RESEARCH ARTICLE SUMMARY

PREBIOTIC CHEMISTRY

Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry

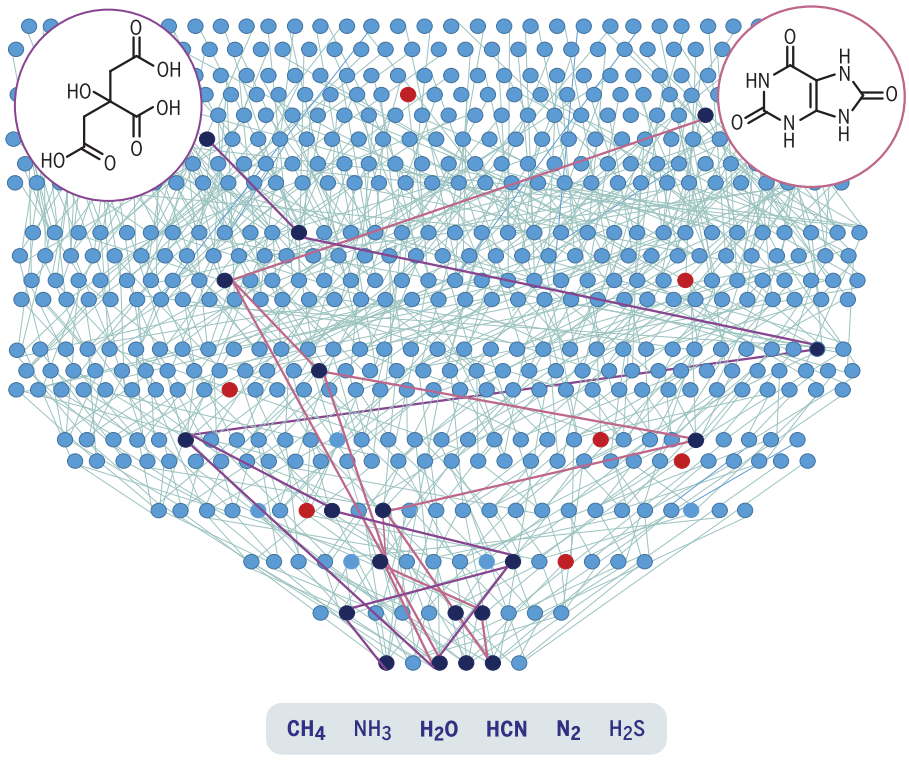
Agnieszka Wołos*, Rafał Roszak*, Anna Żądło-Dobrowolska*, Wiktor Beker, Barbara Mikulak-Klucznik, Grzegorz Spólnik, Mirosław Dygas, Sara Szymkuć†, Bartosz A. Grzybowski†

INTRODUCTION: Although hundreds of organic reactions have been validated under consensus prebiotic conditions, we still have only a fragmentary understanding of how these individual steps combined into complete synthetic pathways to generate life's building blocks, which other abiotic molecules might have also formed, how independent reactions gave rise to chemical systems, and how membranes encapsulating these systems came into being. Answering such questions requires consideration of very large numbers of possible synthetic pathways. Starting with even a few primordial substrates—e.g., H_2O, N_2, HCN, NH_3, CH_4, and H_2S—the number of prebiotically synthesizable molecules grows rapidly into the tens of thousands. Detailed analysis of this space and its synthetic connectivity may be beyond the cognition of individual chemists but can be performed by smart computer algorithms.

RATIONALE: We harnessed the power of computer-assisted organic synthesis to map the network of molecules that are synthesizable from basic prebiotic feedstocks. This was done by encoding currently known prebiotic reactions in a machine-readable format, augmenting these reaction transforms with information about incompatible groups and reaction conditions, and then applying them iteratively to a set of basic prebiotic substrates. The reaction network thus created was queried by algorithms to identify complete synthetic routes as well as those tracing reaction systems—notably, reaction cycles. All calculations were supported by a software application that is freely available to the scientific community.

RESULTS: We demonstrate that this network comprises more abiotic molecules than biotic molecules. The biotic compounds differ from the abiotic compounds in several ways: They are more hydrophilic, more thermodynamically stable, and more balanced in terms of the hydrogen bond donors and acceptors they contain and are synthesizable along routes with fewer changes of conditions. The network contains not only all known syntheses of biotic compounds but also previously unidentified routes, several of which (e.g., prebiotic syntheses of acetaldehyde and diglycine, as well as malic, fumaric, citric, and uric acids) we validated by experiment. We also demonstrate three notable forms of chemical emergence: (i) that the molecules created within the network can themselves enable new types of prebiotic reactions; (ii) that within just a few synthetic generations, simple chemical systems (including self-regenerating cycles) begin to emerge; and (iii) that the network contains prebiotic routes to surfactant species, thus outlining a path to biological compartmentalization. We support these conclusions with experimental results, establishing previously undescribed prebiotic reactions and entire reaction systems—notably, a self-regenerating cycle of iminodiacetic acid.

CONCLUSION: Computer-generated reaction networks are useful in identifying synthetic routes to prebiotically relevant targets and are indispensable for the discovery of prebiotic chemical systems that are otherwise challenging to discern. As our network continues to grow by means of crowd-sourcing of newly validated prebiotic reactions, it will allow continued simulation of chemical genesis, beginning with molecules as simple as water, ammonia, and methane and leading to increasingly complex targets, including those of current interest in the chemical and pharmaceutical industries. ■



Network of prebiotic chemistry. Computer simulation of plausible prebiotic reactions creates a network of molecules that are synthesizable from prebiotic feedstocks and establishes multiple unreported—but now experimentally validated—syntheses of prebiotic targets as well as self-regenerating cycles. In this schematic illustration, light blue nodes represent abiotic molecules, dark blue nodes represent molecules along newly discovered prebiotic syntheses of uric and citric acids, and red nodes represent other biotic molecules.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: saraszymkuc@gmail.com (S.S.); nanogrzybowski@gmail.com (B.A.G.)
Cite this article as A. Wołos *et al.*, *Science* **369**, eaaw1955 (2020). DOI: 10.1126/science.aaw1955

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.aaw1955

RESEARCH ARTICLE

PREBIOTIC CHEMISTRY

Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry

Agnieszka Wołos[1,2]*, Rafał Roszak[1,2]*, Anna Żądło-Dobrowolska[1]*, Wiktor Beker[1,2], Barbara Mikulak-Klucznik[1,2], Grzegorz Spólnik[1], Mirosław Dygas[1], Sara Szymkuć[1,2]†, Bartosz A. Grzybowski[1,2,3,4]†

The challenge of prebiotic chemistry is to trace the syntheses of life's key building blocks from a handful of primordial substrates. Here we report a forward-synthesis algorithm that generates a full network of prebiotic chemical reactions accessible from these substrates under generally accepted conditions. This network contains both reported and previously unidentified routes to biotic targets, as well as plausible syntheses of abiotic molecules. It also exhibits three forms of nontrivial chemical emergence, as the molecules within the network can act as catalysts of downstream reaction types; form functional chemical systems, including self-regenerating cycles; and produce surfactants relevant to primitive forms of biological compartmentalization. To support these claims, computer-predicted, prebiotic syntheses of several biotic molecules as well as a multistep, self-regenerative cycle of iminodiacetic acid were validated by experiment.

Research on the chemical origins of life (OL) is coming of age. The pioneering efforts of Miller (*1*), Oparin (*2*), Oró (*3*, *4*), and Orgel (*5*) by the 1960s; Eschenmoser (*6*) in the 1990s; and Sutherland (*7*, *8*), Carell (*9*), Moran (*10*), and others (*11*–*15*) in recent years have systematized the knowledge about reactions that can be performed under consensus prebiotic conditions, as well as the plausible synthetic routes leading to life's key molecules. On the other hand, we still have only a fragmentary understanding of whether and how other types of molecules formed on primitive Earth and how this entire prebiotic molecular space evolved into systems of chemical reactions (*12*, *16*) and compartments (*17*, *18*) housing them. Such analyses require consideration of very large numbers of putative reaction pathways but are finally becoming possible, owing to recent advances in the study of chemical reaction networks and computer-assisted organic synthesis (*19*, *20*). In this study, we use such large-scale in silico network analyses to map the space of molecules synthesizable from basic prebiotic feedstocks, quantifying the structure of this space as well as the abundances and thermodynamic properties of its members. We then demonstrate three notable forms of chemical emergence: (i) that the molecules created within the network can themselves enable new types of prebiotic reactions, including multicomponent transformations that lead to complex and useful organic scaffolds; (ii) that within just a few synthetic generations, simple chemical systems (such as self-regenerating cycles) begin to emerge; and (iii) that the network contains prebiotic routes to surfactant species (both peptide-based and long-chain carboxylic acids), thus outlining a path to biological compartmentalization. We support these results by experimental validation of previously unappreciated prebiotic syntheses (e.g., of acetaldehyde, diglycine, as well as malic, fumaric, citric, and uric acids) and entire reaction systems—notably, we demonstrate a self-regenerating cycle of iminodiacetic acid (IDA) that complements prebiotic autocatalysis on the basis of the formose cycle (*21*). The web application underlying our calculations is made freely available to the community (https://life.allchemy.net) in the hope that synthetic network analyses will become a useful addition to the toolkit of OL research by supporting accelerated discovery of prebiotic routes, including environmentally friendly syntheses of useful targets from basic feedstocks.

Allchemy's "Life" module uses 614 reaction rules ("transforms") involving C, O, N, S, and P elements, grouped within 72 broader reaction classes. Inclusion of these rules in our set is contingent on the existence of literature-described examples that document their execution under generally accepted prebiotic conditions [for reaction templates as well as conditions and literature references substantiating prebiotic relevance, see supplementary materials (SM) section S2]. All of these transforms generalize, are broader than the underlying literature precedents, and are coded to take into account the underlying reaction mechanisms, as described in our previous works (*19*, *22*). This approach has been recently validated experimentally [by successful execution of numerous computer-designed syntheses of medicinally relevant targets and natural products (*20*, *23*)] and is less prone to yield chemically problematic predictions than either machine rule extraction or ab initio methods (*24*) [for examples, see (*22*) and SM section S1.1]. Our transforms account for reaction byproducts and specify the scope of admissible substituents, structural motifs incompatible with a given reaction (some 400 potentially conflicting groups are considered for each reaction), typical conditions accepted in prebiotic chemistry, solvents, temperatures, and more. They do not consider stereochemistry [because homochirality probably appeared as a result of chemical evolution of racemic mixtures (*25*)] or reaction kinetics [because kinetic data are only sparingly reported in the studies of prebiotic chemistries (*26*)]. On the other hand, yields for each type of reaction are approximated on the basis of statistics collected from relevant publications and are categorized as trace (≤3%), low (>3% to ≤10%), moderate (>10% to <80%), and high (≥80%) (SM section S2).

The analyses described below are for H_2O, N_2, HCN, NH_3, CH_4, and H_2S substrates and the corresponding C-, O-, N-, and S-based transforms. These substrates were chosen because they are the starting points of many OL studies, are thought to be the components of Earth's early atmosphere (*27*), and are sufficient to build many common molecules of life. On the other hand, Allchemy-supported P-based chemistries were not included in this study because resultant searches generate large numbers of chemically redundant species and dilute the pool of chemically distinct scaffolds (e.g., numerous alcohols created via C, O, N, and S reactions serve as substrates for phosphorylation, leading to phosphate esters and rapidly increasing the size of the network; for further comments, see SM section S2.2).

