymers. We found on multiple occasions that the quantitative predictions were several orders of magnitude higher or lower than the expected concentration range, which suggested that the method was not behaving as anticipated. These results can be found in COoL Boigenzahn/Misc MS Tests/2020-08-03_GlyPheHisAnalysis. When we analyzed various dilutions of the same glycine sample to create a calibration curve, we found that the samples which had been diluted less (and therefore should have higher concentrations) had lower concentrations due to the distorted peak shapes (Fig. B.2). These results can be found in COoL Boigenzahn/Misc MS Tests/2020-09-28_AAQuant_redo. Materials present in high concentrations had irregular signals, meaning that samples need to be extensively diluted to produce clear peak shapes. Dilution is not inherently a problem when applied to standards, but such high dilution rates would potentially make lower abundance peptides in real samples impossible to detect.

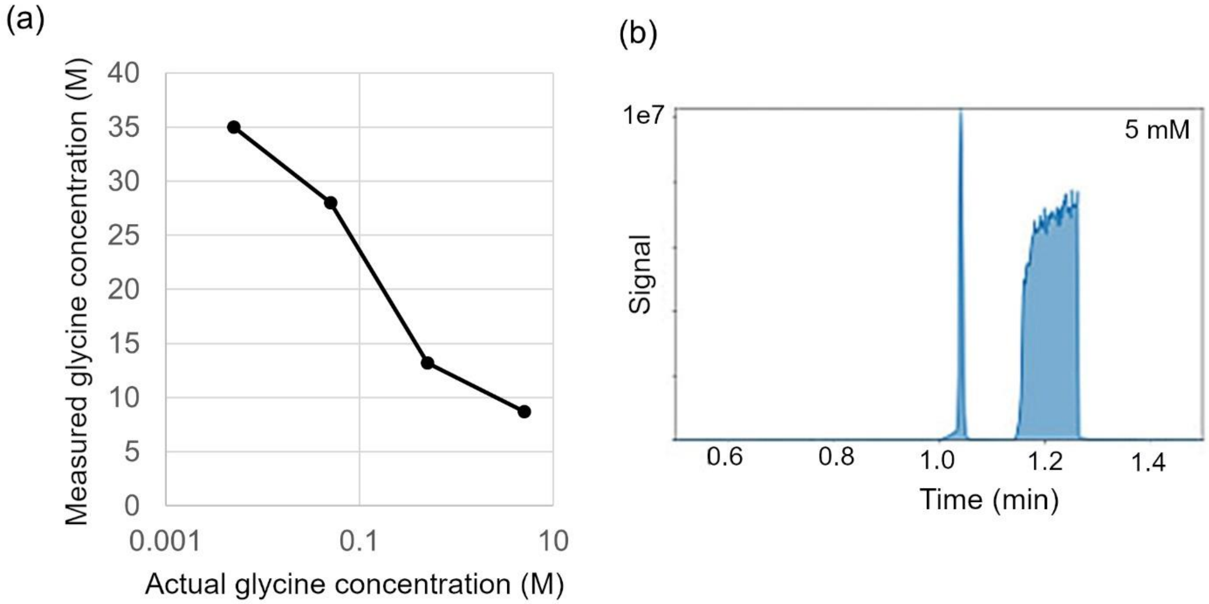


Figure B.2 Quantitation for various concentrations of glycine using histidine as an internal standard. (a) Measurements for various dilutions of 0.5 M glycine. The trend is the inverse of the expected behavior, probably due to irregularities in the peak shape of glycine (1.2 min) at higher concentrations. (b) The extracted ion chromatogram for the sample containing 5 mM glycine demonstrates the irregular peak shape found during these experiments.

It is also important to note that since the ZipChip is a CE device, not a chromatography column, it does not produce a chromatogram with peak intensities that can be used for quantitation. The peaks shown here are based only on the ion abundance detected by the mass spectrometer, which is notoriously unreliable for quantitation without standards of all the molecules being analyzed, since relatively minor differences in molecule composition or instrument setup can create biases in the ionization efficiency of the different molecules (Forsythe et al. 2017; Jeanne Dit Fouque et al. 2018). These biases can lead to wildly off-target concentration estimates, which may have been the problem we were experiencing with the glycine and phenylalanine samples. In proteomics, this issue is often addressed using heavy isotopologues, or molecules that are identical to the targets of interest but contain at least one isotopic amino acid (Liu et al. 2019).

Appendix B.2.3: Quantitation by heavy isotope standards

Heavy isotopologues have the same chemical behavior as their light analogs during sample preparation and analysis, making them much more accurate internal standards than an arbitrary amino acid like histidine (Leis et al. 1998). We were able to generate consistent linear calibration curves for glycine and alanine using 15N-glycine and 15N-alanine (Fig. B.3).

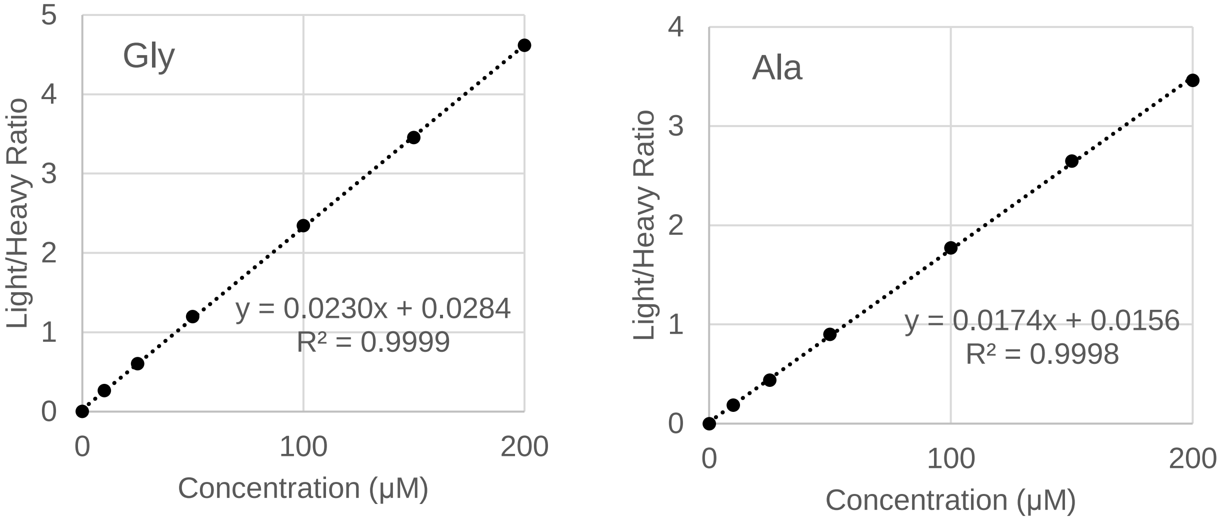


Figure B.3 Calibration curves for glycine and alanine based on internal standards of 15N-Gly and 15N-Ala. The isotoptic standards were added so that their final concentration was always 50 μM.

These calibration curves were used to analyze samples containing glycine and alanine, which replicated the samples analyzed in Chapter 3 of this work. Details can be found in COoL Boigenzahn/Misc MS Tests/2021-01-25_GA-Analysis. These results were consistent with the expected concentrations, with values slightly below the concentration of amino acids in the samples before they had been reacted to form peptides and having high precision (Fig. B.4). However, these results had limited use for the overall kinetic study since they only provide information about the monomer concentrations –an isotopic standard of each individual peptide would be required for rigorous quantitation of all peptides using this method. Isotopic custom peptides are even more expensive than regular standards, so this approach was not viable for complete quantitation. Furthermore, it eliminates one of the major potential benefits of MS over HPLC, which is the lack of a need for extensive standards. However, it could still be a useful method for quantitatively measuring specific amino acids or peptide products of interest.

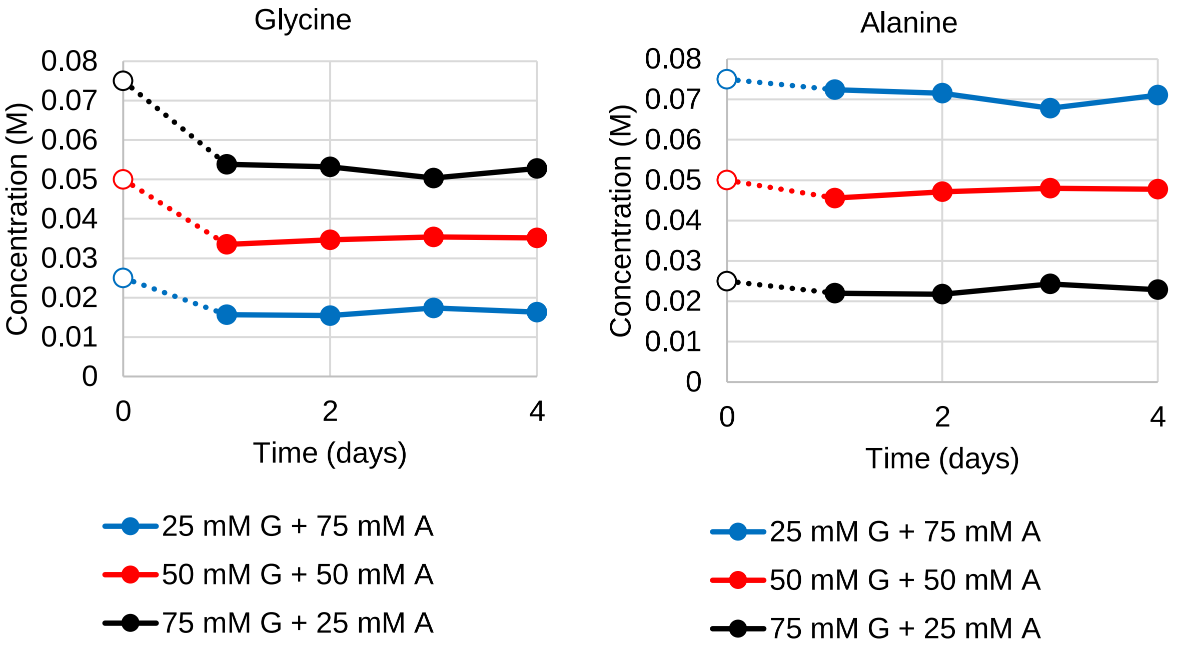


Figure B.4 Monomer concentrations in experiments created with various initial concentrations in glycine and alanine. The actual initial point was not measured since the samples had to be stored before analysis, and it was uncertain how stable the initial mixture would be even in the absence of high heat and drying conditions. Error bars from experimental triplicates representing standard deviations are not visible on this scale.

