Biochemical Networks Across Planets and Scales

by

Harrison Brodsky Smith

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2018 by the
Graduate Supervisory Committee:

Sara I. Walker, Chair
Ariel D. Anbar
Michael R. Line
Jordan G. Okie
Stephen J. Romaniello

ARIZONA STATE UNIVERSITY

December 2018

ABSTRACT

Biochemical reactions underlie all living processes. Their complex web of interactions is difficult to fully capture and quantify with simple mathematical objects. Applying network science to biology has advanced our understanding of the metabolisms of individual organisms and the organization of ecosystems, but has scarcely been applied to life at a planetary scale. To characterize planetary-scale biochemistry, I constructed biochemical networks using global databases of annotated genomes and metagenomes, and biochemical reactions. I uncover scaling laws governing biochemical diversity and network structure shared across levels of organization from individuals to ecosystems, to the biosphere as a whole. Comparing real biochemical reaction networks to random reaction networks reveals the observed biological scaling is not a product of chemistry alone, but instead emerges due to the particular structure of selected reactions commonly participating in living processes. I perform distinguishability tests across properties of individual and ecosystem-level biochemical networks to determine whether or not they share common structure, indicative of common generative mechanisms across levels. My results indicate there is no sharp transition in the organization of biochemistry across distinct levels of the biological hierarchy—a result that holds across different network projections.

Finally, I leverage these large biochemical datasets, in conjunction with planetary observations and computational tools, to provide a methodological foundation for the quantitative assessment of biology's viability amongst other geospheres. Investigating a case study of alkaliphilic prokaryotes in the context of Enceladus, I find that the chemical compounds observed on Enceladus thus far would be insufficient to allow even these extremophiles to produce the compounds necessary to sustain a viable metabolism. The environmental precursors required by these organisms provides a reference for the compounds which should be prioritized for detection in future planetary exploration missions. The results of this framework have further consequences in the context of planetary protection, and hint that forward contamination may prove infeasible without meticulous intent. Taken together these results point to a deeper level of organization in biochemical networks than what has been understood so far, and suggests the existence of common organizing principles operating across different levels of biology and planetary chemistry.

DEDICATION

Dedicated to my grandpa—your jocular disposition and appreciation of nature were an inspiration to me, and they continue to inspire me every day.

ACKNOWLEDGMENTS

There are many people to whom I owe thanks for helping get me to the end of my PhD: My mom and dad, for always encouraging me to pursue my interests with unwavering support; my brother, for the much needed distractions of lighthearted articles and games; Melissa, who was always down to talk about life, science, and field my weird biochem questions; everyone in eLife/Elife/ELife/ELIFE—but special thanks to. . . Doug, an invaluable resource not only for his intellect but his patience; Cole, who was a sounding board for my crazy ideas and shared my love of music and the outdoors; Jake, who stoked my entrepreneurial and artistic spirits; Tucker, who was an enigma; Hyunju, who embraced continually editing and improving our work; and of course Sara, whose curiosity and positivity were infectious—and who brought me to, and sent me to, conferences and workshops of all subjects, and never had to be convinced of the value of that sort of thing.

Shout out to all the coffee shops (and the baristas therein) that caffeinated me through my PhD... Gold Bar, where I came often at the start of my PhD to code and read; Cartel and Sip, where I came to work in mornings and afternoons; The Gelato Spot, where I wrote large parts of my third and fourth chapters late at night; and Royal, where I was a regular every week for most of graduate school.

Lastly, thanks to my partner Liz who has supported me in every way imaginable through most of my PhD—be it with delectable cooking, words of encouragement, or planning fun nights out after long weeks, just to name a few of the ways.

TABLE OF CONTENTS

CHAPTER Page

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

There is increasing interest in whether biology is governed by general principles, not tied to its specific chemical instantiation or contingent upon its evolutionary history [1–3]. Such principles would be strong candidates for being universal to all life [4, 5]. Universal biology, if it exists, would have important implications for our search for life beyond Earth [6–9], for engineering synthetic life in the lab [10, 11], and for solving the origin of life [12, 13]. Systems biology provides promising quantitative tools for uncovering such general principles [14–16]. So far, systems approaches have primarily focused on specific levels of organization within biological hierarchies, such as individual organisms [17, 18] or ecological communities [19, 20], and are rarely applied to the biosphere as a whole. But, biology exhibits some of its most striking regularities moving up in levels of organization from individuals to ecosystems, and these regularities may only truly manifest at the level of ecosystems, and ultimately the biosphere [21, 22]. For example, while individual organismal lineages fluctuate through time and space, the functional and metabolic composition of ecological communities is stable [23, 24]. To understand the general principles governing biology, we must understand how living systems organize across levels, not just within a given level [25–27].

But what is the best way to quantitatively describe the relationship between these living systems existing in embedded hierarchical levels? By what means can we capture how an individual cell's biochemical reactions interface with neighboring cells, and the cellular communities that intertwine to sustain the biosphere?

One option (natural to a physicist) would be to use statistical mechanics. Statistical mechanics was developed in the 19th century for studying and predicting the behavior of systems with many components. It has been hugely successful in its application to those physical systems well-approximated by idealized models of non-interacting particles. However, real-world systems are often much more complex, leading to a realization over the last several decades that new statistical approaches are necessary to describe biological and technological systems. Among the most logical mathematical frameworks for developing the necessary formalism is network theory, which projects the complex set of interactions composing real systems onto an abstract graph representation [16, 17, 28–37]. Such representations are powerful in their capacity to quantitatively describe the relationship between components of complex systems and because they permit inferring function and dynamics from structure [38–42].

Here, a *network* (or alternatively a *graph*) is simply a mathematical structure which can capture the relationship between components in a system [32, 34]. There are two building blocks of a network, the *nodes* and *edges*. Depending on what type of relationships we are hoping to describe, we can vary the components of our system represented by these node and edges. Perhaps the most canonical real-world example of a network is a social network. In this network, nodes often represent people, with edges connecting people who are friends. The algorithms developed around network theory then allow us to easily quantify properties of this social network which are of interest to researchers studying social systems.

For instance, say we are interested in how "cliquey" the physics department at ASU is—that is, does everyone tend to intermingle equally, or does there tend to be more mingling between small groups of people? First, we would want to obtain a list of all the students in the department. This defines all the nodes in our network, since each student is represented as a node. Then we would want to get a list of friends for each of those students. This defines all the edges in our network, since an edge in this case simply represents if two people are friends. Collectively, the arrangement of all the nodes and links in a network is called the network *topology*. Using a network clustering algorithm on the network topology, we could then estimate the number of clusters in the department. If there are many clusters, then we would say the department is cliquey.

As easily as we can represent a social system using a network formalism, we can represent a chemical system. Instead of nodes and edges denoting people and friendships as they do in a social network, a chemical system would denote nodes and edges as chemical compounds and shared reactions. For instance, in using network theory to describe a chemical system, two nodes would be connected if they constituted compounds on opposite sides of a reaction.

The simplicity of network theory's description of a system as complex as biochemistry is exactly why it is such a useful tool for probing the architecture of life. To fully characterize living chemical processes, the collective behavior of *reactions* must be understood—considering only *individual* components (molecules) is inadequate. The structure of how these components interact with one another via reactions is precisely what separates organized biological systems from messy chemical ones [13, 43, 44].

However, the problem of characterizing the structure of real-world systems is compounded by the fact there are often many ways to coarse-grain a real system to generate a network representation, each corresponding to a different way for a set of interactions to be projected onto a graph. For example, metabolic networks—describing the transformation of matter through catalysis of reactions—may be represented in different ways. In our initial chemical network example, nodes (compounds) were connected by edges (reactions) if they appeared on opposite sides of a reaction. This is referred to as a *unipartite* graph—'uni' because there is only one type of node. If we want to capture how enzymes fit into the network, we can represent enzymes and compounds as two disjoint node types—and now we are dealing with a *bipartite* network. Each of these different representations of the the same biochemical system are referred to as network *projections.* In the bipartite projection, we would connect compounds nodes to enzymes nodes if the enzyme requires or produces that compound as part of any reactions it catalyzes. We could also imagine representing both reactions and compounds as nodes in a bipartite network, connecting a reaction node to a compound node if it is involved in the reaction.

These graphs projections can have different large-scale topological properties, even when projected from the same underlying system [45–47]. The challenge of choosing a projection arises because biochemical networks are themselves a multi-layer system consisting of enzymes and their catalyzed reactions; enzymes (often abstracted away in network representations) control the biological organization we aim to characterize through reactions.

There is a rich body of literature which has made use of the fact that biochemical systems can be represented using network theory for describing the complexity of life. In general, this research has focused on a subset of biochemistry—such as metabolism—and a subset of the hierarchy of life—such as individuals or the biosphere.[48–55].

For this dissertation, since I am interested in properties universal across life, and not just subsets of living processes, I instead construct networks inclusive of every known catalyzed reaction (regardless of pathway) at the scale of individuals, ecosystems, and the biosphere as a whole.

In order to do this, I leverage two global databases of genomes and metagenomes, sampled from across life on Earth—the Pathosystems Resource Integration Center (PATRIC) [56], and the Joint Genome Institute's Integrated Microbial Genomes and Microbiomes database (JGI IMG/m) [57]. In conjunction with reaction data cataloged in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [58], it is possible to construct biochemical networks for each individual organism (genome) and ecosystem (metagenome). Building on prior work studying biosphere-level models of metabolism [53–55], I use the database of all $8,000+$ enzymatically catalyzed reactions cataloged in KEGG as a proxy for the biochemistry of the biosphere as a whole, modeled as a 'soup of enzymes' by disregarding the boundaries of individual species [19]. Network representations of ecosystem-level and biosphere-level biochemistry are 'compartment-free' in that no knowledge of individual species is included. Throughout each chapter of this dissertation, I utilize the above global databases of omic data and enzymatically catalyzed reaction data, in conjunction with network theory, to describe the structure of biology.

In the first chapter of this dissertation, I seek to determine whether biochemical networks display scaling laws governing their topology and chemical diversity which are similar across levels, indicative of the existence of self-organizing principles universal across life in the biosphere. A widely implemented framework for assessing commonality across different systems is to look at their scaling behavior [59–64]. If scale-invariant properties are found, it can be suggestive of deep, underlying organizing principles [3, 65, 66]. For example, organismal metabolic rates were observed to vary with organismal body mass to the $\frac{3}{4}$ power, even over enormous scales of 21 orders of magnitude [67]. Later a model based on fundamental physical principles, like minimizing energy dissipation in fractal circulation systems was found to reproduce this scaling [65]. Another example is how distinctive scaling laws emerge in critical systems [68].

The majority of network analyses applied to biochemistry have focused on the structure *within* individual metabolic network degree distributions [41, 48, 69] I instead focus on topological measures such as average shortest path length, average clustering coefficient and assortativity (degree correlation coefficient), which are directly comparable across different networks, allowing me to make statements about regularities existing across biochemistry sampled from different levels of organization in a manner that has not been possible in previous work focusing only on a given level.

I show that biochemical networks share universal organizational properties across levels, characterized by scaling laws determining how topology and biochemical diversity change with network size. These scaling relations exist independent of evolutionary domain or level of organization, applying across the nested hierarchy of individuals, ecosystems, and the biosphere.

For the second chapter of my dissertation, I return to the topic of how best to represent biochemical networks in order to capture properties representative of the system that the network is describing. Within the formalism of network theory, one of the simplest ways to capture insights into the global structure of a network is to analyze the shape of its degree distribution—that is, the distribution of number of connections that each node has. A huge volume of research into various complex biological, technological and social networks has therefore focused on identifying the shape of the corresponding degree distributions for network projections describing those systems. One of the most significant results emerging from these analyses is that many networks describing real-world systems exhibit ostensibly 'scale-free' topology [70–74], characterized by a power-law degree distribution. The allure of scale-free networks is in part driven by the simplicity of their underlying generative mechanisms, for example a power-law degree distribution can be produced by relatively simple preferential attachment algorithms [70], or to a lesser extent through optimization principles [75]. This raises the question of determining which projection to analyze, and whether or not a real-world system should be considered "scale-free" if only some of its network projections exhibit power-law degree distributions.

Applying a newly developed formalism based on rigorous statistical methods, I find that a majority of biochemical networks are not scale-free, independent of projection or level of organization. I also demonstrate how the network properties analyzed herein can be used to distinguish individual and ecosystem level networks, and find that independent of projection, individuals and ecosystems share very similar structure.

For my third chapter, I pivot away from describing the topology of biological networks, and instead use a dynamic biochemical network methodology to answer a complementary question about life's organization—how does it interface with the planet from which it emerges? Despite the biosphere's apparent interminable coexistence with the geosphere, there remain many open questions on the matter of life persisting in Earth's absence. Quantifying the viability of Earth life outside of our own geosphere is also necessarily important for understanding the possibility of terraformation, and for forward contamination in the context of planetary protection [76–78].

To begin to address these topics, I must first lay the framework for determining the environmental conditions required for a species to produce or acquire the chemical compounds necessary to yield a viable metabolism. For this, I utilize the network expansion method [79]: an organism can catalyze a reaction only if it has access to the necessary substrates. The organism catalyzes all the reactions it can based on the compounds available in its network, and then adds the new compounds it can generate to its network. This process proceeds iteratively until the organism can produce no new compounds.

Network expansion models have been used to explore the scope of chemicals accessible to biology across space and time on Earth, and how changing environments and changing biochemical networks impact one another [80]. For example, the models have been utilized to identify how oxygen drastically altered life's biochemical networks during the great oxygenation event [54]; how biochemistry differed before phosphorous was widely available [53]; how organismal scopes vary across the tree of life [80, 81]; and how organismal metabolic variability is impacted both in the presence of diverse environments and the presence of other species [82].

I lay out a framework for using network expansions to address the question of life's viability amongst other planetary chemistries, and then we work through a case study of this framework to determine the viability of varying Earth organisms within Enceladus's planetary context. The results hint that forward contamination from individuals may be much less concerning than contamination by a microbial ecosystem which can emulate the robustness and catalytic capabilities of the biosphere—reinforcing the perspective that the emergence of life on a planet is an extension of the planet's geosphere [21, 83].

Chapter 2

UNIVERSAL SCALING ACROSS BIOCHEMICAL NETWORKS ON EARTH

This chapter was written in collaboration with Hyunju Kim, Cole Mathis, Jason Raymond and Sara I. Walker. It is pending acceptance in Science Advances. An earlier version is posted on bioRxiv [84]

2.1 Abstract

The application of network science to biology has advanced our understanding of the metabolism of individual organisms and the organization of ecosystems, but has scarcely been applied to life at a planetary scale. To characterize planetary-scale biochemistry, we constructed biochemical networks using a global database of 28,146 annotated genomes and metagenomes, and 8,658 cataloged biochemical reactions. We uncover scaling laws governing biochemical diversity and network structure shared across levels of organization from individuals to ecosystems, to the biosphere as a whole. Comparing real biochemical reaction networks to random reaction networks reveals the observed biological scaling is not a product of chemistry alone, but instead emerges due to the particular structure of selected reactions commonly participating in living processes. We show the topology of biochemical networks for the three domains of life is quantitatively distinguishable, with $> 80\%$ accuracy in predicting evolutionary domain based on biochemical network size and average topology. Taken together our results point to a deeper level of organization in biochemical networks than what has been understood so far.

2.2 Introduction

There is increasing interest in whether biology is governed by general principles, not tied to its specific chemical instantiation or contingent upon its evolutionary history [1–3]. Such principles would be strong candidates for being universal to all life [4, 5]. Universal biology, if it exists, would have important implications for our search for life beyond Earth [6–9], for engineering synthetic life in the lab [10, 11], and for solving the origin of life [12, 13]. Systems biology provides promising quantitative tools for uncovering such general principles [14–16]. So far, systems approaches have primarily focused on specific levels of organization within biological hierarchies, such as individual organisms [17, 18] or ecological communities [19, 20], and are rarely applied to the biosphere as a whole. But, biology exhibits some of its most striking regularities moving up in levels of organization from individuals to ecosystems, and these regularities may only truly manifest at the level of ecosystems, and ultimately the biosphere [21, 22]. For example, while individual organismal lineages fluctuate through time and space, the functional and metabolic composition of ecological communities is stable [23, 24]. To understand the general principles governing biology, we must understand how living systems organize across levels, not just within a given level [25–27].

In order to explore regularities within and between levels of organization, we adopt a network view of biochemistry [17],[48, 50, 85] by constructing biochemical reaction networks from genomic and metagenomic data. We show biochemical networks share universal organizational properties across levels, characterized by scaling laws determining how topology and biochemical diversity change with network size. These scaling relations exist independent of evolutionary domain or level of organization, applying across the nested hierarchy of individuals, ecosystems, and the biosphere. The biochemical diversity and network properties driving this scaling behavior are predictive of evolutionary domain, indicating the biochemical network structure for each domain is distinct even though all three conform to the same scaling behavior. Our results provide a first quantitative demonstration that the application of network theory at a planetary scale can uncover properties existing across different levels of organization within the biosphere, and can be predictive of major divisions within a given level (such as evolutionary domains). On the whole, our results provide new paths forward for identifying universal properties of life.

Our analysis begins with a global database of genomes and metagenomes, sampled from across life on Earth. We leverage available existing annotated genomic data representing the three domains of life, including genomes of 21,637 bacterial taxa and 845 archaeal taxa from the Pathosystems Resource Integration Center (PATRIC) [56], and 77 eukaryotic taxa from the Joint Genome Institute (JGI) [57]. Our metagenomic data includes 5,587 metagenomes from JGI cataloging ecosystem-level biochemical diversity across the planet, see Fig. 1.

From this data, we constructed biochemical networks for each individual organism (genome) and ecosystem (metagenome) using reaction data cataloged in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [58]. Building on prior work studying biosphere-level models of metabolism [53–55], we use the database of all 8,658 enzymatically catalyzed reactions cataloged in KEGG (at the time of data retrieval) as a proxy for the biochemistry of the biosphere as a whole, modeled as a 'soup of enzymes' by disregarding the boundaries of individual species [19]. Network representations of ecosystem-level and biosphere-level biochemistry are 'compartment-free' in that no knowledge of individual species is included. Previous topological analyses of

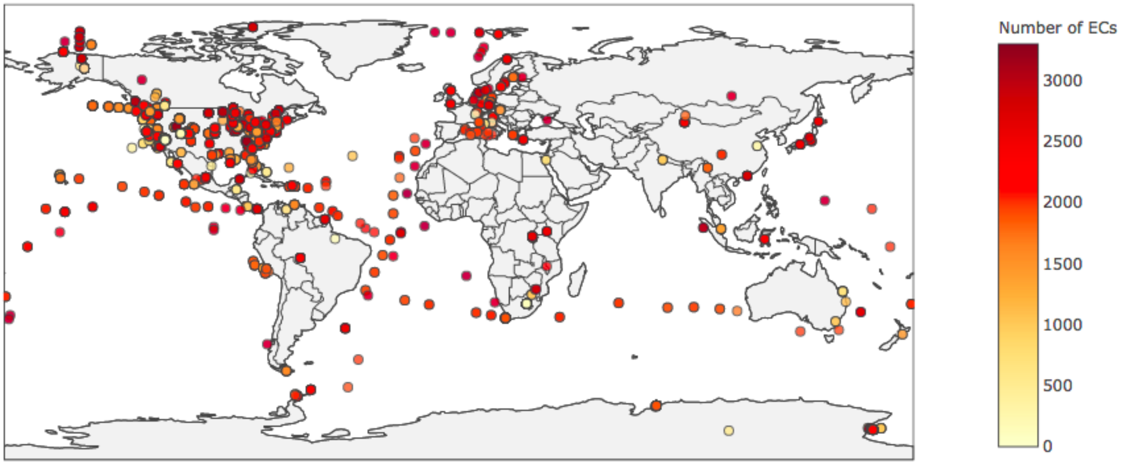


Figure 1. Enzyme diversity of ecosystems across Earth. Shown is the geographical distribution of the 5,587 ecosystems in our study, colored by the number of different enzyme functional classes (enzyme commission (EC) numbers) encoded in sampled metagenomes (data from JGI). Despite large variances in the enzyme diversity and what enzymes are present in each ecosystem, all ecosystems sampled are found to conform to the same scaling behavior for biochemical diversity and topology as a function of biochemical network size, see Fig. 3.

biochemical networks have primarily focused on the subset of biochemical reactions associated with metabolism [50, 85]. Since we are interested in properties universal across life, and not just subsets of living processes, we instead construct networks inclusive of every known catalyzed reaction (regardless of pathway) coded by the respective genome or metagenome, provided the reaction is cataloged in KEGG.

Adopting a network representation allows systematic quantification of topological properties using graph theory and statistical mechanics [16, 17, 29, 30, 34–37]. Using two different graph projections, we compare biochemical networks across levels to test whether they are similar or different and compare to biologically-motivated, randomly sampled networks (see Methods for details on network construction). Our use of the term ‘random’ herein refers to the random sampling procedures we implement to generate ensembles sharing some—but not all—properties with the ensemble of real biochemical networks (described below and in the Methods). These are specifically used to isolate those properties of the real biochemical networks driving the reported scaling behavior, and should be distinguished from more generic random networks, such as Erdös-Rényi random graphs typically contrasted to chemical or biochemical networks (see e.g. [48, 86, 87]). A widely implemented framework for assessing commonality across different systems is to look at their scaling behavior [59–64]. If scale-invariant properties are found, it can be suggestive of deeper, underlying organizing principles [3, 65, 66], such as when distinctive scaling laws emerge in critical systems [68]. We therefore sought to determine whether biochemical networks display scaling laws governing their topology and chemical diversity which are similar across levels, indicative of the existence of self-organizing principles universal across biological levels.

There are three alternative scenarios to be tested relating network structure across individuals, ecosystems and the entire biosphere, each is shown in Fig. 2. In the first, biochemistry does not have shared network structure across levels, and different scaling behaviors emerge at different levels. In the second scenario, biochemistry has shared network structure across levels, but this shared structure can be fully explained by the structure of random chemical networks (generated from random collections of biochemical reactions used by biology). In this case, real biochemical reaction networks would be statistically indistinguishable from random reaction networks, implying the self-organizing principles are solely chemical and not biological in origin. In a third scenario, biochemistry has shared structure across levels, which is different from that of random reaction networks. We find the third scenario to be consistent with our analysis, suggesting the presence of universal organizing principles unique to biology that recur across biological levels of organization. We show these can be explained as an emergent property of the topological structure of the most common reactions participating in living processes.

Before proceeding to the details of our results, it is worth noting the well-known challenges associated with the introduction of statistical artifacts when coarse-graining real-world systems to generate graphical representations [47, 88]. For example, bipartite network representations of biochemistry (treating reactions and substrates as two disjoint sets of nodes) have information that cannot be recovered from unipartite representations (which treat only substrates as nodes). The challenge of choosing a projection arises because biochemical networks are themselves a multi-layer system consisting of enzymes and their catalyzed reactions; enzymes (often abstracted away in network representations) control the biological organization we aim to characterize. To ensure the regularities we report are reflective of the true underlying organization of biochemistry, and are not statistical artifacts introduced by a specific choice of coarse-graining, we therefore consider both a unipartite and bipartite projection in our analysis [47]. We also compare catalytic diversity—quantified in terms of the number of enzymes and reactions—across levels, which is independent of network representation. As we will show, common scaling laws describing biochemical networks across levels of organization are consistently observed in each of these different views of biochemistry, confirming our results are independent of the type of network representation.

One remaining consideration once a network representation is adopted is how to analyze it. So far, the majority of network analyses applied to biochemistry have focused on the 'scale-free' structure of metabolic networks [41, 48, 69]. For example in a seminal paper by Jeong et al. [48], it was shown (using a unipartite representation) that metabolic networks from all three domains of life exhibit the characteristic power-law degree distribution of a scale-free network, with similar scaling

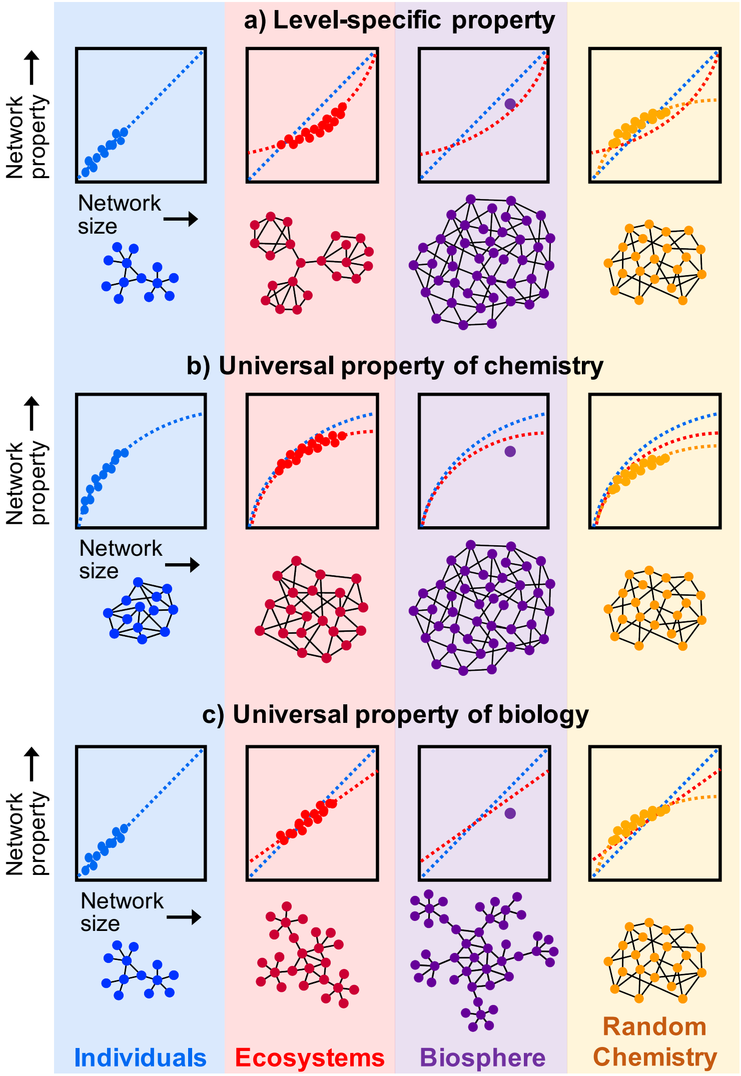


Figure 2. Three alternative scenarios for how biochemical network structure might be similar or dissimilar across levels of organization. For each scenario, illustrative plots show examples of scaling behavior of some network property as function of network size, where each data point corresponds to the measure for a single instance of a network. In the first (A) biochemistry does not exhibit common network structure across levels, and different properties emerge at different levels. In the second (B), biochemistry has a common network structure across all levels, but this structure is also shared by random chemical networks. In the final scenario (C), biochemistry has shared structure across all levels, which is different from that of random chemical networks. Our results are consistent with this third scenario, indicative of universal organizing principles recurring across biological levels, which are unique to biology (not shared by random chemistry), which we show arises due to the network structure of common reactions shared across life on Earth.

exponents for bacteria, archaea and eukaryotes. This and other previous work has focused primarily on properties within single instances of a network (e.g. an individual organism's metabolism), with similar structure to biology (such as the scale-free property) reported in chemical networks more generally [86, 87, 89, 90]. However, as we stated earlier, our interest is in looking at properties across networks (e.g. describing ensemble properties of biochemical networks at the individual and ecosystem-level). We therefore focus on topological measures such as average shortest path length, average clustering coefficient and assortativity (degree correlation coefficient), which are directly comparable across different networks, allowing us to make statements about regularities existing across biochemistry sampled from different levels of organization in a manner that has not been possible in previous work focusing only on a given level.

2.3 Results

2.3.1 Scaling laws describe biochemical networks and catalytic diversity across levels

Organisms can vary widely in their number of genetically encoded reactions, and ecosystems generally include more encoded reactions than individuals do. We therefore compare topological properties relative to the size of biochemical networks as a relevant scaling parameter for our analysis. We define network size as the number of molecules connected through catalyzed reactions within the largest connected component (LCC) for a given biochemical network. We focus analyses on the LCC since some measures are not defined on disconnected networks. The LCC includes $> 90\%$ of compounds for all but the smallest networks in our study, and $> 97\%$ of compounds for the largest (see Supplementary Section on Topological Measures, Supplementary Fig. SI 7, and Supplementary Table SI 1). The fact that the LCC is not 100% of the network could be attributable to missing data in the annotation of genomes and metagenomes. We therefore verified our results are not sensitive to similar magnitude of missing data by confirming the scaling trends reported here are not affected when 10% of nodes are randomly removed (see Supplementary Fig. SI 8). Furthermore, our results reported below suggest larger proportions of missing data (as often occurs for genomes or metagenomes missing many annotated genes) would not significantly affect our results, as we find the reported scaling behavior is primarily driven by the topological structure of the most common reactions found across all of biology. We also verified our results hold when analyzing a more balanced subset of our data, avoiding disproportionately large contributions by genetically similar taxa (see Supplementary Fig. SI 14 and Fig. SI 15).

We calculate several frequently implemented topological measures for the LCC for each network (see Supplementary Section on Topological Measures). We classify properties (e.g. topological or diversity measures) as universal when they scale in the same way across levels. We identify these cases by properties which scale according to the same fit across levels (e.g. network average clustering coefficient scales linearly with network size for both individuals and ecosystems, and we thus identify this scaling as universal across levels). Shared fit functions across levels suggest mechanisms driving the structure of biochemical networks may be independent of level of organization; in such a case individuals and ecosystems could both be subject to the same general principles acting to architect them. That is, we do not require the scaling coefficients to be exactly the same (indicating the tuning of mechanisms generating structure in individuals and ecosystems), but we do require the same fit to be shared across our data (indicating the possibility of shared generative mechanisms) to qualify as universal.

To test whether biology exhibits universal scaling behavior across levels, we first determined how topological properties and biochemical diversity vary with size for all individuals and ecosystems in our data set. Measured values for the unipartite representation and catalytic diversity (enzymes, reactions) are shown in Fig. 3 as a function of network size (see Supplementary Fig. SI 10 for data on bipartite representation which exhibits similar consistency across levels). We find individuals and ecosystems scale according to the same functional form for each network and diversity measure, with similar scaling coefficients (for fits and confidence intervals, see Supplementary Data S1). Scaling for individuals and ecosystems is therefore universal. For some measures (assortativity and betweenness) the biosphere falls within the 95% confidence interval observed for fits of ecosystem level scaling. An exception is clustering coefficient, where the biosphere significantly departs from the observed ecosystem scaling: this could be attributable to missing data on global enzymatic diversity (which falls slightly below what our scaling laws would predict). Topological measures that scale following power-law fits ($y = y_0 x^{\beta}$, where $\beta = \beta_{ind}$ for individuals and $\beta = \beta_{eco}$ for ecosystems) include: average betweenness ($\beta_{ind} = -1.1581$, $\beta_{eco} = 1.136$), average shortest path length ($\beta_{ind} = -0.117$, $\beta_{eco} = -0.084$), and number of edges ($\beta_{ind} = 1.219$, $\beta_{eco} = 1.243$). Both biochemical diversity measures also scale according to power-law fits: number of enzyme classes (a proxy for enzymatic diversity) ($\beta_{ind} = 1.294$, $\beta_{eco} = 1.838$), and number of reactions ($\beta_{ind} = 1.229$, $\beta_{eco} = 1.319$). Average clustering coefficient scales with a linear fit ($y = mx + y_0$, $m = m_{ind}$ for individuals and $m = m_{eco}$ for ecosystems) for individuals and ecosystems ($m_{ind} = 3.77 * 10^{-5}$, $m_{eco} = 3.32 * 10^{-5}$). These results rule out the possibility scaling laws are level-specific (Fig. 2a). The observed scaling laws confirm biochemical networks exhibit shared structure across levels of organization, where network properties and biochemical diversity are largely determined by size as the relevant scaling parameter.