The transforms were iteratively applied to the user-specified substrates. After each iteration, the newly created molecules were combined with the products of preceding iterations and with the original substrates, and the cycle was repeated until a user-defined limit of synthetic generations [typically up to seven (G7)] is reached (Fig. 1A). All calculations described in the following text were performed within Allchemy's Life module (available at https://life.allchemy.net; for login details and user manual, see SM section S1.2), which also allows for additional, user-specified constraints (e.g., only certain reaction types, temperature

[1]Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw, Poland. [2]Allchemy, Inc., Highland, IN, USA. [3]Center for Soft and Living Matter of Korea's Institute for Basic Science (IBS), Ulsan, South Korea. [4]Department of Chemistry, Ulsan National Institute of Science and Technology, Ulsan, South Korea.
*These authors contributed equally to this work.
†Corresponding author. Email: saraszymkuc@gmail.com (S.S.); nanogrzybowski@gmail.com (B.A.G.)

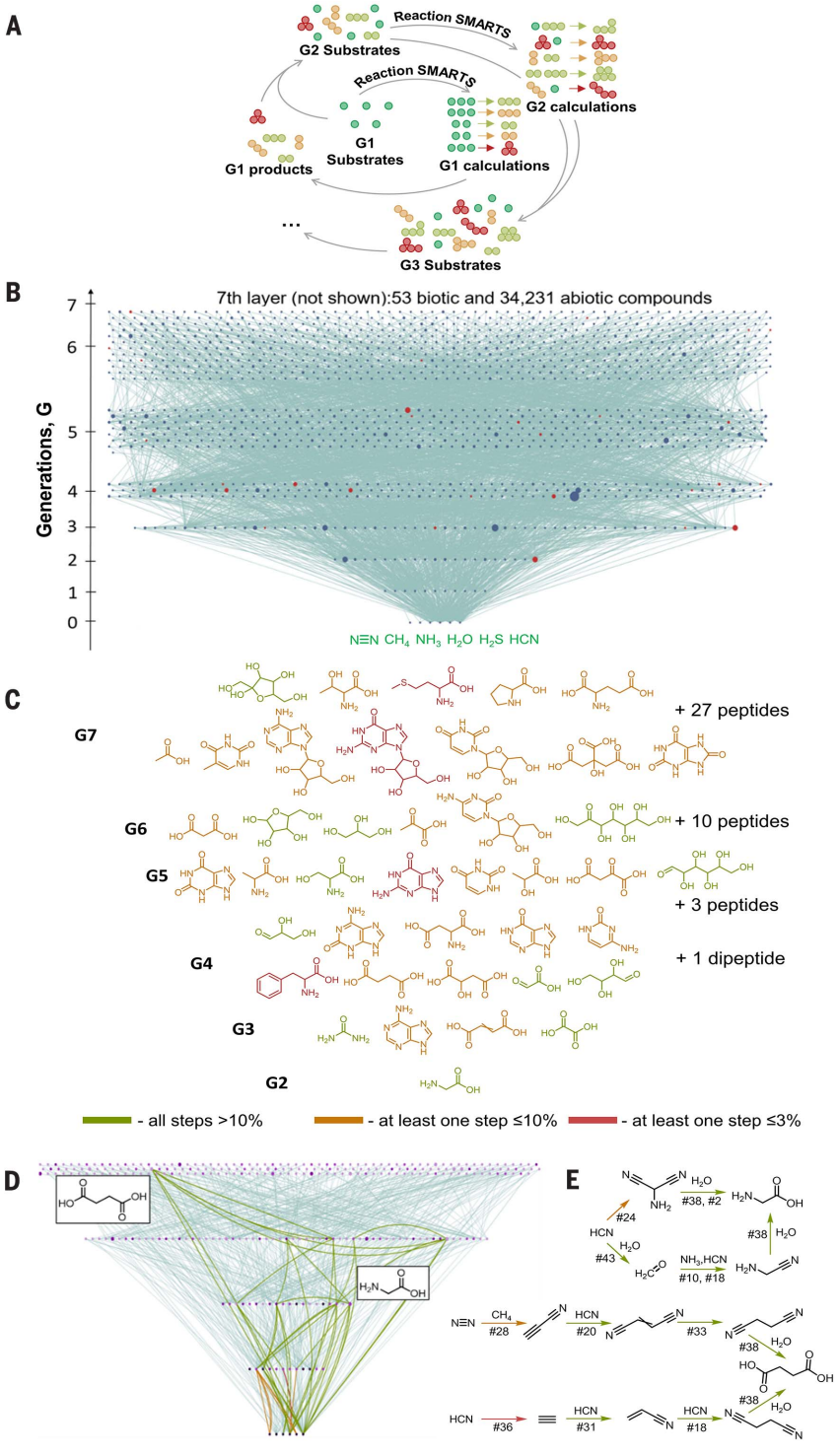


Fig. 1. Biotic and abiotic molecules in the network of prebiotic chemistry. (**A**) Scheme illustrating the synthetic algorithm in which SMARTS-coded (*22*) reaction transforms act on the current pool of reactants to produce the next generation of compounds. Afterward, these products are combined with original reactants and the procedure repeats until a user-specified number of generations is reached. (**B**) Six generations of a synthetic network originating from six primordial substrates—H_2O, N_2, HCN, NH_3, CH_4, and H_2S—and leading to possible biotic products [amino acids and peptides, nucleobases and nucleosides, carbohydrates, and metabolites found in living organisms (*28*); red circles] and abiotic products [other small molecules; blue circles] with molecular mass not exceeding 300 g/mol. Circle size corresponds to the molecule's incoming connectivity, k_{in} (i.e., the number of reactions that produce this molecule as product). (**C**) Forty-one biotic molecules within the network's seven generations [six generations are shown in (B); for the full network with all seven generations, see https://tol.allchemy.net]. Of the biotic class, glycine is in the second generation (G2); urea, adenine, butenedioic acid, and oxalic acid are in G3; glyceraldehyde, isoguanine, aspartic acid, hypoxanthine, cytosine, phenylalanine, succinic acid, malic acid, glyoxylic acid, and aldotetrose are in G4; xanthine, alanine, serine, guanine, uracil, lactic acid, oxaloacetic acid, and aldohexose are in G5; malonic acid, pentofuranose, glycerol, pyruvic acid, cytidine, and ketoheptose are in G6; and ketohexose furanose, threonine, methionine, proline, glutamic acid, citric acid, acetic acid, thymine, adenosine, guanosine, uridine, and uric acid are in G7. Various di-, tri-, and tetrapeptides are also present within G4 to G7 (fig. S55). In addition, histidine is in G8, arginine in G11, valine in G12, and leucine in G16. The molecules shown are colored according to the lowest-yielding step within the shortest pathway: Red, at least one step is predicted to generate only traces of product (≤3%); orange, at least one step is low yielding (≤10%); and green, all steps are predicted to proceed in moderate or high yields. (**D**) Allchemy's screenshot of the G4 tree, with the two shortest prebiotic synthesis pathways of succinic acid and of glycine colored according to the yields of individual steps [color coding as in (C)]. (**E**) Schemes of the pathways. Numbers below reaction arrows correspond to the transform labels in SM section S2; this section also contains details of prebiotically plausible reaction conditions (e.g., CuCN, KCN, H_2O, and irradiation for conversion of hydrogen cyanide into formaldehyde), along with pertinent literature references. Raw Allchemy screenshots of the pathways are provided in SM section S4. If two numbers are given below a single arrow, it means that the software recognizes the product of the first reaction as highly reactive and prone to the second reaction in a tandem sequence [e.g., hydrolysis of a nitrile to a carboxylic acid (#38) creates 2-aminomalonic acid, which readily undergoes elimination of carbon dioxide (#2) under hydrolysis conditions; formation of imines (#10) creates methyleneamine, which then undergoes addition of cyanide (#18)].

ranges, certain classes of solvents, or reaction conditions).

We began by quantifying the molecular composition and synthetic structure of the C-, O-, N-, and S-based network up to the seventh synthetic generation (G7), with molecular mass limited to 300 g/mol (see Fig. 1B for the subnetwork up to G6). Within this synthetic space, which was generated on a standard desktop computer within ~2 hours, there are 82 biotic molecules [amino acids and peptides, nucleobases and nucleosides, carbohydrates, and metabolites found in living organisms (*28*)] and 36,603 abiotic molecules. Of these, 41 non-peptide biotic compounds up to G7 are shown

in Fig. 1C and are color coded according to relative abundances estimated from the approximate individual-reaction yields along the shortest route to a given molecule [e.g., serine can be made efficiently, as confirmed in (*29*, *30*), whereas phenylalanine can be generated only in trace amounts, in agreement with (*31*)]. Of note, the biotic compounds are more thermodynamically stable than the synthesizable abiotic compounds of similar masses (see the distribution of heats of formation in Fig. 2A); are, on average, less hydrophobic (*32*) than those in the abiotic pool (as expected given that life began in water; Fig. 2B); and are more balanced in terms of hydrogen bond donors and acceptors (biotic molecules contain, on average, comparable numbers of donors and acceptors, which may facilitate formation of supramolecular aggregates; for statistics, see SM section S6). The biotic molecules also contain fewer reactive groups (e.g., no highly reactive imines compared with ~8000 such groups in the abiotic pool; see table S3 for other groups) and fewer distinct functional groups per molecule (2.67 in biotic versus 3.52 in abiotic). One way to rationalize this last difference is that the introduction of each new functionality might have required a change in reaction conditions, which was perhaps less likely on early Earth—in fact, pathways leading to biotic molecules entail fewer condition changes than those leading to abiotic ones (Fig. 2, C and E, versus Fig. 2, D and F).

In regard to its synthetic connectivity, the network is robust in the sense that removing as many as 34 of 63 C-, O-, N-, and S-based reaction classes still allows for the synthesis of all biotic molecules via bypass routes (as opposed to only eight removable reaction classes for all abiotic molecules to remain synthetically accessible; see fig. S57). This high degree of robustness (and of synthetic redundancy) is reminiscent of metabolic networks and could indicate that our network has a similar, scale-free architecture also characterizing the internet, airline networks, or even the network of all published organic reactions (*33*–*35*). This assumption is corroborated by the node power-law connectivity distributions shown in Fig. 2G, implying the existence of synthetic hubs (e.g., formic acid, cyanoacetylene, and isocyanic acid), which are becoming increasingly more connected as the network grows (via the so-called preferential attachment; Fig. 2H).

Because the individual reaction rules used to generate the network are derived from the OL literature, we trivially expect the network to contain the known prebiotic pathways leading to all of these biotic compounds. Indeed, all such syntheses are present in the network, as illustrated in Fig. 3, A and B, for adenine, a popular target of prebiotic studies (for syntheses of other targets, see SM section S4). Notably, in addition to cataloging known routes, the network also contains previously unreported syntheses of biotic molecules. As a case in point, consider a computer-generated subnetwork of reactions leading to succinic acid and also involving syntheses of lactic, pyruvic, malic, fumaric, and glyoxylic acids (all biotic molecules are depicted in green in Fig. 3C). Analysis of the network in comparison with known literature reveals that most routes to these biotic molecules are a patchwork of steps reported in different publications (corresponding to different colors of the arrows), some of which are not concerned with OL issues (steps marked as NOL for non-OL), but all performed under prebiotic conditions. In this regard, computational analysis is helpful merely as an aggregator of known but scattered synthetic information. More importantly, the software suggested three reactions, marked with thicker red arrows in Fig. 3C, that lack clear literature precedent and serve to establish new synthetic connections within the network and unlock synthetic pathways that had been overlooked in prior studies. The reaction marked "(1)"—hydrolysis of fumaronitrile—is relatively unimportant, as it exchanges the order of hydrogenation and hydrolysis steps from that in a known method of producing succinic acid (compare the sequences of blue and pink arrows). When carried out, the experimental yield depended on time and acid concentration and ranged from 8% in 0.1 M HCl (9 days) to 54% in prebiotically less likely 5 M HCl (1 day). Reaction (2) is another hydrolysis of nitrile groups (validated in >70% yield in 0.5 M HCl and 9% in 0.1 M HCl) but is more consequential in that it establishes a route to pyruvic acid that does not require sulfur. This route is also shorter, with six steps [starting from Orgel's classic prebiotic conversion of N_2 and CH_4 into cyanoacetylene under discharge (*5*)] versus eight steps for the sulfur-containing route starting from HCN (light blue and orange arrows). Reaction (3) involves tandem hydrolysis and decarboxylation of cyanoacetaldehyde to produce acetaldehyde (validated in 19% yield in 0.1 M HCl and 34% in 5 M, both 7 hours). This reaction is of interest because it enables a sulfur-free synthesis of lactic acid in five steps, as opposed to seven steps from HCN and H_2O.

