

UC Berkeley

UC Berkeley Previously Published Works

Title

Diversity-driven biochemical survey reveals widespread dimerization throughout the rubisco superfamily

Permalink

<https://escholarship.org/uc/item/16j877jr>

Journal

Nature Communications, 17(1)

ISSN

2041-1723

Authors

Kehl, Alexander J
Taylor-Kearney, Leah
Jaffe, Alexander L
[et al.](#)

Publication Date

2026-06-03

DOI

10.1038/s41467-026-73982-5

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Diversity-driven biochemical survey reveals widespread dimerization throughout the rubisco superfamily

Received: 18 November 2025

Accepted: 22 May 2026

Published online: 03 June 2026

 Check for updates

Alexander J. Kehl^{1,2,3,19}, Leah Taylor-Kearney^{2,3,4,19}, Alexander L. Jaffe⁵, Jose Henrique Pereira^{3,6}, Jennifer Lee⁷, Michal Hammel⁶, Lucas M. Waldburger^{2,3,8,9,10}, Caroline Yeow⁷, Luis Valentin-Alvarado¹¹, Paul D. Adams^{6,12}, Jillian F. Banfield^{13,14,15}, Justin B. Siegel^{7,16,17}, Noam Prywes¹⁸  & Patrick M. Shih^{2,3,4,13} 

Rubisco is the entry point of nearly all organic carbon into the biosphere and is present in all domains of life. Despite its global importance, biochemical studies of this enzyme superfamily have been limited to a relatively narrow set of subclades. Recent advances in metagenomics have dramatically reshaped our understanding of both microbial and rubisco diversity; however, biochemical characterization of these sequences has not kept pace with the exponential growth in sequence data. To better survey the functional and structural diversity of rubisco, we systematically sample and synthesize a library of diverse rubisco sequences with an emphasis on clades that are sparsely represented in the biochemical literature. Our updated phylogenetic analysis reveals that many deep-branching rubiscos assemble as dimers, supporting a dimeric origin for the superfamily – in contrast to the ecologically dominant hexadecameric form I. Additionally, we discover and structurally characterize an unusually large catalytic subunit among characterized rubiscos, originating from a early-branching subclade with secondary structural elements not present in canonical rubisco architectures.

Ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) is the primary carboxylating enzyme on earth^{1,2}, facilitating the conversion of atmospheric carbon dioxide (CO₂) into the building blocks of life in autotrophic organisms like plants, algae, and bacteria³. Rubisco evolved at least three billion years ago, and today is found in eukaryotes, bacteria, and archaea⁴. All rubisco sequences form a single clade that is distinct from rubisco-like proteins (RLPs); however, both rubiscos and RLPs share structural folds and a common evolutionary origin^{5,6}. It is thought that rubisco carboxylase activity emerged from a different pre-existing catalytic function in an RLP². Throughout its evolution, rubisco has been adapted to function in all three domains of life and in multiple metabolic pathways, only some of which involve contribute meaningfully to net carbon fixation⁷. Nonetheless, the vast majority of studies have focused on characterizing rubiscos from

clades that are involved in autotrophy (i.e., form I and form II); yet, this represents only a fraction of the full diversity of rubiscos found across the entire tree of life. Thus, our understanding of rubisco is largely biased and there remains large gaps in our understanding of rubisco diversity as a whole.

Beyond sequence diversity, rubisco has also emerged as a structurally diverse enzyme family that can adopt a wide range of homo- and hetero-oligomeric states. All rubiscos share a common fold with a distinct N-terminal domain followed by a TIM barrel and a short C-terminal region¹. The fundamental functional unit of rubisco is composed of an antiparallel dimer of two large subunits (RbcL, ~50 kDa) with a C₂ symmetry axis and two active sites located at the intradimeric interface. Across the tree of rubiscos, a variety of higher-order assemblages of dimers have been found (L₂)_n, including L₂, L₄,

A full list of affiliations appears at the end of the paper.  e-mail: np648@cam.ac.uk; pmsih@berkeley.edu

L₆, L₈ and L₁₀ as well as L₈S₈ hetero-oligomeric structures containing the rubisco small subunit^{8–10}. Given this structural diversity, the prevalence and phylogenetic distribution of rubisco dimers remain uncertain, largely due to incomplete coverage of quaternary structures across the phylogeny.

Given their roles in photosynthetic organisms, form I and form II rubiscos have been the most extensively sampled. Recent studies have thoroughly examined the biochemical diversity across sequences in these two subclades revealing a wide range of catalytic rates in rubiscos involved in the Calvin-Benson-Bassham (CBB) cycle^{11,12}. Moreover, such efforts have revealed that the fastest rubisco enzymes are among the form II clade. Notably, form I and II rubiscos represent only 10.38% of non-redundant rubisco sequence diversity (Table 1). Moreover, 56.8% of sequences are found in regions of the rubisco tree that lack formal clade designation, representing 62.1% of phylogenetic diversity (Table 1). Ultimately, improving our understanding of rubiscos in sparsely covered regions of the phylogeny may provide distinct insights into the evolution of the broader enzyme family, highlighting the need to revisit the classification and naming scheme of under-represented rubisco subclades.

