nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-019-1018-8

Environmental boundary conditions for the origin of life converge to an organo-sulfur metabolism

Joshua E. Goldford[1,2,3*], Hyman Hartman[4], Robert Marsland III[5] and Daniel Segrè[1,3,5,6*]

It has been suggested that a deep memory of early life is hidden in the architecture of metabolic networks, whose reactions could have been catalyzed by small molecules or minerals before genetically encoded enzymes. A major challenge in unravelling these early steps is assessing the plausibility of a connected, thermodynamically consistent proto-metabolism under different geochemical conditions, which are still surrounded by high uncertainty. Here we combine network-based algorithms with physico-chemical constraints on chemical reaction networks to systematically show how different combinations of parameters (temperature, pH, redox potential and availability of molecular precursors) could have affected the evolution of a proto-metabolism. Our analysis of possible trajectories indicates that a subset of boundary conditions converges to an organo-sulfur-based proto-metabolic network fuelled by a thioester- and redox-driven variant of the reductive tricarboxylic acid cycle that is capable of producing lipids and keto acids. Surprisingly, environmental sources of fixed nitrogen and low-potential electron donors are not necessary for the earliest phases of biochemical evolution. We use one of these networks to build a steady-state dynamical metabolic model of a protocell, and find that different combinations of carbon sources and electron donors can support the continuous production of a minimal ancient 'biomass' composed of putative early biopolymers and fatty acids.

The structure of metabolism carries a memory of its evolutionary history that may date back to before the onset of an RNA-based genetic system[1–6]. Decoding this ancient evolutionary record could provide important insights into the early stages of life on our planet[2,5–8], but it constitutes a challenging problem. This challenge is due to the difficulty of interrogating complex biochemical networks under different environmental conditions and also to the uncertainty about these conditions on prebiotic Earth. Estimates of plausible Archaean environments that led to the emergence and evolution of living systems vary dramatically[9,10], ranging from alkaline hydrothermal vents driven by chemical gradients[11–13] to acidic ocean seawater driven by photochemistry[3,4]. Although geochemical data support the availability of mid-potential electron donors (H_2)[14,15], sulfur (for example, hydrogen sulfide, H_2S)[3,4,16–18] and potentially fixed carbon[19–22] in ancient environments, several key molecules used in living systems may have been severely limiting, including a source of fixed nitrogen[23,24] (for example, ammonia, NH_3), low-potential electron donors[25,26] and phosphate[27–29]. Using network-based algorithms we found evidence that thioesters, rather than phosphate, may have supplied ancient metabolism with key energetic and biosynthetic capacity[30]. This raises the question about whether other molecules and physico-chemical conditions may not be as crucial as previously thought for the emergence of a proto-metabolism[31].

A computational method that can help address these questions is the network expansion algorithm, which simulates the growth of a biochemical network by iteratively adding to an initial set of compounds the products of reactions enabled by available substrates, until no additional reactions or metabolites can be added[32,33]. This algorithm, in its application to the study of ancient life[30,34], relies on three key assumptions. The first assumption is that classes of biochemical reactions essential for the rise of living systems were gradually built on, but were rarely lost throughout early evolution. This would imply that the memory encoded in metabolism about its history is sufficiently complete to allow for inferences of ancient states and their evolutionary expansion.

Although this assumption is currently a conjecture, it is supported by the broader evolutionary argument that essential molecules and biological structures tend to be conserved and built on by subsequent molecules and structures. This layered architecture has been extensively studied and observed in ferredoxins[1], the ribosome[35] and also metabolism[2,5,36–40]. This concept is also consistent with recent evidence that early core biochemical pathways, similar to those we see today, may have arisen readily[19,41,42] and become prevalent in the biosphere without further global optimization[43,44]. Further support to this conjecture comes from the observation that early innovations in biochemical functions would spread broadly across the biosphere[45–47], suggesting that the complete loss of fundamental enzymatic capabilities would be very unlikely, even on organismal extinction. It is also implausible that whole categories of reactions would have become extinct in the presence of drastic global changes, such as the great oxygenation event, due to the opportunities seized through fast adaptations in specific environmental niches[34]. The importance of this conjecture for the current work and possible follow up studies is further examined in the Discussion.

This view of metabolism as a biosphere-level phenomenon is an inherent aspect of the network expansion algorithm and could be viewed as its second key assumption. This assumption allows questions to be asked about the rise of metabolism across organismal boundaries. Over long time-scales, horizontal gene-transfers produced abundant shuffling of biochemical reactions across different organisms[45–47], which supports the idea that a global ecosystem-level approach to metabolism may be particularly suitable

[1]Bioinformatics Program, Boston University, Boston, MA, USA. [2]Department of Chemistry, Boston University, Boston, MA, USA. [3]Biological Design Center, Boston University, Boston, MA, USA. [4]Earth, Atmosphere and Planetary Science Department, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Department of Physics, Boston University, Boston, MA, USA. [6]Department of Biomedical Engineering, Department of Biology, Boston University, Boston, MA, USA. *e-mail: goldford@bu.edu; dsegre@bu.edu

for describing ancient biochemistry. The third assumption is that inorganic or small molecular catalysts could catalyze, in a weaker and less specific manner relative to modern enzymes, many metabolic reactions, as confirmed by an increasing body of experimental evidence[19,41,42,48,49].

In this paper we systematically explore a combinatorial set of molecules and parameters associated with possible early Earth environments, and use an enhanced network expansion algorithm to determine which proto-metabolic networks are thermodynamically reachable under each of these initial conditions. We also use constraint-based flux balance modelling to demonstrate the capacity of some of these networks to sustain flux in a way that resembles homeostatic growth of present-day cells. Our results suggest that a thioester-driven organic network may have robustly arisen without phosphate, fixed nitrogen or low-potential electron donors. By supporting the biosynthesis of keto acids and fatty acids, this network may have prompted the rise of complex self-sustaining biochemical pathways, marking a key transition towards the origin of life.

Results

Thermodynamically constrained metabolic network expansion. We wanted to systematically characterize the effect of various geochemical scenarios on the possible structure of ancient metabolism. Building on prior work[30,34] we constructed a model of ancient biosphere-level metabolism based on the KEGG (Kyoto Encyclopedia of Genes and Genomes) database[50]. We first modified the network (as described in Methods) to account for previously proposed primitive thioester-coupling and redox reactions[30]. These modifications included the introduction of reactions whose redox cofactors were substituted by unspecified molecules defined only by their redox potential. Many of the reactions in our set represented whole classes of reactions, with a multitude of possible specific instances. For each possible set of environmental parameters (including temperature, pH and redox potential) we calcualted the thermodynamic feasibility of each reaction and removed infeasible reactions (Methods). This allowed us to implement a thermodynamically constrained network expansion algorithm[30], which iteratively adds metabolites and thermodynamically feasible reactions to a network until no additional reactions and metabolites can be added to the network[30,32–34]. This method ensures that reactions added to the network are locally (rather than globally) thermodynamically feasible (Fig. 1a and Methods).

We performed thermodynamically constrained network expansion (Methods and Fig. 1a) for $n=672$ different geochemical scenarios, where we systematically varied pH, temperature, redox potential of primitive redox systems (analogous to extant nicotinamide adenine dinucleotide (NAD)/nicotinamide adenine dinucleotide phosphate (NADP) and flavin adenine dinucleotide (FAD)-coupled reactions) and the availability of key biomolecules including thiols (that subsequently form thioesters), fixed carbon (formate/acetate), fixed nitrogen (ammonia) and various electron donors and acceptors (Methods, Fig. 1 and Supplementary Data 1). Initial seed sets were chosen to be representative of hypothesized prebiotically available sources of carbon, sulfur, oxygen, hydrogen and nitrogen-containing biomolecules, spanning a range of relevant redox states[11], as discussed in detail previously[30].

A systematic analysis of how environmental conditions affect proto-metabolism. Of the 672 different simulated geochemical scenarios, we found that 288 (43%) expanded to networks containing over 100 metabolites (Fig. 1b). A logistic regression classifier that uses geochemical parameters as predictors (Methods and Fig. 1c) allowed us to quantify the importance of each environmental parameter when determining whether the expanded network would reach such a large size. Surprisingly, removing the variable associated with the presence/absence of ammonia did not affect the predictive power of the classifier, suggesting that a source of fixed nitrogen is not an important determinant of the expansion. Consistent with the relevance of this result to ancient metabolism, we found that the enzymes that catalyze reactions in the expanded networks before the addition of ammonia were depleted in nitrogen-containing coenzymes (one-tailed Wilcoxon signed-rank test: $P < 10^{-24}$; Extended Data Fig. 1c,d) and were depleted in active site amino acids with nitrogenous side-chains (one-tailed Wilcoxon signed-rank test: $P < 10^{-24}$; Extended Data Figs. 1e,f and 2) relative to enzymes added after the addition of ammonia (Supplementary Text). These results suggest that ammonia may have not been essential for the initial expansion of metabolism and indicate a thioester-coupled organo-sulfur metabolic network (Fig. 1) as a core network that deserves further attention.

Beyond the dispensability of nitrogen, the simulations described above revealed a number of relationships between plausible geochemical scenarios and the structure and size of our simulated proto-metabolic networks. First, expansion beyond 100 metabolites was feasible in the absence of a source of fixed carbon, but only when thiols were provided in the seed set, highlighting the importance of thioester-coupling for ancient carbon fixation pathways[2,4,22,25,26,30]. The presence of thiols enabled the production of key biomolecules, including fatty acids and branched-chain keto acids (Extended Data Fig. 3). Second, we explored the effect of the primitive redox system by systematically varying the reduction potential of the electron donor in the seed set (Methods and Fig. 2a). Unexpectedly, we found that as we increased the fixed potential of the electron donor, expansion to a large network was feasible over a broad range of reduction potentials (between −150 and 50 mV). Only when 50 mV was reached did the expanded network collapse to a much smaller solution, suggesting that the generation of low-potential electron donors from H_2 may not have been a necessary condition for the early expansion of a proto-metabolism (Fig. 3a). We also explored conditions with combinations of generic oxidants and reductants (Extended Data Fig. 4), as well as with the addition of fixed carbon to the seed sets (Extended Data Figs. 4 and 5), but did not find any conditions where expansion was selectively dependent on low-potential electron donors. Less stringent constraints, for example the presence of mid-potential redox couples and thioester-forming thiols, could therefore have enabled the emergence of a proto-metabolic network capable of producing key biomolecules.

Autotrophic expansion with thioesters was found to be infeasible at pH 5 and $T=50$ °C (Fig. 2b) due to a blockage in the production of oxalyl-thioesters (Fig. 2c). In our simulations of expansion from autotrophic seed sets, oxalyl-thioester was a critical intermediate in the production of glyoxylate, which was recently proposed to be a key starting material for the production of proto-metabolic networks[42]. This observation prompted us to explore more thoroughly the consequences of removing reactions from the set of feasible reactions used during network expansion. To address this, we systematically removed 236 classes of reactions, grouped by Enzyme Commission (EC) numbers, and performed network expansion using an autotrophic seed set and a mid-potential redox system (−220 mV). Interestingly, we found that three classes of reactions were critical for expansion, including reactions carried out by NAD/NADP-dependent oxidoreductases operating on aldehydes and ketones (1.2.1.X), thioester hydrolases (3.1.2.X) and carboxy-lyases (4.1.1.X). The most perturbed networks were generated by removing reactions catalyzed by enzyme classes involved in fatty acid biosynthesis in (for example, 3.1.2.X and 5.3.3.X) as well as fatty acid degradation (1.1.1.X, 4.2.1.X and 2.3.1.X).