Turning our attention to other classes of prebiotic chemistry, we considered the synthesis of citric acid (CA), illustrated in Fig. 3D. Recently, a prebiotic mimic of the CA cycle was reported (*10*), but it contained only analogs of CA and not CA itself. Our network analysis suggested that CA could emerge under prebiotic conditions in water from two equivalents of oxaloacetic acid [the synthesis of which has already been reported in the OL literature (*10*)] via a tandem aldol self-condensation (H_2O, pH 7.5, 4°C) and decarboxylation sequence followed by a second decarboxylation. This second decarboxylation, promoted by either 0.054 or 0.081 M $FeCl_3$, worked better at room temperature than at 70°C, the temperature used for related compounds in previous work (*10*). Under our milder conditions, we obtained CA in ~5% yield, whereas under harsher conditions, the citroylformic acid substrate gradually decomposed, reducing the yield to ~2% (table S8).

Next, we validated a computer-predicted synthesis of diglycine from *N*-carboxyanhydride (NCA) in a sulfur-rich environment (Fig. 3E)—that is, under conditions that differ from the sulfur-free route described by Bartlett and Jones (*36*). The first step, thiolysis of glycine *N*-carboxyanhydride, was performed at room temperature in the presence of H_2S and potassium carbonate and proceeded in 40% yield (additionally, ~55% of unreacted substrate was recovered). The resulting thioacid was then oxidized with $K_3[Fe(CN)_6]$ to a disulfide (thioacid oxidative dimer), which spontaneously underwent an intramolecular nucleophilic acyl substitution rearrangement (in which an amine group displaced the disulfide from the distal carbonyl) followed by hydrolysis. The yield for the two thioacid-to-diglycine steps was 29%.

Finally, we validated a five-step synthesis of uric acid, which previously has been obtained under prebiotic conditions only in trace amounts, from urea and acetylene subjected to ultraviolet irradiation (*37*). Under optimized reaction conditions, all five steps of our computer-designed plan, shown in Fig. 3F, proceeded in one pot without isolation of intermediates, although their presence was confirmed by high-resolution mass spectroscopy as well as by the execution of partial reaction sequences. Starting from aminomalononitrile and NaOCN, the first two steps were performed at pH 4 at room temperature overnight. Afterward, the reaction mixture was supplemented with an additional portion of NaOCN, acidified (with 1 M HCl), and left at room temperature for 24 hours to complete the sequence. When performed in water, the average per-step yield was ~30%; in a prebiotically plausible 4/1 v/v mixture of water and acetonitrile (*38*, *39*), it was ~40%, translating to an overall, five-step yield of 1% (table S21). For all synthetic details regarding this and other syntheses, see SM sections S7 to S12.

Perhaps the most noteworthy and far-reaching finding of this work is that the network gives rise to three forms of chemical emergence that are not a simple consequence of the primordial substrates and initial reaction transforms.

Emergence of catalysts and reaction types

We first discuss the finding that compounds created within the network can themselves act as catalysts of additional chemical reactions, all operative under prebiotic conditions, thereby

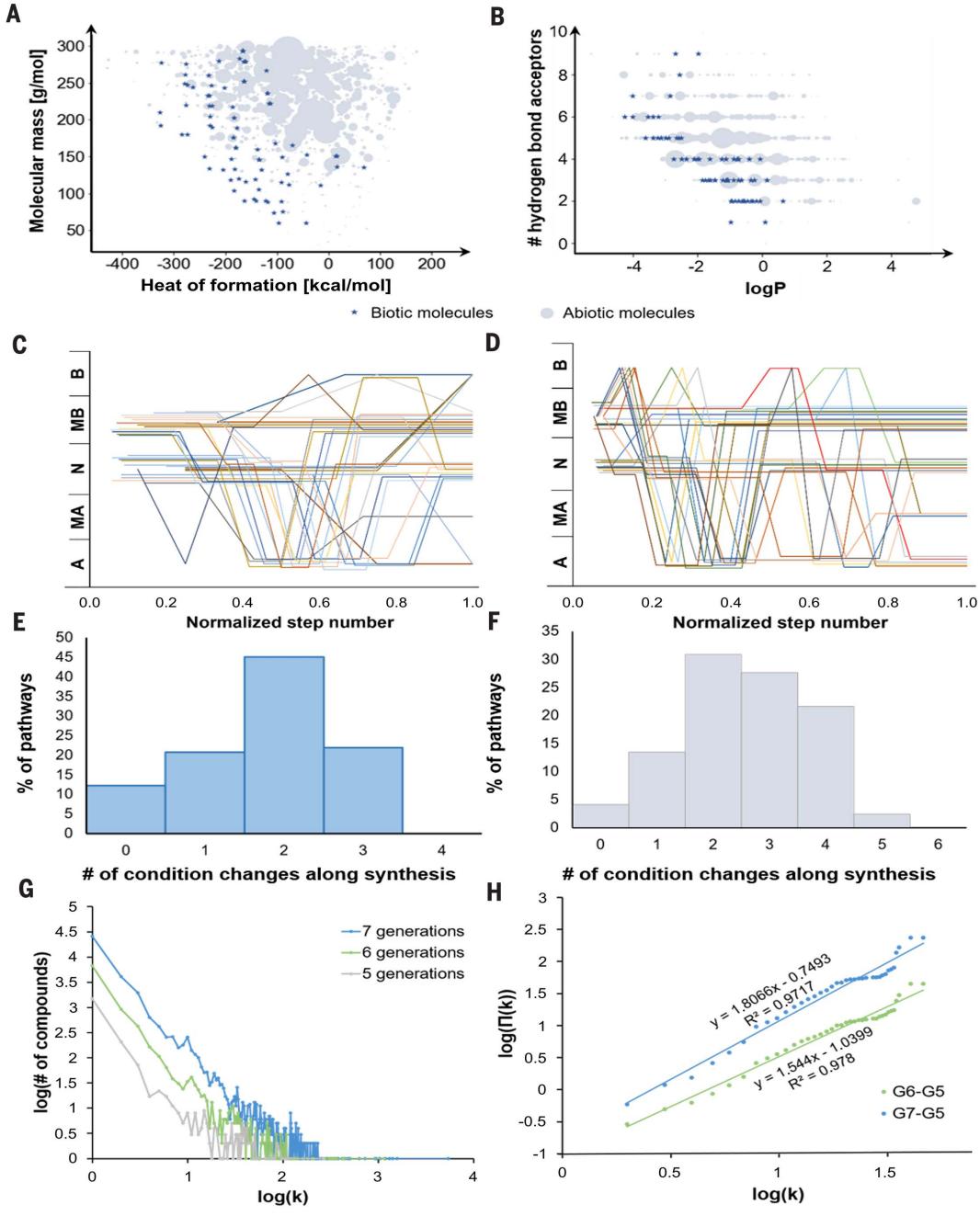


Fig. 2. The network's molecular content and synthetic connectivity. (**A**) Distribution of biotic (blue markers) and abiotic (gray markers) molecules in a plane defined by molecular mass and heat of formation calculated using the PM6-D3H4X (*66*) semiempirical method implemented in the MOPAC2016 software (*67*). To simplify presentation, abiotic compounds were clustered into 1202 groups according to their structural similarity (quantified by Tanimoto coefficients between molecules' ECFP4 fingerprints). Each cluster is represented by a circle of diameter proportional to the number of members, and position is determined by the group's centroid (i.e., a group's "representative" molecule, defined as the molecule with maximum average Tanimoto similarity to other members of the cluster). A similar correlation is observed when larger and unclustered samples of abiotic compounds are considered (see fig. S11 for distributions of >11,000 compounds; also see table S1 for additional thermodynamic considerations). (**B**) Distribution of biotic and abiotic compounds in a plane defined by the logP values calculated from Wildman and Crippen's method (*32*) and the number of hydrogen bond acceptors. Biotic compounds are, on average, less hydrophobic than abiotic compounds for a given number of hydrogen bond acceptors. Further details of the underlying feature selection are described in the SM section S6.2. (**C** and **D**) Graphical illustration of condition changes along synthetic pathways leading to 30 randomly chosen (C) biotic versus (D) abiotic compounds in the G7 network. The horizontal axes quantify the numbers of steps in each pathway: For an *n*-step synthesis, the first step will correspond to location 1/*n*, the second step to 2/*n*, and the final step to *n*/*n* = 1 (i.e., all pathways stop at the scale's value of 1). Conditions on the vertical axis: A, acidic; MA, moderately acidic; N, neutral; MB, moderately basic; and B, basic. (**E** and **F**) Full condition variability statistics are summarized in histograms for the syntheses of (E) 82 biotic molecules and (F) 36,603 abiotic molecules. The difference in the two distributions is statistically significant with P value < 0.001, as evaluated by χ^2, Kolmogorov-Smirnov, and bootstrap tests (SM section S6.1). (**G**) Distribution of node degrees (k) for G5, G6, and G7 networks. Connectivity of a given node is the sum of the numbers of its incoming and outgoing connections ($k = k_{in} + k_{out}$; for distributions of k_{in} and k_{out}, see fig. S58). Linearity of these dependencies shown on the doubly logarithmic scale indicates a power law—$P(k) \propto k^{-\gamma}$, where $\gamma \sim 1.8$—and a scale-free network architecture. (**H**) Cumulative distribution of $\Pi(k) = \sum_{k_i=0}^{k} < \Delta k_i >$ versus k provides evidence for preferential attachment. In this expression, $< \Delta k_i >$ denotes the average increase of the degrees of nodes with $k = i$ between the fifth and sixth generations (green curve) and between the fifth and seventh generations (blue). The plot traces such evolutions of all nodes present in the network's fifth generation (compounds in G5 with only a single incoming connection are not considered). The linearity of the dependence on the doubly logarithmic scale indicates another power law, $\Pi(k) = k^{\alpha}$, and an exponent greater than unity ($\alpha \sim 1.6$ to 1.8) confirms preferential attachment. Notably, the slopes of both power laws are close to the values we previously found (*35*) for the scale-free network of all organic chemistry, indicating that prebiotic and modern organic syntheses are both governed by the same rules of synthetic reactivity.

