The history of enzyme evolution embedded in metabolism

Tatsuya Corlett[1], Harrison B. Smith[1,2,*], Eric Smith[1,3], Joshua Goldford[2,4,5], Liam M. Longo[1,2,*]

[1]Earth-Life Science Institute, Institute of Science Tokyo, Tokyo, Japan

[2]Blue Marble Space Institute of Science, Seattle, USA

[3]Georgia Institute of Technology, School of Chemistry and Biochemistry, Atlanta, USA

[4]Dayhoff Labs, Inc., Cambridge, USA

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

*To whom correspondence should be addressed: hbs@elsi.jp, llongo@elsi.jp

Keywords

metabolism, oxygen, network expansion, KEGG, α/β class, TIM barrel, Rossmann, ECOD

Abstract (200 words)

Phylogenetic reconstructions are a primary record of protein evolution. But what other records can attest to the deep history of enzymes, and what tools are needed to decode their meaning? Here, we demonstrate that the history of enzyme discovery and reuse is embedded within the web of interdependencies that constitute contemporary, biosphere-scale metabolism. Using a simple network analysis approach, we reconstruct both the relative temporal ordering of enzyme domain emergence and, where possible, the first reactions that they catalyzed. These network-based histories were found to be broadly concordant with phyletic information, suggesting that the two approaches reflect related generative processes. When enzyme emergence is initiated after the discovery of nucleotide cofactors, a predominantly stepwise trajectory of domain discovery is recovered. We find that the earliest enzyme-mediated metabolisms were dominated by α/β domains, likely due to their high discoverability and functional potential under constraint. Finally, we quantify how the protein universe responded to a major transition, the biological production of molecular oxygen, by preferentially reusing pre-existing enzyme domains. This work presents a

self-consistent model of metabolic and enzyme evolution, essential progress towards integrating multiple, independent records into a unified history of protein evolution.

Introduction

Biological discovery is strongly constrained by limits on chemical synthetic mechanism and biomolecular folding, reflected in the striking parsimony of core metabolism (Srinivasan and Morowitz 2009; Braakman and Smith 2014) and protein structure convergence, both globally (Medvedev et al. 2021) and of active sites (Hol et al. 1978; Davidi et al. 2018; Longo et al. 2020). Broad trends in enzyme-reaction associations, however, are convolved with a rich history of serendipitous recruitment (Kim et al. 2010) driven by chance functional exploration (Khersonsky and Tawfik 2010; Copley 2020). Consequently, after billions of years of evolution, these associations, which form the basis of contemporary metabolism, cannot be predicted from first principles. Yet, written into the web of known metabolic interdependencies are signatures of the processes through which these associations were formed – and a potential opportunity to characterize biological discovery in earlier eras.

A primary readout when interpreting large networks of metabolic interdependencies is *reachability* (Handorf et al. 2005): Is it possible to achieve a metabolite or set of metabolites by known biochemical reactions under permissive assumptions of reaction thermodynamics and a presumed set of seeding compounds? Analyses of reachability have provided independent support (Goldford et al. 2017; Goldford et al. 2019) for a proposed bioenergetic progression from thioesters to phosphates (de Duve 1995), and demonstrated an unexpectedly direct connection between a small collection of geologically plausible compounds and biosphere-scale metabolism (Goldford et al. 2024). These successes support the view that organosynthetic dependency can report on the temporal organization of metabolic elaboration. However, to date, one of the richest records of metabolic interdependence – enzyme-reaction associations – has not been included. If properly decoded, a history of enzyme evolution derived independently of phylogenetic reconstruction could be revealed.

Here, we present the first explicit treatment of enzymes within a reachability framework. Using a simple scoring function, the *enzyme-gated network expansion* algorithm models the first appearance and subsequent functional diversification of enzyme domains. We find broad agreement with patterns of phyletic distribution and proteome composition, indicating that these two perspectives reflect related generative processes. While it is not possible to order the emergence of the most ancient enzyme lineages by phyletic considerations alone (Winstanley et al. 2005; Edwards et al. 2013), our approach predicts the relative temporal ordering of universally conserved enzyme domains, as well as information about the surrounding metabolic context

within which they emerged. A handoff from geochemical control to enzyme control soon after the availability of nucleotides yields a robust refinement of early expansion sequences (Goldford et al. 2024) and a consistent, realistic progression in enzyme complexity. The results suggest an intriguing transition from easily discovered α/β domains to those dependent on more refined evolutionary information and control systems. Finally, using the biological production of molecular oxygen as an example of metabolic upheaval, domain discovery versus reuse during periods of adaptation could be quantified. We conclude that metabolic network models capture the history of enzyme emergence, providing an orthogonal approach to study protein evolution. This work represents a key advancement toward developing a fully integrated history of protein evolution within the framework of a self-consistent model of metabolic evolution.

Results and Discussion

The protein domains that mediate metabolism suggest a primal role of the α/β class

Protein domains are a fundamental unit of protein structure evolution: Enzymes are composed of domains (often several) that can be shared across multiple enzyme families, resulting in complex patterns of partial homology. The definition of a domain is not fully agreed upon, but is generally meant to entail structural separation, reuse, and independent foldability (Richardson 1981). There are 2,458 protein-domain groupings codified as X-groups in the Evolutionary Classification of Domains (ECOD) database (Cheng et al. 2014; Cheng et al. 2015). We will refer to X-groups as *lineages* because the members within each grouping are hypothesized to have evolved from a single founder domain. The largest connected component of the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto 2000), which represents the extended core of biosphere-scale metabolism, comprises 7,678 reactions, of which 5,730 are associated with one or more lineages. In total, just 396 lineages, the *metabolic lineages*, are associated with the enzymes catalyzing reactions in the largest connected component of KEGG.

Metabolic lineages are more broadly phyletically distributed than non-metabolic lineages (**Figure 1A upper, Figure S1, Table S1**), and 87% of the metabolic set were detected in all three domains of life (archaea, bacteria, or eukarya). All protein structure classes (α, β, α+β, α/β, mixed, other) are present among the metabolic lineages (**Figure 1A lower, Figure S1, Tables S1-2**); however, lineages adopting an α/β structure (in which α and β elements alternate) are enriched 3.2-fold in the metabolic set relative to all lineages. By definition, each metabolic lineage is associated with one or more reactions, though these relationships are not necessarily unique because (a) multiple unrelated enzymes can catalyze the same reaction, (b) a reaction can require more than one enzyme, and (c) an enzyme can be composed of multiple domains from different lineages (**Figure S2, Table S3**). We define a *rule* as a set of lineages that act together to

catalyze a given KEGG reaction (**Figure 1B inset**). The association between a reaction and a set of lineages specifies a unique rule; thus, any given reaction or lineage set on its own can be associated with multiple rules. The majority of reactions have rules with a set of one or two lineages (**Figure 1B**). Among reactions that can be catalyzed by a single lineage, the prevalence of α/β domains is 2.7-fold greater than the fraction of metabolic lineages that adopt an α/β structure.

The metabolic importance of a lineage may be assessed by the number of reactions it is associated with, or by its unique ability to enable reactions that produce key metabolites (**Figure 1C**). We obtain a measure of the second sense of importance within the framework of network expansion (Handorf et al. 2005). Network expansion assigns an ordering to reactions from a given set based on synthetic dependencies among their substrates from an assumed subset of metabolic precursors. Compound reachability and ordering by network expansion have been used as proxies for historical emergence order (Raymond and Segrè 2006; Goldford et al. 2017; Goldford et al. 2024). Expansion proceeds recursively: First the assumed precursors, called *seed compounds*, undergo all possible reactions, and then the products are added to the seeds to give the pool of possible inputs to the next iteration of reactions. The process is repeated until no new compounds are discovered. Once a compound is produced within the network, it is never removed, as network expansion concerns reachability, it is not a kinetic model. The final set of compounds reached is referred to as the scope. By *ablating* the subset of reactions that depend on a given lineage or compound, the role of that lineage or compound is reflected in the resulting contraction in scope size.