Convergence of geochemical scenarios onto a core organo-sulfur metabolism. Analysis of the expanded networks without nitrogen revealed that a large number of different initial conditions converged to similar expanded organo-sulfur proto-metabolic networks,

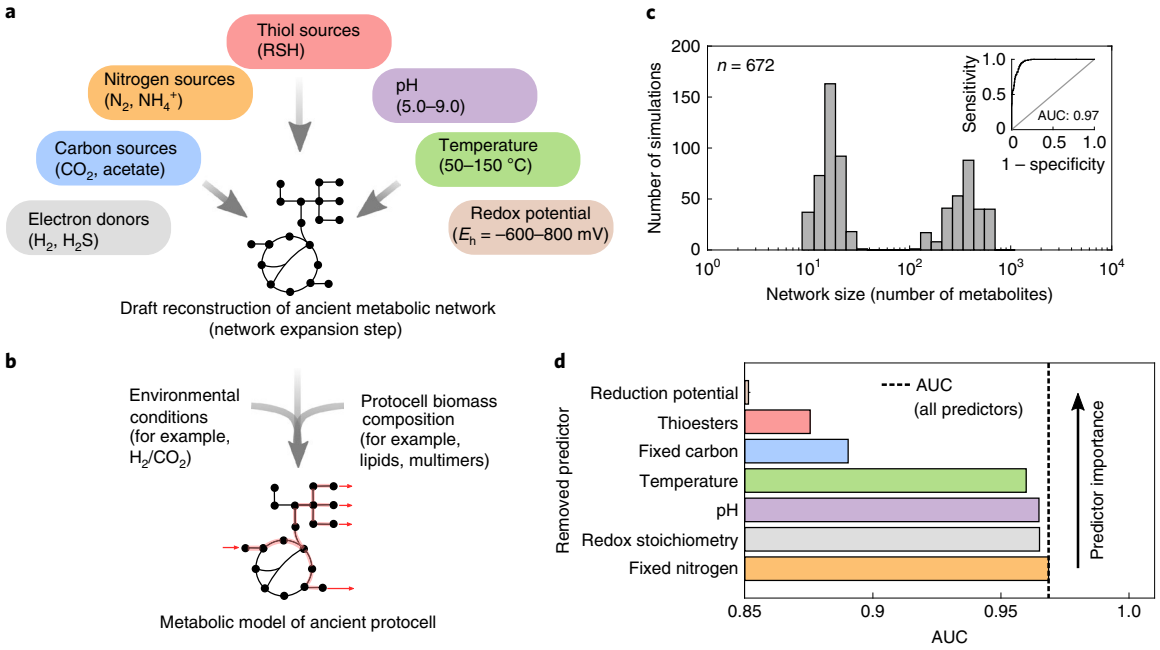


Fig. 1 | Nitrogen is not essential for the initial expansion of metabolism. a, A network expansion algorithm was used to simulate the early expansion of metabolism under 672 scenarios, systematically varying the availability of reductants in the environment, pH, carbon sources, the presence of thiols (RSH), temperature, reduction potential (E_h), and the availability of nitrogen. This process is subject to local thermodynamic feasibility constraints, that is it allows new reactions to occur only if they are individually thermodynamically feasible (Methods). **b**, We implemented detailed stoichiometric model simulations using FBA for a subset of networks obtained from network expansion. Global thermodynamic feasibility constraints were applied (Methods and Fig. 4). **c**, A histogram of network sizes (*x* axis, number of metabolites) revealed a bimodal distribution, where expansion occurred beyond 100 metabolites in 43% (288/672) of scenarios. Inset: a logistic regression classifier was constructed to predict whether a geochemical scenario resulted in a network that exceeded 100 metabolites and a receiver operating curve was plotted. The trained classifier resulted in an area under the curve (AUC) of 0.97 and a leave-one-out cross-validation accuracy of 0.89. **d**, Models were trained without information on specific geochemical variables (*y* axis, ranked by predictor importance) and the ensuing AUC was plotted as a bar chart (*x* axis), revealing that knowledge of the availability of fixed nitrogen offers no information on whether networks expanded.

spanning variants of key pathways in central carbon metabolism (Fig. 3b). For the majority of simulations, variants of modern heterotrophic carbon assimilation pathways, including the glyoxylate cycle and TCA cycle, were well represented in the network (Fig. 3b). Several carbon fixation pathways were also included in the simulated networks. In more than half of the networks expanded beyond 100 metabolites, we found 92% (12/13) of the compounds (or generalized derivatives) that participate in the reductive tricarboxylic acid (rTCA) cycle, with the exception of phosphoenolpyruvate. We also found that under several geochemical conditions, all intermediates were able to be produced for three carbon fixation pathways, including the 3-hydroxypropionate bi-cycle, the hydroxypropionate–hydroxybutylate cycle and the dicarboxylate–hydroxybutyrate cycle (Fig. 3a). Only three of the nine metabolites used in the Wood–Ljungdahl (WL) pathway were observed due to the lack of nitrogen-containing pterins in the network. This does not necessarily rule out the primordial importance of the WL-pathway because its early variants could have been radically different from today's WL-pathway, relying on native metals to facilitate reduction of CO_2 to acetate[19,26]. In addition to observing a large number of metabolites used in carbon fixation pathways, we found that a large fraction of the β-oxidation pathway was represented in our networks, which may have supported the production of fatty acids in ancient living systems by operating in the reverse direction. Interestingly, recent metabolic engineering efforts have demonstrated the feasibility of fatty acid synthesis via a reversible β-oxidation pathway[51]. We also observed that the majority of intermediates involved in the production of branched-chain amino acids were also able to be produced in the expanded networks. To explore the variability of networks generated by the expansion, we provide an interactive visualization as a supplementary data file (Supplementary Software 1) and website (https://prelude.bu.edu/pmne/).

A more detailed analysis of the convergent organo-sulfur proto-metabolic network reveals new possible ancestral metabolic pathways that involve previously unexplored combinations of reactions and metabolites. Figure 3b shows a variant of the (r) TCA cycle that is a component of these expanded networks and may have been the core organo-sulfur network fuelling ancient living systems. Rather than using ATP-dependent reactions found in extant species (for example, succinyl–coenzyme A (CoA) synthetase and ATP citrate lyase), these reactions are substituted with non-ATP-dependent reaction mechanisms. For instance, the production of a succinyl–thioester in the extant rTCA cycle relies on succinyl–CoA synthetase, performing the following reaction: ATP + succinate + CoA → succinyl–CoA + ADP + Pi. However, in the network presented in Fig. 3b, malyl-thioester, produced through alternative reactions, donates a thiol to succinate and subsequently forms a succinyl-thioester. This (r)TCA cycle analogue is able to produce eight keto acids normally serving as key intermediates and precursors to common amino acids in central carbon metabolism (glyoxylate, pyruvate, oxaloacetate, 2-oxoglutarate and hydroxypyruvate), as well as a few branched-chain keto acids. Long-chain fatty acids such as palmitate can also be produced in this network, driven by thioester and redox-coupling rather than ATP, as in extant fatty acid biosynthesis. Despite the simplicity of seed compounds, several small molecular weight keto acids and fatty acids may have been produced in an organo-sulfur proto-metabolism.

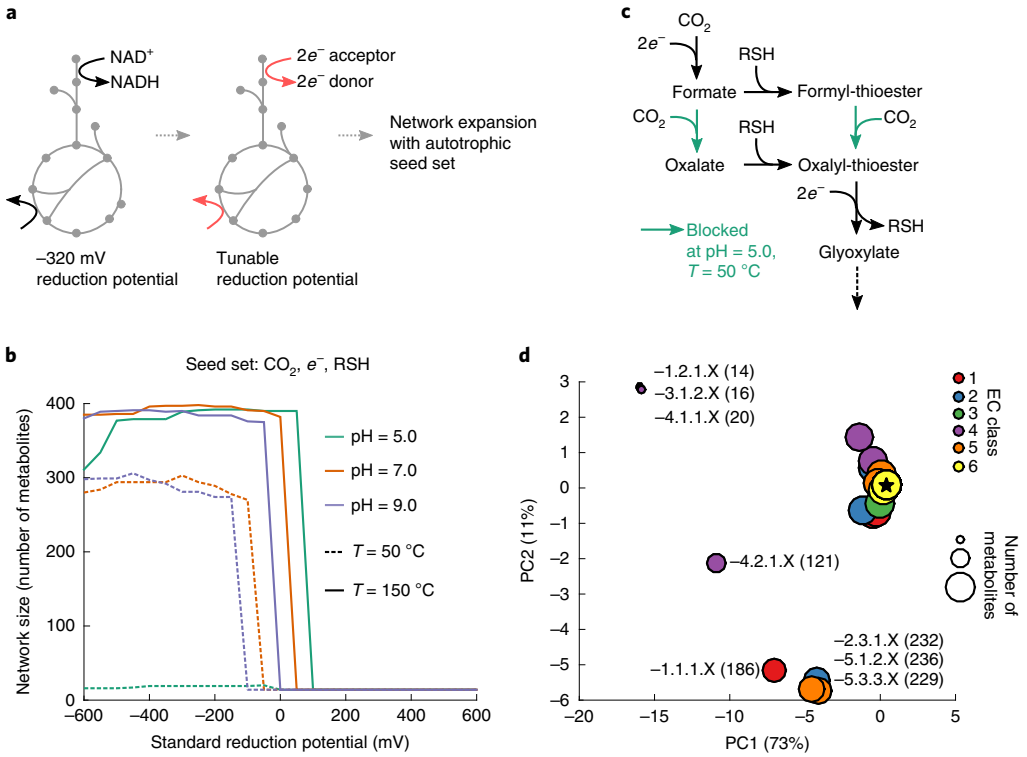


Fig. 2 | Primitive redox systems and reaction classes constrain network expansion from CO_2. **a**, Redox coenzymes (NAD, NADP and FAD) were substituted with an arbitrary electron donor/acceptor at a fixed reduction potential. **b**, We performed thermodynamic network expansion in acidic (pH 5), neutral (pH 7) and alkaline (pH 9) conditions at two temperatures ($T = 50$ and 150 °C) using a two-electron redox couple at a fixed potential (x axis) as a substitute for NAD(P)/FAD coupling in extant metabolic reactions (Methods). We plotted the final network size across all pH and temperatures with no fixed carbon sources (for example, only CO_2) and thiols. Notably, for these simulations we used a base seed set of H_2, H_2S, H_2O, HCO_3^-, H^+ and CO_2. In general, network size is increased at higher temperatures, consistent with Weiss et al.[13]. **c**, Analysis of the expanding networks revealed a critical set of reactions blocked at pH 5 and $T = 50$ °C, preventing the production of oxalyl-thioesters and subsequently glyoxylate. These reactions belong to EC class 4.1.1.X. **d**, Results of network expansion after removing groups of reactions based on EC codes. The ensuing networks are visualized using principal component analysis. Specifically, each network is plotted as a circle in the plane of the first two principal components (PC1 and PC2), which explain 73% and 11% of the variance respectively. The colour of the circle represents the EC code removed from the original network, and the size corresponds to the final number of metabolites in the expanded network. Points are labelled by EC class, and the number of metabolites in the perturbed networks are provided in parentheses. The star represents the location of the unperturbed network.

