The history of enzyme evolution embedded in metabolism

Tatsuya Corlett[1], Harrison B. Smith[1,2,*], Eric Smith[1,3], Joshua E. 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

protein evolution, metabolism, KEGG, ECOD, network expansion

Abstract (250 words)

Whereas phylogenetic reconstructions are a primary record of protein evolution, it is unknown whether the deep history of enzymes are encoded at higher levels of biological organization. Here, we demonstrate that the emergence and reuse history of enzymatic folds is embedded within the web of metabolite-cofactor-enzyme interdependencies that comprise biosphere-scale metabolic reaction networks. Using a simple network analysis approach, we reconstruct the relative ordering of enzymatic fold emergence and, where possible, the first reaction(s) that each enzymatic fold catalyzed. We find that a large majority of enzymatic folds were sufficient as independent additions to open new avenues for metabolic growth. The resulting network-based histories are broadly concordant with enzyme phyletic distribution in prokaryotes, a proxy for enzyme age. Our results suggest that the earliest enzyme-mediated metabolisms were enriched for α/β proteins, likely due to their strong association with cofactor utilization, and that α-proteins preferentially emerge at later stages. The cradle-loop barrel, a member of the small β-barrel metafold, is predicted to be the founding β-fold, in agreement with analyses of ribosome structure. An examination of how the protein universe responded to the biological production of molecular oxygen reveals that the adaptation of existing enzymatic folds, not novel fold emergence, was the primary driver of metabolic evolution. This work presents a self-consistent model of metabolic and enzyme evolution, key progress towards integrating diverse perspectives into a unified history of protein evolution.

Significance Statement (120 words)

Enzyme emergence is an ongoing process that began ~4 billion years ago. Here, we show that the modern biosphere-scale network of metabolic reactions and enzymes is an archive of enzyme history independent from, but concordant with, phylogenetics. Based on this record, we predict the order of enzyme emergence from before the last universal common ancestor up until the biological production and metabolic utilization of molecular oxygen. We find that while α/β proteins dominated primitive enzyme-mediated metabolism, other folds — including the cradle-loop barrel, which is a member of the small β-barrel metafold — were likely important early contributors. This study represents key progress towards building an internally consistent, joint history of metabolic reactions and enzymes.

Introduction

The last universal common ancestor (LUCA) possessed a repertoire of complex protein domains (Moody et al. 2024), indicating that key events in protein evolution occurred even earlier in the history of life. While extant protein sequences are a rich archive of protein evolution, understanding protein fold emergence and evolution prior to the LUCA poses a unique challenge (Winstanley et al. 2005; Edwards et al. 2013). Because the LUCA represents the approximate point at which historical inference by comparative approaches converges, resolution at the earliest stages of protein evolution is poor (Winstanley et al. 2005; Edwards et al. 2013). Thus, alternative records of protein evolution are needed to reconstruct the fold emergence process. What records may be suitable for this purpose?

One alternative record of early protein evolution that has been proposed is the ribosome, in which domains or pieces of domains have been "frozen in time" by accretionary processes (Petrov et al. 2015; Kovacs et al. 2017). By analyzing the layered structure of the ribosome, an early history of protein evolution can be inferred. This approach indicates the early emergence of β-structural elements prior to other structure classes — a surprising result, given that β-protein composition within proteomes increases markedly after the emergence of chaperones (Rebeaud et al. 2021), perhaps suggesting that β-proteins have intrinsic deficits to foldability (Plaxco et al. 1998; Takács et al. 2024). Although the extent to which ribosomal proteins inform the emergence of other protein folds is unknown, this work represents a remarkable insight into the types of pre-LUCA archives that may record the earliest stages of protein evolution.

Another layered system in biology is the structure of metabolism (Handorf et al. 2005; (Raymond and Segrè 2006). Whereas physical stratification suggests an emergence order in the ribosome, metabolic layering results from the network of substrate, cofactor, and enzyme requirements for each metabolic reaction. Viewing metabolism through this lens has produced significant insights into the chemistry of primitive metabolic systems (Goldford et al. 2017), the environment of the earliest life (Goldford et al. 2019), and the existence of forgotten chemical reactions (Goldford et al. 2024). We note that these prior studies were independent of enzyme history, leaving a gap in understanding about the co-evolutionary history of metabolic networks and enzymes. We hypothesize that the early history of enzyme emergence and evolution may also be written into the web of known metabolic interdependencies.

Here, we reconstruct the order of enzyme evolution based on the layered structure of metabolism using a novel network analysis algorithm that we refer to as *enzyme-gated network expansion*. We demonstrate that the complex, multifaceted associations between metabolites, reactions, cofactors, and enzymes can constrain the order of enzyme discovery. We find that the handoff from geochemistry-mediated to enzyme-mediated metabolic exploration soon after the availability of nucleotides yields a robust enzyme emergence trajectory with a consistent and

realistic progression in enzyme complexity. Our reconstructed order of enzymatic fold emergence is broadly concordant with phylogenetic information. The results suggest an intriguing transition from easily discovered α/β domains to those dependent on more refined evolutionary information and control systems. Using the biological production of molecular oxygen as an example of metabolic upheaval, we characterize the interplay between enzyme reuse — in which a previously discovered enzyme performs a new function — and the emergence of novel enzymatic folds. 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.

Materials and Methods

Connecting protein evolution and metabolism. Many, though not all, metabolic reactions are catalyzed by enzymes. Enzymes are composed of structural units called protein domains (often several) that can be shared across multiple enzyme families, resulting in complex patterns of partial homology between enzymes. The definition of a protein domain is not fully agreed upon, but is generally meant to entail structural separation, reuse, and independent foldability (Richardson 1981). Consequently, protein domains are a fundamental unit of protein evolution (Lo Conte et al. 2000; Knudsen and Wiuf 2010; Cheng et al. 2014). To reconstruct the history of enzyme evolution from metabolism, we must first connect each metabolic enzyme with both the reactions that it catalyzes and the protein domains that comprise its structure (**Figure 1A**).

Connecting enzymes to the reactions that they catalyze has been a longstanding challenge in molecular biology (Sorokina et al. 2014; Hadadi et al. 2019). The Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto 2000) is a manually curated database of experimentally validated enzyme-reaction associations (**Figure 1B**). The KEGG database includes reactions and enzymes from archaea, bacteria, and eukaryotes, and diverse metabolic strategies are represented (Kanehisa and Goto 2000). For example, the KEGG database includes the reactions for the seven major carbon fixation pathways. For these reasons, KEGG is considered a repository of biosphere-scale metabolism (Raymond and Segrè 2006). Here, we employ a previously described model of biosphere-level metabolism based on the KEGG database (Goldford et al. 2024). In total, our metabolic model (Goldford et al. 2024) comprises 4,294 metabolites and 7,678 reactions. These reactions are associated with 4,331 enzyme orthologous groups, referred to as KO groups in the KEGG nomenclature.