2.3.2 Real networks exhibit different scaling behavior than random chemical networks

The observed universal scaling across individuals and ecosystems could be unique to biology, or it could arise due to self-organizing principles of chemistry. If the later is true, we should expect randomly sampled chemical networks to exhibit the same fit functions across networks as real biochemical networks do. Testing this requires comparison to randomly sampled chemical networks, which must be generated with an appropriate biologically-relevant control to be informative [91]. Since we are interested in the global organization of biochemistry, we constructed control random chemical reaction networks (henceforth called random reaction networks) by merging randomly sampled reactions from the KEGG database following a flat distribution where all reactions are equally likely to be sampled (see Methods for details on network construction). This random sampling produces ensembles of random reaction networks that globally (over the ensemble) share the same chemical reactions as our biosphere.

We performed the same analyses on the ensemble of random reaction networks as real biochemical networks. We observe random reaction networks do not scale according to the same functional form as biochemical networks for some network topology measures (Fig. 4, first column). The fits for average clustering coefficient of random reaction networks favor a power-law function ($y = y_0 x^{\beta}$, with $\beta_{ran} = 0.6401$), compared

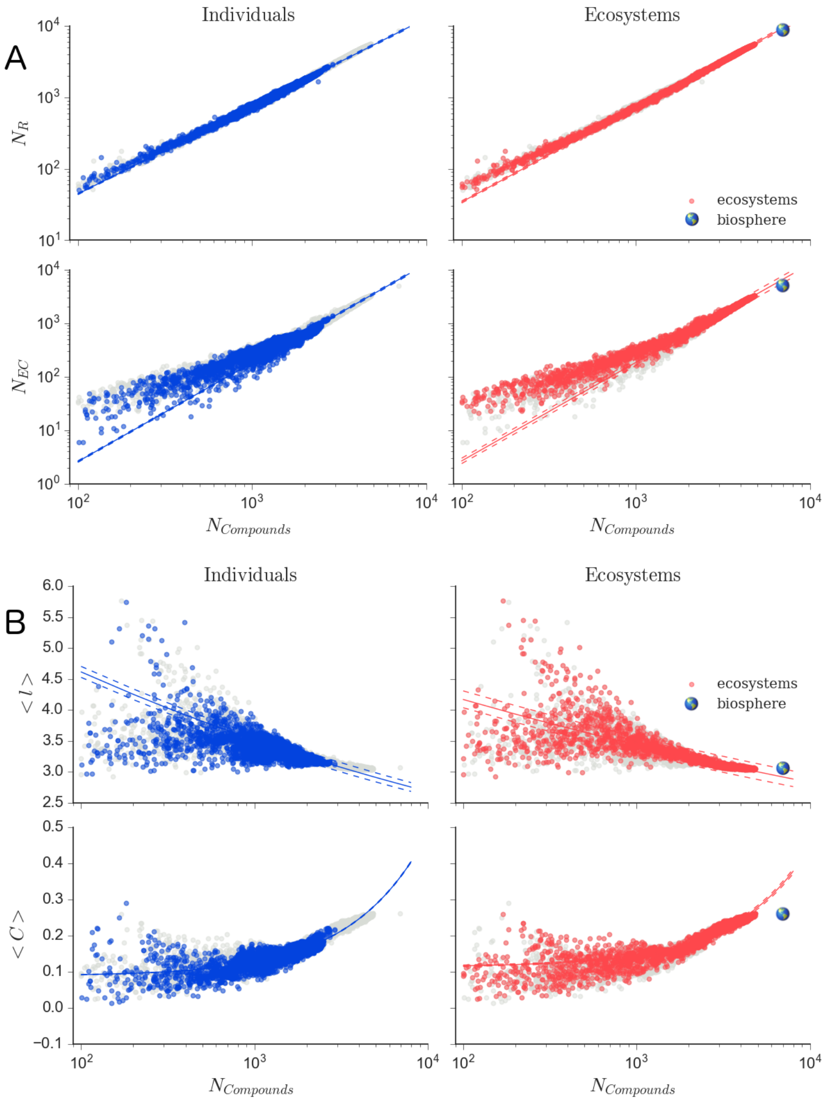


Figure 3. Common scaling laws describe biochemical networks across levels of organization. Scaling of biochemical measures for individuals (left column) and ecosystems (right column) share the same functional form for biochemical diversity (enzyme and reaction diversity) and for topological measures. Shown from top to bottom are: (A) number of reactions (N_R), and number of enzyme classes (N_{EC}). (B) average shortest path ($< l >$), and average clustering coefficient ($< C >$). All measures are as a function of the size of the largest connected component ($N_{Compounds}$). Ecosystems include metagenomes (red) and the biosphere-level network (Earth icon). Fits for each dataset (solid lines) are shown with 95% confidence intervals (dashed lines). For reference, shown in light grey is data for all biochemical networks (individuals, ecosystems, biosphere). Additional measures are shown in Supplementary Fig. SI 9, and scaling for bipartite networks is shown in Supplementary Fig. SI 10.

to the linear function favored by the biochemical networks. Fits for assortativity favor a linear function for random reaction networks ($y = mx + y_0$; $m_{ran} = -4.5255$), whereas for biochemical networks it was found to not scale with size (Supplementary Data S1). That is, there are certain aspects of the topology of random reaction networks that scale with network size in a manner that is entirely distinct from that of real biochemistry. The qualitative differences in scaling behavior indicate the real and random biochemical networks represent different universality classes. In addition to these qualitative differences in scaling behavior, we also observed statistically significant quantitative differences in the random chemical networks: scaling relationships for randomly sampled biochemical networks do not overlap with real biological individuals in many cases. Topological measures in random reaction networks which scale according to power-law fits ($y = y_0 x^{\beta}$, $\beta = \beta_{ran}$ for random reaction networks) include: average betweenness ($\beta_{ran} = -1.0595$), average shortest path length ($\beta_{ran} = -0.0543$), and number of edges ($\beta_{ran} = 1.2459$). Both biochemical diversity measures also scale according to power-law fits: number of enzyme classes ($\beta_{ran} = 1.10156$), and number of reactions ($\beta_{ran} = 1.3590$). We conclude the organizational properties of random chemical networks cannot alone explain the scaling laws observed for real biochemical networks.

2.3.3 Scaling-laws represent shared constraints re-emerging across levels

Our results establish that Earth's biochemistry exhibits universal scaling behavior across levels of organization not explainable by the organizational patterns of randomly sampled chemistry alone. A natural next question is whether ecosystems inherit their properties from individuals, or whether they instead exhibit similar structure due to

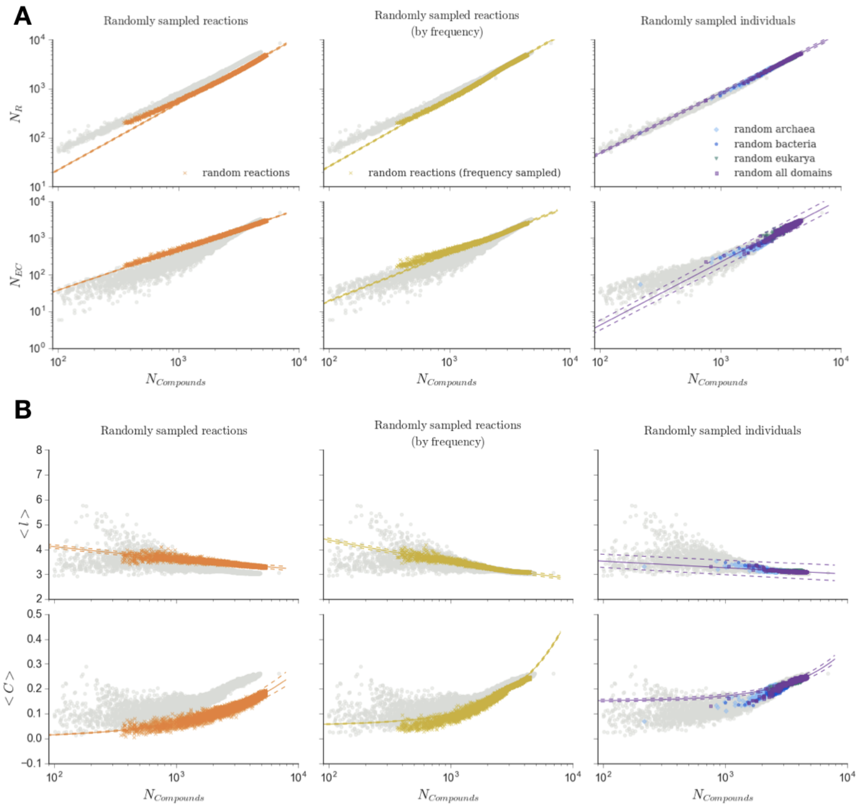


Figure 4. Scaling laws distinguish biochemical networks from random networks across levels of organization. Shown are random reaction networks created by sampling biochemical reactions from a flat distribution (left column), frequency-sampled random reaction networks created by sampling reactions based on the frequency distribution observed across all organisms (center column), and random genome networks (right column). Merged networks composed of individuals include bacteria only (light blue), archaea only (dark blue), eukarya only (blue-green), and all domains combined (purple). (A) Scaling of biochemical diversity. Diversity measures and fit are as described in Fig. 3. For reference, all real biochemical network data from Fig. 3 is shown in light grey. Additional measures are shown in Supplementary Fig. SI 11. (B) Scaling of network structure. Measure and fit descriptions match those described in Fig. 3. For reference, all real biological networks from Fig. 3 are shown in light grey. Additional measures are shown in Supplementary Fig. SI 11, and scaling for bipartite networks shown in Supplementary Fig. SI 12. We find random reaction networks do not recover the same fit functions as real biological networks for assortativity and clustering, whereas frequency-sampled random reaction networks and random genome networks only differ for assortativity, but nonetheless are statistically distinguishable from real biochemical networks some measures.

similar constraints re-emerging at different levels. Addressing this requires determining whether or not scaling behavior for individuals is statistically distinguishable from ecosystems. We assumed as a null hypothesis scaling relationships are consistent across levels of organization, and performed a permutation test [92], using the scaling coefficient as the test statistic (see Methods on Fitting network measure scaling and permutation tests). We find scaling relationships are not distinguishable for individuals and ecosystems when analyzing average node betweenness and average shortest path length (Supplementary Table 2). However, scaling coefficients are distinguishable for number of reactions, number of edges, number of enzyme classes, and mean clustering coefficient, with p-values $< 10^{-5}$ in most cases. Confidence intervals on scaling coefficients for ecosystem topology are narrower than for individuals, indicating ecosystem scaling is more tightly constrained. Although biochemical networks for individuals and ecosystems share similar scaling behavior, they are not drawn from the same distributions; allowing the possibility shared constraints operate at each level separately.

We next sought to identify sufficient constraints for recovering the observed scaling across levels. To do so, we constructed a different ensemble of random reaction networks by merging randomly sampled reactions from the KEGG database, but this time weighting the sampling frequency of biochemical reactions by the number of individual genomes where the reactions are found (henceforth called frequency-sampled random reaction networks, see Methods for details of their construction). This random sampling produces ensembles of random chemical networks that, as before, globally (over the ensemble) share the same reactions as our biosphere, but also has the additional property of sharing same frequency distribution of reactions over ‘individuals’ as our biological dataset. This random ensemble therefore more closely reproduces properties of real biological networks than that introduced in the previous section. Since most highly connected nodes (participating in many reactions) are common to all three domains, e.g. ATP and H_2O [48, 93] this sampling procedure yields random control networks that tend to include the most common compounds used by life. We find the scaling behavior of this ensemble of random networks much more closely matches the observed scaling trends in real biology (Fig. 4, second column). Whereas we observe qualitative differences for scaling of clustering in the previous case, here the fit function is the same as for the clustering coefficient for both the real biochemical networks and the frequency-sampled random reaction networks. In fact, all fit functions are the same as those for real biochemical networks, with the exception of assortativity. Additionally, for measures sharing the same fit functions, fewer measures distinguish biological networks from frequency-sampled random reaction networks than from random reaction networks without this constraint. We can therefore conclude the network structure of the most frequently occurring reactions across life on Earth is an important driver of the observed scaling behavior for the real networks.

To further confirm scaling emerges due to these shared properties across all life, we next generated simulated ecosystem-level networks by merging randomly sampled genome networks from each domain individually and from all three domains together (see Supplementary Materials and Methods for details on network construction). This allows us to determine how scaling behavior could be the same or different for an archaeasphere (archaea alone), bacteriasphere (bacteria alone), eukaryasphere (eukarya alone), or artificial ecosystems (all three domains) (Fig. 4, third column). We find the functional forms of scaling relationships are the same for real ecosystems and these randomly merged organismal networks (hereafter called random genome networks). This is somewhat surprising given it is not in general true randomly selected subnetworks of a network have the same structure as the original network [94] (e.g., individuals as subnetworks of ecosystems do not necessarily have to share the same structure). However, we also checked whether scaling exponents and coefficients are statistically distinguishable for real ecosystems and random genome networks, using the same permutation tests as before, and find they are for most measures (see Supplementary Table SI 2). Random genome networks and real ecosystems exhibit exponents distinguishing their scaling coefficients for most topological measures and for number of enzymes, with p-values $< 10^{-5}$. Scaling of betweenness is indistinguishable between the two datasets. These results indicate random genome networks differ from real ecosystems in many of the same ways individuals do. However, just as scaling of assortativity qualitatively distinguishes individual biochemical networks from the frequency-sampled random reaction networks, scaling of assortativity also distinguishes random genome networks from real ecosystems, whereas it does not distinguish real individuals from real ecosystems (Fig. 5). Taken on the whole, these results suggest scaling behavior for real ecosystems arises due to organizing principles operative at the level of ecosystems, and is not solely an emergent property due to merging individual-level networks.

Combining these results for frequency-sampled random reaction networks with that of random genome networks indicates the existence of individuals sharing a common set of biochemical reactions is a sufficient condition for networks of all sizes (from small individuals to large ecosystems) to exhibit the scaling behavior observed in real living systems. Taken together with the results of the previous section, we can conclude the particular form of the scaling relations observed across life on Earth

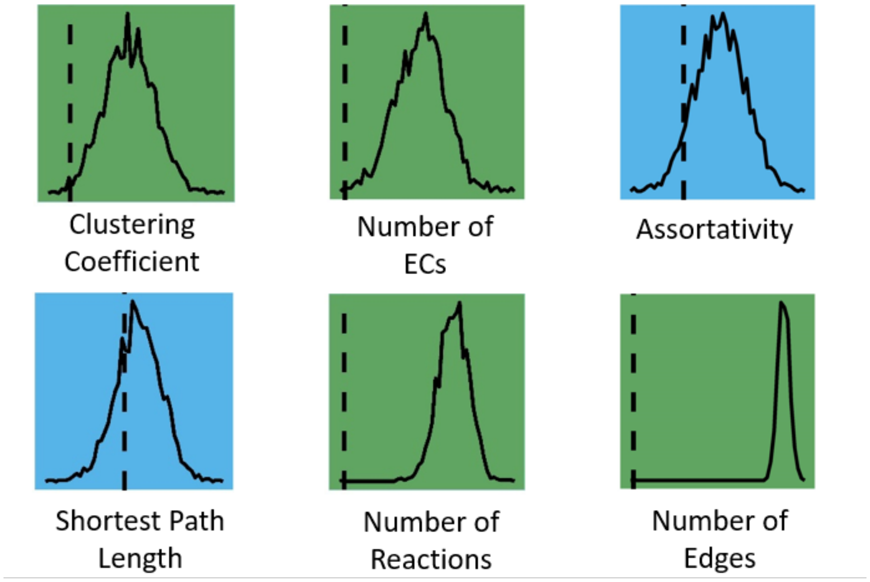


Figure 5. Scaling laws for individuals and ecosystems are statistically distinguishable for some network and catalytic diversity measures. Shown are the results of a permutation test to determine whether properties of biochemical networks constructed from individual genomes scale differently than those constructed from metagenomes (ecosystems). For each network measure the test statistic is shown as a vertical dashed line, while the null distribution is shown as a solid line (see Methods on Fitting network measure scaling and permutation tests for more details). Blue squares indicate scaling behavior is indistinguishable between levels of organization, while green squares show measures which can distinguish scaling of individuals from that of ecosystems.

emerges due to the structure of interactions among compounds common across all life, which is not in general characteristic of non-biological chemical reaction networks.

2.3.4 Network structure predicts evolutionary domain

Any general organizing principles in biology must be consistent with the variation responsible for the diversity of life we are already familiar with. Since the three domains of life represent the most significant evolutionary division in the history of life [95], we therefore also tested whether or not network structure can distinguish individuals sampled from the three domains (see Methods on Predicting evolutionary domain from topology). To approach this question, we first investigated compounds shared across all domains to determine which compounds are distinct to each domain and which are universal to all three. We identified the contributions of each domain to the biosphere as a whole by comparing compounds at the biosphere-level to those across the networks of individuals, identified by evolutionary domain. We do so by identifying which compounds are unique to each domain and which are shared across all three domains, determined from annotated data in the 22,559 genomes in our dataset. At the biosphere-level, 0.44% of compounds are unique to archaea, 3.14% are unique to bacteria, and 17.08% are unique to eukarya, reaffirming each domain represents significantly different metabolic strategies and genetic architectures, as is well established by earlier work [95]. However, it is also well established all life on Earth shares a common set of core-biochemistry [96]: a higher percentage of compounds, constituting 37.23% of the biosphere-level network, are shared across all three domains in our dataset (Fig. 6A,B,C,D), including hubs such as ATP, and H_2O as mentioned previously. Since many more chemical compounds are shared across all three domains than are unique to each, one might a priori expect the organization of these compounds into biochemical networks to not be predictive of domain.

We find the opposite to be true: despite a large fraction of shared biochemical compounds, the organization of those compounds into networks is distinct for each domain. We find in most cases average topology normalized to size can reliably predict evolutionary domain (Fig. 6E). In many cases prediction accuracy is $> 80\%$, when only a single topological measure is used. By contrast, topology or size alone provides significantly less accurate predictions. This demonstrates biochemical network

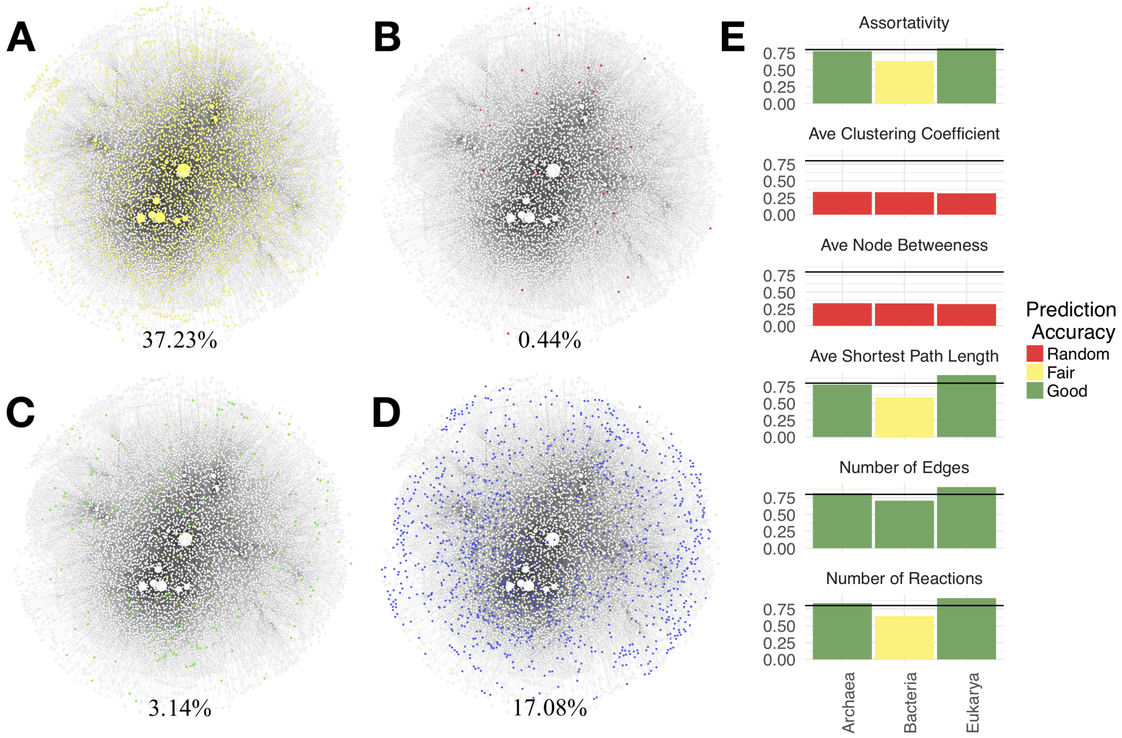


Figure 6. The biosphere-level chemical reaction network. The biosphere-level network is constructed from the union of all 22,559 genomic networks in our study. Each panel shows the same biosphere-level network, with nodes (representing compounds) in white and edges (representing their connections) in grey. Node size indicates degree within the network. Colors indicate biochemical compounds used in (A) all three domains of life (yellow), (B) in archaea only (pink), (C) in eukarya only (green) and (D) in bacteria only (blue). Although many more chemical compounds are shared across all three domains than are unique to each, the organization of these compounds into biochemical networks is distinct for each domain based on statistical testing which shows (E) catalytic diversity and biochemical network topology can predict evolutionary domain. Shown is the estimated prediction accuracy (y-axes) for each measure and each domain. The colors of each bar indicate prediction accuracy of a given measure for a particular domain: red is comparable to random guessing ($y \leq 33\%$ accuracy); yellow is better than random but not completely predictive ($33\% < y \leq 67\%$); green is predictive of domain ($67\% < y$). The horizontal line indicates 80% prediction accuracy.

structure can be predictive of the taxonomic diversity of individuals. Combined with our other results this suggests the same biochemical network properties (topology and catalytic diversity) driving regularity across levels of organization can also be predictive of major evolutionary divisions within a given level, providing evidence the regularities identified herein are indeed consistent with a signature of global organizing principles for biochemistry.

2.4 Discussion

Our analyses reveal biochemical networks display common scaling laws governing their topology and biochemical diversity, which are independent of the level of organization they are sampled from. These scaling laws cannot be fully explained by the structure of random reaction networks that do not account for the structure of the subset of reactions shared across life on Earth. We were also able to confirm the same topological regularities occurring across levels of organization within the biosphere can be predictive of evolutionarily divisions within a level, using the three domains as an exemplar. Collectively, our results indicate a deeper level of organization in biochemical networks than what is understood so far, providing a new framework for understanding the planetary-scale organization of biochemistry and how nested hierarchical levels are structured within it.

A key implication of our analysis is the importance of individuals sharing a common set of biochemical reactions in shaping the universal scaling laws observed across hierarchical levels. Scaling laws often emerge in systems where universal mechanisms operate across different scales, yielding the same effective behavior independent the specific details of the system. It is in this sense scaling laws can uncover universal properties, motivating their widespread use in physics and increasing application to biology [59, 63, 65, 97–100]. Here we have shown the relevant scaling parameter for biochemical organization is the number of biochemical compounds (in a network representation this is the size of the network). Individuals, ecosystems and the biosphere obey much the same scaling behavior for biochemical network structure, indicating the same universal mechanisms could operate across all three levels of organization. In physics, this kind of universality usually implies there is no preferred scale or basic unit. However, in the biological example uncovered here, the presence of specific scaling relations observed in real biochemical networks can be explained by biological individuals (lower-level networks) sharing a common set of reactions as basic 'units'.

Future work should explore the connections between the scaling relationships reported here and other work characterizing scaling behavior across living processes. For example, our results indicate ecosystems are more tightly constrained than individuals, better displaying the regularities of biochemical network architecture. However, projecting ecosystem-level scaling to the biosphere as a whole does not recover the observed network properties for the biosphere-level network. Recently, scaling laws describing microbial diversity were used to predict Earth's global microbial diversity, and in particular to highlight how much diversity remains undiscovered [61]. It could be an analogous case here, where the uncovered scaling relations could be used to predict missing enzymatic diversity in the biosphere. Furthermore, one area of intensive investigation is allometric scaling relations [61, 97, 100], including how shifts in metabolic scaling could be indicative of major transitions in evolutionary hierarchies [59]. Allometric scaling laws are derived by viewing living systems as localized physical objects with energy and power constraints. Here, scaling emerges due to an orthogonal view of living systems as distributed processes transforming matter within the space of chemical reactions. The connections between these different aspects of scaling in living organization remain to be elucidated.

A final implication of our work is the consequences for our understanding of the origin of life, before the emergence of species. The existence of common network structure across all scales and levels of biochemical organization suggests a logic to the planetary-scale organization of biochemistry [101], which—if truly universal—

would have been operative at the origin of life. While our analysis has uncovered universal scaling behavior for extant life, arising due to the structure of connectivity and diversity among the most common biochemical compounds and reactions, it remains to be determined whether the particular scaling reported is a by-product of shared biochemistry across all life, or if fundamental constraints on biochemical network structure, operative across scales from individuals to planets, drives lower-level individuals to necessarily share common reactions. If the latter is true it would have important implications for understanding the processes operative at the time of the last universal common ancestor. If the same global network structure, characterized by the same scaling laws, described Earth's biosphere throughout its evolutionary history, the emergence of individuals (as selectable units) with shared biochemistry would have played an important role in mediating a transition in the organization of Earth's chemical reaction networks. Even if we could assume the same planetary-scale chemistry for a lifeless world, we should expect to see dramatically different scaling for a hierarchically organized biosphere of nested evolutionary units where 'units' are defined by a shared subset of biochemical architecture across all life [102, 103]. An important question for future work is identifying the planetary-drivers of Earth's biosphere-level biochemical network structure and how this has structured living systems across nested levels over geological timescales. This will require characterizing the organization of planetary-scale biochemistry, as developed here, within the broader context of studying a planet's geologic and atmospheric evolution. It remains an open question as to what will ultimately explain the universal structure of Earth's biochemical networks, or whether we should expect all life to exhibit similar scaling behavior, even on other worlds.

2.5 Materials and Methods

2.5.1 Obtaining genomic and metagenomic information

2.5.1.1 Genomes (PATRIC)

Archaea and bacteria genomic datasets were obtained from PATRIC [56]. Enzyme commission (EC) numbers were obtained from `ec_number` column in the pathway data of each taxon. Eukarya genomic datasets were obtained from the Joint Genome Institute's (JGI) integrated microbial genomes database and comparative analysis system (IMG/M)[57]. All eukarya data used in this study was sequenced at JGI. All EC numbers used to construct eukarya biochemical networks were obtained from the list of total enzymes associated with each eukaryote. EC numbers were used in conjunction with KEGG enzyme and reaction data in order to build biochemical networks for each taxon.

2.5.1.2 Metagenomes (JGI)

Metagenomic data was obtained from JGI IMG/M [57]. All metagenomic data used in this study was sequenced at JGI. All EC numbers used to construct metagenomic biochemical networks were obtained from the list of total enzymes associated with each metagenome. These EC numbers were used in conjunction with KEGG enzyme and reaction data in order to build biochemical networks for each metagenome.

Our metagenomic data comes from a wide variety of ecosystems associated with the natural environment, host organisms, and human-made environments from across the globe (see Fig. 1). The metagenomes were sampled from a variety of locations, inclusive of 51 different bodies of water, countries, or Antarctica. The largest categories of sampled ecosystems includes aquatic, terrestrial, plant, wastewater, fungi, insect, mammal, and air, among many others. They come from, for example, soil, marine and freshwater environments, thermal springs, digestive systems, sediments, sludges, and the deep subsurface. Metagenomes were collected over a variety of altitudes, from sea level to a few thousand meters above sea level. Terrestrial and aquatic metagenomes include surface samples as well as those from depths of cms to thousands of meters below the surface. Samples also range in pH from nearly 0 to over 9, and in temperature from just above 0 degrees celsius to 90 degrees celsius. At present it is impossible to say how representative the diversity of life sampled so far is of the total biodiversity of life on Earth (which is presently unknown and not well-constrained). Nonetheless, the breadth of environments in our sample suggests that our dataset includes a reasonable representation of known biodiversity. Additional environmental and omic information is publicly available on JGI's IMG website (`https://img.jgi.doe.gov/cgi-bin/m/main.cgi`).

2.5.1.3 Biosphere

To create the biosphere network, we included all (at the time our data was retrieved) 8,658 enzymatically catalyzed reactions in KEGG.

2.5.2 Network Construction

In this study, all biochemical reaction networks consist of chemical compounds that are involved in biochemical reactions: two chemical compounds are connected to each other when one is a reactant and the other is a product of the same biochemical reaction (see Supplement Section on Network Representations of Catalyzed Biochemical Reaction for more details). The different types of biochemical reaction networks come from how we select a set of reactions to be included in each network, which is described below. Note that all edges in the networks in this paper are represented as undirected and unweighted since our interests lie on the presence or absence of particular reactions in given networks and, in principle, all biochemical reactions can happen in both directions depending on the environment.

2.5.2.1 Biological Networks

For each biological network, we include all catalyzed biochemical reactions annotated in each genome or metagenome. More specifically, we consider three different levels of organization: individual organisms, ecosystems and the biosphere. For the construction of individual networks, we utilize the genome data of 21,637 bacterial taxa and 845 archaeal taxa from the Pathosystems Resource Integration Center (PATRIC)[56], as well as 77 eukaryotic taxa from the Joint Genome Institute (JGI)[57]. From this data, we obtain the set of classes of enzymes for each genome. All reactions catalyzed by this set of enzymes and present in the Kyoto Encyclopedia of Genes and Genomes (KEGG)[58] database are included in the network representation of the corresponding genome. Similarly, for the network representation of each of the

5,587 ecosystems from JGI, we include all reactions catalyzed by the ecosystem's coded enzymes, provided they are catalogued in the KEGG dataset. Finally, for the biosphere network, we include all 8,658 enzymatically catalyzed reactions in KEGG.

2.5.2.2 Parsed Biological Networks

We also analyzed a parsed subset of biological data, in order to reduce the relative size differences between each of our domain datasets. This allows us to test whether our results are consistent with a more balanced representation of biodiversity from each domain. Starting with all bacteria genomes, we selected one representative genome containing the largest number of annotated ECs from each genus. Unique genera (genera only represented by a single genome) were also included in our parsed data. Uncultured/candidate organisms without genera level nomenclature are also included in the parsed dataset. The parsed archaea dataset was created in the same way. Because we have much less extensive data from eukarya, the parsed results include all eukarya (there is no "parsed" eukarya).

