Title: A mechanism of abiogenesis based on complex reaction networks organized by seed-dependent autocatalytic systems

Authors

Zhen Peng[1], Jeff Linderoth[1,2], David A. Baum[1,3*]

Affiliations

[1]Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison WI 53706, USA

[2]Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison WI 53706, USA

[3]Department of Botany, University of Wisconsin-Madison, Madison WI 53706, USA

*Correspondence to: David A. Baum (dbaum@wisc.edu).

Abstract

Life is the canonical example of a complex system, consisting of diverse chemical components that are organized in a specific way that allows perpetuation of the living state. In contrast, the abiotic environment, which life feeds on and originated from, is much simpler and less organized. The complexity gap between the biotic and abiotic worlds, and the lack of direct observation of abiogenesis, has made explaining the origin of life one of the hardest scientific questions. A promising strategy for addressing this problem is to identify features shared by abiotic and biotic chemical systems that permit the stepwise accretion of complexity. We used such a rationale to compare abiotic and biotic reaction networks in order to evaluate the presence of autocatalysis, the underlying basis of biological self-propagation, to see if it is structured in such a way as to permit stepwise complexification. We develop the concept of, and provide an algorithm to detect, seed-dependent autocatalytic systems (SDASs), namely subnetworks that can use food chemicals to self-propagate but cannot emerge without being first seeded by some non-food chemicals. We show that serial activation of SDASs can result in incremental

complexification. Furthermore, we identify life-like features that emerge during the accretion of SDASs that open up new ecological opportunities and improve the efficiency of food utilization. SDAS theory, thus, provides a conceptual roadmap from a simple abiotic environment to primitive forms of life, without the need for linear genetic polymers at the outset (though these may be added later). This framework also suggests new experiments that have the potential to detect the spontaneous emergence of life-like features, such as self-propagation and adaptability.

1. INTRODUCTION

Life is the paradigmatic example of a complex system (Ladyman et al., 2013; Mitchell and Newman, 2001). It consists of a large number of chemical components, some of which are large, energy-expensive molecules, such as nucleic acids, proteins, polysaccharides, organic metabolites, and cofactors. Furthermore, these components are not simply lumped together but are organized and coordinated in such a way that the entire system can resist environmental perturbations and grow or divide to give rise to more life. Such organization is not cheap, of course, which is why life must consume energy to maintain its internal order. It is almost magical that complex organisms can use "untargeted" energy sources such as light to convert simple ultimate food sources (e.g., water, carbon dioxide, and minerals) into more living matter of the same kind. The origin-of-life mystery boils down to the question of how a system with sufficient complexity to conduct such orderly and improbable conversions could emerge spontaneously, when there was no prior design to follow or template to copy.

The solution to this conundrum lies, we believe, not in looking for particular molecules or reactions, but in looking at the emergent properties of networks of chemical reactions. Provided that chemical reaction networks contain many autocatalytic motifs, systems with a pre-existing ability to sustain themselves may arise readily, and then become progressively more complicated through the accretion of more autocatalytic modules. In our recent work (Peng et al., 2020), we used analyses of "toy" reaction networks to show that systems of autocatalytic cycles can exhibit features of ecological interactions, which provides the basis for complex dynamics, including succession and evolution. Here, we sought to extend this analysis to real chemical reaction networks to see if they have features needed for evolutionary dynamics. Specifically, we focused on databases that "bracket" the origin of life, the radiolytic and geochemical reactions assembled by Adam et al., 2021 serving to represent chemistry without life, and a curated subset of biochemical reactions (Xavier et al., 2020) serving to constrain the metabolic network of the Last Universal Common Ancestor (LUCA). Our reasoning is that features shared by both networks are very likely to also apply to systems on the path from non-life to life. Thus, if we find evidence of mechanisms and network structures that allow for the evolutionary accretion of complexity in both databases, then it is reasonable to assume that abiotic reactions could gradually become organized to yield systems deserving the label "life".

In this study, we focus on the concept of a seed-dependent autocatalytic system (SDAS), which is a network motif that can be activated by a rare chemical event but can, once activated, sustain itself. A network with many potential SDASs can gradually incorporate new components and re-organize old components. We suggest that, when actualized in a spatially and temporally structured environment, a multi-SDAS network may be able to show adaptive evolution. We start by formalizing the concept of network expansion, which is the basic procedure needed to map out a network's architecture. Then, we introduce SDASs, and describe an algorithm for identifying them within a stoichiometric matrix. Finally, we show that SDASs are found in both the abiotic and biotic chemical reaction networks and are organized in such a way so as to store information of past environments and allow for stepwise accretion of complexity. Furthermore, the larger, biotic network provides examples where later-activated processes improve the efficiency of the system as a whole and open new ecological opportunities. In combination, this shows that reaction networks can show evolutionary dynamics – periodically finding new, transiently steady states in such a way that existing resources can come to be used more efficiently and additional resources can come to be exploited over time. We end by discussing some implications of our results, several remaining challenges, and how our theory can be used to guide laboratory experiments.

2. RESULTS

2.1 Databases of abiotic and biotic reactions

The abiotic reaction database we analyzed is based on a recently published reaction network assembled from seven decades of publications (Adam et al., 2021). We have also added some additional well-known abiotic organic reactions to the database, including the classical formose reaction (Breslow, 1959), but without formaldehyde dimerization because it is very slow and its reaction mechanism is yet to be determined (Cleaves, 2011; Haas et al., 2020). The reaction network assembled by Adam et al., 2021 includes: free radical reactions, mineral geochemical reactions, amino acid production, chloride radical and polar reactions, nitrile radical and polar reactions, RNA nucleotide assembly, nuclear decay, and physicochemical reactions. Sources of radiation such as X-rays, ultraviolet, and visible light are treated as reactants and products in this database. In light of this, we will use "entity" to refer to both chemicals and electromagnetic

energy sources in this specific database. In part because of these energy sources, most of the reactions in the abiotic database are irreversible. Although only covering a small portion of known abiotic chemical reactions, these reactions may still be used to test the applicability of our theoretical framework to real chemical reaction networks.

The biotic reaction database we analyzed is based on a recently published reaction network (Xavier et al., 2020), which was obtained by removing reactions that only occur in eukaryotes and reactions dependent on O_2 from the KEGG reaction database (Kanehisa, 2019; Kanehisa et al., 2021; Kanehisa and Goto, 2000). Xavier et al., 2020 claimed that the resulting reaction network could be a proxy of the primordial metabolism of LUCA. We also added five spontaneous reversible reactions that are missing from KEGG, such as $H_2O \leftrightarrow H^+ + OH^-$ and $H_2CO_3 \leftrightarrow H^+ + HCO_3^-$. We acknowledge that, because most of the reactions in the KEGG database are catalyzed by enzymes, it is likely that many of them could not occur at sufficiently high rates in a prebiotic world to have been relevant before biological catalysts had evolved. Nonetheless, since all these reactions are chemically feasible, we reasoned that the relationships between reactants and products in such a curated biochemical reaction network is meaningful and that features shared by it and the abiotic network should be relevant to the origin of life.

We preprocessed the databases such that every reaction appears only once and had clear stoichiometry. Each reaction in KEGG is assumed to be reversible, so they were split into two unidirectional reactions (see Materials and Methods). After curation, the abiotic reaction database consists of 277 entities and 717 unidirectional reactions for a reaction density of 2.59 reactions/entity (Table S1). The resulting biotic reaction database consists of 4216 chemical species and 8400 unidirectional reactions for a reaction density of 1.99 reactions/species (Table S2). Fig. S1 shows the histograms of the numbers of reactions that an entity or chemical species is involved in for the two databases. Both distributions fit a power law, being highly left-skewed, meaning that most entities/chemicals are involved in a small number of reactions.

2.2 Network expansion and tier-0 systems

Reaction databases are just collections of reactions allowed to occur. To generate an organized subnetwork for analysis, we define a network expansion operation, $\Xi(\boldsymbol{S_O}, \boldsymbol{R})$, which calculates all

reactions and chemical species that can be accessed given a starting set of chemical species $\boldsymbol{S_O}$ and a set of all allowed reactions $\boldsymbol{R}$. The network expansion (see Materials and Methods for details) starts with $\boldsymbol{S_O}$ and scans $\boldsymbol{R}$ for any reactions that are not yet in the subnetwork, but whose reactants are all present in the current subnetwork. These reactions and their products are added to the subnetwork and the expansion iterates until no more chemical reactions can be added to the subnetwork. The reaction subnetwork resulting from the expansion is described by the tuple ($\boldsymbol{S_E}$, $\boldsymbol{R_E}$), where $\boldsymbol{S_E}$ is the set of chemical species and $\boldsymbol{R_E}$ is the set of reactions. The network expansion operation can be visualized using a stoichiometric matrix (Fig. S2A-D), where each row represents a chemical species and each column represents a unidirectional reaction, with stoichiometric coefficients as the entries (negative values for reactants and positive values for products).

To apply network expansion to explore the properties of chemical systems, we need to specify a set of chemical species that are assumed to be provided by the environment, which we will call the ultimate food set (even if some of these chemicals are also produced by reactions within the subnetwork). A full expansion starting from the ultimate food set generates a reaction subnetwork that would include all chemicals that would be expected to be generated at a non-zero rate in an environment receiving an ongoing flux of the ultimate food species. We will call this subnetwork the tier-0 system, reflecting the fact that no additional events are needed for its generation except for the provision of the ultimate food (Fig. S2A-D).

Although there is still no wide consensus on the environmental conditions of the prebiotic sites where life originated (Damer and Deamer, 2015; Donaldson et al., 2004; Lathe, 2004; Marín-Yaseli et al., 2016; Martin et al., 2008; Maruyama et al., 2019; Sojo et al., 2016; Wächtershäuser, 1988; Westall et al., 2018), to illustrate the approach, we selected sets of simple entities or chemical species as ultimate food. For the abiotic reaction database, we chose {H_2, CH_4, NO, FeS_2, visible light} as the ultimate food set. A full expansion starting from this set generated a tier-0 system with no additional chemicals and zero reactions. For the biotic reaction database, we chose {H_2O, CO_2, NH_3, H_2S, H_2SO_4, H_2SO_3, HSO_3^-, H_3PO_4, $H_4P_2O_7$} as the ultimate food set. The full expansion starting from this set generated a tier-0 system with 30 chemical species and 44 unidirectional reactions (i.e., 22 reversible reactions). This network contains substructures that can be drawn as autocatalytic cycles (Blokhuis et al., 2020; Peng et al., 2020), an example of

which is shown in Fig. S2. However, because all the members in the tier-0 system are either provided as the ultimate food or spontaneously synthesized from the ultimate food, these autocatalytic cycles lack significance beyond suggesting that the system might show non-linear dynamics upon initiation or perturbation.