The Evolutionary Classification of Domains (ECOD) database (Cheng et al. 2014; Cheng et al. 2015) employs a hierarchical classification scheme that groups protein domains into nested tiers based on evolutionary relatedness (**Figure 1C**). In ECOD, closely related domains with highly similar sequences and structures are grouped into protein families (or F-groups). Related protein families are then collected into an X-group, the broadest evolutionary classification in ECOD, which emphasizes more distant evolutionary relationships. X-groups are hypothesized to have evolved from a single founder protein domain. We will refer to X-groups as *fold lineages* to differentiate this grouping from the narrower protein family grouping. There are 2,458 X-groups (fold lineages) and 15,144 F-groups (protein families) in ECOD. Linking fold lineages to KEGG orthologous groups can be done using the ECOD-curated database of hidden Markov models (HMMs). A detailed description of the procedure used to associate biochemical reactions, cofactors, enzyme orthology groups, and fold lineages is provided in the **Supplemental Methods**.

Protein structure classification. Each fold lineage (X-group) in ECOD belongs to one of six structure classes based on the organization of α-helices and β-sheets within the protein domain. α- and β-proteins are comprised predominantly of α-helices or β-sheets, respectively; α+β proteins include both α and β elements but they are structurally segregated from each other; α/β proteins also contain both α and β elements but, in this case, the two elements alternate along the chain; mixed proteins have features of both α+β and α/β proteins; and finally, "other" proteins lack well defined secondary structure elements. **Table S1** lists the structure classification of each fold lineage (X-group), based on the class–architecture mapping in **Table S2**.

Standard metabolic network expansion. To model the stepwise evolution of biosphere-scale metabolism, the network expansion algorithm (Handorf et al. 2005; Goldford et al. 2017) was employed, which is described briefly here: First, a set of precursors, called *seed compounds*, are assumed. These seed compounds then undergo all possible reactions within a pre-defined set of allowed reactions, subject to reactant and cofactor requirements. The products of these reactions are then added to the pool of available compounds used in the next iteration. This process is repeated until no new compounds are discovered. Once a compound is produced within the network, it is never removed. Network expansion concerns reachability; it is not a kinetic model. The final set of compounds reached is referred to as the scope. *Ablation analyses* remove a component of the metabolic model — either a metabolite, a reaction, a fold lineage, or sets thereof — to assess its role in the expanding network. To ablate a single fold lineage, all reactions that require that fold lineage are removed. Likewise, ablation of a compound removes all reactions involving that compound, either as a reactant or a product. The network expansion implementation used in this work is available as a python package on GitHub (https://github.com/jgoldford/networkExpansionPy).

Metabolic model. The reactions in our metabolic model were curated from KEGG (Kanehisa and Goto 2000) with added cofactor requirements (Goldford et al. 2024). Some forward or reverse reactions were removed from the network based on thermodynamic constraints calculated by the component contribution method (Noor et al. 2013) using the eQuilibrator python API (Beber et al, 2022). If a reaction's standard molar free energy is ever negative 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. In the standard KEGG database, purines are required for purine synthesis. This circular causality blocks metabolic expansion early in the trajectory unless a non-autocatalytic pathway for purine synthesis is included (Goldford et al. 2024). Thus, non-autocatalytic purine synthesis reactions were added to the model. These reactions were not given KO group or enzyme associations as none are known. Reactions with elemental inconsistency and reactions producing H_2O_2 before the discovery of molecular oxygen

were removed. Allowed reactions are listed in **Table S3** and removed reactions are listed in **Table S4**.

Seed set. Network expansion requires a set of seeding compounds from which the metabolic trajectory starts. The base seed set used in this work comprises 80 compounds including metals, inorganic molecules, CO_2, H_2S, H_2, orthophosphate, ammonia, and 20 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). These amino acids are predicted to have been part of a primitive genetic code (Copley et al. 2005) and are sufficient to encoding stable, foldable proteins (Akanuma et al. 2002; Longo et al. 2013; Yagi et al. 2021; Giacobelli et al. 2025). Nine additional seed sets, each with successively more compounds, were considered for the enzyme-gated network expansion analysis: Eight 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 from the base seed set (and thus form nested sets); and one seed set that contains all metabolites in our metabolic model. Note that the ATP-expanded seed set also contains the canonical amino acids *L*-asparagine, *L*-cysteine, and *L*-glutamine, as well as the basic amino acid *L*-diaminobutyrate, which is hypothesized to have been part of the primitive genetic code (Copley et al. 2005). The seed sets are listed in **Table S5**.

Enzyme-gated network expansion. An implementation of enzyme-gated network expansion in Python is available at https://github.com/jgoldford/networkExpansionPy. A natural language description of the algorithm is presented in the main text. A pseudocode implementation of the algorithm is presented in **Figure S1**.

Phyletic distribution score. Genome data were obtained from Genome Taxonomy Database (GTDB) (Parks et al. 2022) for archaea and bacteria (version 207) and EukProt (version 3) (Richter et al. 2022) for eukaryotes. HMM profiles from ECOD (version 279) (Cheng et al. 2014) were queried against the genomes using hmmsearch (Mistry et al. 2013) with the search space set to 106,052,079 sequences. Only hits with an independent E-value of less than 1×10^{-5} and an HMM coverage of greater than 60% were kept as potential hits. In cases where there was an overlap of more than 25% between two or more hits, only the hit with lower E-value was retained. Hits were assigned to X-groups (fold lineages) only when unambiguous association was possible. For each fold lineage, a phyletic distribution score, D, was calculated as:

$$D = \frac{1}{N}\sum_{n=1}^{N} f_n$$

where f_n is the fraction of genomes containing the fold lineage within each phylum and N is the number of phyla. Phyletic distribution score was calculated separately for archaea, bacteria, and eukaryotes. Prokaryotic distribution score is the average of the archaea and bacteria phyletic distribution scores. Distribution scores for each fold lineage are provided in **Table S1**.

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 "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 GTDB (Parks et al. 2022). See **Table S6** 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 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 a Zenodo repository (https://doi.org/10.5281/zenodo.15580734) and a GitHub repository (https://github.com/tseamuscorlett/enzyme-gated-network-expansion).

Statistical analysis. Pearson correlations, 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 3.12.4.

Results and Discussion

Enzymatically versatile α/β proteins are key mediators of metabolism

Approximately 75% of the 7,678 reactions in our metabolic model are associated with one or more fold lineages (**Table S7**). However, just 396 fold lineages — less than 17% of those identified in the ECOD database — are represented. Nearly 90% of these "metabolic fold lineages" are detected in both archaeal and bacterial genomes, compared to just 42% for the full set of fold lineages (**Figure 2A, Figure S2, Table S1**). All protein structure classes (*e.g.*, α, β, α+β, α/β; see **Methods** for a description of each structure class) are present among the metabolic fold lineages, though α/β structures are enriched 3.2-fold in the metabolic fold lineages relative to all fold lineages (**Figure 2B, Figure S2**). Likewise, the majority (60%) of enzyme orthology groups that catalyze a reaction in our metabolic model are associated with at least one protein domain that adopts an α/β structure (**Figure 2C**). The number of reactions associated with each fold lineage (**Figure 2D**, y-axis) reveals a minority of exceptionally versatile fold lineages, with the three most versatile ones — Rossmann, TIM barrel, and flavodoxin — each adopting an α/β structure. The outsized role of α/β proteins in metabolism has been noted previously (Caetano-Anollés et al. 2007; Ma et al. 2008), and may relate to an intrinsic preference for, or a historical association with, (di)nucleotide cofactors (Longo et al. 2020; Yanai et al. 2025).