Constraint-based flux modelling of proto-metabolism. So far, we have focused only on the topology and local thermodynamic feasibility of putative ancient metabolic networks. Inspired by recent studies on the molecular budget of present-day cells[31,52,53], we decided to further explore whether proto-metabolic networks could support thermodynamically feasible steady-state fluxes and fuel primitive protocells with internal energy sources (for example, thioesters), redox gradients and primitive biopolymers that are capable of catalysis and compartmentalization. Flux balance analysis (FBA), originally developed for the study of microbial metabolism, enables the prediction of systems-level properties of metabolic networks at steady-state[52]. Fundamentally, FBA computes possible reaction rates in a network constrained by mass and energy balance, usually under the assumption that a specific composition of biomolecules is efficiently produced during a homeostatic growth process. In microbial metabolism, FBA is used to simulate the production of cellular biomass (for example, protein, lipids and nucleic acids) at fixed proportions, which are derived from a known composition of extant cells. The same approach could help test the sustainability of a proto-metabolic biochemical system, provided that we could develop a plausible hypothesis for the 'biomass composition' of ancient protocells. As a starting point, we recalled de Duve's suggestion that the thioester-driven polymerization of monomers produced from ancient proto-metabolism may have led to 'catalytic multimers,' which could have served as catalysts for ancient biochemical reactions[4]. Under nitrogen-limited conditions, keto acids produced from proto-metabolism (Fig. 3b) could have been reduced to α-hydroxy acids and polymerized into polyesters using thioesters as a condensing agent (Extended Data Fig. 6). Recent work has suggested that polymers of α-hydroxy acids may have been produced in geochemical environments[54] and that these molecules could have served as primitive catalysts[55]. These results all point to the intriguing possibility that the thioester-driven polymerization of α-hydroxy acids (produced from keto acid precursors of common amino acids) generated the first metabolically sustainable cache of ancient catalysts, leading to a collectively autocatalytic protocellular system. We employed a variant of FBA to specifically test the feasibility of such a system. Using an expanded metabolic network as a scaffold for network reconstruction (Fig. 3b) we constructed a constraint-based model of an ancient protocell using a biomass composition consisting of fatty acids (for protocellular membranes), 'catalytic multimers' derived from eight keto acids (Fig. 4a), and redox and thioester-based free energy sources (Methods, Fig. 4a). We used thermodynamic metabolic flux analysis (TMFA), a variant of FBA that explicitly includes thermodynamic constraints[53] (Methods), to determine whether homeostatic growth of the whole system was achievable (Methods). We found that to obtain feasible production of each keto acid and fatty acid precursor, the model required an

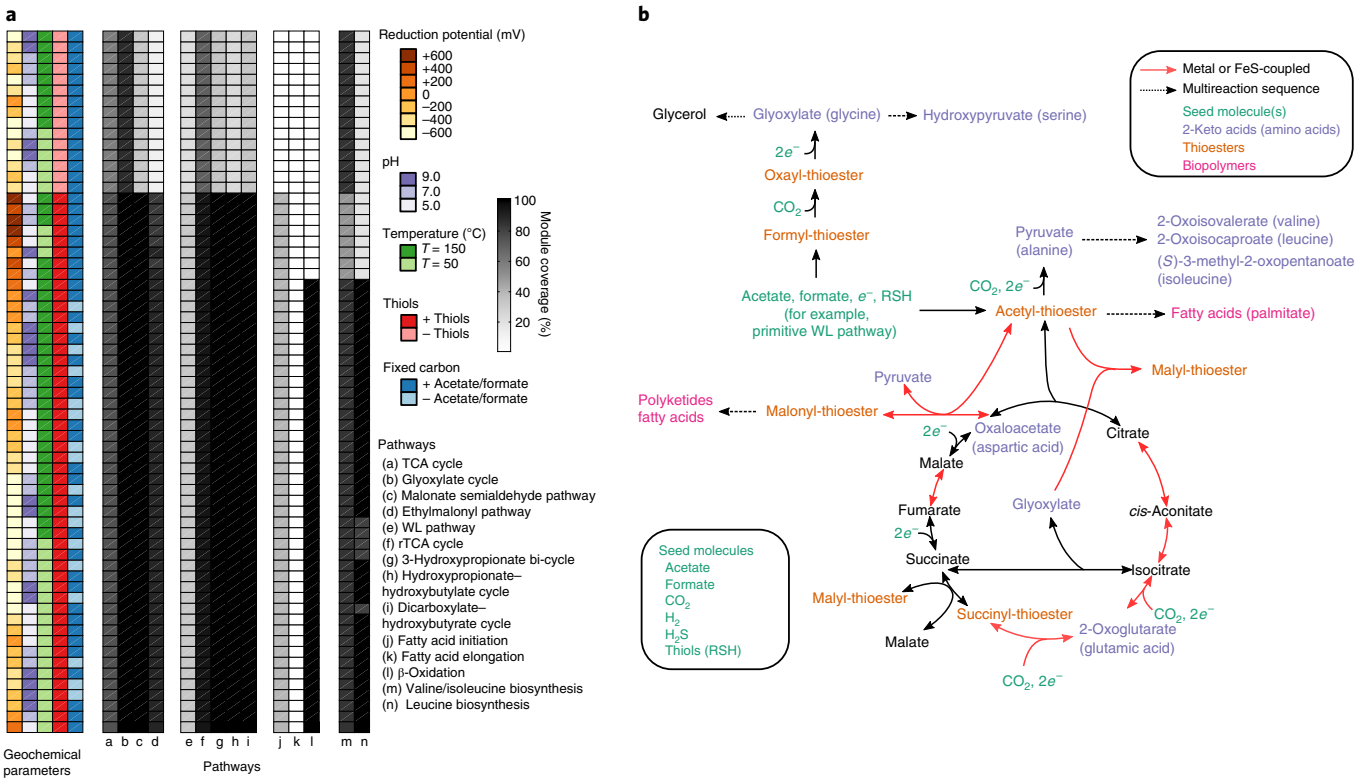


Fig. 3 | Systematic exploration of prebiotic scenarios reveals a core organo-sulfur network. a, A thermodynamically constrained network expansion algorithm was used to simulate the early expansion of proto-metabolism under various scenarios, including the availability of reductants in the environment, pH, temperature and the availability of fixed carbon sources and thiols (represented in the five colour-coded heatmap columns of geochemical parameters). For a selected set of KEGG modules involved in core carbon metabolism (*x* axis), the proportion of molecules within each pathway is plotted as a greyscale heatmap to the right of the geochemical parameters. **b**, A representation of the core network produced from a prebiotically plausible seed set without nitrogen or phosphate (bottom left box). Acetyl-thioesters are first produced, potentially from a primitive WL pathway[8,19] from acetate and thiols provided as seed molecules (green). Acetyl-thioesters enable the production of all intermediates in the rTCA cycle, with the exception of phosphoenolpyruvate. ATP-dependent reactions in the rTCA cycle may have been substituted with a primitive malate synthase and transthioesterification of succinate, as well as the recently discovered reversible citrate synthase[73,74]. The keto acid precursors for eight common amino acids (A, D, E, G, I, L, S, V) are in purple, and routes to thioester-mediated polymerization of fatty acids and polyketides are in pink.

internal redox system with a fixed potential of between −500 and −200 mV, where very low reduction potentials led to an inability to produce 2-oxoisocaprioate (Fig. 4b). We fixed the potential of the internal redox system to −220 mV and determined whether steady-state growth was achievable with a variety of electron donors, electron acceptors and carbon sources (Fig. 4c,d). We found that growth of the protocell metabolic model is feasible under a wide variety of assumptions in relation to macromolecular compositions, carbon sources, electron donors and acceptors (Fig. 4c,d). Notably, growth is achievable in simple chemoautotrophic conditions with either H_2 or glutathione (or a free thiol) as electron donors (Fig. 4c), which is consistent with recent work suggesting the last universal common ancestor was a thermophilic H_2-consuming chemoautotroph[13]. In this model, thiols and thioesters are not supplied as food sources, but rather are recycled during steady-state growth of the protocell. This reflects the possibility that thiols could have been initially supplied abiotically, followed by the rapid takeover of biotic production of mercaptopyruvate, a keto acid that could have been incorporated into primitive multimers.

Discussion

Although most efforts to reconstruct ancient biochemistry have traditionally relied on building qualitative models of small pathways[2,4,5,7,21], we found that quantitative modelling of larger networks can provide substantial new insight into the origin of life. By computationally mapping geochemical scenarios to plausible ancient proto-metabolic structures we estimated which portions of extant biochemistry may have been very sensitive or very robust to initial geochemical conditions. Our approach reveals that, contrary to expectations[8,11,25,26], environmental sources of fixed nitrogen and low-potential electron donors may have not been necessary for early biochemical evolution and a substantial degree of complexity may have emerged prior to incorporation of nitrogen into the biosphere[3]. The key catalytic role played by nitrogen in the active sites of modern enzymes may have been preceded by positively charged surfaces or metal ions[19,21,41,56], which could have been replaced by amino/keto acids with nitrogen side-chains once nitrogen became incorporated into proto-metabolism. Our simulations also cast doubts on the essential role of a low-potential electron donor in early life[8,11,25,26], consistent with the proposal that low-potential electron donors may not be necessary for acetogenesis[57], and with the possibility that energy conservation via electron bifurcation might not have been necessary in primordial metabolism[58]. The independence of our inferred ancestral networks of low-potential electron donors and ATP, both key substrates for nitrogen fixation[59], suggests that nitrogen fixation may have evolved later throughout the history of life[13,60]. A striking feature of our analysis is the convergence of multiple geochemical scenarios towards a core organo-sulfur proto-metabolic network capable of producing various keto acids and fatty acids (Fig. 3b), and potentially providing a metabolic flow of molecular