Appendix B.2.4: Isobaric labeling

Another common method of quantitative MS in proteomics is isobaric labeling, which involves attaching different heavy isotope tags to peptides in different samples. These tags are designed such that when mixed, the peptides from the different samples will have the same MS peak, but when the peptide is fragmented for MS/MS, the unique reporter region of each tag separates and can be used to find what proportion of each peptide sent for fragmentation came from what sample. A simple schematic of a general isobaric tagging procedure is provided in Fig. B.5, but there are many variations with different tags and different capacities for multiplexing (the number of experiments that can be mixed and analyzed in a single run) (Hsu et al. 2003). Isobaric tagging can be used for relative or absolute quantitation (Liu et al. 2017). Our collaborators in the Li lab have a great deal of experience developing isobaric tagging methods for proteomics applications. One of the approaches used in their lab is leucine-based isobaric tags (DiLeu) which are relatively inexpensive to synthesize and have been successfully used to study a variety of biological peptide mixtures (Frost et al. 2015; Frost et al. 2020; Sauer & Li 2022).

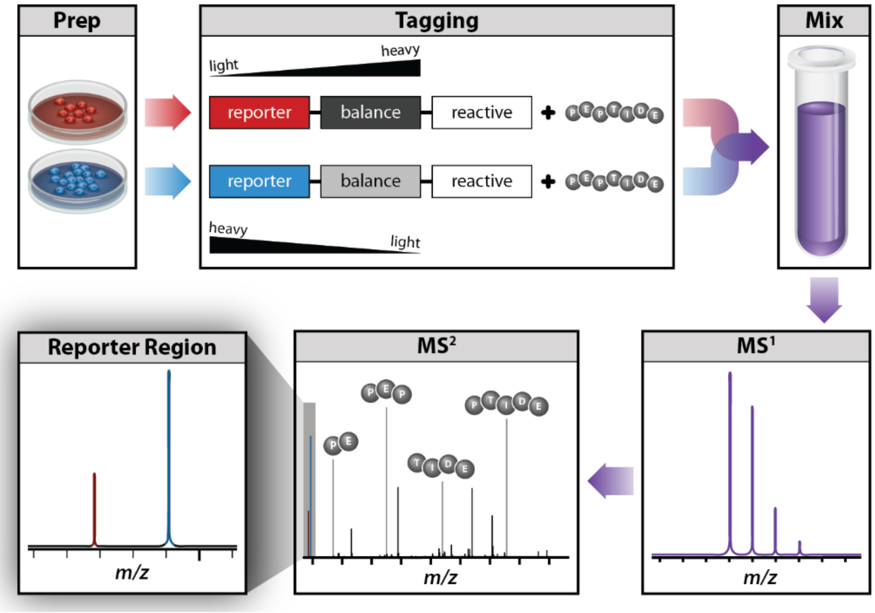


Figure B.5 Schematic diagram of isobaric tagging. Image: "Isobaric labeling" by A.J. Bureta (https://en.wikipedia.org/wiki/File:Isobaric_labeling.png). Accessed July 14, 2023. Licensed under Creative Commons 3.0.

We attempted DiLeu labeling of the peptide samples, but never completed the experiments since the protocol was not compatible with the high concentration of monomers and salts in the samples. The exact protocol for our samples is no longer available, but it was heavily based on the protocol in Frost et al. 2020. This protocol requires that samples be redissolved in the solvent for the LC-MS process after tagging. We found that it was impossible to get the sample back into solution after they were dried, probably due to an excessively high concentration of either monomers or salts. Diluting the samples sufficiently to mitigate the monomer and salt concentrations would drop the mass of any longer peptides well below the recommended level for tagging and detection. We explored several avenues for removal of the salt, monomer, and dimer species from glycine samples (see Section 2.2.1), but ultimately did not find a reliable method.

Fractionation via HPLC or UPLC prior to MS/MS is a promising angle to continue to explore for this approach. However, the combination of fractionation, tagging, and then performing MS/MS is labor intensive and would likely limit throughput. Additionally, a different column may need to be used for hydrophilic amino acids like glycine, since our current columns do not separate underivatized hydrophilic species particularly well, and FMOC tagging is not compatible with DiLeu tagging. FMOC is an MS compatible group in general, but both FMOC and DiLeu react with the amine group of a peptide, therefore the active group needed for DiLeu tagging would be blocked if fractionation was performed with FMOC derivatized samples. Removal of the FMOC group after fractionation but prior to DiLeu labeling is theoretically possible, but the sheer number of intermediate steps involved in the analytical process would probably make establishing reliable quantitation extremely difficult. We also discussed using a different kind of isobaric tag instead of DiLeu, since there are some tags that would react with the C terminus and therefore would not necessarily be affected by FMOC, but these tests were never performed due to time constraints.

Appendix B.2.5: Peptide sequencing

We were able to perform MS/MS to sequence peptides in some of our experimental samples. These tests were able to identify peptides in experiments which were not detectable with our HPLC methods. However, we found we were only able to sequence peptides that were roughly four amino acids long or larger.

In the samples containing glycine and alanine which were studied in Chapter 3, we identified eight peptides in the sample prepared using 75 mM glycine and 25 mM alanine (AGAG, AGGA, GGGG, AAGGA, AGGGG, AGGGGA, AAGGGG, AGGGGG). These results were interesting since the longest peptide being tracked by HPLC was GGGG, and the appearance of peptides up to six amino acids long with a single day of drying was encouraging. However, no peptides were detected using MS/MS in the samples initiated with 50 mM Gly + 50 mM Ala or 25 mM Gly + 75 mM Ala. While it was unsurprising that the samples enriched in glycine formed longer peptides, since glycine tends to form peptide bonds more easily than many other amino acids, we did detect GGGG in the other experiments using HPLC, so it is unclear why that peptide was not detected using MS/MS. Details on these experiments can be found in COoL Boigenzahn/Misc MS Tests/2021-01-25_GA-Analysis.

It also seemed unusual that longer peptides were being detected, but dimers and trimers were not. Graham manually examined the MS1 results for these experiments and located peaks for the dimers and trimers in all experiments, however, these peaks were not selected for MS/MS. Even when we tried using a targeted method, in which the molar mass of the MS1 peaks desired for fragmentation is specified, the peaks of shorter peptides were not selected for fragmentation. This behavior was consistent with observations from previous experiments, in which we had been unable to see short peptide standards in MS/MS, despite there being nothing else in solution that should be selected for fragmentation.

Graham hypothesized that, due to the combination of small size and high charge on the short peptides, the instrument software regarded them as noise and filtered them out. There may be a setting to correct this behavior and allow all peaks to be selected for fragmentation, but we were never able to determine what that setting was, or even confirm if the problem was related to the software. If it is a question of a particular software setting, this problem may be self-correcting if these tests were repeated on a different instrument. Despite the limitations of small polymer sequencing, the existing methods are still useful for certain types of analyses, such as assessing the number of species detected vs. the number of species that could have theoretically formed.

Appendix B.2.6: Data analysis

Mass spectrometers can produce a large amount of data which is difficult to analyze without the appropriate software. We briefly collaborated with Prof. David Baum and his student Tymofii Sokolskyi on some additional MS experiments. The Baum lab uses CompoundDiscoverer, a proprietary ThermoFisher software, to analyze their mass spectrometry results. We prepared a code to look for the masses of peptides given a particular input combination of amino acids based on the output of CompoundDiscoverer. The goal of this code was to automatically clean and search the mass results for peaks of interest, plot the information in a legible way, and eventually perform statistics on the cleaned results. More details on these experiments and code are included in Appendix 3.2.2. Analyzing MS data without some kind of peak identification software like CompoundDiscoverer is technically possible, but it would add a significant software development task onto the analysis, so it may be advisable to look into some of the many available open source or academic software tools for MS analysis (Sturm et al. 2008; Mortensen et al. 2010).

Appendix B.3: Future considerations involving mass spectrometry

Although we had difficulty establishing a quantitative method for mass spectrometry, it is still a promising analytical tool for studying complex peptide mixtures. Using mass spectrometry to identify the presence of certain peptides is straightforward and can be used to study properties like selectivity in complex mixtures. Quantitation for kinetic studies is more challenging and requires more attention to the details of each individual species. The only approach we found to be quantitatively accurate was using isotopologues as internal standards, but there may be other avenues, particularly in isobaric tagging, to explore which would enable quantitation of a greater variety of species. Establishing methods of cleaning up salts and high abundance species, such as leftover monomer, from the samples would probably make establishing a quantitative MS method significantly easier. However, the sample preparation method would have to be checked regularly and evaluated against a variety of amino acids and peptides to determine its efficacy. This type of method development can be performed with a dedicated collaborator, but direct access to a mass spectrometer would be very helpful during this process.

APPENDIX C: STUDIES WITH MORE COMPLEX AMINO ACID MIXTURES

We performed a set of preliminary experiments using mixtures of 5 amino acids and 10 amino acids using replenishment conditions. Each sample was dried for 24 hours, then rehydrated, dissolved, and 90% of the material was replaced with fresh reactants. The purpose of these experiments was to use HPLC and MS methods in which not all species were named and quantified to explore the potential for identifying ongoing or developing behavior in more complex amino acid mixtures. Details and results from this experiment can be found in COoL Boigenzahn/Experiments/2023-03-07_ComplexAminoAcidMixture.