To fill the knowledge gaps in deep-branching, non-CBB associated rubisco sequences across the full phylogeny, we heterologously expressed, systematically synthesized, purified, and carried out biochemical studies on rubiscos from sparsely covered regions and undersampled clades. Our findings clarify the structural origins of the broader rubisco family and reveal secondary structural elements that expand the known structural repertoire of the superfamily within a early-branching clade of exceptionally large rubiscos.

Results

Generation of a comprehensive rubisco diversity library

We compiled a comprehensive set of rubisco superfamily protein sequences by querying large databases—specifically NCBI's non-redundant (nr) and env_nr datasets. Protein sequences were clustered at increasing identity thresholds to account for the high degree of redundancy, particularly within the form I clade. At a 65% identity threshold, we recovered 1859 unique sequence clusters, representing a 14.6% increase over our last effort¹. Representative sequences were selected from all clusters, regardless of form, and subjected to maximum likelihood phylogenetic reconstruction (Materials and Methods). Across our phylogeny, we estimate that only 8.5% of sequence clusters at 65% identity excluding RLPs have been biochemically characterized for kinetic properties. We utilized our phylogeny to

guide the selection of 104 phylogenetically diverse rubisco sequences for biochemical and structural characterization, focusing largely on clades with no previously characterized representatives and/or with evolutionary significance. All 104 genes were synthesized and cloned into an *E. coli* expression plasmid that has previously been used to heterologously express and purify a wide range of rubiscos^{8,12}.

Updated phylogeny resolves form I substructure and uncovers deep-branching clade

Our tree topology of the rubisco superfamily depicts bona fide (i.e., referred to as forms I through III) rubiscos forming a monophyletic clade apart from RLP (Fig. 1a), consistent with previous phylogenetic analyses^{1,13,14}. Similarly, we observed that the forms I and II—the only two rubisco forms currently experimentally associated with CBB Cycle—are separated by several short but well-supported internal branches, and do not resolve as sibling clades (Fig. 1a). Within the form I rubiscos, numerous subforms with different taxonomic, structural, and kinetic attributes are currently recognized¹. Here, we retain the nomenclature for the previously recognized IA, IB, IC, ID and IE subforms (Fig. 1b). However, the phylogenetic relationships between these clades diverge somewhat from previous reconstructions—specifically, we now use form IE to describe a broader, paraphyletic swath of sequences primarily associated with the phylum *Actinobacteria*, overlapping with sequence sets identified in other previous studies^{15,16}. On the other hand, we formally classify the merged, monophyletic clade form IF. Finally, we formally classify the clade Iβ.

The revised phylogenetic placement of rubisco sequences provides insight into the evolutionary trajectory of early branching form I sequences. We also observed intriguing taxonomic patterns among sequences placed within our revised form I topology. Notably, rubisco sequences at the deep base of form I—including those from members of the candidate division WOR-3—derive from bacterial species rather than archaeal ones, distinguishing them from form IIIB; this contrasts with the placement reported by Zeng et al.¹⁷. These basal bacteria also fall outside the Chloroflexi-associated lineages that dominate later-diverging form I groups, diversifying the evolutionary context from which canonical form I emerges.

The majority of deep-branching rubiscos are dimeric

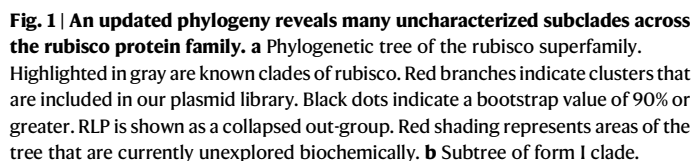
From our synthesized library of 104 diverse rubisco sequences, all were heterologously expressed and purified from *E. coli*, with only 18 that yielded soluble purified enzyme for downstream biochemical characterization and investigation of their oligomeric state. Prior to this study, only four rubiscos have ever been purified from archaea. Here, we purified and characterized seven additional archaeal rubiscos, expanding representation of rubiscos in a poorly sampled domain of life^{9,18–20}. Until recently, rubisco dimers were regarded as anomalous within the broader landscape of rubisco oligomerization, primarily associated with the form II clade⁸. However, the discovery of form Iα and I′ clades positioned dimeric rubisco basal to the form I lineage, suggesting a dimeric origin for the form I clade²¹. Despite these insights, the prevalence of dimeric rubisco across the full phylogenetic spectrum—particularly among deep-branching clades—remained poorly understood.