2.3 Seeds, seed-dependent autocatalytic systems (SDASs), and tier-1 systems

The materials and energy for building a tree already exist in the environment, but without a seed buried in the soil, materials like water, carbon dioxide, nitrate, sodium and energy from sunlight will not spontaneously form a new tree. It is the seed that provides the information for organizing the flow of simple materials and energy into a structure able to grow and self-propagate. That life must come from life has long been appreciated, at least since Louis Pasteur's famous gooseneck flask experiment. How then did life first get started? Given that a seed triggering life must have been generated by something other than prior life, where could these primordial seeds come from? Here we will show that seeding, or at least a prototype of seeding, exists in both abiotic and biotic reaction networks, and can induce systems more complex than tier-0 systems.

Imagine introducing a small amount of a new chemical species, a candidate seed, P, to an activated tier-0 system (Fig. S2E). P may react with some of the chemical species in the tier-0 system to generate new chemical species, which can result in a network expansion to yield a bigger reaction subnetwork, and the chemical species and reactions that are not involved in the tier-0 system are called a tier-1 system (Fig. S2F-H), representing the fact that one seeding event is required to induce such a system. While P allows all chemicals in the tier-1 system to be formed, we will only consider P to be a "seed" if the system it initiates is one that has the potential to sustain itself in an open environment, namely one experiencing constant dilution and influx of the ultimate food. The only way for a tier-1 system to persist in an open environment is to have a network topology that allows materials and energy in the tier-0 system to be converted to P. This means that the tier-1 system induced by P will only be viable if it contains at least one combination of reactions that are collectively autocatalytic. This implies that a net reaction equation can be written such that the reactants and products share no chemical species, only tier-0 chemicals are among the reactants, and all tier-1 chemicals are present in the products. Such a viable tier-1 system is here defined as a seed-dependent autocatalytic system (SDAS). SDASs

are similar to pRAFs in the RAF theory (Steel et al., 2020), because neither can be constructed simply from the food set. However, they are not identical because SDASs require specific stoichiometric relationships among the involved reactions while RAF theory does not consider stoichiometry.

To identify SDASs, we developed a linear programming algorithm inspired by Blokhuis et al., 2020. The key criterion is that for the submatrix consisting of all columns (the *q*th to *n*th columns in Fig. 1) and rows (the *p*th to *m*th rows in Fig. 1) induced by the seed, there exists a vector of non-negative elements $\boldsymbol{x} = (x_q, x_{q+1}, \dots, x_n)$ such that for every row in the submatrix, the dot product of this row and $\boldsymbol{x}$ is positive (Fig. 1), or

$$\sum_{j=q}^{n} x_j s_{ij} > 0 \ (x_j \geq 0) \quad \forall i \in [p, m] \tag{1}$$

where s_{ij} is the entry at the *i*th row and *j*th column of the stoichiometric matrix.

Whether the condition described by Equation (1) can be satisfied can be calculated by linear programming and we may further use integer programming to find autocatalytic motifs within the SDAS satisfying additional constraints, for example the smallest autocatalytic cores, meaning the ones that contain the fewest entities or reactions (see Materials and Methods).

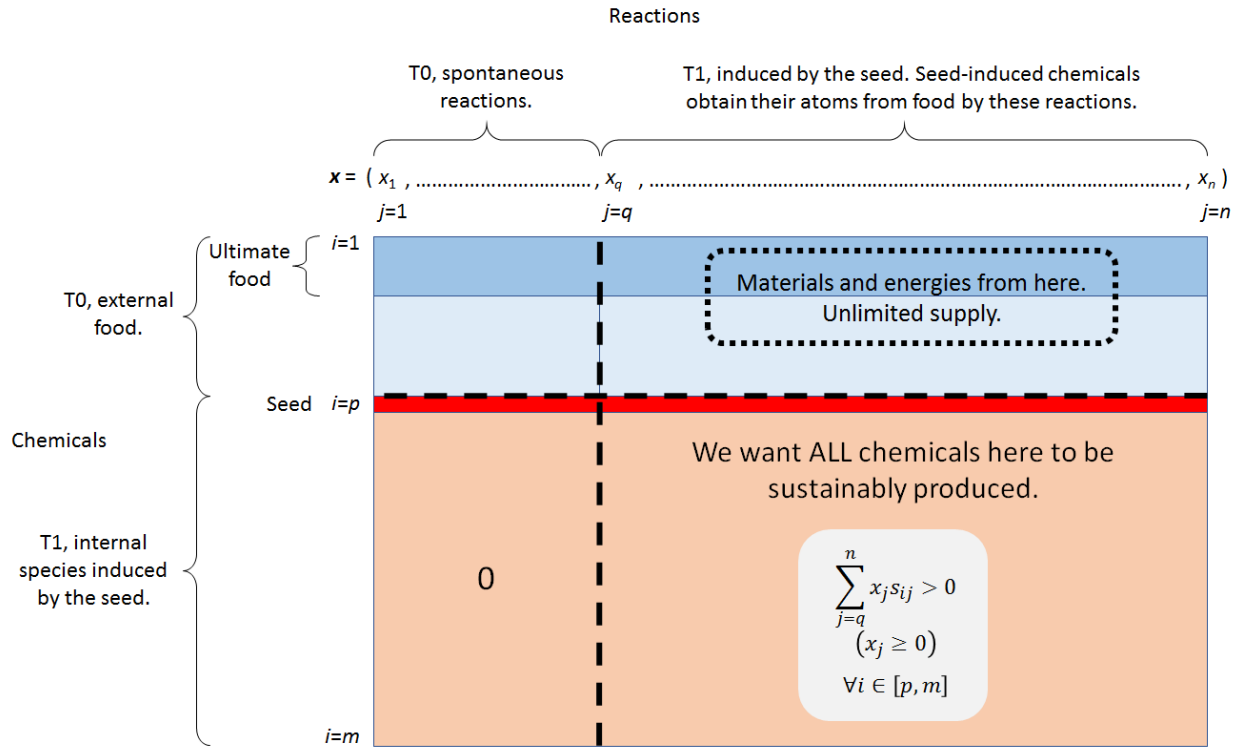


Fig. 1. Detection of SDASs. A stoichiometric matrix resulting from the network expansion starting with an external food set and a seed set can be split into four submatrices: the upper left one with rows representing external food and columns representing reactions involving only the external food, the upper right one with rows representing external food and columns representing reactions induced by seeding, the lower left one with rows representing internal chemical species induced by seeding and columns representing reactions involving only the external food, and the lower right one with rows representing internal chemical species induced by seeding and columns representing reactions induced by seeding. The four submatrices have different importance for detecting a SDAS. The two left ones can be ignored because none of their reactions involve chemical species outside the tier-0 system. The two right submatrices have reactions induced by seeding. Of these, the top submatrix can also be ignored, because these chemicals are provided for free by the environment, either as ultimate food or as products of reactions in the tier-0 system. So, our goal is to determine if all chemical species in the lower right submatrix can be sustainably produced by consuming the external food. Therefore, we want to find a linear combination of the reactions in this submatrix such that every internal chemical species can have a positive net change after this combination of reactions happen.

For the abiotic reaction database, with the ultimate food set {H_2, CH_4, NO, FeS_2, visible light}, 12 of the 272 non-tier-0 entities can serve as seeds capable of inducing a tier-1 SDAS (Table S3). Further analysis showed that all 12 entities induce the same tier-1 system, which includes 91 entities and 220 unidirectional reactions that were absent from the tier-0 system (Table S4). It is worth noting that, apart from the 12 seeds, the other 79 entities in the tier-1 system cannot individually seed the system. We define entities that can each induce the same tier-1 SDAS as a "clique" – these 12 entities comprise a clique of interchangeable seeds. It is worth noting that seeding can, in principle be achieved by simultaneous introduction of multiple entities, but our notion of clique only applies to seeds that can individually induce an SDAS. The smallest combination of reactions within the tier-1 SDAS that satisfies Equation (1) defines a complex autocatalytic core whose food is CH_4, NO, and visible light and whose waste is H_2CNH and infrared light (Fig. 2, Table S5).

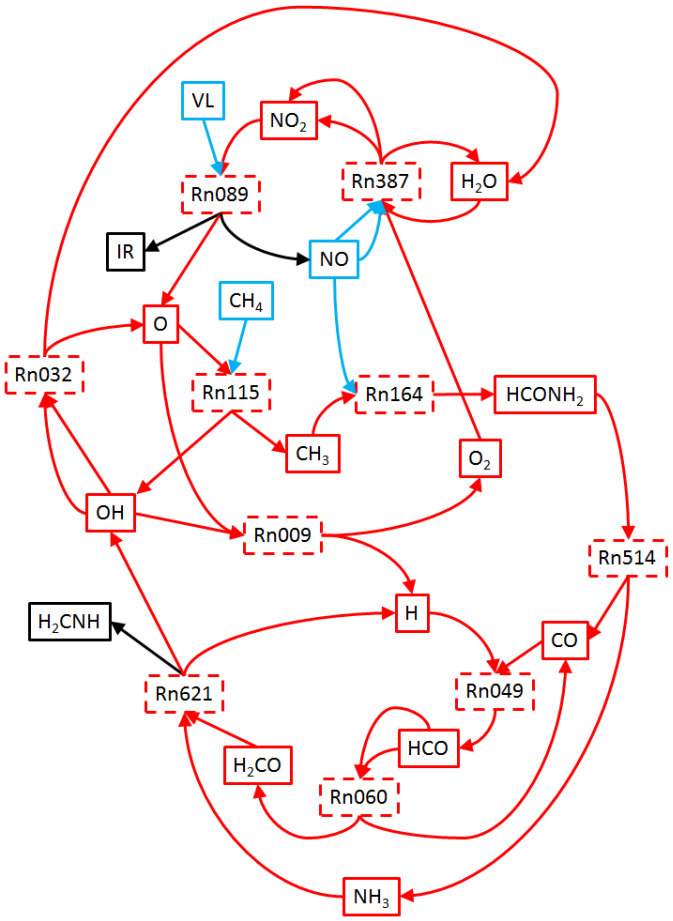


Fig. 2. A minimum autocatalytic core identified within the tier-1 SDAS in the abiotic reaction network. Red solid boxes: members of the autocatalytic core. Cyan solid boxes: food of the net reaction. Black solid box: waste of the net reaction. Red dashed boxes: reactions. Cyan arrows: food consumption. Black arrows: waste production. VL: visible light. IR: infrared.