Both versatile and specialist fold lineages enable metabolic innovation

Counting the number of reactions associated with a fold lineage emphasizes the role of versatility. Some fold lineages, however, may catalyze a small number of essential reactions. To account for these cases, we employed metabolic network expansion (Handorf et al. 2005). Metabolic network expansion predicts the relative discovery order for a set of reactions based on their substrate requirements. Given a set of "seed compounds" that are provided at the beginning of the simulation, reachable reactions are activated to generate new metabolites that, in turn, activate new reactions. The process is repeated until no new metabolites can be discovered — either because all metabolites have been discovered or because a key reactant or cofactor is absent and further expansion is not possible. The set of all discovered metabolites is referred to as the scope. By removing the reactions that are dependent on a fold lineage ("fold lineage ablation") and calculating the resulting scope, specialist enzymes that catalyze a small number of essential reactions can be uncovered.

Using seed compounds based on a primitive Wood-Ljungdahl pathway (Muchowska et al. 2019) and TCA cycle (Braakman and Smith 2013) (see **Methods** for more details), network expansion was performed after the independent ablation of each fold lineage (x-axis of **Figure 2D** and **Table S8**). Foremost, the ablation analysis highlights the hierarchical structure of metabolism

(Raymond and Segrè 2006), with the resulting networks forming nested layers: Small networks that lack both CoA and O_2, medium-sized networks that include CoA but lack O_2, and large networks that include both CoA and O_2 (**Figure 2D**). As expected, both versatile fold lineages (*e.g.*, Rossmann, TIM barrel, and flavodoxin) and specialist fold lineages (*e.g.*, the three *L*-aspartase fold lineages that mediate purine biosynthesis and X-group 3997, which mediates chlorophyll biosynthesis) were essential for complete metabolic expansion (**Figure 2D**). To our surprise, however, ablation of several versatile fold lineages (*e.g.*, the haloacid dehydrogenase lineage, abbreviated HAD) had only a minor impact on the scope. In total, only 37 fold lineages resulted in a contraction of the scope by >20% upon individual ablation. The fact that these 37 fold lineages are more widely distributed across prokaryotes than other fold lineages ($p = 7.9\times10^{-14}$, Mann-Whitney U test; **Figures 2E**), and include both versatile and specialist fold lineages, highlights the key contributions that both kinds of fold lineages make to metabolic innovation.

Biosphere-scale metabolism is rich with alternatives

Does the core of metabolism require just 37 fold lineages, with all others playing a supporting role? Ablation of all fold lineages except the 37 that significantly contract the scope upon individual ablation (*i.e.*, network expansion with just 37 fold lineages) resulted in a network with just 290 metabolites (~7% of full scope). The robustness we observe to single fold lineage ablations, but not to this multi-fold lineage ablation, is due to two main factors: First, approximately 12% of enzyme-catalyzed reactions can be catalyzed by entirely distinct sets of fold lineages (**Figure S3**). Second, the same metabolite can be achieved by alternative pathways, reflecting the composite, biosphere-scale nature of our metabolic model. Thus, ablating a single fold lineage (or a single reaction (Goldford et al. 2024)) reveals the first point along the expansion trajectory where it becomes essential for further network growth despite these two forms of redundancy.

Neither an analysis of enzyme versatility (**Figure 2D**, y-axis) nor of enzyme essentiality upon ablation (**Figure 2D**, x-axis) has provided a satisfying readout of the co-evolution of metabolism and enzymes, as both approaches emphasize the roles of just a small subset of fold lineages. A model that explicitly accounts for the co-evolution of metabolism and fold lineages, as well as the interplay between specialist and versatile fold lineages, is required. It is unclear, however, if a reachability analysis in the spirit of standard network expansion can be made suitable for this task: Is a stepwise trajectory constrained by fold/metabolism co-evolution able to meaningfully inform the emergence order of individual fold lineages? Or will the complex web of relationships between metabolites, reactions, cofactors, and enzymes preclude efforts to constrain co-evolutionary histories in the absence of external information such as phylogenetics?

Enzyme-gated network expansion: A model for the joint evolution of metabolism and enzymes

We propose an algorithm to reconstruct the relative order of fold lineage emergence within the context of a growing metabolic network using an extended network expansion approach. We refer to this algorithm as *enzyme-gated network expansion* (**Figure 3A**) to emphasize the constraining role played by fold lineages.

We model the evolution of reactions and fold lineages as a stepwise process with discrete emergence events. Here, new reactions can only be discovered if the requisite set of fold lineages, reactant compounds, and cofactors are available. However, unlike metabolites and cofactors, fold lineages are not endogenously generated within the reaction network, but must be introduced externally. Thus, each simulation begins with a set of seed compounds and no fold lineages. At first, only enzyme-independent reactions are active, which are insufficient for significant metabolic expansion (yielding just 2% of the scope). To activate more reactions, fold lineages are added in an independent, sequential fashion. After each fold lineage addition, all reactions that can be catalyzed by that fold lineage — alone or in combination with already discovered fold lineages — are activated and network expansion proceeds until no new metabolites can be reached. By allowing as much network expansion as possible before the addition of the next fold lineage, we assume that the discovery of new fold lineages is infrequent compared to their functional diversification. The process is repeated until no new reactions can be activated.

The order of fold lineage addition is selected based on a combination of utility and versatility in the context of the growing network. For a fold lineage to be added at a given point in the trajectory, it must be able to catalyze at least one reaction given the current set of metabolites and fold lineages (utility requiring). Note that the newly catalyzed reaction(s) do not need to produce new metabolites, since (a) new reactions between previously discovered metabolites, (b) previously discovered reactions catalyzed by a different fold or set of folds, and (c) entirely new reactions are treated as equivalent when selecting a fold lineage. If multiple fold lineages meet this criterion, the fold lineage that activates the most reactions is selected (versatility favoring). This second selection criterion stems from the intuition that domain emergence can be supported through interactions with a cognate ligand (Alva et al. 2015; Yagi and Tagami 2024). Thus, the emergence of a fold lineage that can interact with a large pool of cognate ligands (versatile) may be more likely than the emergence of a specialist fold lineage with only a small number of cognate ligands. In the event of a tie, selection is random among the most versatile fold lineages. Consequently, we describe an envelope of possible trajectories given these simple selection criteria.

Fold lineage emergence order is well constrained by enzyme-gated network expansion and comports with phyletic distribution

Using the enzyme-gated network expansion approach, we attempted to reconstruct the co-evolutionary history of fold lineages and metabolism. However, when starting from the base seed set, 93 sequential fold lineage additions yielded only 455 compounds before the simulation terminated, indicating that one or several essential reactions could not be activated. An analysis of fold lineage-reaction associations revealed that nearly 50% of fold lineages cannot operate without a partner fold lineage. As a result, the growing network reaches a point where any single fold lineage addition does not lead to the discovery of a new reaction with this minimal algorithm. To overcome this challenge, reactions that can be catalyzed by two or more fold lineages are considered for activation by the simultaneous addition of the smallest possible set of fold lineages (usually pairs). As with single fold lineages, versatility is favored and fold set selection is random if multiple fold sets are possible. With this adjustment, the full set of metabolites in our metabolic model is achieved (**Figure 3B**). A pseudocode implementation of the enzyme-gated network expansion algorithm is presented in **Figure S1**.