In theory, if a catalytic or autocatalytic relationship develops, the set of molecules involved in that interaction will become enriched in the sample when transferred at high replenishment rates. We sought peptide species whose concentration continued to grow after most other peptide concentrations had stabilized. Most importantly, we were aiming to establish methods of analyzing these types of mixtures so that similar methods and data analysis programs could be used in the future to continue to explore more diverse conditions.

Appendix C.1: Experimental methods

We decided on what amino acids to use by analyzing how frequently the different species were mentioned in various papers discussing amino acid availability on the early Earth. These papers included amino acids found in meteorite compositions (Engel & Nagy 1982; Cronin & Moore 1971; Shinoyama & Ponnamperuma 1979), spark discharge experiments in various atmospheres (Miller & Orgel 1974; Wolman et al. 1972; Keefe et al. 1995; Miyakawa et al. 2002; Glavin et al. 2008; Johnson et al. 2008; Parker et al. 2011), data from hydrothermal vents (Marshall 1994; Hennet et al. 1992; Huber and Wächtershäuser 1998), and various other chemical synthesis experiments (Bar-Nun et al. 1970; Lowe et al. 1963; Yoshino et al. 1971; Sutherland 2016).

Based on these results, we chose G, A, V, D and E for our five amino acid mixture because they appeared the most frequently across the papers we referenced. For the mixture with 10 amino acids, we used those 5, plus S, I, L, P, and T for the additional 5. We used equal concentrations of amino acids with a total amino acid concentration of 0.1 M amino acid (0.02 M each for the mixture with 5 amino acids, 0.01 M each for the mixture with 10 amino acids). The samples also contained 0.1 M TP and 0.15 M NaOH, and were heated with the caps open for 24 hours at 90^{o}C. Every 24 hours, the samples were cycled with 90% replenishment rate, as described above. For future experiments, it may be interesting to adjust the initial concentrations of the amino acids to reflect the concentrations at which they are found in organic synthesis experiments, like those referenced above, since this would probably represent a more realistic prebiotic amino acid distribution. Details for these experiments can be found in COoL Boigenzahn/Experiments/2023-03-07_ComplexAminoAcidMixture.

Appendix C.2: Analytical methods

Appendix C.2.1: HPLC

The peptide mixtures produced from these experiments were too complex to analyze with exhaustive standards, as we had previously done with simpler mixtures. Instead, we found the retention time and intensities for many unlabeled peaks at various time points and compared them to determine how various peaks changed over time. This method has some similarities to molecular fingerprinting, an approach used to characterize peptides in tryptic digests which is often used in tandem with mass spectrometry (Fullmer & Wasserman 1979; Mayes 1984; Henzel et al. 2003).

We performed the HPLC analysis by using a search gradient which slowly covers a wide range of solvent conditions, aiming to separate as many distinct peptide peaks as possible. The gradients used are shown in Table C.1. We also analyzed standards of each individual amino acid to use them as benchmarks to help gauge the quality of separation. For example, D & E and L & I tend to elute very close together, so separation between these amino acids was used as a minimum requirement for the separation quality of the gradient.

5 amino acid mixture (GAVDE)	
Time (min)	%B
0.01	35
5	35
25	45
36	95
40	95
43	35
45	35

10 amino acid mixture (GAVDESILPT)	
Time (min)	%B
0.01	35
5	35
20	45
36	95
40	95
43	35
45	35

Table C.1 Gradient conditions for HPLC analysis of mixtures of 5 or 10 amino acid species.

Since the HPLC peaks do not elute at exactly the same time during each experiment, we needed a method to align the peaks from different analytical runs in order to be able to compare how each peak was changing over time. Peak alignment is a relatively common task, though some packages designed for mass spectrometry require additional information which is not relevant to HPLC results. The R package GCalignR is designed for gas chromatography results, but only requires retention time and intensity information for alignment, so we used this package to process the HPLC results (Ottensmann et al. 2018). This code is located in COoL Boigenzahn/Experiments/2023-03-07_ComplexAminoAcidMixture. There are two files, one for parsing data from the input format suggested on the GCAlignR website to the appropriate R tables required to run the program (2023-03-17_DataParsing_GCAlignR) and one which includes the functions for peak alignment and analysis (2023-03-

21_ComplexAATests_RAnalysisSteps). Various parameters in the second file can be used to adjust the peak alignment. When there is instability in the HPLC retention times, those parameters should be set to be more forgiving in their peak assignments. However, this can also result in grouping peaks together which should be distinct, so the results of the peak alignment still need to be manually evaluated to determine if the parameters were appropriate for the data.

After alignment, various plots can be created, such as heatmaps indicating the presence or absence of certain peaks within a particular threshold (Fig. C.1). Although we had insufficient time to fully develop these methods, statistical approaches such as multivariate analysis of variance (MANOVA) can be used to look for features like differences between experimental triplicates or statistically significant changes over multiple days.

In general, the experimental triplicates were very similar, however, there are certain samples (3d i) which are slightly different from the other versions. These samples are most likely outliers due to experimental or analytical errors, but this idea should be tested by creating additional experimental replicates and comparing them. Based on our preliminary observations, there is no clear trend which indicates that the peptide composition shifts significantly after the first day. However, due to time limitations and the limited availability of samples, especially after removing those that we felt might be outliers, we did not attempt to draw statistical conclusions from these results.

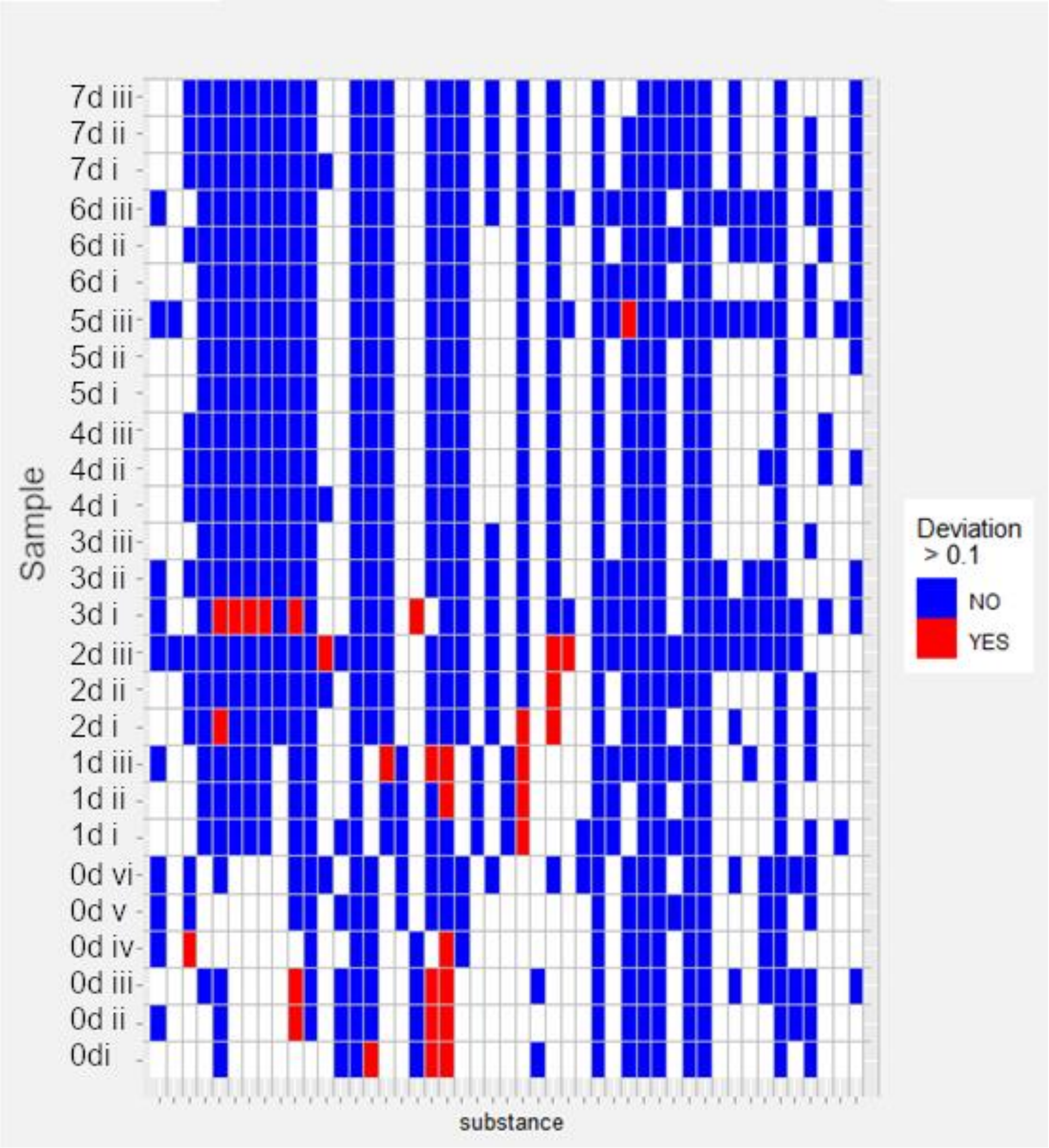


Figure C.1 Example heatmap for multiple days of the 5 amino acid mixture. The threshold can be adjusted to include or exclude peaks that are slightly different from the expected retention time for the substance.