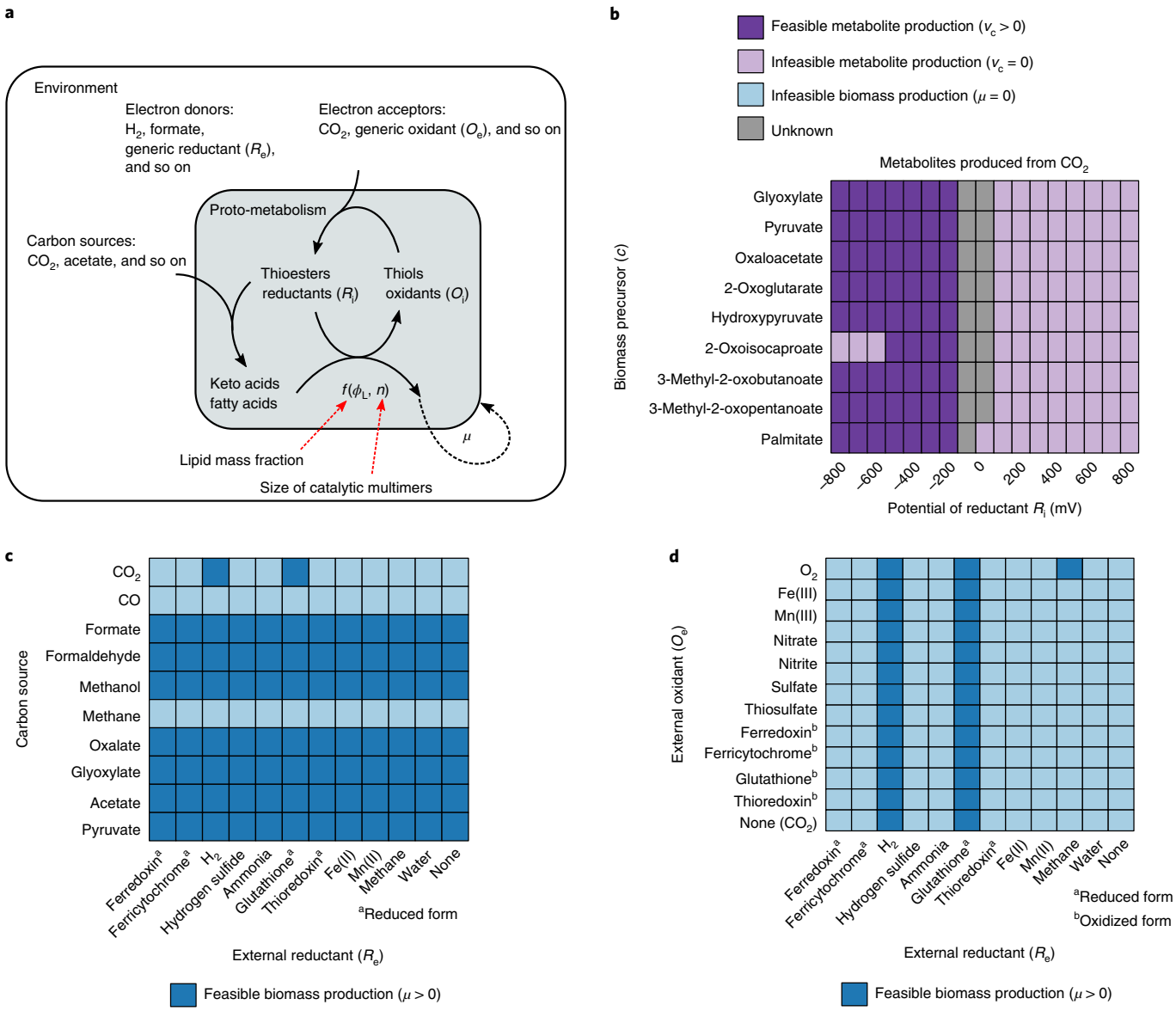


Fig. 4 | Constraint-based modelling of plausible ancient protocells. a, We constructed a metabolic model of a plausible ancient protocell and used TMFA[53] to simulate the feasibility of steady-state growth under a variety of environmental conditions. The metabolic model was constructed using internally generated reductants (R_i), oxidants (O_i) and thioesters that fuelled biomass formation, as well as externally supplied carbon sources, external reductants (R_e) or external oxidants (O_e). The biomass composition was specified as variable fractions of fatty acids and polymerized hydroxy acids from keto acid precursors (Methods). In this model, the internal redox coenzyme was assumed to be at a single fixed standard reduction potential, and the production of biomass was fuelled by the hydrolysis of acetyl-thioesters. We described the biomass composition parameter using a two-parameter model ($f(\Phi_L, n)$ Methods), with the mass fraction of lipids in the protocell set to $\phi_L = 0.1$, and the average size of a catalytic multimer, $n = 10$. **b**, We computed fluxes (v_c) from CO_2 to each biomass precursor c (y axis) using a variety of internally generated reductants at various reduction potentials (x axis). This shows the conditions that led to feasible (dark purple) and infeasible (light purple) flux. Note that some cases did not converge to a solution within the allocated maximal CPU (central processing unit) time, probably due to numerical issues, and were thus classified as 'unknown' (grey). The production of all biomass precursors was feasible if the redox system was between −500 and −200 mV. **c**, Next we simulated growth on a variety of simple carbon sources (y axis) and external electron donors (x axis). Environments supporting non-zero growth are dark blue and those supporting no growth are light blue. Interestingly, H_2 and glutathione were the only reductants capable of supporting fully autotrophic growth on CO_2. Furthermore, CO and methane could not support growth in this model, while the other one-carbon sources, such as methanol, formate and formaldehyde, could support biomass growth. **d**, We then simulated autotrophic growth using both an external oxidant (y axis) and an external reductant (x axis). Environments supporting non-zero growth are shown in dark blue while those displaying no growth are shown in light blue. Feasible growth was entirely dependent on the reductant, rather than the oxidant, except for when methane was the electron donor and oxygen was the electron acceptor. For models in **c** and **d**, the internal redox coenzyme was assumed to be disulfide/dithiol at a standard reduction potential of −220 mV.

substrates for catalysis and self-aggregation[61]. In particular, this feature provides a window into how thioester-driven polymerization of α-hydroxy acid monomers (derived from producible keto acids) could have added primitive macromolecular organic catalysts[4] to initial inorganic minerals or metal ion catalysts[19,41,42]. Further tests of this hypothesis could be pursued by measuring the capacity of these

polymers to catalyze key reactions in the network and by exploring whether these organic compounds are produced in living systems today via mechanisms similar to polyketide or non-ribosomal peptide synthesis. Additionally, the fact that the network expansion is significantly affected by the removal of reactions involved in fatty acid metabolism (Fig. 2d) suggests that future experimental efforts could be directed towards identifying non-enzymatic mechanisms for fatty acid synthesis. Finally, our constraint-based models of this core organo-sulfur proto-metabolism provide an example of how network expansion-based predictions can be translated into dynamical models, whose capacity to estimate sustainable collective growth may drive the search for specific self-reproducing chemical networks and metabolically driven artificial protocells.

Future models of early metabolic systems could be used to estimate the outcome of evolutionary competitions amongst different networks, similar to what has been done for stoichiometric models of bacterial metabolism (where the biomass production is used as a proxy for fitness[62,63]). In order to enable similar simulations, however, it would be necessary to obtain realistic estimates of the kinetics of nutrient inflow, equivalent to uptake rates in present-day cells. Moreover, to perform simulations that are based on thermodynamically feasible metabolic states, one would have to address some of the current challenges we faced with TMFA calculations of proto-metabolic networks due to computational complexity of mixed-integer optimization problems for large networks (Methods). Once these challenges have been addressed, future stoichiometrically based eco-evolutionary models of proto-metabolism could help generate specific testable hypotheses about the ancient biosphere.

Future research could also cover both specific physico-chemical hypotheses presented in this study, as well as fundamental conjectures implicit in our modelling approach. Our approach assumes that the history of metabolic evolution can be reconstructed by the extant biosphere-level metabolic network, which is primarily catalyzed by genome-encoded enzymes. Future studies could refine this assumption by adding potentially 'extinct' reactions to the model reconstructed using alternative computational methods[64,65] or removing kinetically limited reactions using experimental data. Although the removal of reactions could dramatically limit the composition of expanded networks (Fig. 2d), the addition of reactions to the model would not change the principal conclusion that biomolecules previously assumed to be critical for the emergence of living systems (for example, phosphate, fixed nitrogen and low-potential electron donors) may not have been essential for the onset of proto-metabolic systems. However, we would expect the stoichiometric modelling results to be sensitive to additions of new reactions because these could turn currently infeasible states into feasible ones. Overall, the striking concordance between the theory presented in this study and recent experimental models of proto-metabolism[42] suggests that extant metabolism might serve as an approximation of abiotic chemical networks, thus providing a window into the earliest phases of biochemical evolution prior to a genetic coding system.

Methods

Reconstruction of biosphere-level metabolic network. Biosphere-level metabolism was reconstructed from the KEGG database[50] according to a protocol described previously[30]. We modified the network in several ways to model primitive thioester-based metabolic networks without nitrogen or phosphate. First, to simulate the availability of thiols capable of forming thioesters, we included coenzyme A, acyl-carrier protein and glutathione into the seed set. However, to enforce the constraint that these metabolites could only be used in reactions as coenzymes (and not products or substrates), we prevented the degradation by removing KEGG reactions R10747, R02973 and R02972.

We then assigned standard molar free energies to reactions using eQuilibriator at a predefined pH[66]. Next we substituted NAD, NADP and FAD-coupled reactions with an arbitrary redox couple. For example, if the redox reaction $X_{ox} + NADH \rightarrow X_{red} + NAD^+$ was swapped with electron donor with a redox potential of E_0^+ mV, we would use the following formula to adjust the standard molar free energy for the new reaction r':

$$\Delta_r G^{\circ\prime} = \Delta_r G^{\circ\prime} + nF(E_0^+ - E_0) \quad (1)$$

where n is the number of electrons transferred in reaction r and $F = 96.485\,\mathrm{kJ\,V^{-1}}$. Note that if we assumed that the electron donor/acceptor substitute was a two-electron donor/acceptor, we did not change the stoichiometry in the reaction equation. However, in the case where the electron donor/acceptor substitute was a single electron donor/acceptor, we changed the stoichiometric coefficients to $s_{cj} = 2$ for all reactions j, where c represents metabolites NAD(H), NADP(H) and FAD(H_2). For this work, we systematically varied the reduction potential E_0^+ and stoichiometry of the primitive redox coenzyme.

Thermodynamically constrained network expansion. We performed network expansion using thermodynamic constraints in a different way than performed before[30]. Previously, reactions above a predefined free energy threshold of $\tau = 30\,\mathrm{kJ\,mol^{-1}}$ were removed[30]. However, for this work we calculated the lowest reaction free energy possible using estimates for upper (u_i) and lower (l_i) bounds on metabolite concentrations and removed reactions with a positive reaction free energy. For a given biochemical reaction at fixed temperature and pressure, $\Delta_r G'$ is defined as:

$$\Delta_r G' = \Delta_r G^{\circ\prime} + RT\ln\Pi_i a_i^{s_{ir}} \quad (2)$$

where the $\Delta_r G^{\circ\prime}$ 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_{ir} is the stoichiometric coefficient for metabolite i in reaction r. We fixed a_i for each reaction according to the following rules:

$$s_{ir} < 0 \Rightarrow a_i = u_i$$
$$s_{ir} > 0 \Rightarrow a_i = l_i$$

We then removed reactions with a $\Delta_r G' \geq 0$. For all simulations we assumed that $u_i = 10^{-1}$ M and $l_i = 10^{-6}$ M. Note that because we model each reaction independently, metabolite concentrations could be inconsistent. For instance, if metabolite i is the substrate for reaction p and a product for reaction q, then $a_i = u_i$ for reaction p and $a_i = l_i$ for reaction q. Additionally, a fundamental assumption of this algorithm is that over long time-scales, network growth is constrained by 'local' thermodynamic bottlenecks for each reaction individually, rather than 'global' thermodynamic feasibility of the entire network. We also assume that during the expansion the enthalpic portion of each reaction's free energy is constant because the primary physico-chemical changes that could change the enthalpy of formations (for example, pH, ionic strength) are buffered by geochemical boundary conditions.

After using this procedure to systematically remove reactions that were considered to be thermodynamically infeasible, we performed network expansion[32–34] as described in ref. [30].

Parameters for network expansion. We systematically studied the size and composition of networks under precise environmental conditions by varying (1) the reduction potential from the environment, (2) pH, (3) temperature, (4) the presence or absence thiols, (5) the inclusion of fixed carbon into the seed set and (6) the inclusion of fixed nitrogen into the seed set. We now discuss each of these parameters in more detail:

(1) Reduction potential and stoichiometry. A wide range of environmental conditions could have provided electron donors at various potentials: high potential redox pairs, with strong oxidants such as Fe(III), may have been present in oceans at high concentrations, while strong reductants such as H_2, disulfides, protoferredoxin or reductive carboxylation of thioesters may have been produced via serpentinization or geochemical analogues of primitive metabolic pathways[25]. We substituted reactions coupled to NAD, NADP and FAD with a generic single or double electron donor and acceptor pair at a fixed potential. To prevent unbalanced electron transfer, we removed the following transhydrogenase reactions: R10159, R01195, R00112, R09520, R09748, R05705, R05706, R09662 and R09750. We then created a single or double electron donor/acceptor pair with a fixed reduction potential, E_0^+, ranging from –600 to 600 mV. Note that network expansion was performed by adding either the generic oxidant or reductant for NAD(P)/FAD -coupled reactions into the seed set directly, which assumes that this redox system could be produced abiotically.