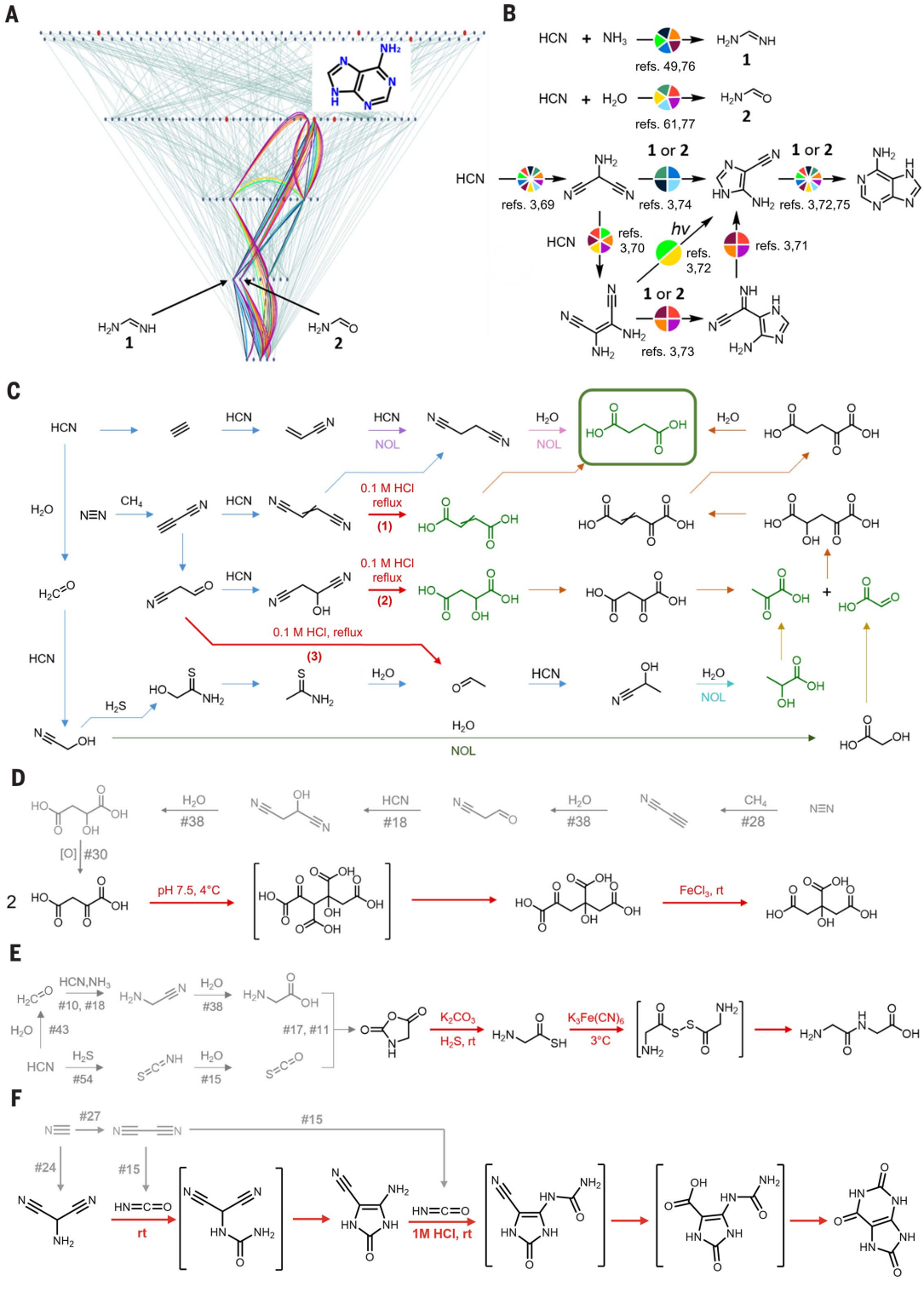


Fig. 3. Examples of known and newly identified syntheses within the network of prebiotic chemistry. (**A**) Ten prebiotic synthetic pathways leading to adenine, all previously described in the OL literature, are highlighted in the network (for clarity, only a subnetwork of C-, O-, and N-based chemistries up to G4 is shown). Identical synthetic connections common to several pathways are indicated by arcs of different curvatures. (**B**) Chemical schemes of adenine's prebiotic syntheses, along with those from pertinent literature (*3*, *49*, *61*, *69*–*77*). Colored circles over the arrows correspond to the colors of pathways shown in (A). Circle segments are used to indicate to which multiple pathways a given step belongs. The first (trimerization of HCN) and last (formation of amides, imides, amidines, or guanidines followed by cyclization) steps are common to all pathways. There are three main strategies in the syntheses of adenine; formamidine (**1**) and formamide (**2**) participate as second substrates in three key, two-component reactions. *hν*, light. (**C**) Subnetwork of reactions that lead to succinic acid and involve syntheses of lactic, pyruvic, malic, fumaric, and glyoxylic acids (biotic molecules are in green). Previously unreported connections now verified by experiment are denoted with red arrows. Previously reported connections share the same color if they come from the same source publication. NOL indicates reactions reported outside of origins research but performed under prebiotic conditions. (**D** to **F**) Syntheses of (D) citric acid, (E) diglycine, and (F) uric acid. Gray arrows and structures denote reactions that have been described previously [the fourth reaction in (D), hydrolysis of malonitrile, is described in (C)]; black structures and red arrows represent the software-predicted reactions that we verified experimentally. When several reactions were performed in one pot, some intermediates were not isolated (but were still confirmed spectroscopically); these are enclosed in square brackets. For all experimental details, see the main text and experimental procedures in SM sections S7 to S12. rt, room temperature.

substantially expanding the accessible prebiotic chemical space. To show this, we queried the network for known organocatalysts or bi- and tridentate metal chelators capable of binding metal cations present on primitive Earth [e.g., Cu(II), Zn(II), and Mn(II) (*40*)] and also used in modern organometallic catalysts. Figure 4A lists eight such catalysts enabling different reaction types and collectively more than doubling the size of the network. All of these reactions were previously carried out under prebiotic conditions (*41*–*48*), but their relevance to OL was unnoticed. In this figure, the arrow next to each reaction indicates how many additional compounds a particular reaction helps generate up to G7 (examples of products are shown on the right), whereas the arrow at the bottom quantifies the network's expansion, by a total of ~56,000 molecules, when all reactions are added to the generative set at once and can act synergistically (with products on some reactions serving as substrates to others).

For example, formaldehyde (Fig. 4A, entry 1), created in the network's second generation (G2), can act as an organocatalyst to enable selective hydrolysis of α-amino nitriles. This reaction was actually carried out under prebiotic conditions by Chitale *et al.* (*41*), but not on substrates that contained another potentially competing nitrile group. As shown in Fig. 4B, we confirmed experimentally that selective hydrolysis (0.2 M NaOH, 0.08 M formaldehyde, H_2O, 1 hour) of such substrates is possible, leading to 2-amino-4-cyanobutanamide in 90% yield. When the reaction was carried out in the absence of the formaldehyde catalyst (other conditions unchanged), the yield was only ~2% (SM section S8).

In entries 2 and 3, acetate (OAc) created in G7 and coordinated to copper in a Cu(II) diacetate complex can catalyze two reactions: oxidation of α-oxoalcohols (*42*) and formation of imidazole from α-oxoalcohols, aldehydes, and ammonia (*43*). Notably, the former reaction, 2, can unlock new prebiotic syntheses of amino acids such as serine or phenylalanine, whereas the latter, 3, can be used to construct the imidazole ring of histidine. Histidine's synthesis, newly found in the network and outlined in Fig. 4C, is appealing because it avoids the use of formamidine {previously obtained in only trace amounts from sodium cyanide and ammonium hydroxide [0.2% in (*49*)] and reacted with aldotetrose (*50*)} and because it reuses the same substrate, formaldehyde, to create glycolonitrile and then to construct the imidazole ring via a multicomponent, $Cu(OAc)_2$-catalyzed reaction of erythrulose, $H_2C{=}O$, and NH_3. We validated this key step, marked by the red arrow, experimentally (room temperature for 48 hours, then 75°C for 2 hours, H_2S bubbling to release imidazole from the copper-imidazole complex) and obtained 1-(1*H*-imidazol-4-yl)ethane-1,2-diol in 22.4% yield. No product was observed when the reaction was performed without $Cu(OAc)_2$ (SM section S13).

Other examples are listed in entries 4 through 9. In 4, when proline coordinates to Zn(II) it can catalyze, in water or without solvent and under elevated temperature (*44*), a classic variant of multicomponent A3 coupling that involves aldehyde, terminal alkyne, and primary or secondary amine. Additionally, in entry 5, a modified variant of A3 coupling that involves heterocyclic azines can be catalyzed by Cu(II) salts in the presence of glucose (*45*) created in G7. In 6, imidazole from G4 can act as an organocatalyst to facilitate formation of primary amides from carboxylic acids and urea (*46*). In 7, phenylalanine created in G4 has been shown to catalyze dehydrative furan ring formation from carbohydrates (*47*); this reaction can enable syntheses of furfural or 5-hydroxymethylfurfural in water under mild conditions. Finally, in entries 8 and 9, IDA from G4 can coordinate to either Mn(II) or Cu(II). The Mn(II) complex has been shown to catalyze (*48*) epoxidation of alkenes not conjugated with a carbonyl group [a much broader scope than prebiotic epoxidation of only acrolein reported by Fernández-García *et al.* (*51*)]. With this reaction included, the network encompasses compounds such as tartaric acid, which could be a plausible prebiotic precursor of pyruvic and oxaloacetic acids (Krebs cycle–related, extant metabolites); unnatural acids such as isoserine or β-hydroxyaspartic acid; and α-sulfanylcarboxylic acids and analogous ethers or thioethers. On the other hand, IDA's Cu(II) complex can catalyze hydrolysis of α-amino acid esters under benign conditions (*52*).

We have restricted the above analyses to exact matches between prebiotic molecules and known catalytic ligands, but the network contains several other candidates for bi-, tetra-, and pentadentate ligands of potentially new organometallic catalysts (fig. S56, B and C). Testing such ligands could lead to the discovery of additional catalysts and channels for prebiotic evolution.

Emergence of chemical systems

The second form of emergence goes beyond individual reactions or even synthetic pathways and relates to primitive chemical systems such as reaction cycles or cascades. Cycles can be difficult to design because they must contain at least one degradation or fragmentation step (to revert to the starting material), and such steps may not be intuitive to chemists accustomed to building up mass during synthesis. Self-regenerating (*53*) cycles—central to many biological processes (such as glycolysis) and often postulated as essential to the emergence of life [e.g., Kauffman's decades-old hypothesis (*54*) of life arising from autocatalytic reaction networks (*55*)]—are particularly difficult to detect because they must also produce by-products identical to one of the cycle's members. Computationally, identification of cycles within a directed graph such as our network is relatively straightforward (*56*), and because our reaction templates are stoichiometrically balanced and keep track of all reaction products, we are also able to discover self-regenerating cycles.

Within just a few synthetic generations from the primordial substrates, the prebiotic synthetic space becomes relatively densely populated with cycles, the statistics and molecular diversity of which are quantified in fig. S59. Notably, within the G7 network, there are already multiple cycles that could be self-regenerating. Figure 5 shows examples of such cycles, fueled by NCA and predicted to produce up to two copies of incoming IDA (Fig. 5A) or even three copies of *N*-(2-cyanoethyl)glycine, albeit over more steps (Fig. 5C). Of course, such idealized diagrams rest on an (unrealistic) assumption that all steps will proceed in quantitative yields. In reality, even one step with <50% yield can push the cycle's yield below 100% and thus prevent self-regeneration. On the other hand, the software estimated that all of the individual reactions within the predicted cycles should proceed in good yields. In addition, for the IDA cycle with NCA aminolysis, Strecker reaction, and hydrolysis (path **1** → **2** → **3** →**1** in Fig. 5A), the algorithm also identified a bypass through which the by-product of reaction **1** → **2** (i.e., **2** reacted with another copy of NCA fuel) undergoes Strecker reaction and hydrolysis but ultimately also regenerates two copies of IDA for each molecule of **4** (path **1** → **2** → **4** → **5** → **1**). This bypass could thus increase IDA's recovery over the cycle and could help us with validation if the software correctly predicted the formation of by-products **4** and **5**.