For the biotic reaction database, with the ultimate food set {H_2O, CO_2, NH_3, H_2S, H_2SO_4, H_2SO_3, HSO_3^-, H_3PO_4, $H_4P_2O_7$}, 301 of the 4186 non-tier-0 chemical species are viable seeds that can induce a tier-1 system (Table S6). These 301 species belong to two cliques: 267 seed the same 301-species (736-reaction) tier-1 SDAS, whereas the other 34 seed a partially overlapping 357-species (916-reaction) tier-1 SDAS (Table S7, Table S8). Every species in the 34-member clique is a pyrimidine nucleoside or a derivative thereof, whereas the 267-member clique includes metabolites of many sizes, from the 2-carbon acetylene to the pyrimidine 5,6-dihydrouracil and even the 42-carbon celloheptaose. The smallest combination of reactions within the 301-species (736-reaction) tier-1 SDAS that satisfies Equation (1) defines a complex autocatalytic core whose food is H_2O, CO_2, and $H_4P_2O_7$ and whose waste is H_3PO_4 and H_2O_2 (Fig. 3, Table S9).

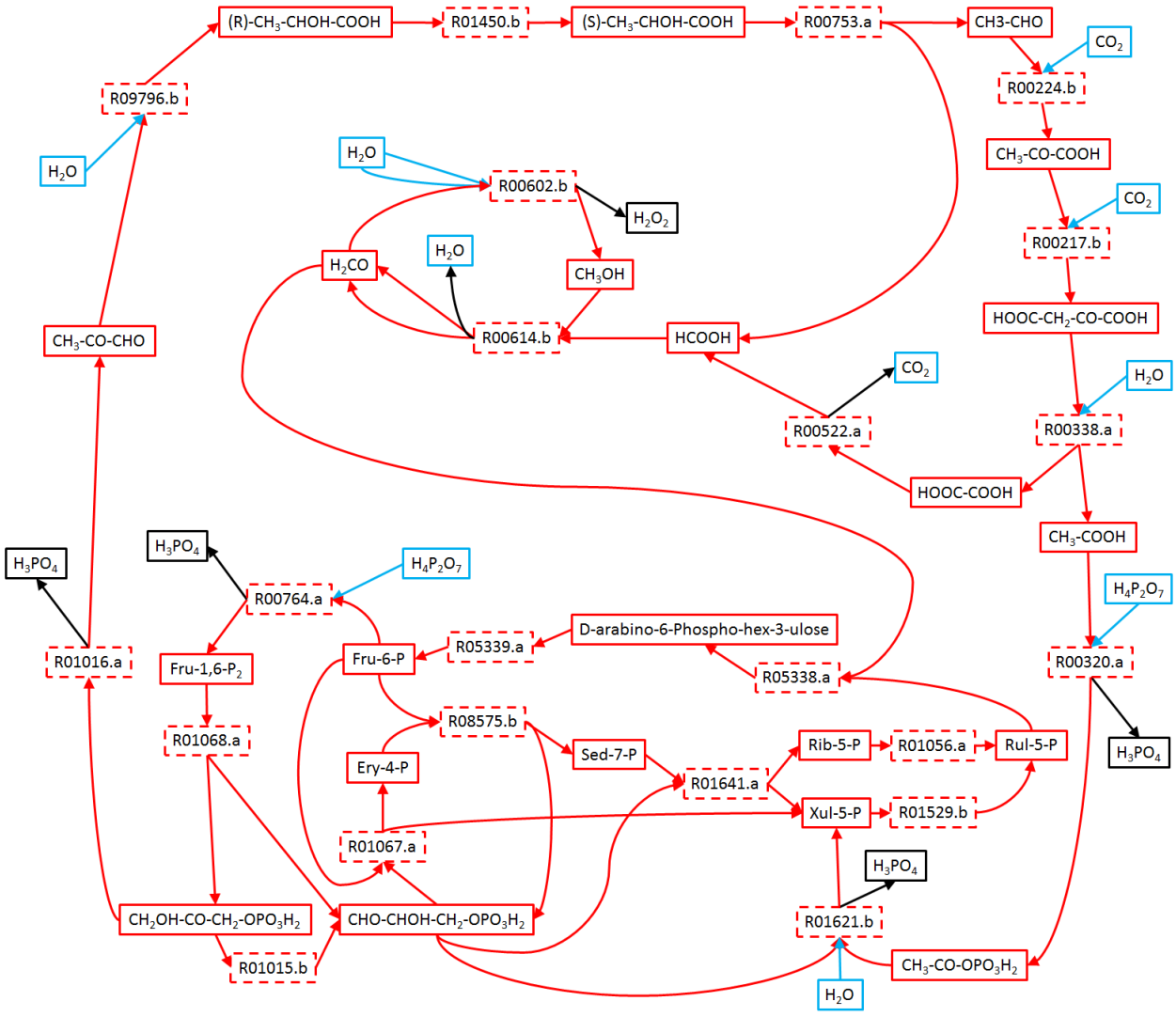


Fig. 3. A minimum autocatalytic core identified within tier-1 SDAS induced by the the 267-member clique in the biotic reactions. Red solid boxes: members of the autocatalytic core. Cyan solid boxes: food of the net reaction. Black solid box: waste of the net reaction. Red dashed boxes: reactions. Cyan arrows: food consumption. Black arrows: waste production.

There is nothing in the concept of seeding that requires a "seed" to be a single chemical. We will label cases in which multiple chemicals are needed to initiate a SDAS as "interdependence." Given a set of chemical species as the external food, interdependence between multiple non-food chemical species that form a set $\boldsymbol{U}$ can be detected if (a) $\boldsymbol{U}$ can induce a viable SDAS that sustainably synthesizes $\boldsymbol{U}$ from the external food and (b) for any non-empty $\boldsymbol{V} \subset \boldsymbol{U}$, $\boldsymbol{V}$ cannot induce a viable SDAS supporting $\boldsymbol{V}$.

We can illustrate the concept of interdependence with a few examples. For the abiotic reaction database with {H_2, CH_4, NO, FeS_2, visible light} as the ultimate food set, neither glycolaldehyde ($C_2H_4O_2$) nor NH_3 can seed a viable tier-1 system (Table S3). Nevertheless, the set {$C_2H_4O_2$, NH_3} can seed a 105-entity (255-reaction) tier-1 system (Table S10). Likewise, for the biotic reaction database with {H_2O, CO_2, NH_3, H_2S, H_2SO_4, H_2SO_3, HSO_3^-, H_3PO_4, $H_4P_2O_7$} as the ultimate food set, neither formaldehyde (H_2CO) nor acetate (CH_3COOH) is a viable seed (Table S6), yet together they can seed a viable tier-1 system containing 301 species and 736 reactions (Table S11). The interdependence between members of a seed set is conceptually linked to the fact that for the life as we know it, multiple chemical species need to be "seeded" together to allow the conversion from abiotic food to more cells – you cannot trigger the formation of a cell by simply seeding DNAs or proteins because the molecules forming a cell are interdependent upon each other.

It is also possible that a single seeding chemical can induce a viable SDAS that does not produce the seeding chemical itself. In an open environment, such chemicals, which we will call "pseudo-seeds," can trigger a SDAS but would be expected to disappear over time. In the abiotic reaction database, for example, when {H_2, CH_4, NO, FeS_2, visible light} is the ultimate food set, C_2H_3 is a pseudo-seed because it can induce a viable seed, OH, but the tier-1 SDAS it triggers cannot produce C_2H_3. Similarly, when {H_2O, CO_2, NH_3, H_2S, H_2SO_4, H_2SO_3, HSO_3^-, H_3PO_4, $H_4P_2O_7$} is the ultimate food set for the biotic reaction network, ATP is a pseudo-seed because it can induce a viable seed, pyruvate, but the tier-1 SDAS cannot produce ATP. The existence of pseudo-seeds is worth noting because it means that key transitions during the origin of life might have been triggered by chemicals that are no longer produced by biochemical systems, potentially confounding historical inference.

These results show that for both the abiotic and biotic reaction networks, seed-dependent autocatalytic systems exist. This is significant because a local chemical ecosystem would tend to be permanently altered by a rare seeding event which implies that the ecosystem "remembers" a seeding event. This perspective leads us to suggest that SDASs may be the earliest and simplest general mechanism of molecular memory, a prerequisite for evolution.

