Primitive purine biosynthesis connects ancient geochemistry to modern metabolism

Joshua E. Goldford[1,2,3,*,#], Harrison B. Smith[3,4,*], Liam M. Longo[3,4,*], Boswell A. Wing[5] and Shawn E. McGlynn[3,4,6,#]

[1]Division of Geophysical and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA

[2]Physics of Living Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[3]Blue Marble Space Institute of Science, Seattle, Washington, USA 98154

[4]Earth-Life Science Institute, Tokyo Institute of Technology, Meguro, Tokyo, Japan 152-8550

[5]Department of Geological Sciences, University of Colorado, Boulder, CO 80309, USA

[6]Biofunctional Catalyst Research Team, RIKEN Center for Sustainable Resource Science, Wako, Saitama, 351-0198, Japan

[*]Co-lead authors

[#]To whom correspondence should be addressed: goldford@caltech.edu, mcglynn@elsi.jp

Running Title: Punctuated evolution of metabolism

Keywords: Ancient metabolism, purine biosynthesis, autocatalysis, metabolic evolution

Abstract

A major unresolved question in the origin and evolution of life is whether a continuous path from geochemical precursors to the majority of molecules in the biosphere can be reconstructed from modern day biochemistry. Here we simulated the emergence of ancient metabolic networks and identified a feasible path from simple geochemically plausible precursors (e.g., phosphate, sulfide, ammonia, simple carboxylic acids, and metals) using only known biochemical reactions and models of primitive coenzymes. We find that purine synthesis constitutes a bottleneck for metabolic expansion, and that non-autocatalytic phosphoryl coupling agents are necessary to enable expansion from geochemistry to modern metabolic networks. Our model predicts punctuated phases of metabolic evolution characterized by the emergence of small molecule coenzymes (e.g., ATP, NAD^+, FAD). Early phases in the resulting expansion are associated with enzymes that are metal dependent and structurally symmetric, supporting models of early biochemical evolution. This expansion trajectory produces distinct hypotheses regarding the timing and mode of metabolic pathway evolution, including a late appearance of methane metabolisms and oxygenic photosynthesis consistent with the geochemical record. The concordance between biological and geological analysis suggests that this trajectory provides a plausible evolutionary history for the vast majority of core biochemistry.

Introduction

Modern metabolism evolved as a consequence of nascent life deriving material and energy from the surrounding geochemical environment [1–3]. However, the transition from geochemistry to extant biochemistry is poorly understood, due in part to a great uncertainty in the structure of ancient metabolic networks [4]. In particular, chemical reactions that are unrelated to biochemistry have been invoked as missing steps in early biosynthetic pathways [5–7], suggesting that records of these chemical transformations were lost throughout the history of evolution; for example, through the emergence and sophistication of protein catalysts. Nevertheless, it is unclear to what degree ancient metabolism has been lost, and whether intermediate stages can be excavated from the extant biosphere.

Many lines of evidence suggest that recovering continuity between ancient geochemistry and extant biochemistry might be impossible without the inclusion of a vast number of abiotic chemical reactions unrelated to modern biology. First, a recent analysis of metabolic networks revealed a high prevalence of autocatalytic subnetworks in which the generation of several key biomolecules, including many coenzymes, are required for their own synthesis [8,9]. Although autocatalysis may have been a necessary feature of early evolutionary processes [8,10–12], the widespread occurrence of such network motifs presents a problem for the initial emergence of ancient metabolism if there are no other routes for biosynthesis. Second, high rates of species extinction, horizontal gene transfer, non-orthogonal displacement, and evolutionary forces like drift could have eroded an early record of ancient biochemistry throughout the course of Earth's history [13–16]. Lastly, several recent studies simulating the emergence of metabolic networks from geochemistry only recovered small or fragmented networks, on the order of 10% of contemporary biochemistry [17–20]. Likewise, models of ancient metabolism with hypothetical non-phosphate alternatives [18,19] drastically limited the potential coverage of modern day metabolic networks, which rely heavily on phosphate-containing molecules. Taken together, it remains unclear to what extent "extinct" biochemistry is necessary to enable the generation of modern metabolism from early Earth environments.

Beyond the question of continuity, constructing a model for the emergence of biochemical networks enables us to address key evolutionary questions. For example, resolving the order that biochemical reactions emerge in time can inform metabolic pathway evolution more broadly, including their relative ages and potential influence on geochemical and isotopic 'biosignatures' in the geologic record. Models of metabolic pathway evolution have invoked several mechanisms, ranging from sequential models, where reactions emerge in the order they appear in the reaction pathway, to mosaic models, where the order of reaction emergence is decoupled from the order of reactions in the extant pathway [21–26 25,27]. Although studies have shown support for various models of evolution for specific pathways [27], a broad, biosphere-scale analysis of the relative occurrence of various modes of metabolic pathway evolution is lacking. Additionally, knowledge of the relative ordering of metabolic pathways that mediate biogeochemical cycling, such as carbon fixation[28], can support efforts to interpret isotopic signatures in the geologic record.

Here, we construct a biosphere-level model of metabolic evolution and show that a single autocatalytic bottleneck in purine synthesis prevents the emergence of metabolism from geochemical precursors. We show that including a hypothetical ATP-independent pathway for purine biosynthesis enables the continuous expansion of metabolism from simple starting material, and that the ensuing trajectory of metabolic network evolution is correlated with features typically associated with the transition from ancient to modern biochemistry. We use this trajectory to resolve key aspects on the nature of metabolic evolution, with a focus on elucidating the mechanisms and order by which metabolic pathways emerged in the biosphere.

Results

Primitive purine production enables expansion to modern biochemistry

To construct a model of the evolutionary history of metabolism at the biosphere scale, we compiled a database of 12,263 biochemical reactions from the KEGG database (Table S1-3, Methods) [29]. Unlike prior studies [18,19], we added detailed organic and inorganic cofactor dependencies for 5,259 reactions from UniProt, Expasy, PDBe, and EBI into the network (Table S1). These dependencies range from inorganic metal ions (e.g., Fe, Mn) to organic molecules

110 (e.g., flavins, quinones) involved in catalysis (see Methods). Using this network, we performed
111 network expansion (Methods, [17–20,30–32]) starting from a set of "seed compounds". Our seed
112 compounds included metals and inorganic material (e.g., Fe, Mn, Zn), CO_2, hydrogen sulfide,
113 molecular hydrogen, orthophosphate, ammonia, and 19 organic substrates that can be produced
114 abiotically with iron, pyruvate, and glyoxylate. The choice of seed set compounds implicitly
115 assumes that ancient carbon fixing reactions, similar to those found in the reductive tricarboxylic
116 acid (rTCA) cycle and reductive acetyl-CoA pathway, are capable of producing simple
117 carboxylic acids from CO_2 and reductants like H_2 [33–38] (Table S4 and see supplemental text on
118 succinate semialdehyde). Consistent with previous studies [18,19], we could generate a network of
119 429 compounds from diverse pathways in central metabolism, including amino acid biosynthesis
120 and some simple organic coenzymes like PLP (Fig. 1a, black line, Table S5). However, as our
121 biosphere-level network consists of >8000 compounds, this scope only constitutes ~5% of all
122 known biochemicals, leaving the vast majority of molecules unreachable from simple seed
123 compounds. Although the inclusion of primitive thioester energy coupling mechanisms and
124 reductants may have been important during the early stages of biochemical evolution [18,19], these
125 modifications marginally increased the scope of the expansion (n=734, Fig. 1d). Hypothesizing
126 that our results were biased and limited by the inclusion of only cataloged biochemical reactions,
127 we explored the possibility that unknown biochemical reactions could enable a more extensive
128 expansion. To investigate this possibility, we included reactions from a database of hypothetical
129 biochemistry [39,40], which added 20,183 new reactions to our network and increased the total size
130 by a factor of ~2.7. Repeating the expansion with this expanded reaction set resulted in only a
131 slight increase in scope to 472 compounds (Extended Data Fig. 1), suggesting that neither
132 currently cataloged nor predicted biochemistry contain transformations required to reach the vast
133 majority of known metabolites.
134
135 Notably, phosphoribosyl pyrophosphate (PRPP), a key precursor to metabolite classes like
136 purines, was not in the expansion scope, suggesting that a bottleneck in purine production limits
137 expansion. Indeed, the addition of adenine to the seed set resulted in a network of 4315
138 compounds, or ~50% network coverage (Fig. 1a, gray line), including all major coenzymes
139 (ATP, NAD, CoA, SAM, flavins, pterins, quinones, and heme; Table S5). To test whether
140 purines were uniquely essential for the expansion to larger networks, we conducted a "rescue"

experiment: for each of the 8314 single compounds from KEGG that were not produced in the original expansion scope (Fig. 1a, black line), we added one to the original seed set and repeated the expansion. We found that only 217 compounds yielded expansion to >4000 compounds, and all of these compounds either contained a purine moiety or were intermediates in *de novo* purine biosynthesis (Fig. 1c).