2.5.2.3 Random Genome Networks

To construct a random genome network, we sample individual networks uniformly at random from the set of all individual organisms in our data set and merged them into one random genome network. When a set of multiple individual networks are merged, every node and edge present in any individual network is added to the resulting network with equal weight regardless of how many individual networks include them. We built four types of random genome networks with individual networks sampled from only archaea, only bacteria, only eukarya and from integration of all the three domains. The number of individual networks merged to form each random genome networks is defined as the sample size. The sample size ranges from 1 to 200 for 845 archaea genomes, from 1 to 200 for 21,637 bacteria genomes, from 1 to 77 for 77 eukarya genomes and from 1 to 477 for all genomes in the three domains. We selected 10 sets of individual networks for every sample size, and merged them to generate 2,000 random genome networks from individual archaea networks, 2,000 from individual bacteria networks, 770 from individual eukarya networks and 4,770 from all individual networks across the three domains.

2.5.2.4 Random Reaction Networks

In this paper, random reaction networks are generated by merging randomly sampled reactions from all biochemical reactions from the KEGG data regardless of whether a known enzyme is cataloged for the reaction. We note 31.46% of chemical compounds in the biosphere network are not included in the genomic data in our study, therefore our construction uniformly sampling the entire KEGG database, the random reaction networks can include enzymatically catalyzed reactions not included in our genomic data. Nonetheless our sampling procedure is biased to generate networks with similar biochemistry to that of the genomic networks (since compounds common to all three domains tend to be highly connected (participate in many reactions) this uniform sampling procedure yields random networks biased to include the most common compounds used by life). Most biological networks for real individual organisms and ecosystems contain 200 - 5000 reactions. To build the random reaction networks with size similar to real individual organisms and ecosystems, we selected a random number between 200 and 5000, sampled that number of reactions from KEGG data uniformly at random and merged these into a random reaction network. Repeating this, we constructed 5,000 random reaction networks in total.

2.5.2.5 Frequency-sampled Random Reaction Networks

With the goal of creating an ensemble of random networks more similar to real biological networks, we also generated random reaction networks by sampling reactions with probability proportional to their frequency across the set of all individual biological networks. We computed the frequency of every reaction as the number of genomes that includes enzymes catalyzing that reaction to generate a frequency distribution for the occurrence of reactions across our genome-level networks. We then selected a random number between 200 and 5000 and sampled that same number of reactions according to this frequency distribution. By repeating this procedure, we generated 5,000 frequency-sampled random networks. As a check to confirm our results are independent of the relative sizes of our domain datasets, we also generated 5,000 random reaction networks with the same size as members of the ensemble of frequency-sampled random reaction networks, but instead sampled reactions according to the sum of domain-frequencies, computed within each domain and normalized by size of the domain (see Supplementary Fig. SI 15).

2.5.3 Fitting network measure scaling and permutation tests

For each network measure, a scaling relationship was fit as a function of the size of the largest connected component (LCC) of the network. For each measure, three different models were tested, a power law of the form $y = y_0 x^{\beta}$, a linear relationship of the form $y = \beta x + y_0$, and a quadratic function of the form $y = \beta_1 x + \beta_2 x^2 + y_0$. For both the assortativity measures, the preferred fit was also compared to a constant $y = \beta$. The preferred model was chosen as the one which minimized cross validation errors, according to 10-fold cross validation, across the entire data set.

Once a model was chosen, a simulated permutation test was performed to determine whether the scaling relationship for a given attribute was the same for ecosystems and individuals or if it was distinct [92]. We took as the null hypothesis that the scaling relationship across different levels of organization is constant, and used the fitted scaling parameters (for individuals and ecosystems) as the test statistic. We used fitted 1,000,000 resamples of the complete dataset to estimate the likelihood of the fit for individuals (or ecosystems) to have been drawn randomly from the complete dataset. We performed this test for both the ecosystem and individuals, if there was a difference in the estimated likelihoods we took the greater of the two. These likelihoods are the (two-sided) p-values reported in Table SI 2. The same procedure was followed to determine the distinguishability of ecosystem networks with the randomized controls (random genome networks, and random reaction networks). Random reaction networks were distinguishable from ecosystems networks for all measures, with p-values $= 10^{-6}$.

To estimate the true scaling parameters, and 95% confidence intervals a bootstrap sample of 100,000 was used for each network attribute [92]. If the permutation test allowed us to reject the hypothesis of a constant scaling relationship across individuals and ecosystems to a confidence greater than 0.01, the scaling parameters were estimated separately for the individuals and ecosystems, otherwise the complete dataset was fit. The scaling parameters (and confidence intervals) for distinct domains were also estimated using a bootstrap of 100,000 samples.

For scaling fits and confidence intervals see Supplementary Data file S1.

2.5.4 Predicting evolutionary domain from topology

To demonstrate topological features of genomes from different domains are distinct, multinomial regression was used. Specifically, we implemented models where the domain of the network was the response class and a single topological feature, normalized by the size of the largest connected component (LCC) of the network was the dependent variable. We found topological features of networks alone were often not predictive of the domain, but the ratio of the topological properties to the size of the network provided a more accurate prediction. Prior to the regression these normalized topological measures were scaled and centered [92]. The regression was implemented in base R using the `glm(..)`, function. In order to control for over fitting the training data was composed of an equal number of samples from each domain. In particular only 35 networks of each domain were sampled and the model was tested on the remaining data. This process was repeated 100 times and the average model error is reported in the main text Fig. 6E.

2.6 Supplementary Information

2.6.1 Network Representations of Catalyzed Biochemical Reaction

The process to encode a biochemical reaction as the network representation can be described with the diagram below (Fig. SI 16) as follows: (a) Suppose that a chemical reaction R catalyzed by an enzyme E is given, which transforms chemical compounds C1 and C2 to C3 and C4. (b) The reaction, R, can be described in a reaction diagram, or a directed bipartite network representation, where the reactants C1 and C2 are connected to the reaction node and the products C3 and C4 are connected as products from the same reaction. In principle, this biochemical reaction, R, can happen in opposite direction depending on the environment. Therefore, in bipartite network representation, the edges connecting chemical compounds and the reactions are considered as bidirected, which is equivalent to undirected for our analysis. (c) The unipartite network representation of the reaction, R, shows how the reaction information is embedded in the network. In the unipartite network representation, nodes are substrates and a reactant is connected directly to a product if they are connected to the same reaction in the corresponding reaction diagram.

2.6.2 Topological Measures

To characterize the structure of biochemical networks, we utilized some of the most frequently used topological measures. The detailed descriptions about these topological measures can be found in [34]. Here, we briefly review these measures. For computing each measure, we used the Python software package, NetworkX [104].

The topological measures implemented in this paper include average degree, average clustering coefficient, average shortest path length, assortativity (degree correlation coefficient), and node betweenness. We calculate all network measures on the largest connected component (LCC) of each network, for the following reasons: 1. Several network measures only make sense to calculate on connected components (e.g. average shortest path), focusing on the LCC therefore permits all network measures implemented in our study to be calculated for all networks; 2. The largest connected component for each network generally contains the vast majority of nodes ($> 90\%$) for the vast majority of networks in each dataset (the only exception is the random reaction networks, of which only $\sim 76\%$ have a largest connected component with at least 90% of a network's nodes). See Table SI 1 and Fig. SI 7 for distribution of sizes of the LCC by dataset.

2.6.2.1 Degree

The degree of a node i, k_i is the total number of connections between i and rest of the network. The average degree $< k >$ in this paper is the average of k_i for all nodes in the LCC of a given network.

2.6.2.2 Clustering coefficient

The local clustering coefficient for a node i, C_i measures the local density of edges in a network by considering the number of connected pairs of neighbors of i. Hence, C_i is defined as,

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \tag{2.1}$$

where k_i is the degree of node i and L_i is the number of connections between neighbors of i. A large value of C_i indicates the highly interconnected neighborhood of i. The variable C_i is measured by using a Networkx method `clustering(..)`. We computed $< C >$, the average of C_i, over all nodes in the LCC of each network.

2.6.2.3 Shortest path length

The shortest path length, l_{ij} between a given pair of two nodes i and j is defined as the minimum number of edges connecting the two nodes in a given network. The variable l_{ij} is computed using the Networkx method `shortest_path_length(..)`. We calculated the average shortest path length, $< l >$ by averaging l_{ij} for every pair of nodes in LCC of a given network.

2.6.2.4 Assortativity (degree correlation coefficient)

Assortativity measures the tendency of two nodes with similar properties to be connected in a given network. The assortativity coefficient proposed by Newman [40] is formulated as follows:

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b} \tag{2.2}$$

where e_{xy} is defined as the fraction of edges between a node with value x and one with value y for a given node attribute, and a_x and by are the fraction of edges coming into and going out from nodes of value x and y respectively. The variables

σ_a and σ_b are the standard deviations of the distributions of a_x and b_y. When the considered attribute of nodes is their degree, the assortativity becomes the degree correlation coefficient, quantifying the correlation between the degrees of nodes on either side of an edge. Hence, for undirected networks in our study, $a_x = b_y$ and $\sigma_a \sigma_b = \sigma^2$. If $r < 0$, the network is assortative, i.e. nodes with similar degree tend to be connected to each other. If $r > 0$, the network is disassortative, i.e. nodes in it tend to be paired to other nodes with different degrees. For an arbitrary network, $-1 \leq r \leq 1$. To measure the assortativity r, we used a Networkx method `degree_assortativity_coefficient(..)`.

2.6.2.5 Betweenness

Betweenness centrality of a node, B_i is defined as [105],

$$B_i = \sum_{s,t \in V} \frac{\sigma(s,t|i)}{\sigma(s,t)} \tag{2.3}$$

where V is the set of all nodes in a network, and $\sigma(s,t)$ and $\sigma(s,t|i)$ denote the number of all shortest paths from s to t , and the number of the shortest paths through a given node i, respectively. Replacing $\sigma(s,t|i)$ with $\sigma(s,t|e)$ for an edge e, one can also formulate the edge betweenness. The variable B_i measures degree of importance of i for the interactions between subsets of a given network. To compute B_i, Networkx methods `betweenness_centrality(..)` is implemented and $< B >$ is average of B_i over every node in LCC of a given network.

2.7 Supplementary Figures

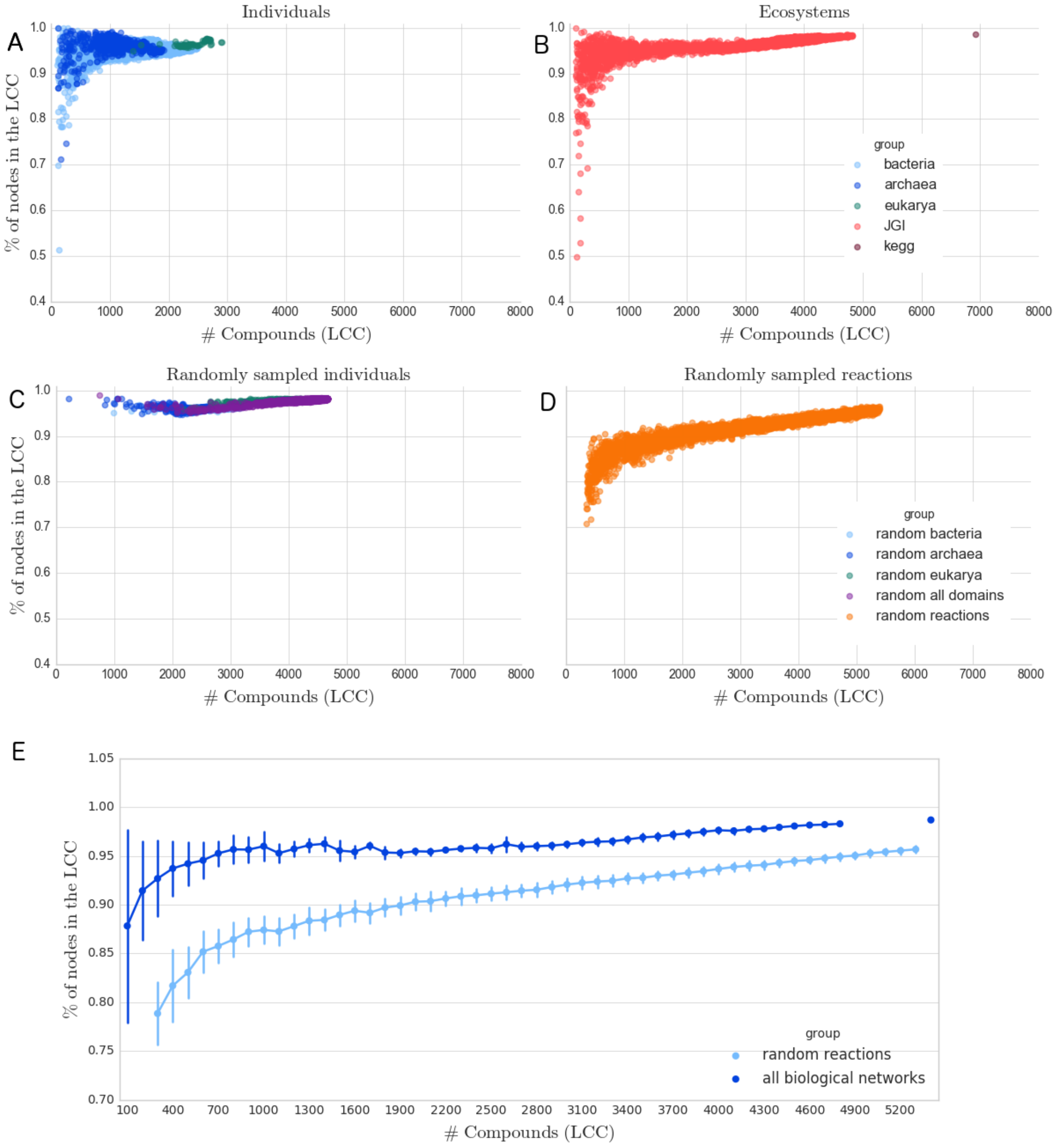


Figure 7. Percentage of nodes in the largest connected component (LCC) of a network versus the size of its LCC. Shown is the percentage of nodes in the LCC, as a function of the size of the network's largest connected component: (A) for all biological individuals (archaea, bacteria, eukarya). (B) for all biological ecosystems (from JGI, KEGG). (C) for randomly sampled individuals (archaea, bacteria, eukarya, and random individuals drawn from all domains). (D) for randomly sampled reactions. (E) Pointplot of biological networks (individuals and ecosystems) and random reaction networks, binned in increments of 100 compound nodes. Bars show one standard deviation of networks within a bin.

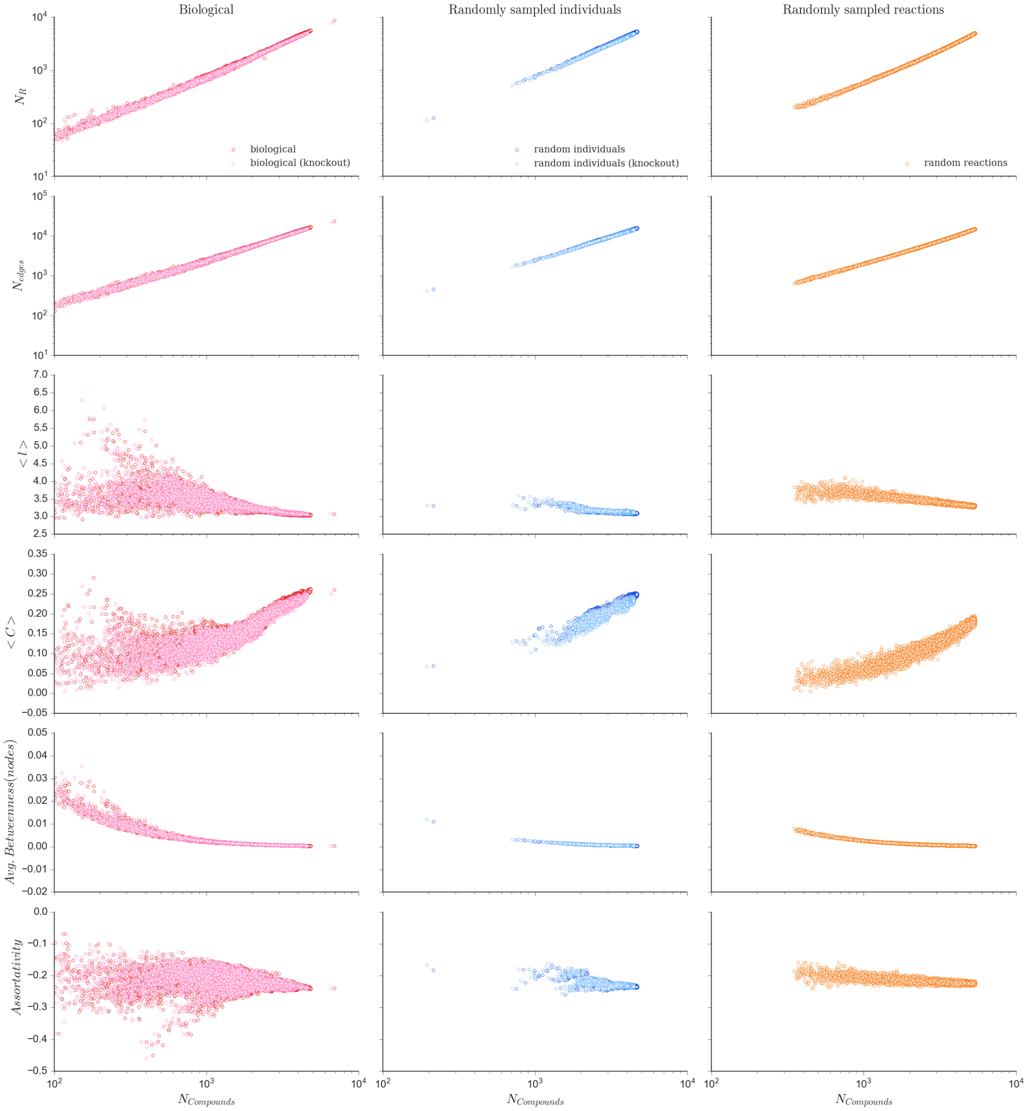


Figure 8. Reaction knockout for unipartite networks. Diversity and topological measures shown for biological networks (left column), randomly sampled individual networks (center column), and randomly sampled reaction networks (right column). Original networks (bold colors) are compared to networks in the same category with 10% of their reactions randomly removed (pale colors). Random reaction networks are shown for comparison, but do not have knocked-out reactions (and cannot, by nature of their construction). Network measure scaling trends are not impacted by the removal of 10% of reactions, indicating our results are robust to missing data. Rows from rom top to bottom show: number of reactions (N_R), number of edges (N_{Edges}), avg. shortest path length ($< l >$), avg. clustering coefficient ($< C >$), avg. betweenness of nodes ($< B >$), assortativity (r).

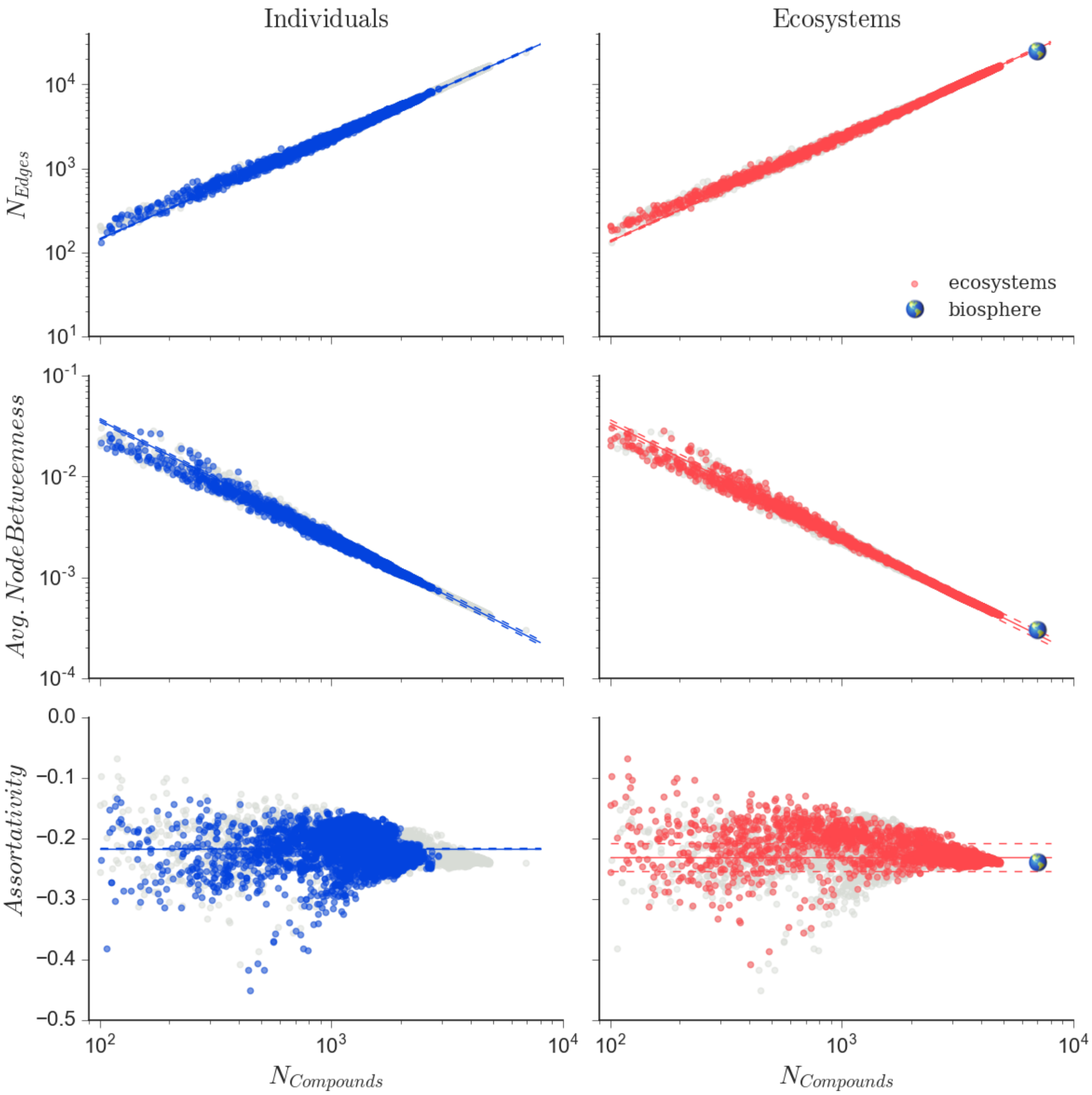


Figure 9. Additional network measures for individuals and ecosystems show universal scaling across levels. Scaling behavior for additional topological measures for unipartite networks, to what is shown in main text Fig. 3. From top to bottom, number of edges (N_{Edges}), average node betweenness ($< B >$), assortativity (r).

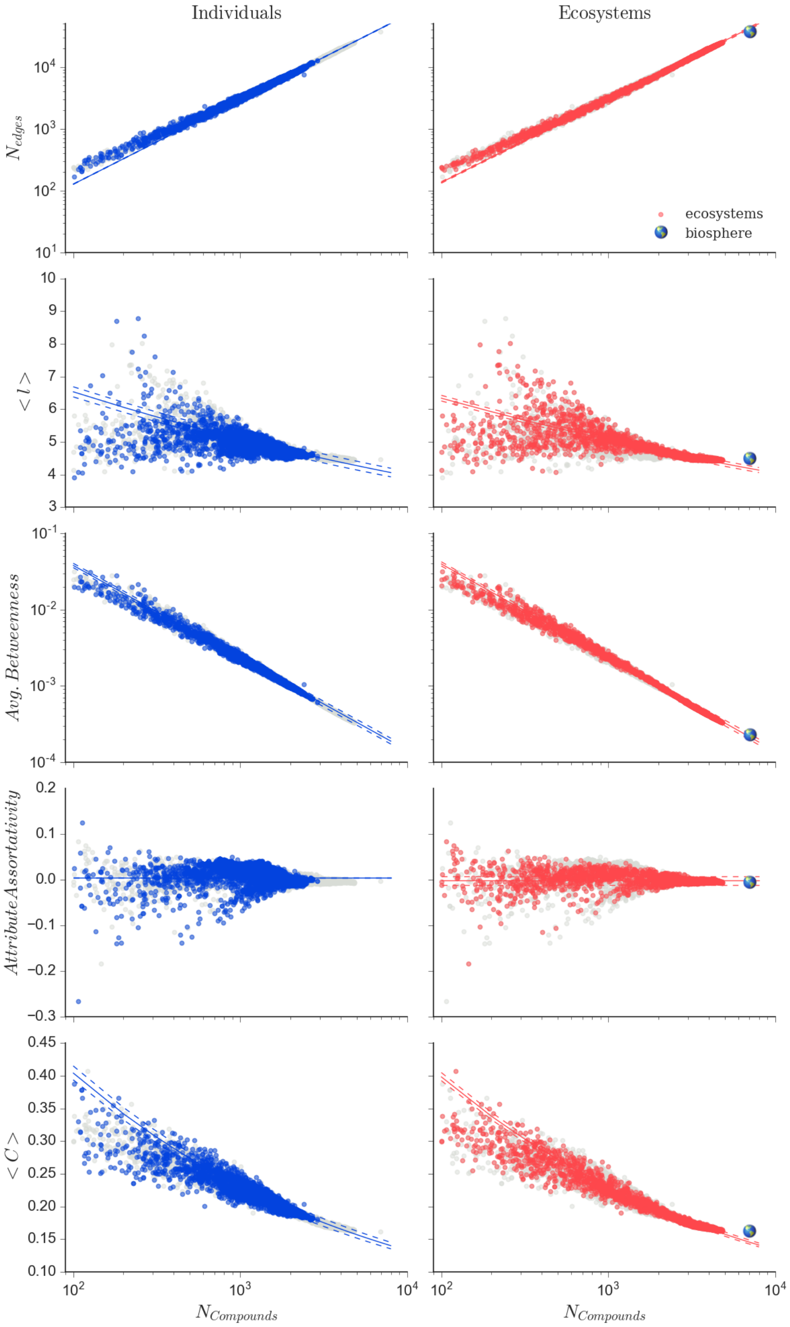


Figure 10. Scaling of bipartite network structure for individuals and ecosystems. Shown are topological measures for bipartite representations of biochemical networks for individuals and ecosystems. Our results demonstrate universal scaling behavior across levels is consistent across both unipartite and bipartite representations. Rows from top to bottom show number of edges (N_{Edges}), average shortest path length ($< l >$), average node betweenness ($< B >$), assortativity (r), and average clustering coefficient ($< C >$).

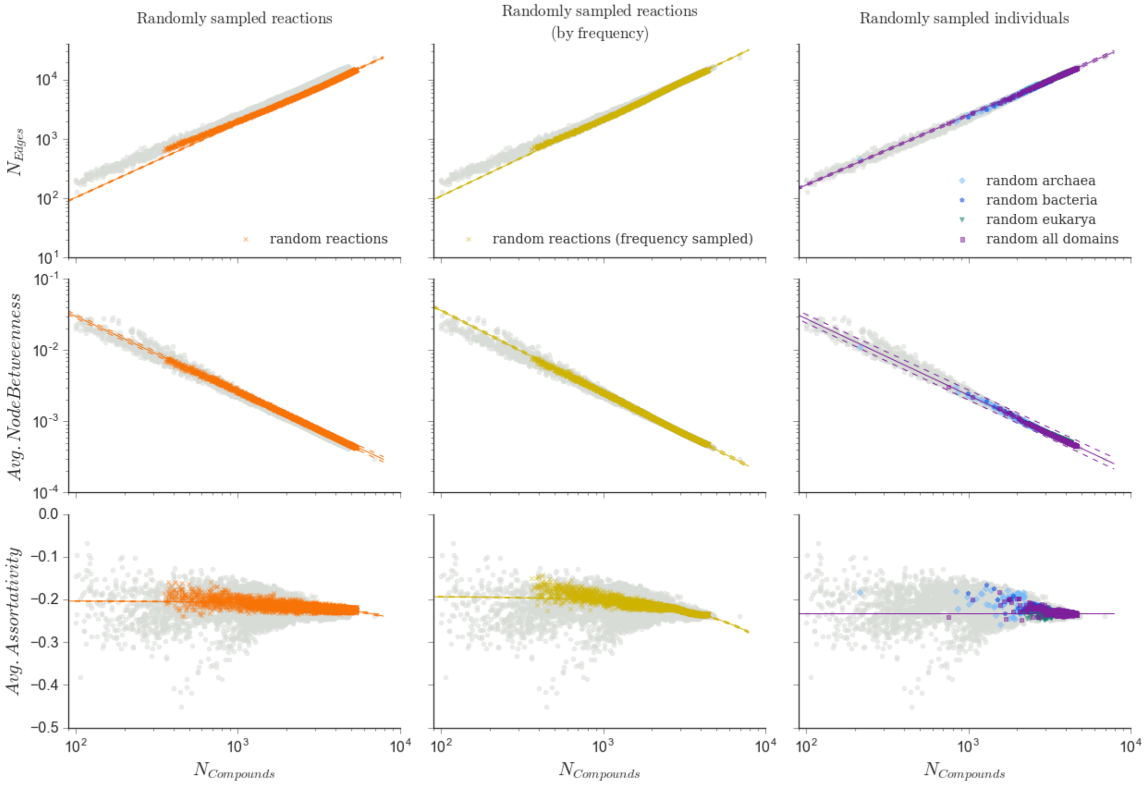


Figure 11. Additional network measures for randomly sampled individuals and randomly sampled reactions. Scaling behavior for additional topological measures to those shown in main text Fig. 4. From top to bottom, number of edges (N_{Edges}), average node betweenness ($< B >$), and assortativity (r).

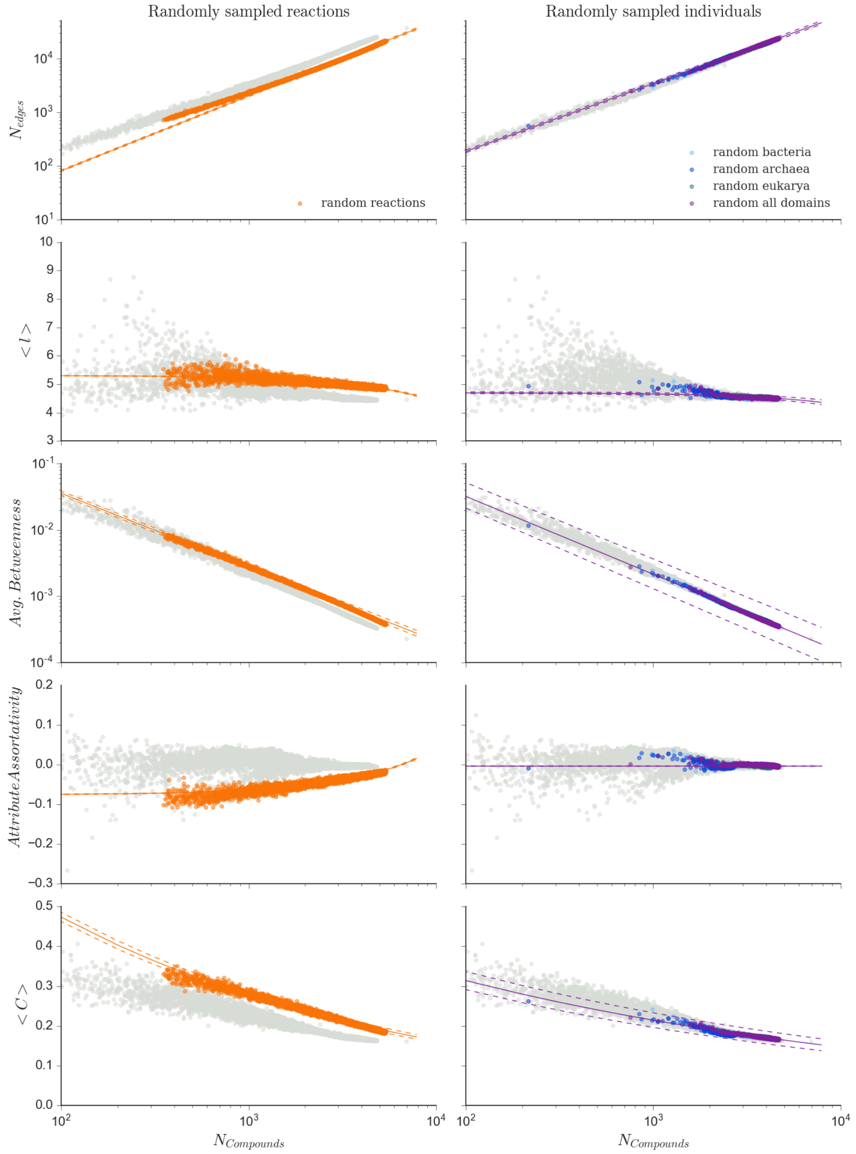


Figure 12. Scaling of bipartite network structure for randomly sampled individuals and randomly sampled reactions. Shown are topological measures for bipartite representations of the random reaction networks and random genome networks. Our results show consistent scaling behavior in comparing the different data sets for both unipartite and bipartite representations. Rows from top to bottom, number of edges (N_{Edges}), average shortest path length ($< l >$), average node betweenness, assortativity (r), and average clustering coefficient ($< C >$).