Using a previously described model of biosphere-level metabolism with a non-autocatalytic pathway for purine biosynthesis (Goldford et al. 2024) and a set of 80 seed compounds – including 20 carboxylic acids, 10 amino acids, various metal ions and inorganic compounds (*base seed set*) (**Tables S4-7**) – the network scope upon *ablation of a single lineage* was calculated for each of the 396 metabolic lineages (**Table S8**). Both generalist lineages with many associated reactions (*e.g.*, Rossmann, TIM barrel, and flavodoxin) and specialist lineages with a small number of associated reactions (*e.g.*, the three L-aspartase lineages that mediate purine biosynthesis and X-group 3997, which mediates chlorophyll biosynthesis) were found to be essential for complete metabolic expansion (**Figure 1C**). Many specialists had little to no impact on network size upon ablation, but several generalists (*e.g.*, the haloacid dehydrogenase lineage, abbreviated HAD) also had a minor impact on the scope. In total, only 37 lineage ablations resulted in a contraction of the scope by more than 20%, in each case to a sub-network that does not include molecular oxygen. Three factors affecting lineage essentiality are: (a) the existence of alternative biosynthetic routes, (b) alternative dependence of the same reaction on different lineages (convergent evolution), and (c) the combining of multiple lineages in an enzyme (joint dependence). The low incidence of lineage essentiality emphasizes the greater role of both forms of alternation over joint dependency, which would amplify essentiality.

Lineages that cause a large contraction of the scope when ablated are more widely distributed across prokaryotes than those that do not ($p = 7.9\times10^{-14}$, Mann-Whitney U test, **Figure 1D, Figure S1**). Only one lineage, which participates in chlorophyll biosynthesis (X-group 3997), is an outlier. The six lineages causing the largest contraction upon ablation – all generalists that are individually associated with more than 100 reactions – each result in a scope contraction to a subnetwork that lacks CoA (**Figure 1C**, in bold). These lineages either adopt α/β structures (TIM barrel, Rossmann, flavodoxin, HUP) or contain α/β character (ATP grasp and α/β hammerhead). The observation that the α/β class is enriched among metabolic lineages (**Figure 1A**), single-lineage reactions (**Figure 1B**), and lineages that cause the greatest scope contraction upon ablation (**Figure 1C**) is consistent with the hypothesis that α/β lineages were key mediators of early metabolic processes (Caetano-Anollés et al. 2007; Ma et al. 2008; Longo et al. 2020).

Disentangling metabolic dependencies to uncover protein histories

Single-lineage ablations (**Figure 1C**) reveal the first point along a network expansion trajectory where a lineage becomes essential for further network growth, setting an *upper limit* on when that lineage may have emerged. Convergent evolution, however, can render emergence order ambiguous. Multi-factor analyses can sometimes resolve ambiguities by performing joint reconstruction of compounds, reactions, and lineages (Braakman and Smith 2013). Here we propose the *enzyme-gated network expansion* algorithm (**Figure 2A**), which integrates the graph structure of metabolism (Kanehisa and Goto 2000), lineage-enzyme associations (Kanehisa and Goto 2000; Cheng et al. 2014), enzyme-reaction associations (Kanehisa and Goto 2000), cofactor-reaction associations (Goldford et al. 2024), and estimates of reaction thermodynamics (Noor et al. 2013; Beber et al. 2022) to reconstruct the shared history of metabolic reactions and protein lineages (see **Methods** for additional details).

Our aim is to understand how dependency in both chemical substrates, and catalyst structures and functions, constrained discovery and adaptation as biochemistry incrementally became a self-referential system autonomous from geochemistry. We frame the evolution of reactions and lineages as a stepwise process with discrete emergence events. Standard network expansion considers only reactions and compounds, and its enforcement of dependency is strict: seed compounds are given geochemically, and all others must be generated within the growing network. Here we introduce lineage availability as a second constraint: new reactions can only be discovered if the requisite set of lineages, reactant compounds, and cofactors are all present. Unlike metabolites and cofactors, lineages are not endogenously generated within the model, but are introduced individually or in the smallest possible groups, as doing so enables an expansion of the reaction set. Each simulation begins with a set of seed compounds, enzyme-independent reactions, and no lineages. At each lineage addition step, the lineage that will enable the most

174 new rules ("top-ranked lineage") is introduced. Both rules that result in the discovery of new
175 compounds and those connecting previously discovered compounds are counted. Lineages are
176 thus ordered so as to maximize their contribution to expansion within the context of the growing
177 network. After the new rules are added, standard metabolic network expansion is performed until
178 no new compounds can be produced. Once a lineage is added, it remains available for all
179 subsequent metabolic expansion steps. This process is repeated until all lineages and reactions
180 are discovered. We define a *metabolic stage* as a set of metabolites, lineages, and reactions that
181 cannot undergo further standard network expansion without addition of another lineage.

182 By partitioning lineage introduction into separate steps and driving standard network
183 expansion to completion between each lineage introduction, we assume that the discovery of
184 new lineages is infrequent compared to their functional diversification and their effect on
185 metabolic production. Our criterion for selecting lineages regards only the effect of new rules on
186 the immediate expansion of metabolites; consequences for the later trajectory of network
187 expansion are not considered. This selection criterion stems from the intuition that domain
188 emergence can be supported through interactions with a cognate ligand (Alva et al. 2015; Yagi and
189 Tagami 2024). In the case of a tie, one among the lineages enabling the most new rules is selected
190 at random, so the expansion algorithm is not deterministic. In rare cases, no single lineage
191 addition enables the discovery of a new rule. At this point, we consider pairs of new lineages, then
192 sets of three lineages, and so on. We add the smallest number of new lineages at each lineage
193 addition step because we deem the near-simultaneous emergence of multiple lineages rare.
194 Finally, we note that emergence here refers to the emergence of *enzyme activity*, as some
195 lineages may have first played non-metabolic roles within the cell prior.

196

197 *Lineage emergence order is well constrained and comports with phyletic distribution*

198 The majority of lineages (79%, averaged over 1,000 simulation runs) are required for expansion to
199 post-O_2 metabolic stages with enzyme gating (**Figure 2B**). In contrast, under single-lineage
200 ablation, less than 10% of lineages resulted in significant contraction (**Figure 1C**). Plotting the
201 results of 1,000 independent runs reveals a narrow distribution of metabolic trajectories (**Figure**
202 **2B**), in which the standard deviation of the addition step for 96% of lineages is less than 20 (**Figure**
203 **S3**). Across these runs, lineage addition order correlates strongly with prokaryote phyletic
204 distribution score (**Figure 2C**), supporting the interpretation of dependency ordering in network
205 expansion as a proxy for evolutionary age.

206 For each run, after the addition of approximately 107 lineages, nucleotide cofactors (*e.g.*,
207 ATP, NAD, or CoA) are produced and rapid metabolic growth follows. If instead lineages are added
208 at random, about 370 additions are required on average before a burst in compound discovery

occurs. Although changing the lineage order can have a large impact on the growth trajectory, compound discovery order remains fairly consistent across all runs, and similar to that of standard network expansion (**Figures S4** and **S5**) with an average Spearman's rank correlation coefficient of 0.85. We conclude that lineage usability is tightly constrained by the strict metabolic dependency enforced in network expansion, so that enzyme-gating fills in interpolating detail left unresolved in substrate-only expansion. As a result, enzyme gating does not significantly alter the modes of pathway emergence (**Figure S6A**, **Table S9**) or the emergence order of carbon fixation pathways (**Figure S6B, Table S10**).

Nucleotide cofactors predate lineage emergence

It is unclear when enzymes emerged as the dominant catalysts in the evolution of metabolism. To identify signatures sensitive to the emergence of enzyme-mediated catalysis at varying degrees of metabolic complexity, we *pre-expanded* the seed set to successively larger fractions of the metabolic network before enzyme gating was initiated (**Figures 2D** and **S7**). We considered 10 points along the standard network expansion trajectory: the base seed set, the inventory of metabolites available upon the discovery of 8 primary cofactors, and all metabolites. Note that pre-expansion regards the inclusion of additional compounds, not reactions or rules. A principal component analysis of the resulting lineage orders (1,000 runs per pre-expansion) revealed just three major clusters: little or no pre-expansion (pre-nucleotide cofactor seed set), intermediate pre-expansion (post-nucleotide seed set, which breaks into two sub-clusters), and complete or near-complete expansion (post-O_2 seed set). All levels of pre-expansion result in a low standard deviation of the lineage addition order across runs (**Figures S3**) and retain strong rank correlation between phyletic distribution and addition order (Spearman's rank correlation, $p < 1.3 \times 10^{-76}$; **Figures 2C** and **S8**). Lineage addition order when starting with all metabolites most strongly reflects the number of rules associated with each lineage with a preference for lineages that can catalyze reactions without co-domains, while successively less pre-expansion emphasizes the metabolic importance of specialist lineages in catalyzing bottleneck reactions (**Figure S9**).