(2) pH. We modified the pH by setting reaction free energies at various pH (5.0–9.0) using eQuilibriator[66], which relies on the component contribution method[67].

(3) Temperature. Temperatures were assumed to have been within a range of 50–150 °C, spanning estimates of ocean seawater temperature in the Archaean[68], up to some alkaline hydrothermal vent systems[11].

(4) Thiols. In our model we provided thiols that were substitutes for coenzymes that form thioester bonds in extant metabolic networks. We provided coenzyme A, acyl-carrier protein and glutathione in the seed set, but removed key

degradation reactions to ensure these compounds only served as coenzymes, rather than material sources, during network expansion[30].

(5) Fixed carbon. We modelled the dependence of the simulated proto-metabolic network on fixed carbon by supplying a set of fixed carbon sources consisting of formate, acetate and CO_2 in the seed set. For simulations with no fixed carbon, we only provided CO_2 as a carbon source.

(6) Fixed nitrogen. To study the consequences of adding or removing a source of fixed nitrogen as a seed compound for network expansion, we either added or removed ammonia from the seed set prior to expansion.

In addition to the parameters we varied as described above, our simulations include two additional parameters that we kept constant in the current analysis, but whose effects could be studied in future work:

(1) Metabolite concentrations. Metabolite concentrations were assumed to be within 1 μM – 100 mM. The upper bound estimate is consistent with recent experimental data showing that key metabolites (formate, methanol, acetate and pyruvate) can be produced near 100 mM[19]. Although we do not have empirical evidence to suggest a reasonable lower bound on metabolite concentrations in ancient metabolic networks, we assumed that 1 μM, the estimated lower bound in today's cells[69], was also the lower bound in our model of ancient metabolism.

(2) Inclusion of reactions with no free energy estimate. We found that 53% of the biosphere-level metabolic network reactions had no free energy estimate (4,851 out of 9,074). For all simulations presented in this paper, we assumed these reactions were blocked and did not include them in the network.

Generalized linear modelling of network expansion results. To assess the effects of various parameters on the outcome of network expansion we used generalized linear models to construct logistic regression classifiers. These were used to predict whether or not the network expanded beyond 100 metabolites based on a combination of predictors, including categorical variables that encoded whether or not ammonia, thiols or fixed carbon was provided in the seed set, and also continuous variables that encoded the reduction potential, pH and temperature in each simulation. We defined the response variable for simulation k as y_k, where $y_k=1$ if the simulation resulted in a network that expanded beyond 100 metabolites, and $y_k=0$ if otherwise. For the set of simulations performed in Fig. 1 we constructed a design matrix consisting of categorical variables representing the following scenarios:

- $x_{N,k}\in\{0,1\}$: 1 if ammonia was included in the seed set, and 0 otherwise.
- $x_{S,k}\in\{0,1\}$: 1 if thiols were included in the seed set, and 0 otherwise.
- $x_{C,k}\in\{0,1\}$: 1 if fixed carbon was included in the seed set, and 0 otherwise.
- $x_{H,k}\in\mathrm{R}_{>0}$: A continuous variable representing the pH. Note for our simulations, we only explored acidic (pH 5), neutral (pH 7) and alkaline (pH 9) regimes.
- $x_{E,k}\in\mathrm{R}$: A continuous variable representing the reduction potential at standard molar conditions (at the specified pH listed above). For our simulations we explored a wide range of standard molar reduction potentials (from −600 mV to +600 mV).
- $x_{Tk}\in\mathrm{R}_{>0}$: A continuous variable representing the temperature. For our simulations, we explored two temperatures: a high temperature regime ($T=150$ °C) and a low temperature regime ($T=50$ °C).

We then constructed the following generalized linear model to determine whether the network expanded beyond metabolites:

$$\mathrm{logit}(y_k)=\beta_0+\beta_N x_{Nk}+\beta_S x_{Sk}+\beta_C x_{Ck}+\beta_H x_{Hk}+\beta_E x_{Ek}+\beta_T x_{Tk} \quad (3)$$

where the subscripts N, S, C, H, E and T correspond to nitrogen, sulfur, carbon, pH, reduction potential and temperature covariates, respectively. We fit the parameters (β_0, β_N, β_S, β_C, β_H, β_E, β_T) using the 'fitglm.m' function in MATLAB 2015a, and a receiver operating curve was generated using the perfcurve.m function. To generate Fig. 1c, individual predictors were removed one by one in the generalized linear model presented above. To assess whether the trained logistic model served as an accurate classifier, we performed leave-one-out cross-validation by removing individual samples from the training set and testing the accuracy of the trained classifier on the removed sample. This procedure resulted in a cross-validation accuracy of 0.89.

Constraint-based modelling. We constructed a model of an autocatalytic network at steady-state using a variant of constraint-based modelling of cellular metabolism called TMFA[53]. TMFA transforms the nonlinear constraints induced by imposing thermodynamic consistency into mixed-integer linear constraints. In the next section, we describe the construction of primitive biomass composition for a model of an ancient protocell and also the TMFA formula used in this analysis.

Prebiotic biomass equation. We constructed a simple model for the macromolecular composition of primitive protocells using empirical knowledge of extant cellular life. Our metabolic model of proto-metabolism does not include macromolecular production of nucleotides (and thus a nucleic acid-based genetic system) and, therefore, we assume that the primary role of protocellular metabolism was to initially produce components for a cellular membrane and catalysts. Building on de Duve's multimer hypothesis[4] we propose that the biomass can be constructed using a simple two parameter model consisting of the mass fraction of lipids (ϕ_L) and the average length of each catalytic multimer (n).

- Lipid mass fraction (Φ_L). The lipid content in modern cells is roughly 10% of the total dry mass (BioNumbers ID: 111209)[70], primarily composed of the fatty acid palmitate. For our analysis we assume that palmitate represents the sole component of lipids. Future models could incorporate glycerol, which enables the production of glycerolipids. Although phosphate is used in cellular membranes as a polar head group to produce amphiphilic molecules, primitive processes may have conjugated negatively charged organic acids (for example, oxalate) to glycerol via a thioester-mediated synthesis mechanism to create amphiphilic lipid molecules resembling modern phospholipids. For our initial model we propose that palmitate was the initial amphiphilic component of primitive membranes, where the negatively charged polar carboxylate ion was sufficient for forming a membrane, and we assume that protocells consisted of a lipid mass fraction of ϕ_L.
- Catalytic multimer mass fraction (Φ_c). We assume here that ancient catalysts were composed of inorganic molecules (for example, iron–sulfur clusters, metal ions, mineral surfaces) chelated with multimers of α-hydroxy acids (Fig. 4a). For our model we assume that the eight keto acid precursors produced from our network were the dominant monomers of ancient multimeric catalysts. We assume that the total mass fraction of these catalysts is $1-\phi_L=\phi_C=\sum_k\phi_k$, where ϕ_k is the mass fraction of polymerized monomer k. For our analysis we assume that each monomer is uniformly distributed within the biomass, so that ϕ_k is constant for all k. Additionally, because each monomer must be reduced to α-hydroxy acids, there is a linear relationship between the electron demand, s_e, and the number of molecules of monomers produced. The stoichiometric equivalents of electron donors are thus: $s_e=2\sum_k\frac{\phi_k}{M_k}$ where M_k is the molar mass of monomer k.
- Average size of catalytic multimers. The average size of multimeric catalysts sets the number of thioester bonds required for synthesis of catalytic multimers. For each polymer of size n, there are $n-1$ thioester bonds required for synthesis. In our model, the total number of monomers are fixed: $\sum_k\frac{\phi_k}{M_k}$, where M_k is the molar mass of monomer k. Thus, for a fixed monomer length n, we can calculate the number of polymers using the following formula: $P(n)=\frac{1}{n}\sum_k\frac{\phi_k}{M_k}$. The thioester demand is $s_t(n)=(n-1)P(n)$ or $s_t(n)=\frac{n-1}{n}\sum_k\frac{\phi_k}{M_k}$. For our analysis we assumed a fixed polymer length of size $n=10$ monomers.

Using these two parameters, we constructed the biomass equation for the protocellular model. Note that the electron source and sink were provided by an unspecified internal redox coenzyme system (analogous to NAD(P)/FAD).

TMFA. To simulate a thermodynamically feasible steady-state of this metabolic network we used a variant of TMFA[53]. Briefly, TMFA transforms the nonlinear constraints induced by imposing thermodynamic consistency into mixed-integer linear constraints. We first converted the model into an irreversible model by separating each reaction into a pair of irreversible forward and backward reactions. We then constructed the following mixed-integer linear programme to find a flux vector, v (with elements v_r for each reaction r), log-transformed metabolite concentrations ($\ln(x)$) and binary variables indicating whether a reaction is feasible (z) given a specific objective function was satisfied.

Past implementations of TMFA in microbial metabolism defined the objective function as maximizing biomass yield. However, for our study we determined whether non-zero growth was feasible and, therefore, we transformed TMFA into a constraint–satisfaction problem by setting a lower bound on the biomass reaction, v_{biomass}, such that $v_{\mathrm{biomass}}\geq\mu_{\min}$ and solving the following mixed-integer linear programme:

$$\text{maximize}_{v,\ln(x),z,e}\ 0$$

subject to

$$Sv=0 \quad (4)$$

$$0\leq v_r\leq z_r ub_r,\forall r\in R \quad (5)$$

$$z_rK-K+\Delta_rG'<0,\forall r\in R \quad (6)$$

$$\Delta_rG^{o\prime}+RT\sum_i s_{ir}\ln(x_i)+\sigma_r e_r<0,\forall r\in R \quad (7)$$

$$\ln(10^{-6})\leq\ln(x_i)\leq\ln(10^{-1}),\forall i\in M \quad (8)$$

$$-\sigma_m\leq\sigma_r\leq\sigma_m,\forall r\in R \quad (9)$$

$$v_{\mathrm{biomass}}\geq\mu_{min} \quad (10)$$

where R and M are the sets of all reactons and metabolites, respectively. As discussed in detail elsewhere[53], equation (4) in the constraint set ensures that intracellular metabolite concentrations are at steady-state and are mass balance constraints for each metabolite. Equation (5) sets the bound on individual reaction fluxes, where the maximum flux through reaction r is ub_r. Note that when $z_r = 0$, the flux through reaction r is constrained to 0. Equation (6) ensures that $z_r = 1$ if and only if $\Delta_r G' < 0$, and $z_r = 0$ otherwise. Note that K is a large number $K > \max_r\{\Delta_r G'\}$ ensuring that this constraint is not violated with $z_r = 0$. Equation (7) is the free energy of each reaction as a function of log-metabolite concentrations, where R is the ideal gas constant, T is temperature and s_{ir} is the stoichiometric coefficient for metabolite i in reaction r. Note that we also add slack variables, e_r, to account for the possible error in estimating the standard molar reaction free energies for each reaction (where σ_r is the standard error for each reaction r), which are bound by a global error tolerance $\sigma_m = 0$ (set in equation (9)). Note that if this global tolerance is >0, thermodynamic infeasible cycles are possible in steady-state solutions. Equation (8) constrains the log-metabolite concentrations to be bounded between 1 μM and 100 mM. For each simulation we constrained the uptake reactions to be 1×10^4 and the lower bound on biomass production to be $\mu_{min} = 1$.