2.4 Higher-tier systems

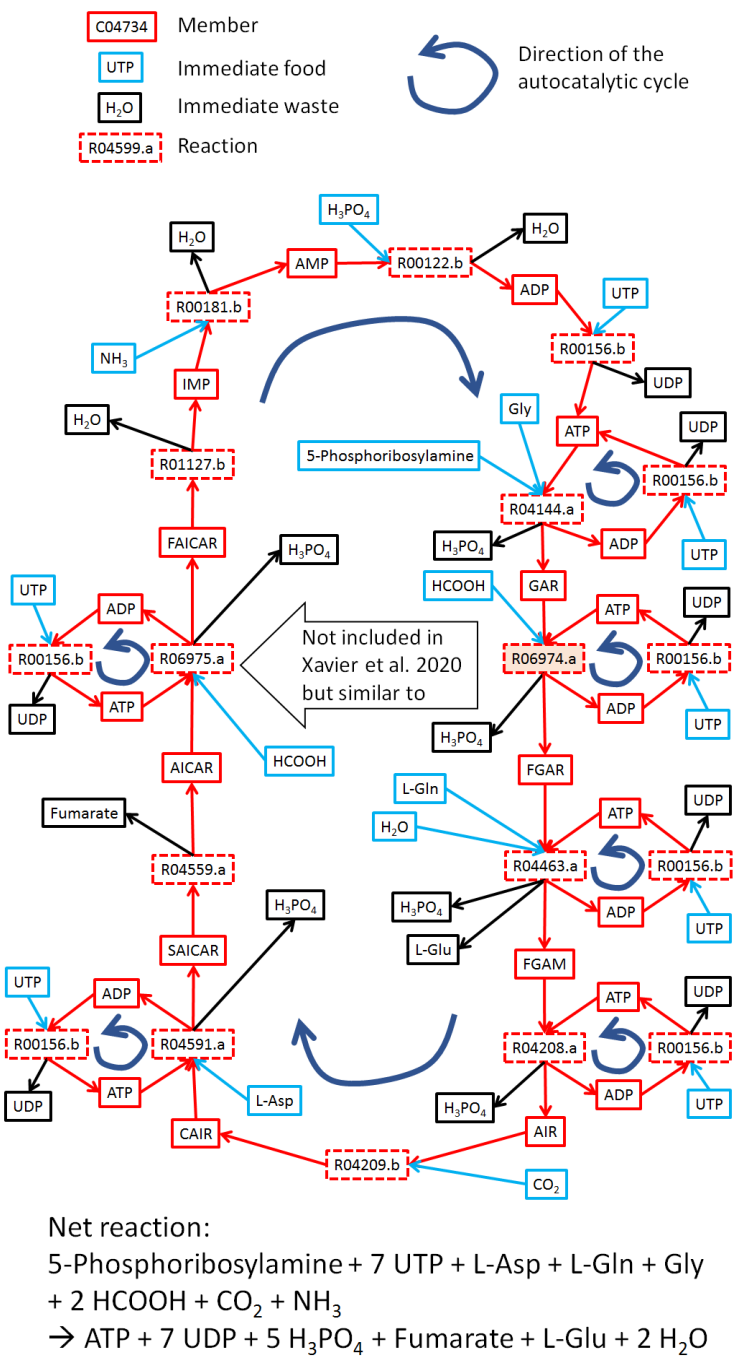
An important feature of biological systems is that some events must happen in a specific temporal order. For example, in primary succession, lichens dominate the environment before grasses and trees, and herbivores can survive only when plant populations are large enough. Such necessary ordering arises when later stages require some conditions that can be provided by the earlier stage.

Such hierarchical structuring is also manifested by reaction networks. Once a tier-1 SDAS is established, all members of the tier-1 system, together with the tier-0 members, are now available to "feed" additional higher-tier SDASs. The same procedures used to detect tier-1 systems can be used to look for viable systems at higher trophic levels. If such additional tiers exist then there is a natural ordering: tier-1 SDASs feed on tier-0 systems, and tier-2 SDASs feeds on the tier-0 and tier-1 systems, etc.

For the abiotic reaction database, once the tier-1 SDAS has been induced by any of its potential seeds, there exists a 13-member clique that induces a 14-entity, 35-reaction tier-2 system containing an autocatalytic core (Table S12, Table S13). This tier-2 SDAS includes the formose reaction, which feeds on H_2CO generated by the tier-1 SDAS. We did not find any viable tier-3 SDAS that can be induced by a single seed.

For the biotic reaction database, we chose to start from the tier-1 system induced by the 267-member clique, for illustrative purposes. We found a 6-member clique that seeds a 56-species, 180-reaction viable tier-2 SDAS (Table S14, Table S15). Not surprisingly, 5 out of the 6 clique members contain pyrimidine moieties, consistent with the fact that the additional reactions and chemicals in this tier-2 system overlap significantly with those activated by the 34-member clique that could trigger the alternative (pyrimidine-nucleoside-containing) tier-1 system.

Similar to the analysis on the abiotic reaction database, we did not find any single seed that can induce a viable tier-3 SDAS. However, we did notice that by adding just one, very plausible, reversible reaction to the reaction database (KEGG: R06974), adenine and 2-aminomuconate semialdehyde can, together, seed a viable tier-3 SDAS, which has 1057 chemicals and 3198 reactions (Table S16). This tier-3 system is able to synthesize purines and their derivatives, including the hydrogen donor and cofactor NADH. The core process of purine synthesis is an autocatalytic cycle feeding on tier-0 and tier-1 chemical species to produce ATP by substrate-level phosphorylation (Fig. 4, Fig. S4). However, it should be noted that including R06974 in the network changes the hierarchical structure in that it allows a single seed, such as NAD^+, to simultaneously trigger all three tiers from the tier-0 system. This observation illustrates that network hierarchy is sensitive to network topology and order of seeding, and that a complex molecule may compactly store information carried by multiple simpler chemical species.



Net reaction:

5-Phosphoribosylamine + 7 UTP + L-Asp + L-Gln + Gly + 2 HCOOH + CO_2 + NH_3

→ ATP + 7 UDP + 5 H_3PO_4 + Fumarate + L-Glu + 2 H_2O

Fig. 4. An autocatalytic core synthesizing ATP within a tier-3 SDAS detected in the biotic database (when reaction KEGG:R06974 is included). This autocatalytic core requires prior establishment of a lower-tier system able to supply specific organic molecules such as 5-phosphoribosylamine and L-glutamine. Red solid boxes: members of the autocatalytic core. Cyan solid boxes: immediate food of the autocatalytic core. Black solid box: immediate waste of the autocatalytic core. Red dashed boxes: reactions. Cyan arrows: food consumption. Black arrows: waste production.

As we have seen, the activation of SDASs at different trophic levels expands the spectrum of food available for future SDASs, providing a mechanism for the complexification of a realized reaction network that receives a flux of simple food inputs. This is significant because it suggests ways in which complexity can accrue contingently, depending on seeding events, which might be very rare. For example, some reactions not included in the network might be possible in the current environment but so slow that there might be a long and uncertain duration before seeding occurs. Alternative, seeds could disperse into the site from some other environments that allows for different chemical reactions for reasons such as a different physical environment (e.g., higher temperature), a different food set (e.g., a surficial *versus* a submarine site), or different catalysts (e.g., different exposed minerals).

2.5 From predators/parasites to mutualists

In a biological ecosystem, species on higher trophic levels typically have direct negative effects on their prey or hosts at a lower trophic level. However, ecological interactions can be complicated by indirect effects. For example, wolves in Yellowstone National Park have a direct negative effect on elk, on whom they predate, but also an indirect positive effect on plants eaten by elk (e.g., willows) and animals that compete for those plants (e.g., beavers) (Ripple et al., 2001). Furthermore, it is generally accepted that the presence or absence of certain species can alter the productivity of an ecosystem (e.g., Smith et al., 1991). We previously suggested that, since autocatalytic cores behave rather like species within a biological ecosystem, we can define spatially localized sets of active autocatalytic cores as a "chemical ecosystem" (Peng et al., 2020). Here we seek to extend this view and ask whether autocatalytic processes at higher trophic levels can provide analogous beneficial feedbacks on chemical ecosystems to those seen in biological ecosystems. To explore this possibility, we examined the structure of the tier-3 SDAS, described above, on overall ecosystem function.

The increased diversity of chemicals, which is maintained once higher-tier SDAS are activated, increases the chances of finding a chemical that catalyzes reactions at lower tiers. Catalysts are molecules that usually use a series of intermediate steps whose collective effect is to lower kinetic barriers to produce the same products as an uncatalyzed reaction while regenerating the catalyst. To identify potential catalysts we can, thus, look for tier-3 chemicals that provide a new,

multi-step path for completing a pre-existing reaction. By this reasoning, the tier-3 chemical FAD could catalyze assimilation of phosphate moieties via the reaction $CH_3COOH + H_4P_2O_7 \rightarrow CH_3COOPO_3H_2 + H_3PO_4$ (Fig. 5). Assuming that the composite reaction rate of the FAD-mediated reaction mechanism is higher than the direct reaction, which seems plausible, FAD would catalytically increase the rate at which the entire ecosystem could acquire phosphate from the environment.

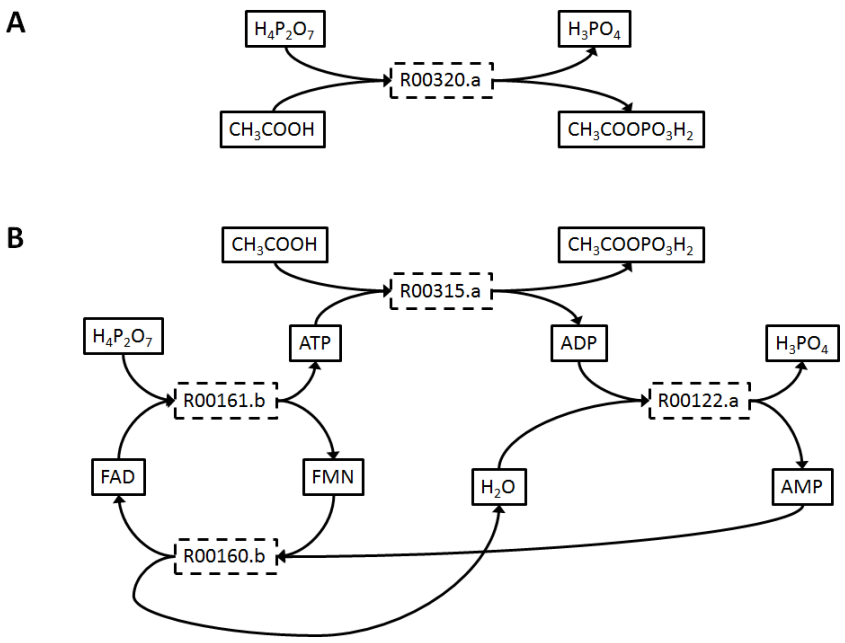


Fig. 5. Higher-tier chemicals may catalyze lower-tier reactions. An example in the biotic reaction network shows that FAD, a tier-3 chemical allows an alternative pathway for the phosphorylation of CH_3COOH, a tier-1 reaction, potentially increasing the rate at which the chemical ecosystem can assimilate phosphate. Solid box: chemical species. Dashed box: reaction.