The most striking indication that there is a "natural" position for enzymatic takeover of metabolism is that pre-expansion to a nucleotide-containing seed set allowed a long interval of single lineage addition – more than 250 addition steps – before the first requirement for concurrent multi-lineage additions (**Figures 3A** and **S10**). Thus, despite the substantial fraction of rules that require multiple domains (**Figure 1B**), relations among their compositions together with the majority subset of single-lineage rules enables expansion almost entirely by single-lineage innovations on seed sets that contain nucleotides. In contrast, the base seed set and the PLP seed set required multiple concurrent multi-lineage additions before 100 steps. The advent of nucleotides almost certainly predates protein coding, and the fact that signatures of this

historical ordering are apparent in our dependency analysis further suggests that we are capturing historical information. We conclude that the intermediate expansion cluster best reflects the earliest enzyme histories.

The first metabolic lineages: TIM Barrel and the Rossmannoids

A joint history of metabolite and lineage discovery based on the ATP pre-expansion seed set is presented in **Figure 3A** (also see **Tables S11-12**). Previously, metabolic evolution was observed to be bursty or punctuated, with periods of innovation interspersed with periods of quiescence (Goldford et al. 2024). Enzyme gating did not fundamentally change this property of the substrate trajectories (see also **Figures 2B** and **S11**). It does, however, assign an emergence ordering to the earliest protein lineages, which have proven recalcitrant to relative dating by other methods due to their near-absolute conservation across the tree of life (Winstanley et al. 2005; Edwards et al. 2013).

The first six lineages (**Figure 3A**) – which appear in the same order across all 1,000 independent ATP pre-expansion runs, and are similarly positioned in other pre-expansions (**Figure S12, Table S13**) – are the TIM barrel, HAD, ribonuclease H-like, Rossmann, flavodoxin, and cradle loop barrel. The profound metabolic importance of the TIM barrel is well established (Goldman et al. 2016), as are its evolutionary relationships to flavodoxin (Farías-Rico et al. 2014) and other Rossmannoids (in this case, HAD and Rossmann) (Kolodny et al. 2021; Qiu et al. 2022). The early trajectory positions of HAD and ribonuclease H-like, which ablation studies classified as non-essential generalists, likely relates to the large repertoire of simple reactions that these lineages catalyze. That 4/6 of the earliest lineages belong to the α/β structure class and one adopts a 'mixed' structure (Ribonuclease H-like), having significant α/β and α+β character, is consistent with hypotheses that α/β structures were the earliest entry point into folded domains based on phyletic distribution (Caetano-Anollés et al. 2007), sequence-structure fragment sharing (Nepomnyachiy et al. 2014), and structure similarity (Choi and Kim 2006). Moreover, TIM barrel, Rossmannoids, and some β-barrel folds adopt repetitive structures (*e.g.*, TIM barrel has a pseudo axis of C_4 symmetry and the Rossmann has a β-strand order of 3-2-1-4-5-6, suggestive of an early duplication event) underscoring symmetry and repetition as organizing principles in early protein evolution (Eck and Dayhoff 1966). Structure class composition as the trajectory proceeds (**Figure 3B**), however, shows that while α/β lineages are preferred early, they are eventually overtaken by all α and α+β lineages. These results are broadly consistent with the history of fold emergence inferred from patterns of protein structure similarity (Choi and Kim 2006). We have previously hypothesized that the early dominance of the α/β structure class likely relates to facile emergence and functional potential under constraint (Romero Romero et al. 2018; Longo et al. 2020; Longo et al. 2020; Vyas et al. 2021), but their eventual dilution challenges the notion that

281 they are fundamentally more evolvable or functionally capable. β-proteins play a more limited
282 role in metabolism (**Figures 1A**, **1B,** and **3B**) but the cradle loop barrel lineage is found to emerge
283 early. This lineage plays essential roles in information processing and encompasses several
284 small, traversable β-barrel folds (Yagi and Tagami 2024). Small β-folds have been suggested to
285 play a founder role, particularly among ribosomal proteins (Petrov et al. 2015; Kovacs et al. 2017).

286

287 *Bursts of metabolic innovation recruit lineages young and old*

288 When new metabolites arise in the biosphere, they can drive the emergence of new lineages or be
289 acted upon by established lineages (*reuse*). The history of lineage incorporation and subsequent
290 reuse upon the emergence of new metabolites, for a representative ATP pre-expansion run, is
291 presented in **Figure 4A**. While approximately half of lineages find use in only a single metabolic
292 stage, 54% (216/396) of lineages mediate rules at multiple points along the trajectory (**Figure**
293 **S14**). Some lineages, particularly the early-emerging generalists, enjoy near constant reuse along
294 the trajectory. The Rossmann lineage, for example, is reused in an average of 30% of metabolic
295 stages, likely due to its close association with various cofactors. Other lineages lie dormant for
296 long periods before appearing in new rules; for example, X-group 211, which was added in lineage
297 addition step 99 to perform carbon-sulfur lyase chemistry, is later reused to catalyze a reaction
298 involving O_2 at step 296. About 70% (152/216) of the lineages with reuse catalyze a reaction with a
299 new 2-digit EC (Enzyme Commission) number sometime later, usually within 30 lineage addition
300 steps of their emergence (**Figure S15A**). Peaks in lineage reuse (**Figure 4A**, upper) correspond to
301 significant events of metabolic growth, and late peaks increase in size as more established
302 lineages are available to draw from. The number of lineages per rule modestly increases over the
303 expansion (**Figure S16**) as key lineages accrete cooperating domains (see **Figure S17** for an
304 analysis of the TIM barrel).

305

306 *Reuse outpaces emergence in the Proterozoic: The case of molecular oxygen*

307 The emergence of molecular oxygen approximately 3 billion years ago (Crowe et al. 2013) added a
308 layer of O_2-associated reactions around a largely anaerobic core, a metabolic organization
309 referred to as hierarchical modularity (Raymond and Segrè 2006). The modest rewiring of interior
310 metabolic layers that resulted from the advent of O_2 stands in stark contrast to its consequence
311 within the protein universe: In absolute terms, molecular oxygen drove one of the largest periods
312 of lineage emergence and reuse in our expansions (**Figures 4A** and **S15B**), affecting lineages along
313 the entire metabolic trajectory. To quantify the impact of metabolite discovery on the protein
314 universe, each lineage was grouped into one of five classifications with respect to a given
315 metabolite (see **Methods** and **Figure S18**): Lineages whose first reactions all require the

316 metabolite (*emerge to use*, ETU); lineages for which some of their first reactions require the
317 metabolite (*first use*, FU); lineages that catalyze reactions involving the metabolite sometime after
318 their emergence (*learn to use*, LTU); lineages that catalyze reactions that are made possible by
319 the metabolite, but do not catalyze reactions with the metabolite directly (*indirect use*, IU); and
320 lineages that are not associated with the metabolite in any way (*never use*, NU).

321 We find that molecular oxygen is one of the strongest drivers of both lineage emergence
322 and lineage adaptation (**Figure 4B**), as evidenced by the number of lineages whose first reactions
323 all require oxygen (ETU) and those that adopt oxygen use later in their evolutionary trajectory
324 (LTU). For 8% of metabolic lineages, oxygen usage is hypothesized to be their primal function, as
325 either all (ETU) or some (FU) of their first reactions require oxygen (**Figure 4C**, **Table S14**). We
326 estimate that 14% of metabolic lineages acquired the ability to use oxygen later in the trajectory,
327 sometime after their emergence as enzymes. About 65% of generalist lineages (those having ≥100
328 associated reactions) fall into this category (**Figure S19A**). About 17% of lineages catalyze
329 reactions that indirectly use (IU) molecular oxygen as some or all of their associated rules
330 become unreachable when oxygen is removed from the network (**Figure S19B**). Lastly, 61% of
331 lineages do not catalyze reactions directly or indirectly associated with molecular oxygen (NU).
332 Taken together, these data indicate that, in the Proterozoic era, the response to metabolic
333 innovation favors reuse of established lineages over *de novo* emergence by a factor of only 2.3.
334 This result implies the existence of a fairly large reservoir of proto-enzymatic forms – including
335 both emergent domains and conserved non-enzymatic domains – and is consistent with the fact
336 that ~2,500 lineages are known. Nevertheless, the fact that metabolic lineages are still
337 preferentially reused suggests that innovation by existing enzymes is likely the dominant mode of
338 metabolic adaptation for advanced cellular life. As the repertoire of metabolic lineages continues
339 to grow, the preference for reuse will likely increase further.