De novo purine production in extant biochemistry exhibits an autocatalytic dependence in the production of adenosine triphosphate (ATP) and PRPP, due to multiple steps requiring ATP as a phosphorylating agent (Fig. 1b)[8,9,41,42]. This autocatalytic dependence may have been relaxed in primitive metabolism, where primitive phosphorylating agents could have been either polyphosphate or acyl-phosphate (e.g., acetyl phosphate) [43–45]. We explored this hypothesis by substituting 8 ATP-coupled reactions involved in purine synthesis with polyphosphate-coupled variants (Methods, Fig. 1b). We repeated the expansion of the original seed set with this modified network and found that expansion led to networks consisting of >4000 molecules with simple high-energy phosphoryl donors (Fig. 1a, blue line, Extended Data Fig. 2). To further explore the sensitivity of the expansion to variations in the seed set, we performed additional simulations where we first varied the composition of organic molecules, nitrogen, and sulfur sources and then re-ran the expansion algorithm. We found that expansion to networks with >4000 compounds required fully reduced nitrogen (e.g., ammonia or glycine) and relatively oxidized carbon source (Extended Data Fig. 3). Taken together, this model demonstrates that expansion from reduced nitrogen and oxidized carbon to most of primary metabolism requires only a small number of additional phosphate-coupled reactions involved in *de novo* purine biosynthesis.

Punctuated metabolic innovation is gated by coenzyme synthesis

Along the primitive purine biosynthesis expansion trajectory (Fig. 1a, blue line), compound addition occurs in bursts, representing periods where large numbers of metabolites were added simultaneously. This observation motivated us to explore the hypothesis that production of key intermediate molecules are bottlenecks throughout the expansion. To test this hypothesis, we removed a single metabolite (and all associated reactions) from the network, performed network expansion on the perturbed network, and then repeated this procedure for each metabolite in the

expansion scope (Fig. 2a). Of the 4,248 compounds produced during the expansion, only the removal of 149 led to scopes less than 4000 compounds, indicating that the vast majority of perturbations had little effect on the network size. By plotting the size of the network perturbed by removing a compound (*y*-axis) versus the iteration where the compound was made in the original expansion (*x*-axis), we found that removing certain compounds produced networks of similar size to networks generated before the compounds were produced in the unperturbed expansion (Fig. 2a, red dots). This result suggests that these compounds represent a specific class of "bottleneck" molecules that restrict expansion and are made exactly when they are essential for further expansion of the network (Fig. 2a, red dots). Interestingly, several of these molecules were coenzymes like PLP, ATP, NAD, CoA, and thiamine diphosphate (ThDP), or intermediates in the biosynthesis of key coenzyme classes like isoprenoids (isopentenyl diphosphate).

As noted in previous studies, the inclusion of various coenzyme classes has a large effect on metabolic network expansion [17,30], motivating the hypothesis that the observed peaks were induced by the production of key metabolites. To explore this possibility, we simultaneously added 8 key coenzymes (ATP, NAD, CoA, PLP, FAD, SAM, ThDP, and cobalamin) as seed molecules and re-ran the expansion. Including these coenzymes collapsed the multiple peaks observed in Fig. 1a into just two peaks, where the later peak corresponds to the production of molecular oxygen (Fig. 2b). We further found that addition of these 8 coenzymes to the seed set enabled expansions with fewer iterations relative to randomly chosen sets of 8 compounds with or without oxygen (Monte Carlo permutation test: $P < 10^{-3}$). Altogether, this analysis shows that the production of major coenzymes are key factors shaping the observed peaks during expansion.

Ancient features of compounds and reactions are correlated with expansion iteration

The last peak shown in Fig. 1a corresponds to the production of molecular oxygen after the emergence of quinones used in modern-day photosystems (Supplemental Text). Since O_2 production was only thought to emerge ~2.4 Gya, after the rise of cyanobacteria [46,47], it is tempting to hypothesize that expansion iteration number may, to a first approximation, represent time. To investigate this possibility, we computed metrics associated with ancient biochemistry and examined how these metrics changed throughout the expansion process.

As suggested recently [48], the complexity of molecules (as well as their abundance) might be a useful metric to quantify the degree of evolution of the biosphere. To explore this possibility, we first computed the Bertz complexity of each molecule with a defined molecular structure that was produced during the expansion (n=3588, Table S6). For each iteration of the expansion, we identified all the compounds produced at that iteration and computed the mean Bertz molecular complexity. The expansion iteration number was positively correlated with mean molecular complexity (Fig. 3a, Pearson's $r = 0.72$, $P < 10^{-16}$) consistent with the hypothesis of increasing molecular complexity during metabolic evolution. This trend was also observed with other measures of molecular complexity, and large expansions were only observable with relatively simple seed sets compared to more complex seed sets (Extended Data Fig. 4). The mean molecular complexity is correlated with expansion iteration even when controlling for the average redox state of molecular carbon (Partial correlation: $r_{partial} = 0.70$, $P < 10^{-15}$).

The proportion of extant prokaryotic genomes using compounds produced during each expansion step decreased throughout the expansion (Fig. 3b, Pearson's $r = -0.83$, $P < 10^{-24}$, see Methods), consistent with the hypothesis that the most ancient parts of metabolism are the most conserved across the tree of life. More broadly, we found that this network is biased for compounds in pathways that are frequently observed in prokaryotic genomes and, on average, compounds in this network are found in 55% of prokaryotic genomes, compared to just 10% for KEGG compounds not included in the expansion scope (Extended Data Fig. 5). At each iteration, we computed the proportion of reactions that are exclusively found in eukaryotic genomes (see Methods). Interestingly, the proportion of reactions unique to eukaryotes increases towards the end of the expansion (Extended Data Fig. 6a). The selective removal of eukaryote-specific reactions (n=1638) led to a network of 3806 compounds (Extended Data Fig. 6b), which is larger than expansions performed after the removal of an equal number of randomly selected reactions (Monte Carlo Permutation test: $P<10^{-3}$).

Models of ancient metabolic evolution suggest that before the evolution of the modern genetic coding system, "pre-enzymatic" catalysts such as minerals, metal ions, peptides or short RNA polymers served as catalysts for the majority of metabolic reactions [34,49–52]. It has been proposed that relics of these pre-enzymatic catalysts may be retained in the active sites of many extant

232 enzymes as metal or iron-sulfur cofactors [4,33,53]. As Archean oceans became increasingly sulfidic
233 and/or oxidized [54], the geochemical supply of various metals like iron may have decreased. This
234 decline in the solubility of iron-bearing minerals may be reflected in the increased deposition of
235 iron formations after ~3 billion years ago [55] and may have provided a selective pressure for new
236 enzymes to be less dependent on specific metals and iron-sulfur cofactors. We thus computed the
237 proportion of reactions at each iteration that were metal cofactor-dependent (metal cofactors are
238 listed in Table S8), and observed that this proportion generally decreased as expansion iteration
239 increased (Fig. 3c, Pearson's $r = -0.75$, $P < 10^{-18}$), consistent with the idea that later networks
240 were less dependent on metal-catalyzed reactions.
241
242 Beyond molecular complexity, taxonomic associations, and metal-dependence in reaction
243 mechanisms, specific types of protein folds have been hypothesized to have served as ancient
244 catalysts in primitive biochemistry [56–59]. In particular, it has been argued that
245 repetitive/symmetric folds may have been among the first protein folds used in metabolism [60],
246 due to their ability to encode complex structures within a short, oligomerizing peptide [61–64] and
247 their ubiquitous association with diverse metabolic processes [56,65,66]. Assuming ancient protein
248 folds had more opportunity to become catalysts for biochemical reactions than more recent
249 protein folds, we hypothesized that iteration number should correlate with enzyme fold
250 repetitiveness, a putative correlate of fold age. To explore this possibility, we computed a
251 structural repetitiveness score for the set of protein folds associated with a reaction [67], and the
252 mean repetitiveness score at each expansion iteration (Table S9, Methods). Consistent with our
253 expectation, we observed a significant decrease in mean repetitiveness throughout the expansion
254 trajectory (Fig. 3d, Pearson's $r = -0.70$, $P < 10^{-15}$). This association is still observed when
255 limiting our analysis to reactions dependent on a single enzyme family (Pearson's $r = -0.49$, $P <$
256 10^{-6}). An alternative explanation for this trend is the potential association between enzyme and
257 substrate complexity. We thus computed the partial correlation between structural repetitiveness
258 and expansion iteration, controlling for molecular complexity, and still observed a strong
259 decrease in repetitiveness throughout the expansion trajectory (Pearson's partial correlation $r_{partial}$
260 $= -0.72$, $P < 10^{-15}$).
261
262 Altogether, these analyses show that many features of ancient metabolism – including molecular

complexity metrics, metal-cofactor dependencies, and protein fold attributes – are associated with expansion iteration, consistent with the hypothesis that expansion iteration can be viewed as a first-order approximation of time along the trajectory of metabolic evolution.

