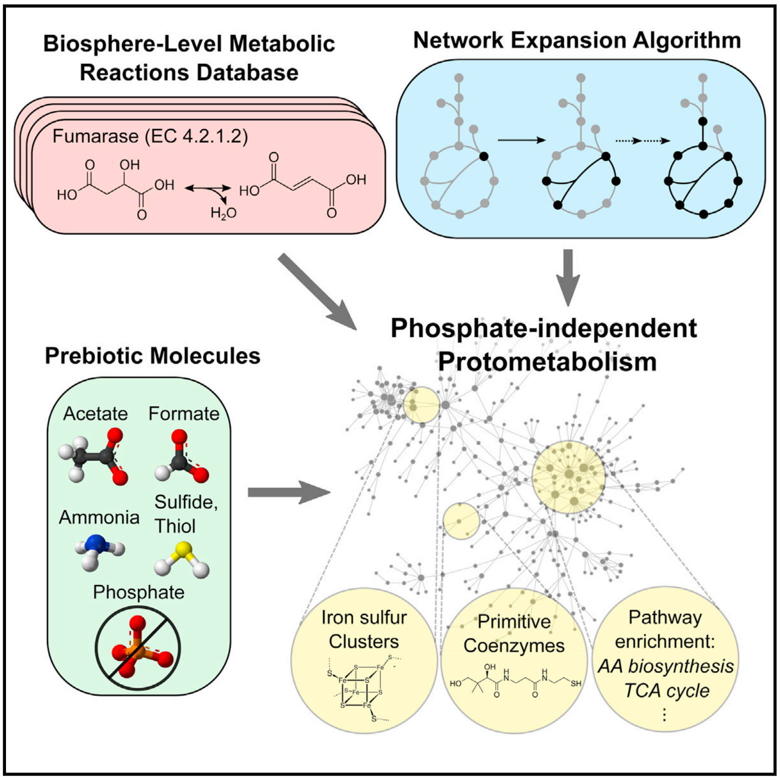
Cell

Theory

Remnants of an Ancient Metabolism without Phosphate

Graphical Abstract



Highlights

- We computationally test the plausibility of an ancient metabolism without phosphate
- A phosphate-independent network exists within biosphere-level metabolism
- This network displays hallmarks of prebiotic chemistry, e.g., iron-sulfur cofactors
- This could represent a "metabolic fossil" of early thioester-driven biochemistry

Authors

Joshua E. Goldford, Hyman Hartman, Temple F. Smith, Daniel Segrè

Correspondence

dsegre@bu.edu

In Brief

Could ancient metabolic networks, prior to the incorporation of phosphate, have led to the emergence of living systems?

Goldford et al., 2017, Cell *168*, 1126–1134
March 9, 2017 © 2017 The Authors. Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2017.02.001

Remnants of an Ancient Metabolism without Phosphate

Joshua E. Goldford,[1] Hyman Hartman,[2] Temple F. Smith,[1,3] and Daniel Segrè[1,3,4,5,*]
[1]Bioinformatics Program, Boston University, Boston, MA 02215, USA
[2]Earth, Atmosphere and Planetary Science Department, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Department of Biomedical Engineering, Boston University, Boston MA, 02215, USA
[4]Department of Biology, Department of Physics, Boston University, Boston MA, 02215, USA
[5]Lead Contact: Daniel Segrè
*Correspondence: dsegre@bu.edu

http://dx.doi.org/10.1016/j.cell.2017.02.001

SUMMARY

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

INTRODUCTION

While most research on the evolution of living systems has been focused on sequences and genomes, some answers to fundamental questions about the emergence of life may be hidden in the architecture of the complex biochemical reaction networks that sustain the cell (Smith and Morowitz, 2016). The field of metabolic network modeling and analysis is expanding as a major research area of relevance to multiple applications (O'Brien et al., 2015; Plata et al., 2015). However, the use of such techniques to address fundamental questions on the emergence of living systems is still highly unexplored.

Among the many unanswered questions on life's origin, the enigma of how phosphate ended up playing a prominent role in cellular biochemistry has been puzzling scientists for decades (Schwartz, 2006), resurfacing in recent years in light of novel discoveries (Adcock et al., 2013; Pasek et al., 2013). Phosphate is present in a large proportion of known biomolecules. It is an essential component of biochemical energy transduction (most notably through ATP), cofactors such as NADH, and information storage (in DNA and RNA polymers) (Nelson and Cox, 2005). However, phosphate is geochemically scarce and difficult to access, often serving as the limiting nutrient in a variety of modern ecosystems (Halmann, 1974). Phosphate is found in terrestrial and marine ecosystems, tightly complexed with rocks and minerals, requiring mechanisms for environmental extraction and transport (Pasek, 2008).

The ensuing dilemma of phosphate's high importance in spite of its poor bioavailability is particularly challenging for early life, as primordial protocells would have needed both a readily available phosphate source and a simple mechanism for early phosphate acquisition. Currently, there is no consensus for a phosphate source in early life, with theories ranging from acid-mediated ion solubilization, high concentrations of reduced phosphorus species in early oceans, or accumulation during late heavy bombardment (Pasek et al., 2013; Schwartz, 2006). Even provided a phosphate source, the mechanisms of phosphate utilization and polymerization in early life remain debated (Keefe and Miller, 1995).

The alternative solution to this dilemma is that primitive forms of life could have initially emerged and endured without a major dependence on phosphate. Multiple scenarios for early metabolic pathways that do not rely on phosphate have been proposed (Deamer and Weber, 2010; de Duve, 1991; Hartman, 1992; Wächtershäuser, 1990). In many of these scenarios, sulfur and iron are conjectured to have fulfilled major catalytic and energetic functions prior to the appearance of phosphate. Most notably, in the thioester world scenario (de Duve, 1991), thioesters are hypothesized to have played a role similar to the one played today by ATP. Thioesters are widespread in modern metabolism, primarily as Coenzyme A (CoA) derivatives (e.g., Acetyl-CoA), and are used as condensing agents, enabling the synthesis of heterogeneous biopolymers.

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

Here we address these questions using computational systems biology approaches originally developed for performing large-scale analyses of complex metabolic networks (Ebenhöh et al., 2004; Handorf et al., 2005). Similar approaches have been previously used to describe the biosphere-level metabolic changes that accompanied the transition to an oxic atmosphere, about 2.2 billion years ago (Raymond and Segrè, 2006). Specifically, we use these and other computational methods to systematically study the size, architecture and physicochemical properties of phosphate-independent biochemical networks. Given that our goal is to shed light on processes that predate the estimated last universal common ancestor (LUCA) (Srinivasan and Morowitz, 2009; Weiss et al., 2016), and given the long-term reshuffling of genes among organisms through horizontal-gene transfer, we focused our analysis on a global, biosphere-level biochemical network, which encompasses all known metabolic reactions across all organisms. In exploring the prebiotic relevance of metabolic reactions that in extant life are catalyzed by highly evolved, efficient, and specific protein-based enzymes, we implicitly formulate the hypothesis that many of such reactions could have been initially catalyzed to a much weaker and less specific extent by a number of small molecules. Such a hypothesis in itself is not new to origin of life research (Martin and Russell, 2007) and is supported by a large body of literature, both pertaining to individual small-molecule catalysts and reactions (Cody et al., 2000, 2004; Gorlero et al., 2009; Metzler and Snell, 1952; Morowitz et al., 2010; Pizzarello and Weber, 2004), as well as to whole networks (Keller et al., 2014; Novikov and Copley, 2013; Semenov et al., 2016).

The major finding we report below is the discovery of a phosphate-independent core metabolism hidden within this biosphere-level network. This core protometabolism is capable of supporting the synthesis of a broad set of biomolecules, including several amino acids and carboxylic acids. Statistical analysis of the physiochemical properties of enzymes within this network show an enrichment for iron-sulfur and transition metal coenzymes. By broadening our analysis of protometabolism with the inclusion of different types of coenzyme precursor couplings, we further show that thioesters, rather than phosphate, could have enabled this core metabolism to overcome energetic bottlenecks, supporting the feasibility of a metabolically rich thioester-based early biochemistry.

RESULTS

Removal of Phosphate from Biosphere-Level Metabolism Leaves Intact a Core Connected Network

The first goal of our analysis was to evaluate the impact of removing all reactions and metabolites involving phosphates (or, more broadly, phosphorus) from metabolism. Rather than analyzing the metabolic networks of individual organisms, we aimed at uncovering effects at the level of the complete collection of all known biochemical reactions (see STAR Methods, Tables S1A and S1B). This "biosphere-level" metabolism (which we inferred from the KEGG database [Kanehisa and Goto, 2000]) allowed us to explore the properties of putative early biochemical networks, beyond the organismal boundaries (Vetsigian et al., 2006).

We started by searching for regions of global metabolism that could be accessible starting from simple molecules likely to have been geochemically abundant on early Earth (Figure 1A). To this end we adopted the network expansion algorithm, which simulates the emergence of metabolic networks from a predefined set of compounds (Ebenhöh et al., 2004; Handorf et al., 2005; Raymond and Segrè, 2006). The algorithm adds metabolites and reactions to an initial seed set, iteratively asking whether any new reaction could take place given the available substrates, until convergence to a final set of reactions and metabolites (or "scope") (see STAR Methods). This algorithm is seed-set dependent, typically resulting in the recovery of a subset of reactions/metabolites within a defined metabolic network (Figure S1). Network expansion was performed with a seed set of eight compounds thought to have been available in prebiotic environments, notably lacking phosphate (Figure 1A, STAR Methods) (Cody et al., 2000; Lang et al., 2010; Martin and Russell, 2007; Russell et al., 2010). Importantly, the set of seed molecules we defined contains simple carboxylic acids in the form of acetate and formate, which could be provided by either an abiotic mechanism or a primitive pathway for carbon fixation (e.g., a primitive variant of the Wood-Ljungdahl pathway [Sousa and Martin, 2014; Sousa et al., 2013; Weiss et al., 2016] or the reductive TCA cycle [Morowitz et al., 2000; Smith and Morowitz, 2004; Wächtershäuser, 1990], see also Discussion). The resulting scope of this seed set consists of a fully connected network of 315 reactions and 260 metabolites (Figure 1A; Tables S2A and S2B), the composition of which is robust to variations of the seed set compounds (Figures S1 and S2). Although this network requires the addition of catalytically accessible carbon, nitrogen and sulfur sources (Figure S1), acetate and formate were substitutable by several alternative carboxylic acids like pyruvate (Figure S2).