340 Finally, as a test of the categorization of lineage-oxygen emergence relationships, we
341 analyzed the domain composition of aerobic and anaerobic bacteria (Madin et al. 2020; Goldford
342 et al. 2023) (**Table S15**). We find that category-wise lineage enrichments (**Figure 4D**) are fully
343 consistent with expectation: ETU+FU lineages are significantly more enriched in aerobes.
344 Likewise, IU lineages, which participate in post-oxygen metabolism, are also enriched in aerobes.
345 Conversely, NU lineages are more enriched in anaerobes. LTU lineages, which presumably
346 emerged before oxygen and only later catalyzed an oxygen-associated reaction, are equally
347 common in both anaerobes and aerobes. Thus, signatures of lineage adaptation to oxygen
348 predicted by the enzyme-gated network expansion algorithm are manifested as distinct
349 distribution patterns across bacteria with different physiologies, suggesting that our approach
350 can estimate the relative order of enzyme functions.

351

352 *Limitations and outlook for models of metabolic evolution with strict causality*

353 The goal of the present approach is to extract protein historical information from the structure of
354 metabolism and domain-metabolism associations in a way that is not derived from phylogenetic
355 information. To do this, metabolite synthetic dependencies are strictly enforced without
356 considering kinetic constraints that often must be satisfied for completing metabolic pathways,
357 which simplifies implementation and interpretability. Nevertheless, strictly enforcing
358 dependencies in the substrate network can lead to fragility in cases where the essential source
359 data is missing or incorrect, as was the case for non-autocatalytic purine synthesis (Goldford et
360 al. 2024). The strengths and limitations of the network expansion algorithm have been discussed
361 in detail (Goldford and Segrè 2018), and apply to this approach as well. It is also essential to
362 emphasize that this approach reflects the emergence of a domain as a metabolic enzyme, and
363 does not account for other types of functional roles. The broad concordance between the inferred
364 domain history presented here and phyletic distribution (**Figure 2C**), diverse perspectives on the
365 order of protein class emergence (**Figure 3B**), and the relationship between physiology and
366 domain composition (**Figure 4D**), suggest that a *bona fide* record of protein emergence is written
367 into the structure of metabolism and can be extracted with a simple algorithm. Future work can
368 begin to integrate alternative historical records, such as by incorporating homology relationships
369 *between* lineages (Farías-Rico et al. 2014; Longo et al. 2020; Kolodny et al. 2021), or by developing
370 frameworks with relaxed causality through careful coarse graining.

371

372 **Concluding Remarks**

373 We introduce a new approach to interpret the record of enzyme evolution embedded within the
374 graph structure and reaction-lineage associations of biosphere-scale metabolism. Using this
375 approach, a stepwise process of lineage emergence and innovation is reconstructed. This record
376 of protein history is independent from (though found to be correlated with) phyletic distribution,
377 and can provide evidence for ancient lineage emergence order. Our data suggest that
378 contemporary lineage emergence began at a stage of metabolism shortly after the emergence of
379 nucleotide cofactors. The first metabolic lineages were likely a group of related α/β lineages – TIM
380 barrels and the Rossmannoids – but the dominance of the α/β class was diluted over time. Finally,
381 we provide a snapshot of how metabolism responds to upheaval, in this case, the emergence of
382 molecular oxygen. We find that the adaptive reuse of pre-existing lineages has thus far outpaced
383 *de novo* emergence, but that a formidable reservoir of proto-enzymatic forms also contributes to
384 metabolic evolution. This work represents essential progress towards synthesizing diverse
385 historical records to reconstruct integrated histories of protein evolution.

386

Acknowledgements

Shawn E. McGlynn helped manually annotate KEGG orthology group interdependence. This work was supported by a Research Grant from HFSP (Ref.-No: RGEC29/2025) to LML with the award DOI https://doi.org/10.52044/HFSP.RGEC292025.pc.gr.230124. JG acknowleges support from National Aeronautics and Space Administration (NASA) Interdisciplinary Consortia for Astrobiology Research (ICAR) award 80NSSC23K1357.

Author Contributions

Conceptualization: TC, HS, JG, LML
Methodology: TC, HS, JG, LML
Software: TC, HS, JG
Validation: TC, HS
Formal Analysis: TC, HS, LML
Investigation: TC, HS
Resources: HS, JG, LML
Data curation: TC, HS, JG, LML
Writing – Original Draft: TC, ES, LML
Writing – Review & Editing: HS, JG
Visualization: TC, LML
Supervision: ES, HS, JG, LML
Project administration: LML
Funding acquisition: LML