Appendix C.2.2: Mass Spectrometry

We also ran a preliminary set of LC-MS analyses using the material from these samples. These tests were performed about a week after the last replenishment cycle was completed, and the samples were stored in the refrigerator in the meantime. Alongside the complex amino acid mixtures, we tested calibration curves of glycine and phenylalanine, as well as some phenylalanine samples reacted in different environmental conditions. It should be noted that these tests were done somewhat spontaneously, and the lab assistant from the Baum lab originally helping me with the lab work for the complex amino acid mixtures became unable to come into lab during this time due to illness. Because of the time constraints, some samples could not be found and therefore were not analyzed, and there may be gaps in the experimental records. The methodology and the complete experimental output from CompoundDiscoverer (a proprietary ThermoFischer software) can be found in Boigenzahn/Experiments/2023-03-07_ComplexAmnioAcidMixture.

These tests were done in collaboration with Tymofii Soloklyski from Prof. David Baum's lab. Tym is accustomed to running higher salt concentrations on the MS he uses, so we did not have to use CE or another method to analyze these samples. We diluted the samples to improve their signal and reduce the total salt concentration. Due to the high sensitivity of the MS method, it tends to find many noise peaks in every sample. CompoundDiscoverer also does not necessarily default to naming peptides when assigning species to molecular weights, meaning that many species that were likely experimentally relevant peptides were mislabeled as other molecules. Before attempting further analysis of this data, we needed a method of cleaning up the noise in the data and finding peaks which aligned with the theoretical peptides that could be formed by the amino acids included in the original samples.

Some code for these tasks can be found at COoL Boigenzahn/Experiments/2023-03-07_ComplexAminoAcidMixture/2023-05-03_MSResults_Processing. It removes all peaks that never exceed a given threshold, since those peaks are either noise or indistinguishable from it.

Candidate monomers can be provided along with their molar masses and recombined to identify peptides for a given set of amino acids. It may be useful to extend this code to verify that there is a significant enrichment in peptides that should theoretically be found over those that should not, based on what amino acids were present in the original sample. This could possibly be done by comparing the intensity of peptides containing only amino acids included in the sample to the intensity of a random selection of peptides which should not theoretically form. This would help indicate how many peptides are being identified purely by chance due to the quantity of noise peaks that are indicated as products. Like the code used for HPLC alignment, another useful extension of this code would be to implement statistical methods to analyze the variation between experimental triplicates and from day to day.

Appendix C.3: Results and conclusions

Additional data processing and statistics may allow us to draw more conclusions from these results, however, these methods are still a work in progress. There were notable outliers in the data that made it difficult to confidently draw any conclusions about how the system changed over time, which should be revisited with duplicate experiments to determine if these may have been analytical errors. One conclusion we can draw from these experiments, particularly regarding the mass spectrometry tests, is that planning the experimental design carefully is of utmost importance when dealing with complex mixtures. Writing out a specific hypothesis and null case, as well as determining number of replicates theoretically required for statistical significance before starting the experiments is highly recommended. Trying to evaluate these mixtures simply by visual inspection, even of simplified diagrams like the heatmap shown above, is very difficult and completely impractical given the amount of time it requires to plan and develop analytical methods for these types of experiments.

Analytical method validation can be somewhat complicated when dealing with complex mixtures and looking for unlabeled peaks. In that regard, we recommend taking the time to use standards of whatever amino acids or peptides are available to gain a sense for how the method behaves both in its day-to-day consistency and in its rate of false positives and false negatives. While nothing can replace a true positive control sample, it is worth the time to characterize and analyze whatever materials are available before designing and performing more complicated protocols.

APPENDIX D: NETWORK INFERENCE MODELS

Analyzing complex experimental data is one of the challenges of working with more diverse sets of amino acids or mixtures of organic molecules. Tools for analyzing and inferring relationships in chemical reaction networks may be useful for extracting useful conclusions from this experimental data. There is some existing work in the field of chemical reaction network analysis, primarily stemming from the area of chemical physics (Crampin et al. 2004). Similarly, a significant amount of work deciphering experimental results generated from systems with complex interactions has been performed in the bioinformatics field to help understand the many interactions in molecular biology (Gauthier et al. 2019; Markowetz & Spring 2007). Adapting these approaches may help generate new hypotheses and goals for experimental design that would allow us to extract insight from the results of complex mixture experiments.

I explored two methods of applying bioinformatics approaches to my prebiotic peptide samples while completing projects for my minor. During the first project (BMI 826), I attempted to use a dynamic network inference algorithm to deduce the most important reaction in the peptide reaction network. I attempted to infer networks for a variety of simulated and experimental networks, the details of which are included in the COoL Boigenzahn/Networks/2019-12-16_BMI826_FinalReport. For the network shown in Fig. D1, the simulated data was meant to imitate results gathered by Izabela Sibilska for experiments prepared as experimental triplicates with diglycine, TP, and base and rehydrated every 24 hours for 7 days. The initial conditions were not fully recorded, but were likely similar to those in Sibilska et al. (2017), and the results included measurements for species up to Gly_{10}. The simulated network was generated similarly to the simulated data in Chapter 3 and Chapter 4 of this work, but used one parameter for all the forward reactions and another parameter for all the hydrolysis reactions, since the parameter estimation method discussed in Chapter 3 had not been completed at the time of this project.

Network inference was performed using an algorithm called dynGENIE3 (Huynh-Thu & Geurts 2018) which was designed to infer gene regulatory networks from dynamic time course data. DynGENIE3 is a semi-parametric model which uses random forests to learn the parameters of a set of ODEs. Unlike the methods we explored in Chapter 3, the terms included in the ODEs for dynGENIE3 are learned by the algorithm, and are not specified by the user. This removes the need to decide on an approximate network model, but also makes it much more difficult to incorporate physical constraints that we know should exist in the system.

For glycine oligomer experiments, I found the networks did create more connections to significant, high abundance species, but also tended to be over-connected and the directionality was unintuitive (Fig. D.1). I also examined the network produced from a mixture of 5 amino acids (also based on experimental data gathered by Izabela Sibilska), but since only the monomers were labelled, I was unable to add very much insight on the accuracy of the larger inferred network. The network being overconnected may be a consequence of the fact that the interactions in these networks are conversion between molecules, not regulatory interactions like those often associated with gene regulatory networks, which may change the nature of the correlations in the network.

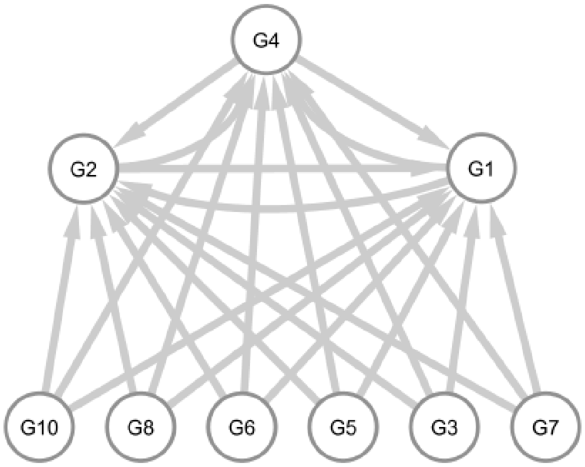


Figure D.1 The inferred reaction network for diglycine forming up to Gly$_{10}$ using simulated data generated from ODEs. See COoL Boigenzahn/Networks/BMI 826 Data for additional detail.

The second project, which was for BMI 776, was similar to the first, but applied two other dynamic network inferences methods found in literature. The report summarizing these results will be available in COoL Boigenzahn/Networks/2020-05-11_BMI776_FinalProject. These experiments were conducted using data from a mixture of glycine and alanine, and the accuracy of the method was evaluated based on whether that edge was likely to physically exist – for example, there should be a connection between G and GG, but there should probably not be a direct connection between AA and GG. All the methods I attempted were relatively inaccurate, either finding only a small percentage of the physically correct edges, or finding an extremely high number of physically unlikely edges along with the correct ones. The dynamic network inference method I used in the first project still had the best results overall, though the direction of the edges was still counterintuitive. One of these methods was specifically intended for use in chemical reaction networks (Burnham et al. 2008), but within the method the additional constraints did not significantly improve the quality of the network inference approach. Because I was using approaches from literature, it was often difficult to explain why the algorithm was performing poorly. An approach developed from basic principles to address a specific hypothesis or data analysis task, rather than just applying existing literature algorithms that claim to be generalizable, is more likely to produce meaningful results.

In conclusion, a high level of modification is required to apply bioinformatics algorithms to complex chemical reaction networks. There is still some overlap in the mathematical foundations of the two, and understanding how the subjects relate to each other and to interesting research questions related to origins of life reaction networks is a potentially interesting and novel direction for this project. Since it is challenging to make significant progress developing models while also doing the lab work to gather the experimental data, this work would be best as a purely theoretical project or in close collaboration with another research group who is familiar with chemical reaction network models or bioinformatics.

APPENDIX E: STUDIES WITH METAL SALTS

Metal salts, particularly divalent cations, are an interesting possible angle to pursue in future experiments. Although they can complicate method development, particularly for MS analysis, metal ions are extremely common the active sites of modern enzymes and would have been available in the minerals and salts of the early Earth (Belmonte & Mansy 2016; Bromberg et al. 2022). Amino acids and short peptides tend to bind to metal cations, and metal ions have been used to generate catalytic activity from relatively short, hypothetical prebiotic peptides (Kim et al. 2018; Wang et al.2019; Timm et al. 2023). We did not pursue these studies extensively but did perform preliminary tests with copper and magnesium salts.

Appendix E.1: Copper

The salt-induced peptide formation approach (SIPF) was proposed and explored by Dr. Bernd Rode in the 1990s (Rode & Schwendinger 1990). It involves a two-to-one ratio of amino acid to Cu(II) and is promoted by high concentrations of NaCl, though a previous member of our lab found that NaCl was not required for peptide bond formation when the samples were fully dried (Napier & Yin 2006). Rode and his group arrived at several interesting conclusions using SIPF, such as finding a correlation between the dimers formed by SIPF to the frequency of dimer sequences in proteins from archaebacteria and prokaryotic cells (Rode et al. 1997). One of the promising elements of SIPF is that the metal ion can act as a true catalyst rather than an activating agent which is consumed during the process. This eliminates the question of how activating agents could have been continuously supplied to an autocatalytic reaction cycle without causing waste accumulation. We performed some experiments using SIPF conditions to assess how they compared to TP during multiple cycles.