The majority of fold lineages (79%, averaged over 1,000 simulation runs) — including both versatile and specialist fold lineages — are required for expansion to an O_2-containing metabolic network with enzyme gating (**Figure 3B**). Plotting the results of 1,000 independent runs reveals a surprisingly narrow distribution of metabolic trajectories (**Figure 3B**, useful/versatile). These trajectories are highly similar to that of the standard network expansion (**Figures S4** and **S5**; average Spearman's rank correlation coefficient of 0.85), and enzyme gating does not significantly alter the patterns of pathway emergence (**Figure S6A**, **Table S9**) or the order of carbon fixation pathway discovery (**Figure S6B, Table S10**) relative to standard network expansion.

Does enzyme-gated network expansion meaningfully constrain the relative emergence order of fold lineages? Several pieces of evidence suggest that it does: First, the standard deviation of the addition step is small (less than 20 fold lineage addition steps) for 96% of fold lineages (**Figure S7**). Second, fold lineages added in a random order — with no requirement for activating a reaction and no multi-fold lineage additions — can serve as a point of comparison to the useful/versatility selection criteria (**Figure 3B**). Whereas random fold addition orders yield significant growth only at the end of the trajectory, application of the useful/versatility selection criteria maintains modest growth of the network throughout the trajectory — though bursts in expansion are still apparent. Finally, the fold lineage sampling along the trajectory is highly correlated with the phyletic distribution of fold lineages in prokaryotes (**Figure 3C**). These results suggest that the predicted order of fold lineage addition by enzyme-gated network expansion meaningfully reflects aspects of the relative ordering of their emergence as enzymes.

At what point in the history of metabolism did enzymes emerge?

The earliest metabolic reactions were likely catalyzed by minerals (Muchowska et al. 2019) and cofactors (Goldman and Kacar 2021) prior to the emergence of enzymes. At what level of metabolic complexity did fold lineage emergence begin, and can our approach reveal signatures of this transition? To explore these questions, ten sets of seed compounds, ranging from 2% to 100% of metabolites in our model, were analyzed using enzyme-gated network expansion (**Figures 3D** and **S8**). The seed sets of intermediate size correspond to the inventory of metabolites available at the point of discovery of eight cofactors in the standard network expansion model: pyridoxal phosphate (PLP, 6% of metabolites), adenosine triphosphate (ATP, 11%), nicotinamide adenine dinucleotide (NADH, 14%), S-adenosyl methionine (SAM, 28%), coenzyme A (CoA, 30%), flavin adenosine dinucleotide (FAD, 32%), thiamine diphosphate (ThDP, 36%), and cobalamin (93%).

To our surprise, a principal component analysis of the fold lineage orders (1,000 runs per seed set) revealed just three major clusters: a pre-nucleotide cofactor seed set, a post-nucleotide seed set (which breaks into two sub-clusters), and a post-O_2 seed set. All seed sets constrain the relative order of fold lineage emergence (**Figure S8**) and have strong rank correlation between phyletic distribution and addition order (Spearman's rank correlation, $p < 1.3 \times 10^{-76}$; **Figures 3C** and **S9**). However, with increasingly larger seed sets, fold lineage addition order more strongly reflects the number of reactions associated with a fold lineage (versatility-emphasizing), suppressing the role of specialist lineages in catalyzing essential reactions (**Figure S10**).

Various lines of evidence highlight the post-nucleotide seed set as the "natural" position for the enzymatic takeover of metabolism. Foremost, coded protein synthesis depends on nucleotides, meaning that pre-nucleotide synthesis of contemporary domains is unlikely. The observation that pre-nucleotide seed sets result in (a) long intervals of single fold lineage additions without significant metabolic expansion and (b) the requirement for concurrent addition of pairs or larger groups of fold lineages to discover new metabolic reactions early in the trajectory is consistent with a too-early initiation of enzyme gating (**Figure S11**). As enzymes likely predate the biological production of molecular oxygen (Sousa et al. 2016), we conclude that the nucleotide-containing seed set is the most reasonable reflection of early enzyme history. Given the strong similarity between nucleotide-containing seed sets (**Figures 3D** and **S11**), we focus on the ATP-containing seed set going forward.

The first metabolic fold lineages: TIM Barrel and the Rossmannoids

A joint history of metabolite and fold lineage discovery based on the ATP seed set is presented in **Figure 4A** (see **Tables S11** and **S12** for the discovery order of metabolites and fold lineages, respectively). 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 3B** and **S12**). It does, however, assign an emergence order to the earliest fold 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 fold lineages (**Figure 4A**) — which appear in the same order across all 1,000 independent ATP seed set runs, and are similarly positioned in other seed sets (**Figure S13, 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 emergence of HAD and ribonuclease H-like, which ablation studies classified as versatile but non-essential (**Figure 2D**), likely relates to the large repertoire of simple reactions that these fold lineages catalyze. That 4/6 of the earliest fold 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 patterns of protein structure similarity (Choi and Kim 2006). Moreover, TIM barrels, 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). We have previously hypothesized that the early dominance of the α/β structure class likely relates to facile emergence and functional potential under the constraints of a limited amino acid alphabet and short peptide length (Romero Romero et al. 2018; Longo et al. 2020; Longo et al. 2020; Vyas et al. 2021). However, we observe a dilution of α/β fold lineages as the trajectory proceeds (**Figure 4B**) — with α and α+β fold lineages being preferred later in the trajectory. This result may suggest that α/β fold lineages are not fundamentally more evolvable or functionally capable under present conditions, or that other structure classes may be uniquely suited to perform emergent metabolic roles.

Although β-proteins play a more limited role in metabolism (**Figures 2B**, **2C,** and **4B**), remarkably, the first β-protein predicted to have emerged by enzyme-gated network expansion is the cradle-loop barrel. This fold lineage plays essential roles in information processing and encompasses several small, highly traversable β-barrel folds (Yagi and Tagami 2024). Small β-

folds have also been suggested to play a founder role, particularly among ribosomal proteins (Petrov et al. 2015; Kovacs et al. 2017). The fact that a ribosome-centric view does not highlight α/β as being ancient may reflect different emergence filters for ribosome-associated and metabolism-associated folds — for example, nucleic acid binding versus independent foldability and nucleotide binding, respectively.

Bursts of metabolic innovation recruit fold lineages young and old

When new metabolites arise in the biosphere, they can drive the emergence of new fold lineages or be acted upon by pre-existing fold lineages performing a new reaction (*reuse*). The history of fold lineage reuse upon the emergence of new metabolites for a representative ATP seed set run is presented in **Figure 5A**. While approximately half of fold lineages are used when they are introduced and never again, 54% (216/396) of lineages are reused at multiple points along the trajectory (**Figure S15**). Some lineages, particularly the early-emerging, versatile fold lineages, exhibit near constant reuse along the trajectory. The Rossmann lineage, for example, is reused after ~30% of fold lineage additions, likely due to its close association with various cofactors and diverse repertoire of cooperating domains. Other lineages lie dormant for long periods before finding a new use; X-group 211, for example, is the 99th fold lineage added and performs carbon-sulfur lyase chemistry. This fold lineage is then later reused to catalyze a reaction involving O_2 upon addition of the 296th fold lineage. About 70% (152/216) of the fold lineages with reuse catalyze a reaction with a new 2-digit EC (Enzyme Commission) number sometime later, usually within 30 fold lineage additions of their emergence (**Figure S16**). Peaks in fold lineage reuse (**Figure 5A**, upper) correspond to significant events of metabolic growth, and late peaks increase in size as more established fold lineages are available to draw from. The number of required fold lineages per discovered reaction modestly increases over the expansion (**Figure S17**) as key lineages accrete cooperating domains (see **Figure S18** for an analysis of the TIM barrel).