This core, phosphate-independent network is significantly enriched with reactions within primary metabolic pathways such as amino acid biosynthesis, pyruvate metabolism, glyoxylate/dicarboxylate, and the TCA cycle, as well as intermediary metabolic pathways such as C5-branched dibasic metabolism (Figure 1B, Fisher's exact test, Bejamini-Hochberg procedure, FDR < 0.05; Table S2C). Further analysis showed significant enrichment for metabolites/reactions involved in various carbon fixation pathways, including the dicarboxylate-hydroxybutyrate cycle, the hydroxypropionate bi-cycle, and the reductive TCA cycle (Table S2D), which has been previously proposed as a primitive carbon fixation pathway in ancient autotrophs (Morowitz et al., 2000). Enrichment for reactions involved in heterotrophic carbon utilization was also observed within pathways for one-carbon (serine pathway) and two-carbon assimilation (Krebs cycle, methylaspartate, and glyoxylate cycle) (Table S2D). In addition to a diverse central carbon metabolism, half of the proteinogenic amino acids (G, A, D, N, E, Q, S, T, C, and H) are producible, representing six of the ten amino acids observed in the Miller-Urey experiment (Parker et al., 2011). In this network, building upon a core carbon, energy and nitrogen metabolism, hydrogen sulfide enables the production of sulfur-containing heterogeneous peptides like glutathione, as well as thioester derivatives

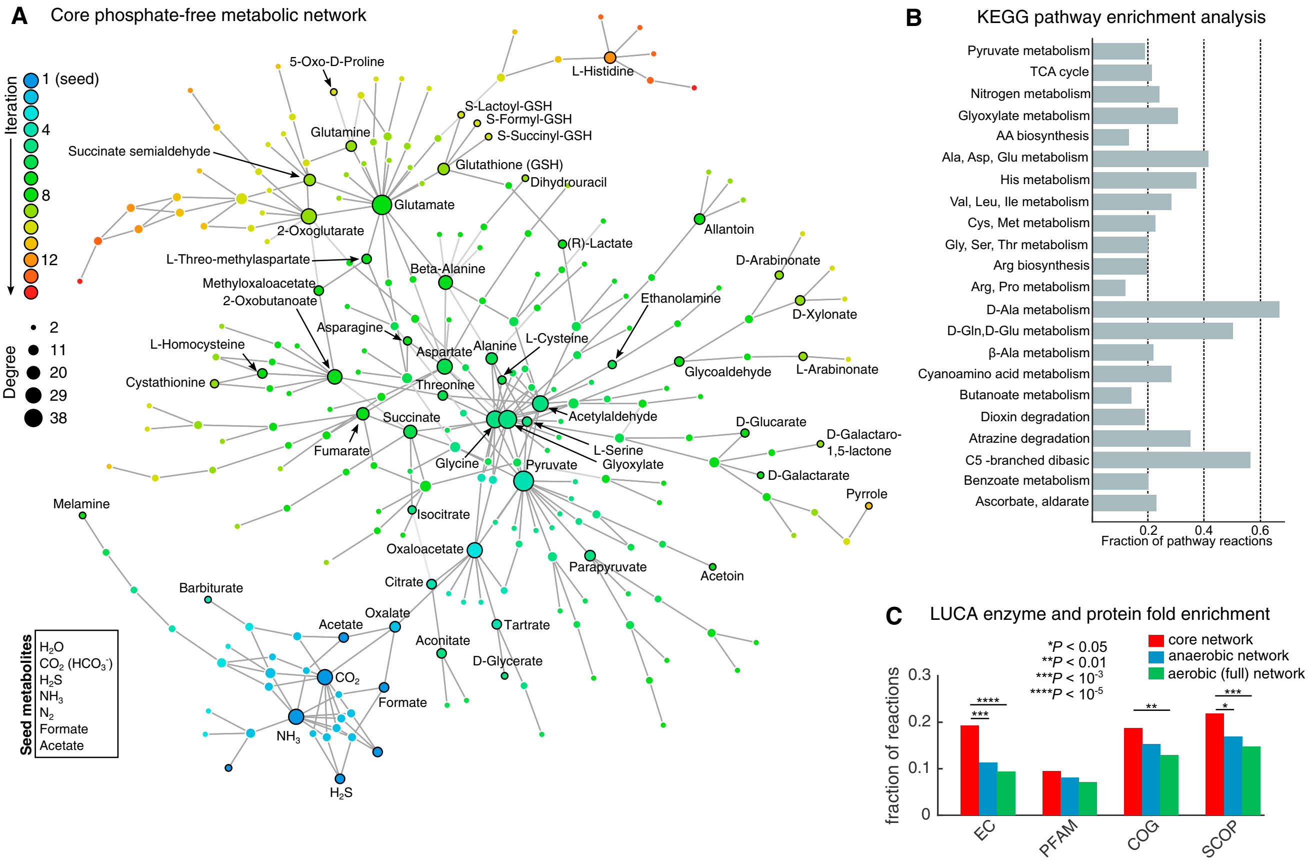


Figure 1. Network Expansion Yields a Core Phosphate-Independent Network

(A) A network expansion algorithm was implemented using a simple set of seed compounds (bottom left box) and all balanced reactions in the KEGG database. The figure displays a simplified view of the resulting network, in which reactions are not explicitly shown, and metabolites are linked if they are interconverted through reactions that are responsible for the expansion. Node color indicates the time (iteration) at which the metabolite appears during the network expansion algorithm, while node size indicates the degree of that node, also indicative of the number of reactions added in the subsequent iteration. Note that major hub metabolites (including pyruvate, glutamate, and glycine—center of the network) are reachable after a few iterations from the seed (blue nodes). Catalytically important amino acids (e.g., His, Ser [Gorlero et al., 2009]) are producible in this network as well.

(B) Pathway enrichment analysis of KEGG pathways within the core network. The fractional abundance of pathway reactions within the core network are plotted for pathways with an FDR < 0.05.

(C) The core network reactions are enriched with enzyme functions (E.C.), protein folds (SCOP) and orthologous genes (COGs) proposed to be present in LUCA, relative to all known metabolic reactions (aerobic network) or to the oxygen-independent (anaerobic) portion of the complete network. (Delaye et al., 2005; Goldman et al., 2013; Mirkin et al., 2003; Srinivasan and Morowitz, 2009; Wang et al., 2007) (Fisher's exact test).

See also Figures S1, S2 and Table S2.

like *S*-formyl and *S*-succinyl glutathione. Intermediates in the degradation and biosynthesis of more complex biomolecules are also observed; 5,6-dihydrouracil is an oxidized catabolic product of uracil and pyrrole is the basic building block for complex heterocyclic aromatic rings like heme (Figure 1A). Thus, we report the existence of a phosphate-independent core metabolic network reachable from simple putative prebiotic compounds.

Core Network Enzymes Are Enriched with Features Associated with Protometabolism

Is there independent evidence that this core phosphate-free network may indeed resemble the very early stages of biochemical processes? The plausibility of this early metabolism relies on the notion that catalysts for these reactions would initially have been much different than they are today, including assortments of short prebiotically-formed peptides (Gorlero et al., 2009; Milner-White and Russell, 2011), metal-ion cofactors (Cody et al., 2000), mineral catalysts (Hazen and Sverjensky, 2010), or iron-rich clays (Hartman, 1975; Laszlo, 1987). Such initial catalysts would have been gradually replaced by longer and more complex genome-encoded protein-enzymes, potentially still retaining properties or components of the early catalysts (Hazen and Sverjensky, 2010; Milner-White and Russell, 2011; Sousa et al., 2013). Thus, we performed multiple analyses to test whether current enzymes within this network contain taxonomic, sequence, and biochemical signals pointing to potential associations with early modes of catalysis.

Taking a taxonomic approach, we found that enzymes in the core network are overrepresented within genomes (Monte Carlo permutation test, $P = 10^{-4}$). The core network is also enriched with enzymes (E.C. numbers) and protein folds (SCOP) previously identified as likely components of the last universal common ancestors (LUCA) proteome (Goldman et al., 2013; Srinivasan and Morowitz, 2009; Wang et al., 2007) (Figure 1C, Fisher's exact test: $P < 10^{-5}$ and $P < 10^{-3}$, respectively), suggesting that a significant fraction of the reactions in this core network appeared in the earliest organisms. One limitation of using comparative phylogenetic analysis is that it only provides information as far back as LUCA. Furthermore, evolutionary processes like horizontal gene transfer (Ochman et al., 2000) and cataclysmic extinction events hamper the elucidation of LUCAs metabolism with certainty. In order to investigate the pre-LUCA features of the phosphate-free core network, we examined the corresponding enzymes in terms of their basic physiochemical properties, with special attention given to properties proposed to be associated with ancient metabolism.

One fundamental property we focused on is the reliance of these enzymes on iron-sulfur or metal coenzymes, reflecting the notion that modern biochemistry emerged from mineral geochemistry (Hazen and Sverjensky, 2010; Martin and Russell, 2007; Wächtershäuser, 1990) and that metal-based cofactors in modern day enzymes represent a living relic of this contingency (Eck and Dayhoff, 1966; Hall et al., 1971; Nitschke et al., 2013). Using a manually curated list of known protein-coenzyme pairs (Goldman et al., 2013), we found that enzymes within the core network were enriched for both zinc and iron-sulfur-dependent coenzymes relative to the full network (Figure 2A, Table S3, Fisher's exact test: $P < 0.05$). For comparison, amino acid derived-coenzymes were observed with comparable frequencies in the core and full KEGG networks, while nucleotide-derived coenzymes (e.g., enzyme-bound FAD, TPP, molybdopterin) were slightly depleted among reactions in the core network, highlighting the coordination between nucleotide and phosphate biochemistry. The occurrence of metal-associated enzymes within the core network was independently corroborated by identifying protein structures with verifiable metal ligands in a separate database (Hemavathi et al., 2009), allowing for the identification of KEGG reactions that rely on enzymes bound to metal ions. Out of the 47% (148/315) of the core network reactions with crystal structures available, 86% (127/148) relied on enzymes with a metal ligand, which constituted a significant enrichment relative to the full KEGG network (Figure 2B, Fisher's exact test: $P < 10^{-5}$).