Figures

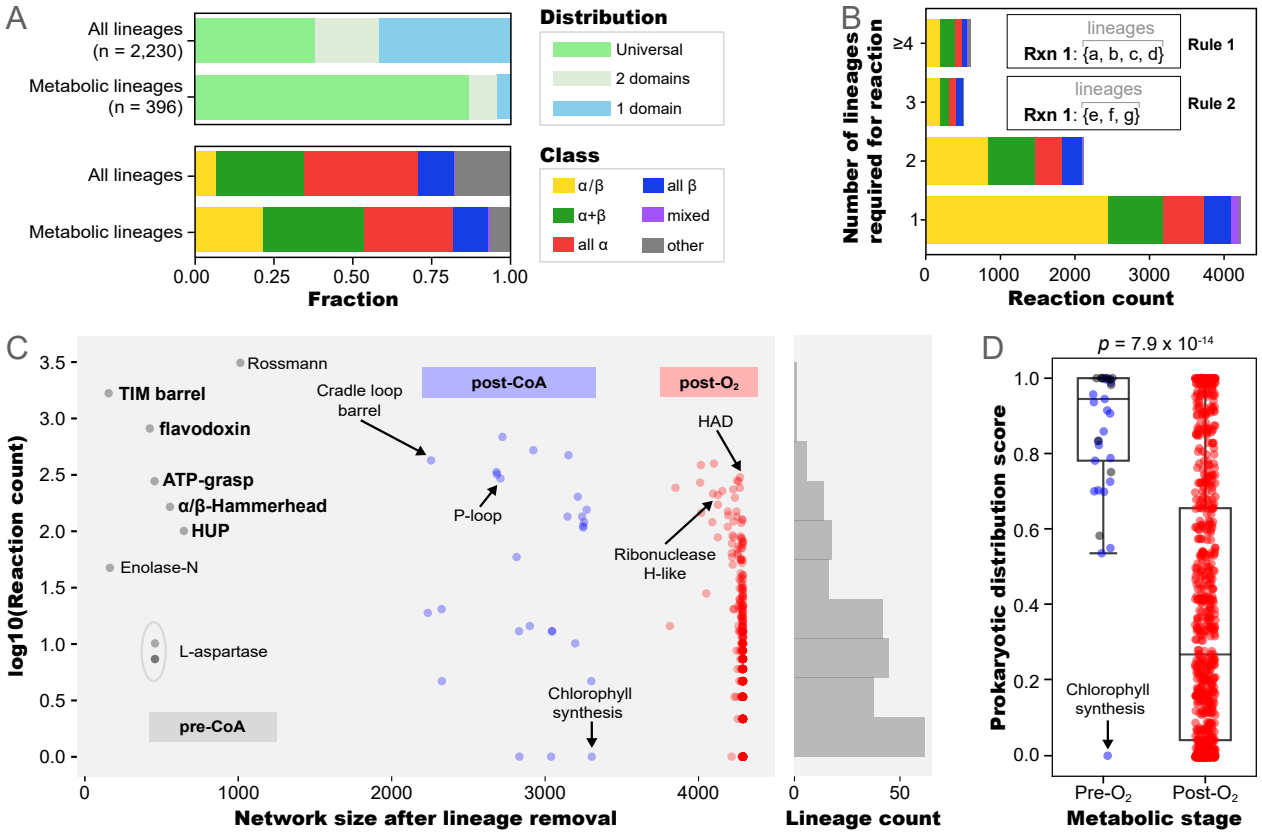

Figure 1. The metabolic lineages. **A**. Distribution and structure class of all lineages in ECOD and the subset of metabolic lineages. Universal refers to lineages present in all three domains of life. **B**. Number of KEGG metabolic reactions in scope (x-axis) that require n (1, 2, 3, or ≥4) distinct lineages to be catalyzed (y-axis). The same reaction can appear more than once if it is associated with rules that have different lineage sets. Lineage structure class for reactions with multiple rules, and rules with two or more lineages, is evenly divided between rules and lineages. Inset: the relationship between reactions, lineage sets, and rules. **C**. The number of associated reactions (y-axis) and the network size from standard network expansion after lineage ablation (x-axis). CoA- (blue) and O_2-containing (red) scopes are indicated. **D**. Prokaryotic distribution scores (see **Methods** for definition) of lineages that are essential for reaching O_2 versus those that are not based on the single lineage ablation analysis in panel C.

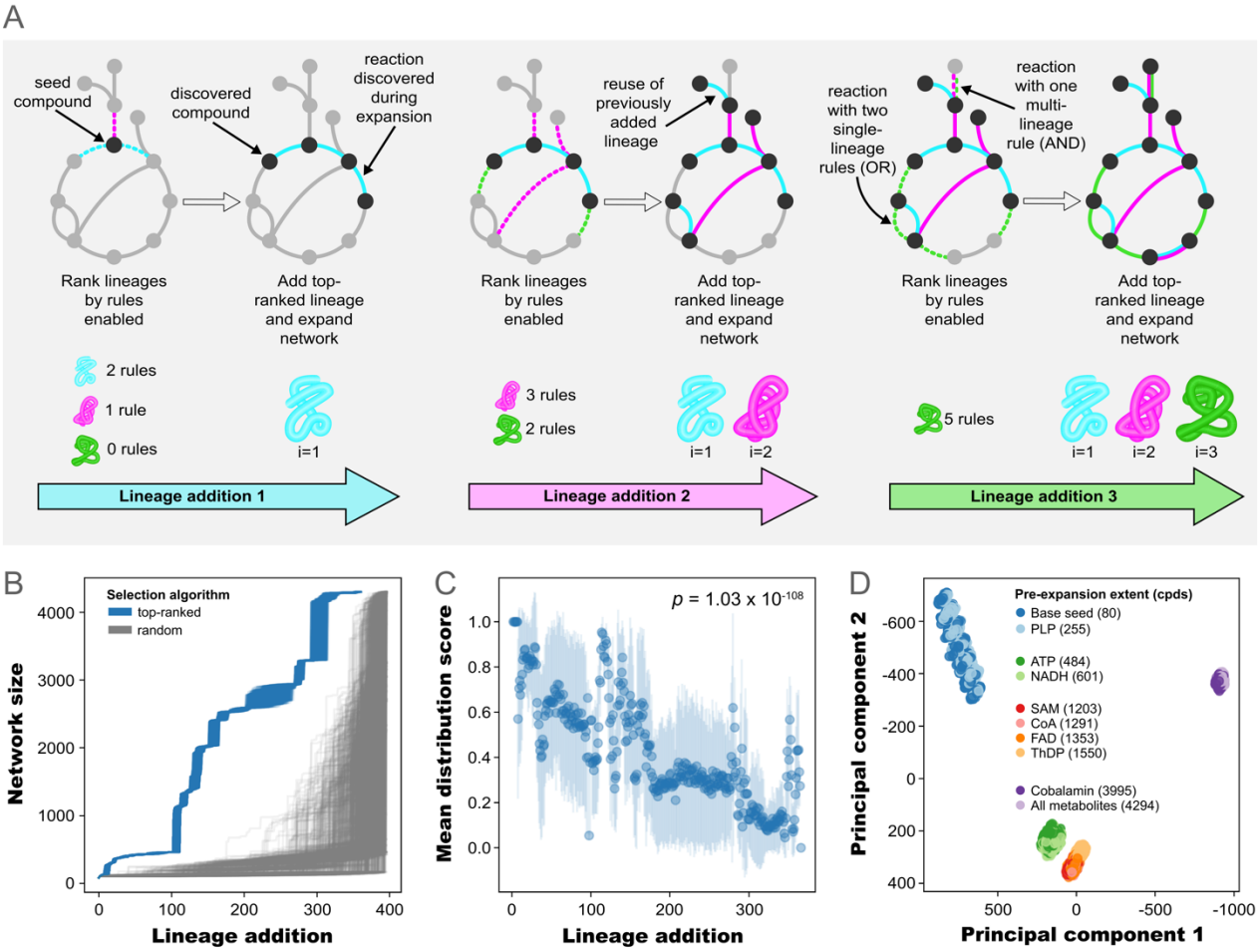

Figure 2. The enzyme-gated network expansion algorithm. **A**. See main text and **Methods** for algorithmic details **B**. Metabolic growth trajectories (network size) when selecting the top-ranked lineage (blue; described above) or a random lineage (grey) for 1,000 runs each. **C**. Average prokaryote distribution score of lineages across 1,000 runs as a function of lineage addition order. Error bars are standard deviations. Indicated p-value calculated from the Spearman's rank correlation. Lineage additions for panels B and C are based on the base seed set of compounds with no pre-expansion. **D**. Principal component analysis of the lineage addition orders from all trajectories across all ten seed sets (1,000 independent runs per seed set, 10,000 runs in total). The seed compounds for each pre-expansion are listed in **Table S4**. Panels B and C for each pre-expansion seed set are provided in **Figures S7** and **S8**.

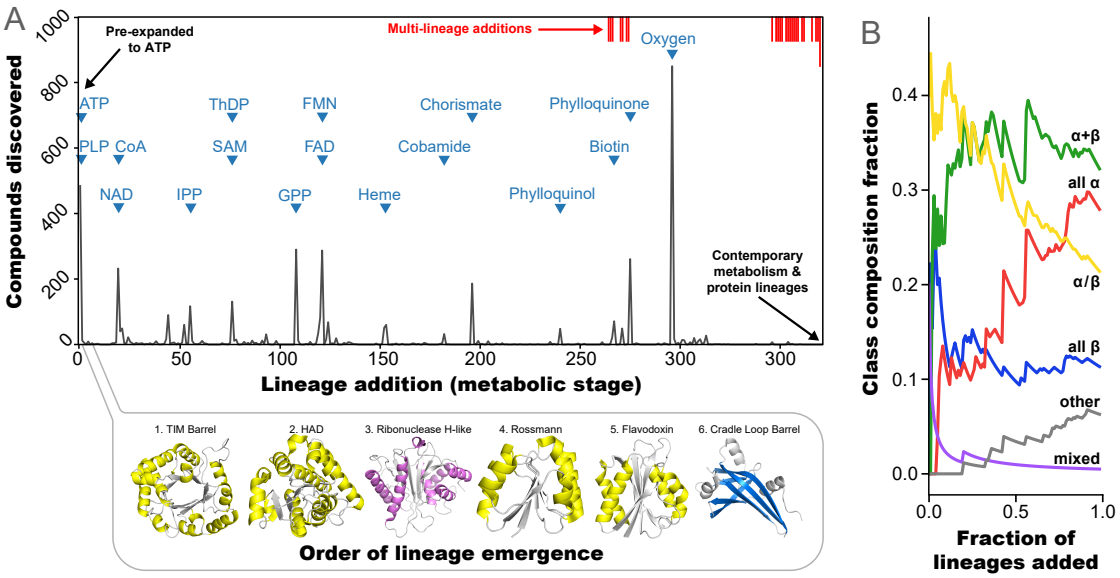

Figure 3. A coevolutionary history of lineages, metabolic reactions, and metabolites. **A**. Fastest growing trajectory (cumulative plot with the greatest integrated area) out of 1,000 runs for the ATP seed set. Select metabolites and the first 6 lineages, colored according to structure class, are indicated. The position of metabolite labels on the y-axis is arbitrary. The red rug plot (along the top of the figure) denotes lineage addition steps where two (short line) or three (long line) lineages had to be added simultaneously for a new rule to be discovered. The average orders of emergence of all lineages across 1,000 runs for each seed sets are provided in **Table S13**. **B**. Average structure class composition along the trajectory for the ATP seed set. Other seed sets are provided in **Figure S13**. The 'mixed' structure class refers to lineages with both α/β and α+β elements.