Reuse outpaces emergence: The case of molecular oxygen

To quantify the impact of metabolite discovery on the protein universe, each fold lineage was grouped into one of five classifications with respect to a given metabolite (see **Supplementary Methods** and **Figure S19**): Fold lineages whose first reactions all require the metabolite (*emerge to use*, ETU); fold lineages for which some of their first reactions require the metabolite (*first use*, FU); fold lineages that catalyze reactions involving the metabolite sometime after their emergence (*learn to use*, LTU); fold lineages that catalyze reactions that are made possible by the metabolite, but do not catalyze reactions with the metabolite directly (*indirect use*, IU); and fold lineages that

are not associated with the metabolite in any way (*never use*, NU). In cases where multiple classifications apply, the order of precedence was ETU/FU>LTU>IU.

We find that molecular oxygen is one of the strongest drivers of both fold lineage emergence and reuse (**Figure 5B**). The emergence of oxygenic photosynthesis occurred approximately 3 billion years ago (Crowe et al. 2013) and added a layer of O_2-associated reactions around a largely anaerobic core, a metabolic organization referred to as hierarchical modularity (Raymond and Segrè 2006) (**Figure 2D**). The modest rewiring of interior metabolic layers that resulted from the advent of O_2, however, stands in stark contrast to its consequence within the protein universe, where many pre-existing fold lineages acquired new functions. Our model predicts that oxygen utilization was a primal function for 8% of metabolic fold lineages (ETU or FU; **Figure 5C**, **Table S14**) whereas almost two times more fold lineages (14%) acquired the ability to use oxygen later in the trajectory, sometime after their emergence as enzymes (LTU). 65% of versatile fold lineages (having ≥100 associated reactions) fall into the LTU category (**Figure S20**). We note, however, that while the consequences of biological O_2 production had sweeping impacts on established fold lineage functions, it was only possible due to the existence of several specialist fold lineages, such as those responsible for synthesizing chlorophyll.

To further test the predicted fold lineage-oxygen emergence relationships, we analyzed the domain composition of aerobic and anaerobic bacteria (Madin et al. 2020; Goldford et al. 2023) (**Table S6**). We find that category-wise fold lineage enrichments (**Figure 5D**) are fully consistent with expectation: ETU+FU and IU fold lineages are enriched in aerobes. NU fold lineages, on the other hand, are more enriched in anaerobes. Finally, LTU fold lineages, which presumably emerged before oxygen and only later catalyzed an oxygen-associated reaction, are equally common in both anaerobes and aerobes. Thus, signatures of fold lineage adaptation to oxygen predicted by enzyme-gated network expansion are manifested as distinct distribution patterns across bacteria with different physiologies, suggesting that our approach can approximate the relative order of enzyme functions.

Limitations and outlook for models of metabolic evolution with strict causality

The present model enforces strict enzyme-reaction and cofactor-reaction dependencies and does not consider the kinetic constraints that must be satisfied for completing metabolic pathways. These decisions simplify implementation and interpretability but can introduce errors in cases where the essential source data (*e.g.*, biochemical reactions) are missing or incorrect, as was the case for non-autocatalytic purine synthesis (Goldford et al. 2024). The strengths and limitations of the network expansion algorithm have been discussed in detail (Goldford and Segrè 2018), and apply to this approach as well. In addition, we note that this approach reflects the

emergence of a fold lineage as a metabolic enzyme, and does not account for other types of functional roles that may have emerged earlier. Future work can begin to integrate alternative historical records, such as by incorporating homology relationships *between* fold lineages (Farías-Rico et al. 2014; Longo et al. 2020; Kolodny et al. 2021), or by developing frameworks with relaxed causality through careful coarse graining.

Concluding Remarks

We introduce a new approach to interpret the record of enzyme evolution embedded within the graph structure and reaction–fold lineage associations of biosphere-scale metabolism. Remarkably, we recover a sequential ordering of fold lineage emergence in which each new fold lineage was largely sufficient on its own to drive metabolic expansion. This record of protein history is independent from (though found to be correlated with) phyletic distribution, and can provide evidence for the emergence order of ancient fold lineages that reflects both fold lineage utility and essentiality. The first metabolic fold lineages were likely a group of related α/β lineages, but the dominance of the α/β class was diluted over time. Differences between metabolic-centering and ribosome-centering perspectives — particularly with respect to the emergence order of protein classes — may provide clues for multiple fold origins. In addition, we provide a snapshot of how the protein universe responds to metabolic upheaval, with preference for reuse of pre-existing fold lineages over the emergence of entirely new fold lineages. This work adds a new archive to the goal of synthesizing diverse historical records to reconstruct the earliest stages of protein evolution.

Acknowledgements

Shawn E. McGlynn helped manually annotate KEGG orthology group interdependence. This study was supported by the Human Frontier Science Program Grant Number RGEC29/2025 (LML) with the award DOI https://doi.org/10.52044/HFSP.RGEC292025.pc.gr.230124 and the National Aeronautics and Space Administration (NASA) award number 80NSSC25K7873 (LML) and 80NSSC23K1357 (JG).