It is also possible that a higher-tier chemical allows for new, more efficient ways for the reaction system to utilize food resources. For example, without the tier-3 chemical NAD^+ (or $NADP^+$), the ultimate food H_2S serves primarily as a source of thiol group rather than a hydrogen donor. However, due to the reaction, $NAD^+ + H_2S \leftrightarrow NADH + H^+ + S$, the presence of NAD^+ allows H_2S to serve as a terminal hydrogen donor. H_2S should provide a stronger driver of carbon fixation than the thermodynamically less favorable carbon fixation pathway within the tier-1 system (Fig. 3). This way of exploiting H_2S was not possible before the "discovery" of NAD^+ (or $NADP^+$), making this another example of feedback on overall ecosystem function. Furthermore, the presence of NADH also allows a new carbon-fixing autocatalytic cycle (Fig. 6) that is shorter, and thus potentially more efficient, than the carbon fixation processes supported by tiers 1 and 2.

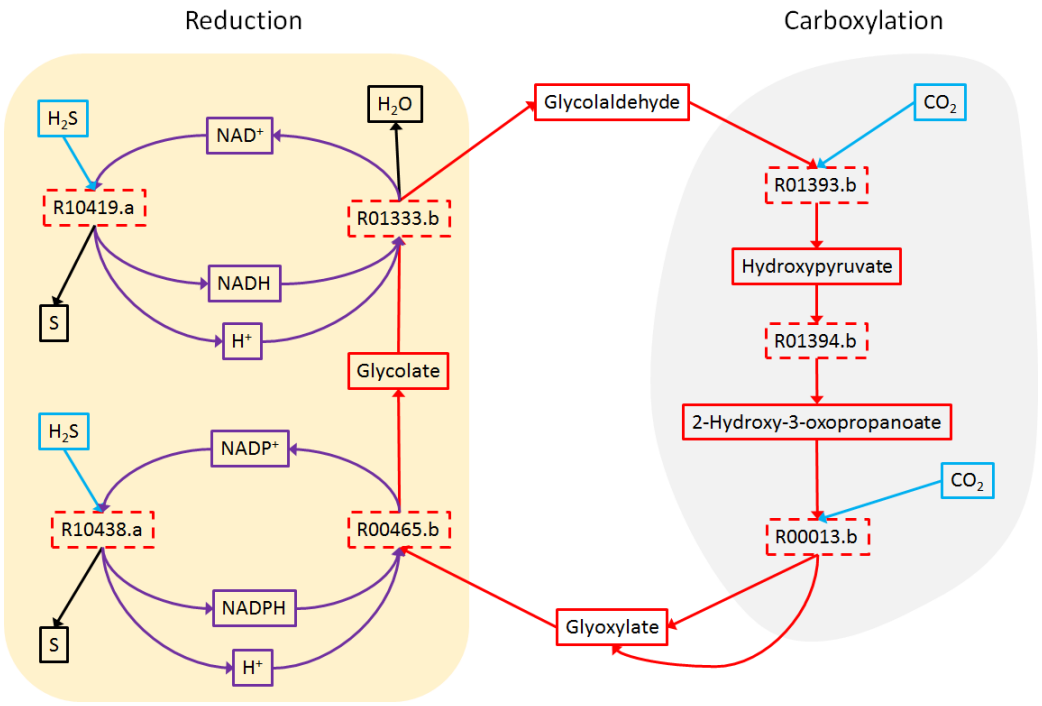


Fig. 6. A carbon-fixing autocatalytic cycle that becomes possible when NADH and NADPH are available. This autocatalytic cycle has two major modules: a reduction module leading from glyoxylate to glycolaldehyde, and a carboxylation module leading from glycolaldehyde to more glyoxylate. Because this cycle entails just seven reactions, it is likely to occur at a higher net rate than the pre-existing carbon fixation mechanisms. Red solid boxes: members of the autocatalytic cycle. Purple solid boxes: catalysts mediating the transfer of hydrogen from H_2S to organic molecules. Cyan solid boxes: immediate food of the autocatalytic cycle. Black solid box: immediate waste of the autocatalytic cycle. Red dashed boxes: reactions. Red arrows: syntheses of organic molecules. Purple arrows: catalysis. Cyan arrows: food consumption. Black arrows: waste production.

To give one final example of the new potential benefit that can arise with higher tiers, it is worth noting that the tier-3 SDAS includes some relatively long-chain amphiphiles such as hexadecanoic acid (Table S16), which might be able to form liposomes. Were this to be the case, the addition of tier-3 might open up the possibility of producing membranes that might alter ecosystem stability by providing protection against mechanical perturbation, by fostering spatial organization, and might enable long distance co-dispersal of interdependent seed chemicals.

These results support the view that, as higher-tier SDASs become activated, a chemical ecosystem can become more resistant to environmental perturbations and may also acquire novel capabilities. And if multiple systems consisting of SDASs coexisted during abiogenesis, it was likely that the systems with higher efficiency of utilizing environmental resources would gradually dominate and cause the entire chemical ecosystem to more efficiently exploit the environment, just as that the competition between species in a community during ecological succession tend to maximize the community's efficiency of utilization of resources over time (MacArthur, 1970, 1969). Thus, it seems plausible that chemical ecosystems may, complexify over time not only in the trivial sense of coming to support the persistence of a greater number of chemical species but also in the sense of acquiring and optimizing new, emergent ecosystem-level properties.

3. DISCUSSION

3.1 Chemical ecosystem theory as a general framework for understanding early steps in abiogenesis

In this paper, we extend our theory of Chemical Ecosystem Ecology (Peng et al., 2020) by proposing that a trophic structure exists within real chemical reaction networks based on the existence of SDASs and that this can lead to a pattern very reminiscent of ecological succession. We provide a mathematical formalism for SDASs and an algorithm for detecting them in networks composed of thousands of stoichiometrically-explicit chemical reactions. Our results confirm that chemical ecosystem theory has heuristic value in the analysis of chemical reaction systems and allows for the detection of common features in abiotic and biotic networks. In the process we were able to document multiple life-like features in the reaction networks analyzed,

including seed-induced autocatalysis, trophic hierarchy, and beneficial feedbacks on ecosystem function.

The model that emerges is that rare chemical reactions (perhaps enabled by chance fluctuations in local chemical concentrations) or import of materials from other environments can seed new autocatalytic systems allowing for the long-term exploration of chemical space in food-driven chemical ecosystems. Furthermore, since the number of chemicals sustained tends to increase as new SDASs are initiated, and since each new chemical has the potential to serve as food for yet more autocatalytic cores, the space of "adjacent possibles" (Kauffman and Gare, 2015) may also be expected to increase as new trophic tiers are added.

An important feature of the SDAS structure is that it provides a mechanism by which a chemical ecosystem can, in a sense, memorize what occurred in the past, a phenomenon that can be equated with heritability. This follows because individual seeding events can trigger a sustained change in the chemical composition of an ecosystem. Additionally, it is possible, though not demonstrated here, that in larger networks the order in which seed molecules are introduced could result in the exploration of different regions of chemical space.

Heritability is a prerequisite for evolution, showing that SDASs may provide a basis for an evolutionary process that could occur even prior to the use of genetic polymers. Such a model suggests that the origin of life might entail the following steps: (1) Planetary processes such as solar irradiation, local redox disequilibria driven by tectonics, or global redox disequilibria driven by the loss of hydrogen to space (Smith and Morowitz, 2016) generate a steady flux of food molecules/entities. (2) One or a few relatively simple chemical species are introduced by rare reactions or stochastic events, triggering low-tier SDASs, which convert the food to chemical species with higher diversity and complexity. (3) Each additional SDAS provides a larger pool of food permitting yet higher tiers to be added. (4) Some newly seeded SDASs, although consisting of energy-expensive molecules, cause the productivity of the network as a whole to increase, for example when complex molecules catalyze more efficient pathways for extracting energy from the food set, resulting in greater irreversibility. (5) Such chemical succession continues until the ecosystem is complex enough to be deserving of the label "life."

This origin-of-life model differs from many prevailing views in suggesting that networks of small organic molecules underwent something akin to biological evolution for some period

before giving rise to a polymer-based genetic system. Polymers are often energy-expensive and, due to the inherent combinatorics of polymerization cascades, individual polymer sequences would be difficult to maintain unless the chemical ecosystem were feeding efficiently on sustained food fluxes. But how could the transition to polymer chemistry arise?

The structure of the 3-tier system that we identified within the biotic network, provides some hints as to key steps in the conversion to polymer-regulated metabolism. Whereas tier-1 and tier-2 SDASs are mostly composed of quite small organic molecules, tier-3 includes some larger and more complex molecules, such as NADH and FAD, which can be seen as short heterogeneous polymers. Moreover, these short polymers show potential catalytic feedbacks on tiers 0-2. This arrangement is, abstractly, similar to the relationship between genetic/catalytic polymers (e.g., DNAs, RNAs, proteins) and metabolism in cellular life (Fig. 7). This illustrates that the catalytic feedback on metabolism achieved by biological cofactors might also have been at play when genetic and/or catalytic polymers, such as nucleic acids and peptides, first arose. However, because functional and non-functional polymers are composed of a similar alphabet of monomers, this pattern of network arrangement alone does not solve the combinatorial problem of producing sufficient functional polymers without most metabolic flux being channeled into the formation of non-functional polymers. Solving the combinatorics of polymerization remains an important focus of future work.

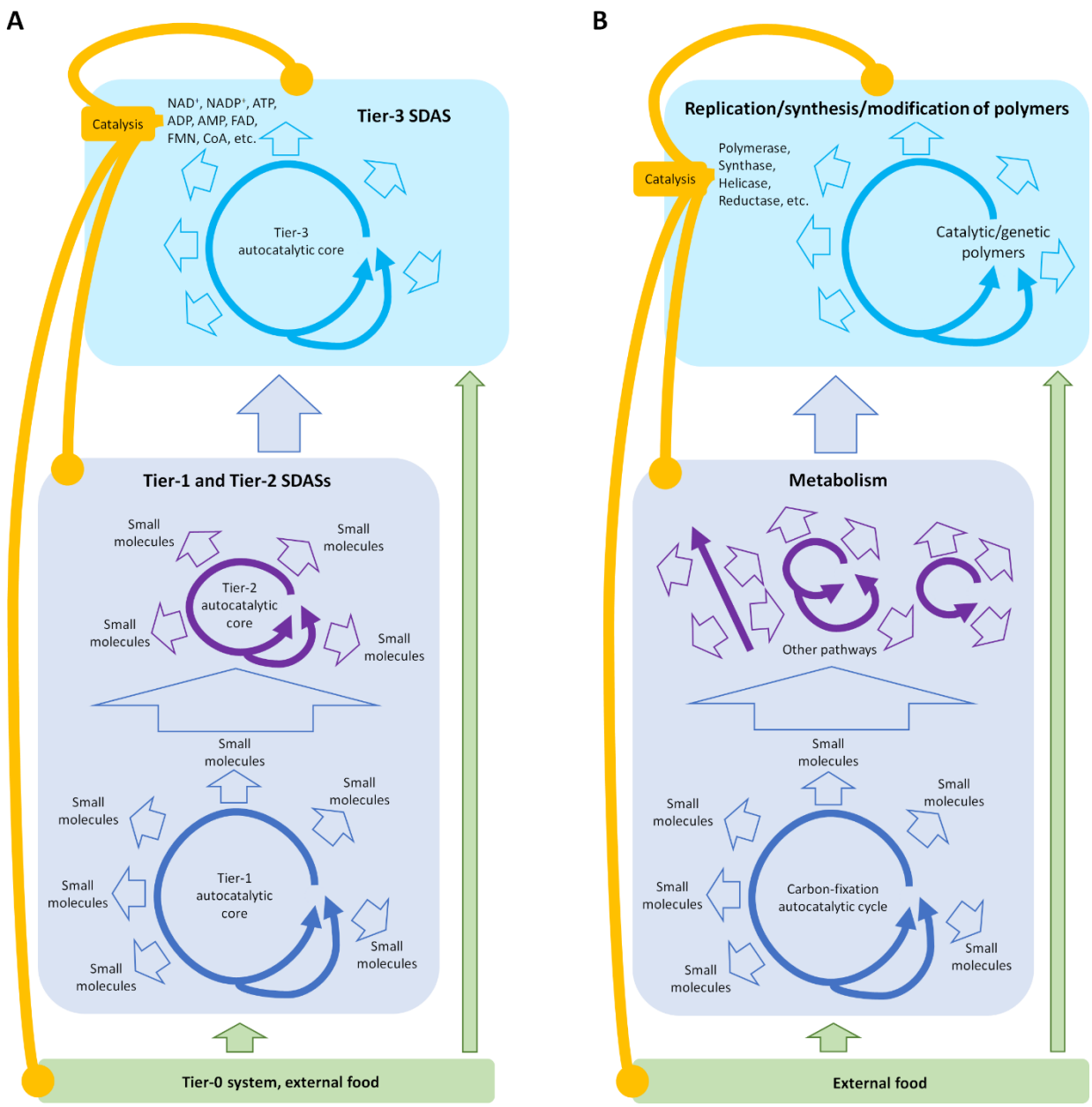


Fig. 7. The interactions between SDASs discovered in the biotic network are topologically similar to the interactions between metabolism, enzymes, and genetic polymers. (A) The summary of the interactions between SDASs on three trophic levels we found in the biotic reaction database. **(B)** A simplified model of the interactions between carbon fixation metabolism, enzymes, and genetic materials in extant life.

3.2 Limitations and future work

We have shown that the SDAS-based mechanism can be applied to databases of realistic abiotic and biotic reactions. Although only a few SDASs were found, it is possible that more SDASs would be found if a different ultimate food set were used and/or a different collection of reactions are allowed to occur. Additionally, both reaction systems are small fragments of a much larger network of potential reactions.

The biotic network has additional limitations, since almost all reactions are catalyzed by evolved enzymes. This implies that only a subset of the reactions in the network would occur at an appreciable rate in the absence of enzymes, which might be taken to imply that the resulting network might be over-connected. Conversely, one might argue that, in the absence of high-efficiency catalysts that direct metabolic flux to a modest number of biochemicals, prebiotic reaction networks might contain many more potential reactions all of which have rate constants within a few orders of magnitude of one another. Under the latter argument, the biotic reaction network would be seen as being under-connected. Thus, we really do not know how representative either the biotic or abiotic databases are of relevant prebiotic chemical networks.

It is non-trivial to obtain such realistic prebiotic reaction networks from empirical data, since only a limited number of reactions can be investigated experimentally, and those that are studied are likely to be sampled non-randomly, for example with a focus on reactions relevant to biology, combustion, or organic synthesis. The best hope is probably to infer reaction networks from *in silico* application of empirically-grounded reaction rules, as done recently by Wołos et al., 2020. However, until such more complete networks are analyzed more closely, it will remain unclear how closely the topology of the network of reactions available to prebiotic chemical ecosystems resembles the two networks studied here. Nonetheless, since the results from the biotic and abiotic networks are qualitatively similar despite differing in size and reaction density, we suspect that our overall conclusions will remain robust.

The approach developed here is based entirely on network topology, ignoring both thermodynamics and kinetics considerations. As a result, detected SDASs are not necessarily feasible because the autocatalytic cores upon which they depend might be non-viable in any realistic environment. Indeed, careful analysis of the carbon fixation reactions within the biotic tier-1 SDAS shows that it depends on a module consisting of two thermodynamically

unfavorable reactions, namely water reducing formaldehyde to methanol ($H_2CO + 2\ H_2O \rightarrow CH_3OH + H_2O_2$) and methanol and formic acid undergoing a reverse Cannizzaro reaction to generate two molecules of formaldehyde ($CH_3OH + HCOOH \rightarrow 2\ H_2CO + H_2O$) (Fig. 3). This fact does not fully undermine the conclusions, because the autocatalytic core can still hold if this module is replaced by a prebiotic process reducing formic acid to formaldehyde and there are several other prebiotically feasible pathways for producing formaldehyde (Cleaves, 2008). However, this case does illustrate that a full development of chemical ecosystem ecology should ultimately consider both thermodynamics and kinetics.

An additional weakness of the SDAS framework is that it does not directly explain the source of seeds. Thus, while the model can identify autocatalytic modules that would be triggered if a particular molecule or a set of particular molecules magically appeared, it says nothing about the likelihood of seeding. This is an oversimplification. We know that a seed molecule can only arise by a chemical reaction happening somewhere, so shouldn't the reactions that generate the seed (prior to SDAS activation) be included in the reaction network? One solution would be to include kinetics and use models, such as the Gillespie algorithm (Gillespie, 2007), that can translate low reaction rates into discrete stochastic events. Another approach might be to utilize two or more distinct networks of allowed reactions, each interpreted as the set of feasible reactions in a particular environmental context, and then allow rare transfer of materials from one environment into another. Finally, the ultimately preferable approach would combine explicit kinetics and an explicit spatial context so that improbable local reactions and rare dispersal events can both play a role in chemical ecosystem dynamics.

3.3 Prospects for experimental validation

Our model suggests that seed-dependent autocatalysis might be a common feature of driven, "messy" chemical reaction systems. That is to say, we should expect that providing a set of small molecules as food to a continuously-stirred tank reactor (CSTR), or its poor-man's equivalent, a serial batch-transfer-with-dilution experiment (Colón-Santos et al., 2019; Surman et al., 2019; Vincent et al., 2019), might lead to different quasistable chemical compositions depending on the addition of different molecular seeds. Can the network analyses conducted be used to guide the design of experiments looking for seed-dependent chemical memory?

The network-topology-based analysis provided here yield some useful information. It suggests, for example, that providing a flux of certain inorganic (or small organic) food to a CSTR might result in a transient steady state, which might then transition irreversibly to a different state through the transient addition of medium-complexity molecules (perhaps up to 300 Daltons), and might then transition to a third state after seeding with high-complexity molecules like NADH. However, given the incompleteness of the network used, and its dependence on enzymatic reactions, it seems premature to confidently predict which food and seed chemicals would show such a pattern. Nonetheless, we hope that experimentalists will either design experiments to test the SDAS hypothesis based on chemical intuition, or that more thorough *in silico* analysis of networks with realistic kinetic parameters can ultimately be used to guide more targeted experiments. Such work would be extremely valuable by empirically grounding chemical ecosystem ecology and providing a basis for systematically exploring the gradual complexification of food-driven chemical networks as might occur during the origin of life.

MATERIALS AND METHODS

Preprocessing databases of reactions

Abiotic reaction database

The reaction network assembled by Adam et al., 2021 includes the following categories: free radical reactions, mineral geochemical reactions, amino acid production, chloride radical and polar reactions, nitrile radical and polar reactions, RNA nucleotide assembly, nuclear decay, and physicochemical reactions. We processed this database by the following steps.

First, we excluded the nuclear decay reactions because we did not plan to put radioactive atoms into the ultimate food set.

Second, with kind help from Dr. Zachary R. Adam and Dr. Albert C. Fahrenbach, we deleted duplicate reactions, added a few new reactions that were not in the original database, balanced some reaction equations, and excluded the reactions without clear stoichiometry. This is because our method requires stoichiometry of reactions.

Third, we added the formose reaction into the database. According to Breslow's mechanism (Breslow, 1959), the formose reaction is driven by aldol and retro-aldol reactions and aldose-ketose isomerization. In combination these reactions allow low-carbon-number monosaccharides to generate high-carbon-number monosaccharides. Therefore, we added reversible aldol reactions and reversible aldose-ketose isomerization among formaldehyde, glycolaldehyde, and monosaccharides with no more than 8 carbon atoms. Optical isomers were not distinguished from each other. Formaldehyde dimerization was not added because it is very slow and its reaction mechanism is unclear but surely neither aldol/retro-aldol reaction nor aldose-ketose isomerization.

Fourth, every reaction labeled reversible was split into two unidirectional reactions.

Biotic reaction database

We processed the reaction database curated by Xavier et al., 2020 to obtain the biotic reaction database by the following steps.

First, we removed all reactions involving chemical species that do not have specific molecular mass, such as reduced ferredoxin (KEGG: C00138), acyl-carrier protein (KEGG: C00229), starch (KEGG: C00369), long-chain aldehyde (KEGG: C00609), and "Glycans" because they sometimes result in "fake" stoichiometric relationships. For example, the reaction: starch + H_2O ↔ dextrin + starch (KEGG: R02108) would make starch an infinite source of starch as long as H_2O is provided.