Numerical simulations were performed using the COBRA (constraint-based reconstruction and analysis) toolbox[71] and the Gurobi optimizer (v.7.0.1). All source code is provided in the following github repository: https://github.com/segrelab/BoundaryConditionsForAncientMetabolism.

Calculation of coenzyme and sequence-level features within enzymes. To determine which reactions were associated with specific coenzymes (for the results presented in Extended Data Fig. 1) we accessed information for each EC number in the KEGG ENZYME database (http://www.genome.jp/kegg/annotation/enzyme.html). We downloaded each page and parsed the 'comment' field for each EC and performed a text-based search to identify coenzymes associated with each EC number. We searched for text indicating that the enzyme mechanisms used one of the following coenzymes (biotin, haem, PLP, TPP, pterin, molybdopterin, flavin), cofactors (Fe, Co, Ni, Cu, Mn, W, Zn, Mo, Mg) or iron-sulfur clusters (FeS, FeFe, Fe_2S_2, Fe_3S_4 and Fe_4S_4). We also searched EC numbers indicating that the reaction mechanisms are non-enzymatic. Text-based searches were reduced manually to remove mis-annotated enzyme–coenzyme relationships.

For Extended Data Fig. 1b, we calculated the fraction of reaction EC numbers associated with a specific coenzyme (Fe, Co, Ni, Cu, Mn, W, Zn, Mo, Mg, FeS, FeFe, Fe_2S_2, Fe_3S_4 and Fe_4S_4) or was marked as non-enzymatic. To obtain the results shown in Extended Data Fig. 1d, we determined the fraction of reaction EC numbers associated with one of the following coenzymes: biotin, haem, PLP, TPP, pterin, molybdopterin and flavin.

For results in Extended Data Fig. 1e, we obtained a database of known enzyme active-site residues[72]. We first mapped the network reactions to EC numbers listed in KEGG, and then identified active sites corresponding to EC numbers within the expanded network. We calculated the fraction of active-site residues containing nitrogenous side chains derived from the following amino acids: arginine (R), lysine (K), glutamine (Q), asparagine (N), histidine (H) and tryptophan (W).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data presented in this paper have been deposited in a public repository and can be accessed at https://github.com/segrelab/BoundaryConditionsForAncientMetabolism.

Code availability

All code presented in this paper can be accessed at https://github.com/segrelab/BoundaryConditionsForAncientMetabolism.

Received: 4 January 2019; Accepted: 27 September 2019;
Published online: 11 November 2019

References

1. 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).
2. Hartman, H. Speculations on the origin and evolution of metabolism. *J. Mol. Evol.* **4**, 359–370 (1975).
3. Hartman, H. Conjectures and reveries. *Photosynth. Res.* **33**, 171–176 (1992).
4. de Duve, C. *Blueprint for a Cell: The Nature and Origin of Life* (Neil Patterson Publishers, 1991).
5. Morowitz, H. J., Kostelnik, J. D., Yang, J. & Cody, G. D. The origin of intermediary metabolism. *Proc. Natl Acad. Sci. USA* **97**, 7704–7708 (2000).
6. Smith, E. & Morowitz, H. J. *The Origin and Nature of Life On Earth* (Cambridge Univ. Press, 2016).
7. Smith, E. & Morowitz, H. J. Universality in intermediary metabolism. *Proc. Natl Acad. Sci. USA* **101**, 13168–13173 (2004).
8. Sousa, F. L. et al. Early bioenergetic evolution. *Phil. Trans. R. Soc. Lond. B* **368**, 20130088 (2013).
9. Lazcano, A. & Miller, S. L. The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. *Cell* **85**, 793–798 (1996).
10. Deamer, D. & Weber, A. L. Bioenergetics and life's origins. *Cold Spring Harb. Perspect. Biol.* **2**, a004929 (2010).
11. Martin, W. & Russell, M. J. On the origin of biochemistry at an alkaline hydrothermal vent. *Phil. Trans. R. Soc. Lond. B* **362**, 1887–1926 (2007).
12. Martin, W., Baross, J., Kelley, D. & Russell, M. J. Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* **6**, 805–814 (2008).
13. Weiss, M. C. et al. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016).
14. Russell, M. J., Hall, A. J. & Martin, W. Serpentinization as a source of energy at the origin of life. *Geobiology* **8**, 355–371 (2010).
15. McDermott, J. M., Seewald, J. S., German, C. R. & Sylva, S. P. Pathways for abiotic organic synthesis at submarine hydrothermal fields. *Proc. Natl Acad. Sci. USA* **112**, 7668–7672 (2015).
16. Parker, E. T. et al. Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark discharge experiment. *Proc. Natl Acad. Sci. USA* **108**, 5526–5531 (2011).
17. Heinen, W. & Lauwers, A. M. Organic sulfur compounds resulting from the interaction of iron sulfide, hydrogen sulfide and carbon dioxide in an anaerobic aqueous environment. *Orig. Life Evol. Biosph.* **26**, 131–150 (1996).
18. Cody, G. D. Primordial carbonylated iron–sulfur compounds and the synthesis of pyruvate. *Science* **289**, 1337–1340 (2000).
19. Varma, S. J., Muchowska, K. B., Chatelain, P. & Moran, J. Native iron reduces CO_2 to intermediates and end-products of the acetyl-CoA pathway. *Nat. Ecol. Evol.* **2**, 1019–1024 (2018).
20. Huber, C. Activated acetic acid by carbon fixation on (Fe,Ni)S under primordial conditions. *Science* **276**, 245–247 (1997).
21. Wachtershauser, G. Evolution of the first metabolic cycles. *Proc. Natl Acad. Sci. USA* **87**, 200–204 (1990).
22. Fuchs, G. Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.* **65**, 631–658 (2011).
23. Dörr, M. et al. A possible prebiotic formation of ammonia from dinitrogen on iron sulfide surfaces. *Angew. Chem. Int. Ed.* **42**, 1540–1543 (2003).
24. Navarro-González, R., McKay, C. P. & Mvondo, D. N. A possible nitrogen crisis for Archaean life due to reduced nitrogen fixation by lightning. *Nature* **412**, 61–64 (2001).
25. Martin, W. F. & Thauer, R. K. Energy in ancient metabolism. *Cell* **168**, 953–955 (2017).
26. Sousa, F. L., Preiner, M. & Martin, W. F. Native metals, electron bifurcation, and CO_2 reduction in early biochemical evolution. *Curr. Opin. Microbiol.* **43**, 77–83 (2018).
27. Halmann, M. in *The Origin of Life and Evolutionary Biochemistry* (eds Dose, K. et al.) 169–182 (Springer, 1974).
28. Schwartz, A. W. Phosphorus in prebiotic chemistry. *Phil. Trans. R. Soc. Lond. B* **361**, 1743–1749 (2006).
29. Keefe, A. D. & Miller, S. L. Are polyphosphates or phosphate esters prebiotic reagents? *J. Mol. Evol.* **41**, 693–702 (1995).
30. Goldford, J. E., Hartman, H., Smith, T. F. & Segrè, D. Remnants of an ancient metabolism without phosphate. *Cell* **168**, 1126–1134 (2017).
31. Goldford, J. E. & Segrè, D. Modern views of ancient metabolic networks. *Curr. Opin. Syst. Biol.* **8**, 117–124 (2018).
32. Ebenhöh, O., Handorf, T. & Heinrich, R. Structural analysis of expanding metabolic networks. *Genome Inform.* **15**, 35–45 (2004).
33. Handorf, T., Ebenhöh, O. & Heinrich, R. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J. Mol. Evol.* **61**, 498–512 (2005).
34. Raymond, J. & Segrè, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**, 1764–1767 (2006).
35. Petrov, A. S. et al. History of the ribosome and the origin of translation. *Proc. Natl Acad. Sci. USA* **112**, 15396–15401 (2015).
36. 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).
37. Barve, A. & Wagner, A. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500**, 203–206 (2013).
38. Szappanos, B. et al. Adaptive evolution of complex innovations through stepwise metabolic niche expansion. *Nat. Commun.* **7**, 11607 (2016).
39. Pál, C. & Papp, B. Evolution of complex adaptations in molecular systems. *Nat. Ecol. Evol.* **1**, 1084–1092 (2017).
40. Lipmann, F. Attempts to map a process evolution of peptide biosynthesis. *Science* **173**, 875–884 (1971).
41. Muchowska, K. B. et al. Metals promote sequences of the reverse Krebs cycle. *Nat. Ecol. Evol.* **1**, 1716–1721 (2017).
42. Muchowska, K. B., Varma, S. J. & Moran, J. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019).

43. Meringer, M. & Cleaves, H. J. Computational exploration of the chemical structure space of possible reverse tricarboxylic acid cycle constituents. *Sci. Rep.* **7**, 17540 (2017).
44. Zubarev, D. Y., Rappoport, D. & Aspuru-Guzik, A. Uncertainty of prebiotic scenarios: the case of the non-enzymatic reverse tricarboxylic acid cycle. *Sci. Rep.* **5**, 8009 (2015).
45. Vetsigian, K., Woese, C. & Goldenfeld, N. Collective evolution and the genetic code. *Proc. Natl Acad. Sci. USA* **103**, 10696–10701 (2006).
46. David, L. A. & Alm, E. J. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93–96 (2011).
47. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
48. Keller, M. A., Kampjut, D., Harrison, S. A. & Ralser, M. Sulfate radicals enable a non-enzymatic Krebs cycle precursor. *Nat. Ecol. Evol.* **1**, 0083 (2017).
49. Keller, M. A., Turchyn, A. V. & Ralser, M. Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. *Mol. Syst. Biol.* **10**, 725 (2014).
50. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
51. Dellomonaco, C., Clomburg, J. M., Miller, E. N. & Gonzalez, R. Engineered reversal of the β-oxidation cycle for the synthesis of fuels and chemicals. *Nature* **476**, 355–359 (2011).
52. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245 (2010).
53. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Thermodynamics-based metabolic flux analysis. *Biophys. J.* **92**, 1792–1805 (2007).
54. Chandru, K. et al. Simple prebiotic synthesis of high diversity dynamic combinatorial polyester libraries. *Commun. Chem.* **1**, 30 (2018).
55. Forsythe, J. G. et al. Ester-mediated amide bond formation driven by wet-dry cycles: a possible path to polypeptides on the prebiotic earth. *Angew. Chem. Int. Ed.* **54**, 9871–9875 (2015).
56. Wächtershäuser, G. Groundworks for an evolutionary biochemistry: the iron–sulphur world. *Prog. Biophys. Mol. Biol.* **58**, 85–201 (1992).
57. Bar-Even, A. Does acetogenesis really require especially low reduction potential? *Biochim. Biophys. Acta* **1827**, 395–400 (2013).
58. Poudel, S. et al. Origin and evolution of flavin-based electron bifurcating enzymes. *Front. Microbiol.* **9**, 1–26 (2018).
59. Duval, S. et al. Electron transfer precedes ATP hydrolysis during nitrogenase catalysis. *Proc. Natl Acad. Sci. USA* **110**, 16414–16419 (2013).
60. Gogarten, J. P. & Deamer, D. Is LUCA a thermophilic progenote? *Nat. Microbiol.* **1**, 16229 (2016).
61. Segré, D., Ben-Eli, D., Deamer, D. W. & Lancet, D. The lipid world. *Orig. Life Evol. Biosph.* **31**, 119–145 (2001).
62. Großkopf, T. et al. Metabolic modelling in a dynamic evolutionary framework predicts adaptive diversification of bacteria in a long-term evolution experiment. *BMC Evol. Biol.* **16**, 163 (2016).
63. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**, 186–189 (2002).
64. Andersen, J. L., Flamm, C., Merkle, D. & Stadler, P. F. *A Software Package for Chemically Inspired Graph Transformation* (Springer, 2016).
65. Banzhaf, W. & Yamamoto, L. *Artificial Chemistries* (MIT Press, 2015).
66. Flamholz, A., Noor, E., Bar-Even, A. & Milo, R. eQuilibrator – the biochemical thermodynamics calculator. *Nucleic Acids Res.* **40**, 770–775 (2012).
67. 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).
68. Halevy, I. & Bachan, A. The geologic history of seawater pH. *Science* **355**, 1069–1071 (2017).
69. Bar-Even, A., Flamholz, A., Noor, E. & Milo, R. Thermodynamic constraints shape the structure of carbon fixation pathways. *Biochim. Biophys. Acta* **1817**, 1646–1659 (2012).
70. Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. BioNumbers – the database of key numbers in molecular and cell biology. *Nucleic Acids Res.* **38**, D750–D753 (2010).
71. Schellenberger, J. et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox v2.0. *Nat. Protoc.* **6**, 1290–1307 (2011).
72. Ribeiro, A. J. M. et al. Mechanism and catalytic site atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* **46**, D618–D623 (2018).
73. Mall, A. et al. Reversibility of citrate synthase allows autotrophic growth of a thermophilic bacterium. *Science* **359**, 563–567 (2018).
74. Nunoura, T. et al. A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile. *Science* **359**, 559–563 (2018).