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

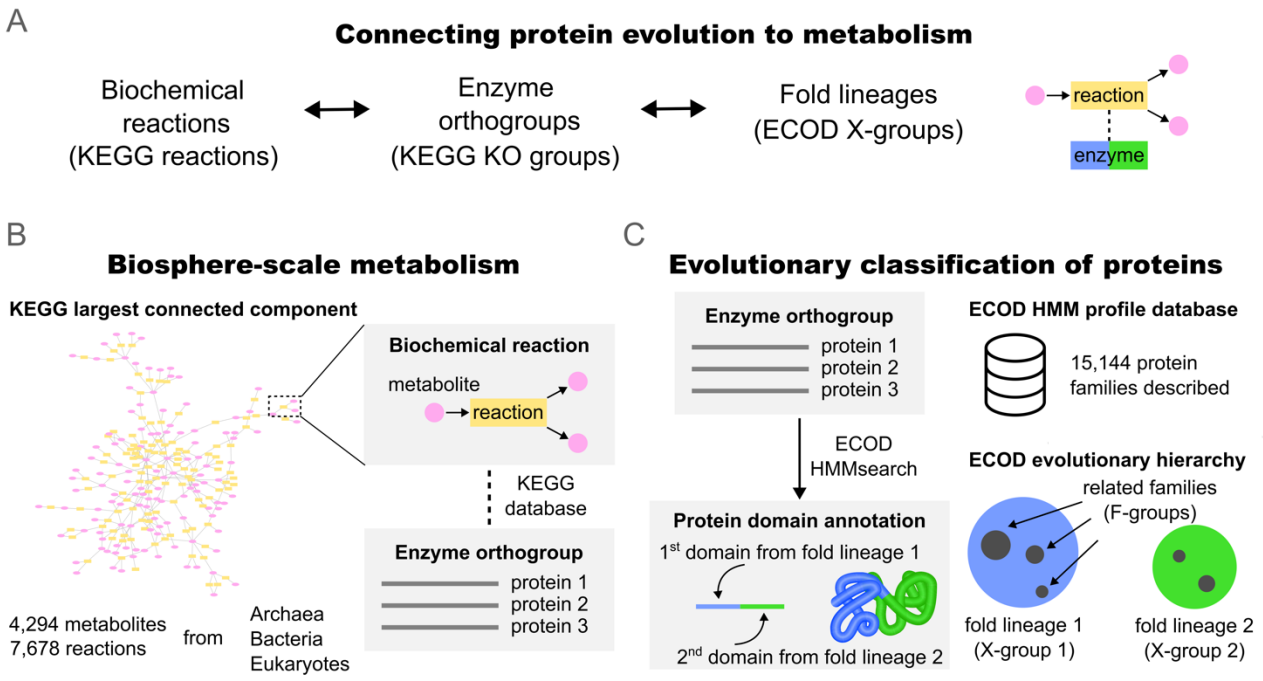

Figure 1. Connecting protein evolution to metabolism. **A**. To simulate the coevolution of metabolism and proteins, biochemical reactions (KEGG reactions) and fold lineages (ECOD X-groups) were connected through mutual association with enzyme orthogroups (KEGG KO groups). **B**. Our metabolic model contains metabolites and reactions encompassing organisms from all branches of the tree of life. Reactions within this network were linked to enzyme orthogroups (KO group) using the manually curated associations within the KEGG database. **C**. For each enzyme orthogroup in KEGG, the evolutionary classifications of the protein domains were assigned based on the ECOD hierarchy. Protein domains can be classified into protein families (F-groups) that are associated with an HMM profile. Protein families can be grouped into fold lineages (X-groups; similar to "clades" in Pfam) that are hypothesized to have emerged from a common founder protein. See **Methods** for detailed explanation of each step.

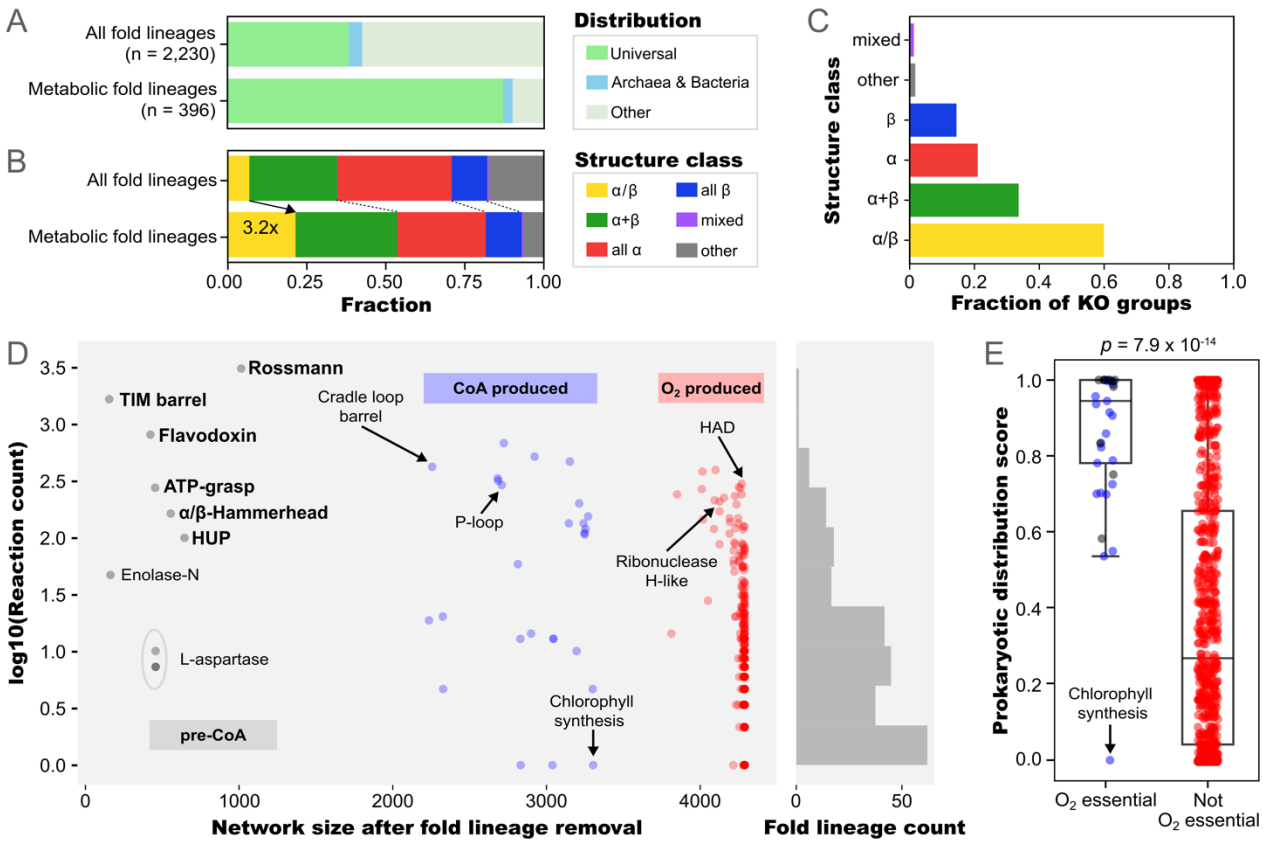

Figure 2. The metabolic fold lineages. **A**. Distribution of all fold lineages in ECOD and the subset of metabolic fold lineages. Universal refers to fold lineages present in all three domains of life: archaea, bacteria, and eukaryotes. **B**. Structure class of all fold lineages in ECOD and the subset of metabolic fold lineages. **C**. The fraction of enzyme orthologous groups associated with metabolic reactions in our model that contain fold lineages from each structure class. **D**. The number of associated reactions (y-axis) and the network size from standard network expansion after single fold lineage ablation (x-axis) for each metabolic fold lineage. CoA- (blue) and O_2-containing (red) networks are indicated. **E**. Prokaryotic distribution scores (see **Methods** for definition) of fold lineages that are essential for reaching O_2 versus those that are not, based on the single fold lineage ablation analysis in panel C.