For these reasons, we proceeded with experimental validation of the IDA cycle in water, under prebiotic conditions for all reactions, and with condition changes (here, basic-acidic-basic) between different steps, not unlike in other multistep prebiotic syntheses [e.g., wet-dry cycles in (*57*) or pH changes in (*8*, *58*, *59*)]. As summarized in the phase diagram in Fig. 5B, we established that the overall yield (efficiency) of the cycle depended on the concentration ratio of the IDA and NCA reagents used in the first aminolysis step, on the pH during the Strecker reaction, and on the concentration of NaOH used for the final hydrolysis. Under optimal conditions, the first step (pH 10.2, 0°C, vigorous stirring, 2.2 equiv of NCA) converted ~70% of IDA (**1**) into **2** and **4**. Subsequently, the reaction mixture was treated with 2.2 equiv of formaldehyde and 2.2 equiv of potassium cyanide (*8*) at pH 6 for 16 hours to produce **3** and **5**. The mixture was then, without purification, subjected to hydrolysis [in prebiotically

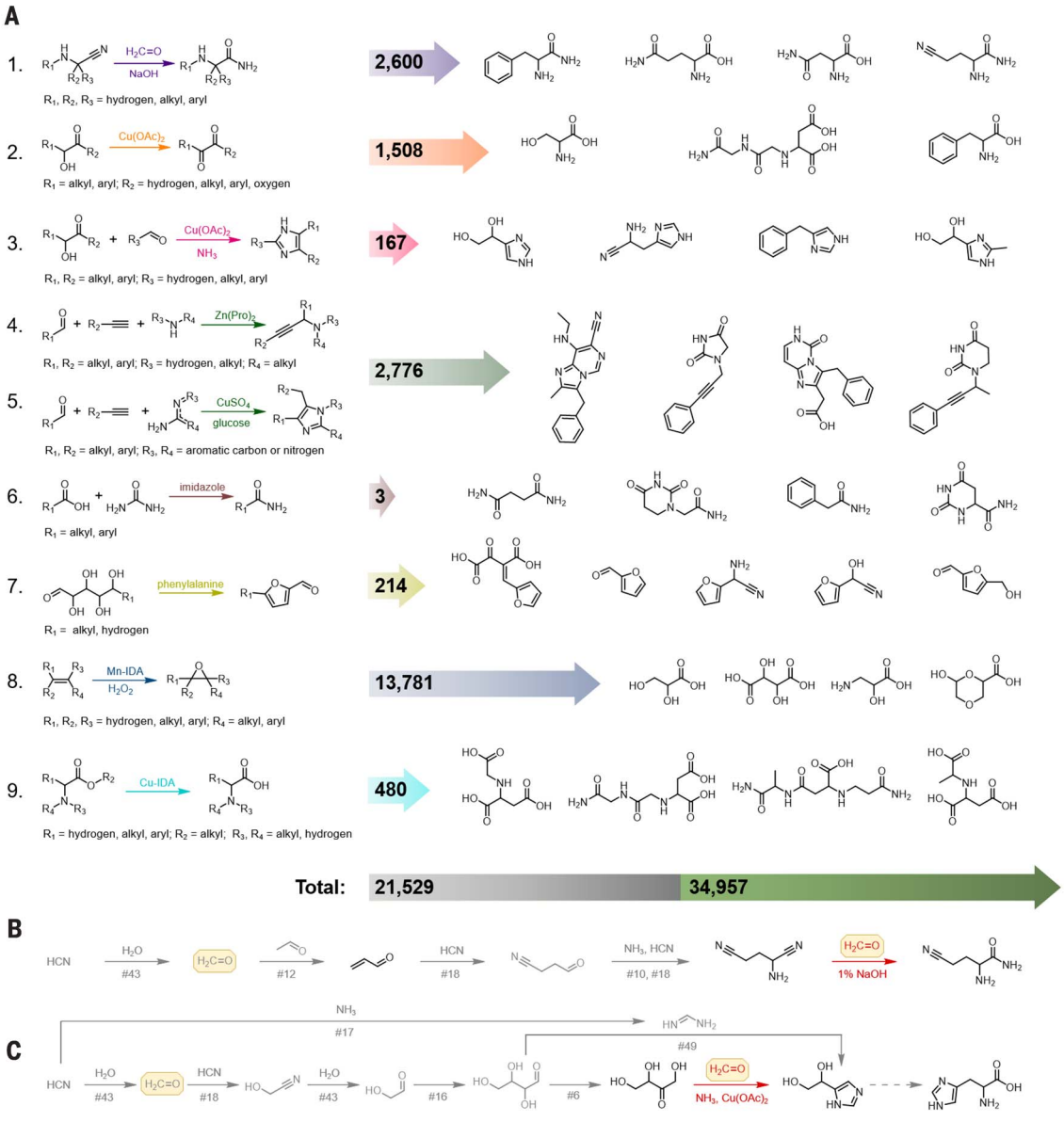


Fig. 4. Chemical emergence in the network of prebiotic chemistry. (**A**) (Left) Eight types of chemical reactions enabled by seven molecules created in the original G7 network (note that OAc and IDA repeat twice in nine entries shown). These molecules are either organocatalysts or components of catalytic complexes with prebiotically plausible metal cations [e.g., Zn(II), Cu(II), and Mn(II)]. All of the reactions shown had been previously performed under prebiotic conditions, but their relevance to origins research was not noted. (Middle) Colored arrows illustrate how many additional compounds can be created in our prebiotic network upon addition of each of the reactions shown. There is one arrow for entries 4 and 5 because two different catalysts enable the same reaction type (A3 coupling). (Right) Examples of molecules that are made synthesizable via these reactions. The gray part of the arrow at the bottom indicates the sum of these molecules (21,529), and its green extension represents the additional 34,957 molecules that are created when all of the reactions (1 through 9) are added to the generative set simultaneously. (**B**) The red arrow corresponds to the selective hydrolysis of 2-aminopentanedinitrile to 2-amino-4-cyanobutanamide catalyzed by formaldehyde, a reaction proposed by the software and validated experimentally. The remaining steps illustrate how the software navigated the synthesis of the aminopentanedinitrile substrate from the HCN primordial feedstock. These steps are shown in gray to indicate that they have already been executed by others and described in the OL literature. (**C**) The red arrow corresponds to the synthesis of 1-(1*H*-imidazol-4-yl)ethane-1,2-diol from ketotetrose, ammonia, and formaldehyde catalyzed by copper(II) acetate. This transformation was proposed by the software and chosen for experimental validation because it establishes an unreported prebiotic route to the histidine amino acid. All downstream and upstream steps (in gray) have been described earlier in OL literature. Previously, histidine was generated along an inefficient bypass (also shown in the scheme) from aldotetrose and formamidine (*50*).

plausible (*60*–*63*) 1.7 M NaOH, 100°C, 3 hours] to regenerate IDA in 126% cycle yield (±2.6% on the basis of three measurements; for yields and intermediate distributions under different conditions, see tables S11 and S12). Electrospray ionization mass spectrometry and high-performance liquid chromatography (SM section S9) confirmed formation of the main-cycle intermediates **2** → **3**, as well as **4** and **5** from the bypass route.

Emergence of surfactants

Finally, the third class of emergence was the formation of surfactant molecules capable of spontaneously forming vesicles that could potentially house reactions and systems such as

those described above. As illustrated in Fig. 6A, straight-chain saturated fatty acid and α-hydroxy acid surfactants can form through repeated four-step cycles of aldehyde homologation. Breaking the cycle, the last step of fatty acid synthesis, may then occur via nitrile or thioamide hydrolysis to carboxylic acid. The aldehyde homologation cycle was proposed by Patel *et al.* (*8*) as a prebiotic method to make hydrophobic amino acids, but its straightforward extension to fatty acid surfactants was not noted in that report. In another and synthetically much shorter approach, peptide surfactants with variable glycine or alanine tails and aspartic acid head groups are available within only a few synthetic generations. Previously, such peptides had been synthesized by modern, nonprebiotic synthetic methods and had been shown (in the context of nanotechnology, not OL research) to form nanotubes and nanovesicles (*64*). In the prebiotic route within our network (Fig. 6B), an initial amino acid (AA; glycine or alanine) reacts with carbonyl sulfide (reaction type 17, SM section S2) followed by cyclization (reaction type 11). Aspartic acid reacts with the thus obtained *N*-carboxyanhydride (Leuchs' anhydride), resulting in dipeptide (reaction type 52), which can be gradually extended by the addition of anhydride molecules, in n +1 overall steps, yielding a $(AA)_n$-Asp chain.

Outlook

Taken together, the above analyses and synthetic examples lead us to suggest that computational reaction network algorithms are useful for identifying new synthetic routes to prebiotically relevant targets and indispensable for the discovery of prebiotic chemical systems that are otherwise challenging to discern. Naturally, the prebiotic reaction networks should and will grow as distinct prebiotically plausible transformations are experimentally validated. As such transformations are added to our generative set (by means of the crowd-sourcing module illustrated in fig. S5, panel xii), we expect that network analyses will be able to trace prebiotic syntheses starting from primitive feedstocks to increasingly complex scaffolds, including those found in modern drugs (e.g., Fig. 6C). In other words, we envision a fruitful junction between prebiotic chemistries, performed in water and often under environmentally friendly conditions, and environmentally friendly pharmaceutical synthesis. This idea echoes the pioneering efforts of Eschenmoser and colleagues (*65*) to synthesize complex targets (e.g., uroporphyrinogens) from prebiotic substrates; with the help of computers, similar efforts can now be streamlined and further extended. Finally, the scope of this work would be broadened if kinetic data for prebiotic reactions became available. It would then be interesting to probe the network for systems of reactions that exhibit rate enhancement and autocatalysis, as well as those that allow for rate-controlled product selection, as in the example of a double cycle described in fig. S60B.

Materials and methods summary

A detailed description of the Allchemy software, its user manual, and examples of output as well as all synthetic details are provided in the supplementary materials. A summary of Allchemy's key routines and theoretical methods is presented here.

General overview of the Allchemy platform

The Allchemy web application is based on the *Django* (www.djangoproject.com) framework, using *PostgreSQL* (https://postgresql.org) for storing calculation results. The web application uses the *d3.js* library (https://d3js.org) for graph representation and *Chemwriter* (https://chemwriter.com/) for visualizing chemical structures. Communication between the web application and Allchemy's back end is supported by the *Redis* (https://redis.io) and *RQ* queue systems (https://python-rq.org/). Cycle-search algorithms were implemented using the *NetworkX* (https://networkx.github.io/) or *graph-tool* (https://graph-tool.skewed.de/) libraries.