The modes and order of emergence of extant biochemical pathways

The trajectory of biochemical network evolution permitted us to explore fundamental questions in the history of metabolism, including how metabolic pathways evolve and the specific ordering of distinct metabolic pathways. We first sought to determine the degree to which extant linear metabolic pathways evolved sequentially (either in the forward or reverse direction, see Methods), or non-consecutively, where pathway steps are realized at various times, independent of pathway structure (i.e., the mosaic model [22–24]). Sequential models of metabolic evolution include the forward direction (also called the Granick, or prograde model) [68], where reactions emerge following the proceeding reaction in the metabolic pathway, and the reverse direction, where reactions are added sequentially in a "retrograde" fashion [69]. For the set of linear metabolic pathways in the KEGG database reachable after expansion (n=110, Table S10), we computed the iteration where each pathway step was feasible and classified these pathways based on the Spearman's rank correlation between iteration number and when each step was feasible (Methods, Fig. 4a). While we found that that ~13% (15/110) of pathways adhere to a retrograde model of pathway evolution, including core pathways like cysteine biosynthesis from serine (Table S10), and ~49% (54/110) follow a prograde model, a large fraction (41/110) of these pathways do not emerge sequentially, but rather emerge in a discontinuous fashion, consistent with a mosaic model of metabolic evolution.

We next sought to order the emergence of specific extant metabolic pathways. We constructed logical relationships between reactions and pathways accounting for redundancies and co-dependencies (see Methods), and computed the iteration number where each of 221 KEGG modules were feasible (Table S11, see Methods). Pathway emergence is distributed throughout the entire expansion (Fig. 4b). For example, before ATP is generated in the network, only the reductive and non-oxidative pentose phosphate modules, PLP biosynthesis, and cysteine biosynthesis are fully enabled (Table S11). In order to make specific predictions testable with the geologic record, we analyzed the ordering of pathways involved in key biogeochemical

cycling in the biosphere, such as carbon fixation and methane metabolism. Surprisingly, of the seven known carbon fixation pathways, the Calvin cycle was predicted to become feasible first, preceding both the reductive tricarboxylic acid cycle (rTCA) cycle and the reductive acetyl-CoA pathway, which have both been predicted to be operative in the last universal common ancestor (Fig. 4c) [70–72]. While the relative order between the rTCA and Calvin cycle was dependent on the seed set composition, we found that the extant Calvin cycle emerged before the extant reductive acetyl-CoA pathway in all simulations with random seed sets (Extended Data Fig. 7). Consistent with this observation, we found that pathways involved in archaeal methane metabolism also emerged after the Calvin cycle and rTCA cycle (Fig. 4d), due to the heavy reliance on pterin cofactors. While methylotrophic and hydrogenotrophic variants of methanogenesis have both been proposed as being ancestral in ancient archaea [73], network expansion predicts that methylotrophic methanogenesis preceded the hydrogenotrophic variant. Fractionations of ^{13}C from ^{12}C in 2.9-2.7 billion year old sediments [74] are consistent with the operation of methanogenesis [75] and/or acetogenesis [76], both of which rely on the reductive acetyl-CoA pathway today. However, the global distribution of older samples (from 3.0 to 4.1 billion years) display C-isotope fractionations akin to those generated by the modern day Calvin cycle [74]. While we cannot rule out alternative metabolic explanations for these isotopic patterns (see Supplemental Note), our metabolic network evolution model, like these geological data, suggests the emergence of the Calvin cycle preceding the modern variant of the reductive acetyl-CoA pathway.

If the Calvin cycle developed early in biosphere evolution, an enzyme with ribulose-1,5-bisphosphate carboxylase (RuBisCo) activity must also have appeared early. While Form I RuBisCo is confined to cyanobacteria, eukaryotic algae, and some proteobacteria, other RuBisCo forms (e.g., Forms II, II/III, III) and RuBisCo-like proteins (RLPs) are more widespread and possibly ancestral [77], suggesting that the earliest variants of the Calvin cycle could have used these variants for carboxylation.

Discussion

Continuity between early Earth geochemistry and modern metabolism is a necessary condition for any theory of the emergence and evolution of the biosphere on Earth. Whereas demonstrating a continuous trajectory does not rule out possibly critical, yet extinct chemistry (for example [5,6]),

323 it does relax the requirement to invoke such transformations as essential for biochemical
324 evolution. Furthermore, whereas many other feasible trajectories may enable this observed
325 continuity, the trajectory in Fig. 1a recapitulates several features associated with time and Earth
326 history. Our model suggests that, while species extinction is likely ubiquitous throughout time,
327 extinction of primary metabolic biochemistry at the biosphere scale may in fact be quite rare.
328 With an approximate time ordering of metabolic exploration, we can explicitly probe the
329 ordering of extant metabolic pathways and key biochemical innovations in the biosphere[78,79]. It
330 is important to note, however, that our model can only predict the timing of extant biochemical
331 pathways and not necessarily their primitive analogues. Our analysis suggests that the evolution
332 of the extant rTCA cycle may have been limited by the emergence of cofactors like CoA and
333 thiamine pyrophosphate. Importantly, our choice of seed compounds suggest that ancient
334 analogues of carbon fixation pathways similar to the rTCA and reductive acetyl-CoA pathways
335 produced many important carboxylic acids, as explored in previous theoretical and experimental
336 work[18,19,25,35–38,71].
337
338 The peaks of metabolic expansion may be characteristic of evolving metabolic networks. Indeed,
339 we found that this punctuated structure is robustly observed when perturbing the initial seed set
340 composition (Extended Data Fig. 4), when we include a database of predicted extant yet
341 undiscovered biochemistry (Extended Data Fig. 1), and when we remove eukaryote-specific
342 biochemical reactions (Extended Data Fig. 6). However, the particular structure and trajectory in
343 Fig. 1a may be sensitive to seed set molecule composition or network structure (Fig. 2b,
344 Extended Data Fig. 3,4 and 7). In contrast, the emergence of metabolic pathways appears to
345 occur with minimal punctuated structure (Fig. 4b), highlighting key differences between the
346 shape of metabolic evolution at the scale of individual reactions and compounds versus
347 pathways. A large fraction of metabolic pathways also appear to occur as a coalescence of
348 biochemistry emerging at different times throughout the expansion, suggesting the "mosaic"
349 model of metabolic evolution might be a major mode of evolution. The core metabolic pathways
350 in extant species might have emerged relatively late, with primitive metabolic functions enabled
351 instead by biochemical transformations from disparate metabolic pathways, like previously
352 proposed ancient carbon fixation pathways [19,71] [34,71,80]. Moreover, it has been suggested that

modern biochemical pathways may contain relics of such an evolutionary model, due to the repetitiveness of their chemical reaction sequences [81].

Although improvements in biochemical annotation and thermodynamic parameter estimation will undoubtedly change details of the trajectory presented in this paper (see Supplemental Information), our analysis suggests that adding an ATP-independent purine biosynthesis pathway to the known catalog of extant biochemistry is necessary to enable the continuous trajectory from geochemical precursors to the majority of core metabolism. Efforts to model metabolic network evolution from geochemically-plausible molecules were unable to generate purines, despite extensive inclusion of non-catalogued hypothetical biochemistry like thioester-coupling [18,19]. While abiotic synthetic routes to purines are well studied both experimentally [82,83] our analysis suggests that these abiotic mechanisms may have not been necessary if analogues to modern-day phosphate-coupling mechanisms were accessible during the development of the biosphere [44].

Prior studies have uncovered several autocatalytic compounds enriched in autocatalytic sets at both organismal and biosphere scales [8,9,84], including ATP, PLP, and NAD. Contrary to what one might expect from these studies, we find purine biosynthesis is the only autocatalytic pathway that limits expansion of biosphere-scale metabolism. Compared to the wide occurrence of strict autocatalytic dependencies at the genome-level [42,84], there appears to be a paucity of obligate autocatalytic dependencies in the biosphere-level metabolism. This observation is consistent with the hypothesis that autocatalytic pathways are strongly selected for at the organismal scale [85,86], yet primitive, non-autocatalytic pathways can be reconstructed by analyzing biochemistry at higher scales of biological organization. These results suggest that biosphere-level metabolic networks, unlike organismal-scale metabolism, may be the appropriate scale to interrogate evolutionary history of metabolism in deep time[8,17,18,32,80]. An intriguing possibility is that this single autocatalytic dependency observed here is due to the incomplete characterization of extant biochemical diversity, and that fragments of an ancient, non-autocatalytic polyphosphate-dependent *de novo* purine synthesis pathway remains hidden in the biosphere.