We analyzed samples containing 0.1 M glycine, 0.05 M $CuCl_2$, and 0.5 M NaCl, which were heated at 90^{o}C with the caps open for 24-hour intervals. At the end of each cycle, 1 mL of water was added to each sample to bring it back to its original volume, then the material was vortexed until dissolved (or until it appeared it would not dissolve further) and heated again. The full experimental procedure and results can be found in COoL Boigenzahn/Experiments/2021-07-05_SIPF-Gly_Kinetics.

We found that within the first two to three cycles, the samples changed from the pale blue color caused by the $CuCl_2$ to greenish-brown, and continued to darken until they were essentially black. We believe this may be because these experiments were performed with the caps open and were not in an anoxic chamber, so the copper may have oxidized over time. CuO is a black powder, which was consistent with what we observed forming in our experiments. Similar oxidation probably would not have occurred on the early Earth, since the oxygen concentrations in the Hadean period were much lower than they are today (Hao et al. 2019). Use of an anoxic chamber should be considered for SIPF experiments because of this effect.

It should be noted that when we originally performed these experiments, we were using the pentane extraction method for FMOC derivatization. This procedure seemed to remove some products from solution and made quantitative analysis virtually impossible. We were only able to generate linear calibration curves for SIPF samples after switching to the alternative FMOC derivatization method which did not require extraction. The extremely high salt concentrations used for SIPF can have a significant impact on analytical methods, so it is particularly important to validate all methods carefully for standards with equivalent salt concentrations.

Once a quantitative method to analyze the SIPF samples was identified, we repeated the experiments. However, we found that the overall concentration of glycine dropped significantly with each day of reaction. We attempted to compare our results to Rode's to determine if they ever found an overall loss to the amino acid concentrations, but their results are always published as yields and therefore do not include the concentration of the monomer. We considered the possibility that the loss of glycine might be due to its coordination to Cu(II), which may have been blocking it from derivatization and therefore preventing it from being detected by our analysis. However, when we tried changing the environmental conditions to promote the release of glycine from the copper complex, we found that there was no improvement. We tried increasing the acidity of the samples, which should protonate the glycine and discourage it from binding to the positively charged copper ion. We also added ethylenediaminetetraacetic acid (EDTA), an acid which is commonly used to bind iron and calcium complexes for descaling. None of these techniques caused any significant change in the amount of glycine recovered (Fig. E.1). Within three days of reactions, the overall quantitation of glycine had dropped to about 10% of its original value.

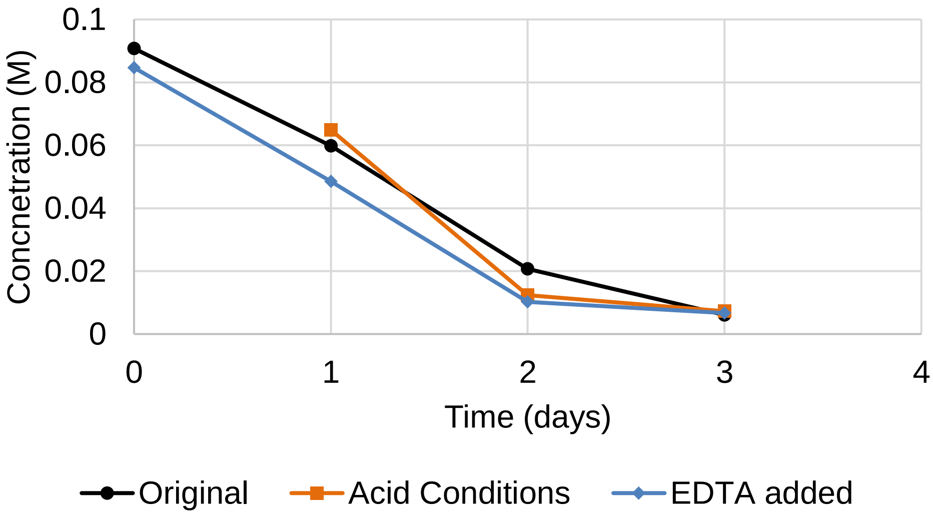


Figure E.1 The total quantity of glycine recovered as a weighted sum of all glycine species for SIPF samples heated over multiple 24-hour cycles. The expected total glycine concentration was 0.1 M. Using acidic conditions or EDTA to try to limit the persistence of glycine-copper complexes prior to analysis did not improve the amount of glycine recovered.

We were never able to fully explain the loss of glycine. No additional unknown peaks on the HPLC chromatogram rose proportionately to the loss of glycine. Longer glycine species did form in yields roughly comparable to those reported by Rode, though those also eventually began to decrease, quite possibly due to the loss of monomer glycine.

There are two possibilities that seem most likely to explain this behavior. First, glycine could still not be derivatizing successfully due to the copper, and the pH and EDTA were unable to disrupt that complex. This could be studied using ion pairing methods previously used by our lab (Sibilska et al. 2017) to determine if there is still a visible loss of glycine over time. Methods for inferring molecular structure, like NMR, could also potentially be used to characterize copper complexes directly. If glycine or glycine complexes still cannot be found, then it is possible that the amino acid is being degraded. If copper degrades amino acids, that behavior may have been noted somewhere in chemical or biochemical research, so related literature should be sought out. Mass spectrometry could also potentially be used to identify the degradation products, although the sample would need to be processed with CE or another clean up method or heavily diluted, considering the high salt concentrations involved.

Appendix E.2: Magnesium

The magnesium-ATP complex is an important step in activating ATP so that it can be used for biological processes (Phillips et al. 1965). Consequently, magnesium is discussed as a possible catalyst relatively frequently in prebiotic chemistry literature, particularly in the context of triphosphate or trimetaphosphate-activated reactions (Sawai & Orgel 1975; Yamagata & Inomata 1997; Oie et al. 1984; Hill & Orgel 2002; Kitadai et al. 2011). The experiments we performed were specifically inspired by van der Gulik et al. (2009), a theoretical study which looked at the conserved motifs in metal binding domains of modern proteins. They identified possible binding motifs involving aspartic acid and magnesium which were found in key enzymes for glycolysis and nucleotide polymerases. The simplest motif proposed was DGDGD.

Since this motif was relatively short and only required two amino acids, it was potentially accessible using our existing experimental and analytical methods. The first study we performed simply aimed to determine how the addition of magnesium affected TP-activated peptide bond formation. The details and complete results of this experiment can be found in COoL Boigenzahn/Experiments/2019-10-28_D+G_MgEffect. Briefly, samples were prepared using 16 mM glycine and 24 mM aspartic acid and heated in the presence of different salts at different pHs. Samples either included 20 mM TP, 20 mM $MgCl_2$, both, or none, and were heated at pH 9, pH 3, or the pH was not adjusted, which left them at about pH 4.5. Samples were heated at 80^{o}C with open caps, then rehydrated with 500 µL water to return them to their original volume and vortexed every 24 hours. Eventually, they were analyzed using HPLC and MS.

This experiment produced several encouraging results. The HPLC peaks were manually aligned, then this table was analyzed using principal component analysis (PCA) (Fig. E.2). We found that the samples including magnesium distinctly separated from the other samples and the samples including both TP and $MgCl_2$ fell between the samples including only TP or only $MgCl_2$, though they were much closer to the samples containing only TP. There was no consistent trend in the grouping of the different pH conditions.

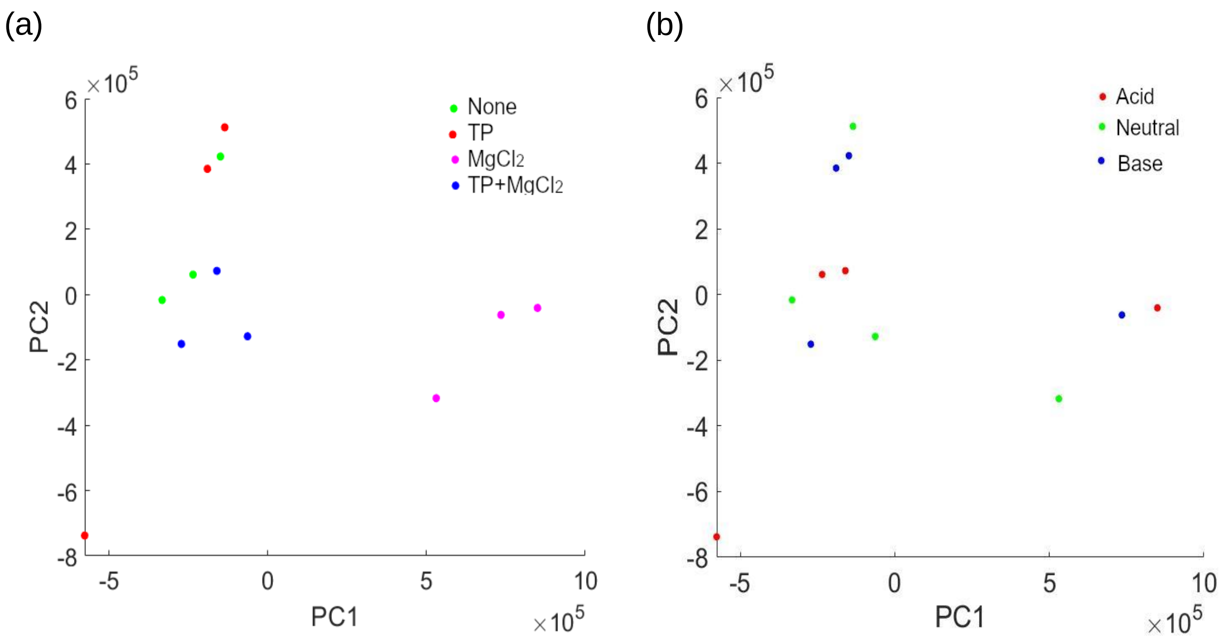


Figure E.2 PCA plot of HPLC peaks for aspartic acid and glycine dried with various salt and pH conditions. (a) The samples did appear to cluster based on the salt content, but there was (b) no apparent clustering based on the pH condition.