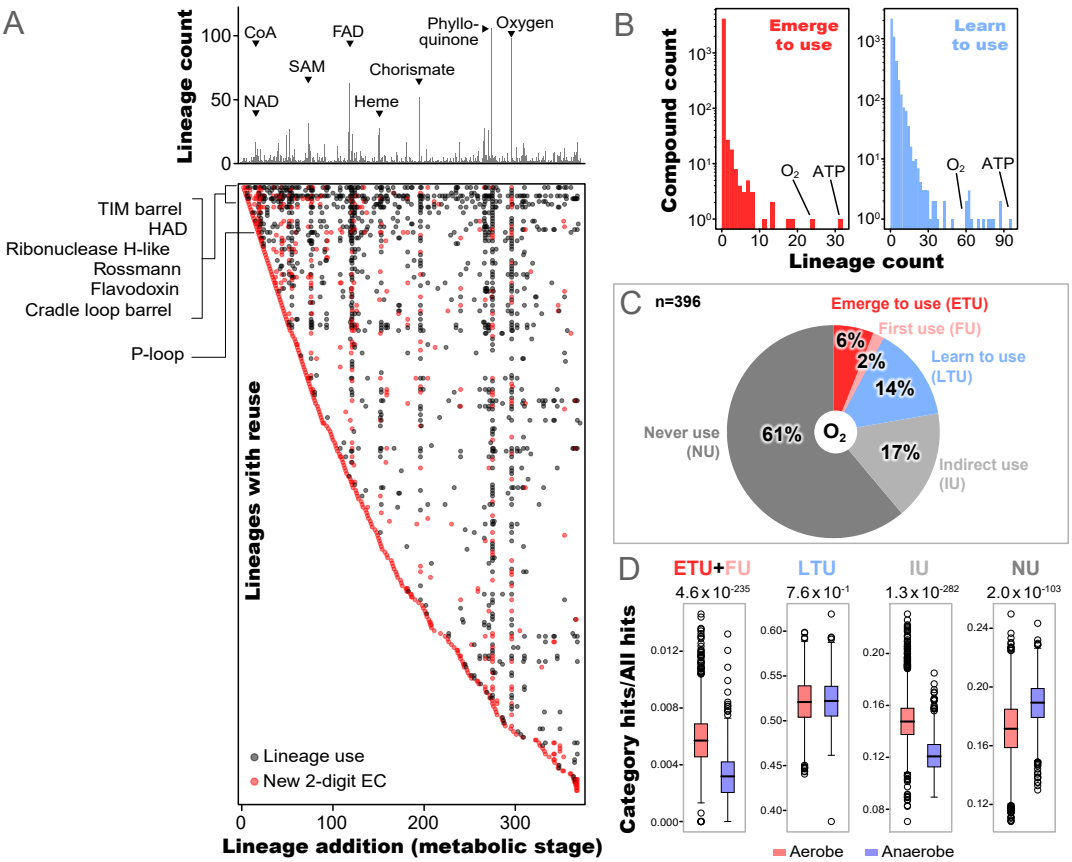

Figure 4. Emergence and reuse of lineages. **A**. A global view of lineage reuse calculated from the trajectory in **Figure 3A**. Total lineage use counts at various stages of metabolism (upper) and the patterns of lineage emergence and reuse along the trajectory (lower). Lineages are shown one per row along the y-axis, with points along the x-axis indicating metabolic stages where a lineage is associated with a rule. The lineages have been ordered from top to bottom by the order of emergence. Only lineages used in 2 or more metabolic stages are shown. Red color indicates metabolic stages where a lineage catalyzes a reaction with a new 2-digit EC number. **B**. Histogram of the number of lineages that emerge to use (left) or learn to use (right) a given compound. ATP and molecular oxygen are both strong drivers of lineage emergence and reuse. **C**. Breakdown of lineage relationships with oxygen. See main text for description of each class. Panels B and C are calculated from 1,000 runs of the ATP seed set. Only classifications consistent across ≥75% of runs are assigned. In the case of oxygen, all lineages satisfy this threshold. A full list of metabolic lineages and their category is available in **Table S14**. **D**. Enrichment of lineages associated with each oxygen-use category in bacteria with predicted aerobic or anaerobic physiologies. For each genome (1,630 aerobes and 1,099 anaerobes), the occurrence of lineages within a given usage category was divided by the occurrence of all ECOD lineages. Indicated p-values were calculated by Mann-Whitney U tests.

462 **Methods**

463 *Lineage-KO group associations*. Associations between KEGG (Kyoto Encyclopedia of Genes and
464 Genomes) orthologous (KO) groups (Kanehisa and Goto 2000) and Evolutionary Classification of
465 Domains (ECOD) X-groups (Cheng et al. 2014) were constructed as follows: Genes from 8,030 KO
466 groups were clustered at 80% sequence identity by CD-HIT (Fu et al. 2012) using a word size of 5.
467 Hidden Markov models (HMMs) from version 279 of the ECOD database were queried against the
468 representative KEGG genes of each KO group using hmmsearch (Mistry et al. 2013) with the search
469 space set to 106,052,079 sequences. Domains with an independent E-value of less than 1×10^{-4}
470 and an HMM coverage of greater than 70% were considered potential hits. Hits were assigned to
471 gene regions in order of increasing E-value using the 'envelope coordinates.' If two hits overlapped
472 by more than 30% of either hit, the hit with the higher E-value was discarded. Finally, reconciled
473 hits associated with a gene were mapped to their corresponding X-groups when such mappings are
474 unambiguous. X-group sets corresponding to at least 20% of the representative genes within a KO
475 group were taken to be catalytically competent. Nested sets, in which one observed X-group set is
476 a subset of another X-group set for a given KO group, were retained. Note that X-groups are referred
477 to as *lineages* throughout the main text.
478
479 *Reaction-KO group associations*. Reaction-KO group associations were extracted from the KEGG
480 database, accessed through the KEGG web interface or the KEGG REST API (either directly or
481 through Biopython) (Kanehisa and Goto 2000; Cock et al. 2009). Three types of annotations were
482 used: a) direct KO-reaction mappings, b) KO-reaction associations linked through KEGG modules,
483 and c) KO-reaction associations linked though EC numbers. For reactions associated to exactly
484 one KO group, no further curation was performed. For reactions that map to multiple KO groups, all
485 KO groups were reviewed to annotate interdependence relationships between KO groups. KO
486 groups were generally assumed to be isozymes capable of catalyzing the same reaction
487 independently unless literature searching indicated that they are interdependent components
488 (such as subunits of a protein complex). KO groups labeled as membrane anchors, effectors,
489 coupling domains, or auxiliary domains were considered nonessential and not included as reaction
490 dependencies.
491
492 *Reaction-lineage associations*. Based on the lineage-KO group and reaction-KO group associations
493 described above, reaction-lineage associations were constructed via their mutual linkage through
494 KO groups. Each reaction-lineage association is referred to as a *rule* (see main text for more details;
495 **Figures 1B** and **S2**).
496
497 *Reaction-cofactor associations*. Detailed cofactor dependencies for 5,259 reactions were
498 manually reconciled from annotations in UniProt (UniProt Consortium 2023), Expasy (Gasteiger et

al. 2003), PDBe (Armstrong et al. 2020), EBI (Rodriguez-Tomé et al. 1996), and KEGG (Kanehisa and Goto 2000) (**Table S5**). Each cofactor is associated with one or several closely related KEGG compounds (**Table S6**) that can, to a reasonable approximation, satisfy the mechanistic role of that cofactor class. A full description of reaction-cofactor associations, and their implementation in the standard network expansion, was described previously (Goldford, 2024).