Acknowledgements

The authors thank Woodward Fisher, Avi Flamholz, and Joan Valentine for valuable discussions. J.E.G. is supported by the Gordon and Betty Moore Foundation as Physics of Living Systems Fellows through grant number GBMF4513, as well as the Simons Foundation. S.E.M. and B.A.W. acknowledge support from NSF Awards 1724300 and 1724393(Collaborative Research: Biochemical, Genetic, Metabolic, and Isotopic Constraints on an Ancient Thiobiosphere). S.E.M. additionally acknowledges support from JSPS KAKENHI (Grant Nos. JP18H01325 and 22H01343). H.B.S acknowledges support from JSPS KAKENHI Grant No. JP19K23459.

Contributions

All authors designed the research. J.E.G, H.B.S. and L.M.L. prepared data. J.E.G. and H.B.S. wrote code and ran simulations. J.E.G. performed analysis. J.E.G., H.B.S. and L.M.L wrote the manuscript. All authors read and approved the final manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

Corresponding Authors

Correspondence to: goldford@caltech.edu, mcglynn@elsi.jp

Figures

Figure 1

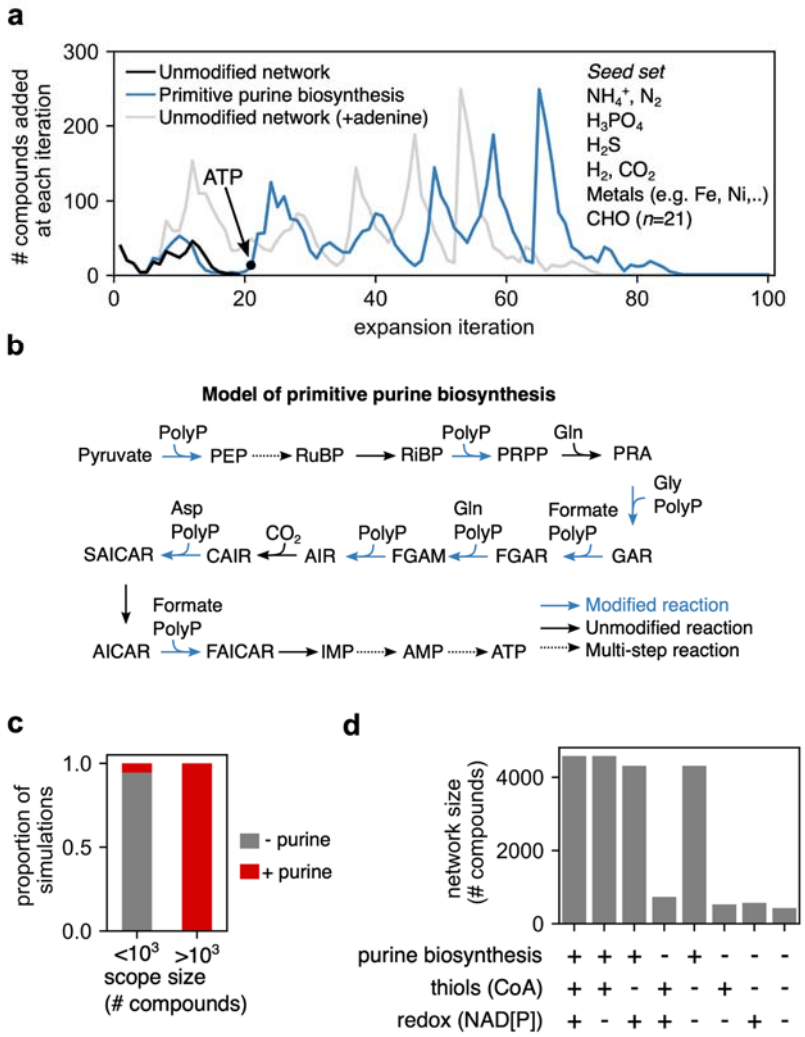

Fig. 1: Primitive purine biosynthesis enables continuity between prebiotic seed molecules and contemporary biochemistry. (a) We performed network expansion with compounds previously hypothesized to have been highly abundant on early Earth (e.g., H_2S, NH_4^+, CO_2, Fe, Mn), as well as molecules producible in simple experiments from pyruvate, glyoxylate, and iron [35] (see Table S4 for complete list). For each simulation, we plotted trajectories of the number of compounds added (*y*-axis) at each expansion iteration (*x*-axis), with either an unmodified network without adenine (black line), with adenine (gray line) or a modified network with a model of a primitive polyphosphate (PolyP)-dependent purine biosynthesis pathway (blue line). The addition of adenine enabled the synthesis of ATP via the reversal of AMP:phosphate (5-phospho-alpha-D-ribosyl)transferase. (b) A model for primitive production of purines. We propose that all ATP-dependent steps in *de novo* purine biosynthesis were preceded by

polyphosphate (PolyP) dependent steps (blue, [43]). Full names for the abbreviations of these compounds are listed in the Materials and Methods. (c) For 8314 compounds, we added each as additional molecules in the seed set, repeated the expansion, and found that compounds containing purine moieties were required for expansion (see Methods). (d) We repeated expansion with models of primitive purine biosynthesis, primitive thioester coupling, or primitive redox systems (in lieu of NAD(P)). While networks increased slightly with primitive redox and thiol coupling, expansion to the >4000 compounds required primitive purine biosynthesis.

Figure 2

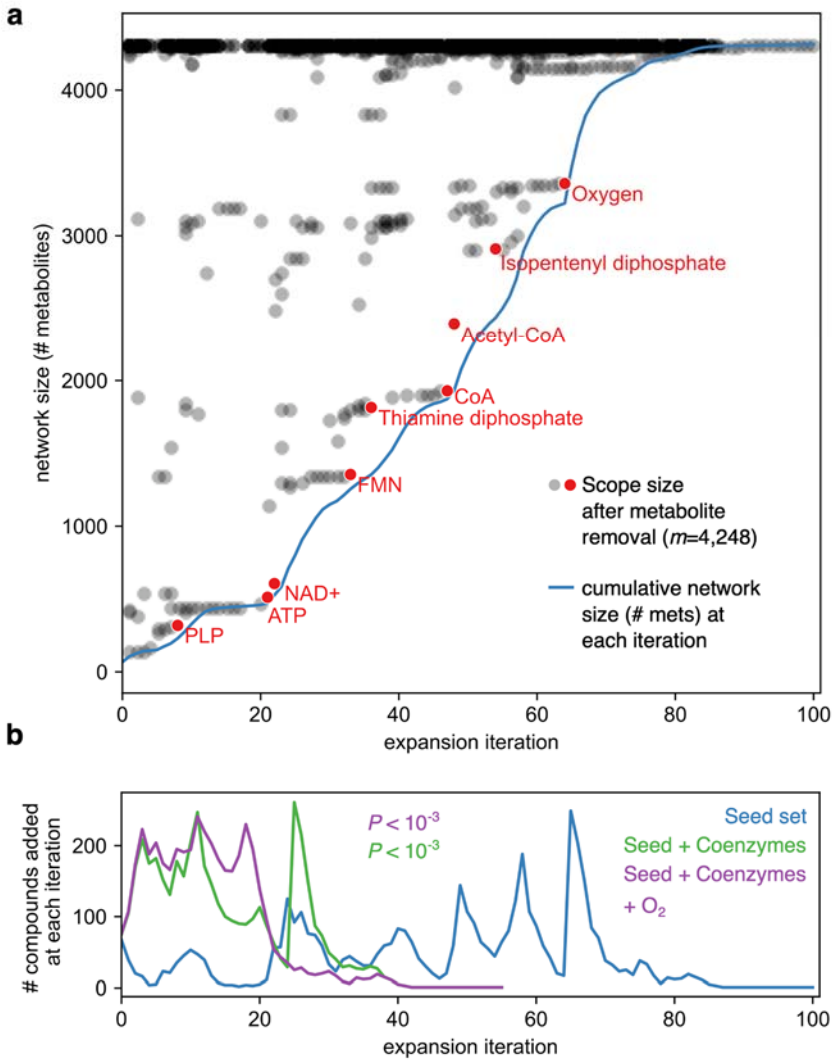

Fig. 2: Peaks during metabolic expansion are associated with coenzyme and oxygen production. (a) For each compound in the expansion scope (*m*=4,248), we removed the compound and all associated reactions from the network used in Fig. 1 and performed expansion. For each compound, we plotted the iteration when the compound is formed in the original expansion (*x*-axis) versus the scope size achieved in the perturbed expansion (*y*-axis). As a reference, we also plotted the cumulative network size (number of compounds) after each iteration for the unperturbed expansion (blue line). Red points are examples of important compounds that become limiting for the expansion close to when they are produced, suggesting they are bottlenecks in the expansion. (b) Trajectories of compounds added at each iteration during network expansion with the base seed set (*n*=70, blue line), base seed set with eight additional coenzymes (ATP, NAD, CoA, PLP, FAD, SAM, ThDP, and Cobalamin), (*n*=78, green line), and the base seed set with additional coenzymes and molecular oxygen (*n*=79, purple line). We measured the iteration where 90% of the network was recovered, and found that this value was significantly less than simulations with eight randomly selected compounds as additional seed molecules (Monte Carlo permutation test: $P < 10^{-3}$).