Acknowledgements

We thank all members of the Segrè Laboratory for helpful discussions. We acknowledge support provided by the Directorates for Biological Sciences and Geosciences at the National Science Foundation and NASA (agreement nos. 80NSSC17K0295, 80NSSC17K0296 and 1724150) issued through the Astrobiology Programme of the Science Mission Directorate; the National Science Foundation (grant no. 1457695, NSFOCE-BSF 1635070); the Human Frontiers Science Programme (grant no. RGP0020/2016) and the Boston University Hariri Institute for Computing and Computational Science and Engineering.

Author contributions

J.E.G., H.H. and D.S. designed the research. J.E.G. wrote the code, ran the simulations and performed the analysis. R.M. contributed to the non-equilibrium steady-state modelling. J.E.G. and D.S. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

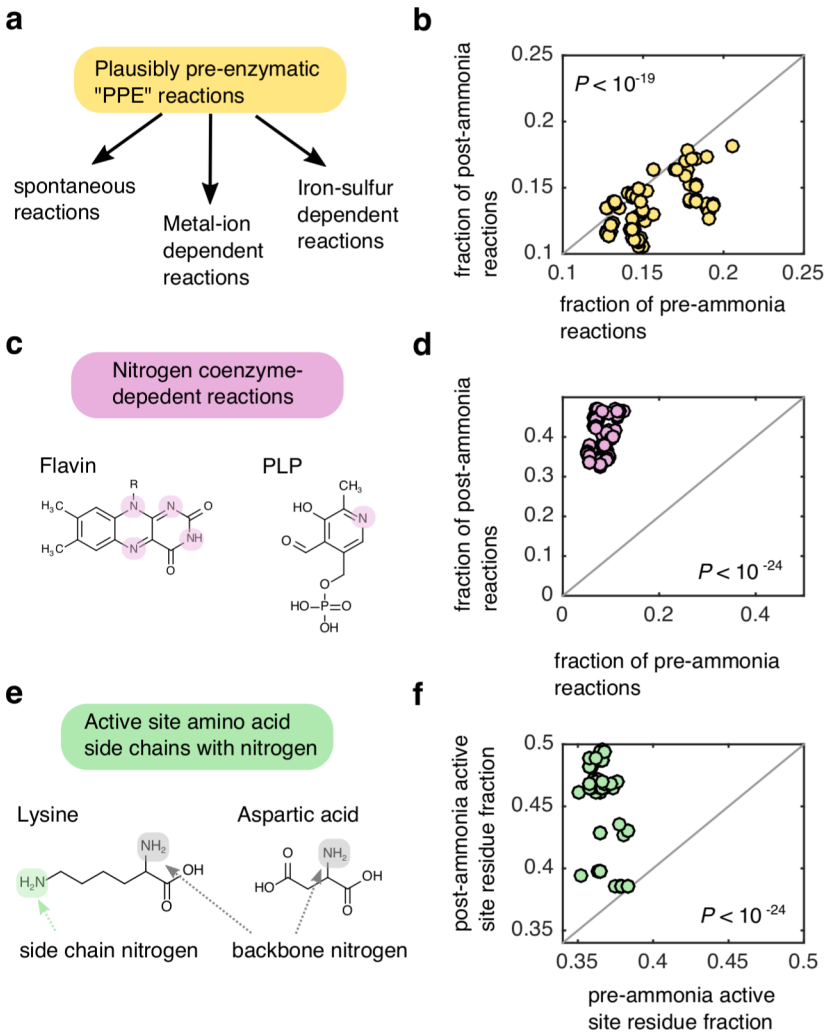
Extended data is available for this paper at https://doi.org/10.1038/s41559-019-1018-8.

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-019-1018-8.

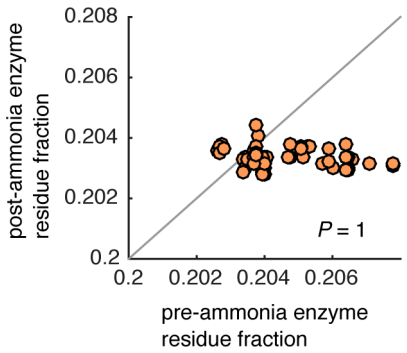
Correspondence and requests for materials should be addressed to J.E.G. or D.S.

Reprints and permissions information is available at www.nature.com/reprints.

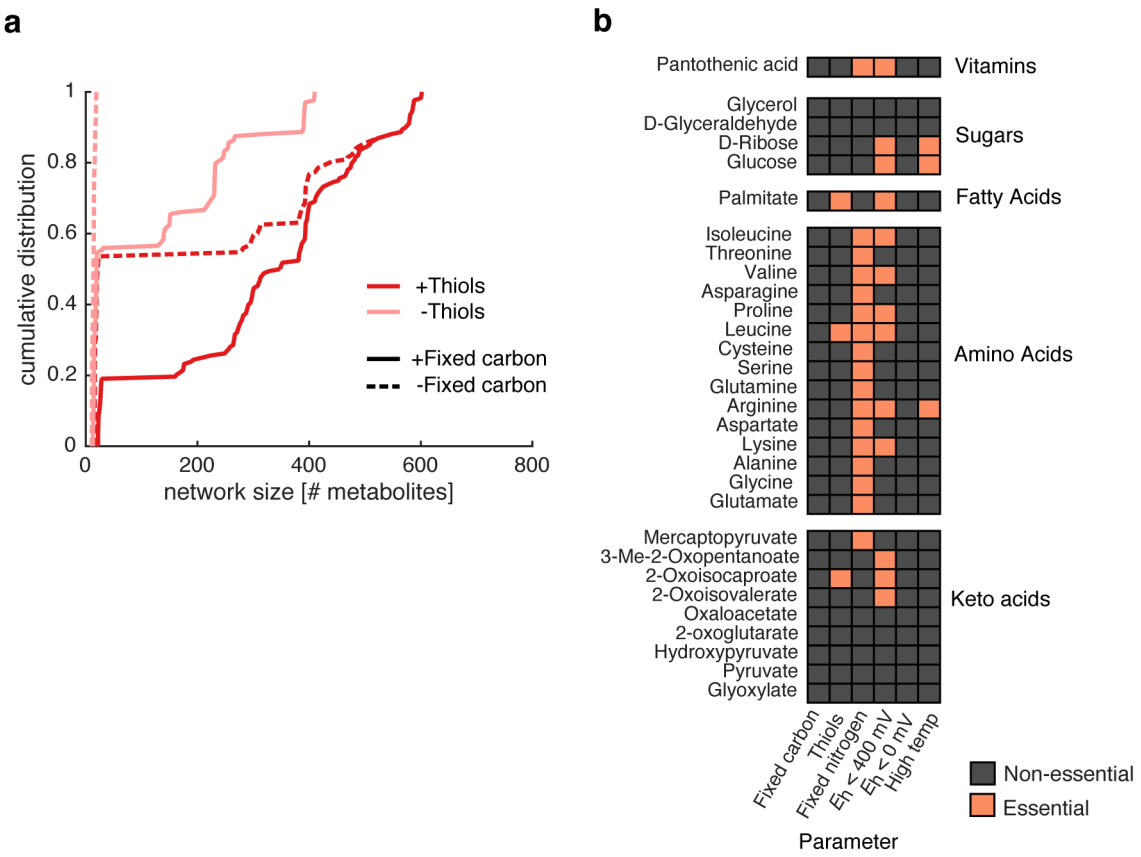
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