Metabolic model. The list of allowed reactions was constructed based on biochemical reactions from KEGG (Kanehisa and Goto 2000) with added cofactor dependencies. Some forward or reverse reactions were removed from the network based on thermodynamic constraints calculated using the component contribution method (Noor et al. 2013) and the eQuilibrator python API (Beber et al, 2022). If the standard molar free energies were above zero at pH = 7.0, pMg = 3.0, an ionic strength of 0.25 M, temperature T = 298.15, and allowing for a concentration range of 10^{-7}-10^{-1} M for all reactants and products, the reaction was permitted. Non-autocatalytic purine synthesis reactions (Goldford et al. 2024) were included to allow expansion to the entire largest connected component in KEGG. These reactions were not given KO group or enzyme associations. Reactions with elemental inconsistency and reactions producing H_2O_2 before the discovery of molecular oxygen were removed. Allowed reactions are listed in **Table S5** and removed reactions are listed in **Table S7**.

Seed set. The base seed set comprises 80 compounds including metals, inorganic molecules, CO_2, H_2S, H_2, orthophosphate, ammonia, and organic substrates that can be produced abiotically when reacting glyoxylate with pyruvate (Muchowska et al. 2019). Many of the organic substrates are fundamental carbon compounds found in the reductive TCA cycle (Braakman and Smith 2013). The base seed set also includes 10 amino acids with plausible prebiotic availability (glycine, *L*-alanine, *L*-aspartate, *L*-glutamate, *L*-isoleucine, *L*-leucine, *L*-proline, *L*-serine, *L*-threonine, *L*-valine) (Higgs and Pudritz 2009; Longo and Blaber 2012) and that are likely sufficienct for protein folding (Akanuma et al. 2002; Longo et al. 2013; Yagi et al. 2021). Note that the ATP pre-expansion seed set also contains *L*-asparagine, *L*-cysteine, and *L*-glutamine as well. Addition of *L*-valine to the seed set accelerates the discovery of CoA but the compound emergence orders with and without the 10 amino acids are otherwise highly rank correlated (**Figure S4**). Nine additional seed sets, each with successively more compounds, were considered for the enzyme-gated network expansion analysis: Eight 'pre-expanded' seed sets which correspond to the points of emergence of various cofactors (PLP, ATP, NADH, SAM, CoA, FAD, ThDP, and cobalamin) in the standard network expansion trajectory, and one with all metabolites in scope. Seed sets are listed in **Table S4**.

Standard metabolic network expansion. The network expansion algorithm was implemented using the networkExpansionPy python package (https://github.com/jgoldford/networkExpansionPy). A

536 detailed explanation of how the algorithm works can be found in previous publications (Handorf et
537 al. 2005; Goldford et al. 2017) and the main text.
538
539 *Ablation analyses*. To simulate the effect of ablating a single lineage, all rules involving that linage
540 were removed and only those reactions for which a rule still exists were allowed. To simulate the
541 effect of ablating a compound (in this work, molecular oxygen), all reactions involving that
542 compound, either as a reactant or a product, were removed from the network.
543
544 *Enzyme-gated network expansion*. An implementation of enzyme-gated network expansion in
545 Python is available at https://github.com/jgoldford/networkExpansionPy. A natural language
546 description of the algorithm is presented in the main text. For a pseudocode implementation of the
547 algorithm, refer to **Figure S20**.
548
549 *Phyletic distribution score*. Genome data were obtained from GTDB (Parks et al. 2022) for archaea
550 and bacteria (version 207) and EukProt (version 3) (Richter et al. 2022) for eukaryotes. HMM profiles
551 from ECOD (version 279) (Cheng et al. 2014) were queried against the genomes using hmmsearch
552 (Mistry et al. 2013) with the search space set to 106,052,079 sequences. Only hits with an
553 independent E-value of less than 1×10^{-5} and an HMM coverage of greater than 60% were kept as
554 potential hits. In cases where there was an overlap of more than 25% between two or more hits,
555 only the hit with lower E-value was retained. Hits were assigned to X-groups (lineages) when
556 unambiguous association was possible. For each lineage, a phyletic distribution score, D, was
557 calculated as:

558 $$\mathrm{D} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{f}_n$$

559 where f_n is the fraction of genomes containing the lineage within each phylum summed over all
560 phyla, and N is the number of phyla. Phyletic distribution score was calculated separately for
561 archaea, bacteria, and eukaryotes. Prokaryotic distribution score is the average of the archaea and
562 bacteria phyletic distribution scores. Distribution scores for each lineage are provided in **Table S1**.
563
564 *Lineage-metabolite emergence relationships*. Each lineage-metabolite pair was classified into one
565 of five emergence relationships (see main text) for a given seed set. The ATP pre-exansion was used
566 for all lineage-metabolitle emergence relationships reported in the main text. The classification
567 proceeded as follows: For each lineage, the *first reactions* – those reactions catalyzed by the
568 lineage in the first network expansion step after lineage addition – were identified. A lineage-
569 metabolite pair was classified as *emerge to use* (ETU) if, in at least 75% of runs, all of the first
570 reactions of that lineage require the metabolite as a *reactant*. A lineage-metabolite pair was
571 classified as *first use* (FU) if, in at least 75% of runs, some but not all of the first reactions of that

lineage require the metabolite as a reactant. If the 75% threshold is not met for either an emerge to use or first use classification, an emergence relationship was not assigned. In the case of molecular oxygen, no lineage falls into the unassigned category (**Figure 4C**). For lineages where none of the first reactions require the metabolite as a reactant across all runs, classification was based on the following criteria: A *learn to use* (LTU) relationship was assigned if a lineage catalyzes a reaction using the metabolite as a reactant sometime later in the trajectory. If a lineage never catalyzes any reaction that requires the metabolite, but does catalyze reactions that become unreachable when the metabolite is ablated from the network, the relationship is classified as *indirect use* (IU). If a lineage never catalyzes reactions that either require the metabolite as a reactant, or become unreachable when the metabolite is ablated from the network, the relationship is classified as *never use* (NU). For any lineage, multiple metabolites can fall into the same category (*e.g.*, Rossmanns learn to use FAD and molecular oxygen), but each metabolite can only be related to a lineage by a single relationship (that is, lineage-metabolite pairs are uniquely classified). **Figure S18** presents a detailed classification of oxygen-lineage emergence relationships.

Bacterial physiology annotation. Bacterial oxygen physiology annotations were obtained from a study that aggregated annotations from the Joint Genome Institute (JGI) Genomes Online Database (GOLD) (Madin et al. 2020; Goldford et al. 2023). Taxa labeled 'obligate aerobe' or just 'aerobe' were classified as aerobes, and taxa labeled 'obligate anaerobe' or 'anaerobe' were classified as anaerobes. NCBI taxonomy IDs were mapped to GTDB genome IDs using metadata provided on the FTP server from the GTDB (Parks et al. 2022). See **Table S15** for the full list of physiology annotations.

Data and code availability. The analyses presented in this study are based on data retrieved from the KEGG database which can be accessed through the URL (https://www.kegg.jp/), subject to their licensing policies and agreements. Processed datasets and code used in this study are available in **Supplementary Tables 1-15** and via the following Zenodo repository: https://doi.org/10.5281/zenodo.15580734 and GitHub repository: https://github.com/tseamuscorlett/enzyme-gated-network-expansion.

Statistical analysis. Pearson's correlation, Spearman's correlations, and two-sided, non-parametric Mann–Whitney U tests were performed using the `pearsonr`, `spearmanr` and `mannwhitneyu` functions, respectively, from the `scipy.stats` python module, version 1.14.1 in python 3.12.0.

606 **References**

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

609 2. Braakman, R. & Smith, E. Metabolic evolution of a deep-branching hyperthermophilic
610 chemoautotrophic bacterium. *PLoS One* **9**, e87950 (2014).

611 3. Medvedev, K. E., Kinch, L. N., Dustin Schaeffer, R., Pei, J. & Grishin, N. V. A Fifth of the
612 Protein World: Rossmann-like Proteins as an Evolutionarily Successful Structural unit. *J Mol Biol*
613 **433**, 166788 (2021).

614 4. Hol, W. G., van Duijnen, P. T. & Berendsen, H. J. The alpha-helix dipole and the properties
615 of proteins. *Nature* **273**, 443–446 (1978).