Figure 3

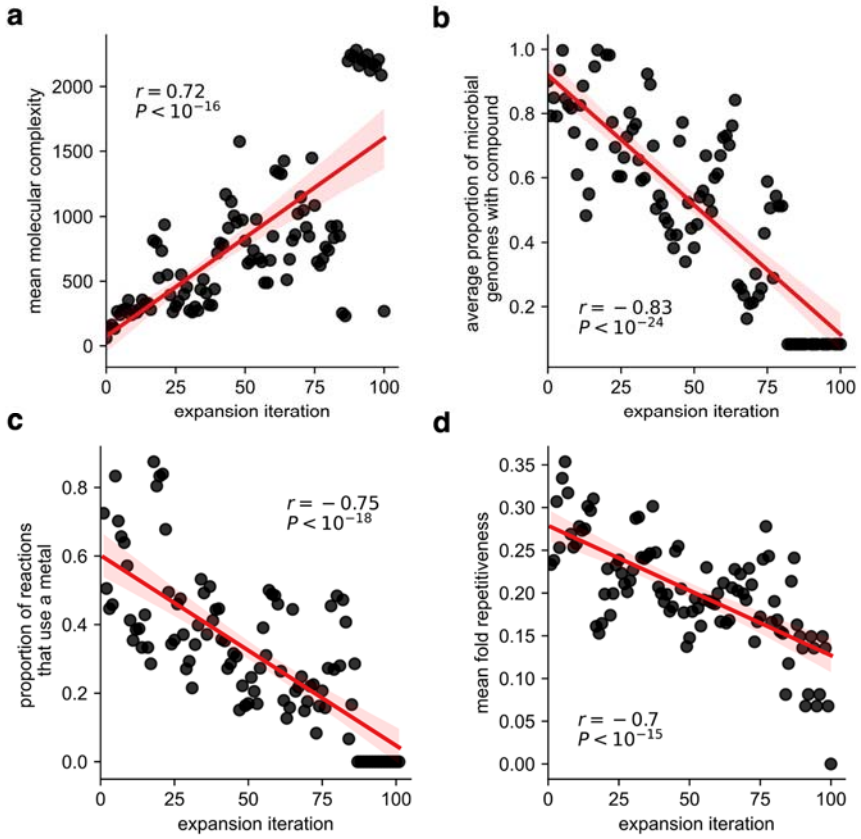

Fig. 3: Ancient biochemical features are correlated with expansion iteration. We computed various biochemical features previously hypothesized to be associated with ancient biochemistry at each iteration during the expansion (*x*-axis, a-d). We computed (a) the average Bertz molecular complexity of molecules produced at each expansion step (*y*-axis), (b) the average proportion of microbial genomes using each compound produced during each expansion step (*y*-axis), (c) the proportion of reactions that use an enzyme dependent on a metal cofactor (*y*-axis), and (d) the mean repetitiveness for protein folds associated with each reaction utilized at each expansion iteration (*y*-axis) (see Methods).

Figure 4

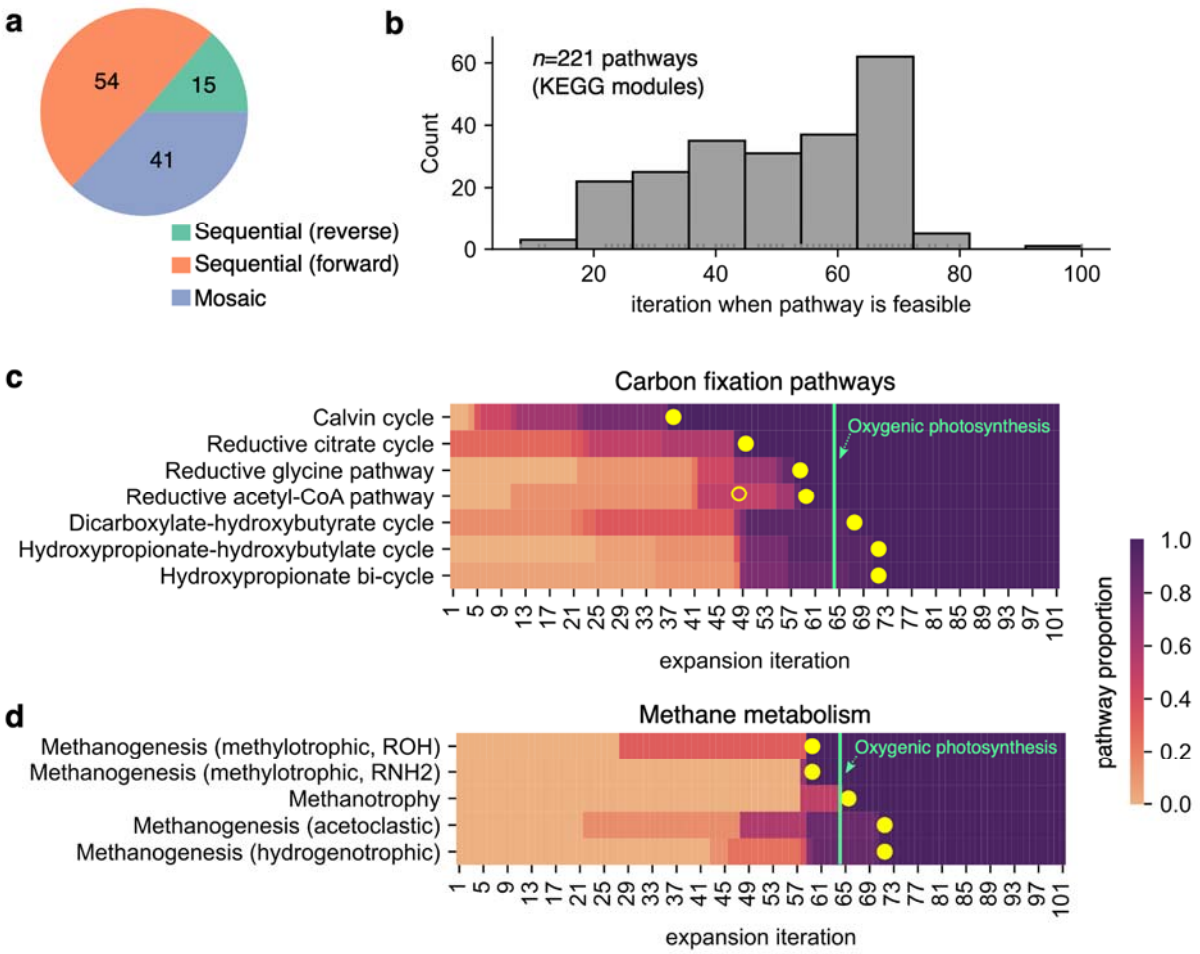

Fig. 4: Timing the emergence of metabolic pathways. (a) For 110 linear metabolic pathways in the KEGG database (Table S10), we classified each pathway as emerging sequentially (either via the forward or reverse direction) or non-sequentially (i.e., Mosaic, blue). We found that 37% (41/110) of these pathways emerged non-sequentially and ~13% (15/110) adhere to a retrograde model of pathway evolution. (b) For 221 KEGG modules, we computed and plotted the distribution of the minimum expansion iteration when each module became feasible (Table S11). (c) A heatmap showing the proportion of feasible steps (color) of seven carbon fixation pathways (*y*-axis) throughout the expansion (*x*-axis). The closed yellow dots represent the iteration when

all steps in the pathway are feasible. The open yellow circle denotes the emergence of the carbonyl-branch of the reductive acetyl-CoA pathway. (d) The same as (c) but for pathways involved in methane metabolism. The green vertical line in both subplots (c) and (d) denote the iteration when oxygenic photosynthesis emerged.