We also analyzed these samples using mass spectrometry. The methodology details are included in the write up in COoL Boigenzahn/Experiments/2019-10-28_D+G_MgEffect. The monomer quantitation for these experiments was later proven to be invalid, since we were using histidine as an internal spike which later turned out to be unreliable, but we were able to identify various peptides using MS/MS. The peptides we identified seemed to be enriched in aspartic acid, which was unexpected since glycine is typically the most reactive amino acid (Table E.1). By far the greatest diversity of peptides was found in samples containing both TP and $MgCl_2$, suggesting that this combination may improve peptide yields above TP alone. The enrichment of aspartic acid in the peptides identified may be related to the presence of magnesium, but it also may be influenced by the fact that, for unknown reasons, smaller peptides were rarely selected for MS/MS in any of our tests. Since glycine is significantly lower molar weight than aspartic acid, the software may have tended to prefer to select peptides including aspartic acid for

MS/MS. Although we found peptides up to six amino acids in length, we did not identify the DGDGD peptide we were originally looking for.

$MgCl_2$ + TP	$MgCl_2$	TP	None
DD	DD	DD	DD
DDD	DDDDG	DDD	GDDDGD
DDDD	GD	DDDGD	
DDDDD	GDDDD	DDGD	
DDDGD		DDGDD	
DDGD		DGD	
DDGDD		DGGD	
DDGDGD		GD	
DDGG		GDDDGD	
DDGGD			
DGD			
DGGD			
GD			
GDDDDG			
GDDDGD			
GDDGDD			
GDGD			
GGD			
GGDG			
GGGD			

Table E.1 **Peptides found in experiments combining D and G in various salt and pH conditions.** Peptides are listed by salt conditions they were identified in, since the effect of the salt condition was much more significant than the effect of the pH. For further breakdown by pH, see the summary of MS results in COoL Boigenzahn/Experiments/2019-10-28_D+G_MgEffect.

We performed several additional experiments studying how the relative concentrations of glycine and aspartic acid, the concentration of the magnesium salts, and the use of other salts besides $MgCl_2$ affected the outcomes. These studies were hindered by the fact that they were not performed in triplicate, and the results are therefore less reliable due to the inherent variability of using single samples. We also did not have peptide standards for most of the products of interest, meaning that even with triplicate samples, it would have been challenging to answer the questions originally proposed by the study.

These experiments were originally stopped because of lab closures due to COVID, and not followed up on afterwards as we explored other directions. To pursue these studies interaction of FMOC and aspartic acid would need to be validated, since the side chain of aspartic acid is more reactive than that of glycine or alanine, and peptide standards would need to be acquired. However, these studies are well within reach since we have successfully performed quantitative analyses of systems with two amino acids (Chapter 3). It may be interesting to revisit the motivating questions behind these studies and pursue similar experiments with our current, more developed HPLC methods.

APPENDIX F: SUGGESTED FUTURE STUDIES

Aside from the extensions of the work presented here, it may also be worthwhile to consider taking a top-down approach to exploring the emergence of autocatalytic peptides. One of the major challenges of the bottom-up approach is the extremely large possible experimental search space. There are many combinations of activating agents and amino acids which could be considered, which will all be affected by environmental conditions such as pH, temperature, or rehydrating steps. Bottom-up methods can help explore how those factors affect peptide formation, but they are an extremely difficult, long-shot approach to finding autocatalytic peptides.

Several autocatalytic peptides have been designed and characterized in literature, and the length of peptides identified as capable of autocatalysis has gotten much shorter in recent years. The longer part of the peptide template identified in Lee et al. (1996) and later extensively explored by the Ashkenasy group (Ashkenasy et al. 2004; Dadon et al. 2015) is 32 amino acids long. Autocatalytic behavior via β-sheet templating has been demonstrated with even shorter peptides; the template in Rubinov et al. (2009) is 12 amino acids long, and the template in Rout et al. (2022) is 8 amino acids long. These peptides could be used as a reference or goal to pursue achieving autocatalysis via bottom-up studies.

It should also be noted that existing literature examples of autocatalytic peptides almost always include capping groups on one or both terminal ends of the peptides. These groups may minimize side reactions, or may actively contribute to the templating behavior. The effect of the capping groups on autocatalysis and potential peptide bond formation would have to be assessed prior to beginning many of the experiments discussed in this section. Depending on the stability of the peptide and the caps on the functional groups, it may be possible for the custom peptide or its degradation products to form peptide bonds with the peptides meant to serve as the opposite half of the template. While this behavior is not necessarily inherently negative, it could substantially complicate the behavior of the system and the difficulty of analysis. The likelihood of these capping groups occurring on the prebiotic Earth should also be noted, though the presence of some prebiotically unlikely features can often be excused when pursuing specific behavior like autocatalysis.

For any of the experiments suggested below, the first step would be to repeat the experiments in the original paper as closely as possible using synthetic peptides to verify that our lab has a suitable analytical method in place for identifying autocatalytic behaviors and that the peptide constructs behave as expected in the conditions used in our lab. Reproducibility of experiments is, frankly, a known problem in scientific research (Stoddart 2016), and pursuing experiments directly inspired by behavior reported in literature without verifying that the expected behavior is reproducible is strongly not recommended. Assuming the autocatalytic behavior can be replicated and identified, there are several approaches that could be pursued.

One possible approach would be to create a steady-state mixture of peptides using replenishment and bottom-up mechanisms like those explored in this work, including all the amino acids required to form the shorter peptides involved in the templating mechanism. That mixture could be seeded with the longer autocatalytic template, then we could look for evidence that the addition of the template triggered autocatalysis. Ideally the environment used to form the bottom-up mixture would be compatible with the conditions required for autocatalysis, however, even if the conditions are slightly different, this experiment can probe how lack of selectivity in bottom-up experiments impact the potential for an autocatalytic system to form. If amino acids do not form the peptides required for autocatalysis due to having insufficient length or selectivity, this experiment could also be pursued by substituting the amino acids with short peptides which represent fragments of the templating peptides. In theory this should increase the length of accessible peptides and bias the resulting peptide distribution towards forming the correct sequence, though differences in amino acid reactivity could have unexpected effects on the peptide distribution.

This experiment would help explore how easily autocatalysis can be established and propagated during the origin of life. If autocatalysis can be achieved from short peptides formed using unguided prebiotic reactions, it suggests that once an autocatalytic reaction network was established, it may have been able to spread or reestablish itself relatively easily. Alternatively, if autocatalysis is difficult to reestablish without using nearly complete sections of the shorter peptide templates as reactants, it suggests that this templating mechanism is unstable, which makes it less to have been involved in the origin of life unless other molecules, reaction networks, or environmental factors helped stabilize it. In this scenario, it would be worthwhile to investigate why the autocatalytic behavior is difficult to recover, examining features like selectivity, waste accumulation, reaction rates or concentration thresholds to determine the limiting factors.

One of the benefits of adding a peptide template into a mixture of peptides formed through unguided synthesis is that it does not necessarily require complete characterization of the mixture – whether the peptide distribution has approximately reached steady state can be determined statistically without knowing the identity of the peaks, and only the peaks of the species involved in the autocatalytic network need to be fully quantified to compare their kinetics in the sample to the kinetics of the system when all templates are known to be present. The relationship between the kinetics and the template concentration could also be investigated and compared to the template dependence of the original system.

A variation of this experiment would be to cleave one or more of the autocatalytic peptides, then attempt to restore autocatalytic behavior using prebiotically plausible methods of peptide bond formation. Splitting the template could be done with varying levels of specificity, ranging from using synthetic peptides to represent breaking the template along a particular bond to allowing the template to freely degrade in conditions that favor hydrolysis. The more easily autocatalysis can be recovered after degradation, the more resilient it suggests the templating behavior to be. Finding the autocatalysis cannot be recovered would be an equally significant result, since it tends to suggest that finding autocatalysis via bottom-up approaches with the mechanism under study is likely to be quite difficult, since autocatalysis is difficult to reach even when it is known to be close by within the experimental space. However, this approach would probably require significant characterization of the degradation products, so it would be more challenging analytically.

Finally, it may be interesting to test shorter variations of the autocatalytic template to search for either the minimal structure required for templating or for intermediate catalytic behavior that precedes the onset of autocatalysis. Evolution often occurs in incremental steps, and autocatalytic behavior that arises gradually due to intermediate catalytic interactions seems more likely to create a stable life-like system than autocatalytic behavior that emerges or collapses abruptly with minor changes (Pascal et al. 2013). Based on the knowledge of what amino acids drive secondary structure formation, the effect of removing different amino acids from the templates could also potentially be studied to determine how the fidelity of secondary structure motifs affects the emergence of catalysis. These experiments would require extensive peptide standards for the different template variations, and do not fully address the question of how autocatalytic peptides could have arisen from unspecific reactions in complex mixtures. However, because the environmental conditions and mechanism of autocatalysis for the system would already be known, specific and quantitative kinetic experiments can be designed relatively easily.

APPENDICES: REFERENCES

Ashkenasy, G., Jagasia, R., Yadav, M., & Ghadiri, M. R. (2004). Design of a directed molecular network. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(30), 10872–10877. https://doi.org/10.1073/pnas.0402674101

Bar-Nun, A., Bar-Nun, N., Bauer, S. H., & Sagan, C. (1970). Shock synthesis of amino acids in simulated primitive environments. *Science*, *168*(3930), 470–473. https://doi.org/10.1126/science.168.3930.470

Belmonte, L., & Mansy, S. S. (2016). Metal catalysts and the origin of life. *Elements*, *12*(6), 413–418. https://doi.org/10.2113/gselements.12.6.413

Bromberg, Y., Aptekmann, A. A., Mahlich, Y., Cook, L., Senn, S., Miller, M., … Falkowski, P. G. (2022). Quantifying structural relationships of metal-binding sites suggests origins of biological electron transfer. *Science Advances*, *3984*(January), 1–14.

Burnham, S. C., Searson, D. P., Willis, M. J., & Wright, A. R. (2008). Inference of chemical reaction networks. *Chemical Engineering Science*, *63*(4), 862–873. https://doi.org/10.1016/j.ces.2007.10.010

Bütikofer, U., Fuchs, D., Bosset, J. O., & Gmür, W. (1991). Automated HPLC-amino acid determination of protein hydrolysates by precolumn derivatization with OPA and FMOC and comparison with classical ion exchange chromatography. *Chromatographia*, *31*(9–10), 441–447. https://doi.org/10.1007/BF02262386

Choi, B. K., Hercules, D. M., & Houalla, M. (2000). Characterization of polyphosphates by electrospray mass spectrometry. *Analytical Chemistry*, *72*(20), 5087–5091. https://doi.org/10.1021/ac000044q

Constantopoulos, T. L., Jackson, G. S., & Enke, C. G. (1999). Effects of salt concentration on analyte response using electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry*, *10*(7), 625–634. https://doi.org/10.1016/S1044-0305(99)00031-8

Crampin, E. J., Schnell, S., & McSharry, P. E. (2004). Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Progress in Biophysics and Molecular Biology*, *86*(1), 77–112. https://doi.org/10.1016/j.pbiomolbio.2004.04.002

Cronin, J. R., & Moore, C. B. (1971). Amino acid analyses of the Murchison, Murray, and allende carbonaceous chondrites. *Science*, *172*(3990), 1327–1329. https://doi.org/10.1126/science.172.3990.1327

Dadon, Z., Wagner, N., Alasibi, S., Samiappan, M., Mukherjee, R., & Ashkenasy, G. (2015). Competition and cooperation in dynamic replication networks. *Chemistry - A European Journal*, *21*(2), 648–654. https://doi.org/10.1002/chem.201405195

Doran, D., Clarke, E., Keenan, G., Carrick, E., Mathis, C., & Cronin, L. (2021). Exploring the sequence space of unknown oligomers and polymers. *Cell Reports Physical Science*, *2*(12), 100685. https://doi.org/10.1016/j.xcrp.2021.100685

Engel, M. H., & Nagy, B. (1982). Distribution and enantiomeric composition of amino acids in the Murchison meteorite. *Nature, 296*(5860), 837–840. https://doi.org/10.1038/296837a0

Forsythe, J. G., Petrov, A. S., Millar, W. C., Yu, S.-S., Krishnamurthy, R., Grover, M. A., Hud, N. V., & Fernández, F. M. (2017). Surveying the sequence diversity of model prebiotic peptides by mass spectrometry. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1711631114

Frost, D. C., Feng, Y., & Li, L. (2020). 21-plex DiLeu Isobaric Tags for High-Throughput Quantitative Proteomics. *Analytical Chemistry*, *92*(12), 8228–8234. https://doi.org/10.1021/acs.analchem.0c00473

Frost, D. C., Greer, T., & Li, L. (2015). High-resolution enabled 12-plex DiLeu isobaric tags for quantitative proteomics. *Analytical Chemistry*, *87*(3), 1646–1654. https://doi.org/10.1021/ac503276z

Fullmer, C. S., & Wasserman, R. H. (1979). Analytical peptide mapping by high performance liquid chromatography. *Journal of Biological Chemistry*, *254*(15), 7208–7212. https://doi.org/10.1016/s0021-9258(18)50305-7

Fürst, P., Pollack, L., Graser, T. A., Godel, H., & Stehle, P. (1990). Appraisal of four pre-column derivatization methods for the high-performance liquid chromatographic determination of free amino acids in biological materials. *Journal of Chromatography A*, *499*(C), 557–569. https://doi.org/10.1016/S0021-9673(00)97000-6

Gartenmann, K., & Kochhar, S. (1999). Short-chain peptide analysis by high-performance liquid chromatography coupled to electrospray ionization mass spectrometer after derivatization with 9-fluorenylmethyl chloroformate. *Journal of Agricultural and Food Chemistry*, *47*(12), 5068–5071. https://doi.org/10.1021/jf990710s

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, *20*(6), 1981–1996. https://doi.org/10.1093/bib/bby063

Glavin, D. P., Dworkin, J. P., & Sandford, S. A. (2008). Detection of cometary amines in samples returned by Stardust. *Meteoritics and Planetary Science*, *43*(1–2), 399–413. https://doi.org/10.1111/j.1945-5100.2008.tb00629.x

Haddad, P. R., Taraji, M., & Szücs, R. (2021). Prediction of Analyte Retention Time in Liquid Chromatography. *Analytical Chemistry*, *93*(1), 228–256. https://doi.org/10.1021/acs.analchem.0c04190

Hao, J., Sverjensky, D. A., & Hazen, R. M. (2019). Redox states of Archean surficial environments: The importance of H2,g instead of O2,g for weathering reactions. *Chemical Geology*, *521*(October 2018), 49–58. https://doi.org/10.1016/j.chemgeo.2019.05.022

Hennet, R. J. C., Holm, N. G., & Engel, M. H. (1992). Abiotic synthesis of amino acids under hydrothermal conditions and the origin of life: A perpetual phenomenon? *Naturwissenschaften*, *79*(8), 361–365. https://doi.org/10.1007/BF01140180

Henzel, W. J., Watanabe, C., & Stults, J. T. (2003). Protein identification: The origins of peptide mass fingerprinting. *Journal of the American Society for Mass Spectrometry*, *14*(9), 931–942. https://doi.org/10.1016/S1044-0305(03)00214-9

Hill, A., & Orgel, L. E. (2002). Trimetaphosphate-Induced Addition of Aspartic Acid to Oligo(glutamic acid)s. *Helvetica Chimica Acta*, *85*, 4244–4254.

Hsu, J. L., Huang, S. Y., Chow, N. H., & Chen, S. H. (2003). Stable-Isotope Dimethyl Labeling for Quantitative Proteomics. *Analytical Chemistry*, *75*(24), 6843–6852. https://doi.org/10.1021/ac0348625

Huber, C., & Wächtershäuser, G. (1998). Peptides by activation of amino acids with CO on (Ni,Fe)S surfaces: Implications for the origin of life. *Science*, *281*(5377), 670–671. https://doi.org/10.1126/science.281.5377.670

Huynh-Thu, V. A., & Geurts, P. (2018). DynGENIE3: Dynamical GENIE3 for the inference of gene networks from time series expression data. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-21715-0

Jámbor, A., & Molnár-Perl, I. (2009a). Amino acid analysis by high-performance liquid chromatography after derivatization with 9-fluorenylmethyloxycarbonyl chloride. *Journal of Chromatography A*, *1216*(15), 3064–3077. https://doi.org/10.1016/j.chroma.2009.01.068

Jámbor, A., & Molnár-Perl, I. (2009b). Quantitation of amino acids in plasma by high performance liquid chromatography: Simultaneous deproteinization and derivatization with 9-fluorenylmethyloxycarbonyl chloride. *Journal of Chromatography A*, *1216*(34), 6218–6223. https://doi.org/10.1016/j.chroma.2009.06.083

Jeanne Dit Fouque, D., Maroto, A., & Memboeuf, A. (2018). Internal Standard Quantification Using Tandem Mass Spectrometry of a Tryptic Peptide in the Presence of an Isobaric Interference. *Analytical Chemistry*, *90*(24), 14126–14130. https://doi.org/10.1021/acs.analchem.8b05016

Johnson, A. P., Cleaves, H. J., Dworkin, J. P., Glavin, D. P., Lazcano, A., & Bada, J. L. (2008). The Miller volcanic spark discharge experiment. *Science*, *322*(5900), 404. https://doi.org/10.1126/science.1161527

Keefe, A. D., Miller, S. L., Mcdonald, G., & Bada, J. (1995). Investigation of the prebiotic synthesis of amino acids and RNA bases from CO2 using FeS/H2S as a reducing agent. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(25), 11904–11906. https://doi.org/10.1073/pnas.92.25.11904

Kim, J. D., Pike, D. H., Tyryshkin, A. M., Swapna, G. V. T., Raanan, H., Montelione, G. T., … Falkowski, P. G. (2018). Minimal Heterochiral de Novo Designed 4Fe-4S Binding Peptide Capable of Robust Electron Transfer. *Journal of the American Chemical Society*, *140*(36), 11210–11213. https://doi.org/10.1021/jacs.8b07553

King, R., Bonfiglio, R., Fernandez-Metzler, C., Miller-Stein, C., & Olah, T. (2000). Mechanistic investigation of ionization suppression in electrospray ionization. *Journal of the American Society for Mass Spectrometry*, *11*(11), 942–950. https://doi.org/10.1016/S1044-0305(00)00163-X

Kitadai, N., Yokoyama, T., & Nakashima, S. (2011). Hydration-dehydration interactions between glycine and anhydrous salts: Implications for a chemical evolution of life. *Geochimica et Cosmochimica Acta*, *75*(21), 6285–6299. https://doi.org/10.1016/j.gca.2011.08.027

Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K., & Ghadiri, M. R. (1996). A self-replicating peptide. *Letters to Nature, 382*, 525–528.

Leis, H. J., Fauler, G., & Windischhofer, W. (1998). Stable isotope labeled target compounds: Preparation and use as internal standards in quantitative mass spectrometry. *Curr Org Chem*, 2(2), 131-144.

Li, W., Zhang, J., & Tse, F. L. S. (2011). Strategies in quantitative LC-MS/MS analysis of unstable small molecules in biological matrices. *Biomedical Chromatography*, *25*(1), 258–277. https://doi.org/10.1002/bmc.1572

Liu, Y., Chen, S., Zhang, J., Li, X., & Gao, B. (2017). Stimulation effects of ciprofloxacin and sulphamethoxazole in Microcystis aeruginosa and isobaric tag for relative and absolute quantitation-based screening of antibiotic targets. *Molecular Ecology*, *26*(2), 689–701. https://doi.org/10.1111/mec.13934

Liu, Z., Tu, M. J., Zhang, C., Jilek, J. L., Zhang, Q. Y., & Yu, A. M. (2019). A reliable LC-MS/MS method for the quantification of natural amino acids in mouse plasma: Method validation and application to a study on amino acid dynamics during hepatocellular carcinoma progression. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, *1124*(March), 72–81. https://doi.org/10.1016/j.jchromb.2019.05.039

Lowe, C. U., Rees, M. W., & Markham, R. (1963). Synthesis of complex organic compounds from simple precursors: Formation of amino acids, amino acid polymers, fatty acids and purines from ammonium cyanide. *Nature, 151*(4890), 219–222.

Markowetz, F., & Spang, R. (2007). Inferring cellular networks - A review. *BMC Bioinformatics*, *8*(SUPPL. 6). https://doi.org/10.1186/1471-2105-8-S6-S5

Marshall, W. L. (1994). Hydrothermal synthesis of amino acids. *Geochimica et Cosmochimica Acta*, *58*(9), 2099-2106.

Mayes, E. L. V. (1984). Peptide Mapping by Reverse-Phase High Pressure Liquid Chromatograph. *Walker, J.M. (Eds) Proteins. Methods in Molecular Biology*^{TM}, *1*. https://doi.org/https://doi.org/10.1385/0-89603-062-8:33

Miller, Stanley L., and Leslie E. Orgel. "*The origins of life on the earth*." (1974). Prentice-Hall.

Miyakawa, S., Yamanashi, H., Kobayashi, K., Cleaves, H. J., & Miller, S. L. (2002). Prebiotic synthesis from CO atmospheres: Implications for the origins of life. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 14628–14631. https://doi.org/10.1073/pnas.192568299

Molnár-Perl, I. (2011). Advancement in the derivatizations of the amino groups with the o-phthaldehyde-thiol and with the 9-fluorenylmethyloxycarbonyl chloride reagents. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, *879*(17–18), 1241–1269. https://doi.org/10.1016/j.jchromb.2011.01.027

Mortensen, P., Gouw, J. W., Olsen, J. V., Ong, S. E., Rigbolt, K. T. G., Bunkenborg, J., Cox, J., Foster, L. J., Heck, A. J. R., Blagoev, B., Andersen, J. S., & Mann, M. (2010). MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *Journal of Proteome Research*, *9*(1), 393–403. https://doi.org/10.1021/pr900721e

Napier, J., & Yin, J. (2006). Formation of peptides in the dry state. *Peptides*, *27*(4), 607–610. https://doi.org/10.1016/j.peptides.2005.07.015

Oie, T., Loew, G. H., Burt, S. K., & MacElroy, R. D. (1984). Quantum Chemical Studies of a Model for Peptide Bond Formation. 3. Role of Magnesium Cation in Formation of Amide and Water from Ammonia and Glycine. *Journal of the American Chemical Society*, *106*(26), 8007–8013. https://doi.org/10.1021/ja00338a001

Ottensmann, M., Stoffel, M. A., Nichols, H. J., & Hoffman, J. I. (2018). GCalignR: An R package for aligning gas-chromatography data for ecological and evolutionary studies. *PLoS ONE*, *13*(6), 1–20. https://doi.org/10.1371/journal.pone.0198311

Parker, E. T., Cleaves, H. J., Callahan, M. P., Dworkin, J. P., Glavin, D. P., Lazcano, A., & Bada, J. L. (2011). Enhanced Synthesis of Alkyl Amino Acids in Miller's 1958 H 2S Experiment. *Origins of Life and Evolution of Biospheres*, *41*(6), 569–574. https://doi.org/10.1007/s11084-011-9253-2

Pascal, R., Pross, A., & Sutherland, J. D. (2013). Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open Biology*, *3*(11), 1–9. https://doi.org/10.1098/rsob.130156

Phillips, R. C., George, P., & Rutman, R. J. (1965). Thermodynamic Studies of the Formation and Ionization of the Mangesium(II) Complex of ADP and ATP over the pH range 5 to 9. *Biological Chemistry*, *265*(2), 2631–2640.

Ramos-Payán, M.; Ocaña-Gonzalez, J. A.; Fernández-Torres, R. M.; Llobera, A.; Bello-López, M. Á., Recent trends in capillary electrophoresis for complex samples analysis: A review. *LECTROPHORESIS* **2018,** *39* (1), 111-125.

Rode, B. M., Eder, A. H., & Yongyai, Y. (1997). Amino acid sequence preferences of the salt-induced peptide formation reaction in comparison to archaic cell protein composition. *Inorganica Chimica Acta*, *254*(2), 309–314. https://doi.org/10.1016/S0020-1693(96)05178-X

Rode, B. M., & Schwendinger, M. G. (1990). Copper-Catalyzed Amino Acid Condensation in Water - A Simple Possible Way of Prebiotic Peptide Formation. *Origins of Life and Evolution of the Biosphere, 20*(Ii), 401–410.

Roturier, J. M., Le Bars, D., & Gripon, J. C. (1995). Separation and identification of hydrophilic peptides in dairy products using FMOC derivatization. *Journal of Chromatography A, 696*(2), 209–217. https://doi.org/10.1016/0021-9673(94)01234-6

Rout, S. K., Rhyner, D., Riek, R., & Greenwald, J. (2022). Prebiotically Plausible Autocatalytic Peptide Amyloids. *Chemistry - A European Journal*, *28*(3). https://doi.org/10.1002/chem.202103841

Rubinov, B., Wagner, N., Rapaport, H., & Ashkenasy, G. (2009). Self-Replicating Amphiphilic β-Sheet Peptides. *Angewandte Chemie*, *121*(36), 6811–6814. https://doi.org/10.1002/ange.200902790

Sauer, C. S., & Li, L. (2022). Multiplexed quantitative neuropeptidomics via DiLeu isobaric tagging. In *Methods in Enzymology* (1st ed., Vol. 663). Elsevier Inc. https://doi.org/10.1016/bs.mie.2021.10.011

Sawai, H., & Orgel, L. E. (1975). Prebiotic peptide-formation in the solid state - III. Condensation Reactions of Glycine in Solid State Mixtures Containing Inorganic Polyphosphates. *Journal of Molecular Evolution*, *6*(3), 185–197. https://doi.org/10.1007/BF01732355

Shimoyama, A., & Ponnamperuma, C. (1979). Amino acids in the Yamato carbonaceous chondrite from Antarctica. *Nature*, *282*(5737), 394–396.

Sibilska, I., Chen, B., Li, L., & Yin, J. (2017). Effects of Trimetaphosphate on Abiotic Formation and Hydrolysis of Peptides. *Life*, *7*(4), 1–11. https://doi.org/10.3390/life7040050

Stoddart, C. (2016). Is there a reproducibility crisis in science? *Nature*, 3–5. https://doi.org/10.1038/d41586-019-00067-3

Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., & Kohlbacher, O. (2008). OpenMS - An open-source software framework for mass spectrometry. *BMC Bioinformatics*, *9*, 1–11. https://doi.org/10.1186/1471-2105-9-163

Sutherland, J. D. (2016). The Origin of Life - Out of the Blue. *Angewandte Chemie - International Edition*, *55*(1), 104–121. https://doi.org/10.1002/anie.201506585

Timm, J., Pike, D. H., Mancini, J. A., Tyryshkin, A. M., Poudel, S., Siess, J. A., … Nanda, V. (2023). Design of a minimal di-nickel hydrogenase peptide. *Science Advances*, *9*(10), eabq1990. https://doi.org/10.1126/sciadv.abq1990

van der Gulik, P., Massar, S., Gilis, D., Buhrman, H., & Rooman, M. (2009). The first peptides: The evolutionary transition between prebiotic amino acids and early proteins. *Journal of Theoretical Biology*, *261*(4), 531–539. https://doi.org/10.1016/j.jtbi.2009.09.004

Wang, M. S., Hoegler, K. J., & Hecht, M. H. (2019). Unevolved de novo proteins have innate tendencies to bind transition metals. *Life*, *9*(1), 1–15. https://doi.org/10.3390/life9010008

Wolman, Y., Haverland, W. J., & Miller, S. L. (1972). Nonprotein Amino Acids from Spark Discharges and Their Comparison with the Murchison Meteorite Amino Acids. *Proceedings of the National Academy of Sciences*, *69*(4), 809–811. https://doi.org/10.1073/pnas.69.4.809

Yamagata, Y., & Inomata, K. (1997). Condensation of glycylglycine to oligoglycines with trimetaphosphate in aqueous solution. II: Catalytic effect of magnesium ion. *Origins of Life and Evolution of the Biosphere*, *27*(4), 339–344. https://doi.org/10.1023/A:1006529421813

Yamamoto, S., Watler, P. K., Feng, D., & Kaltenbrunner, O. (1999). Characterization of unstable ion-exchange chromatographic separation of proteins. *Journal of Chromatography A*, *852*(1), 37–41. https://doi.org/10.1016/S0021-9673(99)00592-0

Yoshino, D., Hayatsu, K., & Anders, E. (1971). Origin of organic matter in early solar system—III. Amino acids: Catalytic synthesis. *Geochimica et cosmochimica acta*, 35(9), 927-938.