Allchemy applies reaction rules coded in the SMARTS (SMILES arbitrary target specification) notation to a set of substrate molecules represented in the SMILES (simplified molecular-input line-entry system) format. The process starts with an initial pool of substrates specified by the user, either in the SMILES format or by drawing molecular structures in *Chemwriter*. During each iteration ("generation"), reaction rules are applied to the current pool of compounds that includes the initial substrates and the products of preceding generations. More specifically, each generation entails the following operations:

1. Matching molecules to reaction templates

Allchemy's reaction rules specify substrates and products to within a specific reaction core (i.e., atoms chemically relevant to a given reaction), as well as a list of functional groups incompatible with this reaction's conditions. Both the core and the incompatible groups are defined using SMARTS (*22*) notation. A molecule is deemed suitable for a given reaction if (i) it contains the core of at least one substrate, as defined by this reaction, but (ii) does not contain any groups incompatible with the reaction. Both matching conditions are evaluated with the *GetSubstructMatches* function from the RDKit library (www.rdkit.org) for all molecules considered in a given synthetic generation. For example, for a reaction involving two substrates, if the first substrate defined in the reaction template matches to five molecules in the current pool of available molecules and the second substrate matches to four other molecules, then the algorithm will identify 20 pairs of molecules on which it will seek to perform the reaction operation. Notably, both substrate templates might be present in one molecule; if so, the algorithm will perform an intramolecular reaction.

2. Reaction run

For each suitable substrate combination, Allchemy applies the chemical transformation and computes possible products. This process is based on the *ChemicalReaction* class from the RDKit library (specifically, the *RunReactants* function) with in-house enhancements (e.g., subroutines enforcing proper tautomeric forms of products). In general, each of Allchemy's chemical transformations requires one or more executions of *RunReactants* (e.g., full esterification of glycol requires Allchemy to execute *RunReactants* of the *ChemicalReaction* object for esterification twice, or three times in the case of glycerin).

3. Post-filtering of products

Reaction products are then filtered by several subroutines to remove chemically invalid molecules (e.g., compounds bearing small rings with triple bonds) and those that do not fulfill user-defined constraints (e.g., those that exceed the mass limit).

4. Construction of reaction paths

A reaction path is stored as an ordered list of reaction steps, each of which is a tuple of reaction SMILES and reaction name. Each compound in the graph is assigned a list of its distinctive reaction paths discovered up to a given generation (as the algorithm may identify multiple paths for each compound). During reaction run (step 3), path lists of each product are either updated (if a compound was already present in the pool) or created; in the latter case, the path list is initialized with an empty sequence. The construction process can be described by three elementary steps: (i) concatenation of path lists of substrates ("ancestors") with a path list of a product (so that each reaction path of each ancestor is also included in a new path list); (ii) appending the current reaction step to each path in the list; and (iii) removal of duplicate entries (if necessary).

5. Identification of reaction cycles

To identify cycles within the network of prebiotic reactions, the network is converted into a directed graph, in which nodes represent molecules and edges denote reactions. Two nodes are connected if there exists a reaction connecting one substrate node to one product node (e.g., cyanoacetylene synthesis from methane and nitrogen is represented as three nodes with two edges between each substrate and the product). In such a graph, a cycle is defined

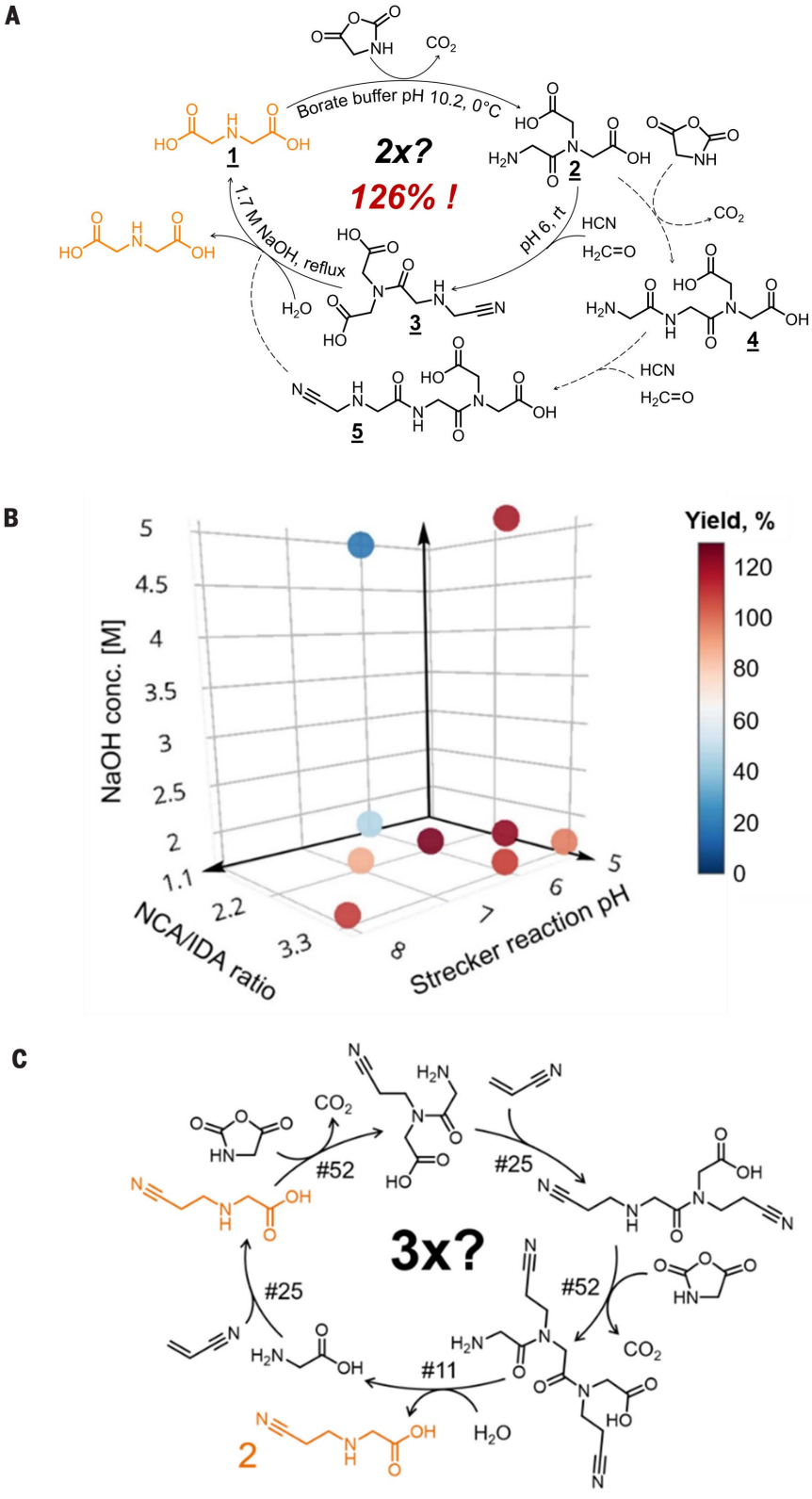


Fig. 5. Emergence of self-regenerating cycles within the network of prebiotic chemistry. (**A**) Self-regenerating cycle in which one molecule of IDA (orange) can produce up to two copies of itself. When the cycle was executed experimentally under prebiotic conditions (indicated next to the arrows), and upon pH changes from basic to slightly acidic to basic, it regenerated 126% of the IDA substrate, confirming autocatalysis. Dashed arrows trace the bypass route (through **4** and **5**) that may also be used to regenerate IDA. (**B**) Plot quantifying the experimentally observed cycle yields for different combinations of the key parameters: the concentration ratio of the IDA and NCA reagents used in the first aminolysis step, the pH during the Strecker reaction, and the concentration of NaOH used for the final hydrolysis [5 M is not a likely prebiotic condition and actually produced suboptimal yield (table S13), but we tested it solely to map the phase space of the cycle]. Circle color corresponds to the yield scale on the right. For all experimental details, see SM section S9. (**C**) Another noteworthy cycle candidate pending experimental validation and producing up to three copies of each incoming (2-cyanoethyl)glycine molecule (orange). For an average 80% yield of each step, the overall cycle yield would still be ~114%.

as a path from a given node to itself. The cycles are then identified by means of a suitably modified breadth-first search algorithm, as implemented in NetworkX or graph-tool libraries.

6. Computation of molecular properties and statistical analyses

All molecular properties were calculated using the RDKit library. In particular, the octanol-water partition coefficient (logP) was calculated with the Wildman-Crippen's method (based on the summation of atomic contributions). Thermodynamic properties were calculated using the PM6-D3H4X (*66*) semiempirical method (PM6 with empirical corrections

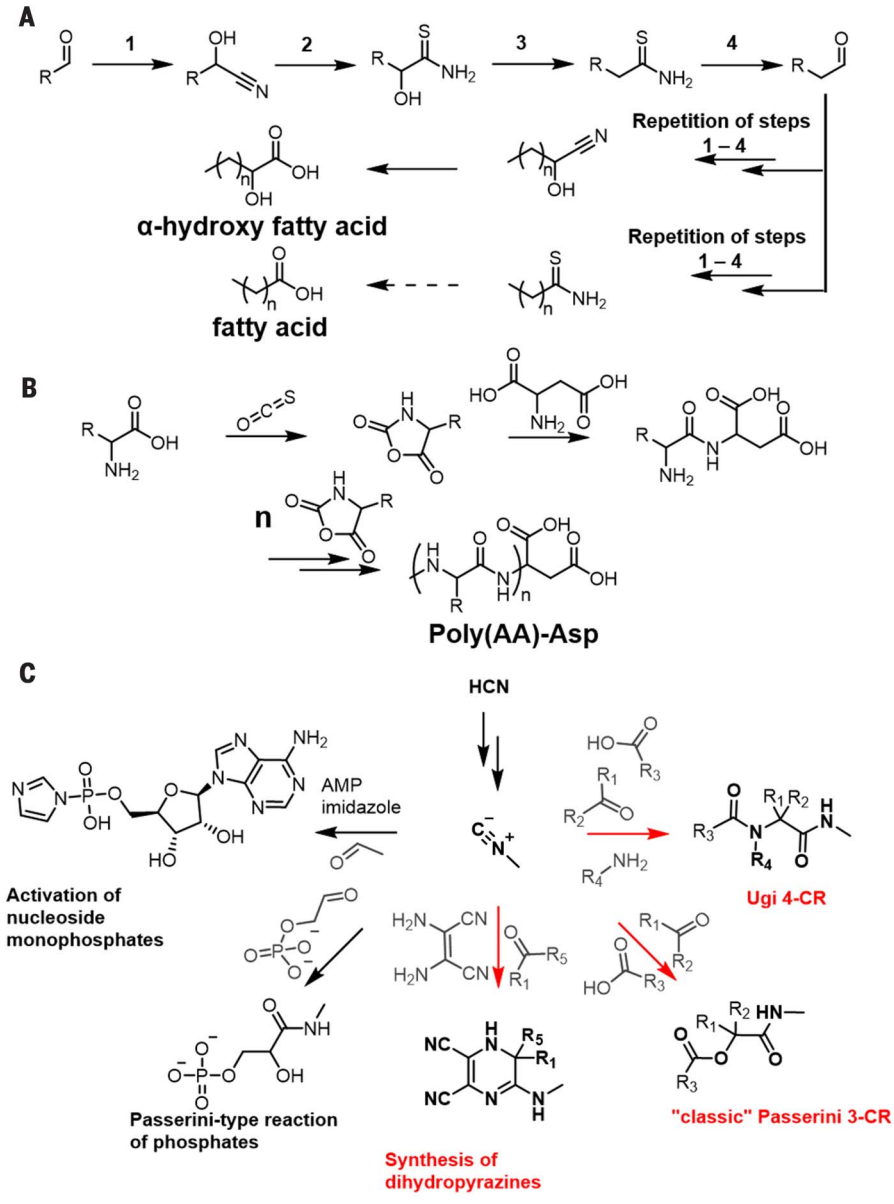


Fig. 6. Biomimetic routes to surfactants and additional pharmaceutically relevant scaffolds. (**A**) Prebiotic synthesis of fatty acid and α-hydroxy fatty acid surfactants by iteration of a known prebiotic sequence of four reactions homologating an aldehyde [for reactions 1 to 4, see (*8*)] followed by a previously unidentified breaking of the cycle via straightforward nitrile or thioamide hydrolysis. (**B**) A much shorter (i.e., fewer reaction steps) synthesis of peptide surfactants with variable glycine or alanine tails and aspartic acid head groups via sequential addition of Leuchs' anhydride. (**C**) Implications of recently reported (*78*) prebiotically plausible methyl isocyanide formation from HCN, ultimately allowing for Passerini-type reactions (black arrows). Addition of methyl isocyanide to our reaction set substantiates prebiotically plausible syntheses of some useful scaffolds (red arrows): α-acyloxycarboxamides via a classic Passerini reaction (3-CR, three-component reaction), peptide mimics via the four-component Ugi reaction, or heterocyclic derivatives of pyrazine via a less obvious three-component reaction, which has been reported in non-OL literature under prebiotic conditions (*79*). In the schemes shown, R_1, R_2 = alkyl, aryl, or hydrogen; R_3 = any carbon; R_4 = alkyl, aryl; and R_5 = alkyl, aryl. AMP, adenosine monophosphate.

for intermolecular interactions) implemented in the MOPAC2016 software (*67*). Statistical tests summarized in SM section S6 were performed with in-house Python scripts using the SciPy and Scikit-learn libraries. These scripts are deposited at Zenodo (*68*).

Overview of synthetic methods

Details of all syntheses described in the text are provided in SM sections S7 to S13. All reagents and solvents were purchased from commercial sources (Sigma-Aldrich, ABCR, POCH, Chempur, and Enamine) and, unless otherwise noted, were used without further purification. Nuclear magnetic resonance spectra were recorded on Bruker 400 MHz Avance III, Bruker 500 MHz, or Varian 600 MHz spectrometers. The liquid chromatography–mass spectrometry quantitative analyses were performed using a High-Performance Liquid Chromatograph Prominence LC-20 instrument (Shimadzu) coupled with a tandem mass spectrometer 4000 Q TRAP (SCIEX) equipped with an electrospray ion source.

REFERENCES AND NOTES

1. S. L. Miller, A production of amino acids under possible primitive earth conditions. *Science* **117**, 528–529 (1953). doi: 10.1126/science.117.3046.528; pmid: 13056598
2. A. I. Oparin, *The Origin of Life* (Academic Press, ed. 3, 1957).
3. J. Oró, Synthesis of adenine from ammonium cyanide. *Biochem. Biophys. Res. Commun.* **2**, 407–412 (1960). doi: 10.1016/0006-291X(60)90138-8
4. J. Oró, Comets and formation of biochemical compound on primitive Earth. *Nature* **190**, 389–390 (1961). doi: 10.1038/190389a0
5. R. A. Sanchez, J. P. Ferris, L. E. Orgel, Cyanoacetylene in prebiotic synthesis. *Science* **154**, 784–785 (1966). doi: 10.1126/science.154.3750.784; pmid: 5919447
6. A. Eschenmoser, E. Loewenthal, Chemistry of potentially prebiological natural products. *Chem. Soc. Rev.* **21**, 1–16 (1992). doi: 10.1039/cs9922100001
7. J. D. Sutherland, The origin of life – out of the blue. *Angew. Chem. Int. Ed.* **55**, 104–121 (2016). doi: 10.1002/anie.201506585; pmid: 26510485
8. B. H. Patel, C. Percivalle, D. J. Ritson, C. D. Duffy, J. D. Sutherland, Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat. Chem.* **7**, 301–307 (2015). doi: 10.1038/nchem.2202; pmid: 25803468
9. S. Becker *et al.*, A high-yielding, strictly regioselective prebiotic purine nucleoside formation pathway. *Science* **352**, 833–836 (2016). doi: 10.1126/science.aad2808; pmid: 27174989
10. K. B. Muchowska, S. J. Varma, J. Moran, Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019). doi: 10.1038/s41586-019-1151-1; pmid: 31043728
11. S. N. Semenov *et al.*, Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature* **537**, 656–660 (2016). doi: 10.1038/nature19776; pmid: 27680939
12. K. Ruiz-Mirazo, C. Briones, A. de la Escosura, Prebiotic systems chemistry: New perspectives for the origins of life. *Chem. Rev.* **114**, 285–366 (2014). doi: 10.1021/cr2004844; pmid: 24171674

13. W. Martin, J. Baross, D. Kelley, M. J. Russell, Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* **6**, 805–814 (2008). doi: 10.1038/nrmicro1991; pmid: 18820700
14. Y. Zhang *et al*., A semi-synthetic organism that stores and retrieves increased genetic information. *Nature* **551**, 644–647 (2017). doi: 10.1038/nature24659; pmid: 29189780
15. M. Frenkel-Pinter, M. Samanta, G. Ashkenasy, L. J. Leman, Prebiotic peptides: Molecular hubs in the origin of life. *Chem. Rev.* **120**, 4707–4765 (2020). doi: 10.1021/acs.chemrev.9b00664; pmid: 32101414
16. R. F. Ludlow, S. Otto, Systems chemistry. *Chem. Soc. Rev.* **37**, 101–108 (2008). doi: 10.1039/B611921M; pmid: 18197336
17. S. Mann, The origins of life: Old problems, new chemistries. *Angew. Chem. Int. Ed.* **52**, 155–162 (2013). doi: 10.1002/anie.201204968; pmid: 23208616
18. J. W. Szostak, D. P. Bartel, P. L. Luisi, Synthesizing life. *Nature* **409**, 387–390 (2001). doi: 10.1038/35053176; pmid: 11201752
19. S. Szymkuć *et al*., Computer-assisted synthetic planning: The end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016). doi: 10.1002/anie.201506101; pmid: 27062365
20. T. Klucznik *et al*., Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 522–532 (2018). doi: 10.1016/j.chempr.2018.02.002
21. L. E. Orgel, Self-organizing biochemical cycles. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 12503–12507 (2000). doi: 10.1073/pnas.220406697; pmid: 11058157
22. K. Molga, E. P. Gajewska, S. Szymkuć, B. A. Grzybowski, The logic of translating chemical knowledge into machine-processable forms: A modern playground for physical-organic chemistry. *React. Chem. Eng.* **4**, 1506–1521 (2019). doi: 10.1039/C9RE00076C
23. B. A. Grzybowski, Computational design and experimental validation of synthetic routes created by Chematica. *Abstr. Pap. Am. Chem. Soc.* **256**, 5 (2018).
24. T. Das, S. Ghule, K. Vanka, Insights into the origin of life: Did it begin from HCN and H_2O? *ACS Cent. Sci.* **5**, 1532–1540 (2019). doi: 10.1021/acscentsci.9b00520; pmid: 31572780
25. D. G. Blackmond, The origin of biological homochirality. *Cold Spring Harb. Perspect. Biol.* **2**, a002147 (2010). doi: 10.1101/cshperspect.a002147; pmid: 20452962
26. R. Larralde, M. P. Robertson, S. L. Miller, Rates of decomposition of ribose and other sugars: Implications for chemical evolution. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 8158–8160 (1995). doi: 10.1073/pnas.92.18.8158; pmid: 7667262
27. K. Zahnle, L. Schaefer, B. Fegley, Earth's earliest atmospheres. *Cold Spring Harb. Perspect. Biol.* **2**, a004895 (2010). doi: 10.1101/cshperspect.a004895; pmid: 20573713
28. M. Fiore, The origin and early evolution of life: Prebiotic chemistry. *Life* **9**, 73 (2019). doi: 10.3390/life9030073; pmid: 31547394
29. D. J. Ritson, J. D. Sutherland, Synthesis of aldehydic ribonucleotide and amino acid precursors by photoredox chemistry. *Angew. Chem. Int. Ed.* **52**, 5845–5847 (2013). doi: 10.1002/anie.201300321; pmid: 23610046
30. S. Islam, D.-K. Bučar, M. W. Powner, Prebiotic selection and assembly of proteinogenic amino acids and natural nucleotides from complex mixtures. *Nat. Chem.* **9**, 584–589 (2017). doi: 10.1038/nchem.2703
31. N. Friedmann, S. L. Miller, Phenylalanine and tyrosine synthesis under primitive earth conditions. *Science* **166**, 766–767 (1969). doi: 10.1126/science.166.3906.766; pmid: 5823319
32. S. A. Wildman, G. M. Crippen, Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999). doi: 10.1021/ci990307l
33. R. Albert, H. Jeong, A. L. Barabasi, Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000). doi: 10.1038/35019019; pmid: 10935628
34. A. L. Barabási, Scale-free networks: A decade and beyond. *Science* **325**, 412–413 (2009). doi: 10.1126/science.1173299; pmid: 19628854
35. M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, B. A. Grzybowski, Architecture and evolution of organic chemistry. *Angew. Chem. Int. Ed.* **44**, 7263–7269 (2005). doi: 10.1002/anie.200502272; pmid: 16276556
36. P. D. Bartlett, R. H. Jones, A kinetic study of the Leuchs anhydrides in aqueous solution. I. *J. Am. Chem. Soc.* **79**, 2153–2159 (1957). doi: 10.1021/ja01566a035
37. C. Menor-Salván, M. R. Marín-Yaseli, A new route for the prebiotic synthesis of nucleobases and hydantoins in water/ice solutions involving the photochemistry of acetylene. *Chemistry* **19**, 6488–6497 (2013). doi: 10.1002/chem.201204313; pmid: 23536286
38. C. Fernandez-Garcia, M. W. Powner, Selective acylation of nucleosides, nucleotides, and glycerol-3-phosphocholine in water. *Synlett* **28**, 78–83 (2017). doi: 10.1055/s-0036-1588626
39. T. Z. Jia *et al*., Membraneless polyester microdroplets as primordial compartments at the origins of life. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15830–15835 (2019). doi: 10.1073/pnas.1902336116; pmid: 31332006
40. R. J. P. Williams, J. J. R. Fraústo da Silva, *The Chemistry of Evolution the Development of our Ecosystem* (Elsevier, ed. 1, 2006).
41. S. Chitale, J. S. Derasp, B. Hussain, K. Tanveer, A. M. Beauchemin, Carbohydrates as efficient catalysts for the hydration of α-amino nitriles. *Chem. Commun.* **52**, 13147–13150 (2016). doi: 10.1039/C6CC07530D; pmid: 27763647
42. L. A. Paquette, E. R. Hickey, The consequences of strain release in the norbornyl subunit of isodicyclopentadiene on cycloaddition stereochemistry. Further evidence that orbital tilting serves as the key determinant of contrasteric π-facial selectivity. *Tetrahedron Lett.* **35**, 2313–2316 (1994). doi: 10.1016/0040-4039(94)85207-3
43. B. H. Lipshutz, M. C. Morey, An approach to the cyclopeptide alkaloids (phencyclopeptines) via heterocyclic diamide/dipeptide equivalents. Preparation and N-alkylation studies of 2,4(5)-disubstituted imidazoles. *J. Org. Chem.* **48**, 3745–3750 (1983). doi: 10.1021/jo00169a027
44. S. Layek, B. Agrahari, S. Kumari, D. D. Anuradha, D. D. Pathak, [Zn(L-proline)$_2$] Catalyzed one-pot synthesis of propargylamines under solvent-free conditions. *Catal. Lett.* **148**, 2675–2682 (2018). doi: 10.1007/s10562-018-2449-6
45. S. K. Guchhait, A. L. Chandgude, G. Priyadarshani, $CuSO_4$-glucose for *in situ* generation of controlled Cu(I)-Cu(II) bicatalysts: Multicomponent reaction of heterocyclic azine and aldehyde with alkyne, and cycloisomerization toward synthesis of *N*-fused imidazoles. *J. Org. Chem.* **77**, 4438–4444 (2012). doi: 10.1021/jo3003024; pmid: 22486279
46. A. Khalafinezhad, B. Mokhtari, M. Soltanirad, Direct preparation of primary amides from carboxylic acids and urea using imidazole under microwave irradiation. *Tetrahedron Lett.* **44**, 7325–7328 (2003). doi: 10.1016/S0040-4039(03)01866-5
47. A. Nakama, E.-H. Kim, K. Shinohara, H. Omura, Formation of furfural derivatives in amino-carbonyl reaction. *Biosci. Biotechnol. Biochem.* **57**, 1757–1759 (1993). doi: 10.1271/bbb.57.1757
48. S. Lymperopoulou *et al*., Synthesis, characterization, magnetic and catalytic properties of a ladder-shaped Mn^{II} coordination polymer. *Eur. J. Inorg. Chem.* **2014**, 3638–3644 (2014). doi: 10.1002/ejic.201402419
49. R. A. Sanchez, J. P. Ferris, L. E. Orgel, Studies in prebiotic synthesis. II. Synthesis of purine precursors and amino acids from aqueous hydrogen cyanide. *J. Mol. Biol.* **30**, 223–253 (1967). pmid: 4297187
50. C. Shen, L. Yang, S. L. Miller, J. Oró, Prebiotic synthesis of histidine. *J. Mol. Evol.* **31**, 167–174 (1990). doi: 10.1007/BF02109492; pmid: 11536478
51. C. Fernández-García, N. M. Grefenstette, M. W. Powner, Prebiotic synthesis of aminooxazoline-5′-phosphates in water by oxidative phosphorylation. *Chem. Commun.* **53**, 4919–4921 (2017). doi: 10.1039/C7CC02183F; pmid: 28401215
52. B. E. Leach, "Metal ion complex catalysis of amino acid esterhydrolysis," thesis, Iowa State University (1968).
53. W. Hordijk, Autocatalytic confusion clarified. *J. Theor. Biol.* **435**, 22–28 (2017). doi: 10.1016/j.jtbi.2017.09.003; pmid: 28888946
54. S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford Univ. Press, ed. 1, 1993).
55. J. L. Andersen, C. Flamm, D. Merkle, P. F. Stadler, In silico support for Eschenmoser's glyoxylate scenario. *Isr. J. Chem.* **55**, 919–933 (2015). doi: 10.1002/ijch.201400187
56. M. D. Bajczyk, P. Dittwald, A. Wołos, S. Szymkuć, B. A. Grzybowski, Discovery and enumeration of organic-chemical and biomimetic reaction cycles within the network of chemistry. *Angew. Chem. Int. Ed.* **57**, 2367–2371 (2018). doi: 10.1002/anie.201712052; pmid: 29405528
57. S. Becker *et al*., Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides. *Science* **366**, 76–82 (2019). doi: 10.1126/science.aax2747; pmid: 31604305
58. A. J. Coggins, M. W. Powner, Prebiotic synthesis of phosphoenol pyruvate by α-phosphorylation-controlled triose glycolysis. *Nat. Chem.* **9**, 310–317 (2017). doi: 10.1038/nchem.2624; pmid: 28338685
59. L. M. R. Keil, F. M. Möller, M. Kieß, P. W. Kudella, C. B. Mast, Proton gradients and pH oscillations emerge from heat flow at the microscale. *Nat. Commun.* **8**, 1897 (2017). doi: 10.1038/s41467-017-02065-3; pmid: 29196673
60. A. W. Schwartz, A. B. Voet, M. Van der Veen, Recent progress in the prebiotic chemistry of HCN. *Orig. Life* **14**, 91–98 (1984). doi: 10.1007/BF00933644; pmid: 6087243
61. S. Miyakawa, H. J. Cleaves, S. L. Miller, The cold origin of life: A. Implications based on the hydrolytic stabilities of hydrogen cyanide and formamide. *Orig. Life Evol. Biosph.* **32**, 195–208 (2002). doi: 10.1023/A:1016514305984; pmid: 12227424
62. P. A. Bachmann, P. L. Luisi, J. Lang, Autocatalytic self-replicating micelles as models for prebiotic structures. *Nature* **357**, 57–59 (1992). doi: 10.1038/357057a0
63. C. Butch *et al*., Production of tartrates by cyanide-mediated dimerization of glyoxylate: A potential abiotic pathway to the citric acid cycle. *J. Am. Chem. Soc.* **135**, 13440–13445 (2013). doi: 10.1021/ja405103r; pmid: 23914725
64. S. Santoso, W. Hwang, H. Hartman, S. Zhang, Self-assembly of surfactant-like peptides with variable glycine tails to form nanotubes and nanovesicles. *Nano Lett.* **2**, 687–691 (2002). doi: 10.1021/nl025563i
65. G. Ksander *et al*., Chemie der α-Aminonitrile 1. Mitteilung Einleitung und Wege zu Uroporphyrinogen-octanitrilen. *Helv. Chim. Acta* **70**, 1115–1172 (1987). doi: 10.1002/hlca.19870700424
66. J. Řezáč, P. Hobza, Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *J. Chem. Theory Comput.* **8**, 141–151 (2012). doi: 10.1021/ct200751e; pmid: 26592877
67. J. J. P. Stewart, MOPAC2016 (Stewart Computational Chemistry, 2016).
68. R. Roszak, W. Beker, rmrmg/tree-of-life: Tree of life, Version 1.2.1, Zenodo (2020); http://doi.org/10.5281/zenodo.4024326.
69. R. Wippermann, Ueber Tricyanwasserstoff, eine der Blausäure polymere Verbindung. *Ber. Dtsch. Chem. Ges.* **7**, 767–772 (1874). doi: 10.1002/cber.187400701244
70. L. De Vries, The evidence for generation of dimethylaminocyanocarbene in the thermolysis of dimethylaminomalononitrile. The dimethylamino(dicyano- and cyano)methyl radicals, carbon analogues of the nitroxides. *J. Am. Chem. Soc.* **100**, 926–933 (1978). doi: 10.1021/ja00471a045
71. M. J. Alves, B. L. Booth, M. F. J. R. P. Proenç, Synthesis of 5-amino-4-(cyanoformimidoyl)-1*H*-imidazole: A reactive intermediate for the synthesis of 6-carbamoyl-1,2-dihydropurines and 6-carbamoylpurines. *J. Chem. Soc. Perkin Trans. 1* **1990**, 1705–1712 (1990). doi: 10.1039/P19900001705
72. J. P. Ferris, L. E. Orgel, An unusual photochemical rearrangement in the synthesis of adenine from hydrogen cyanide. *J. Am. Chem. Soc.* **88**, 1074 (1966). doi: 10.1021/ja00957a050
73. A. Hill, L. E. Orgel, Synthesis of adenine from HCN tetramer and ammonium formate. *Orig. Life Evol. Biosph.* **32**, 99–102 (2002). doi: 10.1023/A:1016070723772; pmid: 12185678
74. J. S. Hudson *et al*., A unified mechanism for abiotic adenine and purine synthesis in formamide. *Angew. Chem. Int. Ed.* **51**, 5134–5137 (2012). doi: 10.1002/anie.201108907; pmid: 22488748
75. G. Zubay, T. Mui, Prebiotic synthesis of nucleotides. *Orig. Life Evol. Biosph.* **31**, 87–102 (2001). doi: 10.1023/A:1006722423070; pmid: 11296526
76. J. Oró, A. P. Kimball, Synthesis of purines under possible primitive earth conditions. II. Purine intermediates from hydrogen cyanide. *Arch. Biochem. Biophys.* **96**, 293–313 (1962). doi: 10.1016/0003-9861(62)90412-5; pmid: 14482339
77. P. L. Magill, Formamide. *Ind. Eng. Chem.* **26**, 611–614 (1934). doi: 10.1021/ie50294a006
78. A. Mariani, D. A. Russell, T. Javelle, J. D. Sutherland, A light-releasable potentially prebiotic nucleotide activating agent. *J. Am. Chem. Soc.* **140**, 8657–8661 (2018). doi: 10.1021/jacs.8b05189; pmid: 29965757
79. S. B. Bhosale, N. H. Naik, R. S. Kusurkar, $AlCl_3$ as an efficient catalyst toward the synthesis of 1,6-dihydropyrazine-2,3-dicarbonitrile derivatives. *Synth. Commun.* **43**, 3163–3169 (2013). doi: 10.1080/00397911.2013.769602

ACKNOWLEDGMENTS

Funding: The experimental synthetic component of this work was supported by the Symfonia Award, grant 2014/12/W/ST5/00592 from the Polish National Science Center (NCN). Development of Allchemy's Life module has been financed by Allchemy, Inc. These internal funds, in particular, supported A.W., R.R., W.B., and S.S. Data analysis by B.A.G. was supported by the Institute for Basic Science, Korea (Project Code IBS-R020-D1). **Author contributions:** A.W., R.R., S.S., and B.A.G. developed the Life module within the Allchemy platform. A.Z.-D. executed the autocatalytic IDA cycle, synthesis of uric acid, and syntheses indicated in Fig. 3C. W.B. performed most of the statistical

analyses. B.M.-K. performed syntheses of CA and formaldehyde-catalyzed selective hydrolysis. G.S. developed analytical methods for the IDA cycle, CA, and uric acid. M.D. synthesized diglycine and performed the synthesis indicated in Fig. 4C, with help from B.M.-K. and A.Z.-D., respectively. S.S. and B.A.G. conceived the project and supervised the research. All authors participated in manuscript writing. **Competing interests:** A.W., R.R., W.B., B.M.-K., S.S., and B.A.G. are contractors and/or stockholders of Allchemy, Inc. **Data and materials availability:** The precalculated networks, up to G6 and G7, are posted at https://tol.allchemy.net. The full, interactive version of the software is freely available at https://life.allchemy.net (users should create their individual accounts). The scripts developed to analyze results, search for cycles within the trees, and so forth are posted as an open-source github project at https://github.com/rmrmg/tree-of-life. The associated data repository (*68*) contains details of the G6 and G7 networks, spreadsheets of calculated molecular properties (heats of formation, logP, numbers of hydrogen bond donors and acceptors, etc.), and clusterization of abiotic molecules. This repository also houses Python scripts used for statistical analyses, as well as example commands to reproduce data used in the article (for more details, refer to the README files and help options of the scripts).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6511/eaaw1955/suppl/DC1
Materials and Methods
Figs. S1 to S139
Tables S1 to S22
References (*80*–*164*)

27 November 2018; resubmitted 28 March 2020
Accepted 24 July 2020
10.1126/science.aaw1955

Science

Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry

Agnieszka Wolos, Rafal Roszak, Anna Zadlo-Dobrowolska, Wiktor Beker, Barbara Mikulak-Klucznik, Grzegorz Spólnik, Miroslaw Dygas, Sara Szymkuc and Bartosz A. Grzybowski

Science **369** (6511), eaaw1955.
DOI: 10.1126/science.aaw1955

Mapping primordial reaction networks

Chemists seeking to understand the origins of life have published a wide range of reactions that may have yielded the building blocks of proteins, nucleic acids, and lipids from simple precursors. Wolos *et al.* scoured the literature to document each such reaction class and then wrote software that applied the reactions first to the simplest compounds such as cyanide, water, and ammonia, and then iteratively to each successive generation of products. The resulting network predicted a variety of previously unappreciated routes to biochemically relevant compounds, several of which the authors validated experimentally.

Science, this issue p. eaaw1955