616 5. Davidi, D., Longo, L. M., Jabłońska, J., Milo, R. & Tawfik, D. S. A Bird's-Eye View of Enzyme
617 Evolution: Chemical, Physicochemical, and Physiological Considerations. *Chem Rev* **118**, 8786–
618 8797 (2018).

619 6. Longo, L. M., Petrović, D., Kamerlin, S. C. L. & Tawfik, D. S. Short and simple sequences
620 favored the emergence of N-helix phospho-ligand binding sites in the first enzymes. *Proc Natl*
621 *Acad Sci U S A* **117**, 5310–5318 (2020).

622 7. Kim, J., Kershner, J. P., Novikov, Y., Shoemaker, R. K. & Copley, S. D. Three serendipitous
623 pathways in E. coli can bypass a block in pyridoxal-5'-phosphate synthesis. *Mol Syst Biol* **6**, 436
624 (2010).

625 8. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary
626 perspective. *Annu Rev Biochem* **79**, 471–505 (2010).

627 9. Copley, S. D. The physical basis and practical consequences of biological promiscuity.
628 *Phys Biol* (2020) doi:10.1088/1478-3975/ab8697.

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

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

633 12. Goldford, J. E., Hartman, H., Marsland, R., 3rd & Segrè, D. Environmental boundary
634 conditions for the origin of life converge to an organo-sulfur metabolism. *Nat Ecol Evol* **3**, 1715–
635 1724 (2019).

636 13. de Duve, C. The Beginnings of Life on Earth. *American Scientist* **83**(5), 428–437 (1995).

637 14. Goldford, J. E., Smith, H. B., Longo, L. M., Wing, B. A. & McGlynn, S. E. Primitive purine
638 biosynthesis connects ancient geochemistry to modern metabolism. *Nat Ecol Evol* **8**, 999–1009
639 (2024).

640 15. Winstanley, H. F., Abeln, S. & Deane, C. M. How old is your fold? *Bioinformatics* **21 Suppl**
641 **1**, i449–58 (2005).

642 16. Edwards, H., Abeln, S. & Deane, C. M. Exploring fold space preferences of new-born and

643 ancient protein superfamilies. *PLoS Comput Biol* **9**, e1003325 (2013).

644 17. Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv Protein Chem* **34**,
645 167–339 (1981).

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

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

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

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

654 22. Caetano-Anollés, G., Kim, H. S. & Mittenthal, J. E. The origin of modern metabolic networks
655 inferred from phylogenomic analysis of protein architecture. *Proc Natl Acad Sci U S A* **104**, 9358–
656 9363 (2007).

657 23. Ma, B.-G. *et al.* Characters of very ancient proteins. *Biochem Biophys Res Commun* **366**,
658 607–611 (2008).

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

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

663 26. Beber, M. E. *et al.* eQuilibrator 3.0: a database solution for thermodynamic constant
664 estimation. *Nucleic Acids Res* **50**, D603–D609 (2022).

665 27. Alva, V., Söding, J. & Lupas, A. N. A vocabulary of ancient peptides at the origin of folded
666 proteins. *Elife* **4**, e09410 (2015).

667 28. Yagi, S. & Tagami, S. An ancestral fold reveals the evolutionary link between RNA
668 polymerase and ribosomal proteins. *Nat Commun* **15**, 5938 (2024).

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

671 30. Farías-Rico, J. A., Schmidt, S. & Höcker, B. Evolutionary relationship of two ancient protein
672 superfolds. *Nat Chem Biol* **10**, 710–715 (2014).

673 31. Kolodny, R., Nepomnyachiy, S., Tawfik, D. S. & Ben-Tal, N. Bridging Themes: Short Protein
674 Segments Found in Different Architectures. *Mol Biol Evol* **38**, 2191–2208 (2021).

675 32. Qiu, K., Ben-Tal, N. & Kolodny, R. Similar protein segments shared between domains of
676 different evolutionary lineages. *Protein Sci* **31**, e4407 (2022).

677 33. Nepomnyachiy, S., Ben-Tal, N. & Kolodny, R. Global view of the protein universe. *Proc Natl*
678 *Acad Sci U S A* **111**, 11691–11696 (2014).

679 34. Choi, I.-G. & Kim, S.-H. Evolution of protein structural classes and protein sequence
680 families. *Proc Natl Acad Sci U S A* **103**, 14056–14061 (2006).

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

683 36. Romero Romero, M. L. *et al.* Simple yet functional phosphate-loop proteins. *Proc Natl*
684 *Acad Sci U S A* **115**, E11943–E11950 (2018).

685 37. Longo, L. M. *et al.* On the emergence of P-Loop NTPase and Rossmann enzymes from a
686 Beta-Alpha-Beta ancestral fragment. *Elife* **9**, (2020).

687 38. Vyas, P. *et al.* Helicase-like functions in phosphate loop containing beta-alpha
688 polypeptides. *Proc Natl Acad Sci U S A* **118**, (2021).

689 39. Petrov, A. S. *et al.* History of the ribosome and the origin of translation. *Proc Natl Acad Sci*
690 *U S A* **112**, 15396–15401 (2015).

691 40. Kovacs, N. A., Petrov, A. S., Lanier, K. A. & Williams, L. D. Frozen in Time: The History of
692 Proteins. *Mol Biol Evol* **34**, 1252–1260 (2017).

693 41. Crowe, S. A. *et al.* Atmospheric oxygenation three billion years ago. *Nature* **501**, 535–538
694 (2013).

695 42. Madin, J. S. *et al.* A synthesis of bacterial and archaeal phenotypic trait data. *Sci Data* **7**,
696 170 (2020).

697 43. Goldford, J. E., Murali, R., Valentine, J. S. & Fischer, W. W. Metabolic evolution of
698 pyranopterin-dependent biochemistry. *bioRxiv* (2023) doi:10.1101/2023.09.01.555371.

699 44. Goldford, J. E. & Segrè, D. Modern views of ancient metabolic networks. *Curr. Opin. Syst.*
700 *Biol.* **8**, 117–124 (2018).

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

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

705 47. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular
706 biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

707 48. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids*
708 *Res* **51**, D523–D531 (2023).

709 49. Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and
710 analysis. *Nucleic Acids Res* **31**, 3784–3788 (2003).

711 50. Armstrong, D. R. *et al.* PDBe: improved findability of macromolecular structure data in the
712 PDB. *Nucleic Acids Res* **48**, D335–D343 (2020).

713 51. Rodriguez-Tomé, P., Stoehr, P. J., Cameron, G. N. & Flores, T. P. The European
714 Bioinformatics Institute (EBI) databases. *Nucleic Acids Res* **24**, 6–12 (1996).

715 52. Muchowska, K. B., Varma, S. J. & Moran, J. Synthesis and breakdown of universal
716 metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019).

717 53. Higgs, P. G. & Pudritz, R. E. A thermodynamic basis for prebiotic amino acid synthesis and
718 the nature of the first genetic code. *Astrobiology* **9**, 483–490 (2009).

719 54. Longo, L. M. & Blaber, M. Protein design at the interface of the pre-biotic and biotic worlds.
720 *Arch Biochem Biophys* **526**, 16–21 (2012).

721 55. Akanuma, S., Kigawa, T. & Yokoyama, S. Combinatorial mutagenesis to restrict amino acid
722 usage in an enzyme to a reduced set. *Proc Natl Acad Sci U S A* **99**, 13549–13553 (2002).

723 56. Longo, L. M., Lee, J. & Blaber, M. Simplified protein design biased for prebiotic amino acids
724 yields a foldable, halophilic protein. *Proc Natl Acad Sci U S A* **110**, 2135–2139 (2013).

725 57. Yagi, S. *et al*. Seven Amino Acid Types Suffice to Create the Core Fold of RNA Polymerase. *J*
726 *Am Chem Soc* **143**, 15998–16006 (2021).

727 58. Parks, D. H. *et al*. GTDB: an ongoing census of bacterial and archaeal diversity through a
728 phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic*
729 *Acids Res* **50**, D785–D794 (2022).

730 59. Richter, D. J. *et al*. EukProt: A database of genome-scale predicted proteins across the
731 diversity of eukaryotes. *Peer Community J*. **2**, (2022).