Second, we added some obviously spontaneous reactions that were missing, such as $H_2O \leftrightarrow H^+ + OH^-$ and $H_2CO_3 \leftrightarrow H^+ + HCO_3^-$.

Third, as all reactions in the KEGG biochemical reaction database are labeled reversible, every reaction was split into two unidirectional reactions. The reaction following the forward direction specified in the KEGG database has a suffix ".a" to its entry, and that of the reverse direction has a suffix ".b".

Fourth, the reactions that are labeled as multi-step were removed because each step is already a reaction in the database. Although keeping these multi-step reactions may not have big impact on the detection of SDAS existence, decreasing the number of reactions in the stoichiometric matrix should help accelerate the computation.

Network expansion

The set $\boldsymbol{R} = \{r_1, r_2, \ldots, r_i, \ldots, r_n\}$ is a set of multiple reactions r_i's that are allowed. Each r_i specifies reactants and products, and the union of all reactants and products across all r_i's is the maximum set of chemical species $\boldsymbol{S} = \{k_1, k_2, \ldots, k_j, \ldots, k_m\}$. We define an operation called full network expansion, $\Xi(\boldsymbol{S_O}, \boldsymbol{R}) = (\boldsymbol{S_E}, \boldsymbol{R_E})$, where $\boldsymbol{S_O}$ the subset of $\boldsymbol{S}$ where the expansion starts, $\boldsymbol{S_E}$ the set of chemical species resulting from the expansion, and $\boldsymbol{R_E}$ the expanded set of reactions resulting from the expansion. The expansion is conducted as follows:

(i) Let $\boldsymbol{R_E} = \emptyset$; define a temporary set of reactions $\boldsymbol{R'} = \boldsymbol{R}$; let $\boldsymbol{S_E} = \boldsymbol{S_O}$.

(ii) Define a temporary set of chemical species $\boldsymbol{S'} = \emptyset$.

(iii) For a reaction r_i in $\boldsymbol{R'}$, check if the reactants required by r_i are all present in $\boldsymbol{S_E}$; if so, move r_i from $\boldsymbol{R'}$ to $\boldsymbol{R_E}$, and scan through the products of r_i to add the chemical species that are not in $\boldsymbol{S_E}$

to $\boldsymbol{S'}$. Do this for all reactions in $\boldsymbol{R'}$. Then add all chemical species in $\boldsymbol{S'}$ to $\boldsymbol{S_E}$. If during this step, no reaction in $\boldsymbol{R'}$ is moved, then the expansion is finished; otherwise, proceed to (ii).

Detecting Seed-Dependent Autocatalytic Systems (SDASs) by linear programming

Let us assume that a (p-1)×(q-1) stoichiometric matrix, where each row represents a chemical species and each column represents a unidirectional reaction, results from a full expansion within the set of allowed reactions $\boldsymbol{R}$. The row labels of this (p-1)×(q-1) stoichiometric matrix form a chemical species set $\boldsymbol{S_F} = \{k_1, k_2, \dots, k_{p\text{-}1}\}$, which serves as the external food. Now we select a non-empty set of non-food chemical species $\boldsymbol{S_P} = \{k_p, k_{p+1}, \dots, k_{p+h}\}$ to serve as the seed set. Then we conduct a full expansion $(\boldsymbol{S_{FP}}, \boldsymbol{R_{FP}}) = \Xi(\boldsymbol{S_F} \cup \boldsymbol{S_P}, \boldsymbol{R})$, generating $\boldsymbol{S_{FP}} = \{k_1, k_2, \dots, k_m\}$ and $\boldsymbol{R_{FP}} = \{r_1, r_2, \dots, r_n\}$.

For some of the reactions, the same chemical species may appear on both sides of the reaction with the same stoichiometric coefficient, meaning that such a chemical species k_i can be viewed as a catalyst for the relevant reaction r_j. Even though the stoichiometric change of k_i is zero after r_j occurs, k_i is necessary for r_j to occur. Therefore, on the reactant side of r_j, we may add a positive value to the stoichiometric coefficient of k_i (i.e., subtract a small positive value from the entry representing k_i and r_j in the stoichiometric matrix) when performing the linear programming; although this manipulation would make the reaction equation unbalanced, if a SDAS does exist, the effects of such unbalance should be compensated by the fact that the SDAS can synthesize excessive k_i from the external food.

A SDAS feeding on $\boldsymbol{S_F}$ exists if there is a vector of non-negative elements $\boldsymbol{x} = (x_q, x_{q+1}, \dots, x_n)$ such that

$$\sum_{j=q}^{n} x_j s_{ij} > 0 \ \left(x_j \geq 0\right) \quad \forall i \in [p, m] \tag{1}$$

where s_{ij} is the entry at the ith row and jth column of the stoichiometric matrix.

We set the linear programming problem as to find

$$\max_{x_q,\ \ldots,\ x_n} \sum_{j=q}^{n} x_j s_{pj} \tag{2}$$

which is constrained by

$$\sum_{j=q}^{n} x_j s_{ij} > 0 \ (x_j \geq 0) \quad \forall i \in [p, m] \tag{1}$$

The linear programming tool provided by SciPy v1.6.2 (https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.linprog.html) is used to solve the problem. If this linear program problem is found to be unbounded, then we know that there must exists some $\boldsymbol{x}$ satisfying the conditions of a SDAS. We used the "highs" method (Huangfu and Hall, 2018) to confirm the existence of SDASs and the "revised simplex" method (Bertsimas and Tsitsiklis, 1997) to enumerate all reactions of a SDAS. Once a SDAS was confirmed to exist, we ran the integer programming process that is described in the next section for finding the autocatalytic cores subject to further specific constraints within the SDAS.

Detecting minimum autocatalytic cores by integer programming

We sought to find small autocatalytic cores within each SDAS because these are easier to visualize and could potentially guide future experimental studies. In this section, we describe a method based on integer linear programming to enumerate small-cardinality SDASs contained in a given SDAS.

For the given SDAS, we know that

$$\sum_{j=q}^{n} x_j s_{ij} > 0 \ (x_j \geq 0) \quad \forall i \in [p, m] \tag{1}$$

To find smaller systems, we seek a set of columns $\boldsymbol{T} \subset [q, n]$ such that if $j \in \boldsymbol{T}$ and $\exists i \in [p, m]$ which makes $s_{ij} \neq 0$, then $\exists \varepsilon_j \in \mathbb{R}_+^{n-q+1}$ such that $\sum_{j=q}^{n} \varepsilon_j s_{ij} \geq 1$. Of course, the set $\boldsymbol{T}$ should have $|\boldsymbol{T}| \geq 1$. Finding such a set $\boldsymbol{T}$ can be accomplished in a systematic manner by seeking solutions to a linear-inequality system wherein some of the variables are required to take integer values.

In the formulation, we use binary variables z_j that take the value 1 if and only if column $j \in [q, n]$ is in the set $\boldsymbol{T}$, and we will minimize $\sum_{j=q}^{n} z_j$ to minimize the cardinality of the selected set. For the selected set to be a SDAS, we must enforce the logic that

$$\sum_{j=q}^{n} z_j |s_{ij}| > 0 \Rightarrow \sum_{j=q}^{n} \varepsilon_j s_{ij} > 0 \quad \forall i \in [p, m] \tag{3}$$

This is done by introducing additional binary variables y_i ($i \in [p, m]$) indicating if species k_i is included in the SDAS represented by $\boldsymbol{z} = (z_q, z_{q+1}, \ldots, z_n)$. Then, the following two sets of linear inequalities accomplish the logic in (3)

$$\sum_{j=q}^{n} z_j |s_{ij}| \leq M_i y_i \quad \forall i \in [p, m] \tag{4}$$

$$\sum_{j=q}^{n} \varepsilon_j s_{ij} \geq 1 - M_i(1 - y_i) \quad \forall i \in [p, m] \tag{5}$$

for a suitably large value of M_i. If β_j is an upper bound on the level of reaction r_j in a SDAS, then $M_i = \sum_{j=q}^{n} \beta_j s_{ij}$ suffices. We also must enforce that if column j is not selected for the SDAS (i.e., $z_j = 0$), then its level of reaction ε_j must also be zero, which is done with the algebraic constraints $\varepsilon_j \leq \beta_j z_j$.

This gives a full integer programming formulation for finding a minimum-cardinality SDAS among the reactions $\{r_q, r_{q+1}, \ldots, r_n\}$. We set the integer programming problem as to find

$$\min_{z_q, \ \ldots, \ z_n} \sum_{j=q}^{n} z_j \tag{6}$$

which is constrained by

$$\begin{cases} \sum_{j=q}^{n} z_j \geq 1 \\ \sum_{j=q}^{n} z_j |s_{ij}| \leq M_i y_i \quad \forall i \in [p,m] \\ \sum_{j=q}^{n} \varepsilon_j s_{ij} \geq 1 - M_i(1-y_i) \quad \forall i \in [p,m] \\ \varepsilon_j \geq 0 \quad \forall j \in [q,n] \\ \varepsilon_j \leq \beta_j z_j \quad \forall j \in [q,n] \\ z_j \in \{0,1\} \quad \forall j \in [q,n] \\ y_i \in \{0,1\} \quad \forall i \in [p,m] \end{cases} \tag{7}$$

This formulation can be solved computationally in a matter of seconds using a state-of-the-art integer programming software, such as Gurobi, for the reaction databases used in this study. The binary solution vector $\boldsymbol{z}$ indicates the reactions $\boldsymbol{T}$ in a minimum cardinality SDAS chosen from reactions $\{r_q, r_{q+1}, \ldots, r_n\}$. To exclude this particular SDAS and seek additional systems, we can add the constraint

$$\sum_{j \in \boldsymbol{T}} z_j \leq |\boldsymbol{T}| - 1 \tag{8}$$

to the integer programming formulation. Our implementation can iteratively add such constraints to enumerate multiple SDASs.