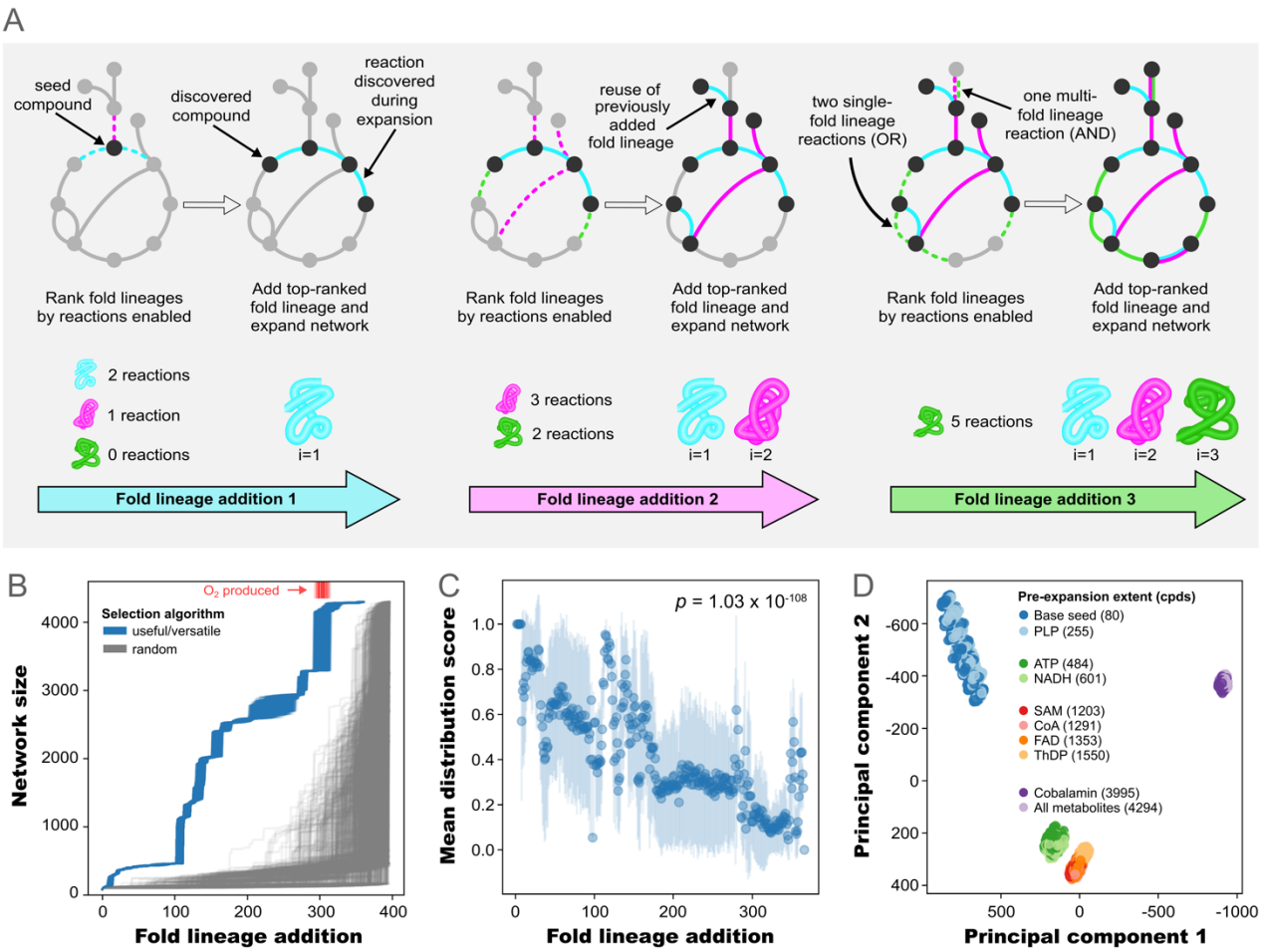

Figure 3. Enzyme-gated network expansion. **A**. The enzyme-gated network expansion algorithm. See main text and **Methods** for additional details. **B**. Metabolic growth trajectories (network size) when selecting the most useful/versatile fold lineage (blue) or a random fold lineage (grey) for 1,000 runs each. The red rug plot at the top indicates the iteration where molecular oxygen is discovered in the useful/versatile runs. **C**. Average prokaryote distribution score of fold lineages across 1,000 runs as a function of fold lineage addition order. Error bars are standard deviations. Indicated p-value calculated from the Spearman's rank correlation. Fold lineage additions for panels B and C are based on the base seed set of compounds. **D**. Principal component analysis of the fold lineage addition orders from all trajectories across all ten seed sets (1,000 independent runs per seed set, 10,000 runs in total). The compounds for each seed set are listed in **Table S5**. Panels B and C for each seed set are provided in **Figures S8** and **S9**.

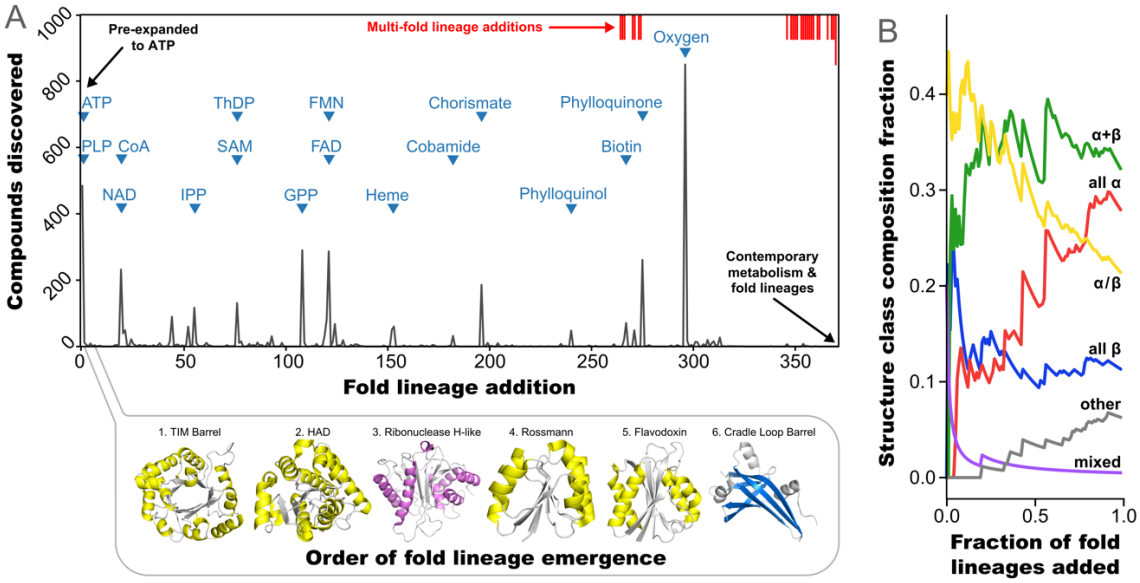

Figure 4. A coevolutionary history of fold lineages, metabolic reactions, and metabolites. **A**. Fastest growing trajectory (cumulative plot with the greatest integrated area) out of 1,000 runs for the ATP seed set. Select metabolites and the first 6 fold lineages, colored according to structure class, are indicated. The position of metabolite labels on the y-axis is arbitrary. The red rug plot along the top of the figure denotes fold lineage addition steps where two (short line) or three (long line) fold lineages had to be added simultaneously for a new reaction to be discovered. The average orders of emergence of all fold 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 S14**. The "mixed" structure class refers to fold lineages with both α/β and α+β elements.

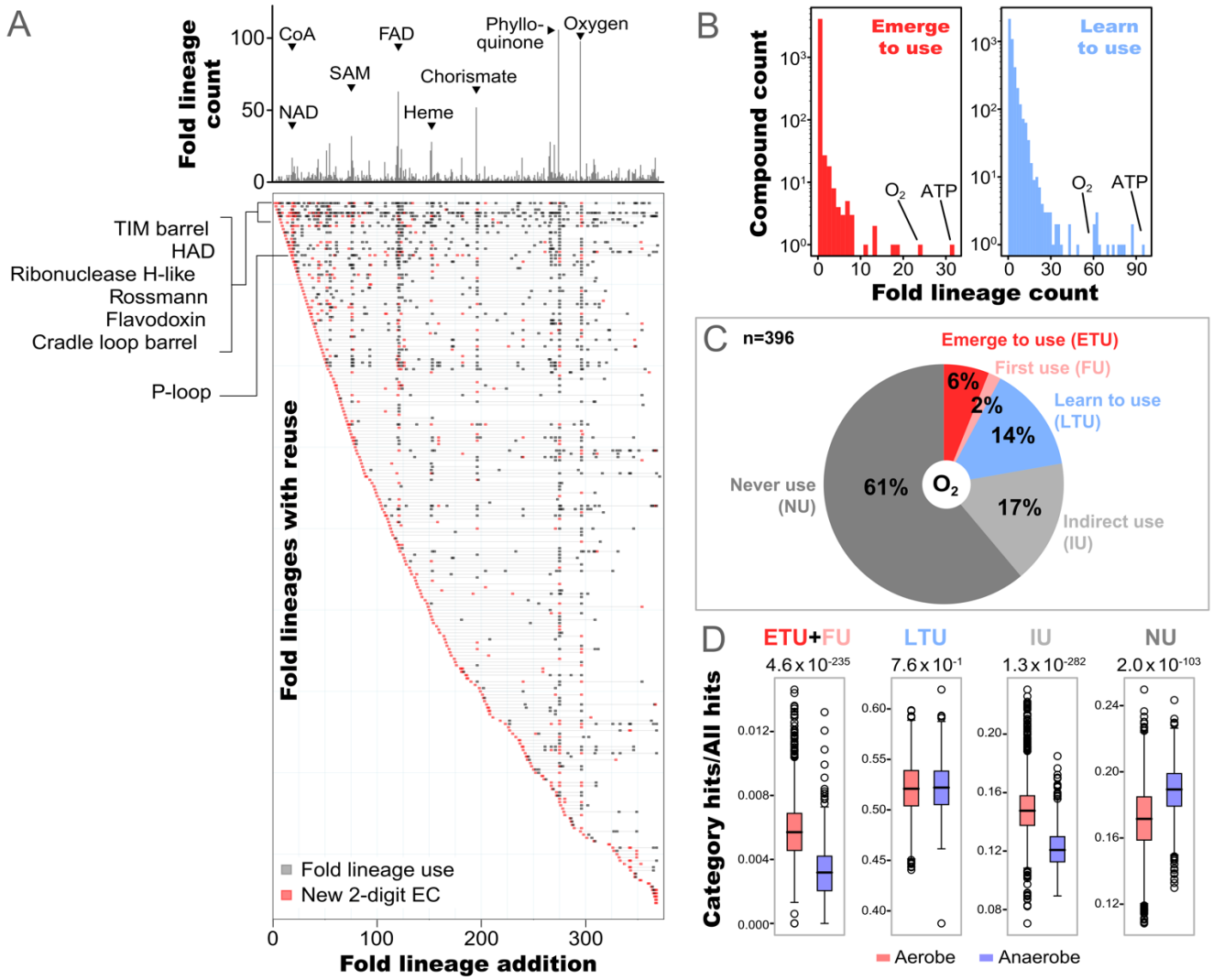

Figure 5. Emergence and reuse of fold lineages. **A**. A global view of fold lineage reuse calculated from the trajectory in **Figure 4A**. Total fold lineage use counts at various points in the expansion (upper) and the patterns of fold lineage emergence and reuse along the trajectory (lower). Fold lineages are shown one per row along the y-axis, while the x-axis indicates points along the trajectory where a fold lineage is associated with a new reaction. The fold lineages have been ordered from top to bottom by their order of emergence. Only fold lineages that catalyze a reaction at 2 or more points in the trajectory are shown. Points where a fold lineage catalyzes a reaction with a new 2-digit EC number are indicated in red. **B**. Histogram of the number of fold lineages that emerge to use (left) or learn to use (right) a given compound. ATP and molecular oxygen are both strong drivers of fold lineage emergence and reuse. **C**. Breakdown of fold 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 fold lineages satisfy this threshold. A full list of metabolic fold lineages and their category is available in **Table S14**. **D**. Enrichment of fold

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 fold lineages within a given usage category was divided by the occurrence of all ECOD fold lineages. Indicated p-values were calculated by Mann-Whitney U tests.

References

1. E. R. R. Moody, *et al.*, The nature of the last universal common ancestor and its impact on the early Earth system. *Nat Ecol Evol* **8**, 1654–1666 (2024).

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

3. H. Edwards, S. Abeln, C. M. Deane, Exploring fold space preferences of new-born and ancient protein superfamilies. *PLoS Comput Biol* **9**, e1003325 (2013).

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

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

6. M. E. Rebeaud, S. Mallik, P. Goloubinoff, D. S. Tawfik, On the evolution of chaperones and cochaperones and the expansion of proteomes across the Tree of Life. *Proc Natl Acad Sci U S A* **118** (2021).

7. K. W. Plaxco, K. T. Simons, D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* **277**, 985–994 (1998).

8. K. Takács, B. Varga, V. Farkas, A. Perczel, V. Grolmusz, Opening Amyloid-Windows to the secondary structure of proteins: The amyloidogenecity increases tenfold inside beta-sheets. *Comput Biol Med* **179**, 108863 (2024).

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

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

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

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

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

14. J. S. Richardson, *The Anatomy and Taxonomy of Protein Structure* (1981).

15. L. Lo Conte, *et al.*, SCOP: a structural classification of proteins database. *Nucleic Acids Res* **28**, 257–259 (2000).

16. M. Knudsen, C. Wiuf, The CATH database. *Hum Genomics* **4**, 207–212 (2010).

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

18. M. Sorokina, M. Stam, C. Médigue, O. Lespinet, D. Vallenet, Profiling the orphan enzymes. *Biol Direct* **9**, 10 (2014).

19. N. Hadadi, H. MohammadiPeyhani, L. Miskovic, M. Seijo, V. Hatzimanikatis, Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. *Proc Natl Acad Sci U S A* **116**, 7298–7307 (2019).

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

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

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

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

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

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

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

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

28. S. D. Copley, E. Smith, H. J. Morowitz, A mechanism for the association of amino acids with their codons and the origin of the genetic code. *Proc Natl Acad Sci U S A* **102**, 4442–4447 (2005).

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

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

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

32. V. G. Giacobelli, *et al.*, Ancient amino acid sets enable stable protein folds. *bioRxiv* (2025).

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

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

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

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

37. J. E. Goldford, R. Murali, J. S. Valentine, W. W. Fischer, Metabolic evolution of pyranopterin-dependent biochemistry. *bioRxiv* (2023).

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

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

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

41. G. Yanai, G. Axel, L. M. Longo, N. Ben-Tal, R. Kolodny, Contrastive learning unites sequence and structure in a global representation of protein space. *bioRxiv* (2025).

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

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

44. A. D. Goldman, B. Kacar, Cofactors are Remnants of Life's Origin and Early Evolution. *J Mol Evol* **89**, 127–133 (2021).

45. F. L. Sousa, S. Nelson-Sathi, W. F. Martin, One step beyond a ribosome: The ancient anaerobic core. *Biochim Biophys Acta* **1857**, 1027–1038 (2016).

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

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

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

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

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

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

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

53. M. L. Romero Romero, *et al.*, Simple yet functional phosphate-loop proteins. *Proc Natl*

Acad Sci U S A **115**, E11943–E11950 (2018).

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

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

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

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