References

1. Smith, E. & Morowitz, H. J. *The Origin and Nature of Life On Earth*. 677 (Cambridge University Press, 2016).
2. Smith, E. & Morowitz, H. J. Universality in intermediary metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13168–13173 (2004).
3. Hartman, H. Speculations on the origin and evolution of metabolism. *J. Mol. Evol.* **4**, 359–370 (1975).
4. Goldford, J. E. & Segrè, D. Modern views of ancient metabolic networks. *Current Opinion in Systems Biology* (2018).
5. Patel, B. H., Percivalle, C., Ritson, D. J., Duffy, C. D. & Sutherland, J. D. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat. Chem.* **7**, 301–307 (2015).
6. Xu, J. *et al.* Selective prebiotic formation of RNA pyrimidine and DNA purine nucleosides. *Nature* **582**, 60–66 (2020).
7. Yadav, M., Pulletikurti, S., Yerabolu, J. R. & Krishnamurthy, R. Cyanide as a primordial reductant enables a protometabolic reductive glyoxylate pathway. *Nat. Chem.* **14**, 170–178 (2022).
8. Xavier, J. C., Hordijk, W., Kauffman, S., Steel, M. & Martin, W. F. Autocatalytic chemical networks at the origin of metabolism. *Proc. Biol. Sci.* **287**, 20192377 (2020).
9. Xavier, J. C. & Kauffman, S. Small-molecule autocatalytic networks are universal metabolic fossils. *Philos. Trans. A Math. Phys. Eng. Sci.* **380**, 20210244 (2022).
10. Kauffman, S. A. *The origins of order : self-organization and selection in evolution*. 709 (Oxford

University Press, 1993).

11. Kauffman, S. A. Autocatalytic sets of proteins. *J. Theor. Biol.* **119**, 1–24 (1986).

12. Blokhuis, A., Lacoste, D. & Nghe, P. Universal motifs and the diversity of autocatalytic systems. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25230–25236 (2020).

13. Petchey, O. L. & Gaston, K. J. Extinction and the loss of functional diversity. *Proc. Biol. Sci.* **269**, 1721–1727 (2002).

14. Marshall, C. R. Five palaeobiological laws needed to understand the evolution of the living biota. *Nat Ecol Evol* **1**, 165 (2017).

15. Williams, T. A. *et al.* Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602–E4611 (2017).

16. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* **372**, (2021).

17. Raymond, J. & Segrè, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**, 1764–1767 (2006).

18. Goldford, J. E., Hartman, H., Smith, T. F. & Segrè, D. Remnants of an Ancient Metabolism without Phosphate. *Cell* **168**, 1126–1134.e9 (2017).

19. Goldford, J. E., Hartman, H., Marsland, R., 3rd & Segrè, D. Environmental boundary conditions for the origin of life converge to an organo-sulfur metabolism. *Nat Ecol Evol* **3**, 1715–1724 (2019).

20. Tian, T. *et al.* Phosphates as Energy Sources to Expand Metabolic Networks. *Life* **9**, (2019).

21. Lazcano, A. & Miller, S. L. On the origin of metabolic pathways. *J. Mol. Evol.* **49**, 424–431 (1999).

22. Teichmann, S. A. *et al.* Small-molecule metabolism: an enzyme mosaic. *Trends Biotechnol.* **19**, 482–486 (2001).

23. Teichmann, S. A. *et al.* The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. *J. Mol. Biol.* **311**, 693–708 (2001).

24. Rison, S. C. G. & Thornton, J. M. Pathway evolution, structurally speaking. *Curr. Opin. Struct. Biol.* **12**, 374–382 (2002).

25. Muchowska, K. B., Varma, S. J. & Moran, J. Nonenzymatic metabolic reactions and life's origins.

Chem. Rev. (2020).

26. Muto-Fujita, A. A Novel model for the Chemical Evolution of Metabolic Networks.
27. Maeda, H. A. & Fernie, A. R. Evolutionary History of Plant Metabolism. *Annu. Rev. Plant Biol.* **72**, 185–216 (2021).
28. Garcia, A. K., Cavanaugh, C. M. & Kacar, B. The curious consistency of carbon biosignatures over billions of years of Earth-life coevolution. *ISME J.* **15**, 2183–2194 (2021).
29. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
30. Ebenhöh, O., Handorf, T. & Heinrich, R. Structural analysis of expanding metabolic networks. *Genome Inform.* **15**, 35–45 (2004).
31. Handorf, T., Ebenhöh, O. & Heinrich, R. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J. Mol. Evol.* **61**, 498–512 (2005).
32. Kim, H., Smith, H. B., Mathis, C., Raymond, J. & Walker, S. I. Universal scaling across biochemical networks on Earth. *Sci Adv* **5**, eaau0149 (2019).
33. Martin, W. & Russell, M. J. On the origin of biochemistry at an alkaline hydrothermal vent. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **362**, 1887–1926 (2007).
34. Camprubi, E., Jordan, S. F., Vasiliadou, R. & Lane, N. Iron catalysis at the origin of life. *IUBMB Life* **69**, 373–381 (2017).
35. Muchowska, K. B., Varma, S. J. & Moran, J. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019).
36. Varma, S. J., Muchowska, K. B., Chatelain, P. & Moran, J. Native iron reduces CO2 to intermediates and end-products of the acetyl-CoA pathway. *Nature Ecology & Evolution* **2**, (2018).
37. Springsteen, G., Yerabolu, J. R., Nelson, J., Rhea, C. J. & Krishnamurthy, R. Linked cycles of oxidative decarboxylation of glyoxylate as protometabolic analogs of the citric acid cycle. *Nat. Commun.* **9**, 91 (2018).
38. Muchowska, K. B. *et al.* Metals promote sequences of the reverse Krebs cycle. *Nature Ecology and*

Evolution **1**, 1716–1721 (2017).

39. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A. & Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* **5**, 1155–1166 (2016).
40. Hafner, J., MohammadiPeyhani, H., Sveshnikova, A., Scheidegger, A. & Hatzimanikatis, V. Updated ATLAS of Biochemistry with New Metabolites and Improved Enzyme Prediction Power. *ACS Synth. Biol.* **9**, 1479–1482 (2020).
41. Kappock, T. J., Ealick, S. E. & Stubbe, J. Modular evolution of the purine biosynthetic pathway. *Curr. Opin. Chem. Biol.* **4**, 567–572 (2000).
42. Kun, A., Papp, B. & Szathmáry, E. Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. *Genome Biol.* **9**, R51 (2008).
43. Whicher, A., Camprubi, E., Pinna, S., Herschy, B. & Lane, N. Acetyl Phosphate as a Primordial Energy Currency at the Origin of Life. *Orig. Life Evol. Biosph.* **48**, 159–179 (2018).
44. Harrison, S. A. & Lane, N. Life as a guide to prebiotic nucleotide synthesis. *Nat. Commun.* **9**, 5176 (2018).
45. Pinna, S. *et al.* A prebiotic basis for ATP as the universal energy currency. *PLoS Biol.* **20**, e3001437 (2022).
46. Catling, D. C. & Zahnle, K. J. The Archean atmosphere. *Sci Adv* **6**, eaax1420 (2020).
47. Fischer, W. W., Hemp, J. & Johnson, J. E. Evolution of oxygenic photosynthesis. *Annual Review of Earth and* (2016).
48. Marshall, S. M. *et al.* Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nat. Commun.* **12**, 3033 (2021).
49. Fontecilla-Camps, J. C. Geochemical Continuity and Catalyst/Cofactor Replacement in the Emergence and Evolution of Life. *Angew. Chem. Int. Ed Engl.* **58**, 42–48 (2019).
50. Milner-White, E. J. & Russell, M. J. Functional capabilities of the earliest peptides and the emergence of life. *Genes* **2**, 671–688 (2011).

51. Goldman, A. D. & Kacar, B. Cofactors are Remnants of Life's Origin and Early Evolution. *J. Mol. Evol.* (2021) doi:10.1007/s00239-020-09988-4.

52. Longo, L. M. *et al.* Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15731–15739 (2020).

53. Nitschke, W., McGlynn, S. E., Milner-White, E. J. & Russell, M. J. On the antiquity of metalloenzymes and their substrates in bioenergetics. *Biochim. Biophys. Acta* **1827**, 871–881 (2013).

54. Saito, M. A., Sigman, D. M. & Morel, F. M. M. The bioinorganic chemistry of the ancient ocean: the co-evolution of cyanobacterial metal requirements and biogeochemical cycles at the Archean–Proterozoic boundary? *Inorganica Chim. Acta* **356**, 308–318 (2003).

55. Johnson, J. E. & Molnar, P. H. Widespread and persistent deposition of iron formations for two billion years. *Geophys. Res. Lett.* **46**, 3327–3339 (2019).

56. Ma, B.-G. *et al.* Characters of very ancient proteins. *Biochem. Biophys. Res. Commun.* **366**, 607–611 (2008).

57. Goldman, A. D., Beatty, J. T. & Landweber, L. F. The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism. *J. Mol. Evol.* **82**, 17–26 (2016).

58. Nath, N., Mitchell, J. B. O. & Caetano-Anollés, G. The Natural History of Biocatalytic Mechanisms. *PLoS Comput. Biol.* **10**, (2014).

59. Aziz, M. F., Caetano-Anollés, K. & Caetano-Anollés, G. The early history and emergence of molecular functions and modular scale-free network behavior. *Sci. Rep.* **6**, 25058 (2016).

60. Eck, R. V. & Dayhoff, M. O. Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences. *Science* **152**, 363–366 (1966).

61. Lee, J. & Blaber, M. Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 126–130 (2011).

62. Smock, R. G., Yadid, I., Dym, O., Clarke, J. & Tawfik, D. S. De Novo Evolutionary Emergence of a Symmetrical Protein Is Shaped by Folding Constraints. *Cell* **164**, 476–486 (2016).

63. Yagi, S. *et al.* Seven Amino Acid Types Suffice to Create the Core Fold of RNA Polymerase. *J. Am.*

Chem. Soc. **143**, 15998–16006 (2021).

64. Seal, M. *et al.* Peptide-RNA Coacervates as a Cradle for the Evolution of Folded Domains. *J. Am. Chem. Soc.* **144**, 14150–14160 (2022).
65. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198 (2001).
66. Longo, L. M. *et al.* On the emergence of P-Loop NTPase and Rossmann enzymes from a Beta-Alpha-Beta ancestral fragment. *Elife* **9**, (2020).
67. Bliven, S. E. *et al.* Analyzing the symmetrical arrangement of structural repeats in proteins with CE-Symm. *PLoS Comput. Biol.* **15**, e1006842 (2019).
68. Granick, S. Evolution of Heme and Chlorophyll. in *Evolving Genes and Proteins* (eds. Bryson, V. & Vogel, H. J.) 67–88 (Academic Press, 1965).
69. Horowitz, N. H. On the Evolution of Biochemical Syntheses. *Proc. Natl. Acad. Sci. U. S. A.* **31**, 153–157 (1945).
70. Fuchs, G. Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.* **65**, 631–658 (2011).
71. Braakman, R. & Smith, E. The emergence and early evolution of biological carbon-fixation. *PLoS Comput. Biol.* **8**, (2012).
72. Weiss, M. C. *et al.* The physiology and habitat of the last universal common ancestor. *Nature Microbiology* **1**, 16116 (2016).
73. Garcia, P. S., Gribaldo, S. & Borrel, G. Diversity and Evolution of Methane-Related Pathways in Archaea. *Annu. Rev. Microbiol.* **76**, 727–755 (2022).
74. Krissansen-Totton, J., Buick, R. & Catling, D. C. A statistical analysis of the carbon isotope record from the Archean to Phanerozoic and implications for the rise of oxygen. *Am. J. Sci.* **315**, 275–316 (2015).
75. Hayes, J. M. Global methanotrophy at the Archean-Proterozoic transition. in *Early Life on Earth. Nobel Symposium* (ed. Bengston, S.) vol. 84 220–236 (Columbia University Press, New York, 1994

1994).

76. Slotznick, S. P. & Fischer, W. W. Examining Archean methanotrophy. *Earth Planet. Sci. Lett.* **441**, 52–59 (2016).

77. Prywes, N., Phillips, N. R., Tuck, O. T., Valentin-Alvarado, L. E. & Savage, D. F. Rubisco Function, Evolution, and Engineering. *Annu. Rev. Biochem.* (2023) doi:10.1146/annurev-biochem-040320-101244.

78. Kędzior, M. *et al.* Resurrected Rubisco suggests uniform carbon isotope signatures over geologic time. *Cell Rep.* **39**, 110726 (2022).

79. Wang, R. Z. *et al.* Evolution of Carbon Isotope Fractionation in Cyanobacteria. *bioRxiv* 2022.06.22.497258 (2022) doi:10.1101/2022.06.22.497258.

80. Braakman, R. & Smith, E. The compositional and evolutionary logic of metabolism. *Phys. Biol.* **10**, 011001 (2013).

81. Lane, N. *Transformer: The Deep Chemistry of Life and Death*. (W. W. Norton & Company, 2022).

82. Shuman, R. F., Shearin, W. E. & Tull, R. J. Chemistry of hydrocyanic acid. 1. Formation and reactions of N-(aminomethylene)diaminomaleonitrile, a hydrocyanic acid pentamer and precursor to adenine. *J. Org. Chem.* **44**, 4532–4536 (1979).

83. Roy, D., Najafian, K. & von Ragué Schleyer, P. Chemical evolution: the mechanism of the formation of adenine under prebiotic conditions. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 17272–17277 (2007).

84. Sousa, F. L., Hordijk, W., Steel, M. & Martin, W. F. Autocatalytic sets in E. coli metabolism. *Journal of systems chemistry* **6**, 4 (2015).

85. Nunes Palmeira, R., Colnaghi, M., Harrison, S. A., Pomiankowski, A. & Lane, N. The limits of metabolic heredity in protocells. *Proc. Biol. Sci.* **289**, 20221469 (2022).

86. Harrison, S. A., Palmeira, R. N., Halpern, A. & Lane, N. A biophysical basis for the emergence of the genetic code in protocells. *Biochim. Biophys. Acta Bioenerg.* 148597 (2022).

87. Noor, E., Haraldsdóttir, H. S., Milo, R. & Fleming, R. M. T. Consistent estimation of Gibbs energy

using component contributions. *PLoS Comput. Biol.* **9**, e1003098 (2013).

88. Beber, M. E. *et al.* eQuilibrator 3.0: a database solution for thermodynamic constant estimation. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab1106.

89. Bertz, S. H. The first general index of molecular complexity. *J. Am. Chem. Soc.* **103**, 3599–3601 (1981).

90. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

91. Cheng, H. *et al.* ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.* **10**, e1003926 (2014).

92. Cheng, H., Liao, Y., Schaeffer, R. D. & Grishin, N. V. Manual classification strategies in the ECOD database. *Proteins* **83**, 1238–1251 (2015).

93. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).

Materials and Methods

Software availability

Code and data are available on the following github repository:
https://github.com/jgoldford/metabolic-continuity

Metabolic network reconstruction

Briefly, we downloaded the KEGG network on May 31, 2021. , We removed *n*=230 reactions that were elementally inconsistent (i.e., the reactants and products did not share the same elements) (Table S3). Reactions that were stoichiometrically imbalanced, or relied on metabolites with unknown “R-groups”, were kept. All reactions are listed in Table S1.

675 Missing reaction and cofactor annotation

676 Cofactors and EC-cofactor dependencies were identified from Expasy (via the `cofactorLabel`
677 field and comments), PDBe (via the cofactor summary), UniProt (via cofactor/EC numbers), and
678 KEGG (via comments and reaction definitions). For more details, see the SI section titled
679 "Cofactor Annotations". We added several missing reactions in KEGG to ensure production of
680 key, generic metabolites, such as quinones, ferricytochrome. We describe each modification in
681 the SI section titled "Addition of missing reactions."

682 Thermodynamic constraints

683 Standard molar free energies were estimated by the component contribution method [87] and the
684 eQuilibrator python API [88] using the following parameters: pH=7.0, pMg=3.0, ionic strength of
685 0.25M and T=298.15. For each reaction with a free energy estimate at standard molar
686 conditions, we estimated the maximum and minimum reaction free energy for both the forward
687 and reverse reactions. We computed these values by assuming that compound concentration
688 ranges were between 10^{-7} and 10^{-1} M, and computed the maximum and minimum reaction free
689 energies as described previously [19]. For all calculations, we assumed activity coefficients were
690 well approximated by concentrations. All forward or reverse reactions with minimum free
691 energies above zero were removed from the network. For KEGG reactions with no free energy
692 estimate (n=839), we kept both the forward and reverse reaction in the network.

693 Modeling primitive purine biosynthesis reactions

694 We modeled the primitive purine biosynthesis by adding 8 new reactions to the model. We
695 replaced all ATP (GTP) with pyrophosphate (PP_i) (C00013) and ADP (GDP) with
696 orthophosphate (C00009) and recomputed standard molar free energies using eQuilibrator for the
697 following KEGG reactions: 5-Phospho-D-ribosylamine:glycine ligase (R04144, EC 6.3.4.13),
698 ATP:pyruvate 2-O-phosphotransferase (R00200, EC 2.7.1.40), , formate:5-amino-1-(5-phospho-
699 D-ribosyl)imidazole-4-carboxamide ligase (R06975, EC 6.3.4.23), ATP:ribose-1,5-bisphosphate
700 phosphotransferase (R06836, EC 2.7.4.23),
701 1-(5-Phosphoribosyl)-5-amino-4-carboxyimidazole:L-aspartate ligase (R04591, EC 6.3.2.6), 5'-
702 Phosphoribosylformylglycinamide:L-glutamine amido-ligase (R04463, EC 6.3.5.3), formate:N1-
703 (5-phospho-beta-D-ribosyl)glycinamide ligase (R06974, EC 6.3.1.21), 2-(Formamido)-N1-(5-
704 phosphoribosyl)acetamidine cyclo-ligase (R04208, EC 6.3.3.1). In Fig. 1b, we provide a

plausible pathway with just 8 of these reactions. The abbreviations for the compounds in this pathway are as follows: Gly = Glycine; Gln = L-Glutamine; Asp = L-Aspartate; PolyP = Polyphosphate (Pyrophosphate); PEP = Phosphoenolpyruvate; RuBP = D-Ribulose 1,5-bisphosphate; RiBP = D-Ribose 1,5-bisphosphate; PRPP = 5-Phospho-alpha-D-ribose 1-diphosphate; PRA = 5-Phosphoribosylamine; GAR = 5'-Phosphoribosylglycinamide; FGAR = 5'-Phosphoribosyl-N-formylglycinamide; FGAM = 5'-Phosphoribosyl-N-formylglycinamidine; AIR = Aminoimidazole ribotide; CAIR = 1-(5-Phospho-D-ribosyl)-5-amino-4-imidazolecarboxylate; SAICAR = 1-(5'-Phosphoribosyl)-5-amino-4-(N-succinocarboxamide)-imidazole; AICAR = 1-(5'-Phosphoribosyl)-5-amino-4-imidazolecarboxamide; FAICAR = 1-(5'-Phosphoribosyl)-5-formamido-4-imidazolecarboxamide; IMP = Inosine monophosphate; AMP = Adenine monophosphate; ATP = Adenine triphosphate.

Models of primitive thioester and redox coenzymes

Previous work suggested that primitive versions of CoA-coupled and NAD(P)-coupled reactions could have enabled the generation of metabolic networks before the availability of phosphate [18,19]. We explored whether primitive coenzyme systems could enable purine production by (i) substituting all CoA-coupled reactions with the thiol mercaptopyruvate and (ii) substituting all NAD(P)H-coupled reactions with hydrogen gas. Note for (i), we were unable to re-compute standard molar free energies because the thioesters formed from mercaptopyruvate were not in the KEGG database. For (ii), we computed standard molar free energies as described previously.

Extending KEGG with additional reactions from ATLAS

We obtained a list of all reactions in the ATLAS database [39,40]. We removed reactions that matched KEGG reactions, resulting in an additional 20,183 reactions. Note that the ATLAS database also contained an additional 3024 metabolites, increasing the total network to 32,449 reactions and 11,768 metabolites. We applied the same thermodynamic methods described above for ATLAS reactions.

Seed set

All seed molecules are listed in Table S4. The seed set consisted of all inorganic divalent and monovalent metal species in KEGG, as well as elemental sources of phosphorus, nitrogen,

sulfur, oxygen, hydrogen, and carbon. While ammonia, orthophosphate and hydrogen sulfide were the only nitrogen, phosphorus, and sulfur sources provided in the seed set, respectively, several organic compounds were included as seed molecules. Since we did not assume a primitive thioester coupling mechanism and include thiols in the seed set, organic molecules such as pyruvate were necessary to enable expansion [19]. We thus included organic molecules producible when reacting glyoxylate with pyruvate [35], resulting in 19 organic compounds (Table S4). Note that when we assumed that succinate semialdehyde could be producible from succinate via geochemically available reductants (e.g., H_2), we were able to generate networks >4000 compounds with only pyruvate as the organic seed molecule.

Network expansion algorithm

The network expansion algorithm was run as described previously [18,19]. Briefly, seed compounds are allowed to react given the reactions in the network, which then produce product compounds. The product compounds are added to the seed set, and the process is repeated until convergence. The network expansion algorithm was implemented using the `networkExpansionPy` python package (https://github.com/jgoldford/networkExpansionPy).

Identification of purine-containing molecules

Purine-containing compounds were identified using the SUBCOMP search function hosted by the KEGG database (https://www.genome.jp/tools/subcomp/). The query structure was purine (i.e., KEGG compound C15587), the database was "COMPOUND", and the search mode was "SUBstructure".

Defining metabolic pathways

We defined metabolic pathways using KEGG modules (https://rest.kegg.jp/list/module). The ordering of reactions and reaction definitions for each module were determined from the "Reaction" field and "Diagram" field in each KEGG module entry. Data was retrieved using the TogoWS REST service via Biopython (https://biopython.org/docs/latest/api/Bio.TogoWS.html). Because the complexity and non-linearity of some modules make it hard to interpret which reactions are strictly necessary, we limit our analysis to linear pathways (Table S10). A full list of the KEGG modules included for analysis can be found in Table S11. Pathways containing less

than 2 reactions are not included. For the linear pathways, we used the reaction to module step rules to compute when each step in the pathway was feasible during the expansion. We then computed the Spearman's rank correlation between iteration number and pathway step (indexed as 0 to *N*, where *N* is the total number of steps in a pathway). A Spearman's rank correlation of 1 indicated the pathway emerged sequentially in the forward direction, while a Spearman's rank correlation of -1 indicated the pathway emerged sequentially in the reverse direction.

Molecular complexity and other ancient features

To compute the molecular complexity for each compound in the expansion scope with a defined molecular formula (*n*=3588), we used RDKit (version 2022.03.1) (https://www.rdkit.org/) to compute the Bertz complexity [89] using the `GraphDescriptors` module with default parameters (Table S6). We computed topological indexes (Extended Data Fig. 6a-b), using custom python scripts, dependent on NetworkX (version 2.5.1). We computed the average state of reduction per carbon [2] (see Extended Data Fig. 3c, Table S7) on all organic molecules with only carbon, oxygen, and hydrogen atoms (*n*=1178).

We computed the proportion of microbial taxa using each compound in the expansion by using the KEGG REST API to download KEGG modules used in each prokaryotic genome (*n*=6416), as well as all KEGG compounds used in each KEGG module. We took 1,314 KEGG compounds that are intermediates in 377 KEGG modules, and computed the proportion of prokaryotic genomes that use each compound as an intermediate in at least one KEGG module. For all compounds produced and a specific expansion iteration, we computed the average proportion of genomes using each compound (Fig. 3b). For Extended Data Fig. 3, we identified KEGG reactions that were unique to Eukaryotes and Prokaryotes. We first identified enzyme classes (EC numbers) in 8652 genomes, 838 of which were Eukaryotes, and computed the EC classes that were unique to either Prokaryotic or Eukaryotic-specific. We converted these KEGG reactions, using KEGG's internal mapping between reactions and EC number. We identified 1638 reactions that were unique to Eukaryotes.

Calculating Reaction-Fold Associations

KEGG orthologous (KO) groups, 8030 in total, were clustered at 80% sequence identity by CD-HIT [90] using the default settings and a word size of 5 characters. Using the curated hidden Markov models (HMMs) from the Evolutionary Classification of Domains (ECOD) database [91,92] (ECOD version develop279), protein families were mapped onto the representative KEGG genes of each orthologous group by the program hmmsearch [93]. HMM searches used the default settings and the search space (the -Z flag) was set to 106,052,079 sequences. Only domains with an independent E-value of less than $1x10^{-4}$ and an HMM coverage of greater than 70% were considered potential hits. Domains were assigned to regions of a gene in order of increasing E-value using the "envelope coordinates" of the associated hit. If two hits overlap by more than 30% of either hit, the hit with the higher E-value is discarded. Finally, reconciled hits associated with a gene were mapped to their corresponding ECOD X-groups when such mappings are unambiguous. An association between an X-group and KO group was identified if at least 20% of the genes contained that X-group domain.

We next constructed a mapping between reactions and KO groups. We used the KEGG REST API to download the associations between KEGG reactions and Enzyme Commission (EC) numbers, and the EC numbers and KEGG orthologous groups. A KEGG orthologous group was associated with a KEGG reaction if they were associated with at least one shared EC number. Finally, X-group and reactions associations were obtained via their mutual associations with KO groups.

Computing structural repetitiveness of protein folds

Protein fold repetitiveness for each representative ECOD domain (70% sequence identity cutoff, ECOD version develop279) was calculated using CE-Symm 2.0 [67] with default settings. All classifications other than C1 (i.e., no internal symmetry) were defined as repetitive. Repetitiveness scores were taken to be the fraction of repetitive representative domains within a fold, defined here as an ECOD X-group (Table S9). The average repetitiveness score for a reaction was calculated by averaging the repetitiveness scores of all associated folds (Table S9).

The average fold repetitiveness per expansion iteration (Fig. 3d) was computed by averaging across all reactions that emerged at each iteration step.

Statistical analysis

Pearson's correlation, Spearman's correlations and two-sided, nonparametric Mann Whitney *U* tests were all performed using the `pearsonr`, `spearmanr` and `mannwhitneyu` functions from the `scipy.stats` python module, version 1.5.4 in python 3.6. For Monte Carlo permutation tests performed in Fig. 2b, we used the `random` function in the python native library `sample` to randomly sample 10^3 combinations of 8 molecules from the expansion scope (but not the seed set in Table S4), and then repeated the expansion. We estimated a *p*-value by computing the proportion of randomly chosen seed sets that led to an expansion with fewer expansion iterations than the original expansion shown in Fig. 2b (blue line). Partial correlation analysis was performed using the `penguin.partial` (version 0.3.12) function in python.