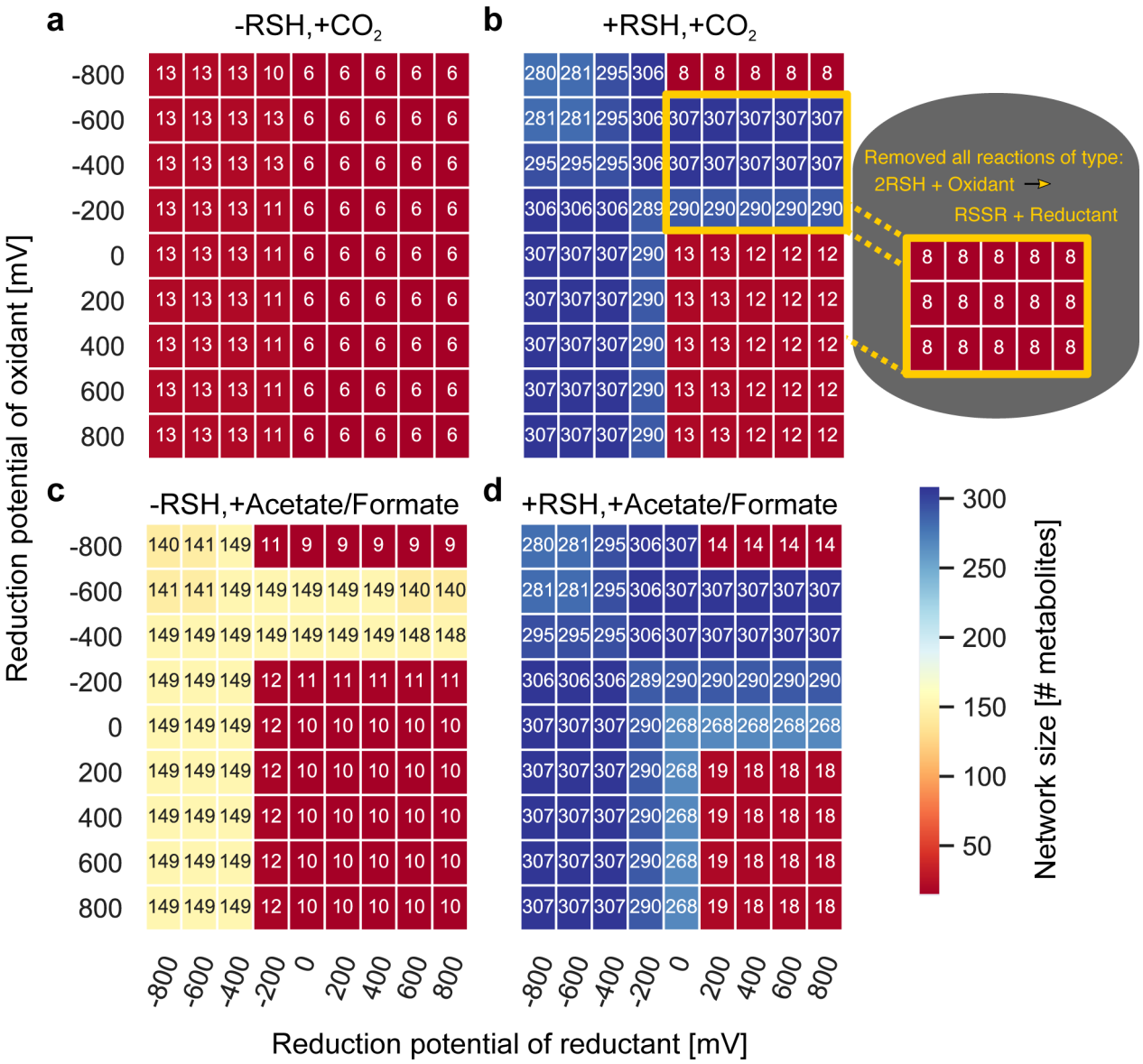
Extended Data Fig. 1 | Enzymes in thioester-driven protometabolism are depleted in nitrogenous compounds. (**a**) We classified reactions in KEGG as being plausibly pre-enzymatic (PPE) reactions if they (i) could proceed spontaneously, (ii) were associated with enzymes that contain at-least one iron-sulfur cluster or (iii) were associated with an enzyme that relied on atleast one metal (Ni, Co, Cu, Mg, Mn, Mo, Zn, Fe, W) cofactor. (**b**) For all scenarios resulting in expansion of more than 100 metabolites (n = 144) we computed the fraction of PPE-reactions amongst the pre-ammonia reactions (*x*-axis) and post-ammonia reactions (*y*-axis). The frequency of PPE-reactions in the pre-ammonia reaction set was on average higher than the frequency of PPE-reactions in the post-ammonia reaction set (one-tailed Wilcoxon signed-rank test: $P < 10^{-19}$). (**c**) We identified KEGG reactions that were dependent on at-least one of the following nitrogen-containing coenzymes: flavin, biotin, thiamine pyrophosphate (TPP) pyridoxal phosphate (PLP), haem, pterin or cobalamin. (d) We compute the fraction of pre- and post- ammonia reactions associated with nitrogen-containing coenzymes in the KEGG database, and found that a much higher proportion of post-ammonia reactions were dependent on these coenzymes relative to pre-ammonia reactions (one-tailed Wilcoxon signed-rank test: $P < 10^{-24}$). (e) We parsed the catalytic active site database[72] to find entries associated with pre and post-ammonia reactions, and compute the fraction of entries associated with amino acids with nitrogen-containing side-chains (Q,N,W,H,K,R). (f) For each scenario, the fraction of active sites with nitrogen-containing amino acids was significantly higher for post-ammonia reactions relative to pre-ammonia reactions one-tailed Wilcoxon signed-rank test: $P < 10^{-24}$).



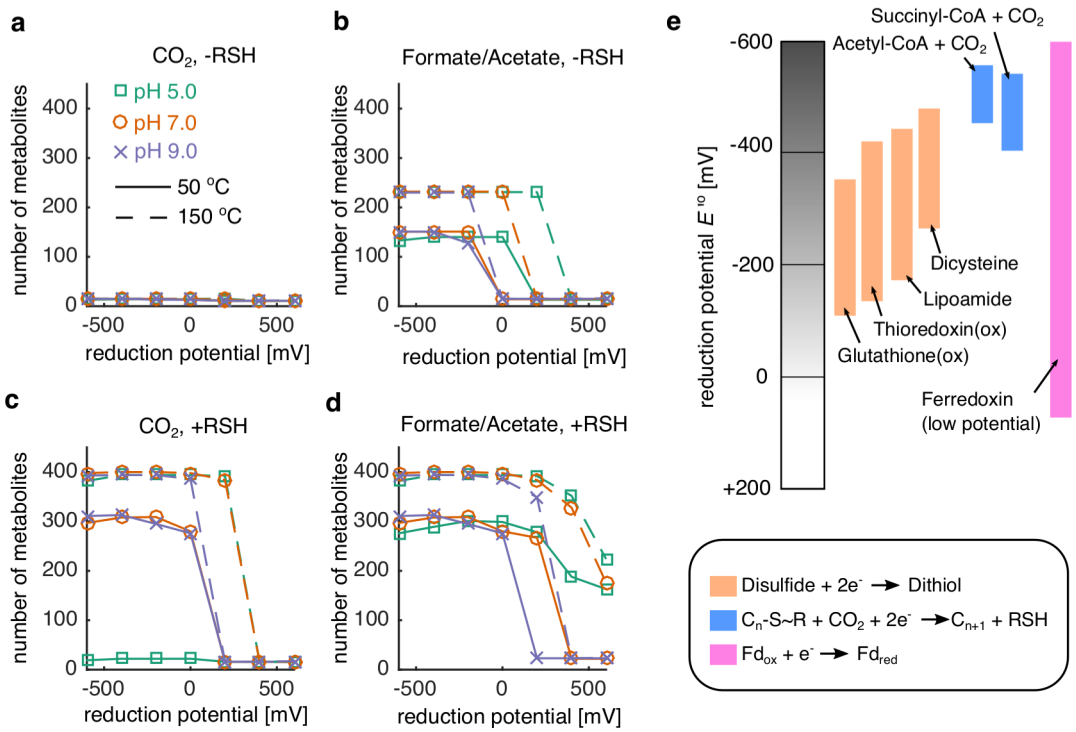
Extended Data Fig. 2 | Enzymes catalyzing reactions before the addition of ammonia are not depleted in nitrogen-containing amino acids relative to enzymes added after ammonia. To see if the amino acid biases in active sites of enzymes catalyzing reactions added to the network without ammonia (see Extended Data Fig. 1e, f) is confounded due to evolutionary selection for reduced nitrogen in these enzymes, we computed the fraction of nitrogen side-chains in enzymes in pre-ammonia reactions (*x*-axis) and in enzymes in post-ammonia reactions (*y*-axis). We found that enzymes in the pre-ammonia networks did not have significantly less nitrogen usage compared to enzymes in post-ammonia reactions (one-tailed Wilcoxon signed-rank test: $P=1$).



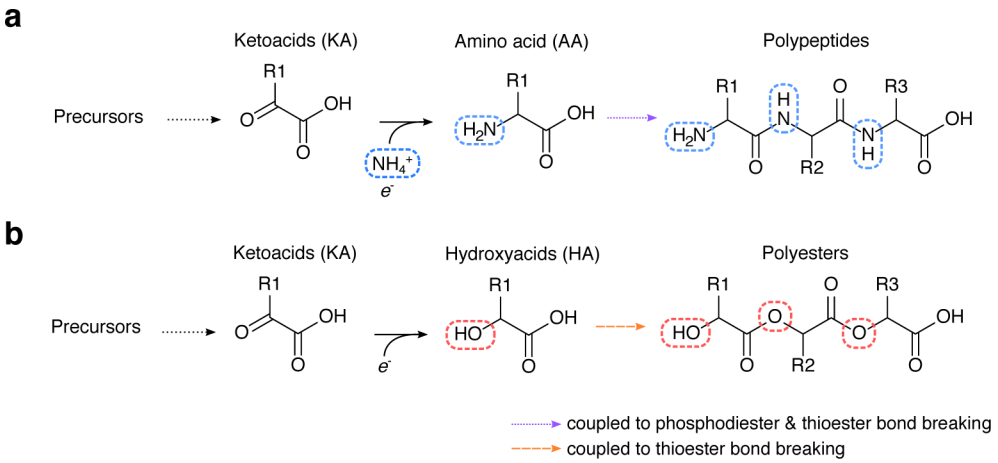
Extended Data Fig. 3 | Thiols are required for autotrophic expansion and fatty acid production. (a) We grouped the $n = 672$ geochemical scenarios into whether a source of fixed carbon and thiols was provided in the seed set. We then plotted the empirical cumulative distributions for each group of scenarios. Notably, when thiols and fixed carbon are not supplied in the seed set, the networks are always below 100 metabolites, indicating that expansion is prohibited without either fixed carbon or thiols in the seed set. (b) We determined what geochemical parameters (*x*-axis) were essential for the production of important biomolecules (*y*-axis). For example, palmitate, a long-chain fatty acid, is producible only if thiols and reductant below 400 mV is provided in the seed set.



Extended Data Fig. 4 | Network expansion with different combinations of carbon sources, thiols, generic reductants and generic oxidants. We performed network expansion using a seed set with both a generic reductant at a fixed potential (*x*-axis) and a generic oxidant at a fixed potential (*y*-axis) with (a) no thiols or fixed carbon, (b) thiols and no fixed carbon, (c) no thiols and fixed carbon, and (d) both thiols and fixed carbon. The colour indicates the size (number of metabolites) in the final expanded network. Interestingly, a strong driving force provided by a strong oxidant (> 0 mV) never sufficiently compensated for the weak driving force provided by a weak reductant (> 0 mV), suggesting that oxidants have little influence on enabling expansion beyond 20 metabolites. The only conditions that led to an expansion that was greater than 20 metabolites with a weak electron donor was when the oxidant was also weak (-200 to -600 mV). We hypothesized that this was due to the ability of thiols or reduced carbon species to reduce the oxidant, enabling the production of a strong reductant. Indeed, when we removed all thiol to disulfide reactions using the generic redox system, expansion was blocked (inset).



Extended Data Fig. 5 | Reduction potential of NAD(P)/FAD substitutes influences the size of expanded networks. We plotted the size (number of metabolites, *y*-axis) of expanded networks as a function of reduction potential of NAD(P)/FAD substitutes (*x*-axis) for different physico-chemical conditions with (a) no fixed carbon or thiols, (b) fixed carbon and no thiols, (c) thiols and no fixed carbon, and (d) both fixed carbon and thiols. (e) We plot the range of physiologically feasible reduction potentials for classes of redox systems potentially relevant for early protometabolic systems, showing that dithiol/disulfide redox systems could potentially have enabled expansion under a variety of conditions.



Extended Data Fig. 6 | Putative ancient catalysts. (a) In extant biochemistry, keto acids are converted to amino acids using transamination or reductive amination reaction mechanisms, which are then polymerized using a phosphate or thioester-coupled mechanism to make polypeptides. (b) If prebiotic environments did not have a source of fixed nitrogen, then keto acids could have been reduced to -hydroxy acids, which could then be polymerized into polyesters either with[4] or without[55] thioester bond breaking.

nature research

Corresponding author(s): Daniel Segre

Last updated by author(s): Sep 2, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	MATLAB2015a, COBRA toolbox V 2.0, Gurobi optimizer (Version 7.0.1), KEGG REST API
Data analysis	MATLAB2015a

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We provide the data as supplemental tables. Figs 1-4 can be re-made using data available in supplemental table Supplementary Dataset 1.xlsx

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our work did not require determination of experimental sample size. For our analysis of biochemical network information from public databases, sample sizes were determined by computing the all combinations of parameters explored in the study.
Data exclusions	No data were excluded from the analysis
Replication	All results were obtained through the computational analysis of biochemical network data from public databases.
Randomization	Our work did not involve allocation of experimental groups; therefore randomization is not relevant to this study.
Blinding	Our work did not involve allocation of experimental groups; therefore blinding is not relevant to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging