

2018

# Multi-scale metabolism: from the origin of life to microbial ecology

---

<https://hdl.handle.net/2144/33239>

*"Downloaded from OpenBU. Boston University's institutional repository."*

BOSTON UNIVERSITY  
GRADUATE SCHOOL OF ARTS AND SCIENCES  
AND  
COLLEGE OF ENGINEERING

Dissertation

**MULTI-SCALE METABOLISM:  
FROM THE ORIGIN OF LIFE TO MICROBIAL ECOLOGY**

by

**JOSHUA E. GOLDFORD**

B.S., University of Minnesota, 2010  
B.S., University of Minnesota, 2010  
M.S., University of Minnesota, 2013

Submitted in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

2018

© 2018 by  
JOSHUA ELLIOT GOLDFORD  
All rights reserved

Approved by

First Reader

---

Daniel Segrè, Ph.D.  
Professor of Biology  
Professor of Bioinformatics  
Boston University, College of Arts and Sciences  
  
Professor of Biomedical Engineering  
Professor of Materials Science and Engineering  
Boston University, College of Engineering

Second Reader

---

Pankaj Mehta, Ph.D.  
Associate Professor of Physics

Third Reader

---

Jennifer Talbot, Ph.D.  
Assistant Professor of Biology

## **Dedication**

This dissertation is dedicated to my father.

## **Acknowledgments**

There are many people I would like to thank for making graduate school at Boston University a positive experience for me.

I'd first like to thank all of the current and former members of the Segrè, Mehta and Sanchez groups involved in various project over the years including: Demetrius Dimucci, David Bernstein, Alan Pacheno, Dr. Ilija Dukovski, Dileep Kishore, Dr. Shany Ofaim, Mike Quintin, Meghan Thommes, Dr. Ali Zomorodi, Dr. Brian Granger, Dr. Ed Reznik, Dr. Arion Stettner, Dr. Christopher Jacobs, Elena Forchielli, Ching-Hao Wang, Alex Day, Wenping Cui, Dr. Robert Marsland, Dr. Nanxi Lu, Dr. Sylvie Estrela, Dr. Djordje Bajic, and Dr. Alicia Sanchez-Gorostiaga. I especially would like to thank Dr. Melisa Osborne for helping me get started in the wet-lab at the Rowland Institute. I am very appreciative of the scientific and professional advice provided by Dr. Jennifer Tablot, Dr. Kirill Korolev and Dr. Temple Smith. I am also indebted to my mentor during my masters degree, Dr. Igor Libourel, for taking time and energy to continually offer advice throughout graduate school even after I moved to Boston to start my PhD.

The completion of my PhD would have not been possible without the help of the amazing administrative staff in the BU Bioinformatics program. I am very thankful to have had consistent administrative support from Dave King, Johanna Vasquez, and Caroline Lyman, as well as amazing technical support from Mary-Ellen Fitzpatrick.

I would like to specifically thank my collaborator Dr. Hyman Hartman for convincing me to start thinking about the origin of life and the ancient evolution of metabolism. Hyman regularly made time to discuss science with me, provided me with papers and text-books in biochemistry, and continues to inspire me in this field of research.

My mentors have helped me tremendously throughout my graduate studies. I am very grateful for being able to work with Dr. Alvaro Sanchez on a highly collaborative project with so many of his lab members. Alvaro gave me space to help drive a project within the field of microbial community ecology, a field I was particularly interested in during graduate school. I also am indebted to Dr. Pankaj Mehta for both scientific and philosophical mentorship, as well as the

countless cups of coffee he purchased for me during our impromptu meetings.

My time in graduate school was such a positive experience thanks to my advisor, Daniel Segrè. Daniel's creative and positive approach has forever shaped my relationship with science and mentorship. I am forever grateful for the time he spent meeting with me as well as his unyielding confidence in my science.

Last but not least, I'd like to thank my friends and family for the support during my time in graduate school. These last five years would have been a lot more challenging without Zach, Lou, Gina, my Mom and Dad listening to me ramble on about metabolism.

**MULTI-SCALE METABOLISM:  
FROM THE ORIGIN OF LIFE TO MICROBIAL ECOLOGY**

**JOSHUA E. GOLDFORD**

Boston University Graduate School of Arts and Sciences and College of Engineering, 2018

Major Professor: Daniel Segrè, Professor of Biology, Professor of Bioinformatics,  
Professor of Biomedical Engineering, Professor of Materials Science  
and Engineering

**ABSTRACT**

Metabolism is a key attribute of life on Earth at multiple spatial and temporal scales, involved in processes ranging from cellular reproduction to biogeochemical cycles. While metabolic network modeling approaches have enabled significant progress at the cellular-scale, extending these techniques to address questions at both the ecosystem and planetary-scales remains highly unexplored. In this thesis, I integrate various multi-scale metabolic network modeling approaches to address key questions with regard to both the long-term evolution of metabolism in the biosphere and the metabolic processes that take place in complex microbial communities.

The first portion of my thesis work, focused on the evolution of ancient metabolic networks, attempts to model the emergence of ecosystem-level metabolism from simple geochemical precursors. By integrating network-based algorithms, physiochemical constraints, and geochemical estimates of ancient Earth, I explored whether a complex metabolic network could have emerged without phosphate, a key molecular component in modern-day living systems, known to be poorly available at the onset of life. We found that phosphate may have not been essential in early living systems, and that thioesters may have been the primitive energy currency in ancient metabolic networks. By generalizing this approach to explore the scope of geochemical scenarios that could have given rise to living systems, I found that other key biomolecules, including fixed nitrogen, may have not been required at the earliest stages in biochemical evolution. The second portion of my thesis deals with a different aspect of ecosystem-level metabolism, namely the role of metabolism in shaping the structure of microbial communities. I studied the relationship between metabolism and microbial community assembly using microbial communities grown in synthetic laboratory

environments. We found that a generalized statistical consumer-resource model recapitulates the emergent phenomena observed in these experiments.

Future work could seek to better clarify the connection between the fundamental rules that led to life's emergence over 4 billion years ago and the laws that shape microbial ecosystems today. An ecosystems-level metabolic perspective may aid in our understanding of both the emergence and maintenance of the biosphere.

# Contents

<b>1</b>	<b>Introduction and background</b>	<b>1</b>
<b>2</b>	<b>Architecture of ancient metabolic networks without phosphate</b>	<b>12</b>
<b>3</b>	<b>Ancient geochemical scenarios converge to an organo-sulfur proto-metabolism</b>	<b>41</b>
<b>4</b>	<b>Emergence of community-level function in microbial community assembly</b>	<b>71</b>
<b>5</b>	<b>Discussion and perspective</b>	<b>115</b>
	<b>Bibliography</b>	<b>120</b>
	<b>Curriculum Vitae</b>	<b>137</b>

## List of Figures

1.1	Metabolism at different scales. . . . .	4
1.2	Towards a model of ancient metabolism . . . . .	6
2.1	Biosphere-level metabolic network without phosphate . . . . .	17
2.2	Robustness of non-phosphate core network to variations in seed compounds . . . . .	18
2.3	Network expansion with various carbon sources . . . . .	19
2.4	Core network enzymes are enriched for iron-sulfur and transition metal coenzymes	22
2.5	Thioesters alleviate thermodynamic bottlenecks . . . . .	25
2.6	Thiols are required for thermodynamically feasible network expansion. . . . .	26
2.7	Global non-phosphate metabolism . . . . .	27
2.8	Coenzymes before phosphate . . . . .	28
3.1	Computational pipeline for the reconstruction of ancient metabolic models . . . . .	44
3.2	Nitrogen is not essential for expansion . . . . .	46
3.3	Reduction potential of redox coenzymes influences network expansion . . . . .	48
3.4	Requirements . . . . .	50
3.5	Constraint-based modeling of plausible ancient proto-cells . . . . .	54
3.6	Catalysts in thioester-driven proto-metabolism are depleted in nitrogen . . . . .	68
3.7	Thiols are required for autotrophic expansion and fatty acid production . . . . .	69
3.8	Putative ancient catalysts . . . . .	70
3.9	Enzymes catalyzing reactions before the addition of ammonia are not depleted in nitrogen containing amino acids relative to enzymes added after ammonia . . . . .	70
4.1	Characterization and diversity of microbiomes isolated from plant and soil samples	74

4.2	Top down assembly of bacterial consortia . . . . .	75
4.3	Dynamics of <i>ex-situ</i> microbial communities . . . . .	76
4.4	Presence of rare taxa in <i>ex-situ</i> assembled microbial communities . . . . .	77
4.5	Low levels of bacterial growth with no externally supplied carbon source . . . . .	78
4.6	Four strains from a representative community coexist in reconstituted communities	79
4.7	The community structure from the same inocula can be highly variable and the genus level, but similar at the family level . . . . .	82
4.8	Strongly deterministic population dynamics of replicate populations . . . . .	83
4.9	The dilution factor likely does not induce substantial variation in community structure	84
4.10	Family-level and metagenomic attractors are associated with different carbon sources	85
4.11	Community structure on citrate minimal media . . . . .	86
4.12	Community structure on leucine minimal media . . . . .	87
4.13	Family-level features associated the carbon source . . . . .	88
4.14	Family-level composition is a strong predictor of the limiting carbon source . . . . .	89
4.15	Widespread metabolic facilitation stabilizes resource competition . . . . .	92
4.16	Assembly of microbial communities in well-shaken cultures . . . . .	93
4.17	Resource abundance on the growth rates of individual species . . . . .	94
4.18	A simple extension of classic ecological models recapitulates several experimental observations . . . . .	97
4.19	Generation of families of consumers in consumer resource models . . . . .	98
4.20	Functional clustering is observed in both consumer resource models and experiments	99

## List of Symbols

CoA	.....	Coenzyme A
FBA	.....	Flux Balance Analysis
$\Delta_r G'^{\circ}$	.....	Free energy of reaction $r$ at standard molar conditions
LUCA	.....	Last universal common ancestor
(M)CRM	...	(Microbial) Consumer Resource Model
MILP	.....	Mixed-integer linear program
NAD(P)	....	Nicotinamide adenine dinucleotide (phosphate)
$P_i$	.....	Inorganic phosphate
$S$	.....	Stoichiometric matrix
rTCA cycle	.	Reductive Tricarboxylic Acid Cycle
TMFA	.....	Thermodynamic Metabolic Flux Analysis
WL-Pathway		Wood-Ljungdahl Pathway

## Chapter 1

### Introduction and background

This introduction was published as the following Review Article:

**Goldford, J. E.** and & Segrè, D. *Modern views of ancient metabolic networks.*

*Current Opinion in Systems Biology.* 2018 Apr; (8) 117-124 [64]

#### Summary

Metabolism is a molecular, cellular, ecological and planetary phenomenon, whose fundamental principles are likely at the heart of what makes living matter different from inanimate one. Systems biology approaches developed for the quantitative analysis of metabolism at multiple scales can help understand metabolism's ancient history. In this review, we highlight work that uses network-level approaches to shed light on key innovations in ancient life, including the emergence of proto-metabolic networks, collective autocatalysis and bioenergetics coupling. Recent experiments and computational analyses have revealed new aspects of this ancient history, paving the way for the use of large datasets to further improve our understanding of life's principles and abiogenesis.

#### Introduction

The metabolic network of a cell transforms free energy and environmentally available molecules into more cells, moving electrons step by step along gradients in a complex energetic landscape [145, 23]. The ability of a cell to efficiently and simultaneously manage hundreds of metabolic processes so as to accurately balance the production of its internal components constitutes a very complex resource allocation problem. In fact, it is only through recent systems biology research

that we have begun to quantitatively assess this resource allocation problem at the whole-cell level [144, 14]. A common perspective in the analysis of cellular self-reproduction is the notion that the genome, with its crucial information-storage role, is the central molecule of the cell, and that everything else can be collectively regarded as the machinery whose role is to produce a copy of the DNA. It is therefore not surprising that, as we struggle with the fascinating question of how life started on a lifeless planet, it is tempting to look for how a single information-containing molecule could arise spontaneously from prebiotic compounds. However, in spite of the appeal of thinking of DNA (or its historically older predecessor, RNA) as the central molecule who is being replicated in the cell, no molecule in the cell really self-replicates: the cell is a network of chemical transformations capable of collective autocatalytic self-reproduction. Collective autocatalysis is the capacity for a collection of chemicals to enhance or catalyze the synthesis or import of its own components, enabling a positive feedback mechanism that can lead to their sustained amplification. Combining this systems-level view of a cell with the argument of what is usually called the metabolism first view of the origin of life, one could propose that the ability of a chemical network to produce more of itself (or to grow autocatalytically) is and has always been a key hallmark of life [171, 45, 90, 6, 175]. An interesting modern version of this very same principle is embedded in one of the most popular systems biology approaches for the study of whole cell metabolism: this approach, based on reaction network stoichiometry and efficient constraint-based optimization algorithms, is commonly known as flux balance analysis (FBA) [144]. FBA solves mathematically the resource allocation problem that every living cell needs to solve in real life in order to transform available nutrients into the macromolecular building blocks that are necessary for maintenance and reproduction. When an FBA calculation estimates the maximal growth capacity of a cell, it essentially computes the set of reaction network fluxes that enable optimally efficient autocatalytic self-reproduction. While in cellular life this process is finely regulated and controlled, ancient life must have gone through many different stages of similar, but much less organized collectively autocatalytic processes. Thus, one of the key problems of the origin of life is the question of how an initially random path in the space of possible chemical transformations driven far from thermodynamic equilibrium could have ended up being dynamically trapped in a collectively autocatalytic

state.

The focus on cellular self-reproduction as the fundamental level at which life and its origin should be understood is however too narrow. An exciting recent development in systems biology of metabolism is the rise of methods to extend FBA models from the genome scale to the ecosystem level [172, 75, 55]. In addition to solving the resource allocation problem of metabolism for individual organisms in a given environment, these approaches take into account the fact that metabolites can be exchanged across species, giving rise to metabolically-driven ecological networks (11). These advances suggest that metabolism may be best understood as an ecosystem-level phenomenon (Fig. 1.1b), where the collective biochemical capabilities of multiple co-existing organisms may reflect better than any individual metabolic network an optimal capacity of life to utilize resources present in a given environment [175, 19]. The ecosystem-level nature of metabolism is another feature of present-day life whose roots likely date back to the early stages of life on our planet. For example, the chemical networks that gradually gave rise to reproducing protocells may have wandered for quite some time in a broader chemical space, effectively generating molecular ecosystems before the rise of spatially and chemically well-defined cellular structures.

At an even larger scale, metabolism could be viewed as operating not just at the level of individual cells or ecosystems, but even as a planetary phenomenon, in which cellular processes collectively affect (and are affected by) the flow of molecules at geological scales (Fig 1c). The strong coupling between the metabolic processes of ecosystems and planetary-scale geochemistry [86, 132] suggest that biosphere-level metabolism should be viewed as one of the natural scales for the study of life's history. A paramount challenge in the study of life's history is thus bridging the gap between material and energy fluxes at the biosphere scale, and detailed molecular mechanisms responsible for the properties of life at the cellular and subcellular level [121]. Bridging this gap could greatly benefit from the use of integrative models similar to the ones used in systems biology research and data science. In this perspective, we will discuss some recent system-level approaches that have provided new important insight into life's ancient history at multiple scales, highlighting the fact that the metabolism and its multiscale nature from the single reaction to the biosphere are taking a center stage role in this endeavor.

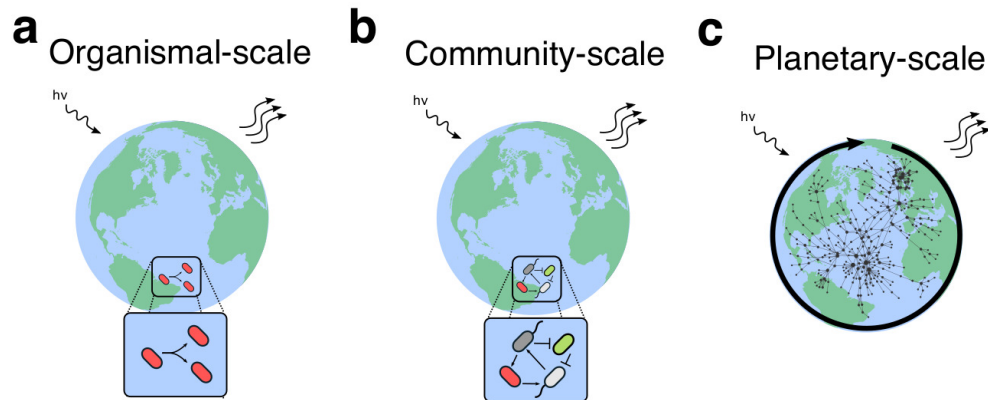


Figure 1.1: **Metabolism at different scales.** (a) metabolic networks can be modeled at the organismal level, where environmentally-supplied resources, under an energy flow, are collectively transformed into biomass of the self-reproducing and evolving organisms. (b) at higher scales, metabolic networks can be viewed as an ecosystem-level phenomenon, where biochemical processes include metabolic exchange and competition between species. (c) metabolism can be also considered a planetary scale phenomenon, whereby the energy flow maintains global biogeochemical cycles.

### Protometabolism before enzymes

A top-down reconstruction of ancient metabolic networks can be achieved based on the inferred history of gene families, using traditional phylogenomic techniques [22, 20, 88]. Leveraging information on the newly mapped genomic diversity of modern life [198], Martin and colleagues recently proposed a comprehensive phylogenetic reconstruction of the metabolic capabilities of the last universal common ancestor (LUCA), suggesting that LUCA was an autotrophic, thermophilic,  $N_2$ -fixing anaerobic prokaryote, living in hydrothermal vents and equipped with life's most complex molecular machines (e.g. ATP synthase) [196]. Although the details of LUCA's specific repertoire of metabolic enzymes are still subject of debate [59, 195], these results corroborate the notion that LUCA was very complex, highlighting a massive gap in knowledge with regard to the transition from prebiotic geochemical processes to the biochemical complexity of LUCA and its progeny. A major challenge in the study of the origin of metabolic networks is to gain insight on the structure of metabolic networks before LUCA and before the rise of genetic coding. At the core of this challenge is the question of whether and how metabolic reactions which depend on genome-encoded

enzymes in modern cells could have been carried out without such enzymes, resulting in a classical origin of life chicken-and-egg problem. One possible way out of this conundrum is the possibility that some of these metabolic reactions were initially catalyzed by less sophisticated and less specific catalysts, such as small organic molecules, metal ions, minerals, short RNA polymers, prebiotic amino acids or peptides. These small molecules could have persisted throughout evolution, gradually becoming incorporated into protein enzymes as catalytic cores or cofactors [192, 197]. Adding to a large body of evidence on individual metabolic reactions being catalyzed by small molecule [166, 106], recent experimental work has demonstrated that several key pathways found in modern day metabolic networks can be catalyzed non-enzymatically [97, 96, 127, 135, 180]. For instance, Ralser and colleagues [97, 96, 127, 95] have shown the feasibility of non-enzymatic networks that resemble modern day biochemical pathways, including the TCA cycle, glycolysis and gluconeogenesis. In addition, Moran and colleagues have demonstrated that metals can selectively catalyze and drive portions of non-enzymatic reductive TCA cycle (rTCA) [135]. These experimental results support the hypothesis that the catalytic cores of some modern enzymes may represent evolved variants of simple geochemically available prebiotic catalysts like transition metals, iron-sulfur clusters or organic cofactors [101]. Despite these important advances, the known instances of non-enzymatic catalysis are still the tip of the iceberg relative to the large number of possible catalyst-reaction pairs. Future high-throughput experiments could greatly expand the scope of possible prebiotic chemical networks and test the limits of non-enzymatic catalysis, shedding important light on the complexity of prebiotic chemistry obtainable before the availability of protein-coded enzymes.

### **From non-enzymatic catalysis to collective autocatalysis**

The above examples illustrate the fact that chemical reactions, and whole pathways, typically viewed in biological context as feasible only in the presence of protein enzymes, could take place under much more primitive conditions, through the catalytic action of small molecules, minerals or even non-covalent supermolecular assemblies [170]. As mentioned above, however, a major leap

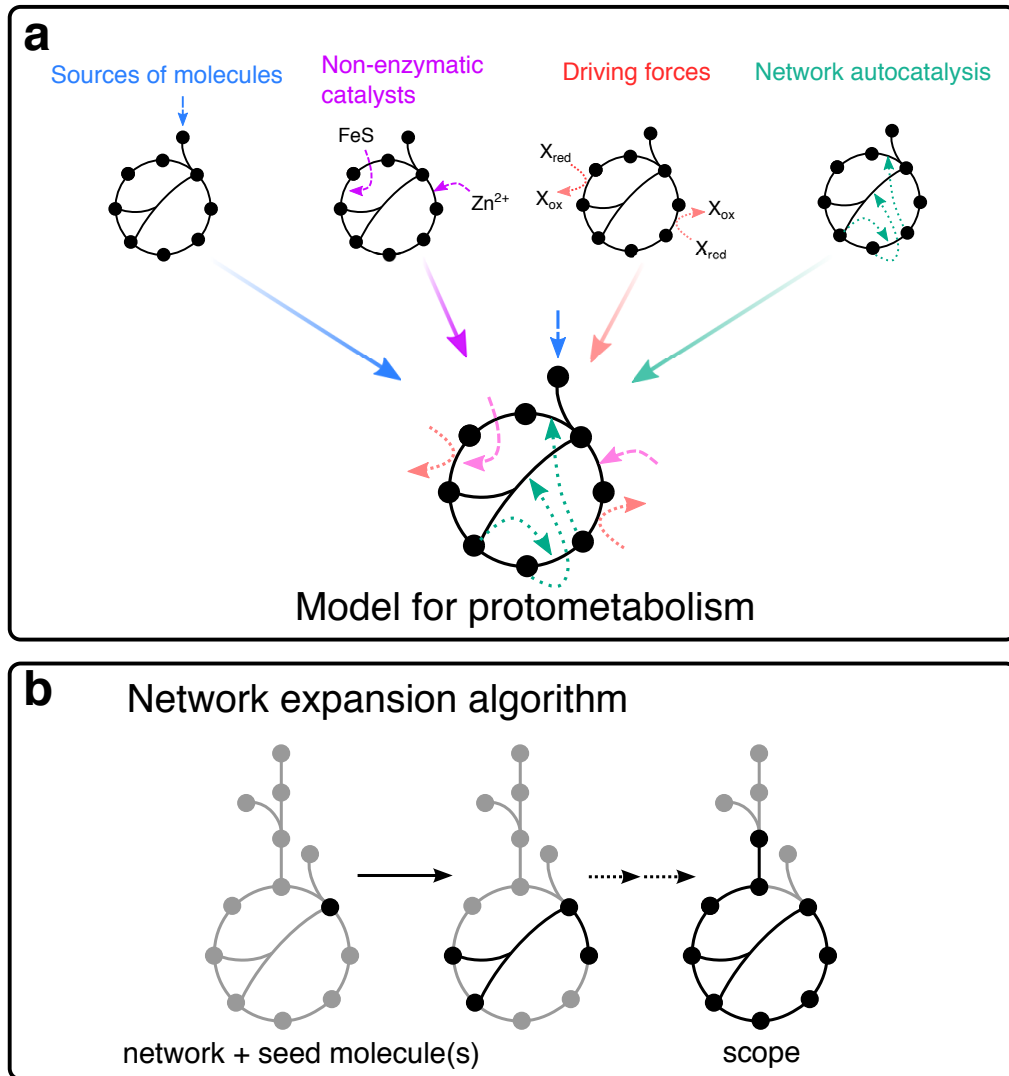


Figure 1.2: **Towards a model of ancient metabolism** (a) models of protometabolism can be constructed using a wide range of data including geochemically-supported data of environments and atmospheres in the early Archaean Eon, knowledge of non-enzymatic chemical reactions, plausible driving forces keeping protometabolism out of equilibrium and a mechanism for sustained growth (e.g. network autocatalysis). (b) The structure of plausible networks can be investigated using the network expansion algorithm which models the integrative expansion of metabolic networks from a set of seed compounds and allowable chemical reactions.

in the history of life must have involved the rise of a collectively autocatalytic chemical system. The feasibility of pre-enzymatic chemistry suggests a possible path for the rise of such collective autocatalysis: if the molecules produced by these reactions are themselves good catalysts, or if these reactions contribute to solubilize from rocks inorganic catalysts, there is a chance that a subset of reactions and molecules will effectively display a dynamic behavior that is equivalent to that of a single autocatalytic, exponentially growing entity [45, 90, 6, 175].

Recent insight into how these autocatalytic sets may have operated has come from both theoretical and experimental work. Recent theoretical work has uncovered generic constraints of autocatalytic networks [13], and offered plausible biophysical mechanisms leading to sustained autocatalysis of biopolymer ensembles [69, 102]. Experimentally, Whitesides and colleagues have constructed an autocatalytic chemical network based on simple, biologically relevant organic compounds [173]. In particular, by using a continuous flow of nutrients into and out of their reaction vessel, they showed that simple mixtures of thiols and thioesters could display a wide range of dynamical properties, such as bistability, oscillations and autocatalysis. Notably, this work demonstrated that dynamical properties observed in biological networks can emerge from simple mixtures of prebiotically plausible chemicals held out of equilibrium. As described recently by Vetsigian and Baum, the time is ripe for experimental explorations of how collectively autocatalytic cycles could spontaneously arise from mixtures of small molecules and mineral surfaces [16].

### **Navigating possible paths from primordial to present-day networks**

Prior studies in evolutionary biology suggest that biological systems evolve by partially building on prior innovations. If this principle extends back to the origin of metabolic networks, then it is reasonable to hypothesize that early proto-metabolic networks were based on previously accessible chemistry. Such logic leads to the conjecture that the structure of metabolic networks encodes the evolutionary history of metabolism, and that the chemistry of core metabolism is similar to the initial abiotic chemical networks that lead to life's emergence [77, 133]. This conjecture is supported, as discussed above, by experimental work demonstrating that a significant portion of

core metabolism is accessible without the use of protein-coded enzymes. While these concepts have been heavily utilized in origin of life research, recent efforts have transformed this conceptual paradigm into an algorithmic and quantitative framework using metabolic network modeling [20, 61]. A modeling approach recently used to explore the plausible evolutionary history of very early stages of biochemistry is the network expansion algorithm, which iteratively simulates the growth of new metabolites and reactions starting from an initial seed set [47, 73, 160]. We used the network expansion algorithm to construct a model for ancient prebiotic metabolism, specifically addressing the question of whether any portion of current biochemistry could have possibly emerged in the absence of phosphate (and thus prior to transcription/translation) [61]. Models of prebiotic networks were constructed starting from minimal sets of compounds thought to have been readily available on early Earth. Notably, even if these initial compounds did not include any phosphate-containing molecule, a surprisingly large expanded network could ensue, covering several pathways that are part of central metabolism today, and of previously proposed models of biogenesis [133]. This finding is consistent with the possibility that thioesters, sulfur-based energy rich chemical moieties, could have predated phosphates as energy carriers in the cell, providing the required thermodynamic driving force. Interestingly, recent work has experimentally demonstrated the possibility that a thioester-based chemistry could fuel autocatalytic networks [173]. Future approaches could extend the use of network expansion models by incorporating additional constraints on metabolic network growth, such as the removal of likely toxic intermediates. Although further experimental and theoretical work is required to fully address the scope, implications and fundamental limitation of an early phosphate-free biosphere, the use of the network expansion algorithm to explore plausible routes of abiogenesis represents an interesting research direction.

Although the majority of chemical reactions important in early living systems may still be encoded in modern day living systems, there is also a possibility that key reactions and compounds initially critical for living systems were lost throughout the course of evolution. Even more broadly, it is plausible that much bigger space of chemically possible reactions could have given rise to an organized metabolism [126, 174]. As shown in recent elegant experimental work, molecules important for life as we know it may in principle be producible through reactions and pathways that are

not part of current biochemistry [150, 184]. On the theoretical side, recent advances in computational chemistry [4] have enabled the construction of chemical network models beyond the scope of modern living systems, paving the way for future broader analyses of possible transient chemistries along the history of life, and of putative alternative outcomes that may have materialized but didn't [126].

Beyond these realistic chemical spaces, biochemical organization has been studied extensively using simplified toy models based on artificial chemical rules, such as the so called string chemistries [8]. Similar approaches were the foundation of some of the early work on collectively autocatalytic networks [91, 6]. More recently, a very simple string chemistry, simulated and analyzed using systems biology approaches (including FBA [144]) yielded a family of optimally efficient pathways, some of which resemble functionally and topologically the rTCA cycle network [163]. An artificial dchemistry which incorporated catalytic polymers with a toy folding process was recently shown to be helpful towards explaining the emergence of polymer-based structures within a compositional inheritance world [69]. In addition to serving as a basis to explore possible scenarios for the emergence of metabolism, abstract chemistry models can be very helpful in the exploration of statistical physics-based models of non-equilibrium chemical systems [84].

### **Overcoming energy barriers, then and now**

Whether realistic or abstract, ancient or modern, any metabolism can operate only if kept far from thermodynamic equilibrium by an external free energy source. Thus, to achieve a working theory for the origin of metabolism, one should identify not only sources of materials, but also sources of free energy consistent with geochemical data. Effectively, even if early life may have extensively used abiotic organic material heterotrophically, this question largely hinges on our understanding of what free energy source could have fueled the production of electron donors capable of reducing abundant gases like CO<sub>2</sub> and N<sub>2</sub> into the reduced forms readily used by biological systems. Two potential sources include chemical energy from hydrothermal vents, and photochemical energy from solar (especially UV) radiation [41]. The former scenario is consistent with recent phyloge-

nomic studies [196], where chemical energy in the form of molecular hydrogen is used to fix carbon dioxide using a variant of the Wood-Ljungdahl pathway in LUCA. However, it is unclear whether this scenario would be compatible with thioesters as a key component for free energy transduction, given that these molecules have been recently shown to be highly unstable in simulated hydrothermal systems [28]. Interestingly, UV light can support the synthesis of organic molecules [76, 11] and iron-sulfur clusters [18] as well as drive the reductive steps in the rTCA cycle [199]. Future work exploring the potential roles of various energy sources to fuel non-enzymatic prebiotic networks will be important in determining plausible models for ancient metabolism.

Even if a source of free energy is available, a major open question in the evolution of bioenergetics is the rise of coupling between driving forces and driven reactions. Through this coupling, currently enabled by large proteins, reactions that dissipate free energy (e.g. thioester or phosphodiester bond breaking) drive reactions that require a free energy input. Such couplings have recently been proposed to universally operate as a Brownian ratchet, in which enzyme complexes rely on the step-wise, gated mechanism of highly coordinated multi-domain enzymes [23]. Martin and colleagues proposed that electron bifurcation, the most recently discovered energy conserving process [179, 123], may have been the first mechanism through which ancient metabolic networks coupled free energy sources to drive endergonic reactions. Electron bifurcation is a mechanism that enables coupling between available, mid-potential electron donors (e.g.  $H_2$ ) and acceptors (e.g.  $CO_2$ ) to generate low-potential electron donors. This mechanism is for example capable of producing reduced ferredoxin, an energy source common in diverse biochemical pathways like photosynthesis and methanogenesis [122]. In general, identifying the scope of non-enzymatic analogues for such free energy coupling processes remains an open challenge, and efforts to this end will undoubtedly shed light on the earliest phases of bioenergetic evolution.

### **Towards data-driven origin of life research**

As the above examples clearly illustrate, origin of life research is a multidisciplinary endeavor, requiring consideration of multiple, increasingly large datasets (chemical, geological, biological,

physical) for both experimental and computational analyses [21]. Currently available databases that may be useful for the study of ancient life range from collections of genetic and phenotypic diversity of microbial species and communities [119], to knowledge-base resources available for exploring metabolites, reactions and biochemical pathways [89]. As origin of life research may require data from broader categories of molecules and reactions beyond present-day biochemistry [150], databases of known organic and inorganic chemicals [100] and reactions [174], will constitute important components of future attempts to reconstruct the first biochemical processes. Other categories of data relevant to the ancient history of metabolism are available on more specialized databases [65, 52]. Future efforts could assemble other databases useful for the computational analysis of prebiotic chemistry, including a database of documented prebiotic chemistry experiments. Furthermore, and most importantly, a standardization of experimental and computational results would enable comparisons across different efforts, allowing researchers to build more systematically on previous work. Integrating data from various sources, ranging from prebiotic chemistry experiments to inferred early Earth geochemical data, could allow for the construction of large-scale models of ancient metabolic states at unprecedented levels of resolution.

Future work aimed at understanding early life will increasingly benefit from ongoing synthetic biology efforts towards the implementation of minimal living systems, and from quantitative approaches developed for systems biology of metabolism [156]. It would be highly beneficial for origin of life research to embrace theory and modeling as essential tools for transforming data and hypotheses into testable, nontrivial predictions, i.e. predictions whose outcome may not be known a priori, and whose validation or falsification may be clearly achievable, even if technologies may be years away from feasibility. Conversely, the study of early metabolism has a chance to provide new tools and ideas for how to move systems biology approaches beyond the current paradigms. For example, the exploration of putative early metabolic pathways not known in present-day organisms bears some similarities with the huge and challenging efforts of annotating metabolic enzyme functions in newly sequenced genomes and metagenomes [29]. Furthermore, biosphere-level analyses of ancient metabolism [61, 160] could inspire new approaches for studying the collective biochemistry of microbial ecosystems.

## Chapter 2

# Architecture of ancient metabolic networks without phosphate

### Summary

This thesis chapter was published as the following Research Article:

**Goldford, J. E.**, Hartman, H., Smith, T.F., & Segrè, D. *Remnants of an ancient metabolism without phosphate*. *Cell*. 2017 Mar 9; 168(6): 1126-1134 [61]

### Abstract

Phosphate is essential for all living systems, serving as a building block of genetic and metabolic machinery. However, it is unclear how phosphate could have assumed these central roles on primordial Earth, given its poor geochemical accessibility. We use systems biology approaches to explore the alternative hypothesis that a protometabolism could have emerged prior to the incorporation of phosphate. Surprisingly, we identified a cryptic phosphate-independent core metabolism producible from simple prebiotic compounds. This network can support the biosynthesis of a broad category of key biomolecules. The enrichment of this network for enzymes utilizing iron-sulfur clusters, and the fact that thermodynamic bottlenecks are more readily overcome by thioester rather than phosphate couplings, suggest that this network may constitute a metabolic fossil of an early phosphate-free nonenzymatic biochemistry. Our results corroborate and expand previous proposals that a putative thioester-based metabolism could have predated the incorporation of phosphate and an RNA-based genetic system.

## Introduction

While most research on the evolution of living systems has been focused on sequences and genomes, some answers to fundamental questions about the emergence of life may be hidden in the architecture of the complex biochemical reaction networks that sustain the cell [175]. The field of metabolic network modeling and analysis is expanding as a major research area of relevance to multiple practical applications [141, 152]. However, the use of such techniques to address fundamental questions on the emergence of living systems is still highly unexplored.

Among the many unanswered questions on life's origin, the enigma of how phosphate ended up playing a prominent role in cellular biochemistry has been puzzling scientists for decades [169], resurfacing in recent years in light of novel discoveries [1, 149]. Phosphate is present in a large proportion of known biomolecules. It is an essential component of biochemical energy transduction (most notably through ATP), cofactors such as NADH, and information storage (in DNA and RNA polymers). However, phosphate is geochemically scarce and difficult to access, often serving as the limiting nutrient in a variety of modern ecosystems [72]. Phosphate is found in terrestrial and marine ecosystems, tightly complexed with rocks and minerals, requiring mechanisms for environmental extraction and transport [148].

The ensuing dilemma of phosphate's high importance in spite of its poor bioavailability is particularly challenging for early life, as primordial protocells would have needed both a readily available phosphate source and a simple mechanism for early phosphate acquisition. Currently, there is no consensus for a phosphate source in early life, with theories ranging from acid-mediated ion solubilization, high concentrations of reduced phosphorus species in early oceans, or accumulation during late heavy bombardment [169, 149]. Even provided a phosphate source, the mechanisms of phosphate utilization and polymerization in early life remain debated [92].

The alternative solution to this dilemma is that primitive forms of life could have initially emerged and endured without major dependence on phosphate. Multiple scenarios for early metabolic pathways that do not rely on phosphate have been proposed [40, 78, 41, 193]. In many of these scenarios, sulfur and iron are conjectured to have fulfilled major catalytic and energetic functions prior

to the appearance of phosphate. Most notably, in the thioester world scenario [40], thioesters are hypothesized to have played a role similar to the one played today by ATP. Thioesters are widespread in modern metabolism, primarily as Coenzyme A (CoA) derivatives (e.g. Acetyl-CoA), and are used as condensing agents, enabling the synthesis of heterogeneous biopolymers.

The thioester world hypothesis, and other phosphate-independent proto-metabolism models, are typically invoked to explain the prebiotic plausibility of general biochemical mechanisms, and are illustrated through specific reactions or pathways. Could systems biology approaches help achieve a more systematic and quantitative understanding of the biosynthetic potential of a putative pre-phosphate metabolic networks? Is it at all possible for a phosphate-independent geochemical setting to support the emergence of a rich and complex organized biochemistry?

Here we address these questions using computational systems biology approaches originally developed for performing large-scale analyses of complex metabolic networks [47, 73]. Similar approaches have been previously used to describe the biosphere-level metabolic changes that accompanied the transition to an oxic atmosphere, about 2.2 billion years ago [160]. Specifically, we use these and other computational methods to study systematically the size, architecture and physico-chemical properties of phosphate-independent biochemical networks. Given that our goal is to shed light on processes that predate the estimated last universal common ancestor (LUCA) [181, 196], and given the long-term reshuffling of genes among organisms through horizontal-gene transfer, we focused our analysis on a global, biosphere-level biochemical network, which encompasses all known metabolic reactions across all organisms. In exploring the prebiotic relevance of metabolic reactions that in extant life are catalyzed by highly evolved, efficient and specific protein-based enzymes, we implicitly formulate the hypothesis that many of such reactions could have been initially catalyzed to a much weaker and less specific extent by a number of small molecules. Such a hypothesis in itself is not new to origin of life research [121], and is supported by a large body of literature, both pertaining to individual small-molecule catalysts and reactions [151, 128, 66, 134, 32, 33], as well as to whole networks [139, 96, 173].

The major finding we report below is the discovery of a phosphate-independent core metabolism hidden within this biosphere-level network. This core protometabolism is capable of support-

ing the synthesis of a broad set of biomolecules, including several amino acids and carboxylic acids. Statistical analysis of the physiochemical properties of enzymes within this network show an enrichment for iron-sulfur and transition metal coenzymes. By broadening our analysis of prometabolism with the inclusion of different types of coenzyme precursor couplings, we further show that thioesters, rather than phosphate, could have enabled this core metabolism to overcome energetic bottlenecks, supporting the feasibility of a metabolically rich thioester-based world.

## Results

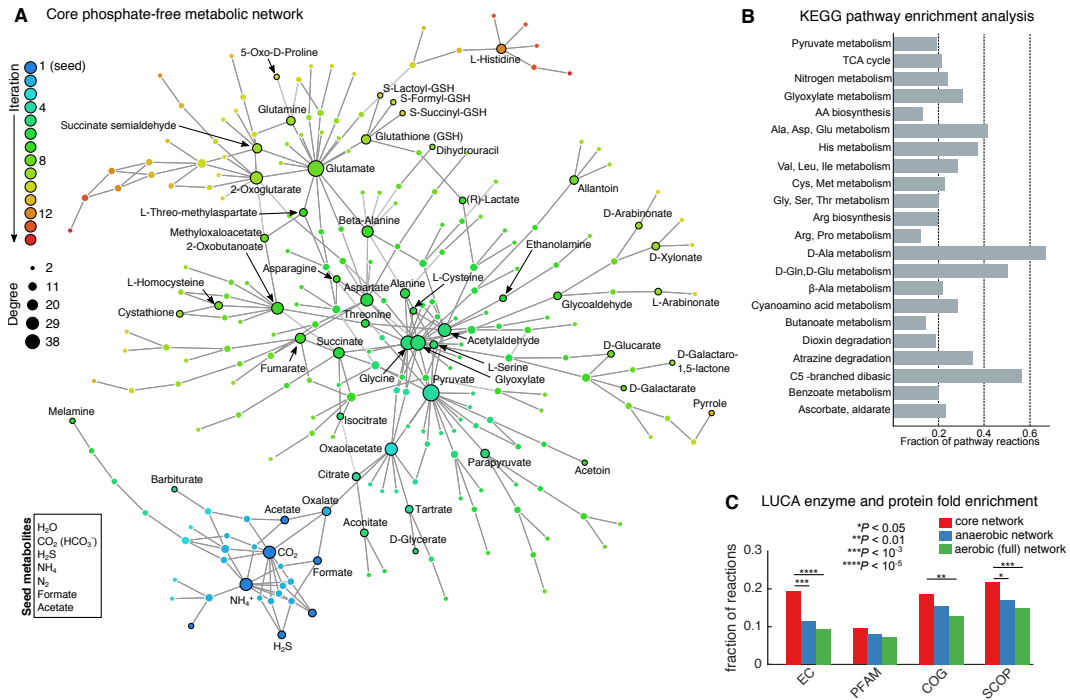
### **Removal of phosphate from biosphere-level metabolism leaves intact a core connected network**

The first goal of our analysis was to evaluate the impact of removing all reactions and metabolites involving phosphates (or, more broadly, phosphorus) from metabolism. Rather than analyzing the metabolic networks of individual organisms, we aimed at uncovering effects at the level of the complete collection of all known biochemical reactions (see Methods). This "biosphere-level" metabolism (which we inferred from the KEGG database [89]) allowed us to explore the properties of putative early biochemical networks, beyond the organismal boundaries [191].

We started by searching for regions of global metabolism that could be accessible starting from simple molecules likely to have been geochemically abundant on early Earth (Fig. 2.1A). To this end we adopted the network expansion algorithm, which simulates the emergence of metabolic networks from a predefined set of compounds [47, 73, 160]. The algorithm adds metabolites and reactions to an initial seed set, iteratively asking whether any new reaction could take place given the available substrates, until convergence to a final set of reactions and metabolites (or "scope") (see Methods). This algorithm is seed-set dependent, typically resulting in the recovery of a subset of reactions/metabolites within a defined metabolic network (Fig. 2.2). Network expansion was performed with a seed set of eight compounds thought to have been available in prebiotic environments, notably lacking phosphate (Fig. 2.1A, see Methods) [32, 121, 167, 103]. Importantly, the set of seed molecules we define contains simple carboxylic acids in the form of acetate and for-

mate, which could be provided by either an abiotic mechanism or a primitive pathway for carbon fixation (e.g. a primitive variant of the Wood-Ljungdahl pathway [179, 177, 196] or the reductive TCA cycle [192, 133, 174], see also Discussion). The resulting scope of this seed set consisted of a fully-connected network of 315 reactions and 260 metabolites (Fig. 2.1A), the composition of which was robust to variations of the seed set compounds (Fig. 2.2-2.3). Although this network requires the addition of catalytically accessible carbon, nitrogen and sulfur sources (Fig. 2.2), acetate and formate were substitutable by several alternative carboxylic acids like pyruvate (Fig. 2.3).

This core, phosphate-independent network is significantly enriched with reactions within primary metabolic pathways such as amino acid biosynthesis, pyruvate metabolism, glyoxylate/dicarboxylate and the TCA cycle, as well as intermediary metabolic pathways such as C5-branched dibasic metabolism (Fig. 2.1B, Fishers exact test, Bejamini-Hochberg procedure, FDR < 0.05). Further analysis showed significant enrichment for metabolites/reactions involved in various carbon fixation pathways, including the dicarboxylate-hydroxybutyrate cycle, the hydroxypropionate bi-cycle, and the reductive TCA cycle, which has been previously proposed as a primitive carbon fixation pathway in ancient autotrophs [133]. Several reactions involved in heterotrophic carbon utilization were also observed within pathways for one-carbon (serine pathway) and two-carbon assimilation (Krebs cycle, methylaspartate, and glyoxylate cycle). In addition to a diverse central carbon metabolism, half of the proteinogenic amino acids (G, A, D, N, E, Q, S, T, C, and H) were producible, representing six of the ten amino acids observed in the Miller-Urey experiment [147]. In this network, building upon a core carbon, energy and nitrogen metabolism, hydrogen sulfide enables the production of sulfur-containing heterogeneous peptides like glutathione, as well as thioester derivatives like *S*-formyl and *S*-succinyl glutathione. Intermediates in the degradation and biosynthesis of more complex biomolecules are also observed; 5,6-dihydrouracil is an oxidized catabolic product of uracil and pyrrole is the basic building block for complex heterocyclic aromatic rings like heme (Fig. 2.1A). Thus, we report the existence of a phosphate-independent core metabolic network reachable from simple putative prebiotic compounds .



**Figure 2.1: Network expansion yields a core phosphate-independent network (A)** A network expansion algorithm was implemented using a simple set of seed compounds (bottom left box) and all balanced reactions in the KEGG database. The figure displays a simplified view of the resulting network, in which reactions are not explicitly shown, and metabolites are linked if they are inter-converted through reactions that are responsible for the expansion. Node color indicates the time (iteration) at which the metabolite appears during the network expansion algorithm, while node size indicates the degree of that node, i.e. the number of reactions added in the subsequent iteration. Note that major hub metabolites (including pyruvate, glutamate and glycine - center of the network) are reachable after a few iterations from the seed (blue nodes). Catalytically important amino acids (e.g. His, Ser [66]) are producible in this network as well. **(B)** Pathway enrichment analysis of KEGG pathways within the core network. The fractional abundance of pathway reactions within the core network are plotted for pathways with an FDR < 0.05. **(C)** The core network reactions are enriched with enzyme functions (E.C.), protein folds (SCOP) and orthologous genes (COGs) proposed to be present in LUCA, relative to all known metabolic reactions (aerobic network) or to the oxygen-independent (anaerobic) portion of the complete network. [42, 65, 131, 181, 194] (Fishers exact test).

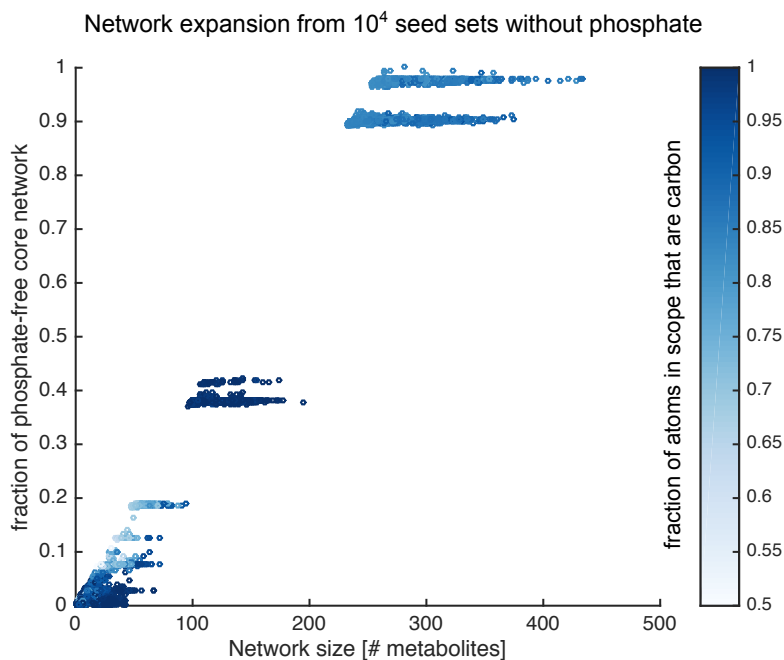


Figure 2.2: **Monte Carlo sampling of seed sets recovers substantial fractions of the non-phosphate core network**  $10^4$  random samples of size  $k = 8$  metabolites were chosen as seeds for network expansion. Each sample was required to contain at least one of the following elements: C, H, O, N and S. Network expansion was first performed using the randomly assembled seed set. For each simulation, the final number of reactions was recorded ( $x$ -axis). Next, the fraction of the the core network recovered after network expansion was computed for each seed set ( $y$ -axis). The color of each point represents the fractional abundance of carbon atoms in the scope of the simulation, highlighting the molecular heterogeneity between simulations. The positive correlation between the network size and the fraction of the core phosphate-free suggests that large ( $> 250$  reactions) networks without phosphate contain a substantial fraction of core network reactions. Note that networks between 100 and 200 metabolites were typically composed of only CHO molecules, networks  $> 250$  metabolites contained a substantial number of molecules with nitrogen and sulfur.

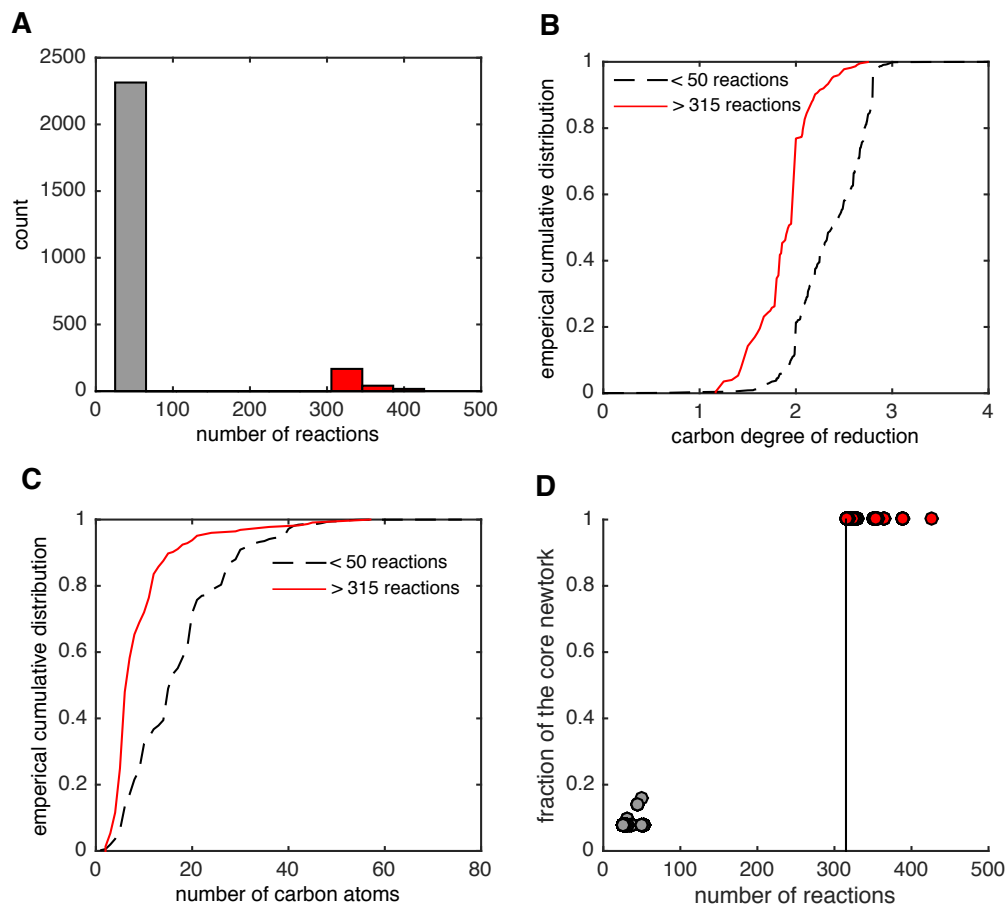


Figure 2.3: **Network expansion with various carbon sources.** Acetate and formate (see Fig 2.1) were replaced as seed compounds with a single organic compound, and network expansion was performed. This was done for each molecule in KEGG exclusively composed of C, H and O. (A) A histogram of network sizes (reaction count) after network expansion. The majority of carbon sources resulted in small networks ( $< 50$  reactions, gray), while 225 carbon sources resulted in networks  $> 315$  reactions (red). (B) Empirical CDFs for the average degree of reduction per carbon atom ( $y/x$  for substrate C, where  $x\text{CO}_2 + y\text{H}_2 \rightarrow \text{C} + z\text{H}_2\text{O}$ , see [174]) for small (black dashed line) and large (red continuous line) networks. Large networks were generated more frequently from more oxidized carbon substrates (two-tailed Kolmogorov-Smirnov test,  $P < 10^{-55}$ ). (C) Empirical CDFs for number of carbons in seed set for small (black dashed line) and large (red continuous line) networks. Large networks were generated more frequently from smaller carbon substrates (two-tailed Kolmogorov-Smirnov test,  $P < 10^{-41}$ ). (D) Scatter plot of the number of reactions (x-axis) of each expansion is plotted vs. the fraction of the core network embedded in the final network (y-axis). All large networks are greater than the 315 reactions obtained using acetate (black line), indicating expansion from acetate represents a suitable lower bound for a phosphate-independent core metabolism. It should be noted that larger network ( $> 350$  reactions) were generated from carbohydrate sources (glucose), while slightly smaller networks were generated from carboxylic acids (acetate, oxaloacetate).

### **Core network enzymes are enriched with features associated with protometabolism**

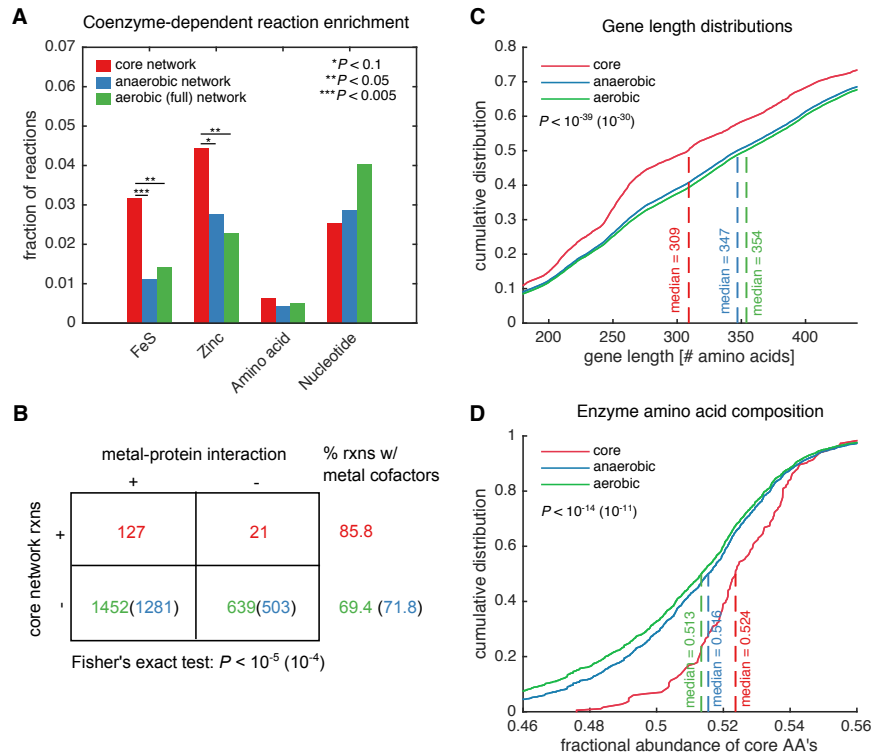
Is there independent evidence that this core phosphate-independent network may indeed resemble the very early stages of biochemical processes? The plausibility of this early metabolism relies on the possibility that catalysts for these reactions would initially have been much different than they are today, composed of short prebiotically-formed peptides [66, 129], metal-ion cofactors [32], mineral catalysts [79], or iron-rich clays [77, 105]. Such initial catalysts would have been gradually replaced by longer and more complex genome-encoded protein-enzymes, potentially still retaining properties or components of the early catalysts [129, 79, 179]. Thus, we performed multiple analyses to test whether current enzymes within this network contain taxonomic, sequence and biochemical signals pointing to potential associations with early modes of catalysis.

Taking a taxonomic approach, we found that enzymes in the core network are overrepresented within genomes (Monte Carlo permutation test,  $P = 10^{-4}$ ). The core network is also enriched with enzymes (E.C. numbers) and protein folds (SCOP) previously identified as likely components of the last universal common ancestors (LUCA) proteome [181, 194, 65] (Fig. 2.1C, Fisher's exact test:  $P < 10^{-5}$  and  $P < 10^{-3}$ , respectively), suggesting that a significant fraction of the reactions in this core network appeared in the earliest organisms. One limitation of using comparative phylogenetic analysis is that it only provides information as far back as LUCA. Furthermore, evolutionary processes like horizontal gene transfer [142] and cataclysmic extinction events hamper the elucidation of LUCAs metabolism with certainty. In order to investigate the pre-LUCA features of the phosphate-free core network, we examined the corresponding enzymes in terms of their basic physiochemical properties, with special attention given to properties proposed to be associated with ancient metabolism.

One fundamental property we focused on is the reliance of these enzymes on iron-sulfur or metal coenzymes, reflecting the notion that modern biochemistry emerged from mineral geochemistry [121, 193, 79] and that metal-based cofactors in modern day enzymes represent a living relic of this contingency [48, 71, 137]. Using a manually curated list of known protein-coenzyme pairs [65], we found that enzymes within the core network were enriched for both zinc and iron-sulfur-

dependent coenzymes relative to the full network (Fig. 2.4A  $P < 0.05$ ). For comparison, amino acid derived-coenzymes were observed with comparable frequencies in the core and full KEGG networks, while nucleotide-derived coenzymes (e.g. enzyme-bound FAD, TPP, molybdopterin) were slightly depleted amongst reactions in the core network, highlighting the coordination between nucleotide and phosphate biochemistry. The occurrence of metal-associated enzymes within the core network was independently corroborated by identifying protein structures with verifiable metal ligands in a separate database [82], allowing for the identification of KEGG reactions that rely on enzymes bound to metal ions. Out of the 47% (148/315) of the core network reactions with crystal structures available, 86% (127/148) relied on enzymes with a metal ligand, which constituted a significant enrichment relative to the full KEGG network (Fig. 2.4b, Fisher's exact test:  $P < 10^{-5}$ ).

In addition to a biased coenzyme usage, we investigated other features that could be associated with an ancient proto-metabolic network. First, motivated by the notion that early catalysts may have been composed of smaller polypeptides relative to present day enzymes [129], we tested if the enzymes in the core network are on average smaller relative to all genome encoded enzymes. We found that sequences are considerably shorter for catalysts in the core network (median = 309) compared to all known metabolic enzymes (median = 354) (Fig. 2.4C; one-tailed Kolmogorov-Smirnov:  $P < 10^{-39}$ ). Second, we thought of checking whether enzymes in the core network are enriched, in their composition, for amino acids producible by the core network itself. Such an enrichment would be consistent with the expectation of self-sustainability and homeostasis in a proto-metabolic network, whereby the network would be capable of producing the building blocks necessary for replenishment and accumulation of its catalysts. We found indeed that core network enzymes are more highly composed of the 10 amino acids found within the core network relative to all known metabolic enzymes (Fig. 2.4D; one-tailed Kolmogorov-Smirnov:  $P < 10^{-14}$ ). One potential simple reason for this enrichment could be attributed to the known sequence bias in FeS-proteins for cysteine, both of which are present in the core phosphate-free network. However, we found no detectable enrichment for cysteine in our core network enzymes (one-tailed Kolmogorov-Smirnov test,  $P = 0.948$ ).



**Figure 2.4: Reactions in the core network are enriched for iron-sulfur and transition metal coenzymes** (A) The fraction of coenzyme-coupled KEGG reactions in the core network (red bars), the anaerobic KEGG network (blue bars) and the aerobic KEGG network (green bars) are compared. Each set of reactions is composed of a manually curated list of coenzyme-coupled reactions in KEGG [65]. We found that a significant number of reactions require iron-sulfur coenzymes (Fishers exact test,  $P < 0.05$ ) and zinc (Fishers exact test,  $P < 0.05$ ) within the core network relative to the aerobic KEGG network. (B) Structural data support metal-protein enrichment in the core network. The number of reactions catalyzed by enzymes with structural data were determined for all KEGG reactions using the MIPS database [82]. For all reactions with crystal structures available, we classified each reaction as either without (-) a metal cofactor, or with (+) a metal cofactor. We performed enrichment tests for metal cofactors within the core network relative to both the aerobic KEGG network (green text), or the anaerobic network (parenthesis, blue text). Core network reactions relied more heavily on enzymes with metal cofactors relative to both the aerobic and anaerobic KEGG reactions. The enzymes in the core network are shorter (C) and bi-ased in their amino acid composition (D) relative to either the aerobic or anaerobic KEGG network. For B-D, we tested for enrichment within the phosphate-free core network enzymes compared to both the aerobic and anaerobic networks. Significance values are reported for the aerobic network, followed by the anaerobic network in parenthesis.

### **Thioesters alleviate thermodynamic bottlenecks**

So far, our analysis was focused on the core network structure, ignoring possible energetic constraints. In extant metabolism, phosphate-mediated group transfer plays a key role by driving unfavorable, or energetically up-hill reactions [41]. To investigate the energetic consequences of phosphate unavailability, we implemented a thermodynamically constrained network expansion algorithm, which blocks endergonic reactions with standard molar free energies above a cutoff value  $\tau$  (Fig. 2.5B, black line). The network becomes dramatically limited to  $< 12\%$  of the core network as  $\tau$  remains below 55 kJ/mol, preventing the condensation of oxalate and acetate to yield oxaloacetate. Energetic constraints of this magnitude would have prohibited the expansion of an early metabolism, given plausible ranges of intracellular metabolites concentrations [17] (see Methods). Consequently, a mechanism to overcome these thermodynamic bottlenecks would be essential for a phosphate-independent metabolism.

Could thioester chemistry [40] serve as a solution to this energetic conundrum? Thioesters, proposed to have served as ancient condensing agents [40, 179], are widespread throughout central metabolic processes (e.g. Coenzyme A (CoA) derivatives in TCA cycle and lipid biosynthesis) and can facilitate energy-rich group transfer. While CoA contains phosphate, this serves mainly as a structural component with no catalytic role, motivating the hypothesis that ancient reactions may have relied on pantetheine, the simpler, phosphate-free variant of CoA [46, 101] thought to be available in prebiotic environments [93]. We explored the energetic consequences of a primitive thioester-based reaction coupling scheme by substituting pantetheine for CoA in modern CoA-coupled reactions, followed by adding pantetheine into the seed set (Fig. 2.5A, see Methods). These changes caused a 33 kJ/mol reduction in the bottlenecks that limited network expansion, enabling the viability of alternative metabolic pathways under physiologically realistic conditions [10] (Fig. 2.5B, red line). Interestingly, these bottlenecks could not be easily overcome through an alternative phosphate-based coupling scheme, in which NTP-coupled reactions are substituted with either pyrophosphate or acetyl-phosphate (Fig. 2.5B, blue line, Fig. 2.6). The uniqueness of this behavior is also emphasized by the fact that removal of elements other than phosphate (e.g.

sulfur or nitrogen) would dramatically limit the possibility of expansion (Fig. 2.6).

### **Primitive coenzymes enable widespread network expansion**

A larger metabolic network may have been reachable if phosphate-free versions of modern day coenzymes drove several primordial reactions. Like CoA, many modern day coenzymes contain nucleotide phosphate groups that are important for enzyme-binding but not directly involved in catalysis. For example, the redox coenzyme NAD contains adenine and phosphate, but facilitates electron transfer at the nicotinamide moiety (Fig. 2.8A). By substituting CoA with pantoic acid and implicitly assuming that oxidoreductase reactions could be coupled to primitive electron donors/acceptors instead of NAD(P)/FAD (Fig. 2.8B), we found that the core network expands to nearly 3 times more metabolites (814), incorporating 5 more amino acids (K, R, L, V and P), uracil and ribose (Fig. 2.7). Addition of these 5 amino acids to the repertoire of amino acids would have enabled broader catalytic capabilities, and paved the way for increased richness of peptides, once suitable and energetically supportable mechanisms for peptide synthesis became available (perhaps, initially driven by thioesters themselves [40]). Further, the formation of pyrimidines, pentoses and vitamins could have set the foundation for the assembly of nucleotides triphosphate and modern coenzymes upon the addition of phosphate.

### **Discussion**

To obtain insight into the early stages of the evolution of metabolism, prior to LUCA, we analyzed the biosphere-level collection of all known metabolic reactions, which throughout the history of life may have greatly shifted their assortment into organisms [142, 191]. By integrating network algorithms and biochemical database analyses at the biosphere-level, we have uncovered a phosphate-independent metabolism that prompts us to revisit models of early biochemistry. This network is enriched with enzyme's requiring inorganic and iron-sulfur cofactors, consistent with the hypothesis that iron-sulfur proteins are among the most ancient in biological systems [48, 71, 193, 79, 137, 179]. This network incorporates several components of the reduc-

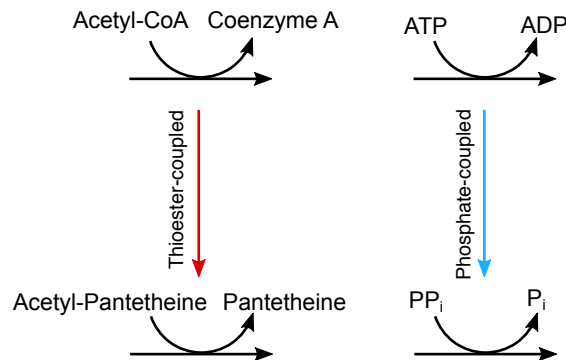
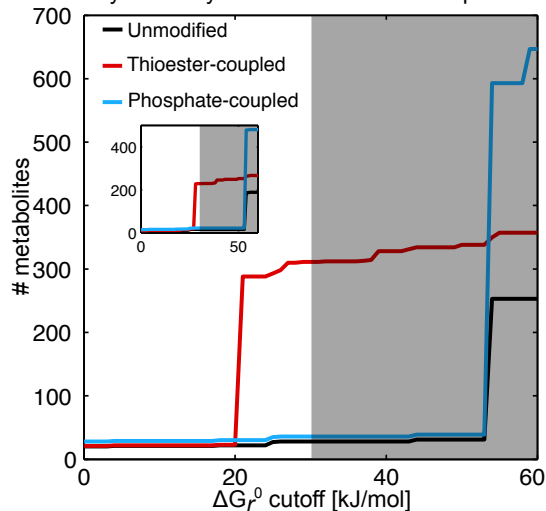
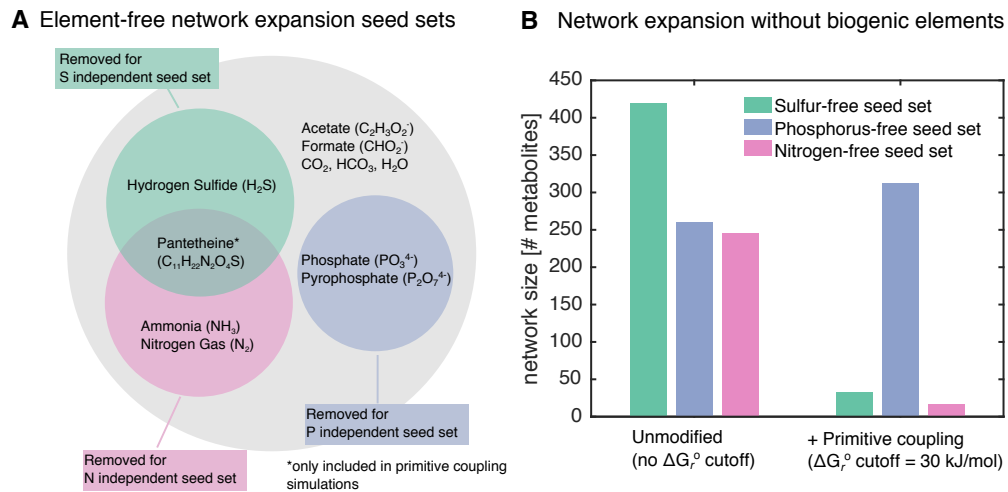
**A** Models of prebiotic metabolic coupling mechanisms**B** Thermodynamically-constrained network expansion

Figure 2.5: **Thioesters alleviate thermodynamic bottlenecks** (A) Models of ancient coenzymes were constructed to simulate the roles of thioesters and phosphates in models of ancient biochemistry (see Methods). (B) Network expansion from the core seed set was performed after removing reactions exceeding a thermodynamic threshold. For each value of this threshold (x-axis) we plot (black line) the size of final network (in terms of the number of metabolites, y-axis). The effect of thioester coupling was simulated by adding a Coenzyme A substitute (pantetheine) to the seed set (red line). For comparison, a phosphate-coupled network was simulated by substituting nucleotide triphosphate-coupled phosphoryl-transfer reactions with pyrophosphate (or acetyl-phosphate) (blue line), followed by adding pyrophosphate (or acetyl-phosphate) to the seed set. Although significantly more metabolites are observed in the phosphate-coupled network with no thermodynamic barrier (due to the addition of sugars and phosphorylated intermediates), network expansion would not be feasible under physiologically realistic conditions (unshaded region) [10]. More than one third of reactions in KEGG lack a free energy estimate. In the main plot, all these reactions with unknown free energies are assumed to be available (equivalent to assuming that they have a free energy barrier lower than then predefined threshold). Results are qualitatively very similar if all such reactions lacking free energy estimates are removed from the network (Top left inset).



**Figure 2.6: Sulfur and nitrogen are required for thermodynamically feasible network expansion.** This analysis aims at testing the uniqueness of the feasibility of a phosphorus-free network, in comparison to other hypothetical scenarios in which other atoms are missing from the initial seed set. Specifically, we compare the size of the expanded network under elimination of phosphorus, sulfur or nitrogen, with and without the thermodynamic feasibility constraints. Network expansion was performed for all KEGG reactions using a seed set without sulfur (no H<sub>2</sub>S and pantetheine, green bars), phosphate (no pyrophosphate, blue bars), or nitrogen (no ammonia, nitrogen gas, and pantetheine, pink bars) (see also Venn diagram for specific seeds and atomic compositions). The left set of bars represent the size of expanded networks without imposition of thermodynamic constraints, while the right plot shows the network sizes when thermodynamic feasibility is imposed. It can be seen that removal of sulfur, without thermodynamic constraints, gives rise to a larger network relative to the P-free core network, due to the appearance of sugars and phosphosugars. However, when taking into account thermodynamic feasibility, the P-independent network is the only one that can reach a large size. Thus, a thermodynamically feasible network expansion is conditional on the presence of sulfur and nitrogen, but not phosphate. One should also stress that while there is debate about the prebiotic availability of phosphorus [40, 169], consensus is much higher among researchers about the availability of sulfur [33, 41, 79, 192]. Thus the significance of our analysis is particularly clear in light of this prior.

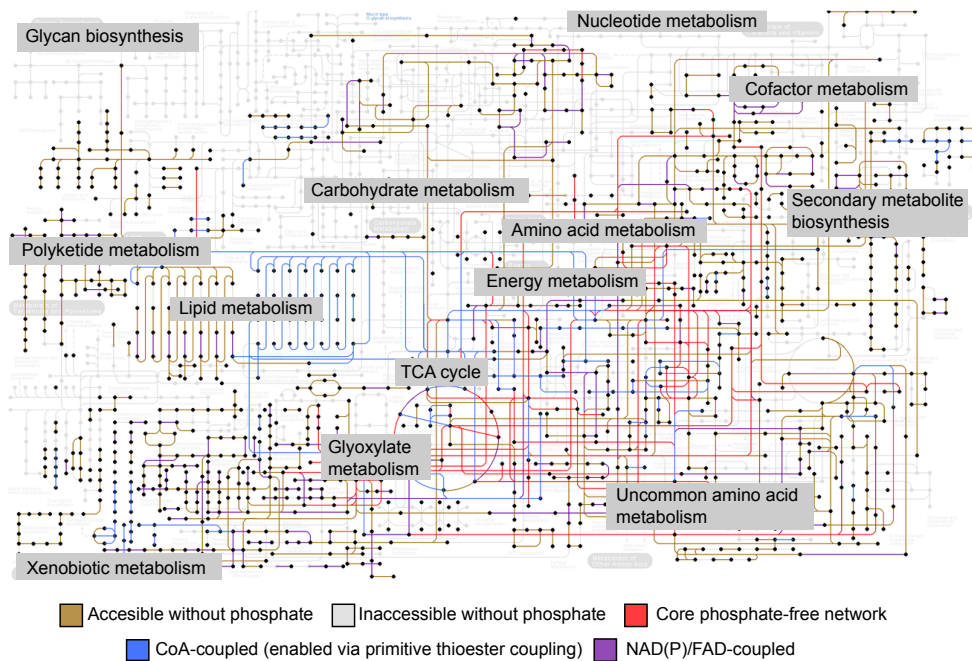
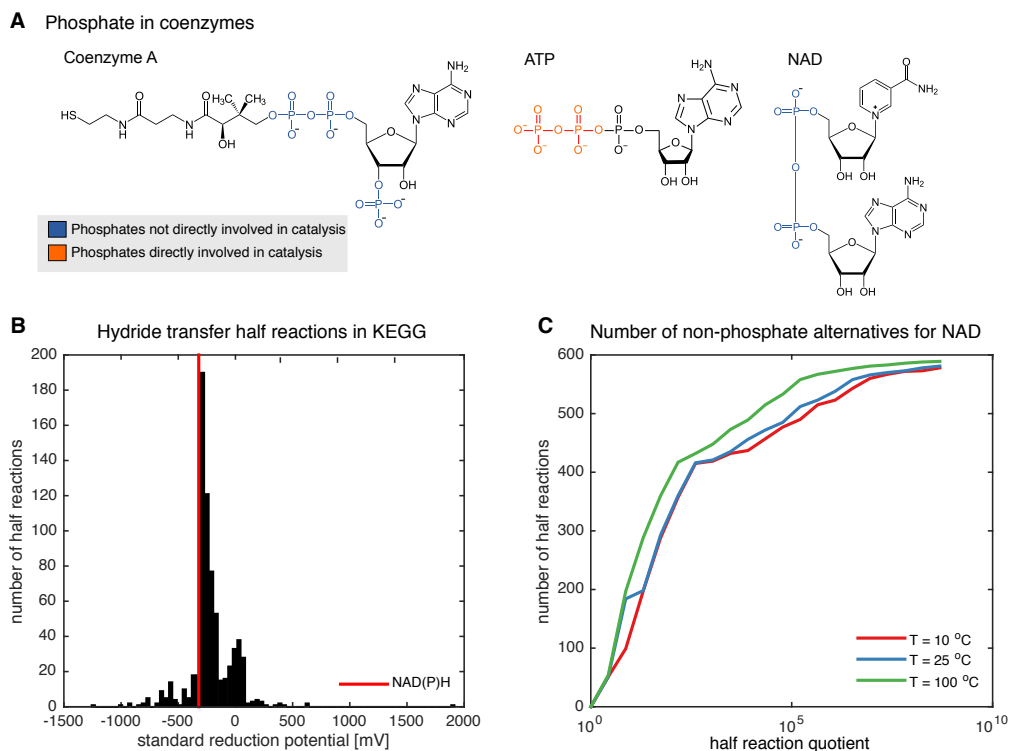


Figure 2.7: **Global non-phosphate metabolism** We removed all phosphate-dependent reactions from biosphere-level metabolism, and the largest connected subnetwork was obtained. The gray lines are unreachable reactions without phosphate, meaning there is no seed set capable of recovering this portion of the network without phosphate. The red lines are the reactions within the core network (identified in Fig. 2.1A) and the brown lines are reactions that are not included in the core, but still accessible without phosphate. The blue lines are all phosphate-free reactions coupled to Coenzyme A, while the purple lines are the phosphate-free reactions coupled to nicotinamide or flavin coenzymes.



**Figure 2.8: There Are Several Suitable Biomolecules that May Have Preceded Modern Phosphate-Dependent Coenzymes** (A) Phosphates in modern day coenzymes. The phosphates in Coenzyme A and NADH play no role in catalysis, while the phosphates in ATP are pivotal in catalysis. (B) Plausible substitutes for NADH. All half reactions representing a two electron reduction via hydrogen transfer (i.e., Oxidized + 2H<sup>+</sup> + 2e<sup>-</sup> → Reduced) were computed using KEGG reaction pairs database. For approximately 3/4 of all suitable half reactions (712/947), a standard reduction potential could be estimated using group contribution estimates of free energies of formations for KEGG metabolites [52, 138]. The red line marks the standard reduction potential of NAD(P)/NAD(P)H. (C) For different maximum allowable concentration ratio between oxidized and reduced species (x axis) we counted the number of half reactions that overlapped with the reduction potential of NADH (y axis). The colored lines represent different temperatures. For example, if a 100-fold concentration ratio was permitted, approximately 200 non-phosphate half-reactions could have sufficed as plausible substitutes for NAD(P).

tive TCA cycle, proposed to be one of the first autotrophic, autocatalytic cycles in metabolism [77, 133, 174]. Its enrichment for enzymes containing iron-sulfur clusters is strikingly consistent with the iron-sulfur world theory [193]. Our results are compatible with the possibility that the iron-sulfur dependent reactions in the TCA cycle and the methylaspartate cycle in haloarchaea [99] may represent modern variants of an iron-sulfur based intermediary metabolism. Upon including phosphate-independent precursors of high-energy and redox cofactors in our model, the resulting network became free of prohibitive thermodynamic bottlenecks, and expanded to a much larger proto-metabolism that includes several precursors for DNA/RNA and modern-day coenzymes. Our work corroborates previous work emphasizing the potential role of thioesters in protometabolic systems [40, 179].

Specific hypotheses generated by our analysis could be testable in future work. For example, it would be interesting to extend currently available evidence of non-enzymatic catalysis of metabolic reactions to a larger set of reactions and potential catalysts, with and without the specific constraint of phosphate availability. In particular, one could test the possible role of previously identified small-molecule catalysts (e.g. amino acids, short peptides and metal sulfides) in enabling reactions within the core network. While our calculations suggest that thioester chemistry had an initial thermodynamic advantage towards generating a surprisingly large and connected metabolism, this set of metabolites constitutes less than 20% of the complete phosphate-dependent set of known metabolites we know today. In future work it will be interesting to search for more evidence that the network dependent on thioesters may have been self-sustaining (i.e. capable of producing its own small-molecule catalysts), and for signatures of a putative thioester-to-phosphodiester transition

The plausibility of a rich phosphate-independent metabolism has a number of implications on important questions about the origin of life. In particular, the expansion of a phosphate-independent metabolism requires the availability of reduced-carbon precursors (i.e. the seed set) and energy (e.g. the driving force for the production of thioesters). Although these could be explainable by purely geochemical (i.e. abiotic) processes (see SI and [167, 103]), a number of scenarios involving ancient variants of modern carbon fixation pathways have been proposed as a source of reduced carbon and thioesters [57]. One such scenario is based on the reductive TCA cycle [133, 174],

which uses several reactions found in our core network, and is compatible with the iron-sulfur enrichment previously discussed. One of the challenges in this scenario is that alternative energy coupling schemes instead of ATP hydrolysis (found in Succinyl-CoA synthetase and ATP-citrate lyase) would have been required to make the process exergonic. An alternative scenario, which could simultaneously explain the availability of formate, acetate and thioesters, is the viability of a primordial Wood-Ljungdahl (WL) pathway, previously suggested to proceed exergonically under prebiotic conditions [179]. One of the appeals of this pathway is that it is the only carbon fixation pathway present within both bacteria and archaea, and that it may have been the first carbon-fixation pathway in LUCA ([196]. For this pathway to be viable in a pre-phosphate world, however, its ancient variants would have to rely on simple coenzyme precursors to pterins, which are currently not known to be synthesized biotically without GTP.

While we cannot rule out possible alternative interpretations of our findings, such as a gradually evolved reliance on metabolic routes that make minimal use of phosphate, it is interesting to ask whether our result could help bridge a fundamental gap between geochemistry and biochemistry. The properties of the core phosphate-free network suggest that a thioester-based proto-metabolism may have started from a few, simple geochemically abundant molecules, and expanded to a surprisingly rich and diverse biochemistry, potentially a network-level fossil of biosphere metabolism, even prior to the appearance of the phosphate based genetic coding system. This network could have enabled the synthesis of a diverse set of (bio)chemical compounds, providing precursors for the subsequent rise of informational nucleic acid polymers. Whether or not such a primordial system could have been endowed with features essential for cellular life as we know it, such as autocatalysis and information processing, remains an open question.

## **Methods**

### **Reconstruction of biosphere-level metabolism**

All metabolic reactions in the KEGG database [89] were downloaded in July of 2015, and an  $m$ -by- $n$  stoichiometric matrix  $S$  was constructed for  $m$  metabolites and  $n$  reactions based on chemical

reaction equations. Reactions that were unbalanced or contained metabolites in unspecified molecular formulas were removed (see SI) resulting in a global network of  $n = 6880$  reactions utilizing  $m = 5944$  metabolites. All reactions were initially assumed to be reversible, except for reactions that utilized molecular oxygen [160]. These reactions were constrained to be irreversible, such that oxygen could never be produced as a byproduct during network expansion, reflecting the expectation that early Earth was anoxic. A more detailed discussion of network construction is provided in the SI.

### Network expansion

Network expansion has been described in detail elsewhere [47, 73, 160]. Briefly, let  $\mathcal{S}$  represent the set of seed metabolites, and the scope of seed set be represented by  $\mathcal{F}(\mathcal{S})$ , which contains all reactions and metabolites reachable from the seed set  $\mathcal{S}$ . At each iteration  $k$ , the set of reactions where all substrates were present,  $\mathcal{R}_k$ , were added to the scope. Reactions were then allowed to produce metabolites,  $\mathcal{M}_k$ . The scope was updated by taking the union of  $\mathcal{F}(\mathcal{S})$ ,  $\mathcal{R}_k$ , and  $\mathcal{M}_k$ , where:

$$\mathcal{F}(\mathcal{S}) \leftarrow \text{union}(\mathcal{F}(\mathcal{S}), \mathcal{R}_k, \mathcal{M}_k).$$

The algorithm terminates when no more reactions or metabolites can be added to the scope, resulting in a final stationary network composition. Although it is possible that an expansion will span the set of all metabolites in the network, a seed set typically is capable of expanding to only a fraction of the network. For more details regarding implementation, see the supporting information.

### Seed set

Seed set compositions were chosen based on previously reported putative molecular composition on prebiotic earth (see e.g. [121]). Fig. 2.1 lists the compounds used in the seed set to generate the core network referenced throughout the paper. Volatiles and gases widely considered to be present on early Earth are dinitrogen, water, hydrogen sulfide and carbon dioxide [159]. Although it is unclear at what time biotic nitrogen fixation emerged, abiotic nitrogen reduction to ammonia has

been demonstrated at high concentrations of hydrogen sulfide [79], and is thought to have been the dominant nitrogen source in early organisms [38], motivating us to include ammonia into the seed set.

Reduced carbon is an essential component of metabolism, requiring either an autotrophic carbon fixation process or the heterotrophic carbon assimilation of abiotically reduced carbon. We tested two scenarios: (i) an autotrophic origin of metabolism from carbon dioxide and hydrogen gas and (ii) a heterotrophic origin of metabolism from formate and acetate. We did not see significant growth from scenario (i), indicating that a reduced form of carbon is required. Acetate and formate were chosen based on previous work suggesting early forms of abiotically fixed carbon may have been of the form of simple carboxylic acids which in principle could have been synthesized at hydrothermal vents from hydrogen and carbon dioxide using the processes of serpentinization [121, 167, 103], or via a primitive variant of a modern carbon fixation pathway such as the Wood-Ljungdahl pathway [57, 179, 177, 196]. We explored variations to this seed set in two ways. First, we performed a Monte Carlo permutation test on the seed set (see Fig. 2.2) and second, we varied the identity of the carbon sources (see Fig. 2.2).

### Reaction thermodynamics

To sustain net flux for a chemical reaction, the laws of thermodynamics require that the difference in free energy between products and reactants,  $\Delta_r G'$ , has to be negative [10]. For a given biochemical reaction at fixed temperature and pressure,  $\Delta_r G'$  is defined as:

$$\Delta_r G' = \Delta_r G'^{\circ} + RT \ln \prod_i a_i^{s_{ir}}$$

where  $\Delta_r G'^{\circ}$  is the free energy change of the reaction at standard molar conditions,  $R$  is the ideal gas constant,  $T$  is temperature,  $a_i$  is the activity of metabolite  $i$  and  $s_{ir}$  is the stoichiometric coefficient for metabolite  $i$  in reaction  $r$ , which is negative for reactants and positive for products. Assuming metabolite concentrations,  $c_i$ , can be substituted for activities, the disequilibrium ratio,  $\Gamma_r$  is defined as  $\Gamma_r = \prod_i c_i^{s_{ir}}$ . The necessary condition of a negative free energy of reaction can be

recast as:

$$\Gamma_r < -\frac{\Delta_r G'^{\circ}}{RT}$$

This indicates that for large  $\Delta_r G'^{\circ}$ , a small  $\Gamma_r$  is required to maintain feasibility.  $\Delta_r G'^{\circ}$  represents a “thermodynamic barrier,” which can be overcome by reducing  $\Gamma_r$  (i.e. increasing the reactants relative to the product concentration).

To identify potential thermodynamic barriers, we performed network expansion without reactions above a predefined free energy threshold,  $\tau$ . In this variant of network expansion, reaction  $r$  was removed if  $\Delta_r G'^{\circ} > \tau$ , for varying levels of  $\tau$ . We obtained estimates for  $\Delta_r G'^{\circ}$  from Equilibrator [52], which uses the component contribution method to estimate free energies of formation of metabolites based on the group decomposition of compounds [138]. We obtained estimates of  $\Delta_r G'^{\circ}$  at various pH values, ranging from pH 5 to pH 9 in increments of 0.5, while assuming a constant ionic strength of 0.1 M and temperature of 298.15 K. We performed network expansion at thresholds varying from 0 to 60 kJ/mol in 1 kJ/mol increments. Final network sizes in all scenarios were insensitive to the choice of pH.

Over one third of all KEGG reactions did not have estimates for  $\Delta_r G'^{\circ}$ , due to the large set of metabolites with no estimate for the free energy of formation. We accounted for this by either assuming (i) all reactions with unknown  $\Delta_r G'^{\circ}$  were feasible regardless of the cutoff or (ii) reactions with no estimate for  $\Delta_r G'^{\circ}$  were infeasible, and subsequently removed altogether. The qualitative results presented in Figure 3 of the main text are unaffected by the treatment of these reactions.

### **Primitive coenzyme coupling**

Coenzymes in modern day metabolism are composed of highly heterogeneous functional units, composed of distinct moieties involved in protein-binding and catalysis (for comprehensive review, see [21]). Two groups of coenzymes are readily observed that contain phosphate: *Phosphoryl-donating/accepting* coenzymes (e.g. ATP and GTP) and *phosphate-containing* coenzymes with no phosphoryl group transfer (e.g. Coenzyme A, TPP, NAD, FAD, Molybdopterin). For Phosphoryl-donating/accepting coenzymes, removing phosphate would clearly abolish the catalytic function

of the coenzymes, while in phosphate-containing coenzymes, removing all phosphate-containing moieties may not eliminate catalytic function. Phosphoryl-donating/accepting coenzymes may have been preceded by non-nucleotide metabolites with phosphodiester bonds, such as phosphoenolpyruvate, acetyl-phosphate, or pyrophosphate [40] while phosphate-containing coenzymes may have been preceded by less complex versions of these coenzymes [101].

The following subsections summarize the introduction in our network of variants of present-day reactions in which current cofactors are substituted with putative primitive alternatives. In particular, this amounts to the addition of putative prebiotic reactions utilizing primitive thioester, phosphate, and redox couplings, as described below (Fig 2.5 and 2.7, main text). Fig. 2.8 provides the structures of Coenzyme A, NAD and ATP, highlighting the role of phosphates in each biomolecule.

**Thioester coenzymes.** For the proposed thioester-coupled network, we directly modified metabolites and reactions such that CoA-mediated acyl transfer reactions were substituted with pantetheine-mediated acyl transfer reactions. This required us to first identify metabolites with CoA thioesters (i.e. Acetyl-CoA, Malonyl-CoA), then substitute the CoA moiety with pantetheine. Second, all reactions typically using these molecules were substituted with the pantetheine thioesters. We also ensured that degradation of pantetheine did not contribute to network growth by blocking degradation pathways. This was achieved by removing (R)-pantetheine amidohydrolase (KEGG ID: R02973) and N-((R)-Pantothenoil)-L-cysteine carboxy-lyase (KEGG ID: R02972), which prevented the hydrolysis of pantetheine into pantothenate and cysteamine. Network expansion was then performed with pantetheine added to the core seed set, resulting in a final network size of 365 metabolites. Simulating the thioester-coupling can also be achieved by (i) blocking degradation of CoA and (ii) adding CoA into the seed set. By removing CoA nucleotido-hydrolase (KEGG ID: R10747), and adding CoA into the core seed set, we obtained the final network observed with the pantetheine substituted network.

**Phosphate coenzymes.** For the proposed model of a primitive phosphate-coupled network, nucleotide (A, G, C, U, T, and I) phosphate-coupled and phosphotransferase reactions were substituted with pyrophosphate. Pyrophosphate has been proposed to have been used for primitive

energy coupling in early protometabolic systems before NTP [40, 121]. Using pyrophosphate coupling instead of NTP prevents network expansion from artificial catabolism of NTP precursors like ribose and nucleobases, which are not assumed *a priori* to be abundant in geochemical models of Hadean environments. In particular, monophosphate transfer reactions were replaced with the diphosphate/monophosphate coenzyme couple, while diphosphate transfers were replaced with the triphosphate/monophosphate coenzyme couple. This model of primitive phosphate-coupled reactions was seeded with orthophosphate, diphosphate and triphosphate, in addition to the “core seed set” listed in Fig. 2.1 of the main text. For monophosphate transfer reactions, we also substituted NTPs with acetyl-phosphate, and found no difference between the network sizes at different free energy threshold cutoffs (Fig 2.5).

**Redox coenzymes.** We found that a much larger network was reachable without phosphate by relaxing the condition that major redox reactions require the phosphate-containing coenzymes NAD(P) or FAD as substrates. For this analysis, in addition to adding modified thioester-coupled reactions (see **Thioester coenzymes**), major redox reactions mediated by NAD(P) and FAD were allowed to proceed using only the half reactions. Reactions utilizing the  $\text{NAD(P)}^+/\text{NAD(P)H}$  or the  $\text{FAD}/\text{FADH}_2$  redox couples were replaced with the associated redox half reactions with no cofactor pair. For example, the reaction  $\text{X} + \text{NAD(P)H} \rightarrow \text{Y} + \text{NAD(P)}^+$  was replaced by the half reaction:  $\text{X} + 2\text{e}^- + 2\text{H}^+ \rightarrow \text{Y}$ , where both  $\text{e}^-$  and  $\text{H}^+$  are in the seed set. Several alternative redox coupling schemes may have been available in proto-metabolic systems, including glutathione or primitive iron-sulfur proteins. For our analysis, we simply decomposed redox reactions into half reactions and allowed for free exchange of electrons. This alteration effectively adds low potential reduced ferredoxin as a seed molecule, potentially producible via electron bifurcation from  $\text{H}_2$  [24].

## Supplemental Methods

### Construction of a biosphere-level metabolic network

The set of all known metabolic reactions was assembled into a biosphere-level (or pangenome) metabolic model using the KEGG database. All KEGG reactions and compounds were downloaded

using the KEGG REST API. We constructed a stoichiometric matrix from the KEGG reaction database using reaction equations. Reactions were removed if they either consumed or produced compounds that (i) did not include a SMILES string or (ii) included an *n*-subunit polymer with undefined molecular formulas. Metabolites with arbitrary, “R” groups were retained as long as “R” groups were balanced in reaction equations. Reactions that were elementally imbalanced for any element except hydrogen were removed. The elementally balanced network consisted of 6880 reactions utilizing 5944 metabolites. This network is reference in the main text as the “full KEGG network” or the “aerobic network” in Figs. 2.1 and 2.4. The KEGG network used throughout the manuscript is depleted in enzymes that catalyze reactions that either chemically unbalanced (e.g. Fatty Acid Elongation reactions, Lysine biosynthesis), and depleted in enzymes with no assigned KEGG reaction like arsenate reductase (EC 1.20.4.4) and ketol-acid reductoisomerase (EC 1.1.1.382). Future versions of KEGG and other knowledgebases will contain a more accurate and comprehensive collection of biosphere-level metabolism.

The results presented in Fig. 2.1, 2.2 and 2.3 were generated using network expansion with all reactions assumed to be reversible, except for reactions utilizing molecular oxygen. Oxygen is thought to have become available in the biosphere after the appearance of the last universal common ancestor, and it was previously shown that widespread network growth is achievable once oxygen is available. To limit network expansion to anoxic metabolism, we prevented the production of molecular oxygen by blocking reactions that produced oxygen, including oxygenic photosynthesis.

For statistical analysis, we compared enrichment features of the phosphate-free core network to the full network (aerobic) and the network accessible without oxygen (anaerobic network). The anaerobic network was generated by removing subsets of reactions and metabolites from the global metabolic network reachable only through reactions that utilize molecular oxygen, resulting in a modified biosphere-level metabolic network we called the “anaerobic network”. This was performed to ensure that statistical enrichment tests were not biased by including reactions and enzymes likely added to the global metabolic network after oxygen accumulated in the atmosphere. We first removed all reactions that utilize oxygen. Second, the stoichiometric matrix was converted into a bipartite undirected graph, where nodes were either reactions or metabolites. In this

Table 2.1: LUCApedia datasets used in this study

Dataset Name	Data type	Examples	Citation
E.C.	Enzyme		Srinivasan et al., 2009 [181]
PFAM	Protein motifs		Delaye et al., 2005 [42]
COG	Genes		Mirkin et al., 2003 [131]
SCOP	Protein folds		Wang et al., 2007 [194]
Iron-sulfur	Coenzyme	4Fe-4S, 2Fe-2S	Goldman et al., 2013 [65]
Zinc	Coenzyme		Goldman et al., 2013[65]
Amino-acid	Coenzyme	Biotin, Coenzyme F430	Goldman et al., 2013[65]
Nucleotide	Coenzyme	TPP, Molybdopterin	Goldman et al., 2013[65]

bipartite graph, an edge exists between a reaction and a metabolite if that reaction either consumes or produces that metabolite. The graph was used as an input into the python package NetworkX (<https://networkx.github.io/>), and all connected components were detected. For this case, a single major connected component was observed that contained the majority of all metabolic reactions.

### Enzyme feature datasets

To determine the plausibility that our portions of metabolism are potential relics of non-enzymatic prebiotic chemistry, we obtained various datasets corresponding to taxonomic, sequence and physiochemical properties of enzymes in modern metabolism. These features of enzymes are independent of our network generation method; the network expansion algorithm simulates the emergence of metabolites via a set of allowable reactions. In our simulation, we assume that all metabolic reactions are feasible. Thus, properties of the enzymes found in simulated networks can be used to as an independent validation of prior assumptions. Below we describe the datasets obtained from previous studies.

### LUCApedia

KEGG genes associated with components in LUCA were downloaded from the LUCApedia webpage. For each dataset (see table 2.1 for list), we obtained a list of genes, and we used the KEGG REST API to map genes to reactions.

## **MIPS**

Data from the MIPS database was downloaded as an HTML page from the website, and an in-house python script was written to parse the webpage into a list of PDB IDs. PDB IDs were mapped to uniprot proteins using PDBWSW [?], followed by the conversion of uniprot to KEGG genes using the KEGG conversion tool ([http://www.genome.jp/kegg/tool/conv\\_id.html](http://www.genome.jp/kegg/tool/conv_id.html)).

## **Gene sequence composition**

Gene lengths and amino acids compositions were obtained from the KEGG database using the REST API. For each reaction in the full KEGG network, we found all orthologous groups (KO) associated with each reaction. For each KO group, we downloaded the amino acid sequence for each gene within the associated orthologous group. For gene lengths, we computed the number of characters in each sequence. For the amino acid composition, we computed the average amino acid composition across all orthologous groups for each reaction, resulting in an averaged number of each amino acid per reaction. For Fig. 2.4D, we computed the fraction of each sequence consisting of the 10 amino acids found within the core network.

## **KEGG reaction to species mapping**

The KEGG REST API was used to identify all reactions in each species ( $n = 3838$ ) in KEGG.

## **Network expansion**

This section provides a formal description of our implementation of the network expansion algorithm. We also provide the MATLAB function used to implement network expansion in the manuscript.

### 2.0.0.1 Definitions

For  $m$  metabolites and  $n$  reactions, let the binary vectors  $x \in \{0, 1\}^m$  and  $y \in \{0, 1\}^n$  represent the states of metabolites and reactions, respectively. The component  $x_i$  ( $y_j$ ) is either 0 or 1, corresponding to whether or not metabolite  $i$  (reaction  $j$ ) is absent or present, respectively. Let  $S$  be the stoichiometric matrix where  $s_{ij}$  is the stoichiometric coefficient of metabolite  $i$  in reaction  $j$ , which is positive for a product and negative for a reactant. Let us define a reactant matrix,  $R$ , and product matrix  $P$ , whose elements are defined respectively as follows:

$$r_{ij} = \begin{cases} 1, & \text{if } s_{ij} < 0 \\ 0, & \text{otherwise} \end{cases}$$

and

$$p_{ij} = \begin{cases} 1, & \text{if } s_{ij} > 0 \\ 0, & \text{otherwise} \end{cases}$$

Let  $B$  represent an  $n$ -dimensional vector containing the total number of reactants within each reaction, such that:  $b_j = \sum_{i=1}^m r_{ij}$ . Let  $\rho(u)$  and  $\phi(u)$  represent vector-valued functions operating on the vector  $u$ , where

$$\rho_i = \begin{cases} 1, & \text{if } u_i = 0 \\ 0, & \text{otherwise} \end{cases}$$

and

$$\phi_i = \begin{cases} 1, & \text{if } u_i > 0 \\ 0, & \text{otherwise} \end{cases}$$

Let a set of seed metabolites,  $C_s$ , contain the indices of metabolites, where for all  $i \in C_s$ ,  $x_i = 1$ .

### 2.0.0.2 Pseudocode

Initialize  $x$  such that  $x_i = 1$  if  $i \in C_s$

$$l_0 = 0, l_1 = \sum_i^m x_i$$

**while:**  $l_k > l_{k-1}$

**do:**

$$y = \rho(R^T x - B)$$

$$x = \phi(Py + x)$$

$$l_{k+1} = \sum_i x_i$$

$$k = k + 1$$

### Data analysis

For the taxonomic enrichment test, we first computed the average number of phosphate-free core network reactions across all species in KEGG. We then randomized the set of 315 reactions and repeated the calculation  $10^5$  times. To test for enrichment for categorical features associated with the core network enzymes (Fig 2.1C, 2.4A-B), a 2x2 contingency table was constructed and a Fisher's exact test was performed. For continuous metrics, we used the nonparametric Kolmogorov-Smirnov test. For all pathway and module enrichment analysis, we used a Benjamini-Hochberg multiple comparison's correction and report only pathways and modules with a False-discovery rate  $< 0.05$ . All statistical tests were performed in MATLAB 2015a, using built-in functions for two-sample Kolmogorov-Smirnov tests (*kstest2.m*), Fisher's exact tests (*fishertest.m*), multiple hypothesis testing (*multcompare.m*). Monte carlo permutation tests were performed using the *randsample.m* function.

## Chapter 3

# Ancient geochemical scenarios converge to an organo-sulfur proto-metabolism

### Summary

This thesis chapter will be published in the following paper:

**Goldford, J. E.**, Hartman, H., Marsland, R., & Segrè, D. *Ancient geochemical scenarios converge to an organo-sulfur proto-metabolism.* manuscript in preparation

### Abstract

Evidence of the key steps that led to the emergence of living systems is likely hidden in the structure of metabolism. One of the main challenges in unraveling these early steps is the difficulty in connecting the uncertain geochemical boundary conditions with the structure and physiology of possible early living systems. Here we first combine network-based algorithms with physiochemical constraints on chemical reaction networks to systematically show how different combinations of boundary conditions (temperature, pH, redox potential and availability of molecular precursors) could have affected the structural evolution of metabolism. We find that a subset of boundary conditions converges to an organo-sulfur-based proto-metabolic network that may have been fueled by a thioester- and redox-driven variant of the reductive TCA cycle, capable of producing lipids and keto acids. Surprisingly, we find that environmental sources of fixed nitrogen and low-potential electron donors (e.g. ferredoxin) may not have been necessary for the earliest phases of biochemical evolution. Next, in analogy with genome-scale models of cellular metabolism, we use one of

these networks to build a steady-state dynamical metabolic model of a proto-cell. We found that different combinations of carbon sources and electron acceptors could support the continuous production of a minimal ancient biomass composed of putative early biopolymers and fatty acids. Our approach highlights the power of multi-scale modeling and systematic analysis to identify plausible geochemical scenarios that could have led to the emergence of ancient living systems.

## **Introduction**

The structure of modern day metabolic networks has been proposed to recapitulate the evolutionary history of metabolism even before the onset of an RNA-based genetic system [48, 77, 76, 40, 133, 175]. Based on this hypothesis, it is possible to propose models for how geochemical initial points may have transitioned into a large and diverse biochemical network [77, 133, 174, 179, 175]. Recent work has translated this hypothesis into a quantitative framework using the network expansion algorithm, which iteratively simulates the emergence of chemical reaction networks from a set of initial compounds [47, 73]. This algorithm has been used to explore the effect of oxygen [160] and phosphate [61] on the structure of biosphere-level metabolism. The network expansion algorithm, when used to reconstruct the earliest history of metabolism prior to a protein translation system, relies on three key assumptions: (1) the majority of enzyme-catalyzed reactions can be catalyzed by inorganic or small molecular catalysts, albeit at much slower rates [61, 64] (due to the large number of plausible non-enzymatic catalysts in ancient living systems [97, 96, 127, 95, 135]), and (2) over long time-scales, there are high rates of horizontal transfer of biochemical reactions [191], and (3) the chemical transformations important in the earliest phases of living systems are still retained in the biosphere. While prior work has focused on either studying the initial expansion of metabolism with a predefined set of compounds [61], or the random exploration of initial geochemically supplied compounds [161], a systematic exploration of plausible geochemical scenarios responsible for the emergence and function of ancient metabolic systems remains unexplored.

Estimates of plausible Archean environments that led to the emergence and evolution of living systems vary dramatically [107, 41], ranging from alkaline hydrothermal vents driven by chemical

gradients [120] to acidic ocean seawater driven by photochemistry [40, 78]. More specifically, there is uncertainty in the composition and availability of geochemically produced substrates for early living systems, including the availability of suitable electron donors and acceptors, fixed carbon (e.g. formate, acetate), fixed nitrogen (e.g. ammonia) and phosphate. Although geochemical data support the availability of mid-potential electron donors ( $H_2$  [167]), sulfur ( $H_2S$ ) and potentially fixed carbon [188] in ancient environments, several key molecules used in living systems may have been severely limiting, including a source of fixed nitrogen [43, 136] (e.g. ammonia), low-potential electron donors [123, 178] and phosphate [72, 169, 93]. Rather than assuming a steady supply of these biomolecules, it is important to ask the question of whether these molecules would have been necessary in ancient proto-metabolic systems. Indeed, we recently used systems biology techniques to ask whether ancient proto-metabolism would have required a source of phosphate, and found evidence that thioesters, rather than phosphate, endowed ancient metabolism with key energetic and biosynthetic capacity [61]. However, it is currently unclear whether other molecules previously proposed to be limiting on early Earth or physiochemical conditions of ancient environments would have prohibited the emergence of proto-metabolism.

In this paper, we systematically explored the environmental conditions that could have led to the emergence of proto-metabolism, and determined the functional capacity of plausible proto-metabolic networks using constraint-based modeling. We found that a thioester-driven network may have lacked phosphate, nitrogen and low-potential electron donors, consisting of only organic molecules and thioesters capable of synthesizing keto acids and fatty acids. Extant enzymes that catalyze reactions in this network are depleted in nitrogen-containing side chains within their active sites, and are depleted in nitrogen-containing coenzymes, suggesting that extant reaction mechanisms may bear similarity to those of primitive organo-sulfur catalysts. We next constructed a constraint-based model of ancient proto-metabolism, and simulated growth in plausible prebiotic environments. Our results suggest that simple thioester-based organic networks embedded in modern-day metabolism may have led to a complex proto-metabolism, potentially capable of sustainability producing proto-cellular components, a critical feature for a metabolism-first theory for the origin of life.

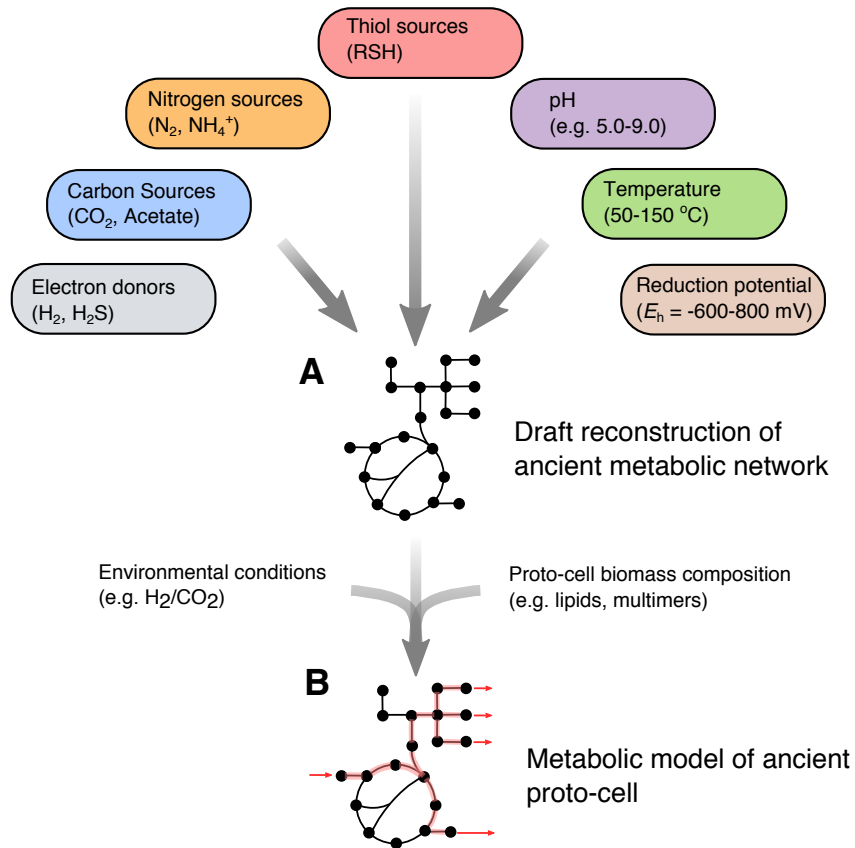


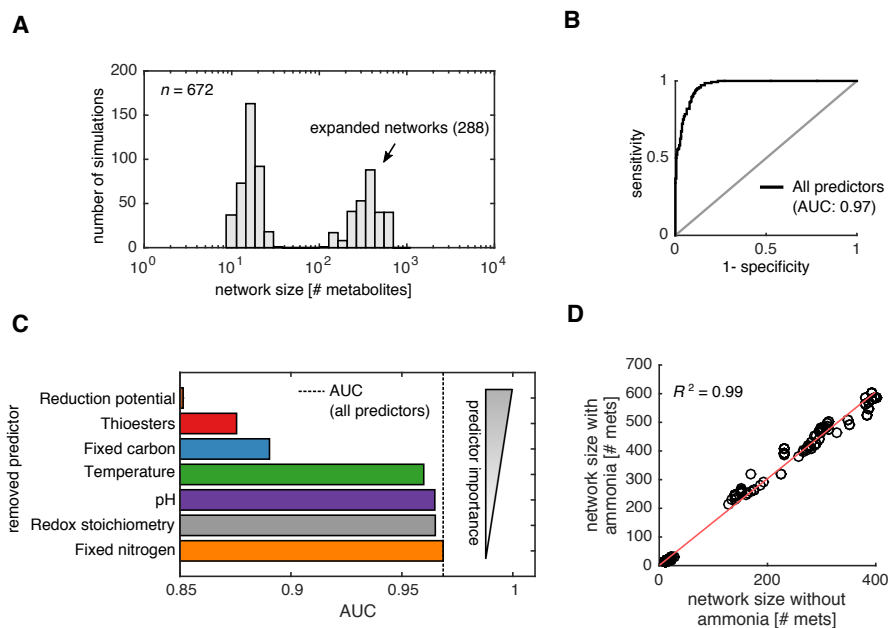
Figure 3.1: **Computational pipeline for the reconstruction and analysis of plausible metabolic networks.** (A) We implemented a thermodynamically-constrained network expansion algorithm using various chemical and physical parameters thought to have been important for the development of ancient proto-metabolism. We varied sources of electrons, carbon, nitrogen, sulfur/thiols, and physical parameters like pH, temperature and the standard reduction potential of primitive electron donors or acceptors, resulting in a draft biosphere-level metabolic network. (B) We then integrated the draft metabolic reconstruction, along with proposed environmental conditions and proposed compositions of ancient proto-cells to construct a metabolic model of ancient proto-cells. We used the model of ancient proto-cells to perform thermodynamic metabolic flux analysis (TMFA) and other constraint-based modeling approaches for a variety of purposes, including determining which reactions are plausibly essential for proto-cell self-replication, as well as estimating which environmental conditions are capable of supporting proto-cellular growth.

## Results

### Results

#### **Nitrogen and low-potential electron donors are not required for initial metabolic expansion**

We first sought to systematically characterize the effect of various geochemical scenarios on the plausible structure of ancient metabolism. Building on prior work [160, 61], we constructed a model of ancient biosphere-level metabolism from the KEGG database [89]. To this end, we modified the network in several ways to account for previously proposed primitive thioester-coupling and redox reactions [61], as well as pruned reactions likely causing local thermodynamic bottlenecks at specific environmental parameters including temperature, pH and redox potential (see Methods). Next, we used a thermodynamically-constrained network expansion algorithm to reconstruct models of ancient biochemical networks [61], which iteratively adds metabolites and thermodynamically-feasible reactions to a network until convergence (see Methods). We performed thermodynamically-constrained network expansion (see Methods and Fig. 3.1A) for  $n = 672$  different geochemical scenarios, systematically varying physical parameters like pH, temperature, the redox potential of primitive redox systems, and the availability of key biomolecules including thiols (that subsequently form thioesters), fixed carbon (formate/acetate) and fixed nitrogen (ammonia) (Methods, Fig. 3.2). Of the 672 different simulated geochemical scenarios, we found that 288 (43%) expanded to networks above 100 metabolites (Fig. 3.2A). To determine which parameters strongly affected whether or not networks expanded beyond 100 metabolites, we trained a logistic regression classifier using geochemical parameters as predictors (see Methods), resulting in a classifier with an area under the receiver operator curve (AUC) of 0.97 with a leave-one out cross validation accuracy of 0.89 (Fig. 3.2B). To determine which parameters were important for the performance of the classifier, we performed backward elimination of variables by removing individual predictors from the classifier, and measured the drop in AUC for each reduced model (Fig. 3.2C). Surprisingly, removing the variable encoding whether ammonia was in the seed set resulted in no drop in AUC, suggesting that a source of fixed nitrogen was not an important parameter leading



**Figure 3.2: Nitrogen is not essential for initial expansion.** A thermodynamically-constrained network expansion algorithm was used to simulate the early expansion of metabolism under 672 scenarios, systematically varying the availability of reductants in the environment, pH, carbon sources, the presence of thiols, temperature and the availability of ammonia. (A) A histogram of network sizes ( $x$ -axis, number of metabolites) revealed that 43 % (288/672) of the scenarios resulted a bimodal distribution, where expansion either occurred beyond 100 metabolites. (B) A logistic regression classifier was constructed to predict whether a geochemical scenario resulted in a network that exceeded 100 metabolites, and a receiver operating curve (ROC) was plotted. The trained classifier resulted in an area under the curve (AUC) of 0.97 and leave-one out cross-validation accuracy of 0.89. (C) Models were trained without information on specific geochemical variables ( $y$ -axis), and the ensuing AUC was plotted as a bar-chart ( $x$ -axis), revealing that knowledge of the availability of fixed nitrogen offers no information on whether networks expanded. (D) We plotted the network sizes (number of metabolites) before the addition of ammonia ( $x$ -axis) verse after the addition of ammonia ( $y$ -axis), revealing a strong linear relationship, fitting a model that a fixed fraction ( $\gamma = 0.517 \pm 0.007$ ) of compounds can react with ammonia.

to initial expansion. Analysis of the network sizes with and without the addition of ammonia into the seed set revealed a strong linear relationship (Fig. 3.2D), suggesting that a fixed proportion of molecules in the network reacted with ammonia. Interestingly, the enzymes that catalyze reactions in the expanded networks before the addition of ammonia were depleted in nitrogen-containing coenzymes (Fig. 3.6), one-tailed Wilcoxon sign rank test:  $P < 10^{-24}$ ) and active site amino acids with nitrogeneous side chains (see Fig. 3.6, one-tailed Wilcoxon sign rank test:  $P < 10^{-24}$ ) relative to enzymes added after the addition of ammonia (see Supplemental Text). Together, these results suggest that the addition of ammonia played no role in the initial expansion of the network, only amending compounds to an thioester-coupled organo-sulfur metabolic network (Fig. 3.2).

The simulations described above revealed a number of relationships between plausible geochemical scenarios and the structure and size of our simulated proto-metabolic networks. First, expansion beyond 100 metabolites occurred without a source of fixed carbon only when thiols were provided in the seed set, highlighting the importance for thioester-coupling for ancient carbon fixation pathways [61, 123, 178]. The presence of thiols enabled the production of key biomolecules, including fatty acids and branched chain keto acids. Second, we explored the effect of the primitive redox system by systematically varying the reduction potential of the electron donor in the seed set (see Methods, Fig. 3.3A). Surprisingly, we found that as we increased the fixed potential of the electron donor, expansion collapsed above reduction potentials between -150 and 50 mV, suggesting that low-potential electron donors were not necessary for expansion (Fig. 3.3B). Overall these results suggest that the emergence of an autotrophic proto-metabolic network capable of producing key biomolecules were contingent on a mid-potential redox couples as well as thiols capable of forming thioesters, two functions potentially carried out by disulfides. Moreover, these results suggest that production of low-potential ferredoxin from  $H_2$  may have not been necessary in ancient proto-metabolic systems.

### **Expanded models converge to a core thioester- and redox-driven rTCA cycle**

Analysis of the expanded networks without nitrogen revealed that a large number of scenarios converged to similar metabolic networks, spanning variants of key pathways in central carbon

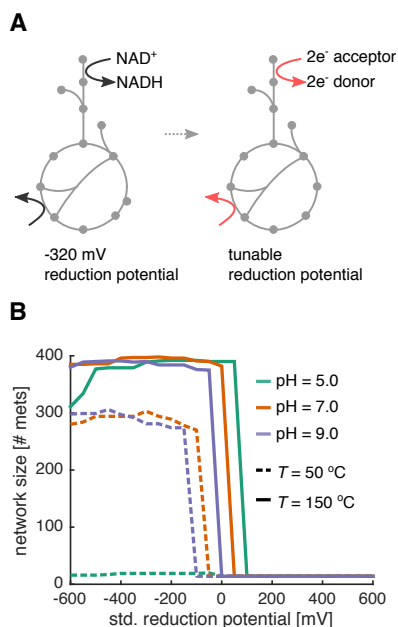
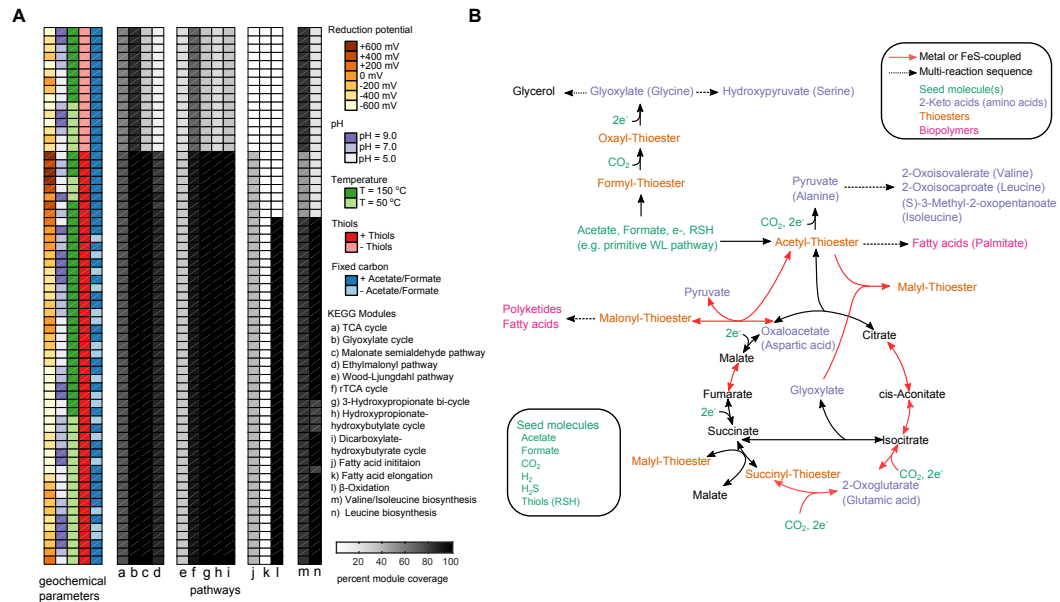


Figure 3.3: **Reduction potential of nicotinamide and flavin substitutes influences network expansion** (A) Redox coenzymes (NAD, NADP, and FAD) were substituted with an arbitrary electron donor/acceptor at a fixed reduction potential. (B) We performed thermodynamic network expansion in acidic (pH 5), neutral (pH 7) and alkaline (pH 9) conditions at two temperatures ( $T = 50$  and  $150$  °C), using a two electron redox couple at a fixed potential ( $x$ -axis) as a substitute for NAD(P)/FAD coupling in extant metabolic reactions (see Methods). We plotted the final network size across all pH and temperatures with no fixed carbon sources (e.g. only  $\text{CO}_2$ ) and thiols. Notably, for these simulations, we used a base seed set of:  $\text{H}_2$ ,  $\text{H}_2\text{S}$ ,  $\text{H}_2\text{O}$ ,  $\text{HCO}_3^-$ ,  $\text{H}^+$  and  $\text{CO}_2$ .

metabolism (Fig. 3.4). For the majority of simulations, variants of modern heterotrophic carbon assimilation pathways, including the glyoxylate cycle and TCA cycle, were highly represented in the network (Fig. 3.4A). Additionally, several carbon fixation pathways were also highly represented in the simulated networks: in over half of the networks that expanded beyond 100 metabolites, 92% (12/13) of the compounds (or generalized derivatives) in the reductive tricarboxylic acid (rTCA) cycle were observed, with the exception of phosphoenolpyruvate. We also found that in several geochemical conditions, all intermediates were producible for three carbon fixation pathways, including the 3-hydroxypropionate bi-cycle, the hydroxypropionate-hydroxybutyrate cycle, and the dicarboxylate-hydroxybutyrate cycle (Fig. 3.4A). At-most, only 3 of 9 metabolites used in the

Wood-Ljungdahl (WL) pathway was observed, due to the lack of nitrogen-containing pterins in the network. Early variants of the WL-pathway could have been radically different than today's WL-pathway, relying on native metals to facilitate reduction of CO<sub>2</sub> to acetate in ancient living systems [188, 178]. In addition to observing a large number of metabolites used in carbon fixation pathways, we found that a large fraction of the  $\beta$ -oxidation pathway was represented in our networks, which may have supported the production of fatty acids in ancient living systems by operating in the reverse direction. Lastly, we also observed that the majority of intermediates involved in the production of branched-chain amino acids were also producible our networks.

We next analyzed the convergent organo-sulfur proto-metabolic network in more detail by comparing the network to extant metabolic pathways, as well as identifying classes of compounds producible from the network. In Fig. 3.4B, we show a variant of the (r)TCA cycle that may have served as the core organo-sulfur network fueling ancient living systems. Rather than using ATP-dependent reactions used in extant species (e.g. Succinyl-CoA synthetase and ATP citrate lyase), these reactions are substituted with non-ATP-dependent reaction mechanisms. For instance, the production of a succinyl-thioester in the extant rTCA cycle relies on Succinyl-CoA synthetase, performing the following reaction:  $\text{ATP} + \text{Succinate} + \text{CoA} \rightarrow \text{Succinyl-CoA} + \text{ADP} + \text{P}_i$ . However, in the network presented in Fig. 3.4B, malyl-thioester, producible through alternative reactions, donates a thiol to succinate, subsequently forming a succinyl-thioester. From this (r)TCA cycle analogue, eight keto acids normally serving as key intermediates and precursors to common amino acids in central carbon metabolism were producible, namely glyoxylate, pyruvate, oxaloacetate, 2-oxoglutarate and hydroxypyruvate, as well as the following branched-chain keto acids: 2-oxoisovalerate, 2-oxoisocaproate, and (S)-3-methyl-2-oxopentanoate. Additionally, long-chain fatty acids like palmitate are producible in this network, driven by thioester and redox-coupling rather than ATP, like in extant fatty acid biosynthesis. Thus, despite the simplicity of seed compounds, several small molecular weight keto acids and fatty acids may have been producible in an organo-sulfur proto-metabolism.



**Figure 3.4: Systematic exploration of prebiotic scenarios reveals a core organo-sulfur network.** (A) A thermodynamically-constrained network expansion algorithm was used to simulate the early expansion of proto-metabolism under various scenarios, including the availability of reductants in the environment, pH, temperature, and the availability of fixed carbon sources and thiols. The proportion of molecules selected KEGG modules involved carbon metabolism are plotted as a heatmap to the right of the parameters. (B) A representation of the core network producible from a prebiotically plausible seed set without both nitrogen and phosphate (bottom left box). Acetyl-thioesters are first produced, potentially from a primitive Wood-Ljungdahl pathway [179, 188] from acetate and thiols provided as seed molecules (green). Acetyl-thioesters enable the production of all intermediates in the reductive tricarboxylic acid (rTCA) cycle, with the exception of phosphoenolpyruvate. ATP-dependent reactions in the rTCA cycle may have been substituted with a primitive malate synthase and transthioesterification of succinate as well as the recently discovered reversible citrate synthase [117, 140]. The keto acid precursors for 8 common amino acids (A,D,E,G,I,L,S,V) are highlighted in purple, while routes to thioester-mediated polymerization of fatty acids and polyketides are highlighted in pink.

### **Flux balance model of primitive proto-cells supports a chemoautotrophic origin of life**

So far, we have studied only the topology of ancient metabolic networks, agnostic to whether and how such a metabolic network could fuel primitive proto-cells with internal energy sources (e.g. thioesters), redox gradients, and proto-cellular materials used for catalysis and compartmentalization. Flux balance analysis (FBA), originally developed for the study of microbial metabolism, enables the prediction of systems-level properties of metabolic networks at steady-state. Fundamentally, FBA simulates the balanced growth of a collection of biomolecules, which are the major material and energy demands of the cell. In microbial metabolism, FBA is used to simulate the production of cellular biomass (e.g. protein, lipids, and nucleic acids) at fixed proportions, which are derived from known composition of extant cells. However, unlike modern day microbial cells, the biomass composition of plausible ancient proto-cells is unknown.

Christian de Duve suggested that the thioester-driven polymerization of monomers producible from ancient proto-metabolism may have produced "catalytic multimers," which were proposed to serve as catalysts for ancient biochemical reactions [40]. If prebiotic environments were severely nitrogen limited, keto acids producible from protometabolism (see Fig. 3.4B) may have been reduced to  $\alpha$ -hydroxy acids, and polymerized into polyesters using thioesters as a condensing agent (Fig. 3.8). Recent work has suggested that polymers of  $\alpha$ -hydroxy acids may have been stably produced in geochemical environments [?], and that these molecules may serve as primitive catalysts [53]. Thus, we propose that the thioester-driven polymerization of  $\alpha$ -hydroxy acids, producible from keto acid precursors for common amino acids, may have served as ancient catalysts.

Using an expanded metabolic network as a scaffold for network reconstruction (Fig. 3.1A), we constructed a constraint-based model of an ancient proto-cell using a biomass composition consisting of fatty acids (for proto-cellular membranes), "catalytic multimers" derived from eight keto acids (Fig. 3.4B), and redox and thioester-based free energy sources (Methods, Fig. 3.5A), and determined if growth was achievable using thermodynamic metabolic flux analysis (TMFA), a variant of FBA that explicitly considers thermodynamic constraints [83] (see Methods). Using this approach, we found that growth of the proto-cell metabolic model is feasible under a wide

variety of assumptions regarding macromolecular compositions, as well as input molecules (Fig. 3.5B). Notably, we found that growth is achievable in simple chemoautotrophic conditions with either H<sub>2</sub> or H<sub>2</sub>S, but not Fe(II) or ferredoxin, as electron donors (Fig. 3.5B). In this model, thiols and thioesters were not supplied as food sources in our model, and are recycled during steady-state growth of the model proto-cell. Initially, thiols could have been supplied abiotically, with the rapid takeover of biotic production of mercaptopyruvate, a keto acid that could have been incorporated into primitive multimers.

In addition to predicting growth yields of metabolic networks, constraint-based modeling of cellular metabolism also enables predictions of flux distributions of intracellular reactions. We hypothesized that if these steady-state flux distributions accurately recapitulated plausible ancient proto-cellular metabolism, then there should be a high flux through reactions relying on inorganic and transition metals relative to random sets of reactions in the network. Indeed, the total flux through reactions relying on inorganic or metal ions was significantly higher than randomly sampled sets of reactions for both chemoautotrophic growth on H<sub>2</sub> (Monte Carlo Permutation test:  $P < 0.05$ ), or chemoautotrophic growth on H<sub>2</sub>S (Monte Carlo Permutation test:  $P < 0.05$ ).

We next used the proto-cell metabolic model to determine which reactions are essential for growth by removing individual reactions from the metabolic model and solving for maximum growth yield. We found that 28 reactions were essential for chemoautotrophic growth on either H<sub>2</sub> or H<sub>2</sub>S. These essential reactions were primarily involved in the synthesis of keto acid precursors to branched chained amino acids, and fatty acids: 18 of the essential reactions were involved in thioester-mediated fatty acid metabolism and biosynthesis, while 9 were involved in the synthesis of branched-chain keto acids. While recent work has showed that several reactions in central metabolism can be catalyzed non-enzymatically, little work has been done to test for the non-enzymatic synthesis of these important molecular components. Thus, constraint-based modeling may serve as a useful tool to identify key metabolic reactions that would have been necessary for proto-cellular growth.

## Discussion

While most efforts to reconstruct the ancient phases of biochemistry have traditionally relied on building qualitative models of small pathways or metabolic reactions [77, 40, 192, 133, 174], we found that quantitative modeling of ancient proto-metabolic networks illuminated key constraints that, if not satisfied in ancient environmental conditions, may have limited the development of ancient living systems. By computationally mapping geochemical scenarios to plausible ancient proto-metabolic structures, we were able to estimate portions of extant biochemistry that may have been very sensitive to initial geochemical conditions, and simultaneously identify the emergent pathways which were robust to variations in environmental conditions. For example, we were able to identify that thiols and thioesters would have been essential for the production of fatty acids and branched-chain keto acids (Fig. 3.4A). Thus, by quantitatively identifying associations between plausible ancient environmental conditions and the plausible structures of ancient proto-metabolism, we can propose key geochemical conditions that may have been required for the earliest phases of biochemistry.

Our approach also revealed that environmental sources of fixed nitrogen and low-potential electron donors, two important molecules typically assumed to have been required in ancient metabolic systems [121, 179, 123], may have not been necessary during the earliest phases of biochemical evolution. The feasibility of nitrogen and phosphate-free proto-metabolic network suggests that a substantial degree of complexity may have evolved prior to incorporation of nitrogen into the biosphere [76]. In addition to being a key component of biomolecules like amino acids and nucleic acids, nitrogen plays critical roles in catalysis within the active sites of modern day enzymes, including several enzymes within the network model constructed here. It is possible that such roles may have been preceded by positively charged surfaces or metal ions [135, 188], which could have been replaced by amino /keto acids with nitrogen side chains once nitrogen became incorporated into proto-metabolism. Furthermore, contrary to our expectations [179, 178, 123], our simulations predict that low-potential ferredoxin was not a necessary component of early living systems, consistent with the proposal that low-potential ferredoxin is not necessary for acetogenesis [9]. If the



ducible keto acids) could have lead to primitive organic catalysts [40], in combination with inorganic minerals or metal ion catalysts [135, 188]. Future experiments could synthesize these putative catalysts to directly test for the catalysis of key reactions in the network. Additionally, it remains interesting to see if these proposed organic compounds are produced in living systems today via mechanisms similar to polyketide or non-ribosomal peptide synthesis.

By building a constraint-based model from this core organo-sulfur proto-metabolism, we were able to simulate non-equilibrium steady-states achievable by the reconstructed proto-metabolic model. This approach allowed us to determine which environmental conditions could have led to collective growth of the chemical network, revealing that high-potential electron donors (e.g. Fe(II)), would have prohibited collective autocatalysis. Simulations using constraint-based modeling thus help refine the scope of environmental conditions worth testing in future experimental studies for the origin of proto-metabolism.

Although detailed predictions regarding the specific mechanisms of catalysis and information storage are not explored in this work, our results highlight the utility of constraint-based modeling in assessing feasible scenarios that may have led to the emergence of living systems. Models could improve iteratively by incorporating new experimental information, including non-enzymatic reactions recently observed experimentally [97, 96, 127, 95, 135, 188]. Thus, beyond simply identifying constraints that might have lead to emergence of living systems, constraint-based models of proto-metabolism offer the ability to concretely synthesize testable models of ancient living systems, a key tool for the ultimate goal of experimentally synthesizing artificial proto-cells capable of collective self-replication.

## **Materials and methods**

### **Reconstruction of biosphere-level metabolic network**

Biosphere-level metabolism was reconstructed from the KEGG database [89] according to protocol described previously [61]. We modified the network in several ways to model primitive thioester-based metabolic network without nitrogen or phosphate. First, to simulate the availability of thiols

capable of forming thioesters, we included Coenzyme A, Acyl-Carrier Protein and Glutathione into the seed set. However, to enforce the constraint these metabolites could only be used in reactions as coenzymes (and not products or substrates), we prevented the degradation by removing KEGG reactions R10747, R02973 and R02972.

We next assigned standard molar free energies to reactions using eQuilibrator at a predefined pH [52]. Next we substituted NAD, NADP and FAD-coupled reactions with an arbitrary redox couple. For example, if the redox reaction  $X_{ox} + \text{NADH} \rightarrow X_{red} + \text{NAD}^+$  was swapped with electron donor with a redox potential of  $E_0^+$  mV, we would use the following formula to adjust the standard molar free energy for the new reaction  $r'$ :

$$\Delta_{r'}G'^{\circ} = \Delta_rG'^{\circ} + nF(E_0^+ - E_0)$$

where  $n$  are the number of electrons transferred in reaction  $r$  and  $F = 96.485$  kJ/V. Note that if we assumed that the electron donor/acceptor substitute was a two electron donor/acceptor, we did not change the stoichiometry in the reaction equation. However, in the case where the electron donor/acceptor substitute was a single electron donor/acceptor, we change the stoichiometric coefficients to  $s_{cj} = 2$  for all reactions  $j$ , where  $c$  represents metabolites NAD(H), NADP(H) and FAD(H<sub>2</sub>). For this work, we systematically varied the reduction potential  $E_0^+$  and stoichiometry of the primitive redox coenzyme.

### **Thermodynamically-constrained network expansion**

We performed network expansion using thermodynamic constraints in a different way than performed previously [61] Previously, we removed reactions above a predefined free energy threshold of  $\tau = 30$ kJ/mol [61]. For this work, we computed the lowest reaction free energy possible using estimates for upper and lower bounds on metabolite concentrations,  $u_i$  and  $l_i$ , and removed reactions with a positive reaction free energy. For a given biochemical reaction at fixed temperature

and pressure,  $\Delta_r G'$  is defined as:

$$\Delta_r G' = \Delta_r G'^{\circ} + RT \ln \prod_i a_i^{s_{ir}}$$

where the  $\Delta_r G'^{\circ}$  is the free energy change of the reaction at standard molar conditions,  $R$  is the ideal gas constant,  $T$  is temperature,  $a_i$  is the activity of metabolite  $i$  and  $s_{ir}$  is the stoichiometric coefficient for metabolite  $i$  in reaction  $r$ . We fixed  $a_i$  for each reactions according to the following rules:

$$a_i = \begin{cases} u_i, & \text{if } s_{ir} < 0, \\ l_i, & \text{if } s_{ir} > 0. \end{cases} \quad (3.1)$$

We then removed reactions with a  $\Delta_r G' \geq 0$ . For all simulations we assumed that  $u_i = 10^{-1}$  M and  $l_i = 10^{-6}$  M. Note that because we model each reaction independently, metabolite concentrations could be inconsistent. For instance, if metabolite  $i$  is the substrate for reaction  $a$  and a product for reaction  $b$ , then  $x_i = u_i$  for reaction  $a$  and  $x_i = l_i$  for reaction  $b$ .

Using this procedure to systematically remove reactions that were considered to be thermodynamically infeasible, we performed network expansion [47, 73, 160] using the computational procedure described in [61].

### Parameters for network expansion

We systematically studied the size and composition of networks under precise environmental conditions by varying (a) the reduction potential from the environment, (b) pH, (c) temperature, (d) the presence or absence thiols, (e) the inclusion of fixed carbon into the seed set and (f) the inclusion of fixed nitrogen into the seed set. We now discuss each of these parameters in more detail:

- *Reduction potential and stoichiometry.* A wide range of environmental conditions could have provided electron donors at various potentials: high potential redox pairs, with strong oxidants like Fe(III), may have been present in oceans at high concentrations, while strong

reductants like  $H_2$ , disulfides, proto-ferredoxin, or reductive carboxylation of thioesters have been produced via serpentinization or geochemical analogues of primitive metabolic pathways [123]. We substituted reactions coupled to NAD, NADP and FAD with a generic single or double electron donor and acceptor pair at a fixed potential. To prevent unbalanced electron transfer, we removed the following transhydrogenase reactions: R10159, R01195, R00112, R09520, R09748, R05705, R05706, R09662, R09750. We then created a single or double electron donor/acceptor pair with a fixed reduction potential,  $E_0^+$ , ranging from -600 to 600 mV.

- *pH* We modified the pH by setting reaction free energies at various pH's (5.0-9.0) using eQuilibrator [52] which relies on the component contribution method [138].
- *Temperature*. Temperatures were assumed to have been within a range of 50-150 °C, spanning estimates of ocean seawater temperature in the Archean [70], up to some alkaline hydrothermal vent systems [121].
- *Thiols*. In our model we provided thiols that serve as substitutes for coenzymes that form thioester bonds in extant metabolic networks. To this end, we provided Coenzyme A, acyl-carrier protein and Glutathione in the seed set, but removed key degradation reactions to ensure these compounds only served as coenzymes, rather than material sources, during network expansion [61].
- *Fixed nitrogen*. To study the consequences of adding or removing a source of fixed nitrogen as a seed compounds for network expansion, we either added or removed ammonia from the seed set prior to expansion.

In addition to parameters we varied, we kept constant two additional parameters that could be studied in future work:

- *Metabolite concentrations*. Metabolite concentrations were assumed to be within 1  $\mu$ M - 100 mM. The upper bound estimate is consistent with recent experimental data showing

that key metabolites (formate, methanol, acetate and pyruvate) can be produced near 100 mM [188]. Although we do not have empirical evidence to suggest a reasonable lower bound on metabolite concentrations in ancient metabolic networks, we assumed that 1  $\mu$ M, the estimated lower bound in today’s cells [10], was also the lower bound in our model of ancient metabolism.

- *Reactions with no free energy estimate.* 53 % of the biosphere-level metabolic network reactions have no free energy estimate (4851 of 9074). For all simulations presented in this paper, we assumed these reactions were blocked and did not include them in the network.

### Generalized linear modeling of network expansion results

To access the effects of various parameters on the outcome of network expansion, we used generalized linear models to construct logistic regression classifiers to predict whether or not the network expanded beyond 100 metabolites using a combination of predictors, including categorical variables encoding whether or not ammonia, thiols or fixed carbon was provided in the seed set, and continuous variables encoding the reduction potential, pH and temperature used in each simulation. We first define the response variable for simulation  $i$  as  $y_i$  where  $y_i = 1$  if the simulation resulted in a network that expanded beyond 100 metabolites, and zero otherwise. For the set of simulations performed in Fig. 2 in the main text, we constructed a design matrix consisting of categorical variables representing the following scenarios:

1.  $x_{N,i} \in \{0, 1\}$ : 1 if ammonia was included in the seed set, and 0 otherwise.
2.  $x_{S,i} \in \{0, 1\}$ : 1 if thiols were included in the seed set, and 0 otherwise.
3.  $x_{C,i} \in \{0, 1\}$ : 1 if fixed carbon (acetate/formate) was included in the seed set, and 0 otherwise.
4.  $x_{H,i} \in \mathbb{R}_{>0}$ : A continuous variable representing the pH. Note for our simulations, we only explored acidic (pH=5), neutral (pH=7) and alkaline (pH=9) regimes.

5.  $x_{E,i} \in \mathbb{R}$ : A continuous variable representing the reduction potential at standard molar conditions (at the specified pH listed above). For our simulations, we explored a wide range of standard molar reduction potentials (from -600 mV to +600 mV).
6.  $x_{T,i} \in \mathbb{R}$ : A continuous variable representing the temperature. For our simulations, we explored two temperatures: a high temperature regime ( $T = 150$  °C), and a low temperature regime ( $T = 50$  °C).

We next constructed the following generalized linear model to model whether the network expanded beyond metabolites:

$$\text{logit}(y_i) = \beta_0 + \beta_N x_{N,i} + \beta_S x_{S,i} + \beta_C x_{C,i} + \beta_H x_{H,i} + \beta_E x_{E,i} + \beta_T x_{T,i} \quad (3.2)$$

We fit the parameters ( $\beta$ ) using the *fitglm.m* function in MATLAB 2015a, and a receiver operating curve (ROC) was generated using the *perfcurve.m* function. For results presented in Fig. 2C in the main text, individual predictors were removed in the generalized linear model presented above. To assess whether the trained logistic model served as an accurate classifier, we performed leave-one out cross-validation by removing individual samples from the training set and testing the accuracy of the trained classifier on the removed sample. This procedure resulted in a cross-validation accuracy of 0.89.

In Fig. 2D in the main text, we plotted the network size before ( $x$ ) and after ( $y$ ) the addition of ammonia, and fitted the following linear to the data  $y = (1 + \gamma)x + \epsilon$ , where  $\gamma$  represents a fixed fraction of compounds that react with ammonia, and  $\epsilon$  is a normally distributed noise term. We used the MATLAB function *fitglm.m* to fit the linear model, revealing that  $\gamma = 0.517 \pm 0.007$ .

### Constraint-based modeling

We constructed a model of an autocatalytic network at steady state using a variant of constraint-based modeling of cellular metabolism called thermodynamic-based metabolic flux analysis (TMFA) [83]. TMFA transforms the non-linear constraints induced by imposing thermodynamic consistency into mixed-integer linear constraints. In this section, we first describe (a) the construction

of primitive biomass composition for a model of an ancient proto-cell and (b) the formulation of TMFA used in this analysis.

### **Prebiotic biomass equation**

We constructed a simple model for the macromolecular composition of primitive proto-cells, using empirical knowledge of extant cellular life. Since our metabolic model of proto-metabolism does not include macromolecular production of nucleotides (and thus a nucleic acid based genetic system), we assume that the primary role of proto-cellular metabolism was to initially produce components for a cellular membrane and catalysts. Building off of Christian de Duve's multimer hypothesis [40], we first propose that the biomass can be constructed using a simple two parameter model consisting of the mass fraction of lipids  $\phi_L$  and the average length of each catalytic multimer  $n$ .

- *Lipid mass fraction.* The lipid content in modern cells is roughly 10% of the total dry mass (Bionumbers ID: 111209) [130], primarily composed of the fatty acid palmitate. For our analysis, we assume that palmitate represents the sole component of lipids. Future models could incorporate glycerol, which enables the production of glycerolipids. While phosphate is used in cellular membranes as a polar head group to produce amphiphilic molecules, primitive processes may have conjugated negatively charged organic acids (e.g. oxalate) to glycerol via a thioester-mediated synthesis mechanism to create amphiphilic lipid molecules resembling modern phospholipids. For our initial model, we simply propose that palmitate was the initial amphiphilic component of primitive membranes, where the negatively charged polar carboxylate ion was sufficient for forming a membrane, and assumed that proto-cells consisted of a lipid mass fraction of  $\phi_L$ .
- *Catalytic multimer mass fraction.* We propose that ancient catalysts were composed of inorganic molecules (e.g. iron-sulfur clusters, metal ions, mineral surfaces) chelated with multimers of  $\alpha$ -hydroxy-acids (see Fig. 5A in main text). For our model, we assume that the eight keto acid precursors producible from our network were the dominant monomers of

ancient multimeric catalysts. We assume that the total mass fraction of these catalysts are  $1 - \phi_L = \phi_C = \sum_k \phi_k$ , where  $\phi_k$  is the mass fraction of polymerized monomer  $k$ . For our analysis, we assumed that each monomers is uniformly distributed, so that  $\phi_k = \text{constant}$  for all  $k$ . Additionally, since each monomer must be reduced to  $\alpha$ -hydroxy acids, there is linear relationship between the electron demand,  $s_e$ , and the number of molecules of monomers produced. The stoichiometric equivalents of electron donors are thus:

$$s_e = 2 \sum_k \frac{\phi_k}{M_k}$$

where  $M_k$  is the molar mass of monomer  $k$ .

- *Average size of catalytic multimers.* The average size of mulimeric catalysts sets the number of thioester bonds required for synthesis of catalytic multimers. For each polymer of size  $n$ , there are  $n-1$  thioester bonds required. In our model, the total number of monomers are fixed to be:  $\sum_k \frac{\phi_k}{M_k}$ , where  $M_k$  is the molecular weight for monomer  $k$ . Thus for a fixed monomer length  $n$ , we can compute the number polymers using the following formula:

$$P(n) = \frac{1}{n} \sum_k \frac{\phi_k}{M_k}$$

The thioester demand is thus  $s_t(n) = (n - 1)P(n)$ , or:

$$s_t(n) = \frac{n - 1}{n} \sum_k \frac{\phi_k}{M_k}$$

For our analysis we assumed a fixed polymer length of size  $n = 10$  monomers.

Using these two parameters, we constructed the biomass equation for the proto-cellular model.

### **Thermodynamic Metabolic Flux Analysis (TMFA)**

To simulate a thermodynamically-feasible steady-state behavior of this metabolic network, we used thermodynamic metabolic flux analysis (TMFA) [83]. Briefly, TMFA transforms the non-linear

constraints induced by imposing thermodynamic consistency into mixed-integer linear constraints. We first converted the model into an irreversible model by modeling each reaction as both forward and backward half reactions. We then constructed the following mixed-integer linear program (MILP) to find a flux vector,  $v$  (with elements  $v_r$  for each reaction  $r$ ), log-transformed metabolite concentrations ( $\ln(x)$ ) and binary variables indicating whether a reaction is feasible ( $z$ ) given a specific objective function was satisfied. The objective function used in this work was to maximize biomass yield, similar to the objectives frequently used in FBA model of microbial metabolism. Thus, the optimization problem was constructed according the following MILP:

$$\begin{aligned}
& \underset{\ln(x), v, z, e}{\text{maximize}} && v_{\text{biomass}} \\
& \text{subject to} && Sv = 0 \\
& && 0 < v_r \leq z_r ub_r, \forall r \in \mathcal{R} \\
& && z_r K - K + \Delta_r G' < 0, \forall r \in \mathcal{R} \\
& && \Delta_r G'^o + RT \sum_i s_{ir} \ln(x_i) + \sigma_r e_r = \Delta_r G' \\
& && \ln(10^{-6}) \leq \ln(x_i) \leq \ln(10^{-1}), \forall i \in \mathcal{M} \\
& && -\sigma_m \leq e_r \leq \sigma_m \forall r \in \mathcal{R}
\end{aligned} \tag{3.3}$$

where  $\mathcal{R}$  and  $\mathcal{M}$  are the sets of all reactions and metabolites, respectively. As discussed in detail elsewhere [83], the first equation in the constraint set ensures that intracellular metabolite concentrations are at steady-state, and are simply mass balance constraints for each metabolite. The second equation sets the bound on individual reaction fluxes, where the maximum flux through reaction  $r$  is  $ub_r$ . Note that when  $z_r = 0$ , the flux through reaction  $r$  is constrained to 0. The third equation sets ensures that  $z_r = 1$  if and only if  $\Delta_r G' < 0$ , and  $z_r = 0$  otherwise. Note that  $K$  is a large number ( $K > \max_r \{\Delta_r G'\}$ ) ensuring that this constraint is not violated with  $z_r = 0$ . The fourth equation is the free energy of each reaction as a function of log-metabolite concentrations. Note that we also add slack variables,  $e_r$ , to account for the possible error in the estimating standard molar reaction free energies for each reaction (where  $\sigma_r$  is the standard error for each reaction  $r$ ),

which are bounded by a global error tolerance  $\sigma_m = 2$  (set in equation 6). Lastly, equation 5 simply constrains the log-metabolite concentrations to be bounded between  $1\mu\text{M}$  and  $100\text{ mM}$ . After each simulation, we performed a secondary optimization to find the minimal set of reactions that achieve the optimal growth rate by minimizing the  $l_1$ -norm of the flux distribution subject to the constraint that  $v_{\text{biomass}} = v_{\text{biomass}}^*$

Numerical simulations were formed using the COBRA toolbox and the Gurobi optimizer (Version 7.0.1). All source code is provided in the following github repository: [http://www.github.com/jgoldford/protometabolic\\_modeling](http://www.github.com/jgoldford/protometabolic_modeling).

### **Calculation of coenzyme and sequence-level features within enzymes**

To determine which reactions were associated with specific coenzymes (for results presented in Fig. 3.6B,D) we downloaded information for each Enzyme Commission number (E.C.) in the KEGG ENZYME database (<http://www.genome.jp/kegg/annotation/enzyme.html>). We downloaded each page and parsed the "comment" field for each E.C. and performed a text-based search to identify coenzymes associated with each E.C. number. We searched for text indicating that the enzyme mechanisms used one of the following coenzymes, cofactors and iron sulfur clusters: biotin, heme, PLP, TPP, pterin, molybdopterin, flavin, Fe, Co, Ni, Cu, Mn, W, Zn, Mo, Mg, FeS, FeFe, Fe<sub>2</sub>S<sub>2</sub>, Fe<sub>3</sub>S<sub>4</sub> and Fe<sub>4</sub>S<sub>4</sub>, respectively. We also searched E.C. numbers indicating that the reaction mechanisms are non-enzymatic.

For results presented in 3.6B, we computed the fraction of reaction E.C. numbers that were associated with one of the following coenzymes: Fe, Co, Ni, Cu, Mn, W, Zn, Mo, Mg, FeS, FeFe, Fe<sub>2</sub>S<sub>2</sub>, Fe<sub>3</sub>S<sub>4</sub> and Fe<sub>4</sub>S<sub>4</sub>, or was marked as non-enzymatic. For results presented in 3.6D, we computed the fraction of reaction E.C. numbers that were associated with one of the following coenzymes: biotin, heme, PLP, TPP, pterin, molybdopterin, and flavin.

For results presented in Fig. 3.6E, we obtained a database of known enzyme active site residues [162]. We first mapped the network reactions to E.C. numbers listed in KEGG, then identified active sites corresponding to E.C. numbers within the the expanded network. We next computed the fraction of active site residues containing nitrogenous side-chains, derived from the following

amino acids: Arginine (R), Lysine (K), Glutamine (Q), Asparagine (N), Histidine (H), and Tryptophan (W).

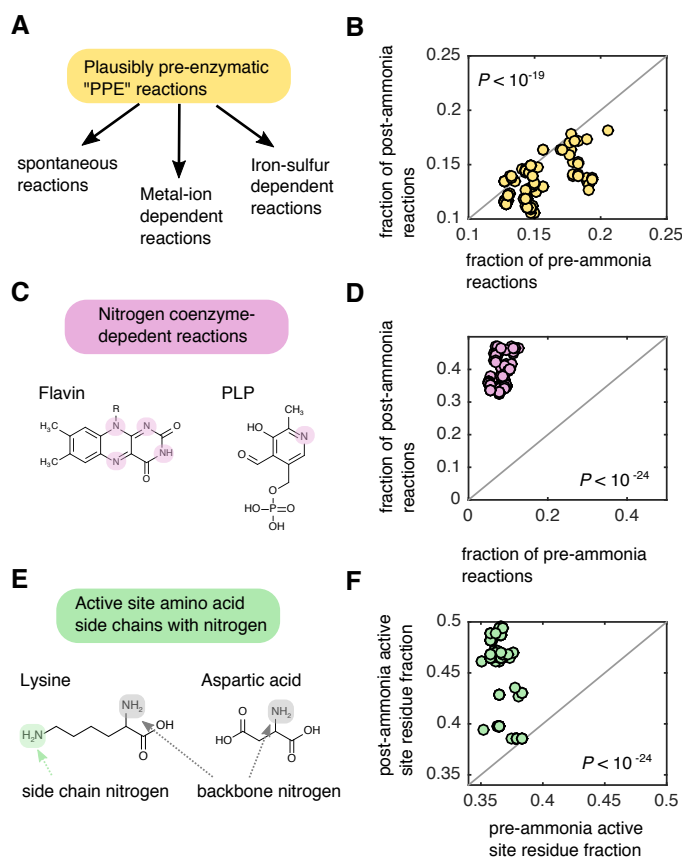
### **Network enzymes retain features of nitrogen-free catalysts**

In order for the expanded networks presented in the main text to have operated in prebiotic conditions, reactions would have been catalyzed non-enzymatically by inorganic or simple organic catalysts available in prebiotic environments. Prior work has suggested that reactions in metabolic networks that proceed spontaneously or depend on enzymes with inorganic coenzymes, such as iron-sulfur or transition metal cofactors, may have operated in prebiotic conditions [179, 135, 188, 178]. We identified reactions in KEGG that could proceed spontaneously or are dependent on one of several inorganic coenzymes (Methods), and defined this set of reactions as *plausibly pre-enzymatic*, or “PPE”-reactions (Fig. 3.6A). For each proposed prebiotic scenario that lead to expansion of at-least 100 metabolites ( $n = 144$ , Fig 3.6A), we partitioned reactions added to the network before the inclusion of ammonia into the seed set (herein called pre-ammonia reactions) and reactions added to the network after ammonia was added to the seed set (or post-ammonia reactions). We then computed the fraction pre- and post-ammonia reactions that were classified as PPE reactions, and found that pre-ammonia reactions contained a higher proportion of PPE-reactions relative to post-ammonia reactions (one-tailed Wilcoxon sign-rank test:  $P < 10^{-19}$ ), suggesting that the pre-ammonia reactions may have been more readily catalyzed by simple inorganic catalysts in prebiotic environments. We next hypothesized that if these enzymes evolved from a thioester-driven proto-metabolism without nitrogen, then enzymes in these networks should be depleted in enzyme-bound nitrogen-containing coenzymes. We thusly computed the fraction of pre- and post-ammonia reactions dependent on enzymes containing TPP, PLP, heme, biotin, flavin, pterin, and cobalamin (Fig. 3.6C, Methods). We found that the proportion of pre-ammonia reactions associated with these coenzymes were significantly less than the proportion of post-ammonia reactions dependent on these coenzymes (Fig. 3.6D, one-tailed Wilcoxon sign-rank test:  $P < 10^{-24}$ ), which is primarily due to the large number of PLP-dependent reactions added to the network after the inclusion of ammonia.

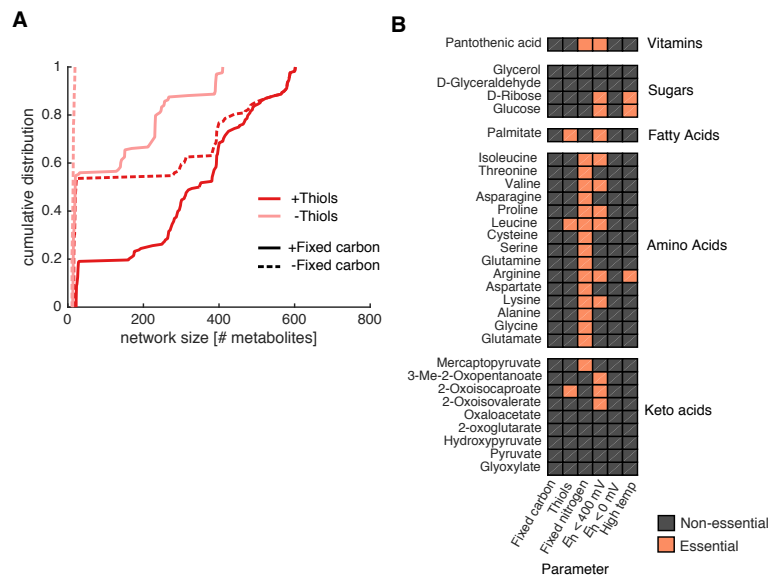
Since only a minority of reactions in this network were categorized as PPE, simple organic or organosulfur catalysts may have been necessary in order for this network to function in prebiotic environments. Christian de Duve suggested that thioester-based polymers may have provided the necessary catalytic components of ancient metabolism in addition to inorganic catalysts [40]. In modern living systems, monomers of keto acids are converted into amino acids, which are then polymerized into polypeptides either with or without the aid of the ribosome and mRNA. If prebiotic environments were severely nitrogen limited, keto acids may have been reduced to hydroxy acids, and polymerized into polyesters using thioesters as a condensing agent. Notably, in such a scenario only the polymer backbone is altered, leaving the side chains (*R*-groups) within today's amino acids intact. Recent work has demonstrated that polyesters may aid in the polymerization of amino acids during dry-wet cycles [53], and that the peptidyl-transferase domain on the ribosome can polymerize hydroxyacylated tRNAs to form polyesters [51, 143], suggesting that ester bond formation may have proceeded amide bond formation in living systems.

It has been proposed that enzymes retain features of early catalysts before the emergence of the genetic code and protein translation systems, and that enzyme active sites may bear resemblance to ancient catalysts. Thus, if this network represents a relic of an ancient metabolism before the biological incorporation of nitrogen, then the active sites of enzymes catalyzing reactions within the network should be depleted in amino acids with side chains containing nitrogen (Fig. 3.6E). To see if the catalytic residues of the enzymes in the pre-ammonia network were depleted in amino acids with nitrogenous side chains, we first obtained a database of catalytic site residues inferred from protein structures [162]. After removing entries with interactions mediated by the peptide backbone, this dataset consisted of 18,721 entries, 1,304 of which were associated with active sites of enzymes in the nitrogen-free network in a representative network. For each putative prebiotic scenario resulting in an expansion with more than 100 metabolites, we computed the fraction of active site residues that contained nitrogen in enzymes associated with both pre- and post-ammonia reactions (Fig. 3.6E). We found that the proportion of nitrogenous catalytic residues associated with pre-ammonia reactions was significantly lower than the proportion of nitrogenous catalytic residues associated with post-ammonia reactions (Fig. 3.6F, Wilcoxon sign-rank test:  $P < 10^{-24}$ ).

One potential alternative explanation for these biases in amino acid composition within the active sites of extant enzymes may be the outcome of evolutionary selection: nitrogen limitation in the environment may have favored mutations that lead to less nitrogen within these enzymes. However, evidence for selection for less nitrogen usage would manifest within the entire protein sequence, rather than just the active sites. Thus, we computed the fraction of amino acids with nitrogenous side chains across the entire coding sequences, rather than specifically the active sites, for enzymes associated with pre- and post-ammonia reactions (see Methods). We found no evidence that enzymes in the pre-ammonia network had a decreased usage of amino acids with nitrogenous side chains relative to enzymes added to the network after ammonia was included in the seed set (one-tailed Wilcoxon sign rank test:  $P = 1$ ), suggesting that the biases within the active sites are not merely a consequence evolutionary selection (see Fig. 3.9).



**Figure 3.6: Enzymes in thioester-driven protometabolism are depleted in nitrogenous compounds** (A) We classified reactions in KEGG as being plausibly pre-enzymatic (PPE) reactions if they could (a) proceed spontaneously, (b) were associated with enzymes that contain at-least one iron-sulfur cluster or (c) were associated with an enzyme that relied on at-least one metal (Ni, Co, Cu, Mg, Mn, Mo, Zn, Fe, W) cofactor. (B) For all scenarios resulting in expansion of  $> 100$  metabolites ( $n = 144$ , Fig. 3.6A) we computed the fraction of PPE-reactions amongst the pre-ammonia reactions (x axis) and post-ammonia reactions (y-axis). The frequency of PPE-reactions in the pre-ammonia reaction set was on average higher than the frequency of PPE-reactions in the post-ammonia reaction set (one-tailed Wilcoxon sign-rank test:  $P < 10^{-19}$ ). (C) We identified KEGG reactions that were dependent on at-least one of the following nitrogen-containing coenzymes: flavin, biotin, thiamine pyrophosphate (TPP) pyridoxal phosphate (PLP), heme, pterin or cobalamin. (D) We compute the fraction of pre- and post- ammonia reactions associated with nitrogen containing coenzymes in the KEGG database, and found that a much higher proportion of post-ammonia reactions were dependent on these coenzymes relative to pre-ammonia reactions (one-tailed Wilcoxon sign-rank test:  $P < 10^{-24}$ ). (E) We parsed the catalytic active site database [162] to find entries associated with pre and post-ammonia reactions, and compute the fraction of entries associated with amino acids with nitrogen-containing side chains (Q,N,W,H,K,R). (F) For each scenario, the fraction of active sites with nitrogen-containing amino acids was significantly higher for post-ammonia reactions relative to pre-ammonia reactions (one-tailed Wilcoxon sign-rank test:  $P < 10^{-24}$ ).



**Figure 3.7: Thiols are required for autotrophic expansion and fatty acid production** (A) We grouped the  $n = 672$  geochemical scenarios into wither a source of fixed carbon and thiols was provided in the seed set. We then plotted the empirical cumulative distributions for each group of scenarios. Notably, when thiols and fixed carbon are not supplied in the seed set, the networks are always below 100 metabolites, indicating that expansion is prohibited without either fixed carbon or thiols in the seed set. (B) We determined what geochemical parameters ( $x$ -axis) were essential for the production of important biomolecules ( $y$ -axis). For example, palmitate, a long chain fatty acid, is producible only if thiols and reductant below 400 mV is provided in the seed set.

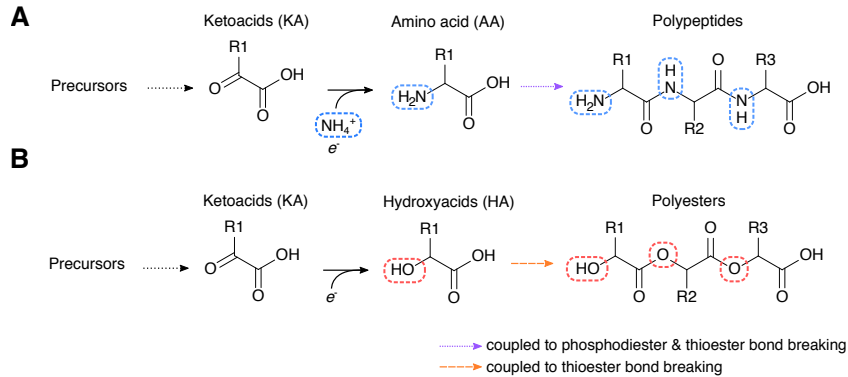


Figure 3.8: **Putative ancient catalysts.** (A) In extant biochemistry, keto acid are converted to amino acids using transamination or reductive amination reaction mechanisms, which are then polymerized using a phosphate or thioester coupled mechanism to make polypeptides. (B) If prebiotic environments did not have a source of fixed nitrogen, then keto acids could have been reduced to  $\alpha$ -hydroxy acids, which could then be polymerized into polyesters either with [40] or without [53] thioester bond breaking.

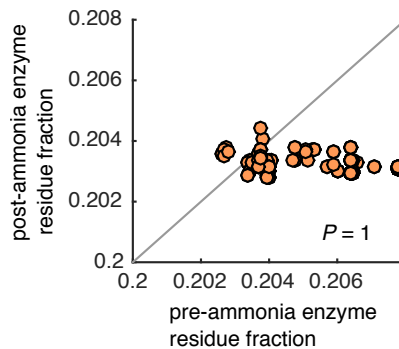


Figure 3.9: **Enzymes catalyzing reactions before the addition of ammonia are not depleted in nitrogen containing amino acids relative to enzymes added after ammonia** To see if the amino acid biases in active sites of enzymes catalyzing reactions added to the network without ammonia (see Fig. 3.6E-F) is confounded due to evolutionary selection for reduced nitrogen in these enzymes, we computed the fraction of nitrogen side chains in enzymes in pre-ammonia reactions ( $x$ -axis) and in enzymes in post-ammonia reactions  $y$ -axis. We found that enzymes in the pre-ammonia networks did not have significantly less nitrogen usage compared to enzymes in post-ammonia reactions (one-tailed Wilcoxon sign-rank test:  $P = 1$ ).

## Chapter 4

# Emergence of community-level function in microbial community assembly

### Summary

This thesis chapter is in the following preprint:

**Goldford J.E.\***, Lu, N.\*, Bajic, D., Estrela, S., Tikhonov M., Gorostiaga, A., Segrè, D., Mehta, P., & Sanchez, A. *Emergent simplicity in microbial community assembly*. (preprint on [bioRxiv](#) [63]; *Science*, in press) (\**Equal contributions*)

The majority of experiments were performed for this paper by Nanxi Lu, Djordje Bajic and Sylvie Estrela. My contributions were the design of experiment, pilot ecological experiments, bioinformatic analysis, development of theory, and numerical implementations of models. I performed all data analysis and made all figures presented in this thesis, and wrote the paper with Alvaro Sanchez.

### Abstract

Microbes assemble into complex, dynamic, and species-rich communities that play critical roles in human health and in the environment. The complexity of natural environments and the large number of niches present in most habitats are often invoked to explain the maintenance of microbial diversity in the presence of competitive exclusion. Here we show that soil and plant-associated microbiota, cultivated *ex situ* in minimal synthetic environments with a single supplied source of carbon, universally re-assemble into large and dynamically stable communities with strikingly

predictable coarse-grained taxonomic and functional compositions. We find that generic, non-specific metabolic cross-feeding leads to the assembly of dense facilitation networks that enable the coexistence of multiple competitors for the supplied carbon source. The inclusion of universal and non-specific cross-feeding in ecological consumer-resource models is sufficient to explain our observations, and predicts a simple determinism in community structure, a property reflected in our experiments.

## **Introduction**

Microbial communities play critical roles in a wide range of natural processes, from animal development and host health to biogeochemical cycles [145, 183, 85]. Recent advances in DNA sequencing have allowed us to map the composition of these communities with an unprecedented high resolution. This has motivated a surge of interest in understanding the ecological mechanisms that govern microbial community assembly and function [34]. A quantitative, predictive understanding of microbiome ecology is required in order to design effective strategies to rationally manipulate microbial communities and steer them away from undesirable or unhealthy states towards beneficial ones.

Survey studies of microbiome composition across a wide range of ecological settings, from the oceans to the human body [183, 85], have revealed intriguing empirical patterns in microbiome organization. These widely observed properties include: high microbial diversity; the coexistence of multiple closely related species within the same functional group; functional stability despite large species turnover; and different degrees of determinism in the association between nutrient availability and taxonomic composition at different phylogenetic levels [85, 187, 112, 111, 125, 25, 39]. These observations have led to the proposed existence of common organizational principles in microbial community assembly [112, 111]. However, the lack of a theory of microbiome assembly is hindering progress towards explaining and interpreting these empirical findings, and it remains unknown which of the functional and structural features exhibited by microbiomes reflect specific local adaptations at the host or microbiome level [39], and which are generic properties of complex,

self-assembled microbial communities.

Efforts to connect theory and experiments for understanding microbiome assembly have typically relied on manipulative bottom-up experiments with a small number of species [56, 190, 81]. While this highly controlled approach is useful to reveal insights into specific mechanisms of interactions, it is unclear to what extent findings from these studies scale up to predict the generic properties of large microbial communities or even the interactions therein. Of note is the ongoing debate about the relative contributions of competition and facilitation [54, 35] and the poorly understood role that high-order interactions play in microbial community assembly [56, 109, 7]. To move beyond empirical observations and connect statistical patterns of microbiome assembly with ecological theory, it is necessary to study the assembly of large numbers of large multi-species microbiomes in simple environments need to be studied, under highly controlled conditions that allow proper comparison between theory and experiment.

## **Results**

### **Assembly of large microbial communities on a single limiting resource**

To meet this challenge, we have followed a high-throughput *ex situ* cultivation protocol to monitor the spontaneous assembly of ecologically stable microbial communities derived from natural habitats in well-controlled environments – using synthetic (M9) minimal media containing a single externally-supplied source of carbon, (Methods) as well as single sources of all of the necessary salts and chemical elements required for microbial life (Fig. 4.2A). Intact microbiota suspensions were extracted from diverse natural ecosystems, such as various soils and plant leaf surfaces (Methods). Suspensions of microbiota from these environments were highly diverse and taxonomically rich (Fig. 4.1), ranging between 110 and 1290 exact sequence variants (ESV). We first inoculated 12 of these suspensions of microbiota into fresh minimal media with glucose as the only added carbon source, and allowed the cultures to grow at 30 °C in static broth. We then passaged the mixed cultures in fresh media every 48 hours with a fixed dilution factor of for a total of 12 transfers (around 84 generations). At the end of each growth cycle, we assayed the community composition

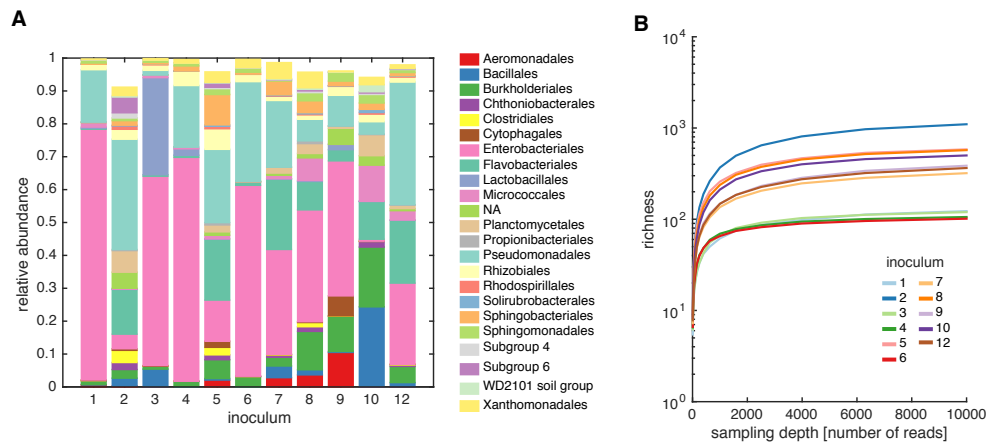


Figure 4.1: **Characterization and diversity of microbiomes isolated from plant and soil samples.** (A) 16S sequencing results for 11/12 initial inocula (labeled 1-10, 12 on the  $x$ -axis). Stacked bar-plots show the community composition at the Order taxonomic level. (B) Rarefaction curves for each inoculum community; the average of 100 random samples of a fixed sampling size ( $x$ -axis) was plotted against the number of unique exact sequence variants (ESV) ( $y$ -axis). The number of unique 16S sequences spanned an order of magnitude, ranging from 110-1290 exact sequence variants. Note that we were unable to generate amplicon libraries for inoculum 11

using 16S rRNA amplicon sequencing (Fig.4.2A, Methods). High resolution sequence denoising allowed us to identify ESVs, which revealed community structure at single nucleotide resolution [26].

Most communities stabilized after 60 generations, reaching stable population equilibria in nearly all cases (Fig. 4.2B, 4.3). For all of the 12 initial ecosystems, we observed large multi-species communities after stabilization that ranged from 4 to 17 ESVs at a sequencing depth of 10,000 reads; further analysis indicates that this is a conservative estimate of the total richness in our communities (Fig. 4.4 - 4.5, see Methods). We confirmed the taxonomic assignments generated from amplicon sequencing by culture-dependent methods, including the isolation and phenotypic characterization of all dominant genera within a representative community (Fig. 4.6).

### Convergence of bacterial community structure at the family taxonomic level

High-throughput isolation and stabilization of microbial consortia allowed us to explore the rules governing the assembly of bacterial communities in well-controlled synthetic environments. At

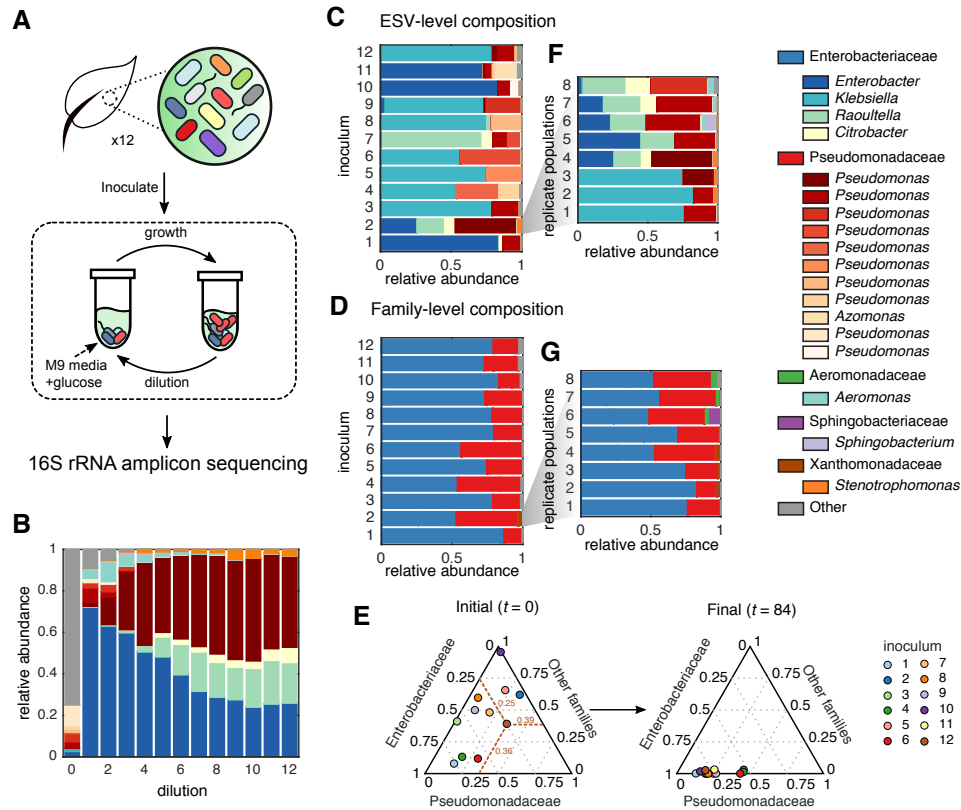


Figure 4.2: **Top down assembly of bacterial consortia.** (A) Experimental scheme: large ensembles of taxa were obtained from 12 leaf and soil samples, and used as inocula in passaged-batch cultures containing synthetic media supplemented with glucose as the sole carbon source. After each transfer, 16S rRNA amplicon sequencing was used to assay bacterial community structure. (B) The community structure of a representative community (from inoculum 2) after every dilution cycle (about 7 generations), revealing a 5-member consortia from the *Enterobacter*, *Raoultella*, *Citrobacter*, *Pseudomonas* and *Stenotrophomonas* genera. The community composition after 84 generations is shown at the exact sequence variant (ESV) level (C) or the family taxonomic level, converging to characteristic fractions of Enterobacteriaceae and Pseudomonadaceae (D). (E) Simplex representation of family-level taxonomy before ( $t = 0$ ) and after ( $t = 84$ ) passaging experiment. (F-G) Experiments were repeated with 8 replicates from a single source (inocula 2), and communities converged to very similar family level distributions (G), but displayed characteristic variability at the genus and species level (F).

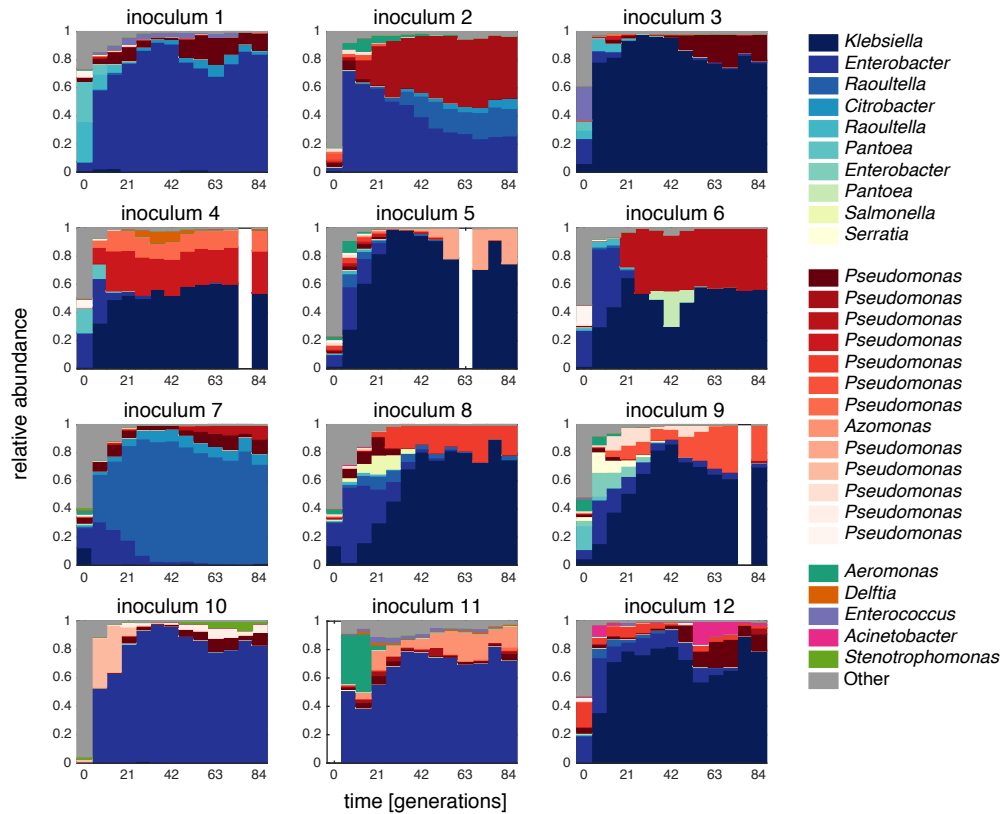
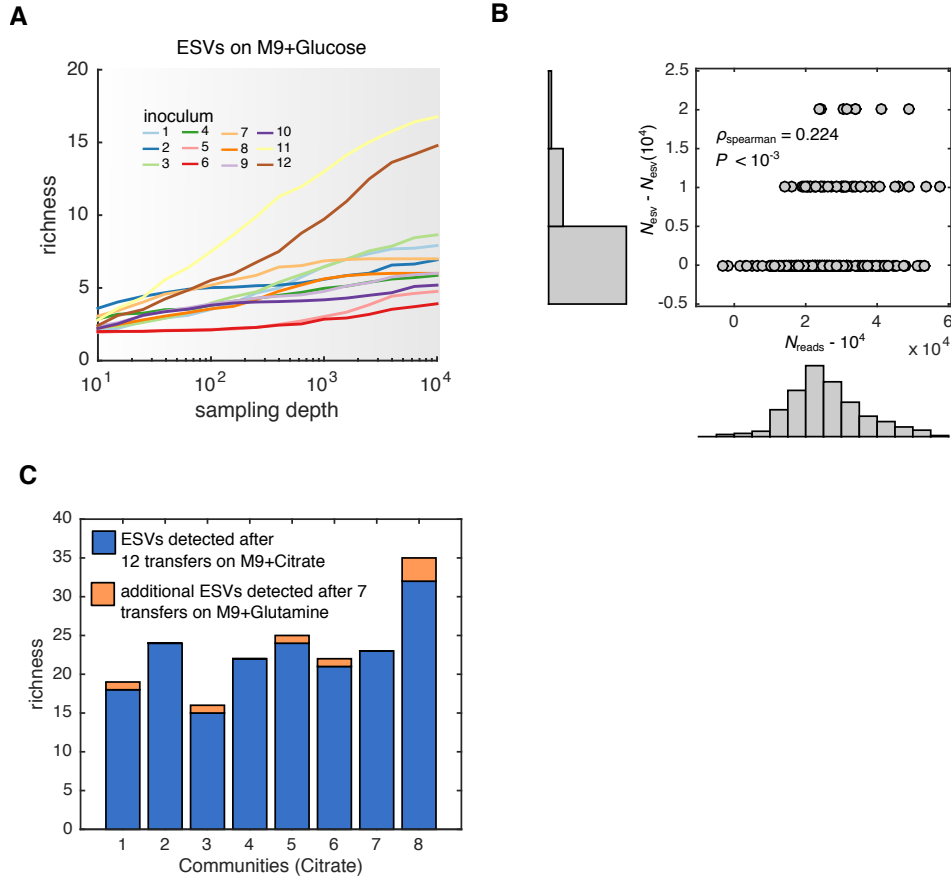
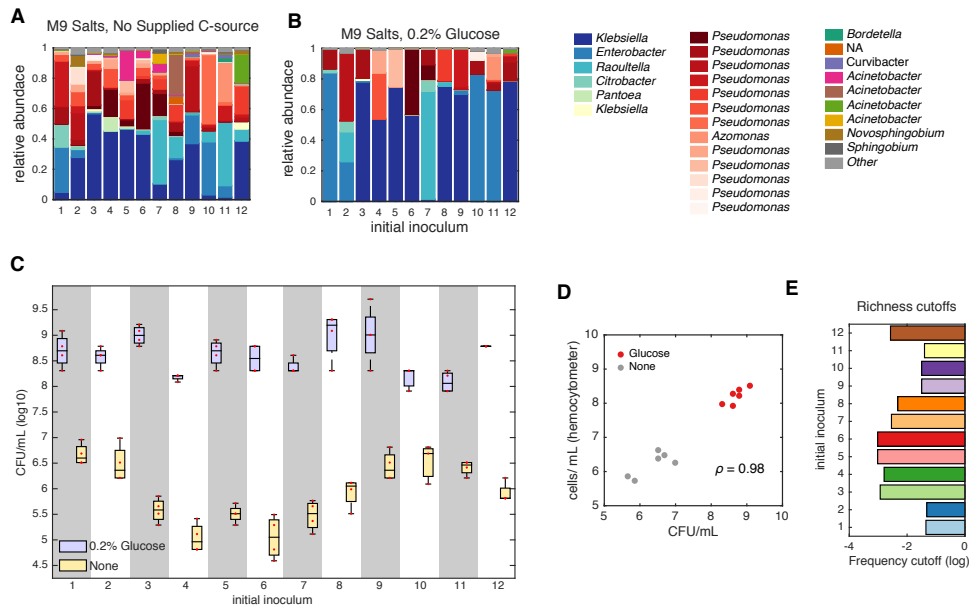


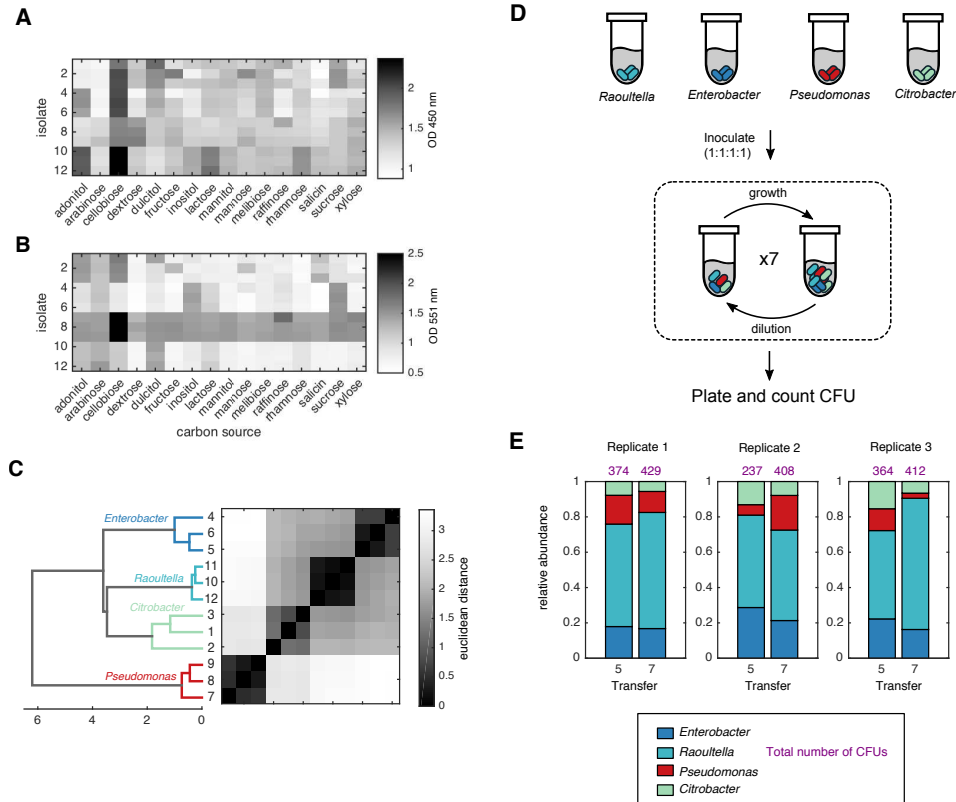
Figure 4.3: **Dynamics of ex-situ community composition over 84 generations in glucose-supplemented media.** Communities were transferred into fresh media every 48 hours, allowing approximately seven growth generations per transfer. After each transfer, we determined the community composition using 16S rRNA amplicon sequencing (see methods). The relative abundance of each taxon was plotted as a function of time (generations). All inocula appear to reach stable community structures by the 60th generation



**Figure 4.4: Presence of sparse rare taxa in ex situ assembled microbial communities.** (A) Rarefaction curves were produced by subsampling a fixed number of reads and computing the number of unique exact sequence variants (ESVs). The plot shows the average over 100 samples at each fixed sampling depth ( $x$ -axis) for each of the 12 inocula. (B) For each stabilized community, we aimed to estimate the prevalence of rare taxa on our stabilized communities by measuring the number of additional ESVs detected at sampling depths above 10,000 reads. We plotted the number additional reads above 10,000 ( $x$ -axis) vs the number of additional ESVs detected at sampling depths above 10,000 reads ( $y$ -axis). Although there appears to be a positive correlation between additional sampling depth and additional reads, at-most 2 additional ESVs were detected at sampling depths of nearly 60,000 reads. (C) To further quantify the presence of rare taxa in our samples, we took eight communities stabilized on M9+citrate and passaged them on M9+glutamine for an additional 7 transfers, and sequenced at an average depth of 25,000 reads. The number of ESVs detected in the communities passaged on M9+citrate are plotted as blue bars, and the additional ESVs detected in the communities passaged on M9+glutamine are plotted as orange bars, where between 0-3 additional ESVs were detected when passaged on glutamine.



**Figure 4.5: Low levels of bacterial growth with no externally supplied carbon source.** (A) We repeated passaging experiments without an externally supplied carbon source, and observed that widespread and diverse communities survive over the course of 84 generations. Communities were similar in structure to glucose supplemented communities (B), but with higher diversity. (C) To determine the richness of communities surviving primarily on the externally supplied resource, we plated the communities after 84 generations and counted colony forming units (CFU). We plotted the CFU/mL in replicates of four for each inoculum passaged either on M9 with 0.2% glucose (blue) or on M9 with no supplemented carbon source (yellow). In all cases, population sizes were orders of magnitude lower when no carbon source was provided compared to population sizes of communities grown on glucose. (D) hemocytometer cell counting was performed to verify that CFU accurately recapitulated cell densities. For 12 samples, hemocytometer counting was performed and compared to CFU counts, exhibiting strong positive correlation (Pearson's correlation,  $\rho = 0.98$ ). (E) Measurements of absolute population sizes allowed us to define relative abundance cutoffs for communities grown on glucose, ensuring that growth of taxa above the relative abundance cutoff was primarily a consequence of the externally supplied glucose.



**Figure 4.6: Four strains from a representative community coexist in reconstituted communities.** 12 isolates were picked from a representative community from inoculum 2 with 4 distinct morphologies. (A-B) Isolates were grown in phenol red broth with the addition of one of 16 carbon sources. Optical density (OD) was measured at 450 nm and 551 nm after 19 hours to track the degree of acidification from fermentation. (C) The O.D. profiles were hierarchically clustered, revealing 4 clusters of isolates with distinct fermentation profiles, corroborating morphology and sequencing results. These results indicate that the 12 isolates belong to one of four taxa, (D) To see if these four taxa could coexist without the presence of other community members, we inoculated M9+0.2% glucose with equal proportions of each taxa, passaged them for seven dilution cycles and plated the final populations. We counted the colony forming units (CFUs) and distinguished each taxa based on morphology. (E) The relative abundance of three replicates at transfers show that all four taxa coexist after seven transfers.

the species (ESV) level of taxonomic resolution, the 12 natural communities assembled into highly variable compositions (Fig. 4.2C). However, when we grouped ESVs by higher taxonomic ranks we found that all 12 stabilized communities, with very diverse environmental origins, converged into similar family-level community structures dominated by Enterobacteriaceae and Pseudomonadaceae (Fig. 4.2D). In other words, a similar family-level composition arose in all communities despite their very different starting points. This is further illustrated in Fig.4.7, where we show that the temporal variability (quantified by the  $\beta$ -diversity) in family level composition is comparable with the variability across independent replicates. The same is not true when we compare taxonomic structure at the sub-family level (i.e. genus).

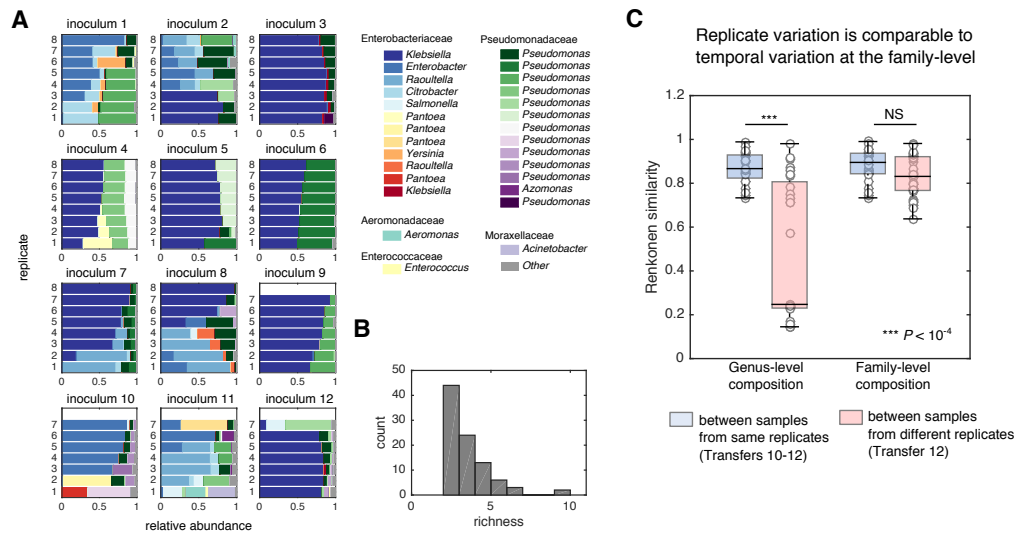
To better understand the origin of the taxonomic variability observed below the family-level, eight replicate communities were started from each one of the 12 starting microbiome suspensions (inocula), and propagated in minimal media with glucose as in the previous experiment. Given that the replicate communities were assembled in identical habitats and were inoculated from the same pool of species, any observed variability in community composition across replicates would suggest that random colonization from the regional pool and microbe-microbe interactions are sufficient to generate alternative species-level community assembly. Indeed, for most of the inocula (nine out of twelve), replicate communities assembled into alternative stable ESV-level compositions, while still converging to the same family-level attractor described in Fig. 4.2E (see also Fig. 4.7). One representative example is shown in Fig. 4.2F-G; all eight replicates from the same starting inoculum assemble into strongly similar family-level structures, which are quantitatively consistent with those found before (Fig. 4.2D). However, different replicates contain alternative Pseudomonadaceae ESVs, and the Enterobacteriaceae fraction is constituted by either an ESV from the *Klebsiella* genus, or a guild consisting of variable subcompositions of *Enterobacter*, *Raoultella*, and/or *Citrobacter* as the dominant taxa. The remaining (three out of twelve) inocula converge to similar community compositions and all replicates exhibit strongly similar population dynamics and population structures at all levels of taxonomic resolution (Fig.4.8). The reproducibility in population dynamics between replicate communities indicates that experimental error is not the main source of variability in community composition. Indeed, the population bottlenecks intro-

duced by the serial dilutions into fresh media have only a modest effect on the observed variability in population dynamics (Fig.4.9).

Despite the observed species level variation in community structure, the existence of family-level attractors suggests the existence of fundamental rules governing community assembly. Recent work on natural communities has consistently found that environmental filtering selects for convergent function across similar habitats, while at the same time allowing for taxonomic variability within each functional class [111, 187, 113]. In our assembled communities in glucose media, fixed proportions of Enterobacteriaceae and Pseudomonadaceae may have emerged due to a competitive advantage, given the well-known glucose uptake capabilities of the phosphotransferase system in Enterobacteriaceae and ABC transporters in Pseudomonadaceae [67]. This suggests that the observed family-level attractor may change if we assemble communities adding a different carbon source to our synthetic media.

To determine the effect of the externally provided carbon source on environmental filtering, we repeated the community assembly experiments with eight replicates of all 12 natural communities, using two alternative single-carbon sources, citrate or leucine, instead of glucose. Consistent with previous experiments on glucose minimal media, communities assembled in citrate or leucine contained large numbers of species: communities stabilized on leucine contained 6-22 ESVs, while communities stabilized on citrate contained 4-22 ESVs at a sequencing depth of 10,000 reads. As was the case for glucose, replicate communities assembled on citrate and leucine also differed widely in their ESV-level compositions, while converging to carbon source-specific family-level attractors (Fig. 4.10A, 4.11, and 4.12).

Family-level community similarity (Renkonen similarity) was on average higher between communities passaged on the same carbon source (median: 0.88) than between communities passaged from the same environmental sample (median 0.77, one-tailed Kolmogorov-Smirnov test,  $P < 10^{-5}$ ; Fig. S11). Communities stabilized in citrate media were composed of a significantly lower fraction of Enterobacteriaceae (MannWhitney U test,  $P < 10^{-5}$ ), and displayed an enrichment of Flavobacteriaceae relative to communities grown on glucose (MannWhitney U test,  $P < 10^{-5}$ ), while communities stabilized in leucine media had no growth of Enterobacteriaceae



**Figure 4.7: The community structure from the same inocula can be highly variable and the genus level, but similar at the family level.** Passing experiments of microbial communities on M9 + 0.2 % glucose were repeated with up to 8 replicates per inoculum. (A) Each subplot is the relative abundance of the exact sequence variants (ESVs) for all replicates originating from the same inoculum. Note that for each inoculum, fixed points range from multiple (e.g inoculum 2) to a single attractor (e.g. inoculum 6). (B) The distribution of richness (see Fig.4.5) estimates across all communities formed in (A) showed that all large-scale competitive experiments retained at-least 2 sequence variants, and the majority (48/92) retained more than four sequence variants. (C) To characterize the variability of community structure across different starting replicates at various levels of taxonomic resolution, we computed the Renkonen similarity (at both genus and family-levels) between replicate communities from inocula 2 after 12 transfers. As a comparison, we computed the Renkonen similarity between samples obtained at the end of the last three transfers (transfer 10-12) within the same replicate. The boxplots are distributions of Renkonen similarities between both within replicates (blue) and between replicates (red) at the genus (left) and family (right) taxonomic levels. Communities are significantly less similar at the genus level when comparing between replicates vs. within replicates (Mann-Whitney U-test:  $P < 10^{-4}$ ), while communities are of comparable similarity at the family level when comparing samples from different replicates vs. samples from the same replicate (Mann-Whitney U-test:  $P = 0.06$ )

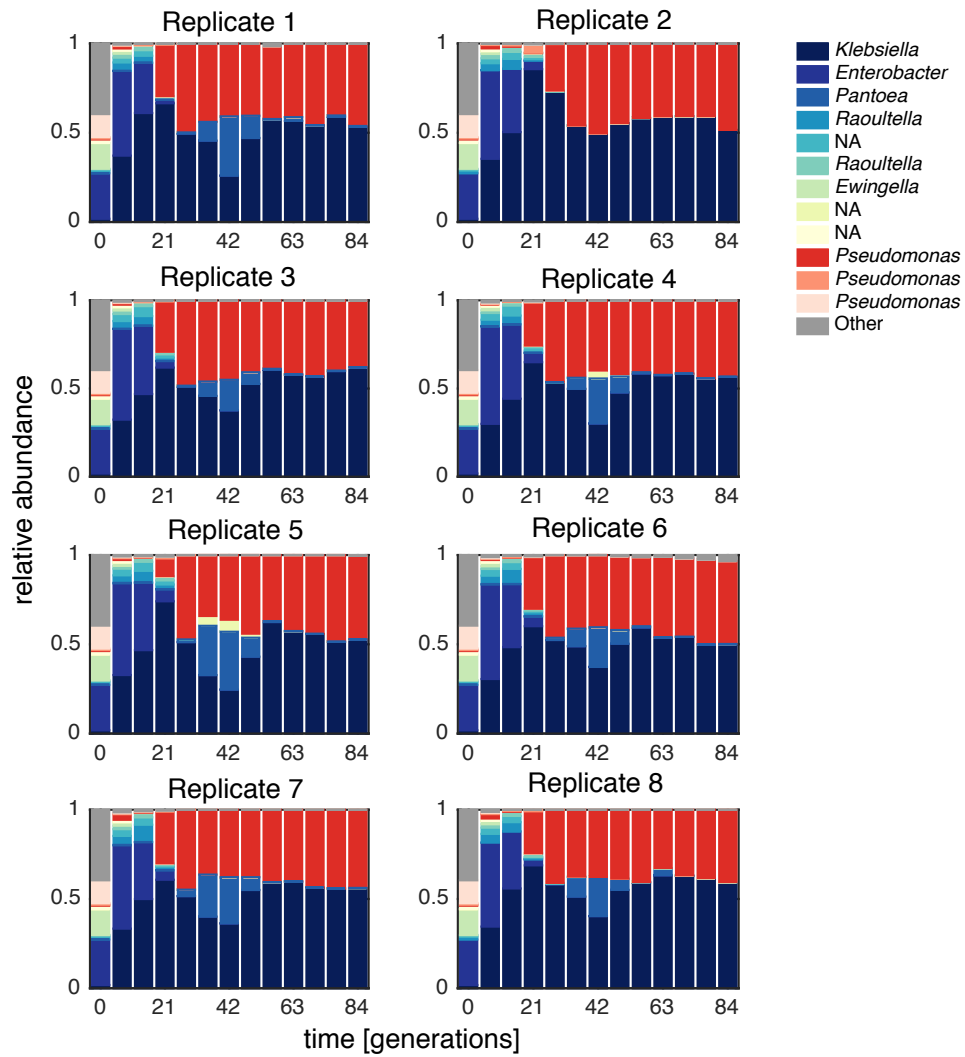


Figure 4.8: **Inoculum 6 exhibits strongly deterministic population dynamics.** We performed replicate passing experiments starting with inoculum 6 and found nearly reproducible population dynamics. Each subplot shows the relative abundance of sequence variants (y-axis) during the course of the passing experiment (x-axis). Notably, in 7/8 replicates, a bloom of a *Pantoea* sequence variant occurred at the 42nd generation.

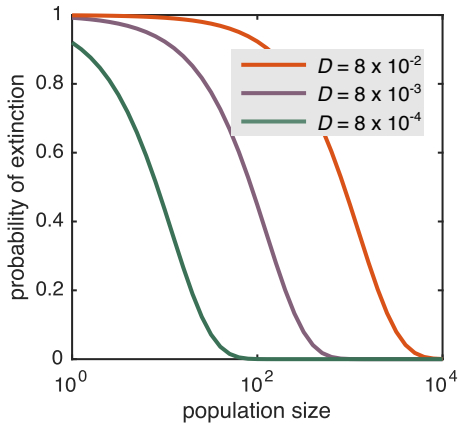


Figure 4.9: **Day-to-day sampling likely does not induce substantial variation in community structure.** We calculate the probability of extinction by stochastic sampling as a function of the size of the population for a given species, for the dilution factor we apply in our experiments ( $D = 0.008$ ; purple line) as well as for 10-fold larger (red) and 10-fold smaller (green) dilution factors. We note that all of the ESVs that we detect in our community 16S sequencing have population sizes of at least 10,000.

and an enrichment for Comamonadaceae relative to communities grown on glucose (MannWhitney U test,  $P < 10^{-5}$ ) or citrate (MannWhitney U test,  $P < 10^{-5}$ ).

These results suggest that the supplied source of carbon governs community assembly. To quantify this effect, we used a machine learning approach and trained a support vector machine (SVM) to predict the identity of the supplied carbon source from the family-level community composition. We obtained a cross-validation accuracy of 97.3% (Fig. 4.10B; Methods). Importantly, we found that considering the tails of the family-level distribution (as opposed to just the two dominant taxa) increases the predictive accuracy (Fig. 4.10B), which indicates that carbon source-mediated determinism in community assembly extends to the entire family-level distribution, including the more rarefied members.

Rather than selecting for the most fit single species, our environments select complex communities that contain fixed fractions of multiple coexisting families whose identity is determined by the carbon source in a strong and predictable manner (Fig. 4.14). We hypothesized that taxonomic convergence might reflect selection by functions that are conserved at the family level. Consistent with this idea, we find that the imputed community metagenomes assembled on each type of carbon

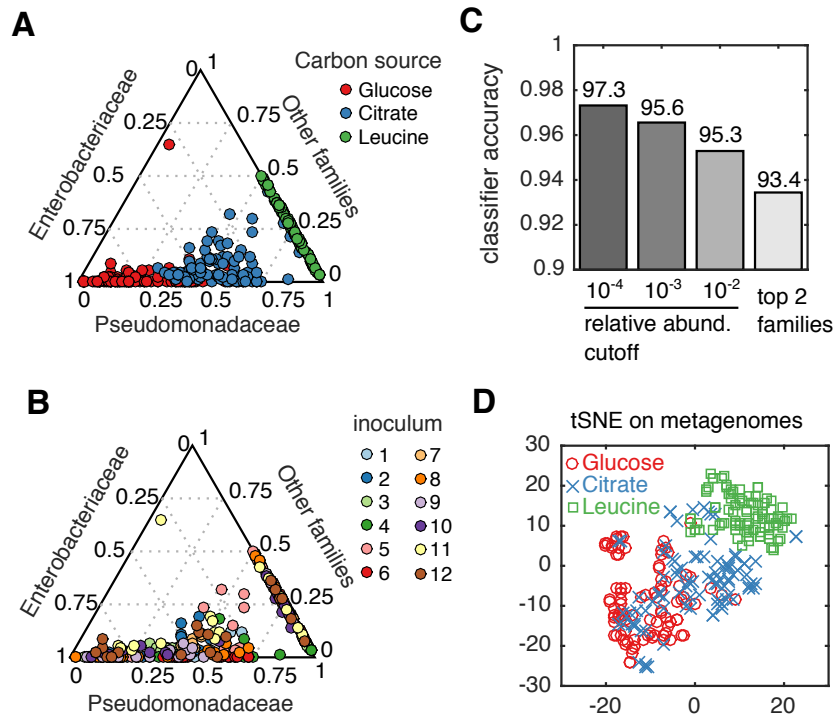


Figure 4.10: **Family-level and metagenomic attractors are associated with different carbon sources.** (A) Family-level community compositions are shown for all replicates across 12 inocula grown on either glucose, citrate or leucine as the limiting carbon source. (b) A support vector machine (see Methods) was trained to classify the carbon source from the family-level community structure. Low abundant taxa were filtered using a predefined cutoff (x-axis) before training and performing 10-fold cross validation (averaged 10 times). Classification accuracy with only Enterobacteriaceae and Pseudomonadaceae resulted in a model with 93% accuracy (right bar), while retaining low abundant taxa (relative abundance cutoff of  $10^{-4}$ ) yielded a classification accuracy of 97% (left most bar). (C) Metagenomes were inferred using PICRUSt [104], and dimensionally reduced using tSNE, revealing that carbon sources are strongly associated with the predicted functional capacity of each community

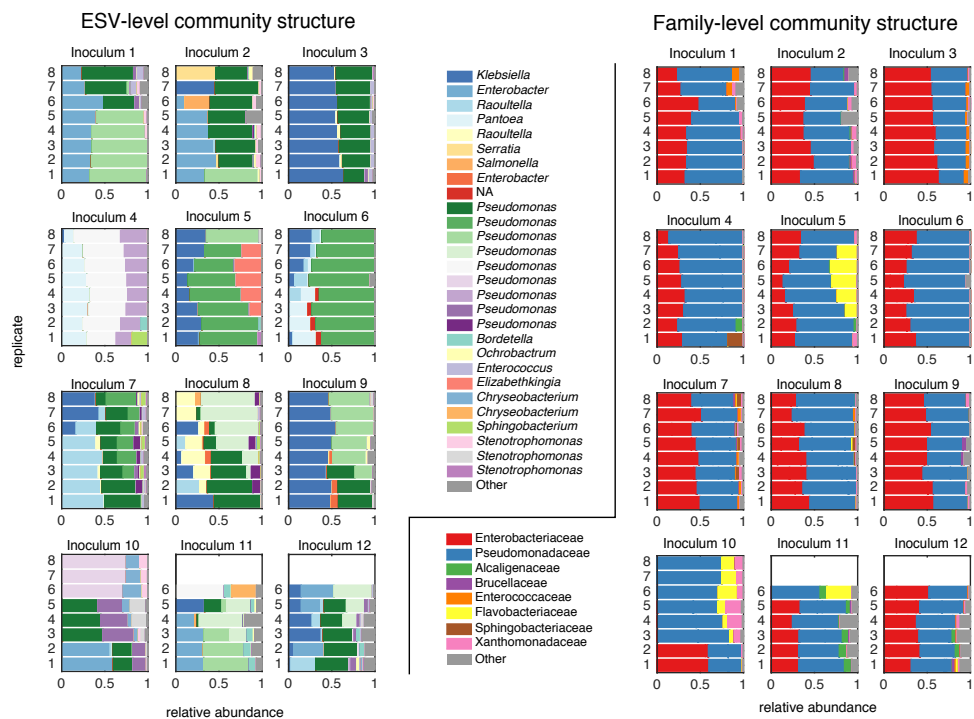


Figure 4.11: **Community structure at ESV and family level on citrate.** Passing experiments of microbial communities on M9 + 0.07 C-mole/L citrate were performed with up to 8 replicates per inoculum, as in the case with glucose.



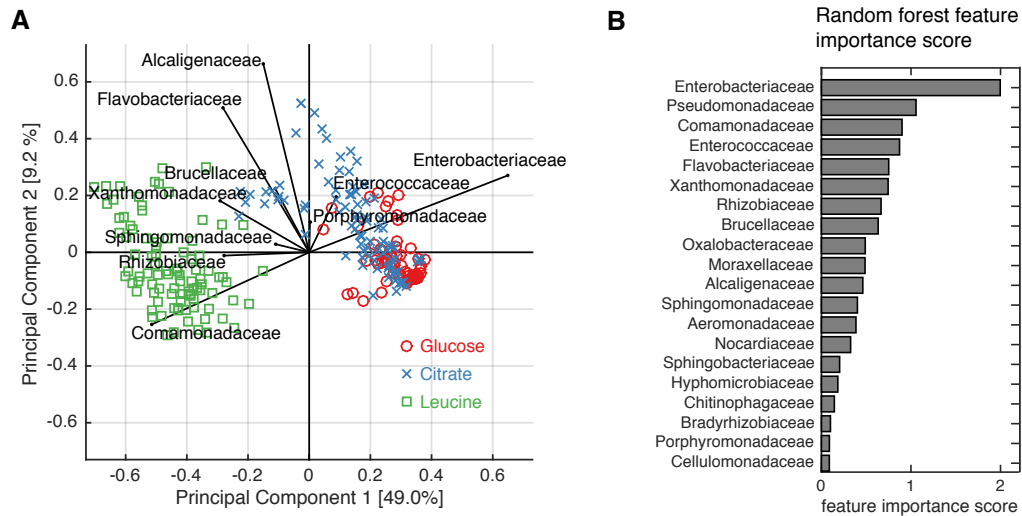
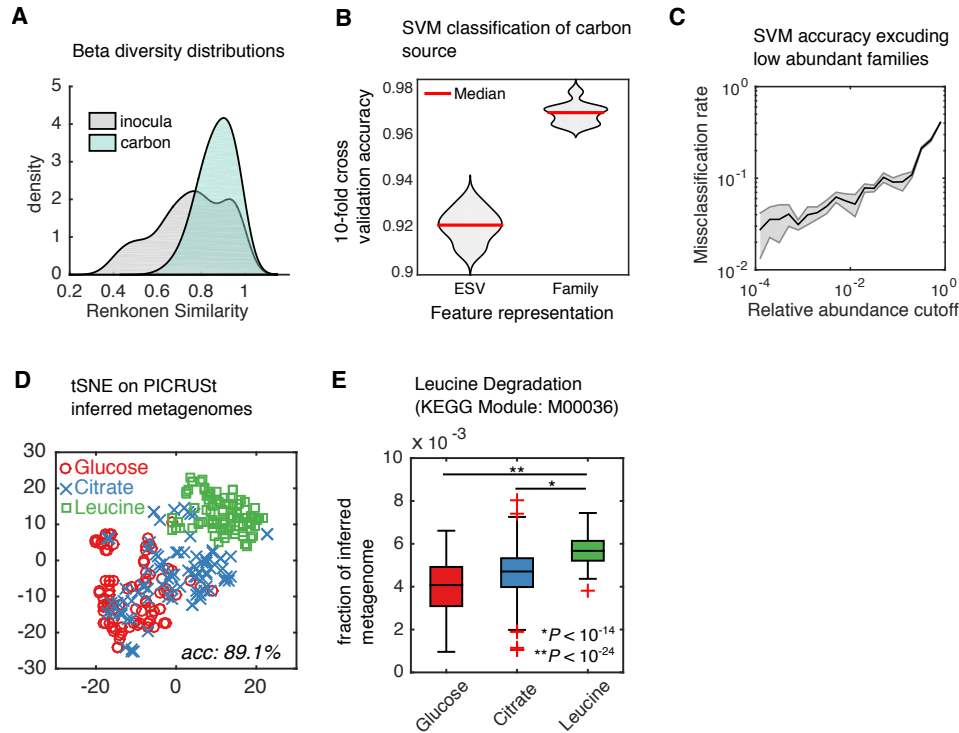


Figure 4.13: **Family-level features associated the carbon source.** (A) The family-level community composition was log-transformed and dimensionally reduced using principal component analysis. Like in Fig 4.10A, family-level community structure was strongly associated with the carbon source in the media. A biplot was used to show which taxa were correlated with the first two principal components. (B) A random forest classifier was trained to predict carbon source from the family-level community structure, and out-of-bag feature importance scores are reported, confirming that the abundance of Enterobacteriaceae and Pseudomonadaceae are important predictors of carbon source.

source exhibit substantial clustering by the supplied carbon source (Fig. 4.10C), and are enriched in pathways for its metabolism (Fig. 4.14). When we spread the stabilized communities on agarose plates, we routinely found multiple identifiable colony morphologies per plate, evidencing that multiple taxa within each community are able to grow independently on (and thus compete for) the single supplied carbon source. This suggests that the genes and pathways that confer each community with the ability to metabolize the single supplied resource are distributed among multiple taxa in the community, rather than being present only in the best competitor species.

### Widespread metabolic facilitation stabilizes competition and promotes coexistence

Classic consumer-resource models indicate that when multiple species compete for a single externally supplied growth-limiting resource, the only possible outcome is competitive exclusion unless specific circumstances apply [115, 182, 186, 164, 98, 68]. However, this situation does not ad-



**Figure 4.14: Family-level composition is a strong taxonomic predictor of the externally-supplied carbon source.** (A) The distributions of Renkonen similarities between family-level compositions between samples either grown on the same carbon source (light blue,  $N = 12558$ ) or between samples from the same inocula (grey  $N = 3056$ ) are plotted, revealing that the communities grown on the same carbon source are more similar than communities grown from the same inocula (one-tailed Kolmogorov-Smirnov test;  $P < 10^{-5}$ ). (B) A support vector machine (SVM) classifier was used to train a model to predict the carbon source (glucose, citrate or leucine) from the clr-transformed community structure at the ESV or family level. Models were trained using different coarse-graining descriptions of community structure based on taxonomy ( $x$ -axis) and the 10-fold cross-validation accuracy (repeated 10 times) for each model is reported on the  $y$ -axis. (C) An SVM was retrained using families above a pre-defined threshold ( $x$ -axis), and the misclassification rate (1-accuracy) is reported on the  $y$ -axis, revealing that low-abundant families aid in model performance. (D) Metagenome compositions were imputed using PICRUSt [104] and embedded in a two-dimensional space using t-distributed stochastic neighbor embedding (tSNE). (E) The summed abundance of genes belonging to the leucine degradation KEGG module (M00036) are plotted for all samples using a boxplot, where samples are grouped by the limiting carbon source ( $x$ -axis). Leucine degradation genes are enriched in communities grown on leucine relative to communities grown on citrate (Mann Whitney U-test:  $P < 10^{-14}$ ) or glucose (Mann Whitney U-test:  $P < 10^{-24}$ ).

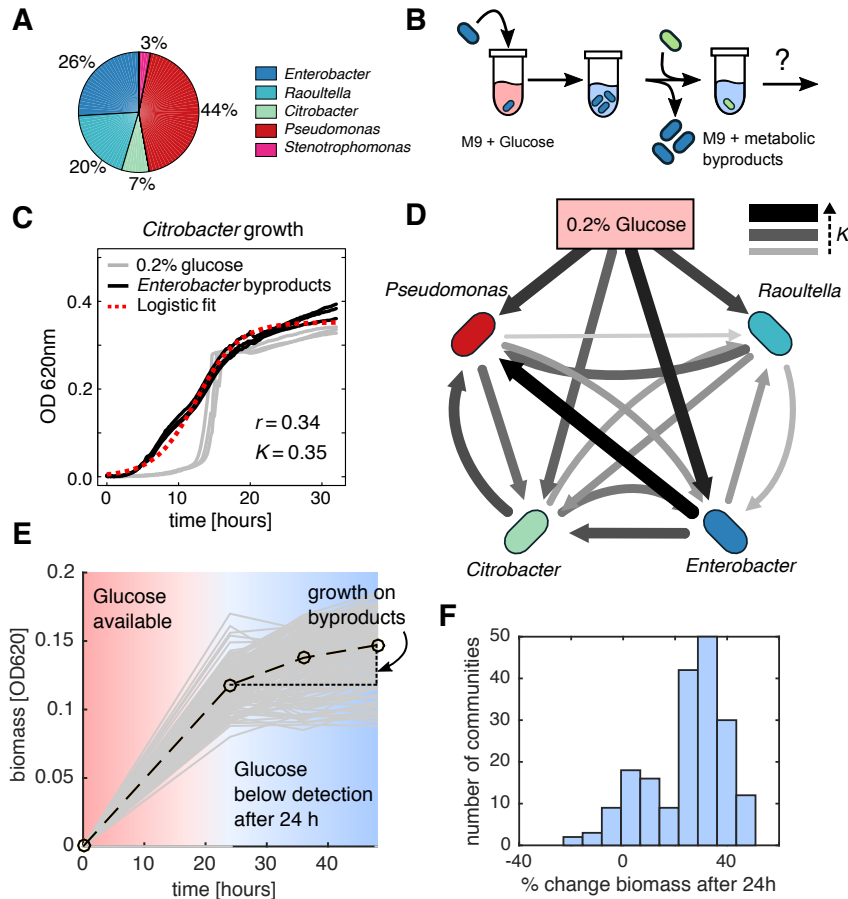
equately reflect microbes, whose ability to engineer their own environments is well documented both in the lab [15, 146, 165, 74] and in nature [12, 37]. Thus, we hypothesized that the observed coexistence of competitor species in our experiments may be attributed to the generic tendency of microbes to secrete metabolic byproducts into the environment, which could then be used by other community members.

To determine the plausibility of niche creation mediated by metabolic byproducts, we analyzed one representative glucose community in more depth. We isolated members of the four most abundant genera in this community (*Pseudomonas*, *Rauoultella*, *Citrobacter* and *Enterobacter*), which together represented 97% of the total population in that community (Fig. 4.15A). These isolates had different colony morphologies and were also phenotypically distinct (Fig. 4.6). All isolates were able to form colonies in glucose agarose plates and all grew independently in glucose as the only carbon source, which indicates that each isolate can compete for the single supplied resource. All four species were able to stably coexist with one another when the community was reconstituted from the bottom up by mixing the isolates together from isolates (Fig. 4.6). To test the potential for cross-feeding interactions in this community, we grew monocultures of the four isolates for 48 hours in synthetic M9 media containing glucose as the only carbon source (Fig. 4.15B). At the end of the growth period the glucose concentration was too low to be detected, indicating that all of the supplied carbon had been consumed and any carbon present in the media originated from metabolic byproducts previously secreted by the cells. To test whether these secretions were enough to support growth of the other species in that community, we filtered the leftover media to remove cells, and added it to fresh M9 media as the only source of carbon (Fig. 4.15B). We found that all isolates were able to grow on every other isolates secretions (e.g. Fig. 4.15C), forming a fully connected facilitation network (Fig. 4.15D). Growth on the secretions of other community members was strong, often including multiple diauxic shifts, and the amount of growth on secretions was comparable to that on glucose, suggesting the pool of secreted byproducts are diverse and abundant in this representative community.

To find out if growth on metabolic byproducts is frequent among our communities, we thawed 95 glucose-stabilized communities (7-8 replicates from 12 initial environmental habitats) and grew

them again on glucose as the only carbon source for an extra 48 hour cycle. In all 95 communities glucose was completely exhausted after 24 hours of growth (Fig. 4.15E); yet, most communities continued growing after glucose had been depleted (Fig. 4.15E), showing that growth on previously secreted byproducts is widespread. Moreover, community growth on the secreted byproducts is strong: on average, communities produce approximately 25% as much biomass on the secretions alone as they did over the first 24 hours when glucose was present (Fig. 4.15F). Both propidium iodide (PI) staining and phase contrast imaging of communities at the single-cell level identified low numbers of permeabilized or obviously lysed cells. This supports the hypothesis that metabolic byproduct secretion (rather than cell lysis) is the dominant source of the observed cross-feeding. However, we note that lytic events that leave no trace behind would not have been detected in our micrographs, so a contribution from cell death to our results cannot be entirely ruled out. Other mechanisms may also operate together with facilitation in specific communities to support high levels of biodiversity [109, 98, 155, 31, 108] In experiments where the environment was well mixed by vigorous shaking, we also found communities containing multiple taxa, indicating that spatial structure is not required for coexistence (Fig. 4.16). In addition we did not observe effects from temporal competitive niches in our experiments (Fig. 4.17).

Recent work has suggested that alteration of the pH by bacterial metabolism may also have important effects on limiting growth [36, 157, 158], and can be a driver of microbial community assembly. Our results suggest that although individual isolates can substantially acidify their environment when grown in glucose as monocultures (e.g. the pH drops to 4.85 in *Citrobacter* and to 5.55 in *Enterobacter* monocultures after 48 hours), our stabilized communities exhibit only modest changes in pH as they grow in glucose minimal media, dropping by less than 1 unit in most communities, and stabilizing to pH 6.5 in all cases after 48 hours of growth. In other carbon sources, such as leucine, the pH is even more stable than in glucose. Altogether, our results suggest that acidification by fermentation may be "buffered" by the community relative to the effect seen by monocultures. Although beyond the scope of this work, efforts to elucidate the roles of other mechanisms that may stabilize competition, like phage predation [164] or non-transitive competition networks [109], will more fully characterize the landscape of interactions in these microcosms.



**Figure 4.15: Non-specific metabolic facilitation stabilizes competition for the supplied resource.** (A) The major taxa in a representative community from inoculum 2 were isolated, grown under conditions with minimal media (M9) and glucose, and the metabolic byproducts were used as the sole carbon source in growth media for other isolates. (B) Experimental set-up: isolates were grown in minimal media with glucose for 48 hours, and cells were filtered out from the suspension. The suspension of byproducts was mixed 1:1 with 2X M9 media and used as the growth media for other isolates (see also Methods). (C) An example growth curve for *Citrobacter* growing either with M9 supplemented with 0.2 % glucose (grey line) or the metabolic byproducts from *Enterobacter* (black line). (D) All isolates were grown on every other isolates metabolic byproducts, and logistic models were used to fit growth curves. We plotted the fitted growth parameters (carrying capacity) as edges on a directed graph, where the edges encode the carrying capacity of the target node isolate when grown using the secreted byproducts from the source node isolate. Edges from the top node encodes the carrying capacity on 0.2 % glucose, which is comparable in edge width/color to several of the other interactions. (E-F) All communities stabilized on glucose were grown in glucose-supplemented M9 media, and optical densities at 620 nm were measured, showing that after glucose was depleted ( 24 hours), communities on average grew an additional 25%.

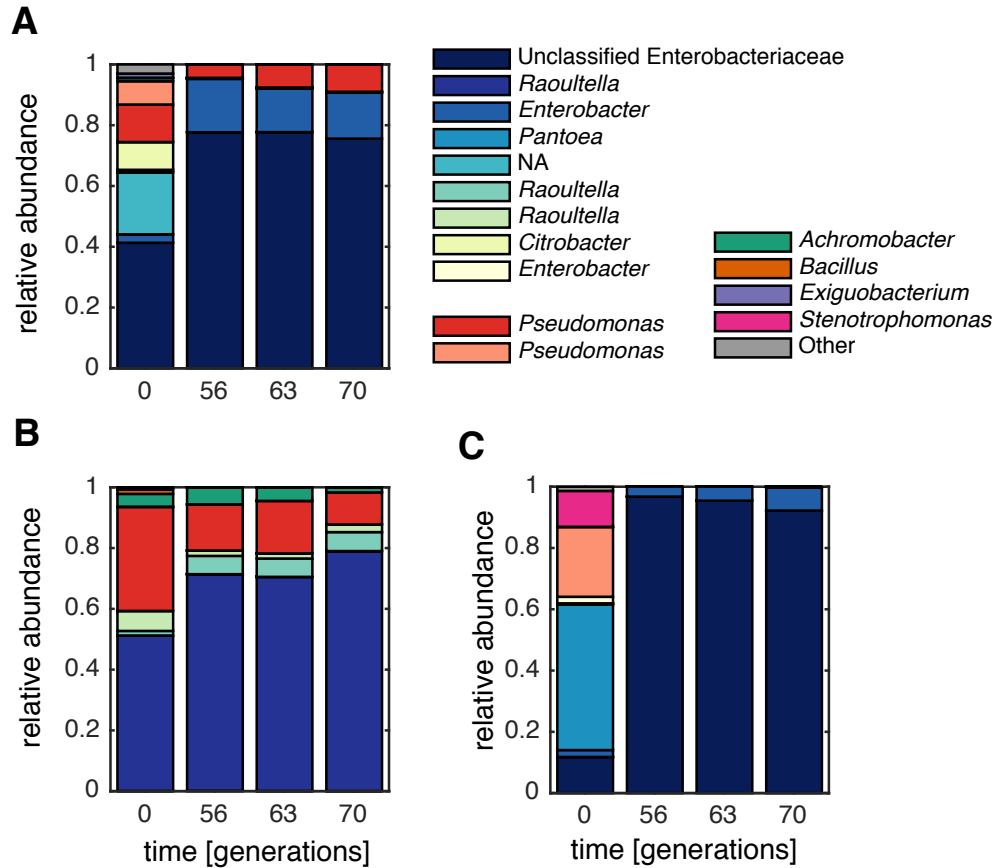


Figure 4.16: **Non-specific metabolic facilitation stabilizes competition for the supplied resource.** Spatial structure in our 96-well plate format could also allow for coexistence of microbial species [31]. Thus, experiments were repeated for three separate inocula passaged on media with M9+0.2% glucose, but while vigorously shaking cultures at 200 RPM. In all cases, no single strain outcompeted all other strains, suggesting that coexistence is stable even without potential spatial heterogeneity.

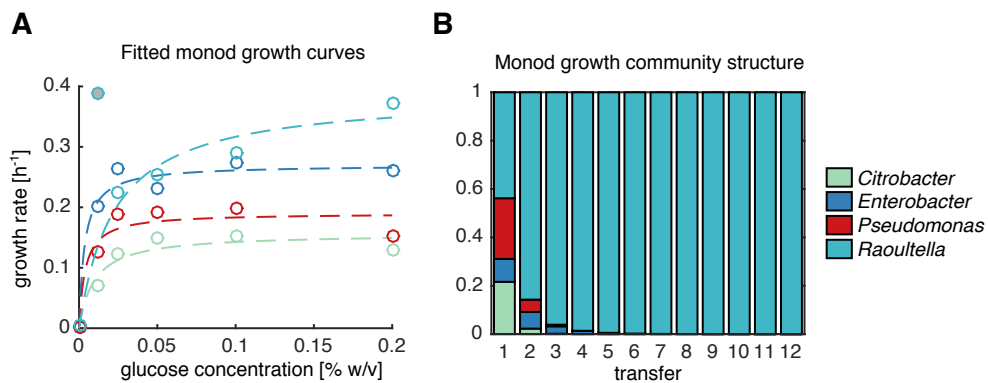


Figure 4.17: **Resource abundance on the growth rates of individual species.** A potential mechanism for coexistence among microbes in an environment with a single limiting nutrient is each species has maximal fitness at least one intermediate level of the limiting nutrient [182]. Thus, isolates from a representative community were grown at various concentrations of glucose (subplot (A),  $x$ -axis), and the initial growth rate was measured (See Monod model section), and fitted to a Monod growth model. *Raoultella* displayed unusually high growth rates at low glucose concentrations. In (A), we removed this outlier (grey dot) at very low resource abundances. We used the Monod parameters to simulate a batch culture passaging experiment (B), and found that *Raoultella* competitively excludes all other species *in silico*. If the outlier observed at low growth rates is retained, *Raoultella* still competitively excludes all other species. Together, these results indicate that there is no supporting evidence of resource abundance-dependent fitness effects that lead to coexistence amongst these strains.

### **A generic consumer resource model recapitulates experimental observations**

Our experiments indicate that competition for a single limiting nutrient may be stabilized by non-specific metabolic facilitation, leading to coexistence. To test whether this feature alone promotes coexistence, we simulated a community assembly process on a single supplied carbon source using a version of the classic MacArthur consumer resource model (CRM) [114], which was modified to include non-specific cross-feeding interactions. Cross-feeding was modeled through a stoichiometric matrix that encodes the proportion of a consumed resource that is secreted back into the environment as a metabolic byproduct (Supporting Information). Setting this matrix to zero results in no byproducts being secreted, and recovers the classic results for the CRM in a minimal environment with one resource: the species with highest consumption rate of the limiting nutrient competitively excludes all others (Fig. 4.18A, inset). However, when we drew the stoichiometric matrix from a uniform distribution (while ensuring energy conservation), and initialized simulations with hundreds of species (each defined by randomly generated rates of uptake of each resource) coexistence was routinely observed (Fig. 4.18A). All of the coexisting species in this simulation were generalists, capable of growing independently on the single supplied resource as well as on each other species secretions.

Our experiments have shown that the family-level community composition is strongly influenced by the nature of the limiting nutrient, which may be attributed to the metabolic capabilities associated with each family. We modeled this scenario by developing a procedure that sampled consumer coefficients from four metabolic families, ensuring that consumers from the same family were metabolically similar (see Fig.4.19). We randomly sampled a set of 100 consumer vectors (or species) from four families, then simulated growth on 20 random subsets of 50 species on one of three resources (labeled here as A, B or C). As in our experimental data (Fig. 4.10A), simulated communities converged to similar family-level structures (Fig.4.18C), despite displaying variation at the species level (Fig. 4.18B). We confirmed the correspondence between family-level convergence and functional convergence by computing the community-wide metabolic capacity per simulation, resulting in a predicted community-wide resource uptake rate for each resource.

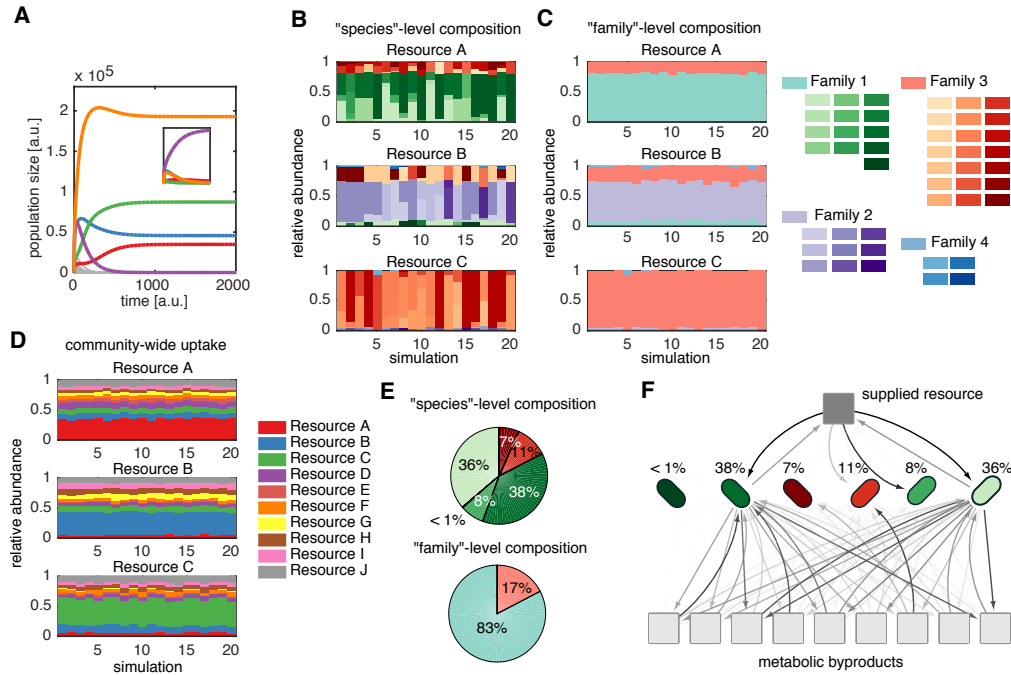
Communities grown on the same resource converged to similar uptake capacities with an enhanced ability to consume the limiting nutrient (Fig. 4.18D). Importantly, this functional convergence is exhibited even when consumers are drawn from uniform distributions, with no enforced family-level consumer structure, suggesting that the emergence of functional structure at the community level is a universal feature of consumer resource models (Fig. 4.20).

We frequently observed that several species belonging to the same metabolic family could coexist at equilibrium. These guilds of coexisting consumers from the same family were capable of supporting the stable growth of rare ( $<1\%$  relative abundance) taxa, rather than a single representative from each family (Fig. 4.18E), similar to our experimental data (Fig. 4.2C,E). Our model suggests that species are stabilized by a dense facilitation network (Fig. 4.18F), consistent with observations of widespread metabolic facilitation in experiments (Fig.4.15D).

Thus, we find that simulations of community dynamics with randomly generated metabolisms and resource uptake capabilities capture a wide range of qualitative observations found in our experiments, and recapitulate previous empirical observations in natural communities [85, 39].

## **Discussion**

In the absence of a theory of microbiome assembly, it is often difficult to determine whether empirically observed features of natural microbiomes are the result of system-specific determinants, such as the evolutionary history and past selective pressures at the host level [39], or whether they are simply emergent generic properties of large self-assembled communities. Our results show that the generic statistical properties of large consumer-resource ecosystems include large taxonomic diversity even in simple environments, a stable community-level function in spite of species turnover, and a mixture of predictability and variability at different taxonomic depths in how nutrients determine community composition. All of these features are not only observed in our experiments, but they have been also been reported in systems as diverse as the human gut [85, 39], plant foliages [111] or the oceans [183, 124]. Our theoretical results thus explain the ubiquity of these empirical findings, and suggest that they may reflect universal and generic properties of large



**Figure 4.18: A simple extension of classic ecological models recapitulates several experimental observations.** MacArthur's consumer resource model was extended to include 10 byproduct secretions along with consumption of a single primary limiting nutrient (see Supporting information), controlled by a global stoichiometric matrix  $D_{\beta,\alpha}$ , which encodes the proportion of the consumed resource  $\alpha$  that is transformed to resource  $\beta$  and secreted back into the environment. Consumer coefficients were sampled from 4 characteristic prior distributions, representing four families of similar consumption vectors. (A) Simulations using a randomly sampled global stoichiometric matrix generically resulted in coexistence of multiple competitors, while setting this matrix zero eliminated coexistence (A, inset). Random ecosystems often converged to very similar family-level structures (C), despite variation in the species-level structure (B). The family-level attractor changed when providing a different resource to the same community (B-C, subplots). (D) Total resource uptake capacity of the community was computed (Supporting Information), analogous to the inferred metagenome (see Fig. 4.10D), and is, like the family-level structure, highly associated with the supplied resource. (E) Communities that formed did not simply consist of single representatives from each family, but often consisted of guilds of several species within each family, similar to experimental data. (F) The topology of the flux distribution shows that surviving communities all compete for the primary nutrient and competition is stabilized by differential consumption of secreted byproducts. The darkness of the arrows corresponds to the magnitude of flux.

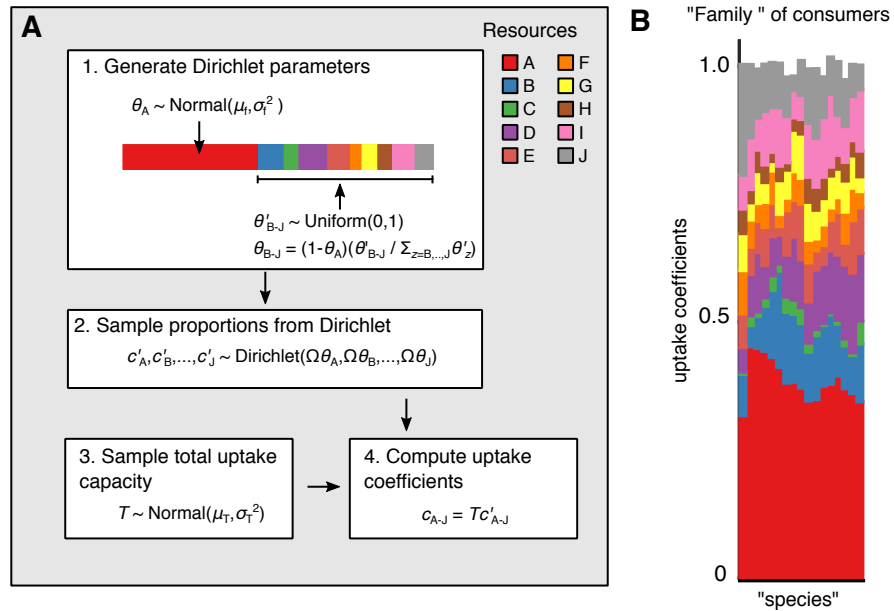


Figure 4.19: **Generation of families of consumers in consumer resource models.** (A) A flow diagram describing the processes of generating families of consumers in consumer resource models. (1) First we define a set of parameters for a Dirichlet distribution specifying the proportion the consumers total uptake used to important each resource ( $\theta_{\text{resource}}$ ), where we each family has a preferred resource (red). (2) We then sample uptake proportions for each resource , from the Dirichlet distribution, and multiply these values with a species dependent the total uptake capacity (Step 3,  $T$ ) to obtain the consumption rate of resource for each consumer. (B) A stacked bar plot showing the uptake coefficients (consumption rates,) for each sampled consumer and resource. Although each species has different uptake rates for each resources, each consumer has a high uptake coefficient for resource A.

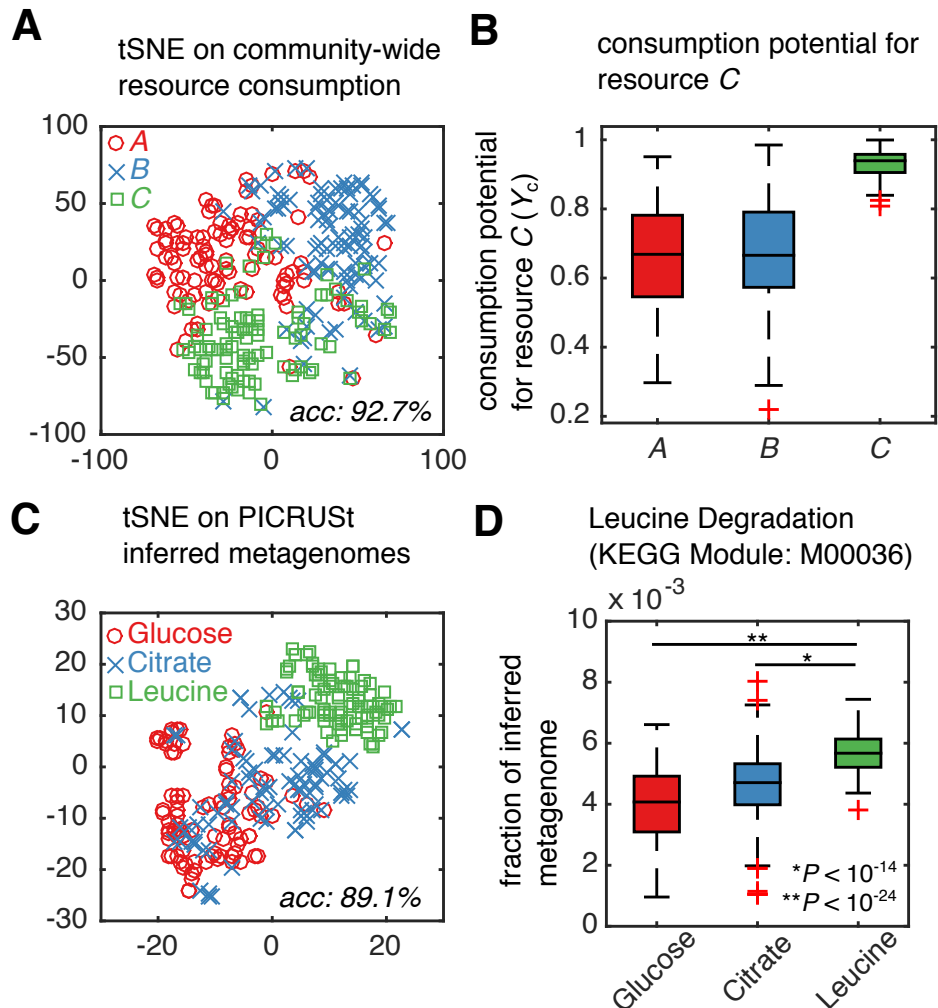


Figure 4.20: **Functional clustering is observed in both consumer resource models and experiments.** (A) Simulations of the microbial consumer resource model (see SI text) was performed by randomly sampling consumer and stoichiometric matrices from uniform distributions, then supplying one of three resources in the environment (denoted as A, B and C here), and the communities capacity to consume each resource was computed (Supplemental information). t-distributed stochastic neighbor embedding (tSNE) was used to reduce dimensionality of the resource uptake vectors and plotted in 2-D, which revealed clustering of uptake capacity based on the identity of the resource in the environment. (B) The distribution of community-wide uptake capacity for resource C when grown on three different resources (x-axis). Note that even in the presence of stabilizing mechanisms like cross-feeding, the dominant signal is the capacity to uptake the primary nutrient. (C) predictions from the model are compared to experiment, where we performed dimensionality reduction on inferred metagenomes. We then computed the relative abundance of genes used for leucine degradation (D), showing that communities grown in leucine are enriched genes involved in leucine degradation relative to communities grown in citrate (Mann Whitney U-test:  $P < 10^{-14}$ ) or glucose (Mann Whitney U-test:  $P < 10^{-24}$ ). Note that in (A) and (C), SVMs were trained to predict the carbon source from either the community-wide uptake rates (in A) or the metagenome (in C), and the leave-one out cross-validation accuracy is reported in the lower right corner.

self-sustained microbial communities rather than specific adaptations. In spite of their simplicity, consumer resource models may not only capture many of the generic qualitative features observed in the experiments, but also can recapitulate more subtle aspects of the experiments including the existence of temporal blooms in species that eventually go extinct and family level similarity of communities (Fig. 4.18A,C). However, the models lack biochemical detail and thus do not have the resolution to explain other experimental results such as pH changes, diauxic shifts, or the fact that glucose and citrate communities are more similar to each other than they are to those stabilized in leucine (Fig. 4.10A).

The theory and simple experimental set-up described above also allowed us to identify widespread mechanisms that lead to the assembly of large, stable communities. We find evidence that densely-connected cross-feeding networks may stabilize competition within guilds of highly related species that are all strong competitors for the supplied carbon source. Such cross-feeding networks naturally lead to collective rather than pairwise interactions, supporting the hypothesis that higher-order interactions play a critical stabilizing role in complex microbiomes [109, 7]. Whether these findings are generic in more complex environments with a larger number of externally supplied resources, or for microbial communities highly evolved in a static environment remains to be elucidated. For instance, the experiments and theory presented in this work indicate that the isolated microbial communities consist of metabolic generalists, rather than metabolic specialists [185], capable of consuming both the supplied resource as well as metabolic byproducts. It is unclear whether these findings are generalizable to highly evolved microbial communities in static environments where metabolic specialization may confer fitness advantages [185]. We propose that high-throughput top-down approaches to community assembly that are amenable to direct mathematical modeling represent an underutilized but highly promising avenue to reveal the existence of generic mechanisms and statistical rules of microbiome assembly, as well as a stepping stone towards developing a quantitative theory of the microbiome.

## **Experimental methods**

### **Isolating microbial communities from natural ecosystems**

Leaf or soil samples (1 g) were collected from natural environments using sterile tweezers and placed in 15 mL falcon tubes. In the lab, 10 mL of 5 % NaCl buffer was added to each sample and allowed to incubate for 48 hours at room temperature. 40% glycerol stock solutions were prepared from aqueous sample suspensions and frozen at -80 °C for storage.

### **Preparation of 96-well media plates**

All media contained 0.07 C-mole/L of carbon source (glucose, citrate or leucine) and was sterile-filtered with a 0.22  $\mu\text{m}$  filter (Millipore). Stock solutions of carbon sources were stored at 4°C for no more than 1 month. M9 media was prepared from concentrated stocks of M9 salts (without  $\text{MgSO}_4$  or  $\text{CaCl}_2$ ) and stock solutions of  $\text{MgSO}_4$  and  $\text{CaCl}_2$ . 500  $\mu\text{L}$  cultures containing 450  $\mu\text{L}$  of sample and 50  $\mu\text{L}$  stock carbon source were grown in 96 deep-well plates (VWR). For the first two cell passages, cycloheximide was added to the media at a concentration of 200  $\mu\text{g}/\text{mL}$  to inhibit eukaryotic growth.

### **Passaging microbial populations**

Starting inocula were obtained directly from the initial buffered solution of microbiota by inoculating 4  $\mu\text{L}$  into 500  $\mu\text{L}$  culture media. For each sample, 4  $\mu\text{L}$  of the culture medium was dispensed into all 60 wells of the fresh media plate. Cultures were allowed to grow for 48 hours at 30 °C in static broth, then each culture was triturated 10 times to ensure communities were homogenized before passaging. Passaging was performed by taking 4  $\mu\text{L}$  from each culture to use as inocula in 500  $\mu\text{L}$  of fresh media, and cells were allowed to grow again. Cultures were passaged 12 times (84 generations). Optical density (OD<sub>620</sub>) was used to measure biomass in cultures after the 48-hour growth cycle. Samples to be sequenced were collected and stored by spinning down in a micro-centrifuge for 10 min at 14,000 RPM at room temperature. Cell pellets were stored at -20°C.

### **DNA extraction, library preparation and sequencing**

Cell pellets were re-suspended and incubated at 37 °C for 30 min in enzymatic lysis buffer (20 mM Tris-HCl, 2mM sodium EDTA, 1.2% Triton X-100) and 20 mg/mL of lysozyme from chicken egg white (Sigma-Aldrich) to lyse the cell walls of Gram-positive bacteria. Following cell lysis, the DNA extractions were performed following the DNeasy 96 protocol for animal tissues (Qiagen). The clean DNA was eluted in 100  $\mu$ L elution buffer of 10 mM Tris-HCl, 0.5 mM EDTA at pH 9.0. DNA concentration was quantified using Quan-iT PicoGreen dsDNA Assay Kit (Molecular Probes, Inc.) and normalized to 5 ng/ $\mu$ L for subsequent 16S rRNA sequencing. 16S rRNA amplicon library preparation was conducted using a dual-index paired-end approach developed by Kozich et al. (53). Briefly, PCR-amplified libraries were prepared using dual-index primers (F515/R806) to generate amplicons spanning the V4 region of the 16S rRNA gene, then pooled and sequenced using the Illumina MiSeq platform. For each sample, a 30-cycle PCR was performed in duplicate in 20  $\mu$ L reaction volumes using 5 ng of DNA, dual index primers, and AccuPrime Pfx SuperMix (Invitrogen). Thermocycling conditions consisted of a 2-min initial denaturation step at 95 °C, followed by 30 cycles of the following PCR scheme: (a) 20-second denaturation at 95 °C, (b) 15-second annealing at 55 °C, and (c) 5-min extension at 72 °C. PCR was terminated after a 10-min extension step at 72 °C. After pooling amplicons from duplicate reactions, the PCR products were purified and normalized using the SequalPrep PCR cleanup and normalization kit (Invitrogen). Libraries were then pooled and sequenced using Illumina MiSeq v2 reagent kit, which generated 2x250 base pair paired-end reads at the Yale Center for Genome Analysis (YCGA). For shaking control experiments (Fig. 4.16), library preparation and sequencing was performed at SeqMatic (Fremont, CA). Sequencing and library preparation were identical when compared to the procedure described above, except primers targeted the V3-V4 region of 16S rRNA gene.

### **16S rRNA sequencing analysis**

QIIME 1.9.0 [27] was used to demultiplex and remove barcodes, indexes and primers from raw files, producing FASTQ files with for both the forward and reverse reads for each sample. Dada2

version 1.1.6 was used to infer exact sequence variants (ESVs) from each sample [26]. Briefly, forward and reverse reads were trimmed to 220 and 160 nucleotides, respectively. All other parameters were set to default values. Sequences below 230 or above 242 nucleotides were discarded (indicative of poor merging of paired reads). Bimeras were removed using the “tableMethod” parameter set to “consensus.” A naive Bayes classifier was used to assign taxonomy to ESVs using the SILVA version 123 database [154]. Metagenome inference was performed using PICRUSt [104]. Denoised ESVs were assigned to OTUs using the greengenes database version 13.5 using the QIIME function `pick_closed_reference_otus.py`, with a 97 % similarity cutoff. Communities were normalized using the `normalize_otus.py` function in PICRUSt, and the metagenomes were estimated using the `estimate_metagenome.py` routine.

### **Fermentation assays and isolation of strains**

Four bacterial strains from a representative community stabilized in glucose were isolated and identified taxonomically. The community was plated onto 0.5% agarose Petri-dishes containing M9 supplemented with 0.2% glucose and were allowed to grow for 48 hours at 30°C. Single colonies were then picked from these plates according to their colony morphologies, re-streaked on fresh agarose plates and grown for another 48 hours at 30°C. Single colonies from each isolate grown for 48 hours at 30°C in liquid M9 supplemented with 0.2% glucose were finally stored at -80 °C in 40% glycerol. Isolates were also identified according to their differential ability to ferment the following 16 carbohydrates: adonitol, arabinose, cellobiose, dextrose, dulcitol, fructose, inositol, lactose, mannitol, mannose, melibiose, raffinose, rhamnose, salicin, sucrose, and xylose (Fig 4.6A-B). Fermentation ability was assessed using a phenol red broth base with an added carbohydrate at a final concentration of 1% w/v, except for cellobiose (0.25%) due to its low solubility. Each isolate was grown on an agarose plate, and a single colony was picked and re-suspended into 100  $\mu$ L 1x PBS. 2  $\mu$ L of each isolate was inoculated into 50  $\mu$ L of Phenol red broth + carbon source (in a 384 well-plate, Corning). Spectrophotometric measurements of phenol red (OD450 and OD551) were measured after 0h, 12, 16, and 19 hours of incubation. Clustering of O.D. profiles after 19 hours revealed 4 distinct phenotypic profiles, consistent with morphologies (Fig. 4.6C). Taxonomic

assignments of isolates were verified using full-length 16S rRNA sequencing of DNA extracted from single colonies grown on agarose plates (GENEWIZ), using the online RDP classifier [194].

### **Reconstitution of isolates from a representative community**

To test whether the dominant species isolated from the glucose stabilized communities are able to coexist, we constructed a four-strain community with four strains isolated from one representative community (C2R4). The four isolates belong to four different genera (*Raoultella*, *Enterobacter*, *Pseudomonas*, and *Citrobacter*) and were chosen because they are the most dominant species in the community and display distinctive morphologies, facilitating plate counting. To ensure that the starting densities were similar for all four isolates, single colonies were picked, resuspended into PBS 1x, and the optical densities were normalized to a OD<sub>620</sub> of 0.15. The initial inoculum was prepared by mixing the four isolates in 1:1:1:1 ratio. 4  $\mu$ L of the initial inoculum was transferred to 500  $\mu$ L fresh media M9 with 0.2% Glucose (3 replicate communities) and cultures were incubated at 30 °C (Fig. 4.6D). Every 48 hours, 4  $\mu$ L from each replicate community was transferred to 500 $\mu$ L of fresh growth media for a total of 7 transfers (14 days). OD<sub>620nm</sub> measurements were conducted every 48 hours and the four isolates were enumerated by colony counts on M9+ 0.2% glucose agar plates on Transfer 5 (day 10) and Transfer 7 (day 14). We found that the four isolates were able to stably coexist after 7 transfers (14 days). *Raoultella* was the most abundant strain, followed by *Enterobacter*, and then *Pseudomonas*, and *Citrobacter* (see Fig 4.6E).

### **Metabolic facilitation assay and measurement of glucose depletion**

To determine whether microbial cross-feeding is a potential mechanism that enables coexistence, four isolates from a single representative community were inoculated in 5 mL of M9 media with 0.2% glucose, then incubated for 48 hours at 30 °C (Fig. 4.15A). Cells were then separated from the spent media (SM) using the following procedure: cells were centrifuged at 3000 rpm for 10 min, and SM was filter-sterilized and stored at 4 °C. Cells were re-suspended in the same volume of PBS, and washed two times times by centrifugation (3000rpm, 10min). Cells were diluted to an OD<sub>620</sub> of 0.24 prior to inoculation. There was no detectable glucose remaining in any SM as measured

using the Glucose GO Assay Kit (Sigma), with the exception of the SM from *Pseudomonas*, which was adequately controlled for (see main text). SM was then mixed 1:1 with fresh 2X M9 media with no carbon source. Each isolate was inoculated in each isolates SM-based M9 in triplicate at 1% v/v in a 384 well plate (Corning). The plate was incubated in a standard plate reader (Thermo 498 Scientific), and OD620 was measured every 10 min at 30 °C.

We sought to determine whether glucose-stabilized communities were able to grow after glucose depletion, which would suggest that biomass accumulation is attributed to consumption of metabolic byproducts. For this, 95 glucose-stabilized communities were inoculated in a 96 deep-well plate from frozen stock in 500  $\mu\text{L}$  of M9 0.2% glucose. Two initial transfers with 48 hours incubation were performed as previously described (30 °C no shaking). The third transfer was performed in duplicate and with final volume 600  $\mu\text{L}$ . From these two plates, 100  $\mu\text{L}$  samples were taken at 24, 36, 48 and 56 hours. OD620 was measured, followed by the measurement of glucose using the Glucose GO Assay Kit (Sigma). Glucose concentrations were inferred using linear regression from the standard curve, although no sample at any time point showed detectable levels.

### **Low abundant growth with no supplied carbon source**

Passaging experiments were performed using M9 synthetic media with no additional carbon sources, which resulted in the stabilization of very low abundant microbial communities (Fig. 4.5). Growth was often several orders of magnitude lower than growth on either the primary nutrient (fig. 4.5C) or secreted byproducts (fig. 4.15E-F), suggesting that metabolic consumption of secreted byproducts is more likely to contribute to stabilizing competition than consumption of low levels of latent resources in the deionized water. To determine community richness resulting from growth on the provided resource, we estimated the abundance of 16S amplicon reads deriving from contamination either by cross-well contamination or microbial growth on the low levels of total organic carbon in deionized water (Fig. 4.5A-B). For each of the 12 initial points, communities propagated for 84 generations with either with M9 and 0.2% glucose, or M9 and no additional carbon source. We plated communities on 0.5% agarose plates containing M9 minimal media and 0.2% D-glucose to determine the colony forming units (CFU) per ml (Fig. 4.5C). CFU/ml was used as a proxy for total

cell number in the community because of the strong correlation with cell counting using a hemocytometer (Fig. 4.5D). The relative contribution of CFU for growth on water alone compared to growth on D-glucose was then used as a relative frequency cutoff for each of the 12 initial communities, respectively (Fig. 4.5E). These values allowed us to estimate lower bounds for community diversity derived from the supplied the carbon source (Fig. 4.7B).

## Statistical and computational methods

### Statistical tests for Beta diversity differences

The co-variables explored in this study are the regional pool of species (initial inocula) and the carbon source supplied in the media. Between samples, we used Renkonen similarity at the family taxonomic level as a measure of beta diversity between communities, which is defined as:

$$D(x, y) = 1 - \frac{1}{2} \times \sum_i |x_i - y_i| \quad (4.1)$$

where  $x_i$  and  $y_i$  are the abundance of taxon  $i$  in sample  $X$  and sample  $Y$ , respectively. We computed the family-level Renkonen similarities between all samples and grouped pairwise similarities if pairs were passaged on the same carbon source, or if pairs of samples originated from the same inocula. We used the one-tailed Kolmogorov-Smirnov test (MATLAB function *kstest2*) to determine if the pairwise similarities grouped by carbon source were on average higher than pairwise similarities grouped by initial inocula (see Fig. 4.14).

### Test of temporal variation and replicate variation

We estimated the variability in community composition from different replicates from inoculum 2 (see Fig 4.2F) and compared this to the variability in community composition between the last three transfers in our passaging experiment. To calculate the variability across replicates, we computed the Renkonen Similarity between each pair of replicates after the last transfer (transfer 12). To calculate the temporal variation within a single replicate, we calculated the Renkonen Similarity

within a replicate at transfers 10,11,12. We used only the final three transfers to ensure that the community composition has had enough transfers to stabilize and to ensure that the number of similarity scores used to assess the temporal variation was similar to the number similarity scores used to assess the replicate variation ( $N = 24$  within time-series, and  $N = 28$  between time-series). We then assessed if replicate variations at the genus and family level were larger than the temporal variations at the same taxonomical resolution using a standard non-parametric test (in this case the Mann-Whitney U test). The statistical test showed that the replicate variation is significantly larger than the temporal variation at the genus level ( $P = 1.1 \times 10^{-5}$ ) while at the family level this was not the case ( $P = 0.0624$ ).

### **Prediction of media carbon source from community structure**

To assess the predictive quality of the community structure and inferred metagenomes, we trained and evaluated multi-class support vector machine (SVM) models. SVMs were constructed using the MATLAB function `fitecoc` and evaluated using 10-fold cross validation in fig. 2b or leave one out cross-validation in Fig. 4.20. Features used in the SVM were either the clr-transformed relative abundances at the family taxonomic level in Fig. 4.10B or the clr-transformed inferred metagenome composition in Fig. 4.20.

## **Theoretical methods**

### **Microbial Consumer Resource Model**

#### **Dynamical equations**

The model presented in the paper is a modification of Robert MacArthur's consumer resource model [116, 30, 31], which models the per-capita growth of species as a function of resource consumption rate. We begin by first re-stating the dynamics of individual species, followed by a modified form of resource dynamics that include environmental modification during bacterial growth.

Let us denote the set of all possible resources by  $R_\alpha$  where  $\alpha = 1 \dots M$ . Furthermore, let us

denote the set of all species by  $N_i$  where  $i = 1 \dots N$ . Each species is characterized by a resource utilization matrix  $C_{i\alpha}$  that measures the rate at which the species uptakes resource  $\alpha$ . Furthermore, there is a resource quality function  $\Delta w_{i\alpha}$  which tells us the growth rate of species  $i$  on resource  $\alpha$ . Assuming that there is a minimum maintenance energy required for growth, this gives us

$$\frac{1}{N_i} \frac{dN_i}{dt} = b_i \left( \sum_{\alpha} \Delta w_{i\alpha} C_{i\alpha} R_{\alpha} - m_i \right) \quad (4.2)$$

This assumes populations die if they cannot achieve minimum growth rate to survive  $m_i$ . The principle modification to the MacArthur's consumer resource model is the addition of a stoichiometric matrix that encodes the proportion of consumed resources that are transformed into new resources and secreted back into the environment. A wide variety of bacterial heterotrophs are capable of excreting a large fraction of the carbon input through overflow metabolism even under aerobic conditions [146, 15].

To model the bacterial secretion of metabolic byproducts, let the matrix  $D_{\beta\alpha}^i$  be a stoichiometric matrix for species  $i$  that encodes the rate at which resource  $\beta$  is produced if species  $i$  is utilizing resource  $\alpha$ . In particular, the rate of production of resource  $\beta$  by species  $i$  is proportional to the rate that a species takes up resource  $\alpha$  times this matrix:

$$\sum_{\alpha,i} D_{\beta\alpha}^i C_{i\alpha} R_{\alpha} N_i \quad (4.3)$$

giving rise to the full dynamical equation for the abundance of resource  $\beta$ :

$$\frac{dR_{\beta}}{dt} = \frac{K_{\beta} - R_{\beta}}{\tau_{\beta}} - \sum_i C_{i\beta} R_{\beta} N_i + \sum_{\alpha,i} D_{\beta\alpha}^i C_{i\alpha} R_{\alpha} N_i \quad (4.4)$$

where  $K_{\beta}$  is the initial resource abundance supplied in fresh media, and  $\tau_{\beta}$  is the replenishing (i.e. transfer) rate during batch culture passaging. Note that we parametrize the growth rate with a function  $\Delta w_{i\alpha} = w_{\alpha} - \sum_{\beta} D_{\beta\alpha}^i w_{\beta}$ , which ensures energy is balanced in our model.

### Ensuring energy conservation

For heterotrophic, aerobic bacteria, energy and carbon sources are often coupled within reduced organic substrates [67]. Following the laws of thermodynamics, the total energy (or free energy) available from resources supplied in the environment constrains the total energy secreted back into the environment. However, energy (or free energy) is not well defined in our far from equilibrium dynamical equations. This quantity is indirectly associated with the resource quality,  $w$ , which is a phenomenological parameter that represents the relative gain in a limiting factor (e.g. carbon or energy) per consumed resource. Our model assumes that the limiting factor is linear in the growth rate, which is expected if species are catabolically-limited, and  $w_\alpha$  is the ATP yield for a resource  $\alpha$ . To ensure energy is not created during the metabolism of a resource, we ensure that the secretion matrix,  $D_{\beta\alpha}^i$  is constrained by the following relation:

$$\sum_{\beta} w_{\beta} D_{\beta\alpha}^i < w_{\alpha} \quad (4.5)$$

### Sampling consumer matrices from metabolic "families"

To simulate the scenario where consumers are non-randomly distributed and taxonomically related, we sampled consumer coefficients from a prior distribution where "families" of consumers share similar consumption coefficients. In this formulation, consumer coefficients are drawn from dirichlet distributions, and the dirichlet concentration parameter encodes the family-level consumption preferences and variability. In our model, sampling from a dirichlet distribution results in stochastically partitioning a fixed amount of cellular resources dedicated for nutrient uptake (e.g. transporters) into groups, and the concentration parameter fixes the average across these samples.

The family-level consumption properties are represented by two parameters,  $\theta_{\alpha,f}$  and  $\Omega_f$ , where  $\theta_{\alpha,f}$  is the concentration parameter for resource  $\alpha$  in family  $f$ , and  $\Omega_f$  is the magnitude of the all concentration parameters, such that:  $\sum_{\alpha} \theta_{\alpha,f} = \Omega_f$ . For family  $f$ , we wish to construct a family of consumers with a tunable degree of preference for resource  $\alpha = f$ . Thus we first sample

$a_{\alpha=f}$  using the following relation:

$$\theta_{\alpha=f}(\mu, \sigma) \sim \text{Normal}(\mu, \sigma^2) \quad (4.6)$$

Note that in all simulations  $\mu$  and  $\sigma$  are chosen to be bounded between 0 and 1. For other concentration parameters we first sample them from a uniform distribution,  $\theta'_{\alpha \neq f} \sim \text{Uniform}(0, 1)$ . The concentration parameters are then normalized using the following formula:

$$\theta_{\alpha \neq f} = (1 - \theta_{\alpha=f}) \frac{\theta'_{\alpha \neq f}}{\sum_{\alpha \neq f} \theta'_{\alpha \neq f}} \quad (4.7)$$

Resulting in a set of concentration parameters  $\theta_{\alpha,f}$ , which can be represented in vector notation as  $\theta_f$ . Note that the parameters  $\mu$  and  $\sigma$  control how “specialist” a family of consumers will be. For all simulations we choose  $\mu = 0.4$  and  $\sigma = 0.01$ .

Let the  $c'_{i,\alpha}$  represent the *relative* specific uptake rate of resource  $\alpha$  for species  $i$ . We next draw a set of relative uptake rates for species  $i$  for all  $M$  resources simultaneously from the following Dirichlet distribution:

$$(c'_{i,1}, c'_{i,2}, \dots, c'_{i,M}) \sim \text{Dirichlet}(\Omega_f \theta_{1,f}, \Omega_f \theta_{2,f}, \dots, \Omega_f \theta_{M,f}) \quad (4.8)$$

where  $\Omega_f$  controls the total variability with each family. A high  $\Omega_f$  ensures that “species” are very similar, where a low  $\Omega_f$  results in “species” that are variable. For our simulations, we chose  $\Omega_f = 100$  for all families. For each species  $i$ , we then sampled a total resource capacity  $T_i \sim \text{Normal}(1, 0.01)$ , to ensure we didn’t obtain an ecosystem with infinite coexistence and solely neutral interactions [153]. Consumer coefficients were then computed using the following function:

$$c_{i,\alpha} = T_i c'_{i,\alpha} \quad (4.9)$$

## Numerical Simulations

For all simulations, we set the number of species to be  $N = 100$  and the number of resources to be  $M = 10$ . The resource qualities, the resource replenishment rates, the maintenance and the growth rate multipliers were set to unity, such that:  $w_{i\alpha} = \tau_\alpha = m_i = b_i = 1$  for all species  $i$  and resources  $\alpha$ . We initialized simulations to model dynamics on a single externally supplied resource  $\gamma$  by setting  $K_\alpha = 10^6$  if  $\alpha = \gamma$  and 0 otherwise. For all simulations, we assumed that the stoichiometric matrix is species-independent, such that  $D_{\beta\alpha}^i = D_{\beta\alpha}$ . Stoichiometric matrices were drawn from uniform distributions, such that:

$$D_{\beta\alpha} \sim \text{uniform}(0, \frac{1}{M}) \quad (4.10)$$

Note that by setting the upper bound of  $D_{\beta\alpha} < \frac{1}{M}$  and  $w_{i\alpha} = 1$ , we ensure that energetic constraints are not violated. For Fig. 4.20, we sampled consumer coefficients from the following uniform distribution:  $C_{i\alpha} \sim \text{uniform}(0, 1)$ .

In Fig 4.18, consumer matrices were drawn from Dirichlet distributions (see previous section), while in Fig. 4.20, consumer matrices were drawn from uniform distributions. Simulations were performed in MATLAB 2015a using ODE solver ode15s. Simulations were performed for at least  $10^4$  timesteps, where the vast majority of simulations resulting in reaching stable equilibria in roughly 500 timesteps. Code is available on the following GitHub repository: <https://github.com/jgoldford/mcrm>.

## Metagenomic analysis and comparison with experiment

Based on our experimental results, we expected that the collection of genes in the community (the metagenome) would associate with the externally-supplied resource (e.g. glucose, citrate, or leucine). To compare to the model, we implicitly assume that the metagenome is associated with the community-wide uptake capability of externally supplied resources. This assumption requires that gene dosage is positively associated with the activity of transporters [80]. From experimental

data, we estimated the metagenome from 16S rRNA amplicon sequencing data using PICRUSt [104]. The gene abundance profiles were normalized to sum to unity, and were transformed using the centered log-ratio transform [2]. Formally, for a composition  $x$ , we define the centered log-ratio transform (clr) as:

$$\text{clr}(x) = z = \left[ \ln \left( \frac{x_1}{g(x)} \right), \dots, \ln \left( \frac{x_D}{g(x)} \right) \right] \quad (4.11)$$

where  $g(x) = \sqrt[D]{\prod_i x_i}^{-1}$ . We then construct a matrix,  $Z$ , where  $z_{i,j}$  represents the clr-transformed abundances for gene  $i$  in sample  $k$ . We then used tSNE to reduce the dimensionality of the clr-transformed metagenome matrix  $Z$ , as seen in Figure 2c and in the main text. In Figure 4.20, the fraction of the metagenome that was dedicated the Leucine degradation (KEGG Module M00036) was computed for each sample, then grouped by the externally-supplied resource ( $x$ -axis), revealing a strong concordance between the presence of a specific limiting nutrient and the community-wide metabolism for that limiting nutrient.

To compare experiments to the model, we first simulated the population dynamics and found the steady state abundance for each species  $i$ ,  $N_i^*$ . We then computed the total uptake of resource  $\alpha$  ( $Y_\alpha$ ) using the following equation:

$$Y_\alpha = \sum_i C_{i\alpha} N_i^* \quad (4.12)$$

For each simulation  $k$  on a resource  $\gamma$ , we constructed a matrix of community wide uptake rates with matrix elements equal to  $Y_{k\gamma}$ . The total uptake capacity per simulation was normalized to sum to unity, and was transformed using the clr transform, just like in the case with inferred metagenomic data. Dimensionality reduction was then performed on this matrix using tSNE, and plotted in the Fig. 4.20.

---

<sup>1</sup>For all metagenome samples, a small value,  $\epsilon = 10^{-20}$  was added to each  $x_i$  to prevent  $g(x)$  from becoming zero.

### Monod model

Microbes in a community can coexist in an environment with a single limiting resource if strains have a peak fitness at some intermediate concentration of the limiting resource [182]. We investigated whether this mechanism may be responsible for coexistence by isolating the dominant taxa from a representative community, and measuring the growth rates at various concentrations to estimate parameters used in a Monod growth model. First, isolates were obtained via plating, then grown in minimal M9 salts media supplemented with glucose at concentrations ranging from 0.01 - 0.2 %. For each strain  $i$  on glucose concentration  $S$ , we fit a curve to the following logistic equation:

$$\frac{1}{N_i} \frac{dN_i}{dt} = r_i(S) \left( 1 - \frac{N_i}{K_i(S)} \right) \quad (4.13)$$

where  $r_i(S)$  is the maximum per capita growth rate, and  $K_i(S)$  is the carrying capacity of strain  $i$  on a carbon source with abundance  $S$ . Monod parameters for each species ( $\mu_i$  and  $\kappa_i$ ), were then fitted using the following function:

$$r_i(S|\mu_i, \kappa_i) = \frac{\mu_i S}{\kappa_i + S} \quad (4.14)$$

These parameters were then used in the following dynamic growth and substrate equations:

$$\begin{aligned} \frac{1}{N_i} \frac{dN_i}{dt} &= \frac{\mu_i S}{\kappa_i + S} - m_i \\ \frac{dS}{dt} &= \frac{\alpha_s - S}{\tau} - \sum_i \frac{N_i}{Y_i} \frac{\mu_i S}{\kappa_i + S} \end{aligned} \quad (4.15)$$

where  $Y_i$  is the yield coefficient for growth on glucose,  $\alpha_s = 0.2\%$  is the supply added every time step  $\tau = 48$  hours. We set  $Y_i = 42$  (in units of O.D. per percent glucose) for each species<sup>2</sup>. We also assume that the maintenance energy is  $7.6 \text{ mmol ATP gCDW}^{-1} \text{ hour}^{-1}$ , which corre-

<sup>2</sup>A yield coefficient of  $0.5 \text{ g/g}$  glucose was used for each species (BNID 105318). Assuming that  $\text{gCDW/cell}$  is roughly  $150 \text{ fg}$  (BNID: 103894), and  $1 \text{ O.D. per mL}$  is  $8 \times 10^8 \text{ cells}$  (BNID: 100985), then  $\frac{0.5 \text{ gCDW}}{1 \text{ g glucose}} \times \frac{0.01 \frac{\text{g glucose}}{\text{mL}}}{\% \text{ glucose}} \times \frac{1 \text{ cell}}{150 \times 10^{-15} \text{ gCDW}} \times \frac{1 \text{ O.D.}}{8 \times 10^8 \frac{\text{cells}}{\text{mL}}} = 42 \frac{\text{O.D.}}{\% \text{ glucose}}$

sponds to a growth rate of approximately  $0.02 \text{ hour}^{-1}$ <sup>3</sup>. Simulations were performed in MATLAB 2015a, using the ode45 solver, and all fitting to experimental data was done using the *fit* function in MATLAB. Fitted Monod curves are plotted in 4.17A, and the outcome of a representative simulation are plotted in 4.17B. Note that in 4.17B, initial conditions were chosen to match experimental relative abundances after the passaging experiment (generation 84). In all simulations, *Raoultella* out-competed all other strains leading to competitive exclusion.

---

<sup>3</sup>The value of maintenance energy was estimated used *E. coli* measurements on glucose minimal media during exponential growth (BNID:111285). This value was converted into the estimated minimum per capita growth rate per hour using the following dimensional analysis:  $\frac{7.6 \times 10^{-3} \text{ mol ATP}}{\text{gCDW h}} \times \frac{1 \text{ mol glucose}}{36 \text{ mol ATP}} \times \frac{1\% \text{ glucose}}{0.01 \text{ g glucose}} \times \frac{0.00012 \text{ gCDW}}{\text{OD}_{600}} \times \frac{42 \text{ OD}_{600}}{\% \text{ glucose}} = 0.0181 \text{ h}^{-1}$

## **Chapter 5**

### **Discussion and perspective**

In this thesis, we explored various problems in the origin of life and microbial ecology from a metabolic perspective at the "ecosystem-scale." This work highlights that metabolism, rather than simply being a mechanism for energy transduction and biosynthesis at the organismal-scale, may be a natural variable for the study of complex living systems across space and time. What follows in this section is a discussion on potential follow-up research, building off the results presented in this thesis, that may enable capturing a more comprehensive and detailed picture of both ancient and modern-day ecosystem-level metabolism.

#### **Metabolism and the origin of life**

Use of network-based approaches to model the emergence of metabolism nearly 4 billion years ago offers a "generative" approach to simulate the relationship between proposed ancient geochemical environments and the structures of ancient metabolic pathways. More deeply, our approach may not merely provide an algorithmic way to map geochemical constraints to early metabolic structures, but may in fact give insight on how life itself evolved [77]. Our interpretation of this hypothesis assumes that, over the course of geological time, biochemical reactions that emerge in the biosphere are, at first-approximation, not lost throughout the course of evolution. This feature is uniquely a function of ecosystem-scale metabolism, where biochemistry is distributed across species and robust to extinction events of individual lineages [191, 60]. Thus, our approach suggests that the evolution of metabolism across large time and spatial scales should be viewed as a process of *time-evolution*, in the dynamical systems sense, rather than solely Darwinian. Future theoretical studies could integrate population genetic models with genome-scale models of metabolism to

concretely address the question how metabolic networks evolve at the ecosystem-level on long time-scales. Such theoretical studies could address the plausibility of our key assumption that metabolic networks represent a historical record of life's history.

In Chapter 3, we used this approach to develop a minimal metabolism that may have fueled ancient living systems before an RNA-based genetic code. However, we explicitly do not address whether such a metabolism could have operated in parallel to information storage mechanisms similar to modern genetic systems. As discussed in the introduction of this thesis, several researchers in the origin of life have suggested that reproduction and evolution may have proceeded the emergence of a genetic system [45, 90, 6, 40, 78, 170, 175]. Future computational studies could look at whether our metabolic model contains autocatalytic cycles [176, 168] capable of amplification and exponential growth [173].

Importantly, there are several short-comings of the presented method that could be improved upon in the future work. First, the network expansion algorithm presented in Chapter 2-3 relies on a database of enzyme-catalyzed reactions, which may or may not reflect important chemical transformations for early living systems. For instance, there could be enzyme-catalyzed reactions that have no non-enzymatic analogue, or non-enzymatic analogues may have not been accessible in ancient living systems. Future studies revealing which portions of enzyme-catalyzed metabolic biochemistry that can be catalyzed non-enzymatically [96, 97, 127, 94, 135, 189] will enable a more accurate reconstructions of ancient metabolic networks. Second, the network expansion algorithm does not explore the inclusion of chemical reactions that have no enzymatic analogue, but may have been important in ancient biochemistry. Third, it is unclear whether non-enzymatic analogues of enzyme-catalyzed reactions would have been kinetically inhibited. Lastly, our current approach does not take into consideration biochemical reactions with no free energy estimate. Future work may leverage emerging quantum chemistry methods [87], which may increase the scope of reactions with estimated free energies relative to component contribution methods [52, 138].

We hope that the model presented in this thesis offers a useful starting point for designing complex chemical networks that resemble plausible ancient metabolic systems. The protocellular metabolic model developed in this thesis represents a simple model for an autotrophic,

self-replicating chemical system. Heavily inspired by Christian de Duve's "thioester world" model [40], our results support the hypothesis that disulfides may have been the first coenzyme system in living systems, providing both redox and energy transduction capabilities. Furthermore, the model also predicts that polymers of hydroxy acids (derived from keto acids that are normally precursors to amino acids), along with inorganic metals and minerals, may have been life's first catalysts. However, determining whether these components are capable of catalyzing proto-metabolic reaction at sufficiently high rates [45] remains an open question.

### **Metabolism and microbial ecology**

In the second part of the thesis, we discuss a new method to re-assemble natural microbial consortia in synthetic environments, revealing patterns commonly observed in various microbial ecosystems. We now discuss interesting applications of the experimental system that can reveal the relationship between metabolism and microbial community structure in more detail.

While our results show that the limiting carbon source is associated with microbial community structure at the family-level (see Fig.4.10), it is unclear why communities grown on glucose are more similar to communities grown on citrate than communities grown on leucine. This result indicates that communities stabilized on limiting-nutrients metabolized by similar metabolic routes may result in similar community structure at the functional level. Indeed, preliminary studies of communities grown on a panel of more than 40 carbon sources consisting of sugars, sugar-alcohols, carboxylic acids and short chain fatty acids suggest that the functional structure (or family-level compositions) of each community cluster by whether the carbon source was consumed via glycolysis (e.g. glucose, fructose, glycerol) or not (e.g. citrate, succinate).

Our experiments indicated that microbes from a representative community, when grown in isolation, secrete several byproducts that can support the growth of every other community member (see Fig. 4.15). However, it is unclear what the structure of the cross-feeding network is when community members are grown together, rather than being inferred based on monoculture studies. We now discuss several computational and experimental techniques that can be used to elucidate the structure of the microbial cross-feeding network *in situ*. Genome-scale metabolic modeling

of microbial metabolism enables the prediction of secreted byproducts resulting from microbial growth [144]. Due to the small size and ease of cultivation of these communities, shotgun metagenomics can be used to generate draft genomes of community members, which can be used to construct metabolic models of individual taxa in the communities. Indeed, we have begun to generate shotgun metagenomic sequencing data from a set of these microbial communities. We have reconstructed draft genomes for all isolated taxa from a representative community (see Fig. 4.15A), and constructed metabolic models of core metabolism using Kbase [5, 49]. Preliminary results of metabolic modeling suggest that each taxa in our representative community secretes organic acids (e.g. lactate, formate, and acetate) that can support the growth the other community members, similar to what was found in Fig. 4.15C-D. Future work will attempt to construct metabolic models directly from metagenomic sequencing data. More generally, we expect that this system will serve as a convenient experimental platform for the assembly of microbial communities amenable for flux balance modeling.

With precise knowledge of the media supporting the stable growth of microbial communities generated in this work, these experimental systems are also highly amenable to more quantitative experimental studies of microbial community nutrient exchange. For instance, with draft genomes available for each community member, new high throughput techniques in transcriptomics and proteomics can help elucidate mechanisms for coexistence at the biochemical level. To determine the topology of cross-feeding networks of these microbial communities, one can extend prior experimental and computational approaches to model and measure the extent of nutrient exchange in these microbial microcosms. Computational predictions of nutrient exchange can be performed using dynamic flux balance analysis [75] or steady-state flux-balance modeling of microbial communities [50]. Models can be improved by integrating high-throughput estimates of protein abundances from ribosome profiling into genome-scale metabolic models [110]. Predicted cross-feeding interactions can be validated using peptide-based metabolic flux analysis [58, 62, 118, 3]. In this assay, isotope-labeling patterns from peptides serve as barcodes for the metabolic states of individual species in a complex community, enabling the identification of metabolic exchanges between community members. Application of these high-throughput ap-

proaches to infer cross-feeding networks in microbial communities will aid in developing a theory of microbial community ecology for large ecosystems.

### **Metabolism as the unifying language bridging scales in the biosphere**

As suggested by previous researchers [175, 60], there may be a deep connection between the fundamental rules that led to life's emergence over 4 billion years ago and what shapes microbial ecosystems today. We argue that this deep connection likely resides in ecosystem-scale metabolism. Our work in microbial ecology suggests that the functional composition of microbial communities, which is analogous to ecosystem-level metabolism, is the stable feature of microbial community assembly in a fixed environment even in short time-scale experiments. This suggests that ecosystem-level metabolism should be considered a natural variable when asking questions about the long-term evolution of the biosphere. We hope that the work presented in this thesis highlights how an ecosystems-level metabolic perspective may aid in our understanding of both the emergence and maintenance of the biosphere.

## Bibliography

- [1] C. T. Adcock, E. M. Hausrath, and P. M. Forster. Readily available phosphate from minerals in early aqueous environments on Mars. *Nature Geoscience*, 6(10):824–827, sep 2013.
- [2] J. Aitchison. *The Statistical Analysis of Compositional Data*. 1986.
- [3] Doug K Allen, Joshua Goldford, James K Gierse, Dominic Mandy, Christine Diepenbrock, and Igor G L Libourel. Quantification of peptide m/z distributions from <sup>13</sup>C-labeled cultures with high-resolution mass spectrometry. *Analytical chemistry*, 86(3):1894–901, feb 2014.
- [4] Jakob L. Andersen, Christoph Flamm, Daniel Merkle, and Peter F. Stadler. A Software Package for Chemically Inspired Graph Transformation. pages 73–88. Springer, Cham, jul 2016.
- [5] Adam P Arkin, Rick L Stevens, Robert W Cottingham, Sergei Maslov, Christopher S Henry, Paramvir Dehal, Doreen Ware, Fernando Perez, Nomi L Harris, Shane Canon, Michael W Sneddon, Matthew L Henderson, William J Riehl, Dan Gunter, Dan Murphy-Olson, Stephen Chan, Roy T Kamimura, Thomas S Brettin, Folker Meyer, Dylan Chivian, David J Weston, Elizabeth M Glass, Brian H Davison, Sunita Kumari, Benjamin H Allen, Jason Baumohl, Aaron A Best, Ben Bowen, Steven E Brenner, Christopher C Bun, John-Marc Chandonia, Jer-Ming Chia, Ric Colasanti, Neal Conrad, James J Davis, Matthew DeJongh, Scott Devoid, Emily Dietrich, Meghan M Drake, Inna Dubchak, Janaka N Edirisinghe, Gang Fang, Jose P Faria, Paul M Frybarger, Wolfgang Gerlach, Mark Gerstein, James Gurtowski, Holly L Haun, Fei He, Rashmi Jain, Marcin P Joachimiak, Kevin P Keegan, Shinnosuke Kondo, Vivek Kumar, Miriam L Land, Marissa Mills, Pavel Novichkov, Taeyun Oh, Gary J Olsen, Bob Olson, Bruce Parrello, Shiran Pasternak, Erik Pearson, Sarah S Poon, Gavin Price, Srividya Ramakrishnan, Priya Ranjan, Pamela C Ronald, Michael C Schatz, Samuel M D Seaver, Maulik Shukla, Roman A Sutormin, Mustafa H Syed, James Thomason, Nathan L Tintle, Daifeng Wang, Fangfang Xia, Hyunseung Yoo, and Shinjae Yoo. The DOE Systems Biology Knowledgebase (KBase). *bioRxiv*, dec 2016.
- [6] RJ Bagley, JD Farmer, and W Fontana. Evolution of a Metabolism. In Langton C G, Taylor C, Farmer J D, and Rasmussen S, editors, *Artificial Life II*, pages 141–158. AddisonWesley, Reading, MA, 1991.
- [7] Eyal Bairey, Eric D. Kelsic, and Roy Kishony. High-order species interactions shape ecosystem diversity. *Nature Communications*, 7:12285, 2016.
- [8] Wolfgang Banzhaf and Lidia Yamamoto. *Artificial Chemistries*. 2015.
- [9] Arren Bar-Even. Does acetogenesis really require especially low reduction potential? *Biochimica et biophysica acta*, 1827(3):395–400, mar 2013.

- [10] Arren Bar-Even, Avi Flamholz, Elad Noor, and Ron Milo. Thermodynamic constraints shape the structure of carbon fixation pathways. *Biochimica et biophysica acta*, 1817(9):1646–59, sep 2012.
- [11] Akiva Bar-nun and Hyman Hartman. Synthesis of organic compounds from carbon monoxide and water by UV photolysis. *Origins of Life*, 9(2):93–101, 1978.
- [12] Richard Baran, Eoin L. Brodie, Jazmine Mayberry-Lewis, Eric Hummel, Ulisses Nunes Da Rocha, Romy Chakraborty, Benjamin P. Bowen, Ulas Karaoz, Hinsby Cadillo-Quiroz, Ferran Garcia-Pichel, and Trent R. Northen. Exometabolite niche partitioning among sympatric soil bacteria. *Nature Communications*, 6:8289, 2015.
- [13] Uri Barenholz, Dan Davidi, Ed Reznik, Yinon Bar-On, Niv Antonovsky, Elad Noor, and Ron Milo. Design principles of autocatalytic cycles constrain enzyme kinetics and force low substrate saturation at flux branch points. *eLife*, 6:1–32, 2017.
- [14] Markus Basan, Sheng Hui, Hiroyuki Okano, Zhongge Zhang, Yang Shen, James R Williamson, and Terence Hwa. Overflow metabolism in *Escherichia coli* results from efficient proteome allocation. *Nature*, 528(7580):99–104, dec 2015.
- [15] Markus Basan, Sheng Hui, Hiroyuki Okano, Zhongge Zhang, Yang Shen, James R. Williamson, and Terence Hwa. Overflow metabolism in *Escherichia coli* results from efficient proteome allocation. *Nature*, 528(7580):99–104, dec 2015.
- [16] David A. Baum and Kalin Vetsigian. An Experimental Framework for Generating Evolvable Chemical Systems in the Laboratory. *Origins of Life and Evolution of Biospheres*, pages 1–17, 2016.
- [17] B D Bennett, E H Kimball, and Melissa Gao. Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nature chemical biology*, 5(8):593–599, 2009.
- [18] Claudia Bonfio, Luca Valer, Simone Scintilla, Sachin Shah, David J Evans, Lin Jin, Jack W Szostak, Dimitar D Sasselov, John D Sutherland, and Sheref S Mansy. UV-light-driven prebiotic synthesis of ironsulfur clusters. *Nature Chemistry*, (July):1–6, 2017.
- [19] Rogier Braakman, Michael J. Follows, and Sallie W. Chisholm. Metabolic evolution and the self-organization of ecosystems. *Proceedings of the National Academy of Sciences*, 114(15):E3091–E3100, 2017.
- [20] Rogier Braakman and Eric Smith. The emergence and early evolution of biological carbon-fixation. *PLoS Computational Biology*, 8(4), 2012.
- [21] Rogier Braakman and Eric Smith. The compositional and evolutionary logic of metabolism. *Physical biology*, 10(1):011001, feb 2013.
- [22] Rogier Braakman and Eric Smith. Metabolic evolution of a deep-branching hyperthermophilic chemoautotrophic bacterium. *PLoS ONE*, 9(2), 2014.

- [23] E. Branscomb, T. Biancalani, N. Goldenfeld, and M. Russell. Escapement mechanisms and the conversion of disequilibria; the engines of creation. *Physics Reports*, 677:1–60, 2017.
- [24] Wolfgang Buckel and Rudolf K. Thauer. Energy conservation via electron bifurcating ferredoxin reduction and proton/Na<sup>+</sup> translocating ferredoxin oxidation. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1827(2):94–113, 2013.
- [25] Catherine Burke, Peter Steinberg, Doug Rusch, Staffan Kjelleberg, and Torsten Thomas. Bacterial community assembly based on functional genes rather than species. *Proceedings of the National Academy of Sciences of the United States of America*, 108(34):14288–93, aug 2011.
- [26] Benjamin J Callahan, Paul J Mcmurdie, Michael J Rosen, Andrew W Han, Amy Jo Johnson, and Susan P Holmes. DADA2 : High resolution sample inference from amplicon data. *Nature methods*, 13(August 2015):0–14, 2015.
- [27] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttenhower, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, may 2010.
- [28] Kuhan Chandru, Alexis Gilbert, Christopher Butch, Masashi Aono, and H James Cleaves. The Abiotic Chemistry of Thiolated Acetate Derivatives and the Origin of Life. *Scientific Reports*, 6(July):29883, 2016.
- [29] Yi-Chien Chang, Zhenjun Hu, John Rachlin, Brian P Anton, Simon Kasif, Richard J Roberts, and Martin Steffen. COMBREX-DB: an experiment centered database of protein function: knowledge, predictions and knowledge gaps. *Nucleic Acids Research*, 44(Database issue):D330–D335, jan 2016.
- [30] Peter Chesson. MacArthur’s consumer-resource model. *Theoretical Population Biology*, 37(1):26–38, 1990.
- [31] Peter Chesson. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol.Syst.*, 31:343–66, 2000.
- [32] G. D. Cody. Primordial Carbonylated Iron-Sulfur Compounds and the Synthesis of Pyruvate. *Science*, 289(5483):1337–1340, aug 2000.
- [33] G.D Cody, N.Z Boctor, J.A Brandes, T.R Filley, R.M Hazen, and H.S Yoder. Assaying the catalytic potential of transition metal sulfides for abiotic carbon fixation. *Geochimica et Cosmochimica Acta*, 68(10):2185–2196, 2004.
- [34] Elizabeth K Costello, Keaton Stagaman, Les Dethlefsen, Brendan J M Bohannan, and David A Relman. The application of ecological theory toward an understanding of the human microbiome. *Science (New York, N.Y.)*, 336(6086):1255–62, jun 2012.

- [35] K. Z. Coyte, J. Schluter, and K. R. Foster. The ecology of the microbiome: Networks, competition, and stability. *Science*, 350(6261):663–666, 2015.
- [36] Jonas Cremer, Markus Arnoldini, and Terence Hwa. Effect of water flow and chemical environment on microbiota growth and composition in the human colon. *Proceedings of the National Academy of Sciences*, 114(25):6438–6443, jun 2017.
- [37] Manoshi S. Datta, Elzbieta Sliwerska, Jeff Gore, Martin F. Polz, and Otto X. Cordero. Microbial interactions lead to rapid micro-scale successions on model marine particles. *Nature Communications*, 7(May):11965, 2016.
- [38] Lawrence a David and Eric J Alm. Rapid evolutionary innovation during an Archaeal genetic expansion. *Nature*, 469(7328):93–96, 2011.
- [39] Lawrence A David, Corinne F Maurice, Rachel N Carmody, David B Gootenberg, Julie E Button, Benjamin E Wolfe, Alisha V Ling, A Sloan Devlin, Yug Varma, Michael A Fischbach, Sudha B Biddinger, Rachel J Dutton, and Peter J Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome Long-term dietary intake influences the structure and activity of the trillions of microorganisms residing in the human gut. *Nature*, 505, 2014.
- [40] Christian de Duve. *Blueprint for a cell: the nature and origin of life*. Neil Patterson Publishers, Carolina Biological Supply Company, Burlington, N.C., 1991.
- [41] David Deamer and Arthur L Weber. Bioenergetics and life’s origins. *Cold Spring Harbor perspectives in biology*, 2(2):a004929, feb 2010.
- [42] Luis Delage, Arturo Becerra, and Antonio Lazcano. The Last Common Ancestor: What’s in a name? *Origins of Life and Evolution of Biospheres*, 35(6):537–554, dec 2005.
- [43] Mark Dörr, Johannes Kässbohrer, Renate Grunert, Günter Kreisel, Willi A Brand, Roland A Werner, Heike Geilmann, Christina Apfel, Christian Robl, and Wolfgang Weigand. A possible prebiotic formation of ammonia from dinitrogen on iron sulfide surfaces. *Angewandte Chemie (International ed. in English)*, 42(13):1540–3, apr 2003.
- [44] S. Duval, K. Danyal, S. Shaw, A. K. Lytle, D. R. Dean, B. M. Hoffman, E. Antony, and L. C. Seefeldt. Electron transfer precedes ATP hydrolysis during nitrogenase catalysis. *Proceedings of the National Academy of Sciences*, 110(41):16414–16419, jun 2013.
- [45] Freeman J. Dyson. A model for the origin of life. *Journal of Molecular Evolution*, 18(5):344–350, sep 1982.
- [46] R E Eakin. An approach to the evolution of metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 49(3):360–6, mar 1963.
- [47] Oliver Ebenhöf, Thomas Handorf, and Reinhart Heinrich. Structural analysis of expanding metabolic networks. *Genome informatics. International Conference on Genome Informatics*, 15(1):35–45, jan 2004.

- [48] R V Eck and M O Dayhoff. Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences. *Science (New York, N.Y.)*, 152(3720):363–366, apr 1966.
- [49] Janaka N. Edirisinghe, Pamela Weisenhorn, Neal Conrad, Fangfang Xia, Ross Overbeek, Rick L. Stevens, and Christopher S. Henry. Modeling central metabolism and energy biosynthesis across microbial life. *BMC Genomics*, 17(1):568, dec 2016.
- [50] Mallory Embree, Joanne K. Liu, Mahmoud M. Al-Bassam, and Karsten Zengler. Networks of energetic and metabolic interactions define dynamics in microbial communities. *Proceedings of the National Academy of Sciences*, 112(50):201506034, 2015.
- [51] S. Fahnstock and A. Rich. Ribosome-Catalyzed Polyester Formation. *Science*, 173(3994):340–343, jul 1971.
- [52] Avi Flamholz, Elad Noor, Arren Bar-Even, and Ron Milo. EQUilibrator - The biochemical thermodynamics calculator. *Nucleic Acids Research*, 40(D1):770–775, 2012.
- [53] Jay G. Forsythe, Sheng Sheng Yu, Irena Mamajanov, Martha A. Grover, Ramanarayanan Krishnamurthy, Facundo M. Fernández, and Nicholas V. Hud. Ester-Mediated Amide Bond Formation Driven by Wet-Dry Cycles: A Possible Path to Polypeptides on the Prebiotic Earth. *Angewandte Chemie - International Edition*, 54(34):9871–9875, 2015.
- [54] Kevin R Foster and Thomas Bell. Competition, not cooperation, dominates interactions among culturable microbial species. *Current biology : CB*, 22(19):1845–1850, oct 2012.
- [55] Shiri Freilich, Raphy Zarecki, Omer Eilam, Ella Shtifman Segal, Christopher S Henry, Martin Kupiec, Uri Gophna, Roded Sharan, and Eytan Ruppín. Competitive and cooperative metabolic interactions in bacterial communities. *Nature communications*, 2:589, dec 2011.
- [56] Jonathan Friedman, Logan M Higgins, and Jeff Gore. Community structure follows simple assembly rules in microbial microcosms. *Nature Ecology & Evolution*, 1(March), 2017.
- [57] Georg Fuchs. Alternative Pathways of Carbon Dioxide Fixation: Insights into the Early Evolution of Life? *Annual Review of Microbiology*, 65(1):631–658, oct 2011.
- [58] Amit Ghosh, Jerome Nilmeier, Daniel Weaver, Paul D Adams, Jay D Keasling, Aindrila Mukhopadhyay, Christopher J Petzold, and Héctor García Martín. A peptide-based method for <sup>13</sup>C Metabolic Flux Analysis in microbial communities. *PLoS computational biology*, 10(9):e1003827, sep 2014.
- [59] Johann Peter Gogarten and David Deamer. Is LUCA a thermophilic progenote? *1(12):16229*, nov 2016.
- [60] Nigel Goldenfeld and Carl Woese. Life is physics: evolution as a collective phenomenon far from equilibrium. *Annual Review of Condensed Matter Physics*, 2:375–399, 2010.
- [61] Joshua E. Goldford, Hyman Hartman, Temple F. Smith, and Daniel Segrè. Remnants of an Ancient Metabolism without Phosphate. *Cell*, 168(6):1126–1134.e9, 2017.

- [62] Joshua E. Goldford and Igor G L Libourel. Unsupervised Identification of Isotope-Labeled Peptides. *Analytical Chemistry*, 88(11):6092–6099, 2016.
- [63] Joshua E Goldford, Nanxi Lu, Djordje Bajic, Sylvie Estrela, Mikhail Tikhonov, Alicia Sanchez-Gorostiaga, Daniel Segre, Pankaj Mehta, and Alvaro Sanchez. Emergent Simplicity in Microbial Community Assembly. *bioRxiv*, oct 2017.
- [64] Joshua E. Goldford and Daniel Segrè. Modern views of ancient metabolic networks. *Current Opinion in Systems Biology*, 8:117–124, apr 2018.
- [65] Aaron David Goldman, Tess M Bernhard, Egor Dolzhenko, and Laura F Landweber. LU-CApedia: a database for the study of ancient life. *Nucleic acids research*, 41(Database issue):D1079–82, jan 2013.
- [66] Maçha Gorlero, Rafal Wieczorek, Katarzyna Adamala, Alessandra Giorgi, Maria Eugenia Schininà, Pasquale Stano, and Pier Luigi Luisi. Ser-His catalyses the formation of peptides and PNAs. *FEBS Letters*, 583(1):153–156, jan 2009.
- [67] Gerhard Gottschalk. *Bacterial Metabolism*. Springer Series in Microbiology. Springer US, New York, NY, 1979.
- [68] Jacopo Grilli, György Barabás, Matthew J. Michalska-Smith, and Stefano Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666):210–213, jul 2017.
- [69] Elizaveta Guseva, Ronald N Zuckermann, and Ken A Dill. Foldamer hypothesis for the growth and sequence differentiation of prebiotic polymers. *Proceedings of the National Academy of Sciences*, 114(36):E7460–E7468, sep 2017.
- [70] I. Halevy and A. Bachan. The geologic history of seawater pH. *Science*, 355(6329):1069–1071, 2017.
- [71] D. O. HALL, R. CAMMACK, and K. K. RAO. Role for Ferredoxins in the Origin of Life and Biological Evolution. *Nature*, 233(5315):136–138, sep 1971.
- [72] M Halmann. Evolution and Ecology of Phosphorus Metabolism. In K. Dose, S. W. Fox, G. A. Deborin, and T. E. Pavlovskaya, editors, *The Origin of Life and Evolutionary Biochemistry*, chapter Evolution, pages 169–182. Springer US, 1974.
- [73] Thomas Handorf, Oliver Ebenhöf, and Reinhart Heinrich. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *Journal of molecular evolution*, 61(4):498–512, oct 2005.
- [74] Susse Kerkelund Hansen, Paul B Rainey, Janus A J Haagensen, and Soren Molin. Evolution of species interactions in a biofilm community. *Nature*, 445(7127):533–536, feb 2007.
- [75] William R Harcombe, William J Riehl, Ilija Dukovski, Brian R Granger, Alex H Lang, Gracia Bonilla, Amrita Kar, Nicholas Leiby, Pankaj Mehta, Christopher J Marx, and Daniel Segrè. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports*, 7(4):1104–1115, 2014.

- [76] H Hartman. Conjectures and reveries. *Photosynthesis research*, 33(2):171–6, aug 1992.
- [77] Hyman Hartman. Speculations on the origin and evolution of metabolism. *Journal of Molecular Evolution*, 4(4):359–370, 1975.
- [78] Hyman Hartman. Conjectures and reveries. *Photosynthesis research*, 2(33):171–176, 1992.
- [79] Robert M Hazen and Dimitri A Sverjensky. Mineral surfaces, geochemical complexities, and the origins of life. *Cold Spring Harbor perspectives in biology*, 2(5):a002162, may 2010.
- [80] Jan-Hendrik Hehemann, Philip Arevalo, Manoshi S. Datta, Xiaoqian Yu, Christopher H. Corzett, Andreas Henschel, Sarah P. Preheim, Sonia Timberlake, Eric J. Alm, and Martin F. Polz. Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. *Nature Communications*, 7:12860, 2016.
- [81] Doeke R Hekstra and Stanislas Leibler. Contingency and statistical laws in replicate microbial closed ecosystems. *Cell*, 149(5):1164–73, may 2012.
- [82] K. Hemavathi, M. Kalaivani, A. Udayakumar, G. Sowmiya, J. Jeyakanthan, and K. Sekar. MIPS: metal interactions in protein structures. *Journal of Applied Crystallography*, 43(1):196–199, dec 2009.
- [83] Christopher S Henry, Linda J Broadbelt, and Vassily Hatzimanikatis. Thermodynamics-based metabolic flux analysis. *Biophysical journal*, 92(5):1792–1805, mar 2007.
- [84] Jordan M Horowitz and Jeremy L England. Spontaneous fine-tuning to environment in many-species chemical reaction networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(29):201700617, 2017.
- [85] The Human Microbiome Project Consortium, Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, Ashlee M. Earl, Michael G. FitzGerald, Robert S. Fulton, Michelle G. Giglio, Kymberlie Hallsworth-Pepin, Elizabeth A. Lobos, Ramana Madupu, Vincent Magrini, John C. Martin, Makedonka Mitreva, Donna M. Muzny, Erica J. Sodergren, James Versalovic, Aye M. Wollam, Kim C. Worley, Jennifer R. Wortman, Sarah K. Young, Qiandong Zeng, Kjersti M. Aagaard, Olukemi O. Abolude, Emma Allen-Vercoe, Eric J. Alm, Lucia Alvarado, Gary L. Andersen, Scott Anderson, Elizabeth Appelbaum, Harindra M. Arachchi, Gary Armitage, Cesar A. Arze, Tulin Ayvaz, Carl C. Baker, Lisa Begg, Tsegahiwot Belachew, Veena Bhongari, Monika Bihan, Martin J. Blaser, Toby Bloom, Vivien Bonazzi, J. Paul Brooks, Gregory A. Buck, Christian J. Buhay, Dana A. Busam, Joseph L. Campbell, Shane R. Canon, Brandi L. Cantarel, Patrick S. G. Chain, I-Min A. Chen, Lei Chen, Shaila Chhibba, Ken Chu, Dawn M. Ciulla, Jose C. Clemente, Sandra W. Clifton, Sean Conlan, Jonathan Crabtree, Mary A. Cutting, Noam J. Davidovics, Catherine C. Davis, Todd Z. DeSantis, Carolyn Deal, Kimberley D. Delehaunty, Floyd E. Dewhirst, Elena Deych, Yan Ding, David J. Dooling, Shannon P. Dugan, Wm Michael Dunne, A. Scott Durkin, Robert C. Edgar, Rachel L. Erlich, Candace N. Farmer, Ruth M. Farrell, Karoline Faust, Michael Feldgarden, Victor M. Felix, Sheila Fisher, Anthony A. Fodor, Larry J. Forney, Leslie Foster, Valentina Di

Francesco, Jonathan Friedman, Dennis C. Friedrich, Catrina C. Fronick, Lucinda L. Fulton, Hongyu Gao, Nathalia Garcia, Georgia Giannoukos, Christina Giblin, Maria Y. Giovanni, Jonathan M. Goldberg, Johannes Goll, Antonio Gonzalez, Allison Griggs, Sharvari Gujja, Susan Kinder Haake, Brian J. Haas, Holli A. Hamilton, Emily L. Harris, Theresa A. Hepburn, Brandi Herter, Diane E. Hoffmann, Michael E. Holder, Clinton Howarth, Katherine H. Huang, Susan M. Huse, Jacques Izard, Janet K. Jansson, Huaiyang Jiang, Catherine Jordan, Vandita Joshi, James A. Katancik, Wendy A. Keitel, Scott T. Kelley, Cristyn Kells, Nicholas B. King, Dan Knights, Heidi H. Kong, Omry Koren, Sergey Koren, Karthik C. Kota, Christie L. Kovar, Nikos C. Kyrpides, Patricio S. La Rosa, Sandra L. Lee, Katherine P. Lemon, Niall Lennon, Cecil M. Lewis, Lora Lewis, Ruth E. Ley, Kelvin Li, Konstantinos Liolios, Bo Liu, Yue Liu, Chien-Chi Lo, Catherine A. Lozupone, R. Dwayne Lunsford, Tessa Madden, Anup A. Mahurkar, Peter J. Mannon, Elaine R. Mardis, Victor M. Markowitz, Konstantinos Mavromatis, Jamison M. McCorrison, Daniel McDonald, Jean McEwen, Amy L. McGuire, Pamela McInnes, Teena Mehta, Kathie A. Mihindukulasuriya, Jason R. Miller, Patrick J. Minx, Irene Newsham, Chad Nusbaum, Michelle O’Laughlin, Joshua Orvis, Ioanna Pagani, Krishna Palaniappan, Shital M. Patel, Matthew Pearson, Jane Peterson, Mircea Podar, Craig Pohl, Katherine S. Pollard, Mihai Pop, Margaret E. Priest, Lita M. Proctor, Xiang Qin, Jeroen Raes, Jacques Ravel, Jeffrey G. Reid, Mina Rho, Rosamond Rhodes, Kevin P. Riehle, Maria C. Rivera, Beltran Rodriguez-Mueller, Yu-Hui Rogers, Matthew C. Ross, Carsten Russ, Ravi K. Sanka, Pamela Sankar, J. Fah Sathirapongsasuti, Jeffery A. Schloss, Patrick D. Schloss, Thomas M. Schmidt, Matthew Scholz, Lynn Schriml, Alyxandria M. Schubert, Nicola Segata, Julia A. Segre, William D. Shannon, Richard R. Sharp, Thomas J. Sharpton, Narmada Shenoy, Nihar U. Sheth, Gina A. Simone, Indresh Singh, Christopher S. Smillie, Jack D. Sobel, Daniel D. Sommer, Paul Spicer, Granger G. Sutton, Sean M. Sykes, Diana G. Tabbaa, Mathangi Thiagarajan, Chad M. Tomlinson, Manolito Torralba, Todd J. Treangen, Rebecca M. Truty, Tatiana A. Vishnivetskaya, Jason Walker, Lu Wang, Zhengyuan Wang, Doyle V. Ward, Wesley Warren, Mark A. Watson, Christopher Wellington, Kris A. Wetterstrand, James R. White, Katarzyna Wilczek-Boney, Yuan-Qing Wu, Kristine M. Wylie, Todd Wylie, Chandri Yandava, Liang Ye, Yuzhen Ye, Shibu Yooseph, Bonnie P. Youmans, Lan Zhang, Yanjiao Zhou, Yiming Zhu, Laurie Zoloth, Jeremy D. Zucker, Bruce W. Birren, Richard A. Gibbs, Sarah K. Highlander, Barbara A. Methé, Karen E. Nelson, Joseph F. Petrosino, George M. Weinstock, Richard K. Wilson, Owen White, and The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

- [86] Benjamin I. Jelen, Donato Giovannelli, and Paul G. Falkowski. The Role of Microbial Electron Transfer in the Coevolution of the Biosphere and Geosphere. *Annual Review of Microbiology*, 70(1):annurev-micro-102215-095521, 2016.
- [87] Adrian Jinich, Dmitriy Rappoport, Ian Dunn, Benjamin Sanchez-Lengeling, Roberto Olivares-Amaya, Elad Noor, Arren Bar Even, and Alán Aspuru-Guzik. Quantum Chemical Approach to Estimating the Thermodynamics of Metabolic Reactions. *Scientific Reports*, 4:7022, nov 2014.
- [88] B. Kacar, V. Hanson-Smith, Z. R. Adam, and N. Boekelheide. Constraining the timing of the Great Oxidation Event within the Rubisco phylogenetic tree. *Geobiology*, 15(5):628–640,

sep 2017.

- [89] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, jan 2000.
- [90] S A Kauffman. Autocatalytic sets of proteins. *Journal of theoretical biology*, 119(1):1–24, mar 1986.
- [91] Stuart A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*, 1993.
- [92] A D Keefe and S L Miller. Are polyphosphates or phosphate esters prebiotic reagents? *Journal of molecular evolution*, 41(6):693–702, dec 1995.
- [93] A D Keefe, G L Newton, and S L Miller. A possible prebiotic synthesis of pantetheine, a precursor to coenzyme A. *Nature*, 373(6516):683–5, feb 1995.
- [94] Markus A. Keller, Domen Kampjut, Stuart A. Harrison, and Markus Ralser. Sulfate radicals enable a non-enzymatic Krebs cycle precursor. *Nature Ecology & Evolution*, 1(4):0083, 2017.
- [95] Markus A. Keller, Gabriel Piedrafita, and Markus Ralser. The widespread role of non-enzymatic reactions in cellular metabolism. *Current Opinion in Biotechnology*, 34:153–161, 2015.
- [96] Markus A Keller, Alexandra V Turchyn, and Markus Ralser. Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. *Molecular systems biology*, 10(4):725, apr 2014.
- [97] Markus A. Keller, Andre Zylstra, Cecilia Castro, Alexandra V. Turchyn, Julian L. Griffin, and Markus Ralser. Conditional iron and pH-dependent activity of a non-enzymatic glycolysis and pentose phosphate pathway. *Science advances*, 2(1):e1501235, 2016.
- [98] Eric D. Kelsic, Jeffrey Zhao, Kalin Vetsigian, and Roy Kishony. Counteraction of antibiotic production and degradation stabilizes microbial communities. *Nature*, 521:516–519, 2015.
- [99] Maria Khomyakova, Özlem Bükmez, Lorenz K Thomas, Tobias J Erb, and Ivan A Berg. A methylaspartate cycle in haloarchaea. *Science (New York, N.Y.)*, 331(6015):334–7, jan 2011.
- [100] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. PubChem Substance and Compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, jan 2016.
- [101] G A King. Evolution of the coenzymes. *Bio Systems*, 13(1-2):23–45, jan 1980.
- [102] Grant Kinsler, Sam Sinai, Nicholas Keone Lee, and Martin A. Nowak. Prebiotic selection for motifs in a model of template-free elongation of polymers within compartments. *PLoS one*, 12(7):e0180208, 2017.

- [103] Susan Q. Lang, David A. Butterfield, Mitch Schulte, Deborah S. Kelley, and Marvin D. Lilley. Elevated concentrations of formate, acetate and dissolved organic carbon found at the Lost City hydrothermal field. *Geochimica et Cosmochimica Acta*, 74(3):941–952, feb 2010.
- [104] Morgan G I Langille, Jesse Zaneveld, J Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A Reyes, Jose C Clemente, Deron E Burkepile, Rebecca L Vega Thurber, Rob Knight, Robert G Beiko, and Curtis Huttenhower. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology*, 31(9):814–821, sep 2013.
- [105] Pierre Laszlo. Chemical Reactions on Clays. *Science (New York, N.Y.)*, 235(4795):1473–1477, 1987.
- [106] Paola Laurino and Dan S. Tawfik. Spontaneous Emergence of S-Adenosylmethionine and the Evolution of Methylation. *Angewandte Chemie - International Edition*, 56(1):343–345, 2017.
- [107] A Lazcano and S L Miller. The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. *Cell*, 85(6):793–8, jun 1996.
- [108] Bruce R Levin. Coexistence of Two Asexual Strains on a Single Resource. *Science*, 175(4027):1272 LP – 1274, mar 1972.
- [109] Jonathan M Levine, Jordi Bascompte, Peter B Adler, and Stefano Allesina. Beyond pairwise mechanisms of species coexistence in complex communities. *Nature*, 546(7656):56–64, jun 2017.
- [110] Gene-Wei Li, David Burkhardt, Carol Gross, Jonathan S Weissman, and Jonathan S Weissman. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell*, 157(3):624–635, apr 2014.
- [111] S. Louca, L. W. Parfrey, and M. Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, sep 2016.
- [112] Stilianos Louca, Saulo M S Jacques, Aliny P F Pires, Juliana S Leal, Diane S Srivastava, Laura Wegener Parfrey, Vinicius F Farjalla, and Michael Doebeli. High taxonomic variability despite stable functional structure across microbial communities. *Nature ecology & evolution*, 1(1):15, dec 2016.
- [113] Stilianos Louca, Martin F. Polz, Florent Mazel, Michaeline B. N. Albright, Julie A. Huber, Mary I. O’Connor, Martin Ackermann, Aria S. Hahn, Diane S. Srivastava, Sean A. Crowe, Michael Doebeli, and Laura Wegener Parfrey. Function and functional redundancy in microbial systems. *Nature Ecology & Evolution*, page 1, apr 2018.
- [114] R MacArthur. Species packing and competitive equilibrium for many species. *Theoretical population biology*, 1(1):1–11, may 1970.

- [115] Robert MacArthur and Richard Levins. Competition, habitat selection, and character displacement in a patch environment. *Proceedings of the National Academy of Sciences*, 51(6):1207–1210, jun 1964.
- [116] Robert Macarthur and Richard Levins. The Limiting Similarity, Convergence, and Divergence of Coexisting Species. *The American Naturalist*, 101(921):377, 1967.
- [117] Achim Mall, Jessica Sobotta, Claudia Huber, Carolin Tschirner, Stefanie Kowarschik, Katarina Bačnik, Mario Mergelsberg, Matthias Boll, Michael Hügler, Wolfgang Eisenreich, and Ivan A. Berg. Reversibility of citrate synthase allows autotrophic growth of a thermophilic bacterium. *Science*, 359(6375):563–567, 2018.
- [118] Dominic E Mandy, Joshua E Goldford, Hong Yang, Doug K Allen, and Igor G L Libourel. Metabolic flux analysis using C peptide label measurements. *The Plant journal : for cell and molecular biology*, 77(3):476–86, feb 2014.
- [119] Victor M Markowitz, I-Min A Chen, Krishna Palaniappan, Ken Chu, Ernest Szeto, Yuri Grechkin, Anna Ratner, Biju Jacob, Jinghua Huang, Peter Williams, Marcel Huntemann, Iain Anderson, Konstantinos Mavromatis, Natalia N Ivanova, and Nikos C Kyrpides. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic acids research*, 40(Database issue):D115–22, jan 2012.
- [120] William Martin, John Baross, Deborah Kelley, and Michael J Russell. Hydrothermal vents and the origin of life. *Nature reviews. Microbiology*, 6(11):805–14, nov 2008.
- [121] William Martin and Michael J Russell. On the origin of biochemistry at an alkaline hydrothermal vent. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1486):1887–1926, oct 2007.
- [122] William F. Martin. Hydrogen, metals, bifurcating electrons, and proton gradients: The early evolution of biological energy conservation. *FEBS Letters*, 586(5):485–493, 2012.
- [123] William F. Martin and Rudolf K. Thauer. Energy in Ancient Metabolism. 168(6):953–955, mar 2017.
- [124] Adam C. Martiny, Amos P.K. Tai, Daniele Veneziano, François Primeau, and Sallie W. Chisholm. Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus. *Environmental Microbiology*, 11(4):823–832, 2009.
- [125] Jennifer B H Martiny, Stuart E. Jones, Jay T. Lennon, and Adam C. Martiny. Microbiomes in light of traits: A phylogenetic perspective. *Science*, 350(6261), 2015.
- [126] Markus Meringer and H. James Cleaves. Computational exploration of the chemical structure space of possible reverse tricarboxylic acid cycle constituents. *Scientific Reports*, 7(1):17540, dec 2017.
- [127] Christoph B. Messner, Paul C. Driscoll, Gabriel Piedrafita, Michael F. L. De Volder, and Markus Ralser. Nonenzymatic gluconeogenesis-like formation of fructose 1,6-bisphosphate in ice. *Proceedings of the National Academy of Sciences*, 114(28):7403–7407, 2017.

- [128] D E Metzler and E.E. Snell. Deamination of serine. I. Catalytic deamination of serine and cysteine by pyridoxal and metal salts. *The Journal of biological chemistry*, 198(1):353–61, sep 1952.
- [129] E James Milner-White and Michael J Russell. Functional capabilities of the earliest peptides and the emergence of life. *Genes*, 2(4):671–88, jan 2011.
- [130] Ron Milo, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic acids research*, 38(Database issue):D750–3, jan 2010.
- [131] Boris G Mirkin, Trevor I Fenner, Michael Y Galperin, and Eugene V Koonin. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC evolutionary biology*, 3:2, jan 2003.
- [132] Eli K. Moore, Benjamin I. Jelen, Donato Giovannelli, Hagai Raanan, and Paul G. Falkowski. Metal availability and the expanding network of microbial metabolisms in the Archaean eon. *Nature Geoscience*, 10(9):629–636, aug 2017.
- [133] H J Morowitz, J D Kostelnik, J Yang, and G D Cody. The origin of intermediary metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 97(14):7704–8, jul 2000.
- [134] Harold J. Morowitz, Vijayarathy Srinivasan, and Eric Smith. Ligand Field Theory and the Origin of Life as an Emergent Feature of the Periodic Table of Elements. *The Biological Bulletin*, 219(1):1–6, aug 2010.
- [135] Kamila B. Muchowska, Sreejith J. Varma, Elodie Chevillot-Beroux, Lucas Lethuillier-Karl, Guang Li, and Joseph Moran. Metals promote sequences of the reverse Krebs cycle. *Nature Ecology and Evolution*, 1(11):1716–1721, nov 2017.
- [136] Rafael Navarro-González, Christopher P. McKay, and Delphine Nna Mvondo. A possible nitrogen crisis for Archaean life due to reduced nitrogen fixation by lightning. *Nature*, 412(6842):61–64, 2001.
- [137] Wolfgang Nitschke and Michael J Russell. Beating the acetyl coenzyme A-pathway to the origin of life. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1622):20120258, jul 2013.
- [138] Elad Noor, Hulda S Haraldsdóttir, Ron Milo, and Ronan M T Fleming. Consistent estimation of Gibbs energy using component contributions. *PLoS computational biology*, 9(7):e1003098, jan 2013.
- [139] Yehor Novikov and Shelley D Copley. Reactivity landscape of pyruvate under simulated hydrothermal vent conditions. *Proceedings of the National Academy of Sciences of the United States of America*, 110(33):13283–8, aug 2013.

- [140] Takuro Nunoura, Yoshito Chikaraishi, Rikihisa Izaki, Takashi Suwa, Takaaki Sato, Takeshi Harada, Koji Mori, Yumiko Kato, Masayuki Miyazaki, Shigeru Shimamura, Katsunori Yanagawa, Aya Shuto, Naohiko Ohkouchi, Nobuyuki Fujita, Yoshihiro Takaki, Haruyuki Atomi, and Ken Takai. A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile. *Science*, 359(6375):559–563, feb 2018.
- [141] Edward J O’Brien, Jonathan M Monk, and Bernhard O Palsson. Using Genome-scale Models to Predict Biological Capabilities. *Cell*, 161(5):971–987, sep 2016.
- [142] H Ochman, J G Lawrence, and E A Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, may 2000.
- [143] Atsushi Ohta, Hiroshi Murakami, and Hiroaki Suga. Polymerization of alpha-hydroxy acids by ribosomes. *Chembiochem : a European journal of chemical biology*, 9(17):2773–2778, nov 2008.
- [144] Jeffrey D Orth, Ines Thiele, and Bernhard O Palsson. What is flux balance analysis? *Nat Biotech*, 28(3):245–248, mar 2010.
- [145] P. G. Falkowski, T. Fenchel, and E. F. Delong. The Microbial Engines That Drive Earth’s Biogeochemical Cycles. *Science*, 320(5879):1034–1039, 2008.
- [146] Nicole Paczia, Anke Nilgen, Tobias Lehmann, Jochem Gatgens, Wolfgang Wiechert, and Stephan Noack. Extensive exometabolome analysis reveals extended overflow metabolism in various microorganisms. *Microbial cell factories*, 11:122, sep 2012.
- [147] Eric T Parker, Henderson J Cleaves, Jason P Dworkin, Daniel P Glavin, Michael Callahan, Andrew Aubrey, Antonio Lazcano, and Jeffrey L Bada. Primordial synthesis of amines and amino acids in a 1958 Miller H<sub>2</sub>S-rich spark discharge experiment. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14):5526–31, apr 2011.
- [148] Matthew A Pasek. Rethinking early Earth phosphorus geochemistry. *Proceedings of the National Academy of Sciences of the United States of America*, 105(3):853–858, jan 2008.
- [149] Matthew A Pasek, Jelte P Harnmeijer, Roger Buick, Maheen Gull, and Zachary Atlas. Evidence for reactive reduced phosphorus species in the early Archean ocean. *Proceedings of the National Academy of Sciences of the United States of America*, 110(25):10089–94, jun 2013.
- [150] Bhavesh H. Patel, Claudia Percivalle, Dougal J. Ritson, Colm D. Duffy, and John D. Sutherland. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nature Chemistry*, 7(4):301–307, mar 2015.
- [151] Sandra Pizzarello and Arthur L. Weber. Prebiotic Amino Acids as Asymmetric Catalysts. *Science*, 303(5661):1151, 2004.
- [152] Germán Plata, Christopher S Henry, and Dennis Vitkup. Long-term phenotypic evolution of bacteria. *Nature*, 517(7534):369–372, jan 2015.

- [153] Anna Posfai, Thibaud Taillefumier, and Ned S. Wingreen. Metabolic Trade-Offs Promote Diversity in a Model Ecosystem. *Physical Review Letters*, 118(2):1–5, 2017.
- [154] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(Database issue):D590–D596, jan 2013.
- [155] Paul B Rainey and Michael Travisano. Adaptive radiation in a heterogeneous environment. *Nature*, 394(6688):69–72, jul 1998.
- [156] Steen Rasmussen, Adi Constantinescu, and Carsten Svaneborg. Generating minimal living systems from non-living materials and increasing their evolutionary abilities. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 371(1701):20150440–, aug 2016.
- [157] Christoph Ratzke, Jonas Denk, and Jeff Gore. Ecological suicide in microbes. *Nature Ecology & Evolution*, 2(5):867–872, may 2018.
- [158] Christoph Ratzke and Jeff Gore. Modifying and reacting to the environmental pH can drive bacterial interactions. *PLoS Biology*, 16(3):e2004248, mar 2018.
- [159] Horst Rauchfuss. *Chemical Evolution and the Origin of Life*. Springer Science {&} Business Media, 2008.
- [160] Jason Raymond and Daniel Segrè. The effect of oxygen on biochemical networks and the evolution of complex life. *Science (New York, N.Y.)*, 311(5768):1764–7, mar 2006.
- [161] Jason Raymond, Janet L. Siefert, Christopher R. Staples, and Robert E. Blankenship. The Natural History of Nitrogen Fixation. *Molecular Biology and Evolution*, 21(3):541–554, 2004.
- [162] António J M Ribeiro, Gemma L Holliday, Nicholas Furnham, Jonathan D Tyzack, Katherine Ferris, and Janet M Thornton. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research*, 46(D1):D618–D623, jan 2018.
- [163] William J. Riehl, Paul L. Krapivsky, Sidney Redner, and Daniel Segrè. Signatures of arithmetic simplicity in metabolic network architecture. *PLoS Computational Biology*, 6(4), 2010.
- [164] Francisco Rodriguez-Valera, Ana-Belen Martin-Cuadrado, Beltran Rodriguez-Brito, Lejla Pasić, T Frede Thingstad, Forest Rohwer, and Alex Mira. Explaining microbial population genomics through phage predation. *Nature reviews. Microbiology*, 7(11):828–36, 2009.
- [165] R F Rosenzweig, R R Sharp, D S Treves, and J Adams. Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*. *Genetics*, 137(4):903 – 917, aug 1994.

- [166] Kepa Ruiz-Mirazo, Carlos Briones, and Andrés de la Escosura. Prebiotic Systems Chemistry: New Perspectives for the Origins of Life. *Chemical Reviews*, 114(1):285–366, jan 2014.
- [167] M J Russell, A J Hall, and W Martin. Serpentinization as a source of energy at the origin of life. *Geobiology*, 8(5):355–71, dec 2010.
- [168] Sumantra Sarkar, Bin Wang, and Jeremy L. England. Design of conditions for emergence of self-replicators. sep 2017.
- [169] Alan W Schwartz. Phosphorus in prebiotic chemistry. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 361(1474):1743–9, oct 2006.
- [170] D Segré, D Ben-Eli, D W Deamer, and D Lancet. The lipid world. *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life*, 31(1-2):119–45.
- [171] D Segré, D Ben-Eli, and D Lancet. Compositional genomes: prebiotic information transfer in mutually catalytic noncovalent assemblies. *Proceedings of the National Academy of Sciences*, 97(8):4112–7, 2000.
- [172] Daniel Segrè, Niels Klitgord, and Daniel Segrè. Environments that Induce Synthetic Microbial Ecosystems. 6(11), 2010.
- [173] Sergey N. Semenov, Lewis J. Kraft, Alar Ainla, Mengxia Zhao, Mostafa Baghbanzadeh, Victoria E. Campbell, Kyungtae Kang, Jerome M. Fox, and George M. Whitesides. Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature*, 537(7622):656–660, 2016.
- [174] Eric Smith and Harold J Morowitz. Universality in intermediary metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36):13168–73, sep 2004.
- [175] Eric. Smith and Harold J. Morowitz. *The Origin and Nature of Life On Earth*. Cambridge University Press, Cambridge, United Kingdom, 1st edition, 2016.
- [176] Filipa L Sousa, Wim Hordijk, Mike Steel, and William F Martin. Autocatalytic sets in E. coli metabolism. *Journal of systems chemistry*, 6(1):4, 2015.
- [177] Filipa L Sousa and William F Martin. Biochemical fossils of the ancient transition from geoenenergetics to bioenergetics in prokaryotic one carbon compound metabolism. *Biochimica et biophysica acta*, 1837(7):964–81, jul 2014.
- [178] Filipa L Sousa, Martina Preiner, and William F Martin. Native metals, electron bifurcation, and CO<sub>2</sub> reduction in early biochemical evolution. *Current opinion in microbiology*, 43:77–83, 2018.

- [179] Filipa L Sousa, Thorsten Thiergart, Giddy Landan, Shijulal Nelson-Sathi, Inês a C Pereira, John F Allen, Nick Lane, and William F Martin. Early bioenergetic evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1622):20130088, 2013.
- [180] Greg Springsteen, Jayasudhan Reddy Yerabolu, Julia Nelson, Chandler Joel Rhea, and Ramanarayanan Krishnamurthy. Linked cycles of oxidative decarboxylation of glyoxylate as protometabolic analogs of the citric acid cycle. *Nature Communications*, 9(1):91, dec 2018.
- [181] Vijayasarathy Srinivasan and Harold J Morowitz. The canonical network of autotrophic intermediary metabolism: minimal metabolome of a reductive chemoautotroph. *The Biological bulletin*, 216(2):126–30, apr 2009.
- [182] Frank M. Stewart and Bruce R. Levin. Partitioning of Resources and the Outcome of Interspecific Competition: A Model and Some General Considerations. *The American Naturalist*, 107(954):171–198, 1973.
- [183] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, Francisco M Cornejo-Castillo, Paul I Costea, Corinne Cruaud, Francesco D’Ovidio, Stefan Engelen, Isabel Ferrera, Josep M Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T Poulos, Marta Royo-Llonch, Hugo Sarmiento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stephane Pesant, Sabrina Speich, Lars Stemmann, Matthew B Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G Acinas, and Peer Bork. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, 2015.
- [184] John D. Sutherland. Opinion: Studies on the origin of life the end of the beginning. *Nature Reviews Chemistry*, 1:0012, 2017.
- [185] Thibaud Tallefumier, Anna Posfai, Yigal Meir, and Ned S. Wingreen. Microbial consortia at steady supply. *eLife*, 6:1–65, 2017.
- [186] David Tilman. *Resource competition and community structure*. Princeton University Press, 1982.
- [187] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, Michael Egholm, Bernard Henrissat, Andrew C Heath, Rob Knight, and Jeffrey I Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–4, jan 2009.
- [188] Sreejith J. Varma, Kamila B. Muchowska, Paul Chatelain, and Joseph Moran. Native iron reduces CO<sub>2</sub> to intermediates and end-products of the acetyl-CoA pathway. *Nature Ecology & Evolution*, 2, apr 2018.

- [189] Sreejith J. Varma, Kamila B. Muchowska, Paul Chatelain, and Joseph Moran. Native iron reduces CO<sub>2</sub> to intermediates and end-products of the acetyl-CoA pathway. *Nature Ecology & Evolution*, apr 2018.
- [190] Nicole M. Vega, Jeff Gore, M Janczyk, R Dale, JB Freeman, and G Casadei. Stochastic assembly produces heterogeneous communities in the *Caenorhabditis elegans* intestine. *PLOS Biology*, 15(3):e2000633, mar 2017.
- [191] Kalin Vetsigian, Carl Woese, and Nigel Goldenfeld. Collective evolution and the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, 103(28):10696–701, jul 2006.
- [192] G. Wachtershauser. Evolution of the first metabolic cycles. *Proceedings of the National Academy of Sciences*, 87(1):200–204, jan 1990.
- [193] G Wächtershäuser. Evolution of the first metabolic cycles. *Proceedings of the National Academy of Sciences of the United States of America*, 87(1):200–204, jan 1990.
- [194] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy . *Applied and Environmental Microbiology*, 73(16):5261–5267, aug 2007.
- [195] Madeline C. Weiss, Sinje Neukirchen, Mayo Roettger, Natalia Mrnjavac, Shijulal Nelson-Sathi, William F. Martin, and Filipa L. Sousa. Reply to 'Is LUCA a thermophilic progenote?'. 1(12):16230, nov 2016.
- [196] Madeline C Weiss, Filipa L Sousa, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-sathi, and William F Martin. The physiology and habitat of the last universal common ancestor. *Nature Microbiology*, 1(July):1–8, 2016.
- [197] Rafal Wieczorek, Katarzyna Adamala, Tecla Gasperi, Fabio Polticelli, and Pasquale Stano. Small and Random Peptides: An Unexplored Reservoir of Potentially Functional Primitive Organocatalysts. The Case of Seryl-Histidine. *Life*, 7(2):19, 2017.
- [198] Tom A. Williams, Peter G. Foster, Cymon J. Cox, and T. Martin Embley. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, 504(7479):231–236, 2013.
- [199] Ruixin Zhou and Marcelo I. Guzman. Photocatalytic Reduction of Fumarate to Succinate on ZnS Mineral Surfaces. *The Journal of Physical Chemistry C*, 120(13):7349–7357, apr 2016.

## **Curriculum Vitae**