In addition to a biased coenzyme usage, we investigated other features that could be associated with an ancient protometabolic network. First, motivated by the notion that early catalysts may have been composed of smaller polypeptides relative to present day enzymes (Milner-White and Russell, 2011), we tested if the enzymes in the core network are on average smaller relative to all genome encoded enzymes. We found that sequences are considerably shorter for catalysts in the core network (median = 309) compared to all known metabolic enzymes (median = 354) (Figure 2C; one-tailed Kolmogorov-Smirnov: $P < 10^{-39}$). Second, we thought of checking whether enzymes in the core network are enriched, in their composition, for amino acids producible by the core network itself. Such enrichment would be consistent with the expectation of self-sustainability and homeostasis in a protometabolic network, whereby the network would be capable of producing the building blocks necessary for replenishment and accumulation of its catalysts. We found indeed that core network enzymes are more highly composed of the 10 amino acids found within the core network relative to all known metabolic enzymes (Figure 2D; one-tailed Kolmogorov-Smirnov: $P < 10^{-14}$). One potential simple reason for this enrichment could be attributed to the known sequence bias in FeS-proteins for cysteine, both of which are present in the core phosphate-free network. However, we found no detectable enrichment for cysteine in our core network enzymes (one-tailed Kolmogorov-Smirnov test, $P = 0.948$).

Thioesters Alleviate Thermodynamic Bottlenecks

So far, our analysis was focused on the core network structure, ignoring possible energetic constraints. In extant metabolism, phosphate-mediated group transfer plays a key role by driving unfavorable or energetically uphill reactions (Deamer and Weber, 2010). To investigate the energetic consequences of phosphate unavailability, we implemented a thermodynamically constrained network expansion algorithm, which blocks endergonic reactions with standard molar free energies above a cutoff value, τ (Figure 3B, black line). The network becomes dramatically limited to <12% of the core network as τ remains below 55 kJ/mol, preventing the condensation of oxalate and acetate to yield oxaloacetate. Energetic constraints of this magnitude would have prohibited the expansion of an early metabolism, given plausible ranges of intracellular metabolites concentrations (Bennett et al., 2009) (see STAR Methods). Consequently, a mechanism to overcome these thermodynamic bottlenecks would be essential for a phosphate-independent metabolism.

Could thioester chemistry (de Duve, 1991) serve as a solution to this energetic conundrum? Thioesters, proposed to have served as ancient condensing agents (de Duve, 1991; Sousa et al., 2013), are widespread throughout central metabolic processes (e.g., Coenzyme A [CoA] derivatives in TCA cycle and lipid biosynthesis) and can facilitate energy-rich group transfer. While CoA contains phosphate, this serves mainly as a structural component with no catalytic role, motivating the hypothesis that ancient reactions may have relied on pantetheine, the simpler, phosphate-free variant of CoA (Eakin, 1963; King, 1980) thought to be available in prebiotic environments (Keefe et al., 1995). We explored the energetic consequences of a primitive thioester-based reaction coupling scheme by substituting pantetheine for CoA in modern CoA-coupled reactions, followed by adding pantetheine into the seed set (Figure 3A, see STAR Methods). These changes caused a 33 kJ/mol reduction in the bottlenecks that limited network expansion, enabling the viability of alternative metabolic pathways under physiologically realistic conditions (Bar-Even et al., 2012) (Figure 3B, red line). Interestingly, these bottlenecks could not be easily overcome through an alternative phosphate-based coupling scheme, in which NTP-coupled reactions are substituted with either pyrophosphate or acetyl-phosphate (Figure 3B, blue line, Figure S3). The uniqueness of this behavior is also emphasized by the fact that removal of elements other than phosphate (e.g., sulfur or nitrogen) would dramatically limit the possibility of expansion (Figure S3).

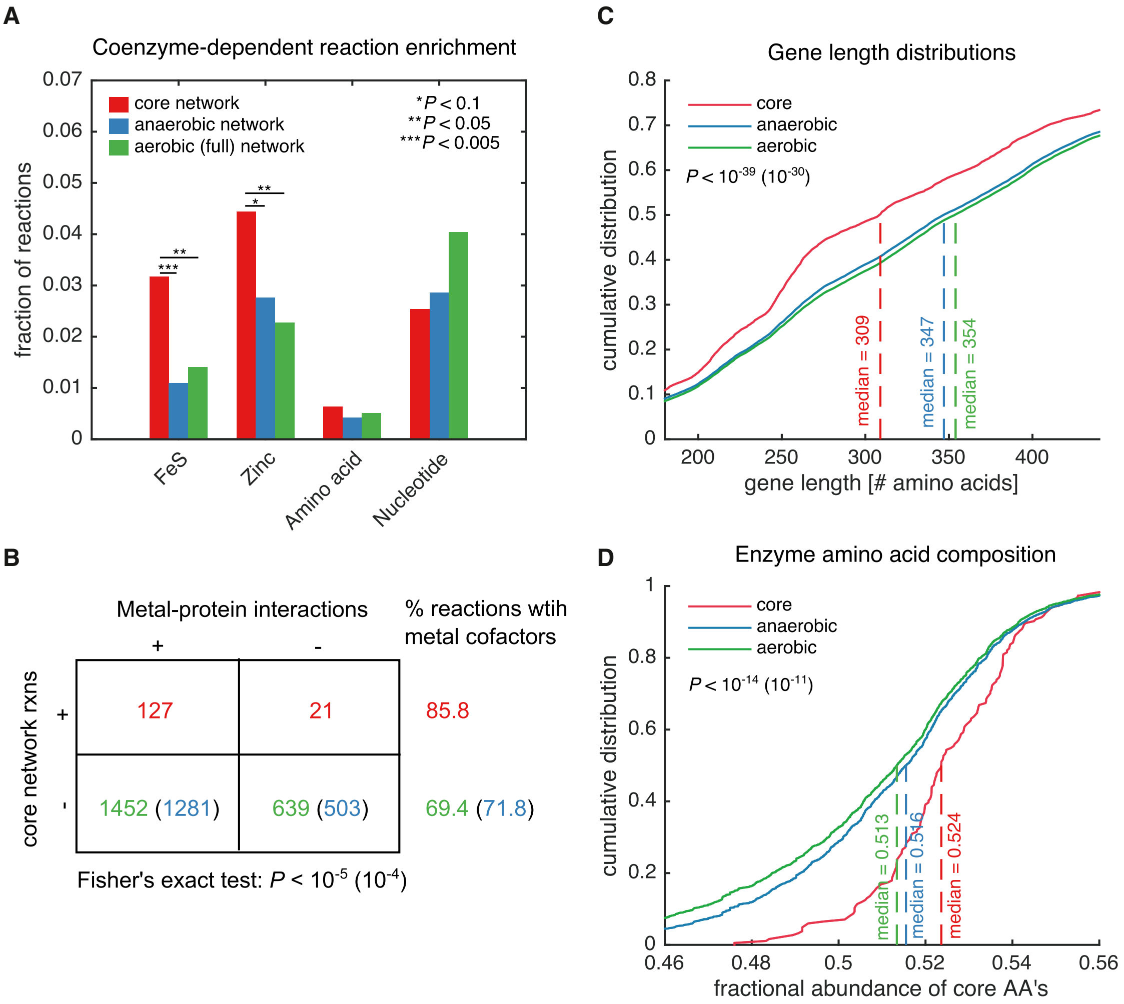


Figure 2. Reactions in the Core Network Are Enriched for Iron-Sulfur and Transition Metal Coenzymes

(A) The fraction of coenzyme-coupled KEGG reactions in the core network (red bars), the anaerobic KEGG network (blue bars) and the aerobic KEGG network (green bars) are compared. Each set of reactions is composed of a manually curated list of coenzyme-coupled reactions in KEGG (Goldman et al., 2013). We found that a significant number of reactions require iron-sulfur coenzymes (Fisher's exact test, $p < 0.05$) and zinc (Fisher's exact test, $p < 0.05$) within the core network relative to the aerobic KEGG network.

(B) Structural data support metal-protein enrichment in the core network. The number of reactions catalyzed by enzymes with available structural data was determined for all KEGG reactions using the MIPS database (Hemavathi et al., 2009). For all reactions with crystal structures available, we classified each reaction as either not having (−) a metal cofactor, or having (+) a metal cofactor. We tested for the enrichment of enzymes containing metal cofactors within the core network relative to both the aerobic KEGG network (green text), or the anaerobic network (parenthesis, blue text).

(C and D) Core network reactions relied more heavily on enzymes with metal cofactors relative to both the aerobic and anaerobic KEGG reactions. The enzymes in the core network are shorter (C) and biased in their amino acid composition (D) relative to either the aerobic or anaerobic KEGG network.

For (B–D), we tested for enrichment within the phosphate-free core network enzymes compared to both the aerobic and anaerobic networks. Significance values are reported for the aerobic network, followed by the anaerobic network in parenthesis.

See also Table S3.