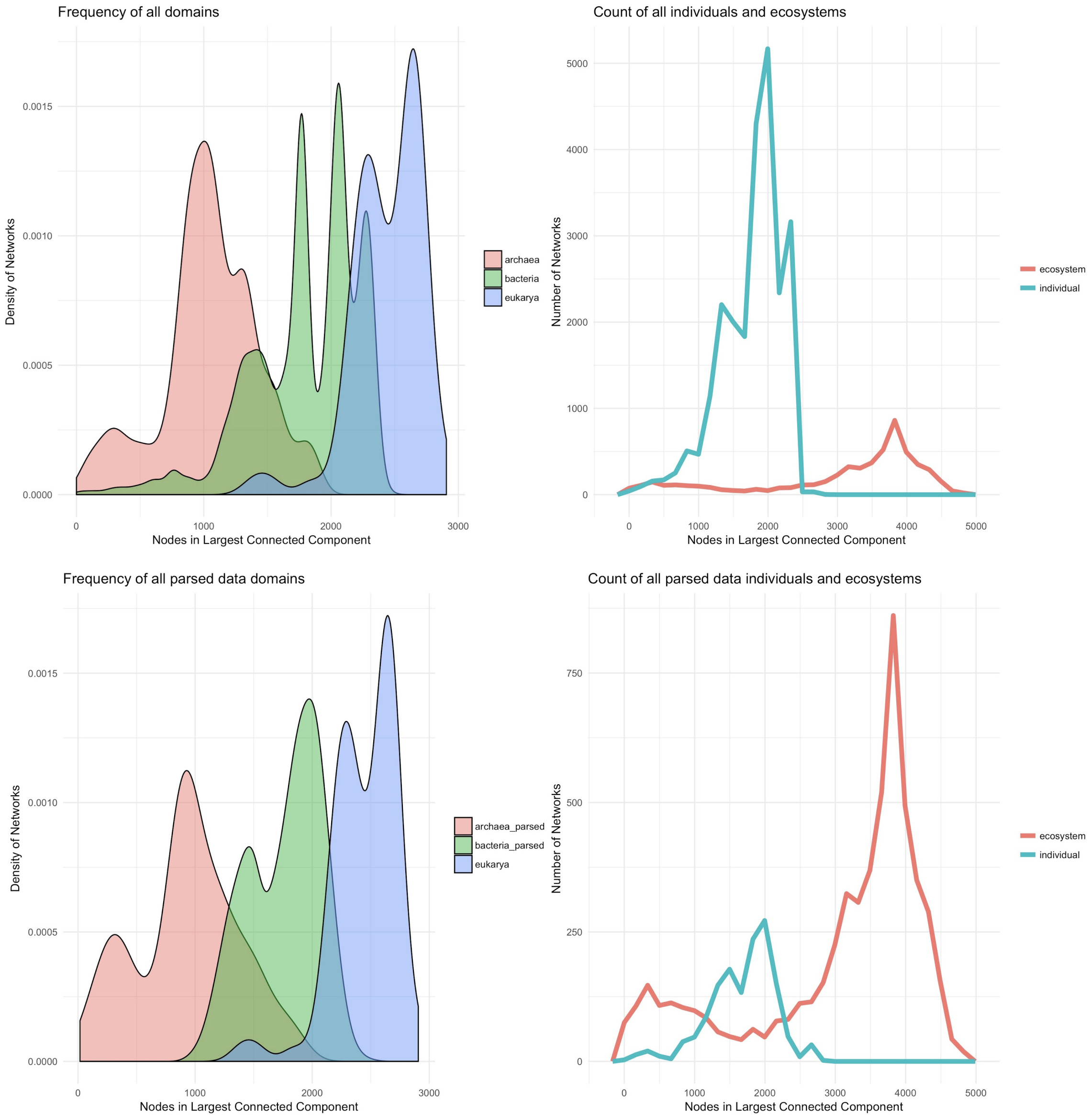


Figure 13. Distributions of network sizes for each domain and across levels of organization. Top row: The relative distribution of network sizes (normalized to 1 over all networks of a given type) for networks in each domain (left), and the total number of networks in individuals and ecosystems (right). Bottom row: The relative distribution of network sizes for networks in each domain, for parsed datasets (left), and the total number of networks in individuals and ecosystems, for parsed datasets (right).

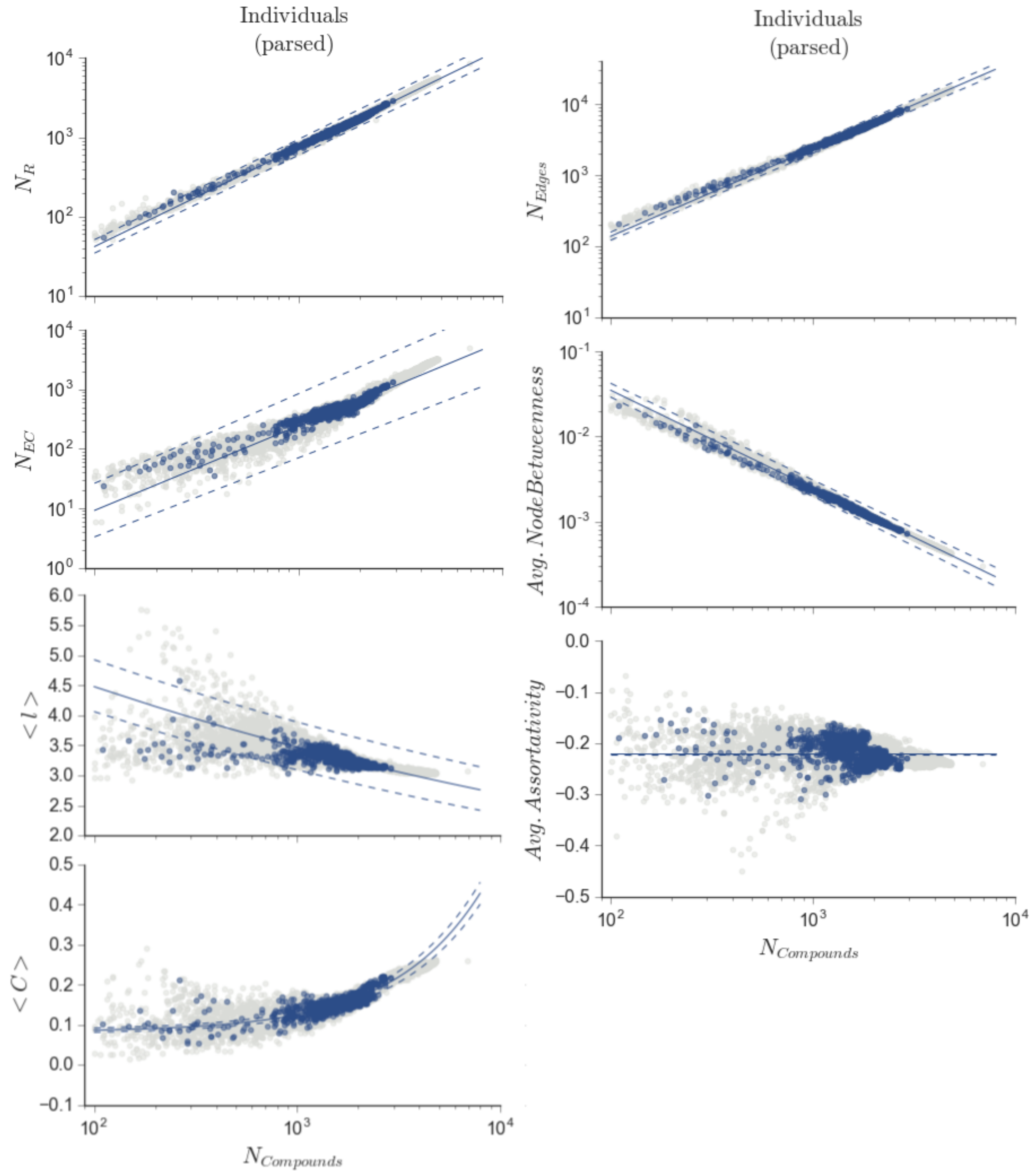


Figure 14. Biochemical diversity and network topology measures for parsed datasets. Shown are data for unipartite representations of the parsed networks we analyzed. Left column, from top to bottom: number of reactions (N_R), number of ECs (N_{EC}), average shortest path length ($< l >$), and average clustering coefficient ($< C >$). Right column, from top to bottom: number of edges (N_{Edges}), average node betweenness ($< B >$), and assortativity (r).

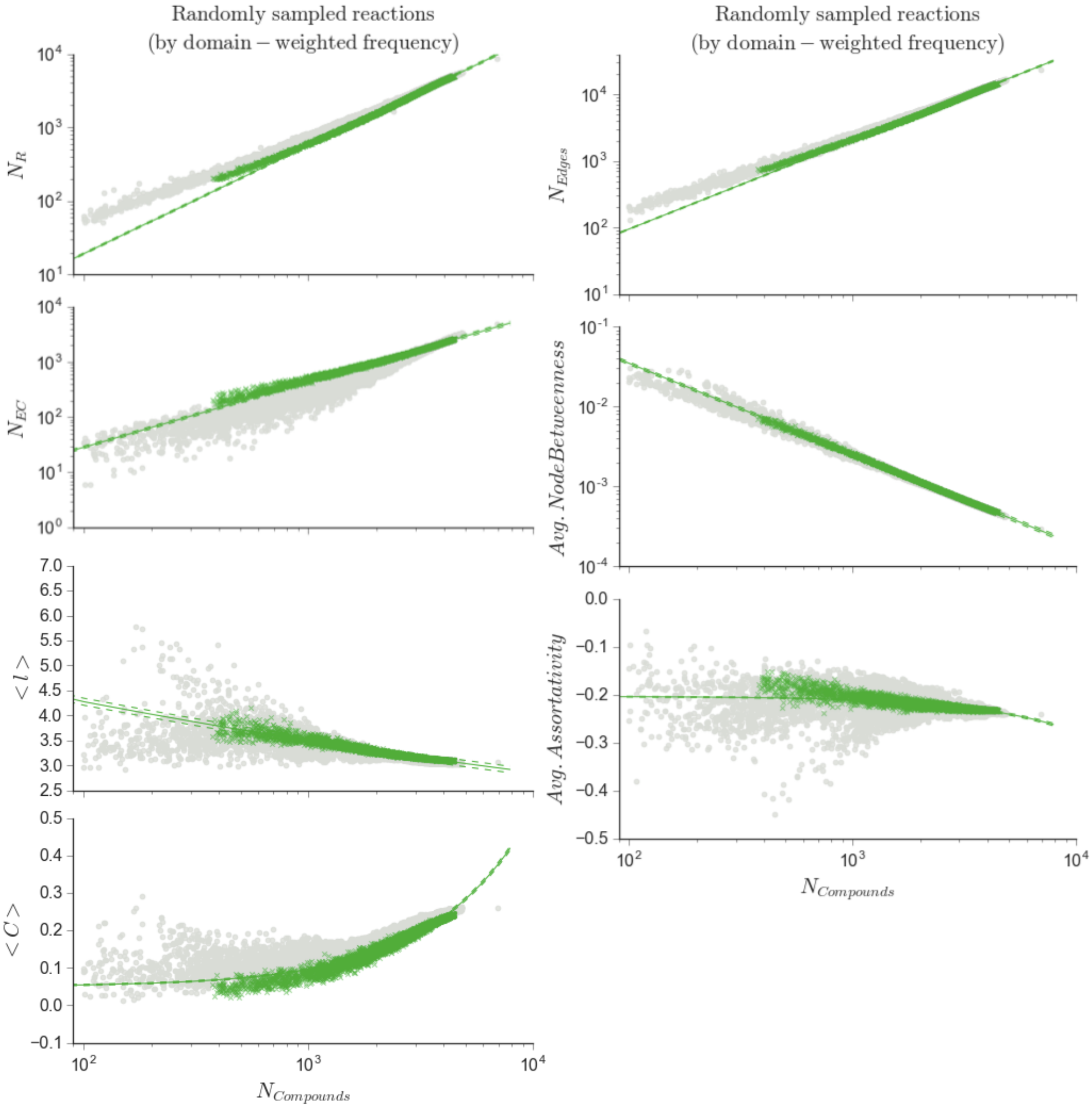


Figure 15. Biochemical diversity and network topology measures for domain-weighted frequency-sampled random reaction networks. Shown are data for unipartite representations of the domain-weighted frequency-sampled random reaction networks we analyzed. These were created by sampling reactions based on the frequency distribution observed within each domain, with reactions from each domain given an equal probability to be sampled. Left column, from top to bottom: number of reactions (N_R), number of ECs (N_{EC}), average shortest path length ($< l >$), and average clustering coefficient ($< C >$). Right column, from top to bottom: number of edges (N_{Edges}), average node betweenness ($< B >$), and assortativity (r).

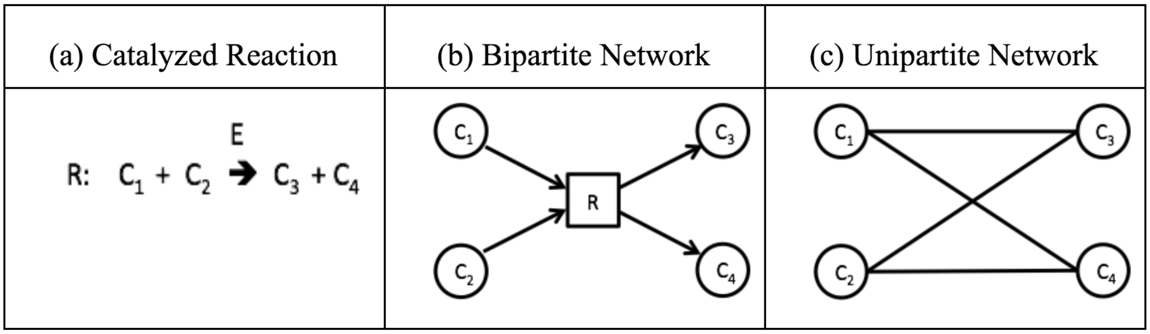


Figure 16. The process to encode a biochemical reaction as a network representation.

2.8 Supplementary Tables

Table 1. Percentage of networks in each dataset with x% of nodes in the LCC.

		x		
	Group	$> 85\%$	$> 90\%$	$> 95\%$
Biological individuals and ecosystems	Archaea	99.17	97.75	86.39
	Bacteria	99.84	99.65	87.53
	Eukarya	100.00	100.00	98.70
	JGI	98.10	97.06	88.42
	KEGG	100.00	100.00	100.00
Random genome	Archaea	100.00	100.00	99.75
	Bacteria	100.00	100.00	100.00
	Eukarya	100.00	100.00	100.00
	JGI	100.00	100.00	100.00
	KEGG	100.00	100.00	100.00
Random reaction	KEGG	95.72	76.86	13.54

Table 2. Distinguishability of individuals and ecosystems, and ecosystems and random genome networks.

Property	Distinguishable Levels of Organization (p-value)	Distinguishability of Ecosystems and Random Genome Networks (p-value)
Number of Reactions, N_R	Yes (10^{-6})	Yes (10^{-5})
Number of Enzyme classes, N_{EC}	Yes (10^{-6})	NA
Average Betweenness (nodes), $< B >$	No (0.272)	No (0.14)
Average Betweenness (edges), $< B_{Edges} >$	No (0.185)	No (0.08)
Number of Edges (LCC), N_{Edges}	Yes (10^{-6})	Yes (10^{-5})
Mean Degree (LCC), $< k >$	Yes (10^{-5})	Yes (10^{-5})
Mean Clustering Coefficient (LCC), $< C >$	Yes (0.00853)	Yes (10^{-5})
Average Shortest Path Length (LCC), $< l >$	No (0.26893)	Yes (10^{-5})
Assortativity (LCC), r	No (0.0761)	No (0.210)
Assortativity for bipartite graphs (LCC), $r_{bipartite}$	No (0.0563)	No (0.256)

2.9 Supplementary Data

Data is not included as part of dissertation, but is available upon reasonable request.

2.9.1 Data file S1

Scaling parameters for topological measures with 95% confidence intervals. Data file S1 contains information for the scaling laws described in the main text. These data describe how various network and enzymatic properties scale with network size (the number of nodes in the largest connected component). This file has 11 columns (plus an index column) which identify the parameters of the fits. Each row is a different fit and each column contains information about the fit. The column entitled `y.var` indicates which network/enzymatic measure is being compared to network size. The column entitled `projection` indicates whether the network measure was applied to the unipartite or bipartite graph representation. The column `level` indicates the biological level of organization, value of `individual` corresponds to a network constructed from genomic data, `ecosystem` indicates a network constructed from metagenomic data, `ranRxn_individual` indicates networks of random biochemical reactions, `syn_individual_all` indicates networks constructed from random combinations of individual networks, `parsed` indicates networks constructed from parsed datasets (except for eukarya), `bio_rand_uni` indicates networks of random biochemical reactions weighted by their occurrence across all individual datasets, `bio_rand_domain` indicates networks of random biochemical reactions weighted by their occurrence within each domain, with each domain's reactions given an equal probability to be included. The column labeled `group` indicates which part of the data set was used, this column only matters for the `individual` level columns. A group value of `bacteria` indicates scaling values for bacterial networks, similarly for the other two domains. The column entitled `scaling` indicates how the measure scales with size, with `powerlaw` meaning that measure scales following a power law, while `linear` means the measure scales linearly. A value of `mean` in the scaling column is used to show the measure does not scale with size. The remaining 6 columns contain numerical values for the scaling fits and their 95% confidence intervals. The mathematical meaning of these values depends on the scaling behavior of that measure (i.e. the corresponding value in the `scaling` column). The value of `alpha` is always related to how the measure changes with size, while `beta` is always related to the intercept. If the scaling behavior is linear, then the measure scales according to $y.var \sim alpha * (size) + beta$, such that `alpha` is the slope of the line and `beta` is the intercept. If the scaling behavior is a power law, then the measure scales according to $y.var \sim exp(beta) * (size)^{alpha}$, such that `alpha` is the scaling exponent and `exp(beta)` is the intercept. The 95% confidence intervals have the same interpretation with the `alphaP` column indicating the upper bound on `alpha` and the `alphaM` column indicating the lower bound on `alpha`, the same convention is used for `betaP` and `betaM`. Measures that do not scale with size have values of zero in the alpha column, and the mean value is given in the beta column, with 95% of the distribution falling between `betaM` and `betaP`.

2.9.2 Data file S2A.

Summary of measured network properties, by domain. Data file S2A contains a statistical summary of network properties, grouped by domain. Each row in the first column contains the name of the partition of data being described in that row, with “JGI” indicating the metagenomic data. Each column in the first row identifies the property which is being summarized in the rows below. The properties are as follows: nbr_rxn is the number of reactions encoded by the genome/metagenome; nbr_nodes is the number of nodes in the network; nbr_edges is the number of edges in the network; nbr_connected_components is the number of connected components in the network; nbr_nodes_lcc is the number of nodes in the largest connected component (LCC) of the network; nbr_edges_lcc is the number of edges in the LCC of the network; ave_degree_lcc is the average node degree in the LCC of the network; ave_clustering_coeff_lcc is the average clustering coefficient in the LCC of the network; ave_shortest_path_length_lcc is the average shortest path length in the LCC of the network; ave_betweenness_nodes_lcc is the average node betweenness in the LCC of the network; ave_betweenness_edges_lcc is the average edge betweenness in the LCC of the network; assortativity_lcc is the average assortativity in the LCC of the network; attribute_assortativity_lcc is the average attribute assortativity of the LCC of the network; diameter_lcc is the diameter of the LCC of the network; nbr_ecs is the number of enzyme commission numbers in the network. The statistical property being measured over all networks in a group, for a particular measure, are listed in each cell. The statistical properties are the count (number of networks), mean, std (standard deviation), minimum, maximum, and the quartiles.

2.9.3 Data file S2B.

Summary of measured network properties, by levels (parsed data only). Data file S2B contains a statistical summary of network values, grouped together for the parsed networks (parsed archaea, parsed bacteria, and all eukarya). The format of the csv is otherwise identical to S2A (see description above).

2.9.4 Data file S2C.

Summary of measured network properties, by levels (parsed data excluded). Data file S2C contains a statistical summary of network values, grouped together by level for all data, excluding the parsed networks. “Individual” includes archaea, bacteria, and eukarya, and “ecosystem” includes all JGI metagenomic networks. The format of the csv is otherwise identical to S2A (see description above).

Chapter 3

NO LOVE FOR POWER-LAWS IN BIOCHEMICAL NETWORKS

This chapter was written in collaboration with Hyunju Kim, and Sara I. Walker. It has been submitted to PLoS Computational Biology.

3.1 Abstract

Biochemical reactions underlie all living processes. Like many biological and technological systems, their complex web of interactions is difficult to fully capture and quantify with simple mathematical objects. Nonetheless, a huge volume of research has suggested many real-world biological and technological systems—including biochemical systems—can be described rather simply as 'scale-free' networks, characterized by a power-law degree distribution. More recently, rigorous statistical analyses across a variety of systems have upended this view, suggesting truly scale-free networks may be rare. We provide a first application of these newer methods across two distinct levels of biological organization: analyzing a large ensemble of biochemical networks generated from the reactions encoded in 785 ecosystem-level metagenomes and 1082 individual-level genomes (representing all three domains of life). Our results confirm only a few percent of individual and ecosystem-level biochemical networks meet the criteria necessary to be anything more than super-weakly scale-free. Leveraging the simultaneous analysis of the multiple coarse-grained projections of biochemistry, we perform distinguishability tests across properties of individual and ecosystem-level biochemical networks to determine whether or not they share common structure,

indicative of common generative mechanisms across levels. Our results indicate there is no sharp transition in the organization of biochemistry across distinct levels of the biological hierarchy—a result that holds across different network projections. This suggests the existence of common organizing principles operating across different levels of organization in biochemical networks, which can best be elucidated by analyzing all possible coarse-grained projections of biochemistry across all scales in tandem.

3.2 Author Summary

Fully characterizing living systems requires rigorous analysis of the complex webs of interactions governing living processes. Here we apply new statistical approaches to analyze a large data set of biochemical networks across two levels of organization: individuals and ecosystems. We find that independent of level of organization, the standard 'scale-free' model is not a good description of the data. Interestingly, there is no sharp transition in the shape of degree distributions for biochemical networks when comparing those of individuals to ecosystems. This suggests the existence of common organizing principles operating across different levels of biochemical organization that can best be elucidated by considering multiple coarse-grained representations in tandem, warranting further research to explain.

3.3 Introduction

Statistical mechanics was developed in the 19th century for studying and predicting the behavior of systems with many components. It has been hugely successful in its application to those physical systems well-approximated by idealized models of non-interacting particles. However, real-world systems are often much more complex, leading to a realization over the last several decades that new statistical approaches are necessary to describe biological and technological systems. Among the most natural mathematical frameworks for developing the necessary formalism is network theory, which projects the complex set of interactions composing real systems onto an abstract graph representation [17, 28–33]. Such representations are powerful in their capacity to quantitatively describe the relationship between components of complex systems and because they permit inferring function and dynamics from structure [38–42].

Network theory has been especially useful for studying metabolism. Metabolism consist of catalyzed reactions that transform matter along specific pathways, creating a complex web of interactions among the set of molecular species that collectively compose living things [48–52]. It is the collective behavior of this system of reactions that must be understood in order to fully characterize living chemical processes–counting only individual components (molecules) is inadequate. The structure of how those components interact with one another (via reactions) really matters: in fact it is precisely what separates organized biological systems from messy chemical ones [13, 43, 44].

Within the formalism of network theory, one of the simplest ways to capture insights into the global structure of a network is to analyze the shape of its degree distribution. A huge volume of research into various complex biological, technological and social networks has therefore focused on identifying the scaling behavior of the corresponding degree distributions for network projections describing those systems. One of the most significant results emerging from these analyses is that many networks describing real-world systems exhibit ostensibly “scale-free” topology [70–74], characterized by a power-law degree distribution. The allure of scale-free networks is in part driven by the simplicity of their underlying generative mechanisms, for example a power-law degree distribution can be produced by relatively simple preferential attachment algorithms [70], or to a lesser extent through optimization principles [75]. For truly scale-free networks the probability to find a node with degree x should scale as:

$$f(x) = x^{-\alpha} \ . \tag{3.1}$$

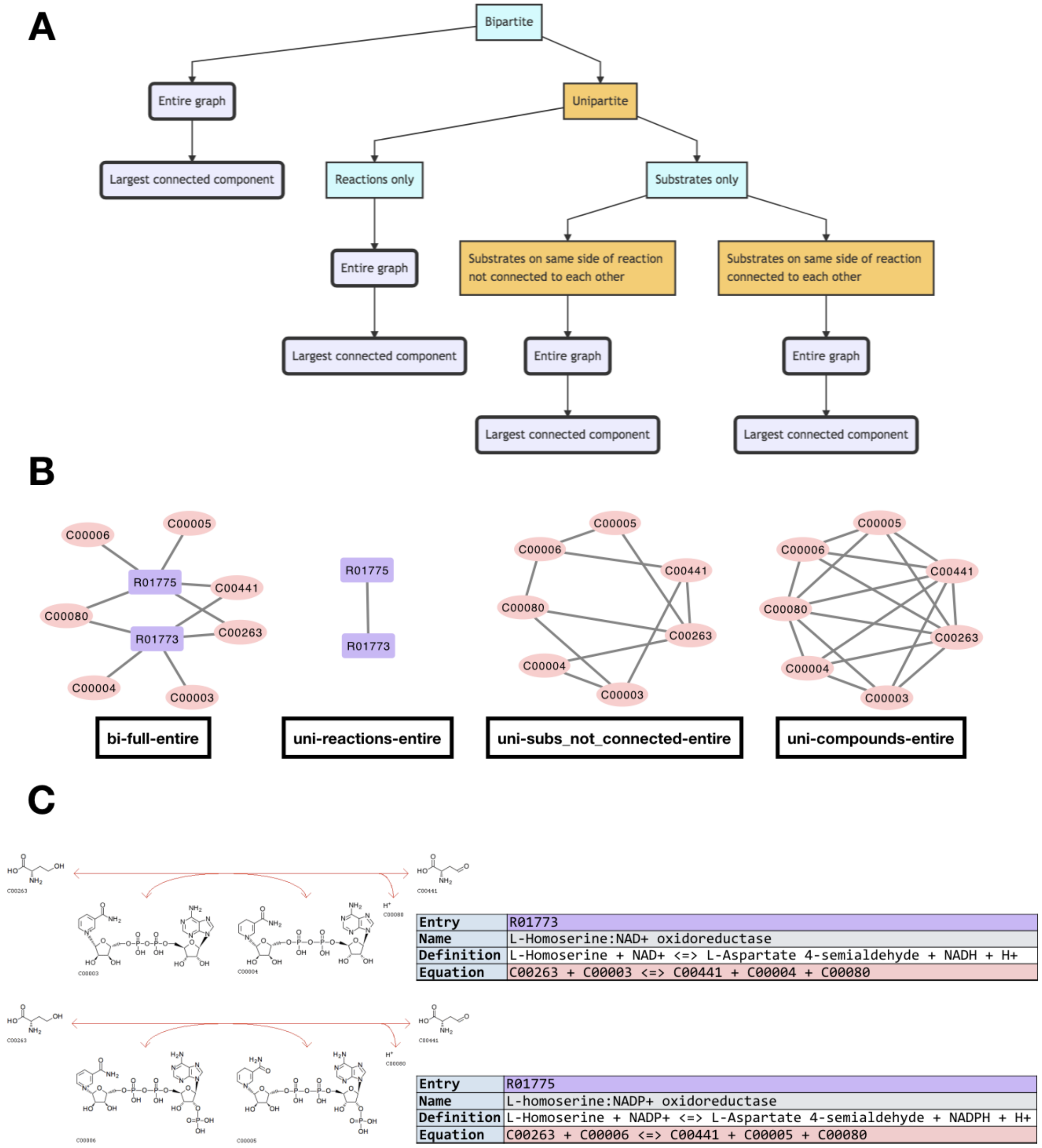
For numerous biological and technological systems, including metabolic networks, the scaling exponent, α, is reported with values in the range $2 < \alpha < 3$. The apparent ubiquity of scale-free networks across biological, technological and social networks has fueled some to conjecture scale-free topology as a unifying framework for understanding all such systems, with the enticing possibility these seemingly diverse examples could in reality arise from relatively simple, universal generating mechanisms [70, 74, 75, 106, 107].

However, this story is far from complete. Recently developed statistical tests to rigorously examine whether observed distributions share characteristics with a power-law, or are instead more similar to other heavy tailed distributions, have revealed that true scale-free networks may not be as ubiquitous as previously supposed [69, 108]. These tests reveal that while it is superficially possible for a network to appear scale-free, more rigorous analysis can reveal a structure more similar to other heavy-tailed distributions such as the log-normal distribution, or even non heavy-tailed distributions like the exponential distribution [69, 107–109].

The problem of characterizing the global structure of real-world systems is further compounded by the fact there are often many ways to coarse-grain a real system to generate a network representation, each corresponding to a different way for set of interactions to be projected onto a graph. For example, metabolic networks may be represented as unipartite or bipartite graphs, depending on whether one chooses to focus solely on the statistics over molecules (or reactions) and their interactions (requiring a unipartite representation) or instead to include both molecules and reactions as explicit nodes in the graph (where molecules and reactions represent two classes of nodes in a bipartite representation) [45–47]. These graphs can have different large-scale topological properties, even when projected from the same underlying system. This raises the question of determining which projection to analyze, and whether or not a real-world system should be considered "scale-free" if only some of its network projections exhibit power-law degree distributions. Broido and Clauset recently developed a methodology to compare the degree distributions of network projections of different complexities, classifying the degree to which they are scale-free on a scale from "Not scale-free" all the way to "strongest" [108]. This provides a framework for statistically analyzing many projections of a given system to determine how well scale-free structure describes the real underlying system when projected onto its different coarse-grained representations.

Herein, we build from the work of Broido and Clauset with specific application to the problem of characterizing biochemical systems. A novelty in our approach is recognizing that in order to really understand the structure of real-world biological (and technological) systems, the relevant scale(s) for performing such analysis must also be considered. In particular, many biological and technological systems are hierarchical, with networks describing interactions across multiple levels. For example, one may study the biochemistry of individual species, but ultimately the function of an individual in a natural system depends on a complex interplay of interactions among the many species comprising its host ecosystem. In this way, biochemistry is hierarchically organized into individuals and ecosystems. Indeed, much discussion about universal properties of life has shifted focus from individuals to ecosystems as the relevant scale best capturing the regularities of biological organization [21, 84]. It is unclear at present whether analysis of biochemical networks at the level of individuals or ecosystems will best uncover their structure and permit identifying generative mechanisms for biology, or whether all levels must be considered simultaneously.

In what follows, we perform statistical analysis of an ensemble of biochemical systems generated from 785 ecosystem-level metagenomes and 1082 individual-level genomes (representing all three domains of life). Our results include the first analysis of scale-free network structure for the different projections of ecosystem-level biochemistry, significantly expanding on on earlier work focusing on the large-scale organization of individual metabolic networks only [45–48, 69, 108]. Like Broido and Clauset, we consider all possible projections of biochemical systems to graphs simultaneously, whereas most prior work on the organization of biochemistry has only considered one or at most a few projections [52, 110–113]. We find a majority of biochemical networks are not scale-free, independent of projection or level of organization. We also demonstrate how the network properties analyzed herein can be used to distinguish individual and ecosystem level networks, and find that independent of projection, individuals and ecosystems share very similar structure. These results have potentially deep implications for identifying underlying rules of biochemical organization at both the individual and ecosystem-level by providing constraints on whether the same or different generative mechanisms could operate to organize biochemistry across multiple scales.



Entry	R01773
Name	L-Homoserine:NAD+ oxidoreductase
Definition	L-Homoserine + NAD+ <=> L-Aspartate 4-semialdehyde + NADH + H+
Equation	C00263 + C00003 <=> C00441 + C00004 + C00080

Entry	R01775
Name	L-homoserine:NADP+ oxidoreductase
Definition	L-Homoserine + NADP+ <=> L-Aspartate 4-semialdehyde + NADPH + H+
Equation	C00263 + C00006 <=> C00441 + C00005 + C00080

Figure 17. How biochemical datasets are decomposed into network projections. (A) Networks are generated from the set of reactions encoded in each genome/metagenome starting from a bipartite representation, and projecting different combinations of attributes. The bold, rounded flowchart nodes show the result of each combination of projections applied in this study. (B) The different network projection types of a simple example dataset, composed of two KEGG reactions: R01773 & R01775. The nomenclature used in this paper's figures is below each network visualization (in this example the entire graph is the same as the largest connected component). (C) How the reactions used in the network visualization example above appear in the KEGG database [58, 114, 115]

3.4 Results

Utilizing the framework developed by Broido and Clauset [108], we used the full set of biochemical reactions encoded in each genome and metagenome to construct eight distinct network representations of each respective biochemical system. This resulted in 8656 network projections for the 1082 individual-level biochemical datasets, and 6280 network projections for the 785 ecosystem-level biochemical datasets. Each representation corresponds to a different coarse-graining of the underlying system of reactions (i.e. the underlying dataset) (Fig. 17). We determine whether or not these datasets are scale-free, and analyze the aspects of them, and their diverse projections, that tend to lend themselves to be more or less scale-free. The alternative distributions that we compare to the power-law are: The exponential distribution, the log-normal distribution, the stretched exponential distribution, and the power-law distribution with a cutoff (see [69, 108] for more details on these distributions).

We first classified each dataset in terms of how scale-free it is. A dataset is classified as: *Super-Weak* if for at least 50% of network projections, none of the alternative distributions are favored over the power-law; *Weakest* if for at least 50% of network projections, the power-law hypothesis cannot be rejected ($p \geq 0.1$); *Weak* if it meets the requirements of the Weakest set, and there are at least 50 nodes in the distribution's tail ($n_{\text{tail}} > 50$); *Strong* if it meets the requirements of both the Super-Weak and Weak set, and that the median scaling exponent is between two and three ($2 < \hat{\alpha} < 3$); and *Strongest* if it meets the requirements of the Strong set for at least 90% of graphs, rather than 50%, and for at least 95% of graphs none of the alternative distributions are favored over the power-law.

Our results indicate most biochemistry at the individual and ecosystem-level is characterized by networks that are “super-weakly” scale-free (Fig. 18). That is, while the power-law is better than other models for fitting the shape of their degree distributions, the power-law is not itself a good model. When doing a goodness-of-fit test, we find that the majority of network representations of each genomic/metagenomic dataset have $p < 0.1$, indicating there is less than a 10% chance that our data is truly power-law distributed. This effectively rules out the possibility that our data is drawn from a power-law shaped degree distribution, despite the fact that, when compared to other distributions through log-likelihood ratios, 99% of all datasets do not favor alternative heavy tail distributions for the majority of their network-projections (Fig. 19, top row).

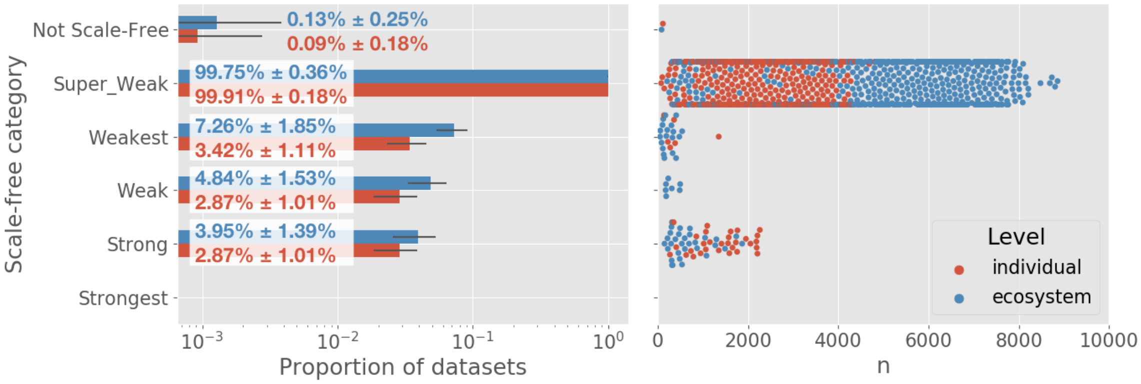


Figure 18. The vast majority of individual and ecosystem level networks are not “scale-free”. Left: Most datasets are super weak, indicating that when compared to other models, a power-law distribution is a better fit. However, the power-law distribution is not a “good” fit for most dataset network representations. No networks meet the “Strongest” criteria defined by Broido and Clauset al.[108]. Overlaid values show the percent of networks of each level which fall into each category, ±2SD. Right: The relationship between scale-freeness and largest network size across projections n. All datasets containing networks larger than approximately 2100 nodes have degree distributions that rule out fitting well to a power-law.

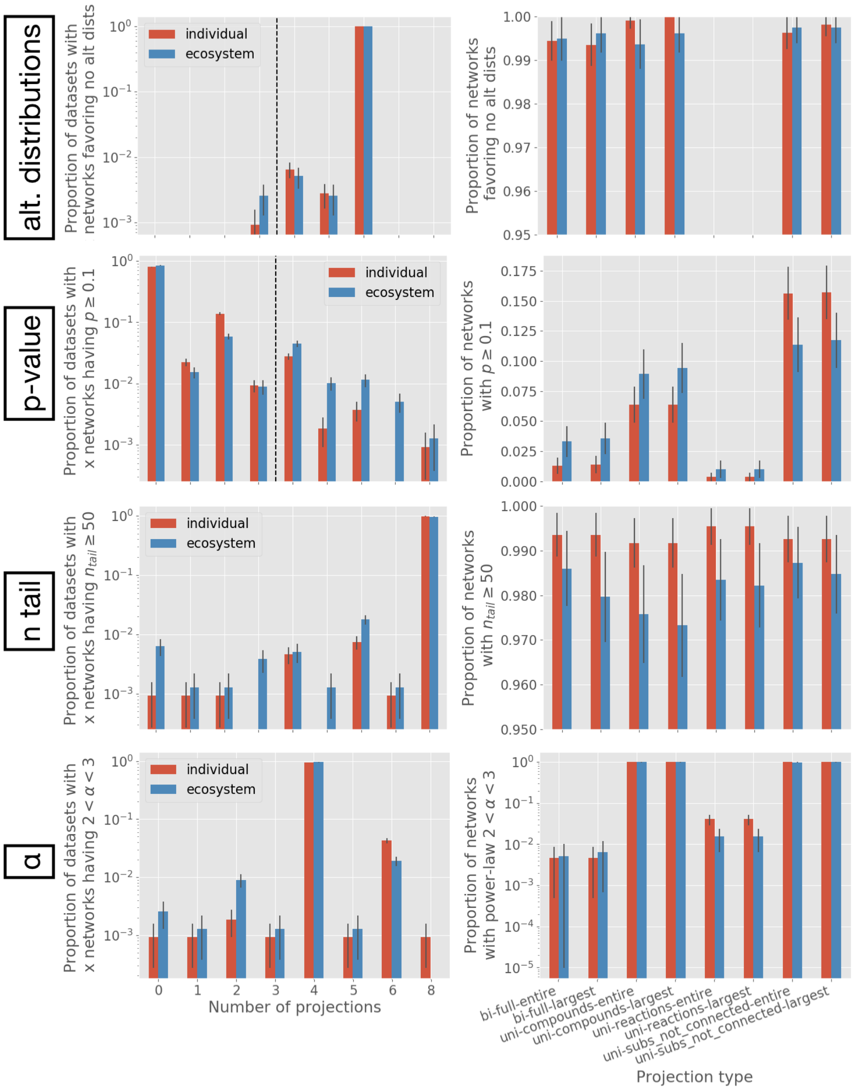


Figure 19. The number of network projections within each dataset which meet some scale-free criteria. Left column: The number of network projections within each dataset which meet some scale free-criteria, where each dataset falls into one of nine bins. Normalized to total number of datasets in a level. Criteria from top to bottom: No alternative distributions favored over power-law in log-likelihood ratio (1st row); $p \geq 0.1$ (2nd row); $n_{tail} > 50$ (3rd row); $2 < \alpha < 3$ (4th row). Dashed lines show: the cutoff for number of networks in a dataset required to meet the threshold criteria for "Super-Weak" (1st row), and "Weakest" (2nd row). Right column: The number of network projections, across all datasets, which meet some scale-free criteria, binned by projection type. Normalized to the total number of each projection within a level. Criteria same as left column. Red bars indicate individual-level datasets/networks, and blue bars indicate ecosystem-level datasets/networks. Black error bars show ±2SD.

3.4.1 Where biochemical systems succeed and fail scale-free classifications

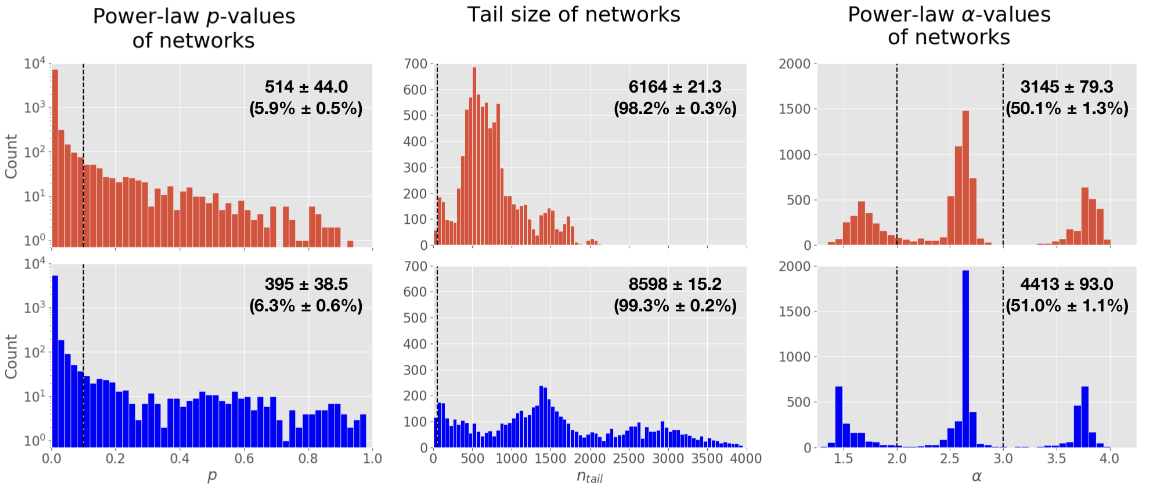


Figure 20. The distribution of p-values, tail-sizes, and power-law alpha values for biochemical network degree distributions, over all network projections. Left column: The goodness-of-fit p-values of networks. When $p \geq 0.1$ (dashed line), it indicates that there is at least a 10% chance of the power-law distribution being a plausibly good fit to a network's degree distribution. Center column: Tail size of networks. When $n_{tail} \geq 50$ (dashed line), it indicates that the tail of distribution is large enough to reliably fit. Right column: Power-law exponent α values of networks. When $2 < \alpha < 3$ (between dashed lines), it indicates that a network meets the criteria of having a power-law exponent which falls into scale-free territory. The top row shows distributions for individuals in red. The bottom row shows distributions for ecosystems in blue. Insets indicate the number (and percent) of networks which meet the criteria, ±2SD.

3.4.1.0.1 Goodness-of-fit p-value.

The "weakest" requirement for a scale-free network introduced by Broido and Clauset stipulates over 50% of a dataset's network-projections must have a power-law goodness-of-fit $p \geq 0.1$. For both individuals and ecosystems, only 6% of network-projections meet this requirement (Fig. 20, left column). This goodness-of-fit p-value requirement is the most restrictive of all scale-free requirements.

3.4.1.0.2 Tail size.

Setting aside the fact each subsequent scale-free requirement builds on the requirement(s) of the preceding one, we find 98% of individual networks and 99% of ecosystem networks *do* meet the requirement of $n_{tail} > 50$ for a scale-free degree distribution (Fig. 20, center column).

3.4.1.0.3 The power-law exponent, α.

Only 50% of individual-level networks and 51% of ecosystem-level networks meet the requirement that $2 < \alpha < 3$ for their degree distribution. The goodness-of-fit p value requirement, followed by the requirement constraining values of α, are the most restrictive when determining whether a biochemical network's degree distribution should be considered scale free (Fig. 20, right column).

3.4.2 Meeting the threshold for scale-free classification is dependent on the network representation

We find the results of each requirement listed above for classifying topology as scale-free differ across the eight network projection types for each dataset. Unsurprisingly, for most requirements, there exists a minute difference between the values observed for the largest connected component and entire graph of a given network projection type (Fig. 19, right column). Depending on the measure, there is a noticeably larger difference between the major network projection types, e.g., between bipartite, unipartite-reactions, unipartite-compounds (where all substrates participating in the same reaction are connected), and unipartite-compounds (where substrates on the same side of a reaction are not connected) (Fig. 19, right column).

3.4.2.0.1 Comparing to alternative distributions.

Over 99% of individual and ecosystem-level datasets have 6 projections which do not favor any other distribution over the power-law (Fig. 19, top row, left column). No datasets have more than 6. The other two projections nearly always favor at least one other distribution over the power-law distribution–either the log-normal, exponential, stretched exponential, or power-law with exponential cutoff (Fig. 19, top row, right column). There are only 3 of the 6280 ecosystem-level network projections (across the 785 ecosystem-level datasets) that do not favor at least one of the alternative distributions. Oftentimes all four are favored over the power-law distribution (Fig. 21, rows 3-4). These results are identical, within 95% confidence, for both individuals and ecosystems.

3.4.2.0.2 Goodness-of-fit p-value.

Out of all datasets, 80% of individuals and 84% of ecosystems have only a single projection type with $p \geq 0.1$ for a power-law fit to their degree distribution. This indicates the majority of datasets would still not meet the “weakest” requirement for scale-free even with a threshold that lowered the percent of a dataset’s projections needed to 25% (2 networks) instead of 50% (4 networks) (Fig. 19, 2nd row, left column). The unipartite projection where substrates on the same side of a reaction are not connected (unipartite-subs_not_connected) was the most likely to satisfy $p \geq 0.1$. For the two unipartite-compound projections, the difference between individuals and ecosystems is within the error. The unipartite-reaction projections were the least likely to satisfy $p \geq 0.1$, which is consistent with the observation that these networks always favor an alternative distribution as a better fit to the data than the power-law (Fig. 19, 2nd row, right column). As we initially reported, the majority of datasets do not meet the p-value threshold for being considered scale-free, although ecosystems-level datasets are more likely to meet the threshold.

3.4.2.0.3 Tail size.

Out of all datasets, 98% of individuals and 96% of ecosystems meet $n_{tail} > 50$ for all projection types (Fig. 19, 3rd row, left column). For 7 of the projection types, there is no difference between individuals and ecosystems, within 95% confidence (Fig. 19, 3rd row, right column).

3.4.2.0.4 The power-law exponent, α.

Out of all datasets, 95% of individuals and 97% of ecosystems meet $2 < \alpha < 3$ for 4 of 8 projection types (Fig. 19, bottom row, left column). The two types of unipartite-compound networks contribute to the datasets which meet the alpha-range requirement the majority of the time. That is, chances are if a dataset has at least 4 projection types meeting $2 < \alpha < 3$, two of them are going to be unipartite-compound network projections (Fig. 19, bottom row, right column). The results are similar for both individuals and ecosystems.

3.4.2.0.5 Correlation of results between projections.

Because 8 different network projections are derived from a single biochemical dataset, there is reason to expect the proportions of each projection type meeting any given scale-free criteria are correlated. We therefore constructed a Pearson correlation matrix to test whether there are correlations between projections (Fig. 22). Unsurprisingly, we find that values from projections of a network's LCC and entire graph are highly correlated. All types of unipartite compound networks tend to be correlated. Values across many other projection types are barely correlated for the p-value and n_{tail} criteria. Ecosystems tend to show more correlation, across all projection types, than individuals.

3.4.3 Distinguishing individuals and ecosystems based on their degree distributions

3.4.3.0.1 Multinomial regression.

We used multinomial regression on network and degree distribution data from the above analyses to attempt to distinguish individuals from ecosystems. Most measures cannot reliably distinguish between these two levels of organization, with only network size and network tail size data distinguishing the two levels better than chance. Using only network size, ecosystems could be correctly identified in test data 72.23% of

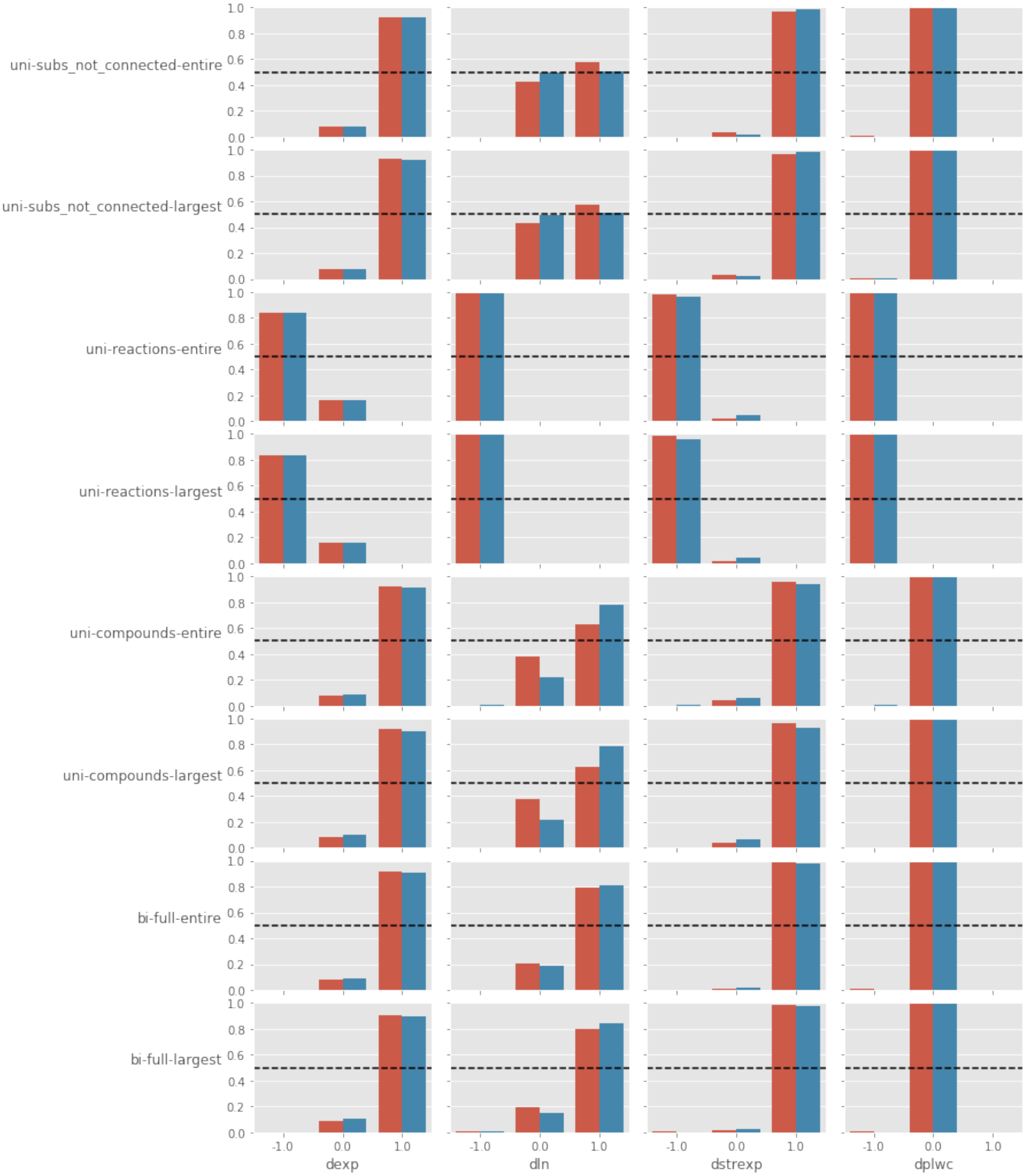


Figure 21. How alternative distributions compare to the powerlaw across each network projection type. The proportion of network projections, across all datasets, that favor either the power-law distribution (1.0), an alternative distribution (-1.0), or are inconclusive (0.0). Each row shows a different network projection type. Each column is a different distribution with which the power-law is being compared to. From left to right is the exponential; log-normal; stretched exponential; and power-law with cutoff. Dashed line is constant at proportion = 0.5 across all subplots. Red bars indicate individual-level networks, and blue bars indicate ecosystem-level networks.

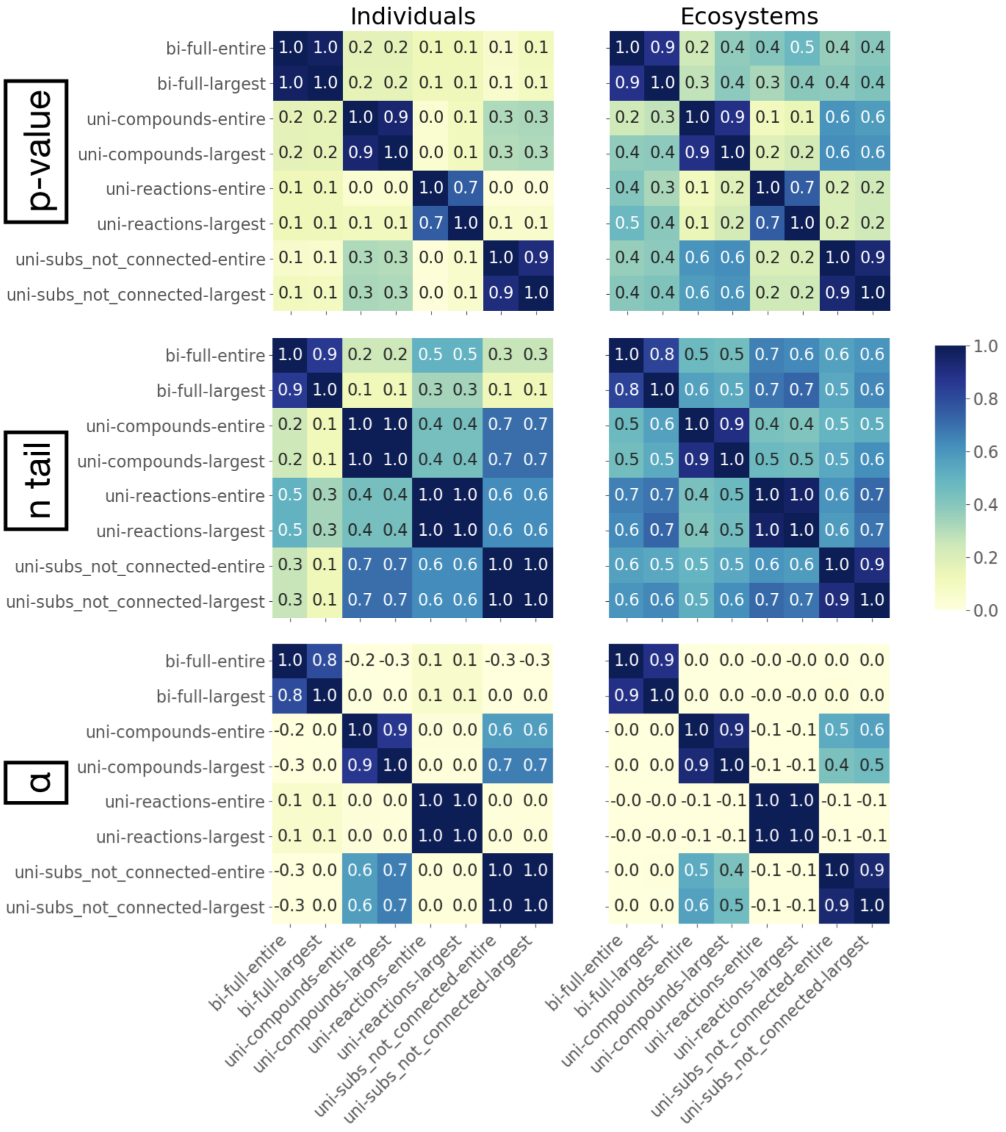


Figure 22. Correlations between network projections which meet scale-free criteria. Correlation matrix heatmaps show type how different types of network projections correlate in their proportions of networks which meet some scale-free criteria. Rows are for each of the different scale-free criteria (p-value, n_{tail} and α), and columns are for individual and ecosystem-level networks. Heatmaps show the correlation between values for each projection type, where the values are of the proportion of networks which meet the scale-free threshold criteria of: $p \geq 0.1$ (top row); $n_{tail} > 50$ (center row); $2 < \alpha < 3$ (bottom row). Values from projections of a network's LCC and entire graph are highly correlated. All types of unipartite compound networks tend to be correlated. Values across many other projection types are barely correlated for the p-value and n_{tail} criteria. Ecosystems tend to show more correlation, across all projection types, than individuals.

the time, whereas individuals could be correctly identified 85.33% of the time (Fig. 23, left columns). When normalizing other measures to network size, the only one that improved in distinguishing individuals and ecosystems to be better than chance was *dexp* (Fig. S24). This is a measure of which type of distribution is favored (or neither) when doing a log-likelihood ratio test between the power-law and exponential distribution.

3.4.3.0.2 Random Forest.

Random forest classifiers are a supervised machine learning technique that use decision trees to make classifications. When using random forests to try and distinguish individuals and ecosystems based on network and distribution data, we find ecosystems can be correctly predicted 87.01% of the time, and individuals can be correctly predicted 95.82% of the time (Out of bag, OOB, error rate is 7.91%). However, the size of the network and size of the degree distribution tail once again are the best relative predictors. Without network size and tail size, the prediction accuracy drops to 79.27% for ecosystems and 94.81% for individuals (OOB error rate of 11.80%). When doing random forest classification by projection type, the prediction accuracies are still above 75% for ecosystems and 91% for individuals across all projections, which is better than multinomial regression models even when information about network size is included (Fig. 23, left columns; Table S3). Mean degree was the best predictor across all network projection types.

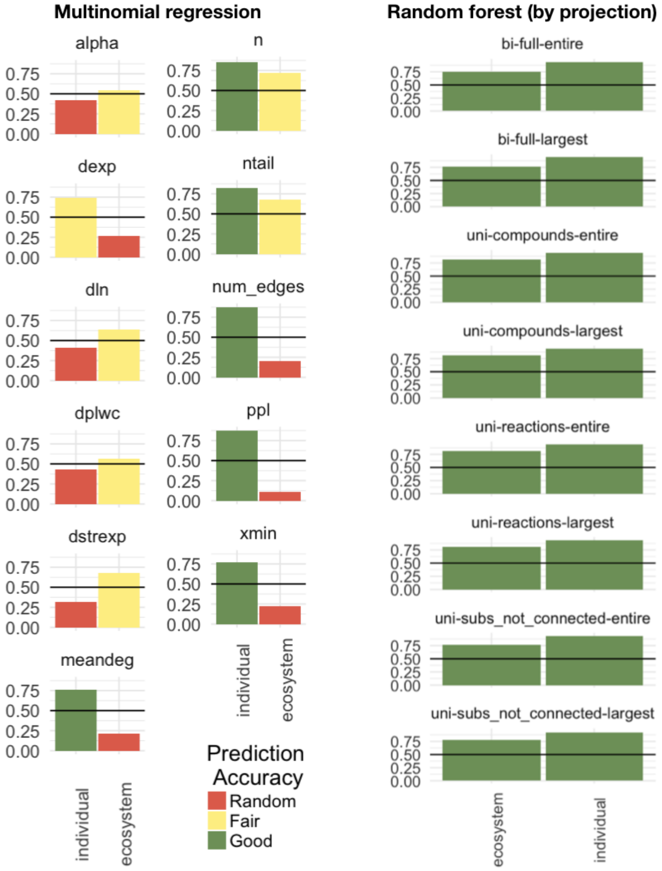


Figure 23. Predicting individuals and ecosystems from degree distribution data using multinomial regression vs. random forest. Each subplot shows the accuracy of using a particular network or statistical measure to predict whether that network data came from an biological individual or ecosystem. The left plots show prediction accuracy from using multinomial regression across all network projection types, and the right plots show prediction accuracy using random forest on each type of projection. The random forest classifier is much better at predicting individuals and ecosystems correctly from network data, even without direct access to network size. All random forest predictions have an accuracy of at least 75% across all projection types. Subplots measures are: power-law alpha value; log-likelihood result from power-law vs. exponential; log-likelihood result from power-law vs. log-normal; log-likelihood result from power-law vs. power-law with exponential cutoff; log-likelihood result from power-law vs. stretched exponential; the network mean degree; network node size; degree distribution tail size; network edge size; the p-value of the goodness-of-fit test for the power-law model; cutoff degree value for network tail. These are the only predictors used in the random forest classifier. Prediction accuracy is random if $\leq 50\%$, Fair if $> 50\%$, and Good if $\geq 75\%$.

3.5 Discussion

Our results indicate biochemical systems across individuals and ecosystems are, at best, only weakly scale-free. This is revealed by studying all possible projections of biochemical systems in tandem: only six of the eight network projection types analyzed favor power-law distributions over alternatives and in all cases the power-law is not itself a good fit to the data. Nonetheless, we can conclude individuals and ecosystems both share qualitatively similar degree distribution characteristics, and while this is a very coarse-grained measure of network structure, it suggests the possibility of shared principles operating across levels of organization to architect biochemical systems. The random forest distinguishability analyses demonstrate using a combination of all the results of scale-free analyses completed in this paper can predict, better than chance, whether the data comes from individuals or ecosystems. Individuals are perhaps more tightly constrained in coarse-grained network structure, based on being able to more accurately predict them based on simple network characteristics. Whether or not this structure is truly a universal property of life's chemical systems is more difficult to conclude. Based on the sample sizes, we are confident our results hold over the population of genomes and metagenomes in the JGI and PATRIC databases. However, the observed scaling is only reflective of biology universally if the databases are unbiased in sampling from all of biology on Earth, and this is impossible to know with certainty (see textit e.g. proposals of 'shadow life' and reports of missing biota [61, 116]). Nonetheless, the fact that multiple levels and multiple projections of biochemistry reveal common structure suggests universal principles may be within reach if cast within an ensemble theory of biochemical network organization.

Achieving such a theory requires recognizing that, unlike simple physical systems where statistics over individual components is sufficient to describe and predict their behavior, biological and technological systems require additional information about the structure of interactions among their many components. This is well-known, but how to project this structure onto simple mathematical objects that can be quantifiably characterized and compared remains a central problem of complex systems science. In physics, the relevant coarse-graining procedure is well understood, but we are not so far in complexity science: the first hurdle we must traverse is to identify the proper coarse-grained network representations for analysis. Existing literature cautions against using unipartite network projections, as it is argued they can lead to "wrong" interpretations of system properties such as degree in biochemical networks [47, 117]. We find instead that whether or not this conclusion should be drawn is highly dependent on the particular characteristics of degree or the degree distribution under consideration. For example, all network projection types, aside from unipartite reaction networks, favor power-law degree distributions over other heavy-tailed alternatives (Fig. 19, top row).

For power-law α ranges $2 < \alpha < 3$, bipartite networks show similar results to the unipartite reaction networks of individuals, but different results for ecosystems and unipartite compound networks (Fig. 19, fourth row). Almost all projections show differing results for meeting the scale-free p-value cutoff (Fig. 19, second row). While other literature [45, 46] has advocated for unipartite networks (with all compounds participating in a reaction connected–called *uni-compounds* here), we find that these networks overestimate power-law goodness-of-fit p-values and α values compared to reaction and bipartite networks (Fig. 19). The similarities and differences in the structure of different projections provides insight into the actual structure of the underlying system of interest. Given that there is no obvious answer for whether a system is scale-free, we advocate for studying all projections possible: regardless of whether or not a given projection is scale-free, all projections provide insights into the structure of the underlying system. In physics we have become accustomed to one unique coarse-grained descriptor providing insight into the structure of a system. It may be that to really understand complex interacting systems, such as the systems of reactions underlying all life on Earth, we must forget the allure of simple, singular models. Instead, to characterize the regularities associated with living processes, we should perform statistical analyses over many (still relatively simple) coarse-grained projections.

3.6 Materials and Methods

3.6.1 Obtaining biological data

Bacteria and Archaea data were obtained through PATRIC [118]. Starting with the 21,637 bacterial genomes available from the 2014 version of PATRIC, we created a parsed dataset by selecting one representative genome containing the largest number of annotated ECs from each genus. Unique genera (genera only represented by a single genome) were also included in our parsed data. Uncultured/candidate organisms without genera level nomenclature are left in the parsed dataset. This left us with 1152 parsed bacteria, from which we chose 361 randomly to use in this analysis. Starting with 845 archaeal genomes available from the 2014 version of PATRIC, we randomly chose 358 to use in this analysis. Enzyme Commission (EC) numbers associated with each genome were extracted from the `ec_number` column of each genome's `.pathway.tab` file.

Eukarya and Metagenome data were obtained through JGI IMG/m [119]. All 363 eukaryotic genomes available from JGI IMG/m as of Dec. 01, 2017 were used. Starting with the 5586 metagenomes available from JGI IMG/m as of June 20, 2017, 785 metagenomes were randomly chosen for this paper's analyses. Enzyme Commission (EC) numbers associated with each genome/metagenome were extracted from the list of *Protein coding genes with enzymes*, and metagenome EC numbers were obtained from the *total* category. All JGI IMG/m data used in this study were sequenced at JGI.

Because each EC number corresponds to a unique set of reactions that an enzyme catalyzes, the list of EC numbers associated with each genome and metagenome can be used to identify the reactions that are catalyzed by enzymes coded for in each genome/metagenome. We use the Kyoto Encyclopedia of Genes and Genomes (KEGG) ENZYME database to match EC numbers to reactions, and the KEGG REACTION database to identify the substrates and products of each reaction [58, 114, 115]. This provides us with a list of all chemical reactions that a genome/metagenome's enzymes can catalyze.

3.6.2 Generating Networks

Each genomic/metagenomic dataset is used to construct eight representations of biochemical reaction networks. We refer to each type of representation as a "network projection type" throughout the text:

1. *Bipartite graph with reaction and compound nodes.* A compound node C_i is connected to a reaction node R_i if it is involved in the reaction as a reactant or a product. Abbreviated in figures as *bi-full*.

2. *Unipartite graph with compound nodes only.* Two compound nodes C_i and C_j are connected if they are both present in the same reaction. A reaction's reactant compounds are connected to each other; a reaction's product compounds are connected to each other; and a reaction's reactant and product compounds are connected. Abbreviated in figures as *uni-compounds.*textit
3. *Unipartite graph with reaction nodes only.* Two reaction nodes R_i and R_j are connected if they involve a common compound. Abbreviated in figures as *uni-reactions.*
4. *Unipartite graph with compound nodes only (alternate).* Two compound nodes C_i and C_j are connected only if they are both present on opposite sides of the same reaction. A reaction's reactant compounds are *not* connected to each other; a reaction's product compounds are *not* connected to each other; but a reaction's reactant and product compounds are connected. Abbreviated in figures as *uni-subs_not_connected.*

There exists a version of each of these four network construction methods for the largest connected component (LCC), and for the entire graph, yielding a total of eight network projections for each dataset (Fig. 17). These network projection types are signified in the figured by appending *-largest* and *-entire* to the network projection abbreviations. Some datasets may yield identical networks for their LCC and entire graph, if there is exists only a single connected component.

3.6.3 Assessing the power-law fit on degree distributions

As defined in Clauset, 2009 [69], a quantity x obeys a power law if it is drawn from a probability distribution

$$f(x) = x^{-\alpha}, \tag{3.2}$$

where α, the exponent/scaling parameter of the distribution, is a constant. In order to estimate α, we follow the methods described in Clauset, 2009 [69], and use an approximation of the discrete maximum likelihood estimator (MLE)

$$\hat{\alpha} \simeq 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min} - \frac{1}{2}} \right]^{-1}, \tag{3.3}$$

where x_{min} is the lower bound of power-law behavior in our data, and x_i, i=1,2,...,n, are the observed values x such that $x_i \geq x_{min}$. The standard error of our calculated α is given by

$$\sigma = \frac{\hat{\alpha} - 1}{\sqrt{n}} + \mathrm{O}(1/n), \tag{3.4}$$

where the higher-order correction is positive [69]. Because many quantities only obey a power-law for values greater than some x_{min}, the optimal x_{min} value must be calculated. The importance of choosing the correct value for x_{min} is discussed in detail in Clauset et al, 2009 [69]. If it is chosen too low, data points which deviate from a power-law distribution are incorporated. If it is chosen too high, the sample size decreases. Both can change the accuracy of the MLE, but it is better to err too high than too low.

In order to determine x_{min}, we use the method first proposed by Clauset et al., 2007 [120], and elaborated on in Clauset et al., 2009 [69]: we choose the value of x_{min} that makes the probability distributions of the measured data and the best-fit power-law model as similar as possible above x_{min}. The similarity between the distributions is quantified using the Kolmogorov-Smirnov or KS statistic, given by

$$D = \max_{x \geq x_{min}} |S(x) - P(x)|, \tag{3.5}$$

where $S(x)$ is the cumulative density function (CDF) of the data for the observations with value at least x_{min}, and $P(x)$ is the CDF for the power-law model that best fits the data in the region $x \geq x_{min}$. Our estimate of x_{min} is the one that minimizes D.

We used the github respository made available in Broido and Clauset [108] to determine the optimal x_{min} of all our degree distributions, and to subsequently calculate the MLE in order to determine the *scaling exponent* α and the *standard error* on α, σ [121].

A power-law can always be fit to data, regardless of the true distribution from which it is drawn from, so we need to determine whether the power-law fit is a good match to the data. We do this by sampling many synthetic data sets from a true power-law distribution, recording their fluctuation from power-law form, and comparing this to similar measurements on the empirical data in question. If the empirical data has similar form to the synthetic data drawn from a true-power law distribution, then the power-law fit is plausible. We use the KS statistic to measure the distance between distributions.

We use a goodness-of-fit test to generate a p-value which indicates the plausibility of a hypothesis. The p-value is defined as the fraction of the synthetic distances that are larger than the empirical distance. If p is large (close to 1), then the difference between the empirical data and the model can be attributed to statistical fluctuations alone; if it is small, the model is not a plausible fit to the data [69]. We follow the methods in Clauset et al., 2009 [69]–and implement them with the github package used in Broido and Clauset [108]–to generate synthetic datasets and measure the distance between distributions. Following these methods, we chose to generate 1000 synthetic datasets in order to optimize the trade-off between having an accurate estimation of the p-value and computational efficiency. If p is small enough ($p < 0.1$) the power law is ruled out. Put another way, it is ruled out if there is a probability of 1 in 10 or less that we would by chance get data that agree as poorly with the model as the data we have [69]. However, measuring a $p \geq 0.1$ does not guarantee that the power-law is the most likely distribution for the data. Other distributions may match equally well or better. Additionally, it is harder to rule out distributions when working with small sample sizes.

A better way to determine whether or not data is drawn from a power-law distribution is to compare its likelihood of being drawn from a power-law distribution directly to a competing distribution [69, 122]. We use the exponential, stretched-exponential, log-normal, and power-law-with-cutoff distributions as four competing distributions to the power-law. While we cannot compare how the data fits between every possible distribution, comparing the power-law distribution to these four similarly shaped competing distributions helps us ensure that our results are valid.

We use the log-likelihood ratio test $\mathcal{R}$ [69, 122] to compare the power-law distribution to other candidate distributions,

$$\mathcal{R} = \mathcal{L}_{\text{PL}} - \mathcal{L}_{\text{Alt}}, \tag{3.6}$$

where $\mathcal{L}_{\text{PL}}$ and $\mathcal{L}_{\text{Alt}}$ are the log-likelihoods of the best fits for the power-law and alternative distributions, respectively. This can be rewritten as a summation over individual observations,

$$\mathcal{R} = \sum_{i}^{n_{\text{tail}}} [\ell_i^{(\text{PL})} - \ell_i^{(\text{Alt})}], \tag{3.7}$$

with the log-likelihood of single observed degree values under the power-law distribution, $\ell_i^{(\mathrm{PL})}$, and alternative distribution, $\ell_i^{(\mathrm{Alt})}$, are summed over the number of model observations, n_{tail}.

If $\mathcal{R} > 0$, the power-law distribution is more likely; if $\mathcal{R} < 0$, the competing candidate distribution is more likely; if $\mathcal{R} = 0$, they are equally likely. Just like with the goodness of fit test, we need to make sure our result is statistically significant ($p < 0.01$). The methodology described here summarizes the methodology introduced by Clauset et al., 2009, and described again in Broido and Clauset, 2018 [69, 108] and more details such as the exact formulas for alternative distributions, and derivation of the p-value for $\mathcal{R}$ can be obtained therein.

3.6.4 Classifying network scaling

We classify each genomic/metagenomic dataset, as represented by the set of eight network projection types, as having some categorical degree of "scale-freeness" from "super-weak" to "strongest". This classification scheme was introduced by Broido and Clauset, 2018 [108] in order to compare many networks with different degrees of complexity, and the definitions below were extracted from therein:

- Super-Weak: For at least 50% of graphs, none of the alternative distributions are favored over the power law.

The four remaining definitions are nested, and represent increasing levels of direct evidence that the degree structure of the network data set is scale free:

- Weakest: For at least 50% of graphs, the power-law hypothesis cannot be rejected ($p \geq 0.1$).

- Weak: The requirements of the Weakest set, and there are at least 50 nodes in the distribution's tail ($n_{\text{tail}} > 50$).
- Strong: The requirements of the Weak and Super-Weak sets, and that $2 < \hat{\alpha} < 3$.
- Strongest: The requirements of the Strong set for at least 90% of graphs, rather than 50%, and for at least 95% of graphs none of the alternative distributions are favored over the power-law.

Categorizing a network as "Super-Weak" is in effect saying that that network's degree distribution data is *better* modeled by a power-law fit than alternative distributions. This is independent of whether or not the power-law model is a *good* fit to the data, which is what is what the "Weakest" and "Weak" definitions emphasize. A network may be classified as "Super-Weak" without meeting any of the nested definition's criteria. Similarly, a network may be classified as "Weak" without meeting the criteria in the "Super-Weak" definition. We believe this framework is a proper way to classify the degree-distributions of biochemical networks, given that there are many different accepted ways to represent biochemical reactions as networks, and each has their pros and cons [45–47].

3.6.4.0.1 Standard error and correlation.

The black error bars on each plot represent 2 standard deviation (2SD) around the sample proportion $\hat{p}$ (the height of the bar, which we also refer to as the mean). This is equivalent to 2 standard error around the mean (2SEM), or a 95% confidence interval for the true population proportion p (true population mean). Standard deviation was calculated by treating each category as a binomial distribution, meaning the standard deviation is given by:

$$\sqrt{\frac{p(1-p)}{n}} \tag{3.8}$$

Although the errors for each plot's categories are calculated independently, there is co-variance between many of them. This is especially true for the right column of Fig. 19, where all bars of a color total to a fixed number of datasets, with each dataset falling into one of the 8 network projection type bins. Because of this, we also calculated the correlations between each network projection type, across both individuals and ecosystems (Fig. 22). The correlation matrices were calculated by using the pandas function `DataFrame.correlation(method='pearson')` on a matrix of binomially distributed True/False values representing whether each dataset passed or failed specific scale-free criteria for p-value, tail size, or power-law exponent value (α), for each network-projection.

3.6.5 Classifying levels of biology using degree distribution data

We used two different statistical methods, multinomial regression and random forest classifiers, in conjunction with the scale-free classification scheme above in order to test if individuals and ecosystems were distinguishable based on their degree distribution characteristics.

3.6.5.0.1 Multinomial regression.

For our multinomial regression, the response class is the biological level (individual or ecosytem), and a single network or statistical measure is the dependent variable. In order to control for over fitting the training data was composed of an equal number of samples from each level. The number of networks used for training data was chosen to be equal in size to 80% of all ecosystem projections, because there were less ecosystem datasets used than individual datasets. This corresponded to 80% of 6280 networks (of all projection types), or 5024 networks. The model was tested on the 20% of the data that it was not trained on. This process was repeated 100 times and the average model error is reported in the results and Fig. 23, left columns. The `multinom` and `predict` functions from the R-package `nnet` were used to do the multinomial regression.

3.6.5.0.2 Random forest classifiers.

We used a random forest to attempt to classify networks as falling into the category of individuals or ecosystems. In the first scenario, we used 11 predictors: power-law alpha value (α); log-likelihood result from power-law vs. exponential ($dexp$); log-likelihood result from power-law vs. log-normal (dln); log-likelihood result from power-law vs. power-law with exponential cutoff ($dplwc$); log-likelihood result from power-law vs. stretched exponential ($dstrexp$); the network mean degree ($< k >$); network node size (n); degree distribution tail size (n_{tail}); network edge size (n_{edges}); the p-value of the goodness-of-fit test for the power-law model (p); and cutoff degree value for network tail (x_{min}). In the second scenario, we repeated the random forest without the three predictors which can be directly used to quantify the size of a network (n, n_{tail}, and n_{edges}). In the third scenario, we repeated the random forest without the three predictors on each network projection type independently. For each scenario, we randomly split our data in two halves: one for training, and one for testing (for the third scenario, each training and testing set is 1/8 as large as for the first two scenarios, since we run the classifier on each network projection type independently). In all scenarios, we use the `randomForest` function from the R-package `randomForest` for classification. Three features were used to construct each tree (`mtry=3`), which is $\approx \sqrt{n_{features}}$, with 100 trees generated each time (enough time for the out-of-bag, or OOB, estimate of the error rate to level off).

3.7 Supporting Information

3.7.1 Supporting Figure

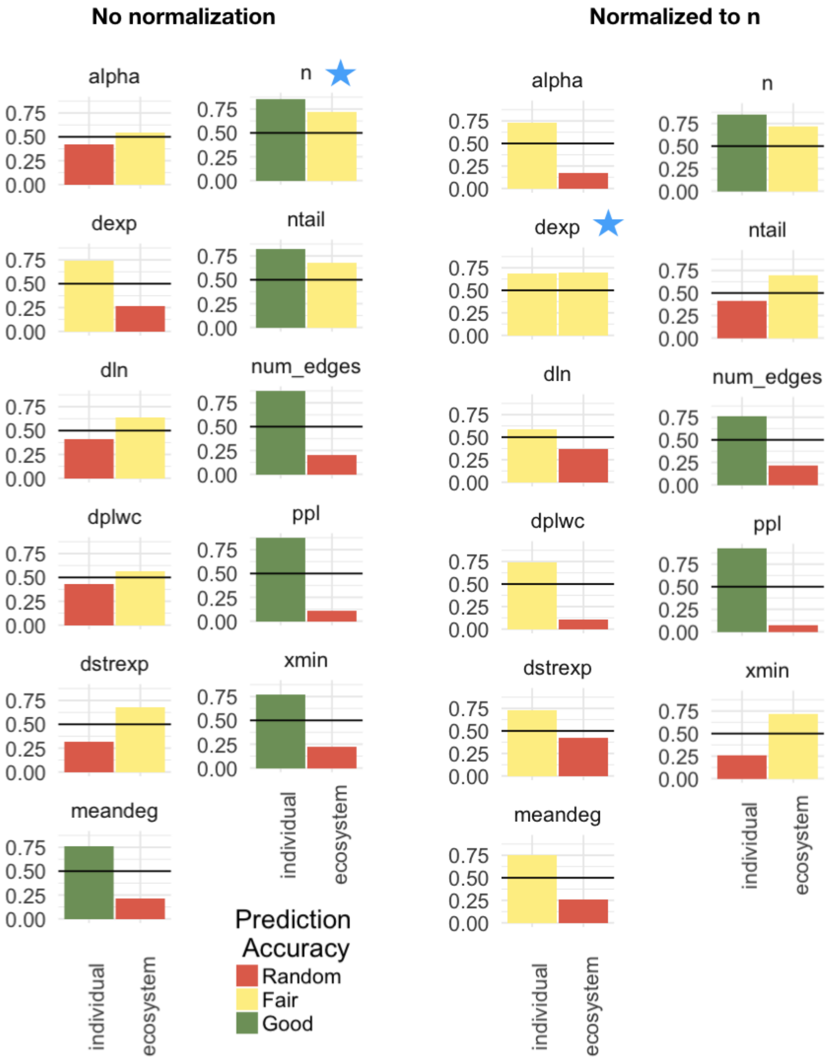


Figure 24. Predicting individuals and ecosystems from degree distribution data using multinomial regression. Each subplot shows the accuracy of using a particular network or statistical measure to predict whether that network data came from an biological individual or ecosystem. The subplots in the right column are the accuracy of using a measure after being normalized to network size. Unsurprisingly, network size is by far the best way to accurately predict whether data comes from an individual or ecosystem (left blue star). Once normalized to size, whether or not a degree distribution favors an exponential fit compared to a power-law fit becomes a decent predictor (right blue star). Subplots measures are: power-law alpha value; log-likelihood result from power-law vs. exponential; log-likelihood result from power-law vs. log-normal; log-likelihood result from power-law vs. power-law with exponential cutoff; log-likelihood result from power-law vs. stretched exponential; the network mean degree; network node size; degree distribution tail size; network edge size; the p-value of the goodness-of-fit test for the power-law model; cutoff degree value for network tail. Prediction accuracy is random if $\leq 50\%$, Fair if $> 50\%$, and Good if $> 75\%$.

3.7.2 Supporting Table

Table 3. Random forest accuracy by network projection type.

Network projection type	Prediction accuracy (%)		OOB error (%)
	Ecosystem	**Individual**	
bi-full-entire	75.58	93.75	13.83
bi-full-largest	75.45	94.83	13.29
uni-compounds-entire	81.44	93.76	11.36
uni-compounds-largest	80.81	93.67	11.79
uni-reactions-entire	80.95	93.67	11.36
uni-reactions-largest	80.40	93.08	12.33
uni-subs_not_connected-entire	76.12	92.93	13.93
uni-subs_not_connected-largest	77.47	91.64	14.36

The predictors used in the random forest are the same predictors used in the multinomial regression: power-law alpha value; log-likelihood result from power-law vs. exponential; log-likelihood result from power-law vs. log-normal; log-likelihood result from power-law vs. power-law with exponential cutoff; log-likelihood result from power-law vs. stretched exponential; the network mean degree; network node size; degree distribution tail size; network edge size; the *p*-value of the goodness-of-fit test for the power-law model; cutoff degree value for network tail. See methods for description of network projection types.

Chapter 4

ASSESSING THE VIABILITY OF BIOCHEMICAL NETWORKS ACROSS PLANETS

4.1 Abstract

The concept of the origin of life implies that initially, life emerged from a non-living medium. If this medium was Earth's geochemistry, then that would make life, by definition, a geochemical process. The extent to which life on Earth today could subsist outside of the geochemistry from which it is embedded is poorly quantified. By leveraging large biochemical datasets in conjunction with planetary observations and computational tools, this research provides a methodological foundation for the quantitative assessment of our biology's viability in the context of other geospheres. Investigating a case study of alkaline prokaryotes in the context of Enceladus, we find that the chemical compounds observed on Enceladus thus far would be insufficient to allow even these extremophiles to produce the compounds necessary to sustain a viable metabolism. The environmental precursors required by these organisms provides a map for the compounds which should be prioritized for detection in future planetary exploration missions. The results of this framework have further consequences in the context of planetary protection, and hint that forward contamination may prove infeasible without meticulous intent.

4.2 Introduction

It is probable that the geochemical process known as life had already commenced when today's oldest minerals began to crystallize. While there is widely accepted evidence that the process of life has been present on Earth continuously for the past 3.4Gy [123], the lack of evidence prior to this date has more to do with the paucity of fossil-preserving rocks than concrete evidence of life's absence [124, 125]. Despite the biosphere's apparent interminable coexistence with the geosphere, there remain many open questions on the matter of life persisting in Earth's absence [21, 101], not to mention the questions of Earth persisting in life's absence [126–128]. For example, Visionaries dream of terraforming planets while program officers fret over "contaminating" them [76–78]. While the terraformers tend to believe that seeding another planet would require careful human or robotic (and usually Earth-assisted) cultivation, planetary protection officers take the more conservative stance that a small, semi-sterilized spacecraft of Earth origin could cause life to spill onto a planet in the same way that a small perturbation to a super cooled liquid would cause the entire volume to quickly crystallize. In both cases, there is the predominately implicit assumption that Earth-life would be viable outside of the Earth.

When life is viewed as a geologic process, this is a somewhat surprising assumption. In the words of Morowitz et al., "the metabolic character of life is a planetary phenomenon, no less than the atmosphere, hydrosphere, or geosphere" [129]. If this "metabolic character of life" is truly a planetary phenomenon, does that imply that life is inextricable from the planet through which it emerged? Or is it possible that an infinitesimal component of our biosphere—a sliver of a sliver of Earth's biochemical diversity captured in a few species—could be enough to imbue another world with Earth's vitality?

To begin to address these questions, we must first lay the framework for determining the environmental conditions required for a species to produce or acquire the chemical compounds necessary to yield a viable metabolism. For this, we utilize the network expansion method [79]: an organism can catalyze a reaction only if it has access to the necessary substrates. The initial substrates, called the *seed set*, are the compounds available to the organism from the environment. Initially, these are the only compounds in the organism's *network*—an abstract representation of the biochemistry able to be utilized by the organism with the given compounds. The organism catalyzes all the reactions it can based on the compounds available in its network, and then adds the new compounds it can generate to its network. This process proceeds iteratively until the organism can produce no new compounds. The state of the organism's network when expansion ceases is referred to as the organism's *scope*—and it contains all of the compounds which can be synthesized by an organism, plus the compounds provided by the environment (the seed set).

While there are other methods which can be used to computationally assess organismal viability, relying on some combination of integer linear programming, kinetic modeling using differential equations, elementary mode analysis, and flux balance analysis (FBA), they require catalytic rates which are difficult to acquire and sparsely catalogued, or a curated list of stoichiometrically balanced reactions [130]. FBA is perhaps the most common method for assessing organismal viability, and operates by solving for the relative fluxes of reactions needed in order for steady state production of compounds identified necessary for organismal growth. Despite FBA requiring more constrained information and computational resources, network expansion has been shown to give near identical results for identifying compounds produced (the network scope) [130, 131].

Network expansion models have been used to explore the scope of chemicals accessible to biology across space and time on Earth, and how changing environments and changing biochemical networks impact one another [80]. For example, the models have been utilized to identify how oxygen drastically altered life's biochemical networks during the great oxygenation event [54]; how biochemistry differed before phosphorous was widely available [53]; how organismal scopes vary across the tree of life [80, 81]; and how organismal metabolic variability is impacted both in the presence of diverse environments and the presence of other species [82].

We propose using network expansions to address the question of life's viability amongst other planetary chemistries in two fundamental ways: For a set of organisms and a set of planetary environments, how many target substrates can each organism produce across the environments? The inverse question—For a set of organisms and a set of planetary environments, what chemical seed sets must be provided in order to produce the substrates which are necessary to the organism's viability?

We work through a case study of this framework to determine the viability of varying Earth organisms within Enceladus's planetary context. Because Enceladus has an ocean with high pH (11-12) [132], we choose to focus on the viability of prokaryotic alkaliphiles. Because other environmental factors are less well constrained, and parameters like temperature and salinity could vary substantially across locations, we do not place any further restrictions on the organismal metabolisms that we run network expansions on [133]. We show that based on the compounds we currently know to be present in Enceladus's subsurface ocean [134], none of the analyzed organismal metabolisms are viable. In order to verify that this is not solely due to the lack of phosphate, a prominent bioessential compound on Earth which has not been detected on Enceladus (likely due to Cassini instrument detection thresholds), we show that adding phosphate as a seed compound still results in no viable organisms. Using an algorithm developed to solve the inverse network expansion problem [135], we identify minimal sets of substrates that satisfy the requirements of what these alkaliphilic organisms would have to acquire externally in order to produce the target substrates. We find that these organisms tend to require complex molecules and coenzymes, lowering the likelihood that the organisms could be viable on Enceladus, given their lack of detection. Nonetheless, when the full catalytic repertoire of Earth's biosphere is available, we find that nearly all target substrates are able to be synthesized from a seed set consisting only of the compounds currently observed on Enceladus (plus phosphate). Although these reactions are not the product of organisms which are solely alkaliphilic, these results hint that forward contamination from individuals may be much less concerning than contamination by a microbial ecosystem which can emulate the robustness and catalytic capabilities of the biosphere—reinforcing the perspective that the emergence of life on a planet is an extension of the planet's geosphere [21, 83]. More importantly, by leveraging large biochemical datasets in conjunction with planetary observations and computational tools, this research provides a methodological foundation for the quantitative assessment of our biology's viability in the context of other geospheres.

4.3 Results

Based on target metabolites necessary for many living organisms, we first sought to determine if the compounds which have thus far been identified on Enceladus were sufficient to produce the target metabolites in a set of organisms which would be viable in an environment with the alkalinity present on Enceladus [132].

We ran the network expansion algorithm on the subset of archaea and bacteria with documented environmental pH in the ranges of 9-11 [58, 114, 115], using a seed set of compounds which have been identified on Enceladus from observations aboard Cassini's Ion and Neutral Mass Spectrometer (INMS) [134] (Table 4).

Name	Formula	KEGG Compound ID
Water	(H2O)	C00001
Carbon Dioxide	(CO2)	C00011
Carbon Monoxide	(CO)	C00237
Hydrogen	(H2)	C00282
Formaldehyde	(H2CO)	C00067
Methanol	(CH3OH)	C00132
Ethylene oxide	(C2H4O)	C06548
Ethanol	(C2H6O)	C00469
Hydrogen sulfide	(H2S)	C00283
Ammonia	(NH3)	C00014
Nitrogen	(N2)	C00697
Hydrogen Cyanide	(HCN)	C01326
Methane	(CH4)	C01438
Acetylene	(C2H2)	C01548
Ethylene	(C2H4)	C06547
Propene	(C3H6)	C11505
Propane	(C3H8)	C20783
Benzene	(C6H6)	C01407
Phosphate	(H3PO4)	C00009

Table 4. Compounds used for Enceladus seed set. All compounds from Waite et al., 2009 [134] that were present in the Kyoto Encyclopedia of Genes and Genomes (KEGG) were included. Phosphate was added to the seed set for additional analyses.

We deem an organism or network to be fully viable if, given a set of environmental seed compounds, it has the catalytic repertoire to produce all the compounds in its network which intersect with a pre-defined set of target metabolites. For this study, we adopt the list of target metabolites defined by Freilich et al (2009) [82], (Table 5).

In that study, the authors found that the organisms which were found to be viable, based on these target metabolites, accurately predicted the ecological compositions of known environments across many habitats and bacterial metabolisms.

Name	KEGG Compound ID	Name	KEGG Compound ID
ATP	C00002	NAD+	C00003
NADH	C00004	NADPH	C00005
NADP+	C00006	ADP	C00008
UDP	C00015	FAD	C00016
AMP	C00020	Acetyl-CoA	C00024
L-Glutamate	C00025	GDP	C00035
Glycine	C00037	L-Alanine	C00041
UDP-N-acetyl-D-glucosamine	C00043	GTP	C00044
L-Lysine	C00047	L-Aspartate	C00049
Adenosine 3',5'-bisphosphate	C00054	CMP	C00055
L-Arginine	C00062	CTP	C00063
L-Glutamine	C00064	L-Serine	C00065
L-Methionine	C00073	UTP	C00075
L-Tryptophan	C00078	L-Phenylalanine	C00079
L-Tyrosine	C00082	L-Cysteine	C00097
UMP	C00105	CDP	C00112
Glycerol	C00116	L-Leucine	C00123
dATP	C00131	L-Histidine	C00135
GMP	C00144	L-Proline	C00148
L-Asparagine	C00152	L-Valine	C00183
L-Threonine	C00188	10-Formyltetrahydrofolate	C00234
dCMP	C00239	Hexadecanoic acid	C00249
Riboflavin	C00255	dGTP	C00286
Phosphatidylethanolamine	C00350	dAMP	C00360
dGMP	C00362	dTMP	C00364
Ubiquinone	C00399	L-Isoleucine	C00407
dCTP	C00458	dTTP	C00459
1,2-Diacyl-sn-glycerol	C00641	Siroheme	C00748
UDP-N-acetylmuramate	C01050	Hexadecanoyl-[acp]	C05764
Cardiolipin	C05980	Diglucosyl-diacylglycerol	C06040
Heme O	C15672	(2E)-Octadecenoyl-[acp	C16221
Undecaprenyl-diphospho-...	C05890		
N-acetylmuramoyl-...	C05894		
(N-acetylglucosamine)-L	C05899		

Table 5. Compounds in the target metabolite set. Target list adopted from Freilich et al (2009) [82]

4.3.1 Prokaryotic viability on Enceladus

We find that none of these organisms, across bacteria and archaea, can produce any target metabolites with the few identified organic and inorganic compounds on Enceladus. In fact, they are found to produce only a fraction of the compounds possible given their reaction network (Fig. 25). However, this was not surprising given the lack of detection of any phosphorous containing compounds. Because of this, we repeated the expansion with the addition of phosphate. While this increased the scope of the organismal seed sets, again, no target compounds were able to be produced. Although in the latter case, we note that the organismal scopes increased in size (Fig. 25A).

4.3.2 Identifying the compounds necessary to make prokaryotes viable

Running network expansions on pre-established seed sets are useful for determining the set of compounds which can be part of an organism's scope. However, as we found in the section above, if we are aiming to produce a specific set of target compounds, there is no guarantee that a chosen seed set will do that. For this reason, it is useful to identify an algorithm which can identify the seed set needed to produce a target set, given a reaction network. We thus sought to identify subsets of all compounds involved in each organism's network which could feasibly produce all the target compounds in that network.

There are three obvious ways to go about this. We could imagine searching for: 1) a single minimal seed set (no subsets of which can produce all target metabolites), 2) the smallest minimal seed set (where there are no sets with fewer elements which can

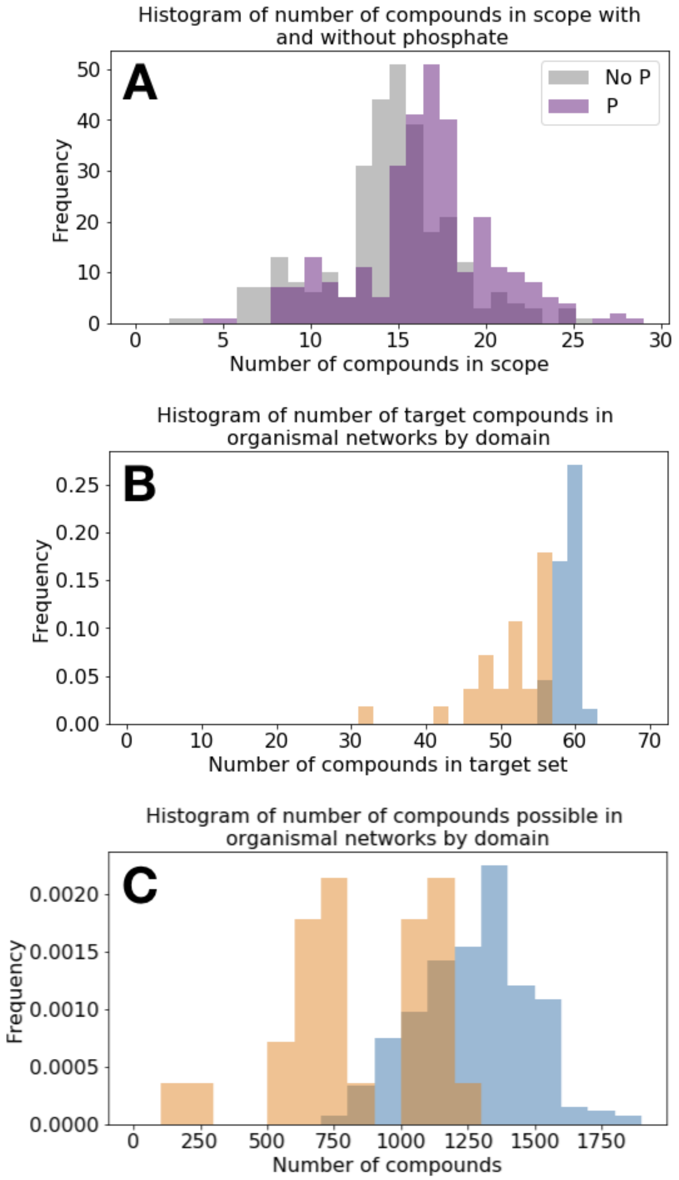


Figure 25. Histograms from the network expansions for prokaryotes using the Enceladus seed set. (A) How the scope size changes for all organisms when adding phosphate to the seed set adopted from Waite et al. [134]. In neither case do any target compounds get produced for any organisms. (B) An overview of the distribution of number of target compounds across all organisms (out of 65 possible based on the target set from Freilich et al. [82]. (C) The maximum theoretical sizes of networks, if scopes were able to take advantage of full organismal reaction networks. Orange bars are for archaea, and blue bars are for bacteria.

produce all target metabolites), or 3) all minimal seed sets (the set of all sets that can produce all target metabolites).

We chose to identify a subset of all minimal seed sets for the archaea and bacteria under consideration, because finding the smallest minimal seed set is an NP-hard problem (Cottret et al., 2008), and because it would result in only a single environment in which a target set could be produced. Finding any given minimal seed set requires a polynomial-time algorithm, so for computational tractability we chose to identify 100 random minimal seed sets for each of the 28 aforementioned archaea, and for 36 of the aforementioned 266 bacteria. We follow the algorithm described in Handorf et al., 2008 to create random minimal seed sets which attempt to minimize the likelihood of obtaining seed sets with large complex biomolecules where possible (see methods).

We first take an overview of the minimal seed sets we find which produce target compounds for each of the analyzed organisms. We find that the environmental seed sets needed are often smaller in size, but more complex (as quantified by the mean molecular weight of the seed sets needed) (Fig. 26). This is especially true for the bacteria, while for archaea the seed sets tend to be composed both of more complex molecules and more of them. Interestingly, there no seeds identified which require more than four of the compounds which have been identified as part of the Enceladus seed set.

Next we look at how similar each of the 100 minimal seeds sets for each organism are to one another. We find that across all organisms, the archaea seed sets tend to have more self-similarity compared to the bacteria. Two archaea share about a quarter of the compounds across all their seed sets, on average (Fig. 27).

We then turn to examine how seed sets necessary to produce viable organisms differ

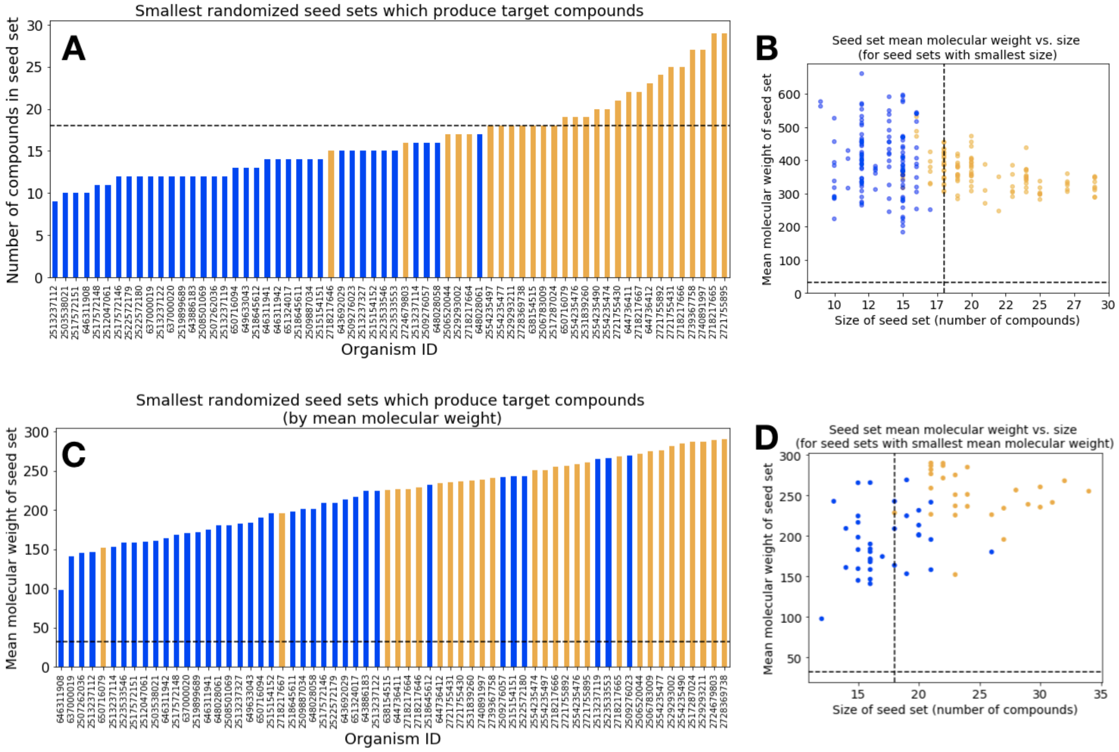


Figure 26. Characteristics of minimal seed sets which produce target metabolites. (A) A rank ordered plot of the smallest minimal seed sets, by number of compounds involved in each seed set. (B) The mean molecular weights of the smallest seed sets, by size, of the seed sets with the smallest size. Note that many organisms have multiple minimal seed sets of the same size, but of different mean molecular weights. (C) A rank ordered plot of the smallest minimal seed sets, by weight. (D) The mean molecular weights of the smallest seed sets, by size, of the seed sets with smallest mean molecular weight. Orange bars are archaea, and blue lines are bacteria, with each organism represented on the x-axis. The black dashed lines in each case shows the size and weight values for the Enceladus seed set.

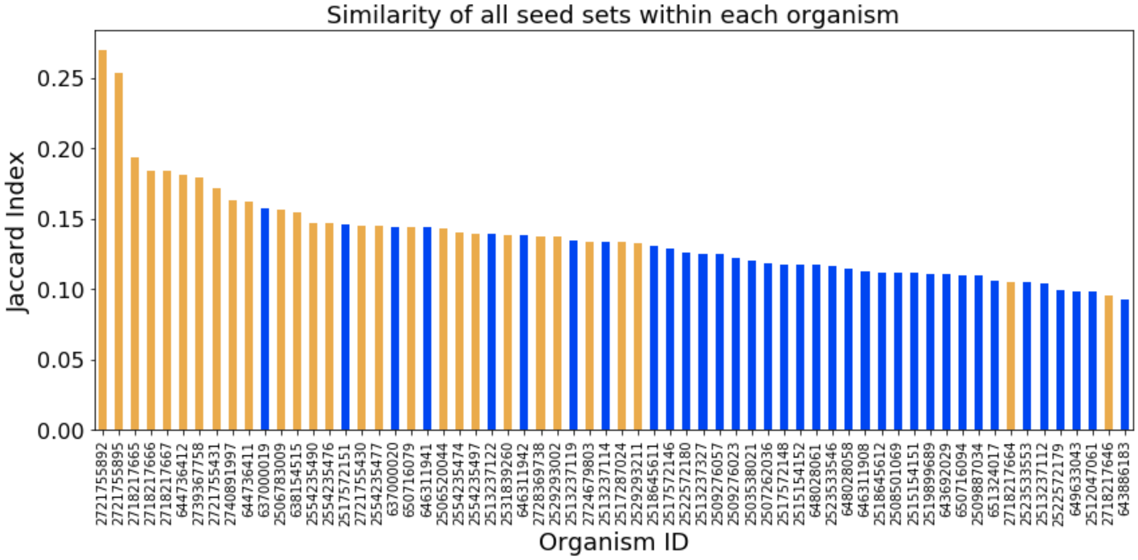


Figure 27. Similarity of all seed sets within each organism. The rank ordered mean jaccard index is shown for all 100 minimal seed sets we calculated for each organism. Bacteria are shown in blue and archaea are shown in orange.

between organisms. We find that archaea seed sets tend to be more similar to one another than bacteria seed sets. Nonetheless, comparing organisms within domains leads to similar seed sets much more often than comparing organisms between domains (Fig. 28). This result holds true even when, instead of comparing the union of seed sets of organism 1 to the union of seed sets of organism 2, we compare the minimum seed set of organism 1 to organism 2. In this case we are looking at the minimal seed set of each organism that has the smallest mean molecular weight (Fig. 28B). However, we find that clustering the jaccard similarity between the union of organism seed sets results in more accurate clustering of the two domains we investigate (orange and blue squares above and to the left of the cluster maps show whether the row is an archaea or bacteria, respectively). The hierarchical clustering produced from unions shows that is is possible to correctly group archaea and bacteria from only their minimal seed sets necessary for viability. This is an interesting result, complementary to that of Ebenhoh et al (2006), who showed that organisms which are more closely related appear to have more similar reaction *scopes*, as measured by the Jaccard distance [136]. Such distinguishability in seed sets might be useful in identifying a relationship with taxonomy, for the purpose of expeditiously discerning the organisms which could be most likely to be risks for planetary contamination, or beneficial for terraformation.

We turn to looking at the 100 most common seed compounds, to get some idea of the types of molecules we would expect to need to detect on Enceladus for this alkaliphiles to be viable. As might be expected, the majority of these compounds fall into common biochemical categories such as coenzymes, cofactors, amino acids, compound used for fatty acid synthesis, and other key metabolic pathways. It is notable that some of these compounds are target compound themselves, implying that these compounds are less likely to be synthesized by simpler compounds within these organismal metabolisms, and instead must be provided by the environment where possible.

Finally, we return to the initial set of seed compounds identified on Enceladus to examine if, with the full catalytic repoiroire of Earth's biosphere utilizing the geochemistry of Enceladus, it is possible to produce the compounds essential for prokaryotic organismal viability. Using only the compounds identified on Enceladus, plus phosphate, leads to the ability to produce nearly all target metabolites, and those needed for most prokaryotic life. The expansion is missing siroheme, a cofactor used for sulfur reduction in metabolic pathways, as well as heme, a complex used for a variety of biological functions including electron transfer and redox reactions.

This would seem to indicate that if it was possible to transplant the entire catalytic repertoire of the Earth to Enceladus, it would be possible to maintain

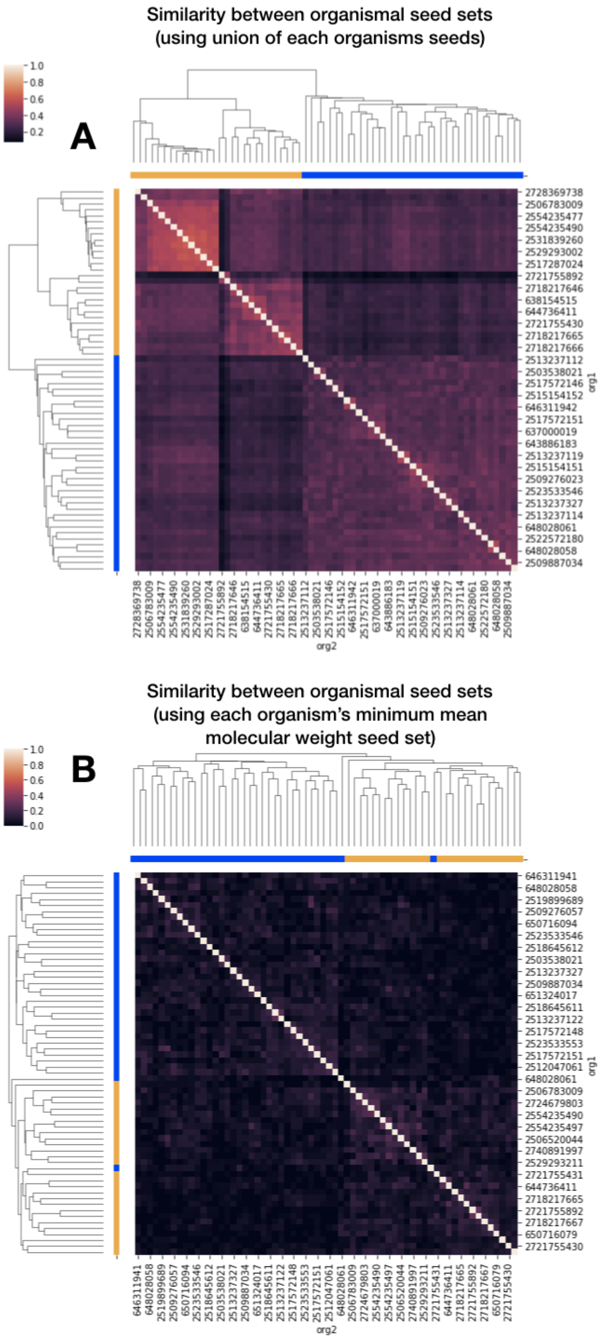


Figure 28. The similarity of seed sets between organisms. The clusters of two methods of organism comparisons are shown. (A) We take the union of all 100 seed sets within each organism, and compare them to one another using the jaccard index. (B) We take the minimal seed set of the smallest mean molecular weight of all 100 seed sets within each organism, and compare them to one another using the jaccard index. In both cases, the clustering separates out the domains (domain of each organism shown as blue squares for bacteria and orange squares for archaea above and below the cluster map.

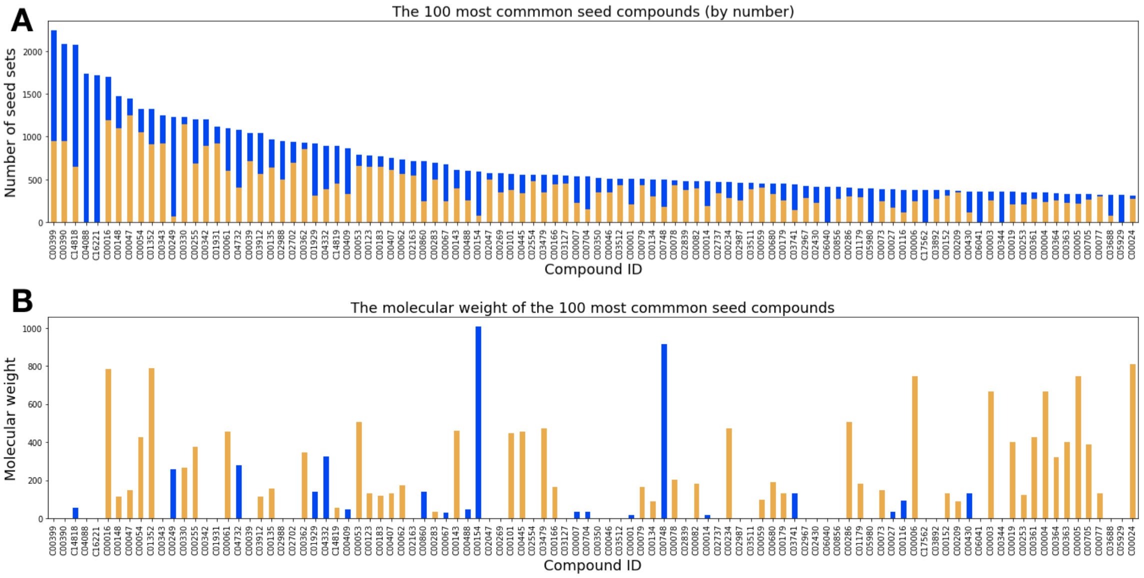


Figure 29. The top 100 most common seed compounds. (A) Rank ordered. The proportion of each found in archaea (orange) vs. bacteria (blue) seed sets are shown. (B) The molecular weights of each of the top 100 most common seed compounds. The domain of organism which most often contains seed sets with the compounds are shown as the color of the bar (archaea is orange and bacteria is blue).

minimal metabolic viability for most prokaryotic organisms, provided that most of the reactions could be catalyzed in the high pressure alkaline environment. However, this is dependent on the exact structure of the individual organismal networks present. One strategy for terraforming might be to try and produce the minimal ecosystem which can reproduce the catalytic potential of the biosphere to send to another planet. Conversely, one potential strategy for making sure that a spacecraft is adequately sterilized might be to take a biological sample from a clean room spacecraft and annotate its metagenome. Then a network expansion could be run on the metagenomic network, with a conservative seed set, to ensure that none of the biochemistry would be viable at the spacecrafts destination.

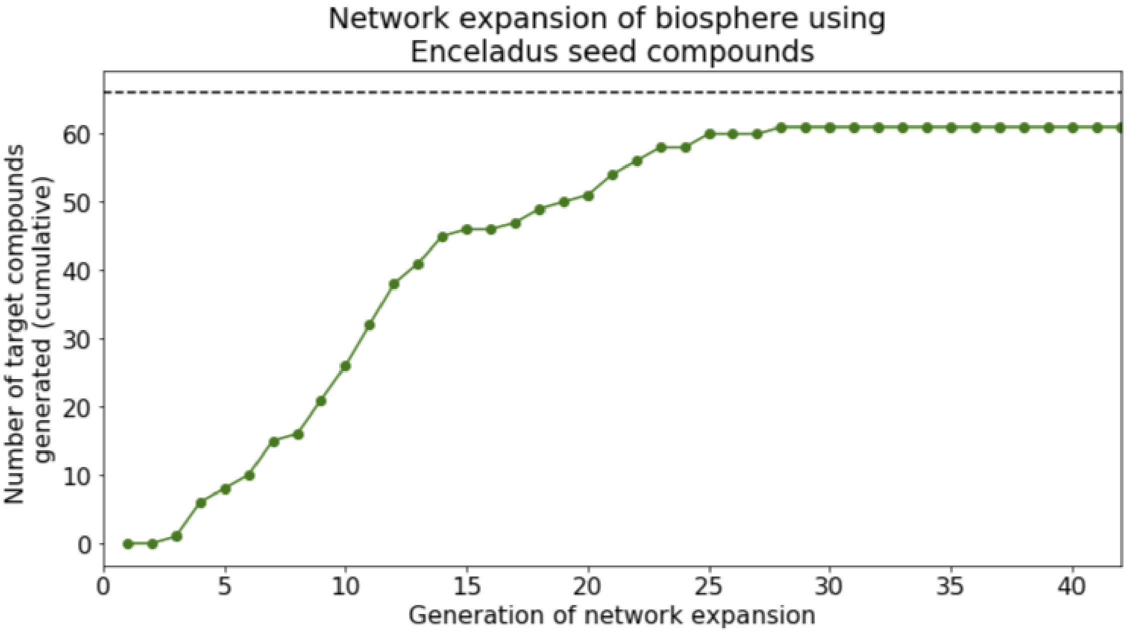


Figure 30. The network expansion of Earth's biosphere using compounds available on Enceladus. Nearly all possible target metabolites are produced in this circumstance, with siroheme and heme being the notable missing compounds.

4.4 Discussion

In this research, we laid out a framework to quantify the chemical compounds necessary to assess the viability of Earth's biochemistry in the context of other geospheres. We examine this framework as applied to Enceladus, executing the network expansion algorithm across metabolic networks of alkaliphilic bacteria and archaea in the chemical environment of Enceladus' subsurface ocean. We find that no organisms analyzed can produce any of the pre-established target metabolites in this environment. However, a key element of life on Earth, phosphorous, has not yet been detected on Enceladus. We determine that incorporating phosphorous, by adding phosphate as a seed compound, does not change our results—there are still no target metabolites produced from any of the prokaryotes analyzed.

Next we investigated what it would take for these organisms to be viable, finding that the chemical complexity of the seed sets, or number of seeds present, has to be much higher. In many scenarios, both number of compounds and mean molecular weight of the compounds present must increase. By analyzing the jaccard index within minimal seed sets of organisms, we find that there are many unique seed sets which produce equally viable organisms across archaea and bacteria. We also find that between different taxa, seeds are more similar between two bacteria and between two archaea than when comparing organisms of different domains. The similarity of seed sets needed for organismal viability clusters organisms into their domains, indicating that there may be further ways to identify environments suitable to specific taxonomies across planets.

Finally we showed that when the catalytic capability of the entire biosphere is expanded around the Enceladus seed set (including phosphate), the target compounds necessary for viability are produced. This could indicate that, in principle, if the bulk biochemical diversity of Earth life could be transplanted to another planet via simple prokaryotic organisms, these organisms might be able to sustain a viable metabolism. Thus, embedding themselves into a planet from which they did not emerge with consequences for both life and the planet.

It is worth noting that the above study provides only a basic proof of concept for the idea of utilizing the well-developed technique of network expansion to quantitatively addressing the most pressing questions of astrobiology. There are many ways that this work could be expanded in order to better reflect geochemical reality as well as incorporate more theoretical considerations. For example, we could permute the initial conditions of the network expansions to force the inclusion of the observed Enceladus compounds into the randomized seed sets. Or we could more strictly constrain the shuffling between high/low molecular mass compounds in these seed sets.

There are further details which could help direct our search for compounds on other planets if we wish to improve this framework's accuracy. For instance, we know that the presence/absence of cofactors is a big influence on the scope size for a seed set [79], so prioritizing our search for these compounds would provide high scientific returns. We could also measure viability as a gradient [82], and compare viability of organisms in other planetary contexts to the average viability of organisms across environments on Earth. We could investigate the specific metabolic pathways which are enriched or depleted in these environments [53]. Laboratory work here on Earth could also focus on better identifying reaction reversibility within organismal metabolic networks, as irreversible reaction networks would allow for more efficient algorithms used to identify minimal seed sets [80, 137, 138].

We might additionally included more statistical or theoretical constraints. For instance, can we identify distributions of molecular weights of compounds which tend to support biochemistry? Or link the expansions with knowledge of biochemical network topology, in order to find structural gaps in organismal networks which need to be filled to produce viable organisms [84]?

Moreover, the subset of organisms analyzed could be expanded to include organisms with greater metabolic diversity, or contracted to attempt to provide a better match between what we know about organismal environments on Earth with what we know about Enceladus. We could analyze the metabolisms of ecosystems, through metagenomic data, in addition to simple genomes. If we were specifically focused on the question of planetary contamination, we might also rerun these analyses on organisms which are known to exist in spacecraft sterilized clean rooms, like *Bacillus* *pumilus* SAFR-032 [139]. Further network expansion analyses could even be used to guide development of the composition of spacecraft materials to avoid metals which, if in contact with certain environments, could provide rich sources of cofactors or other compounds.

To summarize, the results from our network expansion analyses of alkaliphiles on Enceladus shows that there appears to be little risk of viability of these organisms, based on what we know about the chemical composition of the oceans. However, forward contamination, jeopardizing planetary protection, could be a much bigger risk if larger proportions of life's catalytic potential are transported to other planets unintentionally. This seems remarkably less likely, although spacecraft clean room microbial ecosystems are not well characterized. Intentionally seeding a planet with life seems likely only in the circumstance where a metabolism is specifically tailored to the environment, and even then there are questions about how well it could be self-sustaining. We believe that because life on Earth was a product of Earth's geochemistry, there is a significant bias to be viable only in a geochemical environment similar to the Earth's. While there is much more work to be done to quantify the risks, or possibilities, of Earth life being viable amongst other geospheres, we believe that we have laid significant groundwork for exciting research in this domain.

4.5 Materials and Methods

4.5.1 Defining the networks

In order to run the network expansion algorithm from a seed set, we first had to define our networks. To identify the reactions and compounds present in the metabolic networks of individual organisms, we collected data from the Joint Genome Institute's Integrated Microbial Genomes and Microbiomes database (JGI IMG/m) [119]. We located all archaea and bacteria which contained metadata on environmental pH, and filtered to those organisms with pH in the range of 9-11, approximately what might be expected in Enceladus's ocean [132]. For our case study, we extracted data from all 28 archaea and 266 bacteria matching this criteria. We downloaded the Enzyme Commission (EC) numbers associated with each genome from the organism's list of 'Protein coding genes with enzymes'. Each organisms list of EC numbers was mapped to the reactions which they catalyze using the Kyoto Encyclopedia of Genes and Genomes [58, 114, 115]. Using a combination of `Biopython` [140], the `KEGG REST API`, and `TogoWS` [141] to collect all KEGG `ENZYME`, `REACTION`, and `COMPOUND` data, we created reaction-compound networks for each organism. Each organisms network contains all of the reactions which all of its catalogued enzymes can catalyze, and all of the compounds involved in those reactions.

4.5.2 Executing the network expansion

As outlined in the introduction, the network expansion process works as follows: An organism, defined by a fixed set of reactions which it has the ability to catalyze, can catalyze a reaction only if it has access to the necessary substrates. The initial substrates, called the seed set, are the compounds available to the organism from the environment. Initially, these are the only compounds in the organism's network. The organism catalyzes all the reactions it can based on the reactions and compounds available in its network, and then adds the new compounds it can generate to its network. This process proceeds iteratively until the organism can produce no new compounds. The state of the organism's network when expansion ceases is referred to as the organism's scope—and it contains all of the compounds which can be synthesized by an organism, plus the seed set provided by the environment.

We assume that all reactions are reversible, both because the KEGG database recommends to not trust its reaction reversibility field, and because reaction directionality in nature depends on the concentrations of products and reactants, which we do not track here.

We ran the network expansion algorithm on the aforementioned subset of archaea and bacteria with documented environmental pH in the ranges of 9-11, using a seed set of compounds which have been identified on Enceladus from observations aboard CASSINI's Ion and Neutral Mass Spectrometer (INMS) [134]. We additionally ran this seed set when including phosphate, which is likely present in small amounts from water-rock interactions, despite the lack of detection from Cassini's INMS [142].

We also ran the network expansion of KEGG in its entirety (incorporating all catalogued compounds and reactions), representing the full catalytic and metabolic potential of the biosphere, on the seed set of Enceladus with phosphate (Table 5).

4.5.3 Identifying minimal seed sets

We follow the algorithm described in Handorf et al., 2008 [135] to create random minimal seed sets which attempt to minimize the likelihood of obtaining seed sets with large complex biomolecules where possible:

A seed S is minimal if its scope ΣS contains the target compounds T and no proper subset of S fulfills this condition. S is a minimal seed set if:

$$T \subseteq \Sigma(S) \quad \text{and} \quad \forall S' \subset S : T \not\subset \Sigma(S') \tag{4.1}$$

To find minimal seed sets for each organism, we start by creating a list of all the compounds involved in all the reactions that the organism can catalyze. Because the target compounds are by definition the intersection of an organisms compounds with the target metabolites, the target compounds must be present in this list. Going down the list, we check if removing a substrate will cause a network expansion seeded with the remaining substrates to successfully produce all target compounds. If the removal does not impact the target compounds produced, the substrate stays removed. Else, we add it back to the list. Then we move onto the next substrate in the list, repeating until the entire list is traversed.

In this algorithm, the order of the list affects the minimal seed set which gets identified, so it is necessary to permute the list and repeat the algorithm to identify each of the 100 minimal seeds. However, we do not want to start with a completely randomized list for each organism, because ideally we want to remove large complex compounds, as to be left with seed sets composed preferentially with simpler compounds which are more abiogenically plausible to find in a uninhabited environment. Previous research has shown that the scopes of single complex biochemicals tend to be reachable by sets of simpler molecules [79]. Because of this, we initially order every list from largest to smallest molecular weight, but then perturb them such that heavier compounds tend to stay near the top, thus getting preferentially removed. Compounds without associated weights were added in random locations in the list.

We again follow the method laid out by Handorf et al. [135]. From the list, two randomly chosen compounds with mass difference Δm get exchanged with probability p:

$$p = \begin{cases} \exp(\frac{\Delta m}{\beta}) & \text{if } \Delta m > 0 \\ 1 & \text{if } \Delta m \leq 0 \end{cases}$$

The only exception to this rule is that if one of the compounds does not contain weight information, then $p = 0.5$. The parameter beta represents the degree of disorder allowed in the list, where $\beta = 0$ forbids disorder and $\beta = \infty$ ignores disorder. We follow the choice of Handorf et al. [135] and choose $\beta = 20$ amu.

4.5.4 Comparing and clustering seed sets

Similarity of seed sets were calculated using the Jaccard index. Clustering was computed using `scipy.cluster.hierarchy.linkage(method='average')`, where average refers to the unweighted pair group method with arithmetic mean (UPG-MA) algorithm.

Chapter 5

CONCLUSION

Biology is a menagerie of processes across scales on Earth. It is microscopic cells feeding on indivisible particles, and global ecosystems driven by tectonic motions. Despite this, my analyses reveal biochemical networks display common scaling laws governing their topology and biochemical diversity that cannot be fully explained by the structure of random reaction networks. These laws are independent of the level of organization they are sampled from, and seem to persist across different coarse grainings, as characterized by varying network projections. When considered alongside geochemical data, I find that these biochemical networks provide a medium for quantitatively investigating the viability of Earth life outside of our planet. An initial case study of biochemistry's viability on Enceladus suggests that individual organisms cannot attain essential compounds, and instead might require the support of biological functions at the scale of an ecosystem or biosphere. Collectively, my results indicate a deeper level of organization in biochemical networks than what is understood so far, providing a new framework for understanding the planetary-scale organization of biochemistry and how nested hierarchical levels are structured within it.

A key implication of my analysis is the importance of individuals sharing a common set of biochemical reactions in shaping the universal scaling laws observed across hierarchical levels. As described in the introduction, scaling laws often emerge in systems where universal mechanisms operate across different scales, yielding the same effective behavior independent the specific details of the system. It is in this sense scaling laws can uncover universal properties, motivating their widespread use in physics and increasing application to biology [59, 63, 65, 97–100]. In Chapter 2, I show that the relevant scaling parameter for biochemical organization is the number of biochemical compounds (in a network representation this is the size of the network). Individuals, ecosystems and the biosphere obey much the same scaling behavior for biochemical network structure, indicating the same universal mechanisms could operate across all three levels of organization. In physics, this kind of universality usually implies there is no preferred scale or basic unit. However, in the biological example uncovered here, the presence of specific scaling relations observed in real biochemical networks can be explained by biological individuals (lower-level networks) sharing a common set of reactions as basic 'units'.

Future work should explore the connections between the scaling relationships reported here and other work characterizing scaling behavior across living processes. For example, individuals are perhaps more tightly constrained in coarse-grained network structure, based on being able to more accurately predict them based on simple network characteristics. But ecosystems are more tightly constrained than individuals using more descriptive topological measures that can describe paths and clustering within biochemical networks. Whether or not this structure is truly a universal property of life's chemical systems is more difficult to conclude. Projecting ecosystem-level scaling to the biosphere as a whole does not recover the observed network properties for the biosphere-level network. Recently, scaling laws describing microbial diversity were used to predict Earth's global microbial diversity, and in particular to highlight how much diversity remains undiscovered [61]. It could be an analogous case here, where the uncovered scaling relations could be used to predict missing enzymatic diversity in the biosphere. Allometric scaling laws are derived by viewing living systems as localized physical objects with energy and power constraints. Here, scaling emerges due to an orthogonal view of living systems as distributed processes transforming matter within the space of chemical reactions.

How to project this structure onto simple mathematical objects that can be quantifiably characterized and compared remains a central problem of complex systems science. In physics, the relevant coarse-graining procedure is well understood, but we are not so far in complexity science: the first hurdle we must traverse is to identify the proper coarse-grained network representations for analysis. Existing literature cautions against using unipartite network projections, as it is argued they can lead to ”wrong“ interpretations of system properties such as degree in biochemical networks [47, 117]. In chapter 3, I find instead that whether or not this conclusion should be drawn is highly dependent on the particular characteristics of degree or the degree distribution under consideration. For example, all network projection types, aside from unipartite reaction networks, favor power-law degree distributions over other heavy-tailed alternatives to simply describe biochemical systems. The similarities and differences in the structure of different projections provides insight into the actual structure of the underlying system of interest. Given that there is no obvious answer for whether a system is scale-free, we advocate for studying all projections possible: regardless of whether or not a given projection is scale-free, all projections provide insights into the structure of the underlying system.

An important task, stemming from this work, is identifying the planetary-drivers of Earth’s biosphere-level biochemical network structure and how this has structured living systems across nested levels considering their geochemical context. It remains an open question as to what will ultimately explain the universal structure of Earth’s biochemical networks, or whether we should expect Earth life to exhibit similar scaling behavior, even on other worlds. In order to address this important question, in chapter 4 I lay out a framework to quantify the chemical compounds necessary to assess the viability of Earth's biochemistry in the context of other geospheres. I examine this framework as applied to Enceladus, executing the network expansion algorithm across metabolic networks of alkaliphilic bacteria and archaea in the chemical environment of Enceladus' subsurface ocean. I find that no organisms analyzed can produce any of the pre-established target metabolites in this environment. I determine that incorporating phosphorous by adding phosphate as a seed compound does not change my results. Instead, it appears that for these organisms to be viable, the chemical complexity of the seed sets, or number of seeds present, has to be much higher.

While chapter 4 provides these results, it is meant to act more of a proof of principle and proposed procedure to increase the quantitative nature of research investigating the possibility of planetary contamination as well as terraformation. There are many ways that this work could be expanded in order to better reflect geochemical reality as well as incorporate more theoretical considerations. For instance, can we identify distributions of molecular weights of compounds which tend to support biochemistry? Or link the expansions with knowledge of biochemical network topology, that we describe in Chapters 2 and 3, in order to find structure gaps in organismal networks which need to be filled to produce viable organisms [84]?

Intentionally seeding a planet with life seems likely only in the circumstance where a metabolism is specifically tailored to the environment, and even then there are questions about how well it could be self-sustaining. I believe that because life on Earth was a product of Earth's geochemistry, there is a significant bias to be viable only in a geochemical environment similar to the Earth's. While there is much more work to be done to quantify the risks, or possibilities, of Earth life being viable amongst other geospheres, I believe that we have laid significant groundwork for exciting research in this domain.

A final implication of my work is the consequence for our understanding of the origin of life, before the emergence of species. The existence of common network structure across all scales and levels of biochemical organization suggests a logic to the planetary-scale organization of biochemistry [101], which—if truly universal—would have been operative at the origin of life. While my analysis has uncovered universal scaling behavior for extant life, arising due to the structure of connectivity and diversity among the most common biochemical compounds and reactions, it remains to be determined whether the particular scaling reported herein is a by-product of shared biochemistry across all life, or if fundamental constraints on biochemical network structure, operative across scales from individuals to planets, drives lower-level individuals to necessarily share common reactions. If the latter is true it would have important implications for understanding the processes operative at the time of the last universal common ancestor. If the same global network structure, characterized by the same scaling laws, described Earth's biosphere throughout its evolutionary history, the emergence of individuals (as selectable units) with shared biochemistry would have played an important role in mediating a transition in the organization of Earth's chemical reaction networks.

REFERENCES

[1] Nigel Goldenfeld and Carl Woese. ‘Life is Physics: Evolution as a Collective Phenomenon Far From Equilibrium’. In: *Annu. Rev. Condens. Matter Phys.* 2.1 (Feb. 2011), pp. 375–399.

[2] Paul C W Davies and Sara Imari Walker. ‘The hidden simplicity of biology’. en. In: *Rep. Prog. Phys.* 79.10 (Oct. 2016), p. 102601.

[3] T Gisiger. ‘Scale invariance in biology: coincidence or footprint of a universal mechanism?’ en. In: *Biol. Rev. Camb. Philos. Soc.* 76.2 (May 2001), pp. 161–209.

[4] Nigel Goldenfeld, Tommaso Biancalani, and Farshid Jafarpour. ‘Universal biology and the statistical mechanics of early life’. In: *Philos. Trans. A Math. Phys. Eng. Sci.* 375.2109 (Dec. 2017).

[5] Kim Sterelny. *Universal biology.* 1997.

[6] Sara I Walker et al. ‘Exoplanet biosignatures: future directions’. In: *Astrobiology* 18.6 (2018), pp. 779–824.

[7] J E Lovelock. ‘A physical basis for life detection experiments’. en. In: *Nature* 207.997 (Aug. 1965), pp. 568–570.

[8] Evan D Dorn, Kenneth H Nealson, and Christoph Adami. ‘Monomer abundance distribution patterns as a universal biosignature: examples from terrestrial and digital life’. en. In: *J. Mol. Evol.* 72.3 (Mar. 2011), pp. 283–295.

[9] K H Nealson, A Tsapin, and M Storrie-Lombardi. ‘Searching for life in the Universe: unconventional methods for an unconventional problem’. en. In: *Int. Microbiol.* 5.4 (Dec. 2002), pp. 223–230.

[10] Priscilla E M Purnick and Ron Weiss. ‘The second wave of synthetic biology: from modules to systems’. en. In: *Nat. Rev. Mol. Cell Biol.* 10.6 (June 2009), pp. 410–422.

[11] Caleb J Bashor et al. ‘Rewiring cells: synthetic biology as a tool to interrogate the organizational principles of living systems’. en. In: *Annu. Rev. Biophys.* 39 (2010), pp. 515–537.

[12] Caleb Scharf et al. ‘A Strategy for Origins of Life Research’. en. In: *Astrobiology* 15.12 (Dec. 2015), pp. 1031–1042.

[13] Leroy Cronin and Sara Imari Walker. 'Beyond prebiotic chemistry'. In: *Science* 352.6290 (2016), pp. 1174–1175.

[14] L H Hartwell et al. 'From molecular to modular cell biology'. en. In: *Nature* 402.6761 Suppl (Dec. 1999), pp. C47–52.

[15] Fritjof Capra and Pier Luigi Luisi. *The Systems View of Life: A Unifying Vision*. en. Cambridge University Press, Apr. 2014.

[16] Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. en. CRC Press, July 2006.

[17] Albert-Laszlo Barabasi and Zoltan N Oltvai. 'Network biology: understanding the cell's functional organization'. In: *Nature reviews genetics* 5.2 (2004), p. 101.

[18] Hiroaki Kitano. 'Computational systems biology'. en. In: *Nature* 420.6912 (Nov. 2002), pp. 206–210.

[19] Niels Klitgord and Daniel Segrè. 'Ecosystems biology of microbial metabolism'. en. In: *Curr. Opin. Biotechnol.* 22.4 (Aug. 2011), pp. 541–546.

[20] Stephen R Proulx, Daniel E L Promislow, and Patrick C Phillips. 'Network thinking in ecology and evolution'. en. In: *Trends Ecol. Evol.* 20.6 (June 2005), pp. 345–353.

[21] Eric Smith and Harold J Morowitz. *The origin and nature of life on Earth: the emergence of the fourth geosphere*. Cambridge University Press, 2016.

[22] Paul G Falkowski, Tom Fenchel, and Edward F Delong. 'The microbial engines that drive Earth's biogeochemical cycles'. en. In: *Science* 320.5879 (May 2008), pp. 1034–1039.

[23] Rogier Braakman, Michael J Follows, and Sallie W Chisholm. 'Metabolic evolution and the self-organization of ecosystems'. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 114.15 (Apr. 2017), E3091–E3100.

[24] A Fernández et al. 'How stable is stable? Function versus community composition'. en. In: *Appl. Environ. Microbiol.* 65.8 (Aug. 1999), pp. 3697–3704.

[25] Jessica Flack. '12 Life's Information Hierarchy'. In: *From Matter to Life: Information and Causality* (2017), p. 283.

[26] Jordi van Gestel and Corina E Tarnita. ‘On the origin of biological construction, with a focus on multicellularity’. In: *Proceedings of the National Academy of Sciences* 114.42 (2017), pp. 11018–11026.

[27] Bernat Corominas-Murtra et al. ‘On the origins of hierarchy in complex networks’. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 110.33 (Aug. 2013), pp. 13316–13321.

[28] Steven H Strogatz. ‘Exploring complex networks’. In: *nature* 410.6825 (2001), p. 268.

[29] Réka Albert and Albert-László Barabási. ‘Statistical mechanics of complex networks’. In: *Reviews of modern physics* 74.1 (2002), p. 47.

[30] Sergey N Dorogovtsev and Jose FF Mendes. ‘Evolution of networks’. In: *Advances in physics* 51.4 (2002), pp. 1079–1187.

[31] Mark Newman, Albert-Laszlo Barabasi, and Duncan J Watts. *The structure and dynamics of networks.* Vol. 19. Princeton University Press, 2011.

[32] Albert-László Barabási et al. *Network science.* Cambridge university press, 2016.

[33] Mark Newman. *Networks.* Oxford university press, 2018.

[34] Mark Newman. *Networks: An Introduction.* en. Oxford University Press, Mar. 2010.

[35] M E J Newman. ‘Mixing patterns in networks’. en. In: *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 67.2 Pt 2 (Feb. 2003), p. 026126.

[36] Sergey V Buldyrev et al. ‘Catastrophic cascade of failures in interdependent networks’. In: *Nature* 464.7291 (2010), pp. 1025–1028.

[37] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks.* en. Cambridge University Press, Oct. 2008.

[38] Michelle Girvan and Mark EJ Newman. ‘Community structure in social and biological networks’. In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826.

[39] Ron Milo et al. ‘Network motifs: simple building blocks of complex networks’. In: *Science* 298.5594 (2002), pp. 824–827.

[40] Mark EJ Newman. 'The structure and function of complex networks'. In: *SIAM review* 45.2 (2003), pp. 167–256.

[41] Reka Albert. 'Scale-free networks in cell biology'. In: *Journal of cell science* 118.21 (2005), pp. 4947–4957.

[42] Stefano Boccaletti et al. 'Complex networks: Structure and dynamics'. In: *Physics reports* 424.4-5 (2006), pp. 175–308.

[43] George M Whitesides. 'Is the focus on "molecules" obsolete?' In: *Annual Review of Analytical Chemistry* 6 (2013), pp. 1–29.

[44] Sara Imari Walker and Cole Mathis. 'Network Theory in Prebiotic Evolution'. In: *Prebiotic Chemistry and Chemical Evolution of Nucleic Acids*. Springer, 2018, pp. 263–291.

[45] Petter Holme. 'Model validation of simple-graph representations of metabolism'. In: *Journal of The Royal Society Interface* 6.40 (2009), pp. 1027–1034.

[46] Petter Holme and Mikael Huss. 'Substance graphs are optimal simple-graph representations of metabolism'. In: *Chinese Science Bulletin* 55.27-28 (2010), pp. 3161–3168.

[47] Raul Montanez et al. 'When metabolism meets topology: Reconciling metabolite and reaction networks'. In: *Bioessays* 32.3 (2010), pp. 246–256.

[48] Hawoong Jeong et al. 'The large-scale organization of metabolic networks'. In: *Nature* 407.6804 (2000), pp. 651–654.

[49] Eugene V Koonin, Yuri I Wolf, and Georgy P Karev. *Power laws, scale-free networks and genome biology*. Springer, 2006.

[50] Roger Guimera and Luis A Nunes Amaral. 'Functional cartography of complex metabolic networks'. In: *nature* 433.7028 (2005), p. 895.

[51] Reiko Tanaka. 'Scale-rich metabolic networks'. In: *Physical review letters* 94.16 (2005), p. 168101.

[52] Marko Gosak et al. 'Network science of biological systems at different scales: a review'. In: *Physics of life reviews* (2017).

[53] Joshua E Goldford et al. 'Remnants of an ancient metabolism without phosphate'. In: *Cell* 168.6 (2017), pp. 1126–1134.

[54] Jason Raymond and Daniel Segrè. ‘The effect of oxygen on biochemical networks and the evolution of complex life’. In: *Science* 311.5768 (2006), pp. 1764–1767.

[55] Oliver Ebenhöh, Thomas Handorf, and Reinhart Heinrich. ‘Structural analysis of expanding metabolic networks’. en. In: *Genome Inform.* 15.1 (2004), pp. 35–45.

[56] Alice R Wattam et al. ‘Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center’. en. In: *Nucleic Acids Res.* 45.D1 (Jan. 2017), pp. D535–D542.

[57] Victor M Markowitz et al. ‘IMG/M: the integrated metagenome data management and comparative analysis system’. en. In: *Nucleic Acids Res.* 40.Database issue (Jan. 2012), pp. D123–9.

[58] Minoru Kanehisa and Susumu Goto. ‘KEGG: kyoto encyclopedia of genes and genomes’. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[59] John P DeLong et al. ‘Shifts in metabolic scaling, production, and efficiency across major evolutionary transitions of life’. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 107.29 (July 2010), pp. 12941–12945.

[60] Ian A Hatton et al. ‘The predator-prey power law: Biomass scaling across terrestrial and aquatic biomes’. en. In: *Science* 349.6252 (Sept. 2015), aac6284.

[61] Kenneth J Locey and Jay T Lennon. ‘Scaling laws predict global microbial diversity’. In: *Proceedings of the National Academy of Sciences* 113.21 (2016), pp. 5970–5975.

[62] M E Ritchie and H Olff. ‘Spatial scaling laws yield a synthetic theory of biodiversity’. en. In: *Nature* 400.6744 (Aug. 1999), pp. 557–560.

[63] Diego Garlaschelli, Guido Caldarelli, and Luciano Pietronero. ‘Universal scaling relations in food webs’. en. In: *Nature* 423.6936 (May 2003), pp. 165–168.

[64] Geoffrey B West, William H Woodruff, and James H Brown. ‘Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals’. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 99 (suppl 1) (Feb. 2002), pp. 2473–2478.

[65] G B West, J H Brown, and B J Enquist. ‘A general model for the origin of allometric scaling laws in biology’. en. In: *Science* 276.5309 (Apr. 1997), pp. 122–126.

[66] Miguel A Muñoz. 'Colloquium: Criticality and dynamical scaling in living systems'. In: (Dec. 2017). arXiv: 1712.04499 `[cond-mat.stat-mech]`.

[67] Thomas A McMahon and John Tyler Bonner. *On size and life*. Scientific American Library, 1983.

[68] H Eugene Stanley. *Phase transitions and critical phenomena*. Vol. 9. Oxford University Press, 1971.

[69] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 'Power-law distributions in empirical data'. In: *SIAM review* 51.4 (2009), pp. 661–703.

[70] Albert-László Barabási and Réka Albert. 'Emergence of scaling in random networks'. In: *science* 286.5439 (1999), pp. 509–512.

[71] Réka Albert, Hawoong Jeong, and Albert-László Barabási. 'Internet: Diameter of the world-wide web'. In: *nature* 401.6749 (1999), p. 130.

[72] Jean M Carlson and John Doyle. 'Highly optimized tolerance: A mechanism for power laws in designed systems'. In: *Physical Review E* 60.2 (1999), p. 1412.

[73] Réka Albert, Hawoong Jeong, and Albert-László Barabási. 'Error and attack tolerance of complex networks'. In: *nature* 406.6794 (2000), p. 378.

[74] Albert-László Barabási. 'Scale-free networks: a decade and beyond'. In: *science* 325.5939 (2009), pp. 412–413.

[75] Michael Mitzenmacher. 'A brief history of generative models for power law and lognormal distributions'. In: *Internet mathematics* 1.2 (2004), pp. 226–251.

[76] Elon Musk. 'Making Life Multi-Planetary'. In: *New Space* 6.1 (2018), pp. 2–11.

[77] John D Rummel. 'Planetary exploration in the time of astrobiology: protecting against biological contamination'. In: *Proceedings of the National Academy of Sciences* 98.5 (2001), pp. 2128–2131.

[78] Rocco L Mancinelli. 'Planetary protection and the search for life beneath the surface of Mars'. In: *Advances in Space Research* 31.1 (2003), pp. 103–107.

[79] Thomas Handorf, Oliver Ebenhöh, and Reinhart Heinrich. 'Expanding metabolic networks: scopes of compounds, robustness, and evolution'. In: *Journal of molecular evolution* 61.4 (2005), pp. 498–512.

[80] Elhanan Borenstein et al. ‘Large-scale reconstruction and phylogenetic analysis of metabolic environments’. In: *Proceedings of the National Academy of Sciences* (2008).

[81] Oliver Ebenhöh, Thomas Handorf, and Reinhart Heinrich. ‘A cross species comparison of metabolic network functions’. In: *Genome Informatics* 16.1 (2005), pp. 203–213.

[82] Shiri Freilich et al. ‘Metabolic-network-driven analysis of bacterial ecological strategies’. In: *Genome biology* 10.6 (2009), R61.

[83] Harold Morowitz and Eric Smith. ‘Energy flow and the organization of life’. In: *Complexity* 13.1 (2007), pp. 51–59.

[84] Hyunju Kim et al. ‘Universal scaling across biochemical networks on Earth’. In: *bioRxiv* (2018), p. 212118.

[85] Jörg Stelling et al. ‘Metabolic network structure determines key aspects of functionality and regulation’. en. In: *Nature* 420.6912 (Nov. 2002), pp. 190–193.

[86] Gil Benkö, Christoph Flamm, and Peter F Stadler. ‘Generic Properties of Chemical Networks: Artificial Chemistry Based on Graph Rewriting’. In: *Advances in Artificial Life*. Springer Berlin Heidelberg, 2003, pp. 10–19.

[87] Philipp-Maximilian Jacob and Alexei Lapkin. ‘Statistics of the network of organic chemistry’. en. In: *Reaction Chemistry & Engineering* 3.1 (2018), pp. 102–118.

[88] Conner I Sandefur, Maya Mincheva, and Santiago Schnell. ‘Network representations and methods for the analysis of chemical and biochemical pathways’. en. In: *Mol. Biosyst.* 9.9 (Sept. 2013), pp. 2189–2200.

[89] Florian Centler and Peter Dittrich. ‘Chemical organizations in atmospheric photochemistries—A new method to analyze chemical reaction networks’. In: *Planet. Space Sci.* 55.4 (Mar. 2007), pp. 413–428.

[90] Ricard V Sole and Andreea Munteanu. ‘The large-scale organization of chemical reaction networks in astrophysics’. In: *EPL* 68.2 (2004), p. 170.

[91] Georg Basler et al. ‘Evolutionary significance of metabolic network properties’. en. In: *J. R. Soc. Interface* 9.71 (June 2012), pp. 1168–1176.

[92] S T Buckland, B Efron, and R J Tibshirani. ‘An Introduction to the Bootstrap’. In: *Biometrics* 50.3 (1994), p. 890.

[93] D A Fell and A Wagner. 'The small world of metabolism'. en. In: *Nat. Biotechnol.* 18.11 (Nov. 2000), pp. 1121–1122.

[94] Michael P H Stumpf, Carsten Wiuf, and Robert M May. 'Subnets of scale-free networks are not scale-free: sampling properties of networks'. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.12 (Mar. 2005), pp. 4221–4224.

[95] Laura A Hug et al. 'A new view of the tree of life'. en. In: *Nat Microbiol* 1 (Apr. 2016), p. 16048.

[96] N R Pace. 'The universal nature of biochemistry'. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 98.3 (Jan. 2001), pp. 805–808.

[97] Christopher P Kempes et al. 'Predicting Maximum Tree Heights and Other Traits from Allometric Scaling and Resource Limitations'. In: *PLoS One* 6.6 (2011), e20551.

[98] Christopher P Kempes et al. 'Drivers of Bacterial Maintenance and Minimal Energy Requirements'. In: *Front. Microbiol.* 8 (2017).

[99] R V Solé et al. 'Criticality and scaling in evolutionary ecology'. en. In: *Trends Ecol. Evol.* 14.4 (Apr. 1999), pp. 156–160.

[100] Brown, James H., and Geoffrey B. West, ed. *Scaling in Biology.* Oxford University Press, 2000.

[101] Rogier Braakman and Eric Smith. 'The compositional and evolutionary logic of metabolism'. In: *Physical biology* 10.1 (2012), p. 011001.

[102] Paulien Hogeweg. 'From population dynamics to ecoinformatics: Ecosystems as multilevel information processing systems'. In: *Ecol. Inform.* 2.2 (2007), pp. 103–111.

[103] Olof Görnerup and James P Crutchfield. 'Hierarchical self-organization in the finitary process soup'. en. In: *Artif. Life* 14.3 (2008), pp. 245–254.

[104] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX.* Tech. rep. Los Alamos National Laboratory (LANL), 2008.

[105] Ulrik Brandes. 'A faster algorithm for betweenness centrality'. In: *J. Math. Sociol.* 25.2 (June 2001), pp. 163–177.

[106] Albert-László Barabási and Eric Bonabeau. 'Scale-free networks'. In: *Scientific american* 288.5 (2003), pp. 60–69.

[107] Lun Li et al. 'Towards a theory of scale-free graphs: Definition, properties, and implications'. In: *Internet Mathematics* 2.4 (2005), pp. 431–523.

[108] Anna D Broido and Aaron Clauset. 'Scale-free networks are rare'. In: *arXiv preprint arXiv:1801.03400* (2018).

[109] Raya Khanin and Ernst Wit. 'How scale-free are biological networks'. In: *Journal of computational biology* 13.3 (2006), pp. 810–818.

[110] David E Featherstone and Kendal Broadie. 'Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network'. In: *Bioessays* 24.3 (2002), pp. 267–274.

[111] Nabil Guelzim et al. 'Topological and causal structure of the yeast transcriptional regulatory network'. In: *Nature genetics* 31.1 (2002), p. 60.

[112] Siming Li et al. 'A map of the interactome network of the metazoan C. elegans'. In: *Science* (2004).

[113] Marcus Kaiser. 'A tutorial in connectome analysis: topological and spatial features of brain networks'. In: *Neuroimage* 57.3 (2011), pp. 892–907.

[114] Minoru Kanehisa et al. 'KEGG as a reference resource for gene and protein annotation'. In: *Nucleic acids research* 44.D1 (2015), pp. D457–D462.

[115] Minoru Kanehisa et al. 'KEGG: new perspectives on genomes, pathways, diseases and drugs'. In: *Nucleic acids research* 45.D1 (2016), pp. D353–D361.

[116] Paul CW Davies et al. 'Signatures of a shadow biosphere'. In: *Astrobiology* 9.2 (2009), pp. 241–249.

[117] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. 'Hypergraphs and cellular networks'. In: *PLoS computational biology* 5.5 (2009), e1000385.

[118] Alice R Wattam et al. 'Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center'. In: *Nucleic acids research* 45.D1 (2016), pp. D535–D542.

[119] Victor M Markowitz et al. 'IMG: the integrated microbial genomes database and comparative analysis system'. In: *Nucleic acids research* 40.D1 (2011), pp. D115–D122.

[120] Aaron Clauset, Maxwell Young, and Kristian Skrede Gleditsch. 'On the frequency of severe terrorist events'. In: *Journal of Conflict Resolution* 51.1 (2007), pp. 58–87.

[121] Anna D. Broido. *SFAnalysis*. `https://github.com/adbroido/SFAnalysis`. 2017.

[122] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. 'powerlaw: a Python package for analysis of heavy-tailed distributions'. In: *PloS one* 9.1 (2014), e85777.

[123] Abigail C Allwood et al. 'Stromatolite reef from the Early Archaean era of Australia'. In: *Nature* 441.7094 (2006), p. 714.

[124] Allen P Nutman et al. 'Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures'. In: *Nature* 537.7621 (2016), p. 535.

[125] S Blair Hedges. 'The origin and evolution of model organisms'. In: *Nature Reviews Genetics* 3.11 (2002), p. 838.

[126] Timothy M Lenton. 'Testing Gaia: the effect of life on Earth's habitability and regulation'. In: *Climatic Change* 52.4 (2002), pp. 409–422.

[127] Timothy M Lenton and Marcel van Oijen. 'Gaia as a complex adaptive system'. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 357.1421 (2002), pp. 683–695.

[128] Axel Kleidon. 'Beyond Gaia: thermodynamics of life and earth system functioning'. In: *Climatic Change* 66.3 (2004), pp. 271–319.

[129] Harold J Morowitz, Eric Smith, and Vijayasarathi Srinivasan. 'Selfish metabolism'. In: *Complexity* 14.2 (2008), pp. 7–9.

[130] Patrick May et al. 'Integration of proteomic and metabolomic profiling as well as metabolic modeling for the functional analysis of metabolic networks'. In: *Bioinformatics for Comparative Proteomics*. Springer, 2011, pp. 341–363.

[131] Kai Kruse and Oliver Ebenhöh. 'Comparing flux balance analysis to network expansion: producibility, sustainability and the scope of compounds'. In: *Genome Informatics 2008: Genome Informatics Series Vol. 20*. World Scientific, 2008, pp. 91–101.

[132] Christopher R Glein, John A Baross, and J Hunter Waite Jr. 'The pH of Enceladus' ocean'. In: *Geochimica et Cosmochimica Acta* 162 (2015), pp. 202–219.

[133] Christopher P McKay et al. ‘The possible origin and persistence of life on Enceladus and detection of biomarkers in the plume’. In: *Astrobiology* 8.5 (2008), pp. 909–919.

[134] J Hunter Waite Jr et al. ‘Liquid water on Enceladus from observations of ammonia and 40 Ar in the plume’. In: *Nature* 460.7254 (2009), p. 487.

[135] Thomas Handorf et al. ‘An environmental perspective on metabolism’. In: *Journal of theoretical biology* 252.3 (2008), pp. 530–537.

[136] Oliver Ebenhoeh, T Handorf, and Daniel Kahn. ‘Evolutionary changes of metabolic networks and their biosynthetic capacities’. In: *IEE Proceedings-Systems Biology* 153.5 (2006), pp. 354–358.

[137] Ludovic Cottret et al. ‘Enumerating precursor sets of target metabolites in a metabolic network’. In: *International Workshop on Algorithms in Bioinformatics*. Springer. 2008, pp. 233–244.

[138] Sarath Chandra Janga and M Madan Babu. ‘Network-based approaches for linking metabolism with environment’. In: *Genome biology* 9.11 (2008), p. 239.

[139] Victor G Stepanov et al. ‘Bacillus pumilus SAFR-032 genome revisited: sequence update and re-annotation’. In: *PloS one* 11.6 (2016), e0157331.

[140] Peter JA Cock et al. ‘Biopython: freely available Python tools for computational molecular biology and bioinformatics’. In: *Bioinformatics* 25.11 (2009), pp. 1422–1423.

[141] Toshiaki Katayama, Mitsuteru Nakao, and Toshihisa Takagi. ‘TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services’. In: *Nucleic acids research* 38.suppl_2 (2010), W706–W711.

[142] Melissa Guzman et al. ‘Collecting amino acids in the Enceladus plume’. In: *International Journal of Astrobiology* (2018), pp. 1–13.

APPENDIX A

STATEMENT OF CO-AUTHOR PERMISSIONS

All co-authors have granted their permissions to use articles Kim and Smith et al. 2018 (in review) and Smith et al. 2018 (submitted) for Chapters 2 and 3 respectively.