Identifying cliques

Let us assume that $\boldsymbol{S_F}$ is a set of chemical species resulting from a full expansion within the set of allowed reactions $\boldsymbol{R}$ from a set of ultimate food $\boldsymbol{S_{UF}}$, then $\boldsymbol{S_F}$ is the set of external food. Two non-empty seed sets of non-food chemical species $\boldsymbol{S_{P1}}$ and $\boldsymbol{S_{P2}}$ are said to be in the same clique if (a) $\Xi(\boldsymbol{S_F} \cup \boldsymbol{S_{P1}}, \boldsymbol{R}) = \Xi(\boldsymbol{S_F} \cup \boldsymbol{S_{P2}}, \boldsymbol{R})$, and (b) for any proper subset $\boldsymbol{S'_{P1}}$ of $\boldsymbol{S_{P1}}$ and any proper subset $\boldsymbol{S'_{P2}}$ of $\boldsymbol{S_{P2}}$, $\Xi(\boldsymbol{S_F} \cup \boldsymbol{S'_{P1}}, \boldsymbol{R}) \neq \Xi(\boldsymbol{S_F} \cup \boldsymbol{S_{P1}}, \boldsymbol{R})$ and $\Xi(\boldsymbol{S_F} \cup \boldsymbol{S'_{P2}}, \boldsymbol{R}) \neq \Xi(\boldsymbol{S_F} \cup \boldsymbol{S_{P2}}, \boldsymbol{R})$.

In this paper, when talking about cliques, we only focus on cases where seeds are individual chemical species. Nonetheless, the principle could be expanded to potential seed sets comprising more than one chemical species.

The reasons for adding the reaction KEGG: R06974

In the main text, we stated that it is reasonable to add the reaction KEGG: R06974 into the biotic reaction database. This is because a closer look at this reaction revealed that it is actually very similar to the reaction R06975 (Fig. S5): both reactions use HCOOH as the carbon donor to add a -CHO to -NH2 and form a -NH-CHO with ATP hydrolysis providing energy for the reaction. However, R06975 is in the network curated by Xavier et al., 2020 while R06974 is not, presumably because the annotations of R06974 in the KEGG database are not as detailed as those of R06975, and thus R06974 was filtered out due to some strict criterion.

FUNDING

This project is supported by NASA-NSF CESPOoL (Chemical Ecosystem Selection Paradigm for the Origins of Life) Ideas Lab grant (NASA-80NSSC17K0296) and University of Wisconsin Vice-Chancellor for Research and Graduate Education.

ACKNOWLEDGEMENTS

Funding was provided by NASA grant (80NSSC17K0296) and University of Wisconsin Vice-Chancellor for Research and Graduate Education. We thank Dr. Zachary R. Adam and Dr. Albert C. Fahrenbach for their kind help with curating the abiotic reaction database and the following for useful discussions: Alyssa Adams, Stephanie Colón-Santos, Emily Dolson, Praful Gagrani, Chris Kempes, Juan Perez Mercader, Alex Plum, Daniel Segrè, D. Eric Smith, and Lena Vincent.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

Adam, Z.R., Fahrenbach, A.C., Jacobson, S.M., Kacar, B., Zubarev, D.Y., 2021. Radiolysis generates a complex organosynthetic chemical network. Scientific Reports 11, 1743. https://doi.org/10.1038/s41598-021-81293-6

Bertsimas, D., Tsitsiklis, J.N., 1997. Introduction to linear optimization, Athena Scientific series in optimization and neural computation. Athena Scientific, Belmont, Mass.

Blokhuis, A., Lacoste, D., Nghe, P., 2020. Universal motifs and the diversity of autocatalytic systems. PNAS 117, 25230–25236. https://doi.org/10.1073/pnas.2013527117

Breslow, R., 1959. On the mechanism of the formose reaction. Tetrahedron Letters 1, 22–26. https://doi.org/10.1016/S0040-4039(01)99487-0

Cleaves, H.J., 2008. The prebiotic geochemistry of formaldehyde. Precambrian Research 164, 111–118. https://doi.org/10.1016/j.precamres.2008.04.002

Cleaves, H.J. (Jim), 2011. Formose Reaction, in: Gargaud, M., Amils, R., Quintanilla, J.C., Cleaves, H.J. (Jim), Irvine, W.M., Pinti, D.L., Viso, M. (Eds.), Encyclopedia of Astrobiology. Springer, Berlin, Heidelberg, pp. 600–605. https://doi.org/10.1007/978-3-642-11274-4_587

Colón-Santos, S., Cooper, G.J.T., Cronin, L., 2019. Taming the Combinatorial Explosion of the Formose Reaction via Recursion within Mineral Environments. ChemSystemsChem 1, e1900033. https://doi.org/10.1002/syst.201900033

Damer, B., Deamer, D., 2015. Coupled Phases and Combinatorial Selection in Fluctuating Hydrothermal Pools: A Scenario to Guide Experimental Approaches to the Origin of Cellular Life. Life 5, 872–887. https://doi.org/10.3390/life5010872

Donaldson, D.J., Tervahattu, H., Tuck, A.F., Vaida, V., 2004. Organic Aerosols and the Origin of Life: An Hypothesis. Orig Life Evol Biosph 34, 57–67. https://doi.org/10.1023/B:ORIG.0000009828.40846.b3

Gillespie, D.T., 2007. Stochastic Simulation of Chemical Kinetics. Annual Review of Physical Chemistry 58, 35–55. https://doi.org/10.1146/annurev.physchem.58.032806.104637

Haas, M., Lamour, S., Christ, S.B., Trapp, O., 2020. Mineral-mediated carbohydrate synthesis by mechanical forces in a primordial geochemical setting. Commun Chem 3, 1–6. https://doi.org/10.1038/s42004-020-00387-w

Huangfu, Q., Hall, J.A.J., 2018. Parallelizing the dual revised simplex method. Math. Prog. Comp. 10, 119–142. https://doi.org/10.1007/s12532-017-0130-5

Kanehisa, M., 2019. Toward understanding the origin and evolution of cellular organisms. Protein Science 28, 1947–1951. https://doi.org/10.1002/pro.3715

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., Tanabe, M., 2021. KEGG: integrating viruses and cellular organisms. Nucleic Acids Research 49, D545–D551. https://doi.org/10.1093/nar/gkaa970

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28, 27–30. https://doi.org/10.1093/nar/28.1.27

Kauffman, S.A., Gare, A., 2015. Beyond Descartes and Newton: Recovering life and humanity. Progress in Biophysics and Molecular Biology, Integral Biomathics: Life Sciences, Mathematics, and Phenomenological Philosophy 119, 219–244. https://doi.org/10.1016/j.pbiomolbio.2015.06.003

Ladyman, J., Lambert, J., Wiesner, K., 2013. What is a complex system? Euro Jnl Phil Sci 3, 33–67. https://doi.org/10.1007/s13194-012-0056-8

Lathe, R., 2004. Fast tidal cycling and the origin of life. Icarus 168, 18–22. https://doi.org/10.1016/j.icarus.2003.10.018

MacArthur, R., 1970. Species packing and competitive equilibrium for many species. Theoretical Population Biology 1, 1–11. https://doi.org/10.1016/0040-5809(70)90039-0

MacArthur, R., 1969. Species Packing, and What Competition Minimizes. PNAS 64, 1369–1371. https://doi.org/10.1073/pnas.64.4.1369

Marín-Yaseli, M.R., González-Toril, E., Mompeán, C., Ruiz-Bermejo, M., 2016. The Role of Aqueous Aerosols in the “Glyoxylate Scenario”: An Experimental Approach. Chemistry – A European Journal 22, 12785–12799. https://doi.org/10.1002/chem.201602195

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

Maruyama, S., Kurokawa, K., Ebisuzaki, T., Sawaki, Y., Suda, K., Santosh, M., 2019. Nine requirements for the origin of Earth’s life: Not at the hydrothermal vent, but in a nuclear geyser system. Geoscience Frontiers 10, 1337–1357. https://doi.org/10.1016/j.gsf.2018.09.011

Mitchell, M., Newman, M., 2001. Complex systems theory and evolution 5.

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

Ripple, W.J., Larsen, E.J., Renkin, R.A., Smith, D.W., 2001. Trophic cascades among wolves, elk and aspen on Yellowstone National Park's northern range. Biological Conservation 102, 227–234. https://doi.org/10.1016/S0006-3207(01)00107-0

Smith, E., Morowitz, H.J., 2016. The Origin and Nature of Life on Earth: The Emergence of the Fourth Geosphere. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781316348772

Smith, T.J., Boto, K.G., Frusher, S.D., Giddins, R.L., 1991. Keystone species and mangrove forest dynamics: the influence of burrowing by crabs on soil nutrient status and forest productivity. Estuarine, Coastal and Shelf Science 33, 419–432. https://doi.org/10.1016/0272-7714(91)90081-L

Sojo, V., Herschy, B., Whicher, A., Camprubí, E., Lane, N., 2016. The Origin of Life in Alkaline Hydrothermal Vents. Astrobiology 16, 181–197. https://doi.org/10.1089/ast.2015.1406

Steel, M., Xavier, J.C., Huson, D.H., 2020. The structure of autocatalytic networks, with application to early biochemistry. Journal of The Royal Society Interface 17, 20200488. https://doi.org/10.1098/rsif.2020.0488

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

Vincent, Berg, Krismer, Saghafi, Cosby, Sankari, Vetsigian, Ii, Baum, 2019. Chemical Ecosystem Selection on Mineral Surfaces Reveals Long-Term Dynamics Consistent with the Spontaneous Emergence of Mutual Catalysis. Life 9, 80. https://doi.org/10.3390/life9040080

Wächtershäuser, G., 1988. Before enzymes and templates: theory of surface metabolism. Microbiol Rev 52, 452–484.

Westall, F., Hickman-Lewis, K., Hinman, N., Gautret, P., Campbell, K. a., Bréhéret, J. g., Foucher, F., Hubert, A., Sorieul, S., Dass, A. v., Kee, T. p., Georgelin, T., Brack, A.,

2018. A Hydrothermal-Sedimentary Context for the Origin of Life. Astrobiology 18, 259–293. https://doi.org/10.1089/ast.2017.1680

Wołos, A., Roszak, R., Żądło-Dobrowolska, A., Beker, W., Mikulak-Klucznik, B., Spólnik, G., Dygas, M., Szymkuć, S., Grzybowski, B.A., 2020. Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. Science 369. https://doi.org/10.1126/science.aaw1955

Xavier, J.C., Hordijk, W., Kauffman, S., Steel, M., Martin, W.F., 2020. Autocatalytic chemical networks at the origin of metabolism. Proc. R. Soc. B. 287, 20192377. https://doi.org/10.1098/rspb.2019.2377