Primitive Coenzymes Enable Widespread Network Expansion

A larger metabolic network may have been reachable if phosphate-free versions of modern day coenzymes drove several primordial reactions. Like CoA, many modern day coenzymes contain nucleotide phosphate groups that are important for enzyme-binding but not directly involved in catalysis. For example, the redox coenzyme NAD contains adenine and phosphate, but facilitates electron transfer at the nicotinamide moiety (Figure S4A). By substituting CoA with pantheteine and implicitly assuming that oxidoreductase reactions could be coupled to primitive electron donors/acceptors instead of NAD(P)/FAD (Figures S4B and S4C), we found that the core network expands to nearly three times more metabolites (814), incorporating five more amino acids (K, R, L, V, and P), uracil and ribose (Figure 4, Tables S4A and S4B). Addition of these five amino acids to the repertoire of amino acids would have enabled broader catalytic capabilities, and paved the way for increased richness of

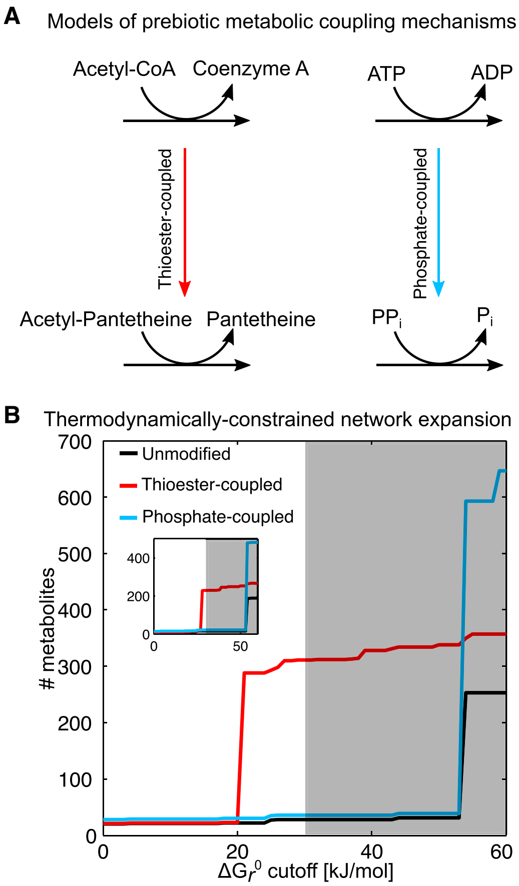


Figure 3. Thioesters Alleviate Thermodynamic Bottlenecks

(A) Models of ancient coenzymes (bottom reactions), based on present-day versions (top reactions), were constructed to simulate the roles of thioesters and phosphates in models of ancient biochemistry (see STAR Methods).

(B) Network expansion from the core seed set was performed after constraining reversibility of reactions exceeding a thermodynamic threshold. For each value of this threshold (x axis) we plot the size (black line) of the final expanded network, in terms of the number of metabolites (y axis). The effect of thioester coupling was simulated by adding a Coenzyme A substitute (pantetheine) to the seed set (red line). For comparison, a phosphate-coupled network was simulated by substituting nucleotide triphosphate-coupled phosphoryl-transfer reactions with pyrophosphate (or acetyl-phosphate) (blue line), followed by adding pyrophosphate (or acetyl-phosphate) to the seed set. Although significantly more metabolites are observed in the phosphate-coupled network with no thermodynamic barrier (due to the addition of sugars and phosphorylated intermediates), the network expansion process would not be thermodynamically feasible under physiologically realistic conditions (non-shaded region) (Bar-Even et al., 2012). Note that more than one third of reactions in KEGG lack a free energy estimate. In the main plot, all reactions with unknown free energies are assumed to be available (equivalent to assuming that they have a free energy barrier lower than then predefined threshold). Results are qualitatively very similar if all such reactions lacking free energy estimates are removed from the network (Top left inset).

See also Figure S3.

peptides, once suitable and energetically supportable mechanisms for peptide synthesis became available (perhaps, initially driven by thioesters themselves [de Duve, 1991]). Further, the formation of pyrimidines, pentoses and vitamins could have set the foundation for the assembly of nucleotide triphosphates and modern coenzymes upon the addition of phosphate (Table S4C).

DISCUSSION

To obtain insight into the early stages of the evolution of metabolism, prior to LUCA, we analyzed the biosphere-level collection of all known metabolic reactions, which throughout the history of life may have greatly shifted their assortment into organisms (Ochman et al., 2000; Vetsigian et al., 2006). By integrating network algorithms and biochemical database analyses at the biosphere-level, we have uncovered a phosphate-independent metabolism that prompts us to revisit models of early biochemistry. This network is enriched with enzymes requiring inorganic and iron-sulfur cofactors, consistent with the hypothesis that iron-sulfur proteins are among the most ancient in biological systems (Eck and Dayhoff, 1966; Hall et al., 1971; Hazen and Sverjensky, 2010; Nitschke et al., 2013; Sousa et al., 2013; Wächtershäuser, 1990). This network incorporates several components of the reductive TCA cycle, proposed to be one of the first autotrophic, autocatalytic cycles in metabolism (Hartman, 1975; Morowitz et al., 2000; Smith and Morowitz, 2004). Its enrichment for enzymes containing iron-sulfur clusters is strikingly consistent with the iron-sulfur world theory (Wächtershäuser, 1990). Our results are compatible with the possibility that the iron-sulfur dependent reactions in the TCA cycle and the methylaspartate cycle in haloarchaea (Khomyakova et al., 2011) may represent modern variants of an iron-sulfur based intermediary metabolism. Upon including phosphate-independent precursors of high-energy and redox cofactors in our model, the resulting network became free of prohibitive thermodynamic bottlenecks, and expanded to a much larger protometabolism that includes several precursors for DNA/RNA and modern-day coenzymes. Our work corroborates previous work emphasizing the potential role of thioesters in protometabolic systems (de Duve, 1991; Sousa et al., 2013).

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

The plausibility of a rich phosphate-independent metabolism has a number of implications on important questions about the origin of life. In particular, the expansion of a phosphate-independent metabolism requires the availability of reduced-carbon precursors (i.e., the seed set) and energy (e.g., the driving force

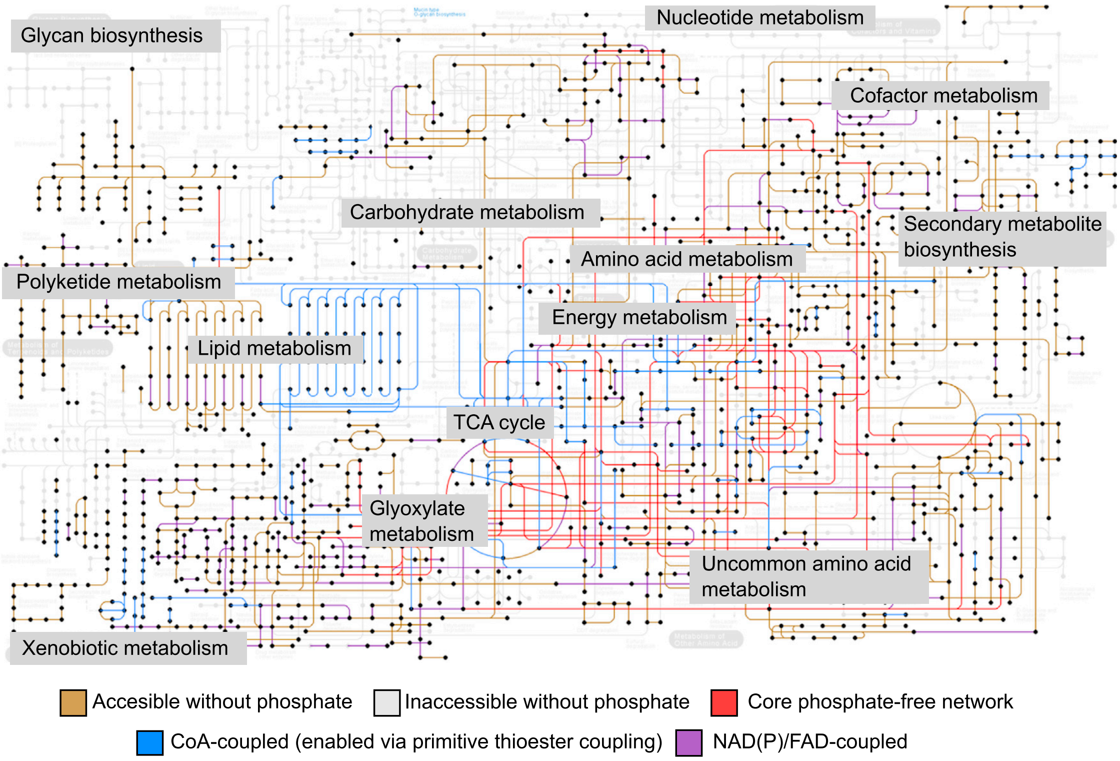


Figure 4. Global Non-phosphate Metabolism

We removed all phosphate-dependent reactions from biosphere-level metabolism, and identified the largest connected subnetwork. The gray lines represent reactions that are unreachable without phosphate, meaning there is no seed set capable of recovering this portion of the network without phosphate. The red lines are the reactions belonging to the phosphate-independent core network (identified in Figure 1A) and the brown lines are reactions that are not included in the core, but still accessible from a phosphate-free seed. The blue lines correspond to phosphate-free reactions coupled to Coenzyme A, while the purple lines are the phosphate-free reactions coupled to nicotinamide or flavin coenzymes.
See also Figure S4 and Table S4.

for the production of thioesters). Although these could be explainable by purely geochemical (i.e., abiotic) processes (see STAR Methods and [Lang et al., 2010; Russell et al., 2010]), a number of scenarios involving ancient variants of modern carbon fixation pathways have been proposed as a source of reduced carbon and thioesters (Fuchs, 2011). One such scenario is based on the reductive TCA cycle (Morowitz et al., 2000; Smith and Morowitz, 2004), which uses several reactions found in our core network, and is compatible with the iron-sulfur enrichment previously discussed. One of the challenges in this scenario is that alternative energy coupling schemes instead of ATP hydrolysis (found in Succinyl-CoA synthetase and ATP-citrate lyase) would have been required to make the process exergonic. An alternative scenario, which could simultaneously explain the availability of formate, acetate and thioesters, is the viability of a primordial Wood-Ljungdahl (WL) pathway, previously suggested to proceed exergonically under prebiotic conditions (Sousa et al., 2013). One of the appeals of this pathway is that it is the only carbon fixation pathway present within both bacteria and archaea, and that it may have been the first carbon-fixation pathway in LUCA (Weiss et al., 2016). For this pathway to be viable in a pre-phosphate world, however, its ancient variants would have to rely on simple coenzyme precursors to pterins, which are currently not known to be synthesized biotically without GTP.

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

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - ○ Reconstruction of biosphere-level metabolism
  - ○ Network expansion
  - ○ Pseudocode
  - ○ Seed set
  - ○ Reaction thermodynamics
  - ○ Primitive coenzyme coupling
  - ○ Enzyme feature datasets
  - ○ Quantification and statistical analysis
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2017.02.001.

An audio PaperClip is available at http://dx.doi.org/10.1016/j.cell.2017.02.001#mmc5.

AUTHOR CONTRIBUTIONS

J.E.G., H.H, T.F.S., and D.S. conceived the project, J.E.G. performed simulations and computational experiments, J.E.G. and D.S. analyzed data, and J.E.G., H.H, T.F.S., and D.S. wrote the paper.

ACKNOWLEDGMENTS

We are grateful for the helpful feedback from Pankaj Mehta, Jennifer Talbot, Kirill Korolev, Jeff Gore, and all members of the Segrè lab. We are also very grateful for the helpful insights provided by anonymous reviewers. This work was partially funded by the NSF under grants DEB-1457695 and NSFOCE-BSF-1635070, by US DOE grant DE-SC0012627, by the Human Frontier Science Program and by the National Institute of General Medical Sciences of the National Institutes of Health under award number T32GM100842.

Received: September 12, 2016
Revised: December 16, 2016
Accepted: January 31, 2017
Published: March 2, 2017

REFERENCES

Adcock, C.T., Hausrath, E.M., and Forster, P.M. (2013). Readily available phosphate from minerals in early aqueous environments on Mars. Nat. Geosci. *6*, 824–827.

Bar-Even, A., Flamholz, A., Noor, E., and Milo, R. (2012). Thermodynamic constraints shape the structure of carbon fixation pathways. Biochim. Biophys. Acta *1817*, 1646–1659.

Bennett, B.D., Kimball, E.H., Gao, M., Osterhout, R., Van Dien, S.J., and Rabinowitz, J.D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat. Chem. Biol. *5*, 593–599.

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

Buckel, W., and Thauer, R.K. (2013). Energy conservation via electron bifurcating ferredoxin reduction and proton/Na(+) translocating ferredoxin oxidation. Biochim. Biophys. Acta *1827*, 94–113.

Cody, G.D. (2004). Transition metal sulfides and the origins of metabolism. Annu. Rev. Earth Planet. Sci. *32*, 569–599.

Cody, G.D., Boctor, N.Z., Filley, T.R., Hazen, R.M., Scott, J.H., Sharma, A., and Yoder, H.S., Jr. (2000). Primordial carbonylated iron-sulfur compounds and the synthesis of pyruvate. Science *289*, 1337–1340.

Cody, G., Boctor, N., Brandes, J., Filley, T., Hazen, R., and Yoder, H. (2004). Assaying the catalytic potential of transition metal sulfides for abiotic carbon fixation. Geochim. Cosmochim. Acta *68*, 2185–2196.

David, L.A., and Alm, E.J. (2011). Rapid evolutionary innovation during an Archaean genetic expansion. Nature *469*, 93–96.

de Duve, C. (1991). Blueprint for a cell: the nature and origin of life (Burlington, N.C.: Neil Patterson Publishers, Carolina Biological Supply Company).

Deamer, D., and Weber, A.L. (2010). Bioenergetics and life's origins. Cold Spring Harb. Perspect. Biol. *2*, a004929.

Delaye, L., Becerra, A., and Lazcano, A. (2005). The last common ancestor: what's in a name? Orig. Life Evol. Biosph. *35*, 537–554.

Eakin, R.E. (1963). An approach to the evolution of metabolism. Proc. Natl. Acad. Sci. USA *49*, 360–366.

Ebenhöh, O., Handorf, T., and Heinrich, R. (2004). Structural analysis of expanding metabolic networks. Genome Inform *15*, 35–45.

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

Flamholz, A., Noor, E., Bar-Even, A., and Milo, R. (2012). eQuilibrator–the biochemical thermodynamics calculator. Nucleic Acids Res. *40*, D770–D775.

Fuchs, G. (2011). Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? Annu. Rev. Microbiol. *65*, 631–658.

Goldman, A.D., Bernhard, T.M., Dolzhenko, E., and Landweber, L.F. (2013). LUCApedia: a database for the study of ancient life. Nucleic Acids Res. *41*, D1079–D1082.

Gorlero, M., Wieczorek, R., Adamala, K., Giorgi, A., Schininà, M.E., Stano, P., and Luisi, P.L. (2009). Ser-His catalyses the formation of peptides and PNAs. FEBS Lett. *583*, 153–156.

Hall, D.O., Cammack, R., and Rao, K.K. (1971). Role for ferredoxins in the origin of life and biological evolution. Nature *233*, 136–138.

Halmann, M. (1974). Evolution and Ecology of Phosphorus Metabolism. In The Origin of Life and Evolutionary Biochemistry, K. Dose, S.W. Fox, G.A. Deborin, and T.E. Pavlovskaya, eds. (Springer), pp. 169–182.

Handorf, T., Ebenhöh, O., and Heinrich, R. (2005). Expanding metabolic networks: scopes of compounds, robustness, and evolution. J. Mol. Evol. *61*, 498–512.

Hartman, H. (1975). Speculations on the origin and evolution of metabolism. J. Mol. Evol. *4*, 359–370.

Hartman, H. (1992). Conjectures and reveries. Photosynth. Res. *33*, 171–176.

Hazen, R.M., and Sverjensky, D.A. (2010). Mineral surfaces, geochemical complexities, and the origins of life. Cold Spring Harb. Perspect. Biol. *2*, a002162.

Hemavathi, K., Kalaivani, M., Udayakumar, A., Sowmiya, G., Jeyakanthan, J., and Sekar, K. (2009). MIPS: metal interactions in protein structures. J. Appl. Cryst. *43*, 196–199.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Keefe, A.D., and Miller, S.L. (1995). Are polyphosphates or phosphate esters prebiotic reagents? J. Mol. Evol. *41*, 693–702.

Keefe, A.D., Newton, G.L., and Miller, S.L. (1995). A possible prebiotic synthesis of pantetheine, a precursor to coenzyme A. Nature *373*, 683–685.

Keller, M.A., Turchyn, A.V., and Ralser, M. (2014). Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. Mol. Syst. Biol. *10*, 725.

Khomyakova, M., Bükmez, Ö., Thomas, L.K., Erb, T.J., and Berg, I.A. (2011). A methylaspartate cycle in haloarchaea. Science *331*, 334–337.

King, G.A. (1980). Evolution of the coenzymes. Biosystems *13*, 23–45.

Lang, S.Q., Butterfield, D.A., Schulte, M., Kelley, D.S., and Lilley, M.D. (2010). Elevated concentrations of formate, acetate and dissolved organic carbon found at the Lost City hydrothermal field. Geochim. Cosmochim. Acta *74*, 941–952.

Laszlo, P. (1987). Chemical reactions on clays. Science *235*, 1473–1477.

Martin, A.C.R. (2005). Mapping PDB chains to UniProtKB entries. Bioinformatics *21*, 4297–4301.

Martin, W., and Russell, M.J. (2007). On the origin of biochemistry at an alkaline hydrothermal vent. Philos. Trans. R. Soc. Lond. B Biol. Sci. *362*, 1887–1925.

Metzler, D.E., and Snell, E.E. (1952). Deamination of serine. I. Catalytic deamination of serine and cysteine by pyridoxal and metal salts. J. Biol. Chem. *198*, 353–361.

Milner-White, E.J., and Russell, M.J. (2011). Functional capabilities of the earliest peptides and the emergence of life. Genes (Basel) *2*, 671–688.

Mirkin, B.G., Fenner, T.I., Galperin, M.Y., and Koonin, E.V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol. Biol. *3*, 2.

Morowitz, H.J., Kostelnik, J.D., Yang, J., and Cody, G.D. (2000). The origin of intermediary metabolism. Proc. Natl. Acad. Sci. USA *97*, 7704–7708.

Morowitz, H.J., Srinivasan, V., and Smith, E. (2010). Ligand field theory and the origin of life as an emergent feature of the periodic table of elements. Biol. Bull. *219*, 1–6.

Nelson, D.L., and Cox, M.M. (2005). Lehninger Principles of Biochemistry, Fourth Edition (W. H. Freeman).

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

Noor, E., Haraldsdóttir, H.S., Milo, R., and Fleming, R.M.T. (2013). Consistent estimation of Gibbs energy using component contributions. PLoS Comput. Biol. *9*, e1003098.

Novikov, Y., and Copley, S.D. (2013). Reactivity landscape of pyruvate under simulated hydrothermal vent conditions. Proc Natl Acad Sci USA *110*, 13283-8.

O'Brien, E.J., Monk, J.M., Palsson, B.O., et al. (2015). Using Genome-scale Models to Predict Biological Capabilities. Cell *161*, 971–987.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299–304.

Parker, E.T., Cleaves, H.J., Dworkin, J.P., Glavin, D.P., Callahan, M., Aubrey, A., Lazcano, A., and Bada, J.L. (2011). Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark discharge experiment. Proc. Natl. Acad. Sci. USA *108*, 5526–5531.

Pasek, M.A. (2008). Rethinking early Earth phosphorus geochemistry. Proc. Natl. Acad. Sci. USA *105*, 853–858.

Pasek, M.A., Harnmeijer, J.P., Buick, R., Gull, M., and Atlas, Z. (2013). Evidence for reactive reduced phosphorus species in the early Archean ocean. Proc. Natl. Acad. Sci. USA *110*, 10089–10094.

Pizzarello, S., and Weber, A.L. (2004). Prebiotic Amino Acids as Asymmetric Catalysts. Science. *303*, 1151.

Plata, G., Henry, C.S., and Vitkup, D. (2015). Long-term phenotypic evolution of bacteria. Nature *517*, 369–372.

Rauchfuss, H. (2008). Chemical Evolution and the Origin of Life (Springer).

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

Russell, M.J., Hall, A.J., and Martin, W. (2010). Serpentinization as a source of energy at the origin of life. Geobiology *8*, 355–371.

Schwartz, A.W. (2006). Phosphorus in prebiotic chemistry. Philos. Trans. R. Soc. Lond. B Biol. Sci. *361*, 1743–1749, discussion 1749.

Semenov, S.N., Kraft, L.J., Ainla, A., Zhao, M., Baghbanzadeh, M., Campbell, V.E., Kang, K., Fox, J.M., and Whitesides, G.M. (2016). Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. Nature *537*, 656–660.

Smith, E., and Morowitz, H.J. (2004). Universality in intermediary metabolism. Proc. Natl. Acad. Sci. USA *101*, 13168–13173.

Smith, E., and Morowitz, H.J. (2016). The Origin and Nature of Life on Earth: The Emergence of the Fourth Geosphere (Cambridge, United Kingdom: Cambridge University Press).

Sousa, F.L., and Martin, W.F. (2014). Biochemical fossils of the ancient transition from geoenergetics to bioenergetics in prokaryotic one carbon compound metabolism. Biochim. Biophys. Acta *1837*, 964–981.

Sousa, F.L., Thiergart, T., Landan, G., Nelson-Sathi, S., Pereira, I.A., Allen, J.F., Lane, N., and Martin, W.F. (2013). Early bioenergetic evolution. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20130088.

Srinivasan, V., and Morowitz, H.J. (2009). The canonical network of autotrophic intermediary metabolism: minimal metabolome of a reductive chemoautotroph. Biol. Bull. *216*, 126–130.

Vetsigian, K., Woese, C., and Goldenfeld, N. (2006). Collective evolution and the genetic code. Proc. Natl. Acad. Sci. USA *103*, 10696–10701.

Wächtershäuser, G. (1990). Evolution of the first metabolic cycles. Proc. Natl. Acad. Sci. USA *87*, 200–204.

Wang, M., Yafremava, L.S., Caetano-Anollés, D., Mittenthal, J.E., and Caetano-Anollés, G. (2007). Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res. *17*, 1572–1585.

Weiss, M.C., Sousa, F.L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W.F. (2016). The physiology and habitat of the last universal common ancestor. Nat. Microbiol. *1*, 16116.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Database		
KEGG	(Kanehisa and Goto, 2000)	http://www.genome.jp/kegg/kegg1.html
LUCApedia	(Goldman et al., 2013)	http://eeb.princeton.edu/lucapedia/
MIPS	(Hemavathi et al., 2009)	http://dicsoft2.physics.iisc.ernet.in/cgi-bin/mips/query.pl
eQuilibrator	(Flamholz et al., 2012)	http://equilibrator.weizmann.ac.il/
Software and Algorithms		
MATLAB 2015a	Mathworks	https://www.mathworks.com/
Python v. 2.7.13	Python	https://www.python.org/
Inkscape 0.91	Inkscape	https://inkscape.org/
NetworkX 1.11	Network X	https://networkx.github.io/
D3.js	Mike Bostock	https://d3js.org/
Webweb.js 3.2	Daniel B. Larremore	http://danlarremore.com/webweb/
Network expansion algorithm	(Ebenhöh et al., 2004)	https://github.com/segrelab/networkExpansion

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for data and software may be directed to, and will be fulfilled by the Lead Contact Daniel Segrè (dsegre@bu.edu).

METHOD DETAILS

Reconstruction of biosphere-level metabolism

The set of all known metabolic reactions was assembled into a biosphere-level (or pangenome) metabolic model using the KEGG database. All KEGG reactions and compounds were downloaded using the KEGG REST API (http://www.kegg.jp/kegg/docs/keggapi.html). We constructed a stoichiometric matrix from the KEGG reaction database using reaction equations. Reactions were removed if they either consumed or produced compounds that (i) did not include a SMILES string or (ii) included an *n*-subunit polymer with undefined molecular formulas. Metabolites with arbitrary, ‘‘R’’ groups were retained as long as ‘‘R’’ groups were balanced in the reaction equations. Reactions that were elementally imbalanced for any element except hydrogen were removed. The ensuing elementally balanced network consisted of 6880 reactions and 5944 metabolites. The final list of chemical reactions is provided in Tables S1A and S1B. This network is referred to as the ‘‘full KEGG network’’ (or simply the KEGG network) in the main text and in the rest of the STAR Methods below, and also as the ‘‘aerobic network’’ in Figures 1 and 2. Note that due to the filtering of KEGG reactions described above, essential for an accurate accounting of atoms, the KEGG networks used throughout the manuscript are depleted in enzymes that catalyze reactions that are chemically unbalanced (e.g., fatty acid elongation reactions, lysine biosynthesis), and in enzymes with no assigned KEGG reaction, like arsenate reductase (EC 1.20.4.4) and ketol-acidreductoisomerase (EC 1.1.1.382). We expect that the accuracy of future analyses will increase with further improvement of KEGG and other metabolic databases, which will better reflect the collection of biosphere-level metabolism.

The results presented in the Figure 1 and supplemental figures S1 and S2 were generated by applying the network expansion algorithm to different variants of the KEGG network, in which we generally assume all reactions to be reversible (by representing each reaction twice in both forward and backward directions). The only exception is the set of reactions that involve molecular oxygen, which were not allowed to proceed in the oxygen-producing direction, in order to best mimic the conditions of the biosphere prior to the great oxidation, as explored previously (Raymond and Segrè, 2006).

While the above constraints on oxygen-involving reactions can be used to study the growth of metabolism through the network expansion algorithm, they cannot easily used for imposing pre-oxic conditions in enrichment analyses that use sets of KEGG reactions and enzymes . In particular, as shown in Figure 2, for establishing the uniqueness of some properties of the phosphate-free network, we performed tests of statistical significance for enrichment of these properties relative to both the full KEGG network (aerobic) and the network accessible without oxygen (anaerobic network). The comparisons with the anaerobic network were performed in order to ensure that statistical enrichment tests were not biased by including reactions and enzymes likely added to the global metabolic network after oxygen accumulated in the atmosphere. The anaerobic network was generated by removing subsets of reactions and metabolites from the global metabolic network reachable only through reactions that utilize molecular oxygen, resulting in a modified biosphere-level metabolic network. This network was obtained through the following steps: We first removed all reactions that utilize oxygen. Second, the stoichiometric matrix was converted into a bipartite undirected graph, where nodes were either reactions or metabolites. In this bipartite graph, an edge exists between a reaction and a metabolite if that reaction either consumes or produces that metabolite. The graph was used as an input into the python package *NetworkX* (https://networkx.github.io/), and all connected components were detected. In our specific analysis, this algorithm identified a single major connected component that contained the majority of metabolic reactions. The final anaerobic network, containing 5,651 reactions and 5,252 metabolites, is provided in the Tables S1A and S1B.

Network expansion

The network expansion algorithm has been described in detail elsewhere (Ebenhöh et al., 2004; Handorf et al., 2005; Raymond and Segrè, 2006). Briefly, let S represent the set of seed metabolites. We will denote by $F(S)$ the scope of the seed S, i.e., the set of all reactions and metabolites reachable from the seed set S. At each iteration k, the set of reactions R_k, whose substrates are present in the current set of metabolites, is added to the scope. The products of these newly added reactions constitute a set of metabolites, M_k. The scope is then updated by taking the union of $F(S)$, R_k, and M_k. This update can also be concisely described as the following operation: $F(S) \leftarrow F(S) \cup R_k \cup M_k$. The algorithm terminates when no more reactions or metabolites can be added to the scope, resulting in a final stationary network composition. Although expansion from a given seed set can in principle end up spanning the complete set of reactions and metabolites in the network, the typical scenario is that a given seed set gives rise to a scope that spans only a fraction of the complete network. Note that in order to take into account thermodynamic feasibility of reactions during network expansion, we model each reaction as a pair of distinct forward and backward irreversible reactions. For example, a reaction listed in KEGG as A → B (reversible), will be represented as a pair of irreversible reactions (A → B and B→A) which we will refer to later as "unidirectional reactions." Thermodynamic infeasibility of the reaction in a given direction is then implemented simply by removing the corresponding unidirectional reaction.

Pseudocode

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

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

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

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

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

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

Let a set of seed metabolites, C_s, contain the indices of metabolites, where for all $i \in C_s$, $x_i = 1$. Initialize x such that for all $i \in C_s$, $x_i = 1$. Define $l_0 = 0$, $l_1 = \sum_i x_i$ and $k = 1$.

```
**While** $l_k > l_{k-1}$:
**Do**:
```

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

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

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

$$k = k + 1$$

Seed set

Seed set compositions were chosen based on previously reported putative molecular compositions on prebiotic Earth (see Martin and Russell, 2007). Figure 1 of the main text lists the compounds used in the seed set to generate the core network referenced throughout the paper. Volatiles and gases widely considered to be present on early Earth are dinitrogen, water, hydrogen sulfide and carbon dioxide (Rauchfuss, 2008). Although it is unclear at what time biotic nitrogen fixation emerged, abiotic nitrogen reduction to ammonia has been demonstrated at high concentrations of hydrogen sulfide (Hazen and Sverjensky, 2010), and is thought to have been the dominant nitrogen source in early organisms (David and Alm, 2011), motivating us to include ammonia into the seed set.

Reduced carbon is an essential component of metabolism, requiring either an autotrophic carbon fixation process or the heterotrophic carbon assimilation of abiotically reduced carbon. We tested two scenarios: (i) an autotrophic origin of metabolism from carbon dioxide and hydrogen gas and (ii) a heterotrophic origin of metabolism from formate and acetate. We did not see significant growth from scenario (i), indicating that a reduced form of carbon is required. Acetate and formate were chosen based on previous work suggesting that early forms of abiotically fixed carbon may have existed in the form of simple carboxylic acids. These acids could have been in principle synthesized at hydrothermal vents from hydrogen and carbon dioxide using the processes of serpentinization (Lang et al., 2010; Martin and Russell, 2007; Russell et al., 2010), or via a primitive variant of a modern carbon fixation pathway such as the Wood-Ljungdahl pathway (Fuchs, 2011; Sousa and Martin, 2014; Sousa et al., 2013; Weiss et al., 2016). We explored variations to this seed set in two ways. First, we performed a Monte Carlo permutation test on the seed set (see Figure S1) and second, we varied the identity of the carbon sources (see Figure S2).

Reaction thermodynamics

To sustain net flux for a chemical reaction, the laws of thermodynamics require that the difference in free energy between products and reactants, $\Delta \mathrm{G}'_r$, has to be negative (Bar-Even et al., 2012; Nelson and Cox, 2005). For a given biochemical reaction at fixed temperature and pressure, $\Delta \mathrm{G}'_r$ is defined as:

$$\Delta G'_r = \Delta G'^{o}_r + RT \ln \prod_i a_i^{s_{i,r}}$$

where $\Delta G'^{o}_r$ is the free energy change of the reaction at standard molar conditions, R is the ideal gas constant, T is temperature, a_i is the activity of metabolite i and $s_{i,r}$ is the stoichiometric coefficient for metabolite i in reaction r, which is negative for reactants and positive for products. Assuming metabolite concentrations, c_i, can be substituted for activities, the disequilibrium ratio, Γ_r is defined as $\Gamma_r = \prod_i c_i^{s_{i,r}}$. The necessary condition for a negative free energy of a reaction can be recast as: $\Gamma_r < -\Delta G'^{o}_r / RT$. This indicates that for large $\Delta G'^{o}_r$, a small Γ_r is required to maintain feasibility. $\Delta G'^{o}_r$ represents a “thermodynamic barrier,” which can be overcome by reducing Γ_r (i.e. increasing the reactants relative to the product concentration).

To identify potential thermodynamic barriers, we performed network expansion without unidirectional reactions (see Network Expansion Algorithm section) above a predefined free energy threshold, τ. In this variant of network expansion, reaction r was removed if $\Delta G'^{o}_r > \tau$, for varying levels of τ. We obtained estimates for $\Delta G'^{o}_r$ from eQuilibrator (Flamholz et al., 2012), which uses the component contribution method to estimate free energies of formation of metabolites based on the group decomposition of compounds (Noor et al., 2013). We obtained estimates of $\Delta G'^{o}_r$ at various pH values, ranging from pH 5 to pH 9 in increments of 0.5, while assuming a constant ionic strength of 0.1 M and temperature of 298.15 K. We performed network expansion at thresholds varying from 0 to 60 kJ/mol in 1 kJ/mol increments. Final network sizes in all scenarios were insensitive to the choice of pH.

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

Primitive coenzyme coupling

Coenzymes in modern day metabolism are composed of highly heterogenous functional units, composed of distinct moieties involved in protein binding and catalysis (for a comprehensive review, see (Braakman and Smith, 2013)). Two groups of coenzymes are readily observed that contain phosphate: Phosphoryl-donating/accepting coenzymes (e.g., ATP and GTP) and phosphate-containing coenzymes with no phosphoryl group transfer (e.g., Coenzyme A, TPP, NAD, FAD, Molybdopterin). For Phosphoryl-donating/accepting coenzymes, removing phosphate would clearly abolish the catalytic function of the coenzymes, while in phosphate-containing coenzymes, removing all phosphate-containing moieties may not eliminate catalytic function. Phosphoryl-donating/accepting coenzymes may have been preceded by non-nucleotide metabolites with phosphodiester bonds, such as phosphoenolpyruvate, acetyl-phosphate, or pyrophosphate (de Duve, 1991) while phosphate-containing coenzymes may have been preceded by less complex versions of these coenzymes (King, 1980).

The following subsections summarize the introduction of variants of present-day reactions into our network, in which current cofactors are substituted with putative primitive alternatives. In particular, this amounts to the addition of putative prebiotic reactions utilizing primitive thioester, phosphate, and redox couplings, as described below (see also Figure 3 and 4). Figure S4A provides the structures of Coenzyme A, NAD and ATP, highlighting the role of phosphates in each biomolecule.

Thioester coenzymes

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

Phosphate coenzymes

For the proposed model of a primitive phosphate-coupled network, nucleotide (A, G, C, U, T, and I) phosphate-coupled and phosphotransferase reactions were substituted with pyrophosphate. Pyrophosphate has been proposed to have been used for primitive energy coupling in early protometabolic systems before NTP (de Duve, 1991; Martin and Russell, 2007). Using pyrophosphate coupling instead of NTP prevents network expansion from artificial catabolism of NTP precursors like ribose and nucleobases, which are not assumed a priori to be abundant in geochemical models of Hadean environments. In particular, monophosphate transfer reactions were replaced with the disphosphate/monophosphate coenzyme couple, while diphosphate transfers were replaced with the triphosphate/monophosphate coenzyme couple. This model of primitive phosphate-coupled reactions was seeded with orthophosphate, diphosphate and triphosphate, in addition to the ‘‘core seed set’’ listed in Figure 1 of the main text. For monophosphate transfer reactions, we also substituted NTPs with acetyl-phosphate, and found no difference between the network sizes at different free energy threshold cutoffs (Fig 3).

Redox coenzymes

We found that a much larger network was reachable without phosphate by relaxing the condition that major redox reactions require the phosphate-containing coenzymes NAD(P) or FAD as substrates (see Tables S4A and S4B). For this analysis, in addition to adding modified thioester-coupled reactions (see Thioester coenzymes), major redox reactions mediated by NAD(P) and FAD were allowed to proceed using only the half reactions. Reactions utilizing the $NAD(P)^+$/NAD(P)H or the FAD/FADH2 redox couples were replaced with the associated redox half reactions with no cofactor pair. For example, the reaction $X + NAD(P)H \rightarrow Y + NAD(P)^+$ was replaced by the half reaction: $X + 2e^- + 2H^+ \rightarrow Y$, where both e^- and H^+ are in the seed set. Several alternative redox coupling schemes may have been available in protometabolic systems (Figures S4B and S4C), including glutathione or primitive iron-sulfur proteins. For our analysis, we simply decomposed redox reactions into half reactions and allowed for free exchange of electrons. This alteration effectively adds low potential reduced ferrodoxin as a seed molecule, potentially producible via electron bifurcation from H_2 (Buckel and Thauer, 2013).

Enzyme feature datasets

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

LUCApedia

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

MIPS database

Data from the MIPS database was downloaded as an HTML page from the website, and an in-house python script was written to parse the webpage into a list of PDB IDs. PDB IDs were mapped to Uniprot proteins using PDBSWS (Martin, 2005), followed by the conversion of Uniprot to KEGG genes using the KEGG conversion tool (http://www.genome.jp/kegg/tool/conv_id.html).

KEGG

Gene lengths and amino acids compositions were obtained from the KEGG database using the REST API. For each reaction in the full KEGG network, we found all orthologous groups (KO) associated with each reaction. For each KO group, we downloaded the amino acid sequence for each gene within the associated orthologous group. For gene lengths, we computed the number of characters in each sequence. For the amino acid composition, we computed the average amino acid composition across all orthologous groups for each reaction, resulting in an averaged number of each amino acid per reaction. For Fig 2D, we computed the fraction of each sequence consisting of the 10 amino acids found within the core network. The KEGG REST API was used to identify all reactions in each species (n = 3838) in KEGG.

Quantification and statistical analysis

For the taxonomic enrichment test, we first computed the average number of phosphate-free core network reactions across all species in KEGG. We then randomized the set of 315 reactions and repeated the calculation 10^5 times. To test for enrichment for categorical features associated with the core network enzymes (Figure 1C, 2A-B), a 2x2 contingency table was constructed and a Fisher's exact test was performed. For continuous metrics, we used the nonparametric Kolmogorov-Smirnov test. For all pathway and module enrichment analysis, we used a Benjamini-Hochberg multiple comparison's correction and report only pathways and modules with a false discovery rate < 0.05. All statistical tests were performed in MATLAB 2015a, using built-in functions for two-sample Kolmogorov-Smirnov tests (*kstest2.m*), Fisher's exact tests (*fishertest.m*), multiple hypothesis testing (*multcompare.m*). Monte Carlo permutation tests were performed using the *randsample.m* function.

DATA AND SOFTWARE AVAILABILITY

MATLAB scripts written for network expansion can be found on Github (https://github.com/segrelab/networkExpansion).

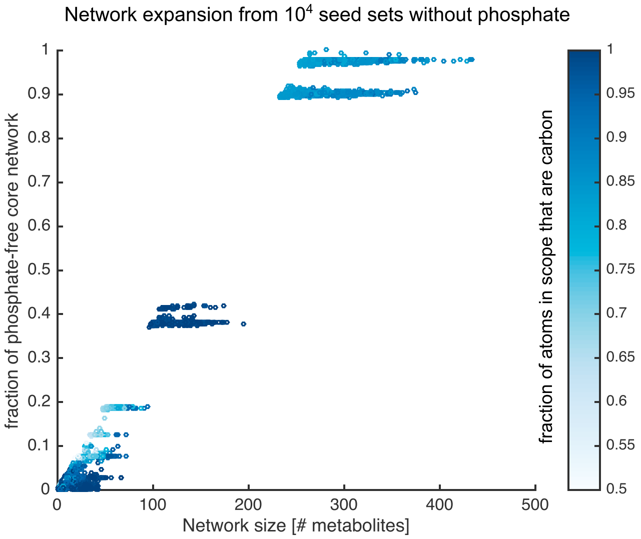


Figure S1. Monte Carlo Sampling of Seed Sets Recovers Substantial Fractions of the Non-phosphate Core Network, Related to Figure 1

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

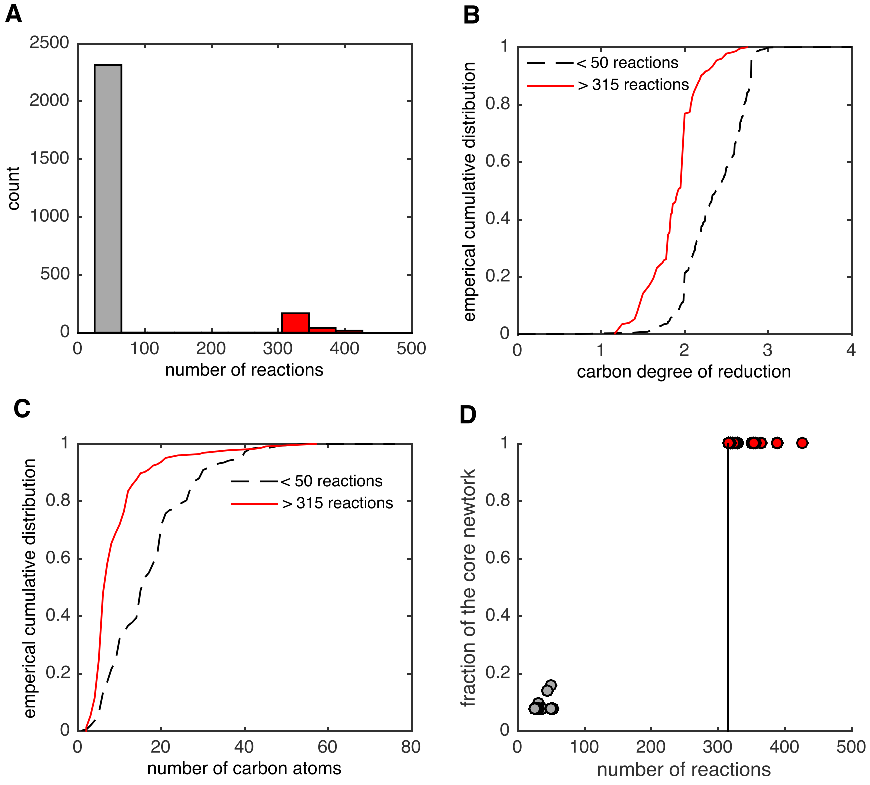


Figure S2. Network Expansion with Various Carbon Sources, Related to Figure 1

Network expansion was repeated with acetate and formate (see Figure 1, main text) replaced by a single organic compound. This process was repeated for each KEGG molecule composed exclusively of C, H and O.

(A) A histogram of network size (reaction count) after network expansion. The majority of carbon sources resulted in small networks (< 50 reactions, gray), while 225 carbon sources resulted in networks > 315 reactions (red).

(B) Empirical CDFs for the average degree of reduction per carbon atom (y/x for substrate C, where $x\mathrm{CO_2} + y\mathrm{H_2} \rightarrow \mathrm{C} + z\mathrm{H2O}$, see (Smith and Morowitz, 2004) in main text) for small (black dashed line) and large (red continuous line) networks. It can be seen that more highly oxidized carbon substrates led, with increased frequency, to larger networks (two-tailed Kolmogorov-Smirnov test, $p < 10^{-55}$).

(C) Empirical CDFs for number of carbons in the seed set that give rise to small (black dashed line) and large (red continuous line) networks. Large networks were generated more frequently from smaller carbon substrates (two-tailed Kolmogorov-Smirnov test, $p < 10^{-41}$).

(D) Scatter plot of the number of reactions (x axis) in each network expansion versus the fraction of the core network embedded in the final network (y axis). All large networks are greater than the 315 reactions obtained using acetate (black line), indicating that expansion from acetate represents a suitable lower bound for a phosphate-independent core metabolism. It should be noted that larger network (> 350 reactions) resulted from carbohydrate sources in the seed (glucose), while slightly smaller networks were generated from carboxylic acids (acetate, oxaloacetate).

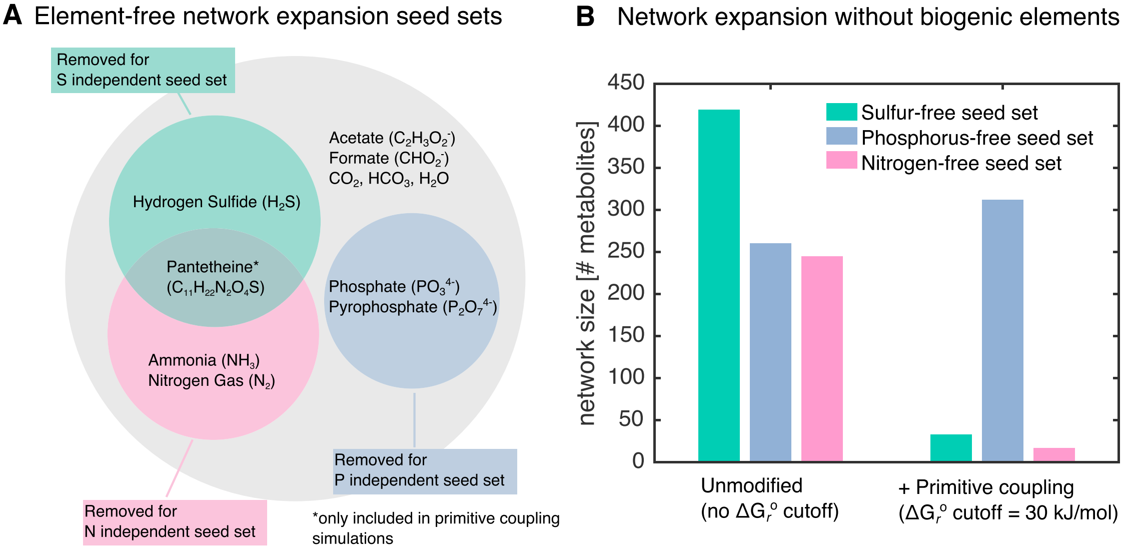


Figure S3. Sulfur and Nitrogen Are Required for Thermodynamically Feasible Network Expansion, Related to Figure 3

(A and B) This analysis aims at testing the uniqueness of the feasibility of a phosphorus-free network, in comparison to other hypothetical scenarios in which other atoms are missing from the initial seed set. Specifically, we compare the size of the expanded network under elimination of phosphorus, sulfur or nitrogen, with and without the thermodynamic feasibility constraints. Network expansion was performed for all KEGG reactions using a seed set without sulfur (no H_2S and pantetheine, green bars), phosphate (no pyrophosphate, blue bars), or nitrogen (no ammonia, nitrogen gas, and pantetheine, pink bars) (see also Venn diagram for specific seeds and atomic compositions). The left set of bars represents the size of expanded networks without imposition of thermodynamic constraints, while the right plot shows the network sizes when thermodynamic feasibility is imposed. It can be seen that removal of sulfur, without thermodynamic constraints, gives rise to a larger network relative to the P-free core network, due to the appearance of sugars and phosphosugars. However, when taking into account thermodynamic feasibility, the phosphate-independent network is the only one that can reach a large size. Thus, a thermodynamically feasible network expansion is conditional on the presence of sulfur and nitrogen, but not phosphate. This observation is in line with broad consensus regarding the prebiotic availability of sulfur (Cody, 2004; Deamer and Weber, 2010; Hazen and Sverjensky, 2010; Wächtershäuser, 1990), as compared to the uncertain and debated prebiotic availability of phosphorus (de Duve, 1991; Schwartz, 2006).

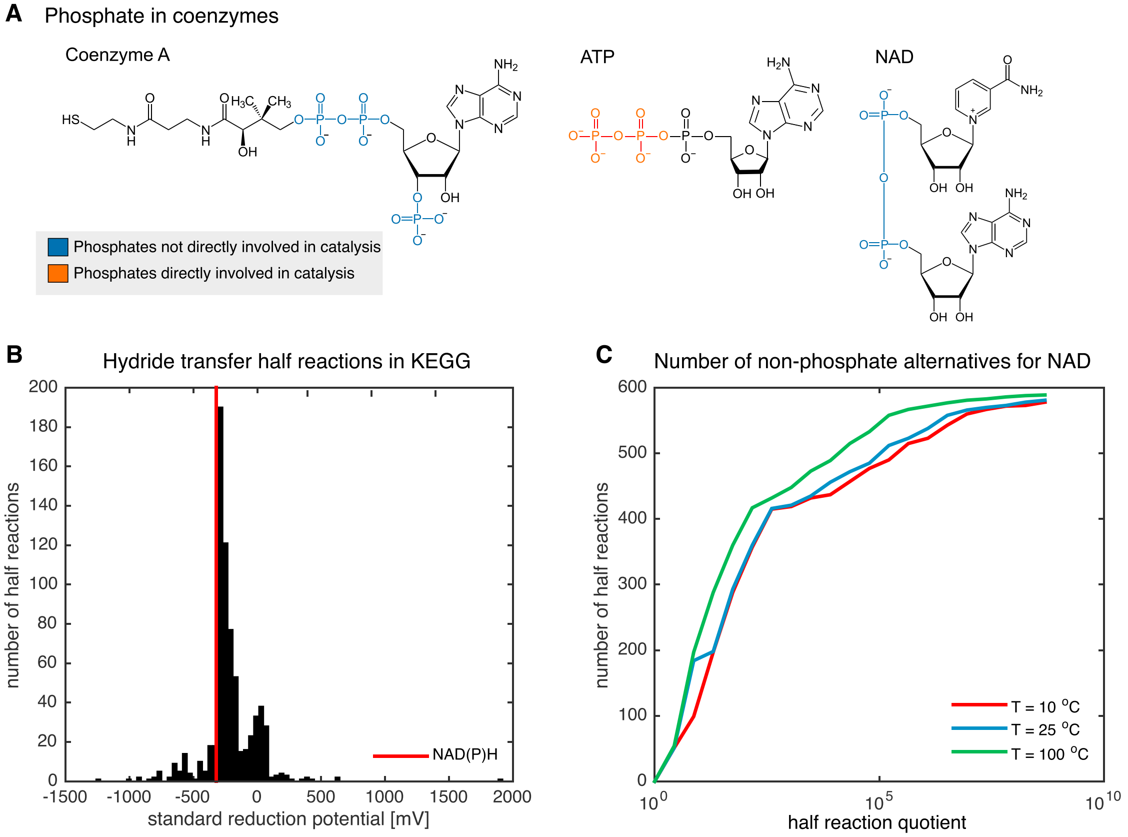


Figure S4. There Are Several Suitable Biomolecules that May Have Preceded Modern Phosphate-Dependent Coenzymes, Related to Figure 4

(A) Phosphates in modern day coenzymes. The phosphates in Coenzyme A and NADH play no role in catalysis, while the phosphates in ATP are pivotal in catalysis.

(B) Plausible substitutes for NADH. All half reactions representing a two electron reduction via hydrogen transfer (i.e., Oxidized + $2H^+$ + $2e^-$ → Reduced) were computed using KEGG reaction pairs database. For approximately 3/4 of all suitable half reactions (712/947), a standard reduction potential could be estimated using group contribution estimates of free energies of formations for KEGG metabolites (Flamholz et al., 2012; Noor et al., 2013). The red line marks the standard reduction potential of NAD(P)/NAD(P)H.

(C) For different maximum allowable concentration ratio between oxidized and reduced species (x axis) we counted the number of half reactions that overlapped with the reduction potential of NADH (y axis). The colored lines represent different temperatures. For example, if a 100-fold concentration ratio was permitted, approximately 200 non-phosphate half-reactions could have sufficed as plausible substitutes for NAD(P).