To address this gap, we characterized all 18 soluble rubisco variants in our library using Size-exclusion chromatography-small-angle X-ray-scattering (SEC-SAXS). This technique provided bound and unbound scattering profiles for each variant, enabling precise oligomeric state determination (Supplementary Figs. 1, 2). Remarkably, 15 of the 18 rubiscos analyzed—originating from sparsely sampled regions of the phylogenetic—exhibited dimeric quaternary structures (Fig. 2, Supplementary Tables 1, 2). Two of the three variants that exhibited higher-order oligomerization underwent an L2-to-L10 shift upon CABP binding; both are members of the Form II–III clade and contain the RAD domain, which is thought to mediate this oligomeric transition⁹.

Table 1 | A large fraction of rubiscos remain uncharacterized

Form	Sequence diversity (% of total sequences)	Phylogenetic diversity (% of branch length across phylogeny)
Form I (<i>sensu stricto</i>)	6.88%	5.85%
Form II	3.50%	2.82%
Form II/III	2.75%	2.18%
Form IIIa	3.73%	3.61%
Form IIIb	17.5%	15.5%
Form IIIc	6.29%	5.87%
Form III-like	2.55%	2.07%
All others	56.8%	62.1%
Total	100%	100%

Sequence and phylogenetic diversity are measured from a dereplicated phylogeny (1859 clusters) described in Fig. 1 generated from all rubisco sequences in NCBI (~87k). Sequence diversity is measured as the percentage of sequences of each subclade compared to the total number of sequences excluding RLPs (*n* = 435). Phylogenetic diversity is calculated as the sum of the length of all branches, and the percentage of branch length attributable to each subclade. Bold represents areas of the phylogeny that are currently unexplored.

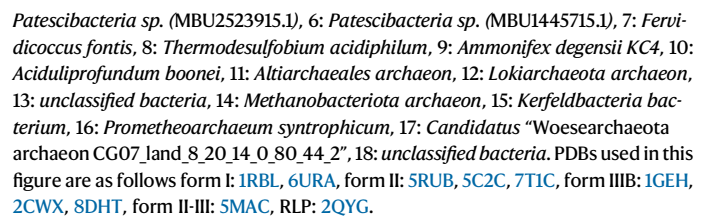


Highlighted in gray are known subclades of form I. A color bar around the tree indicates the kingdom that the rubisco was found in, purple: Bacteria, orange: Archaea, green: Eukaryota, gray: Unknown. The orange colour is not well visible (zoom in) because only 2 branches have an Archaea lineage host organisms are indicated over the form I tree but are not exhaustive. Black dots indicate a bootstrap value of >90%. Red shading newly defined clades.

Taken together, our survey of a 104-sequence library—yielding 18 soluble enzymes for SEC-SAXS analysis—shows that 15 variants from deep-branching, sparsely sampled clades adopt dimeric quaternary structures, supporting a dimeric ancestral state and a broader prevalence of dimers than reflected in the existing literature.

In prokaryotes, rubisco genes are often found in the genome co-locating with a variety of functionally related genes like rubisco activases and other chaperones, regulators, genes involved in the synthesis of carbon concentrating mechanisms, and those involved in the CBB Cycle^{22,23}. As an alternative approach to summarizing rubisco diversity and exploring its functional associations, we undertook a genomic context analysis of diverse rubiscos drawn from our non-redundant sequence set. For each rubisco locus with adequate genomic sequence context, we retrieved all protein sequences within 10

Most strikingly, rubisco loci across the tree of life are organized into distinct clusters that align with their phylogenetic form, even though they share some neighboring functional genes. When sampled at a higher identity threshold (95%), we observed numerous form I associated clusters from the IA-IB subclades. However, genomic context could not be reliably retrieved for sequences assigned to the I' clade. (Supplementary Fig. 3c). Form I loci are mostly, though not entirely, distinct from loci with form II rubiscos from bacteria. Loci with form II-III, IIIA, IIIB IIIC, and III-Like formed one large, dense, cluster alongside several smaller ones associated with form II rubiscos (Fig. 3c), suggesting that these rubiscos are likely functionally similar despite substantial phylogenetic divergence (Fig. 3c). Alternatively, divergence from more canonical form I and form II loci could drive clustering of diverse form II-III-Like rubiscos.



Previous studies have shown that reducing oligomeric state—by mutating dimer–dimer interface residues in higher-order assemblies (e.g., L10, L6)—can lead to decreased thermal stability, supporting a potential link between oligomerization and thermostability^{8,18,27}. However, our data from phylogenetically diverse rubiscos revealed no such correlation between oligomerization and thermostability (Supplementary Fig. 6a). Notably, we identified three dimeric rubiscos

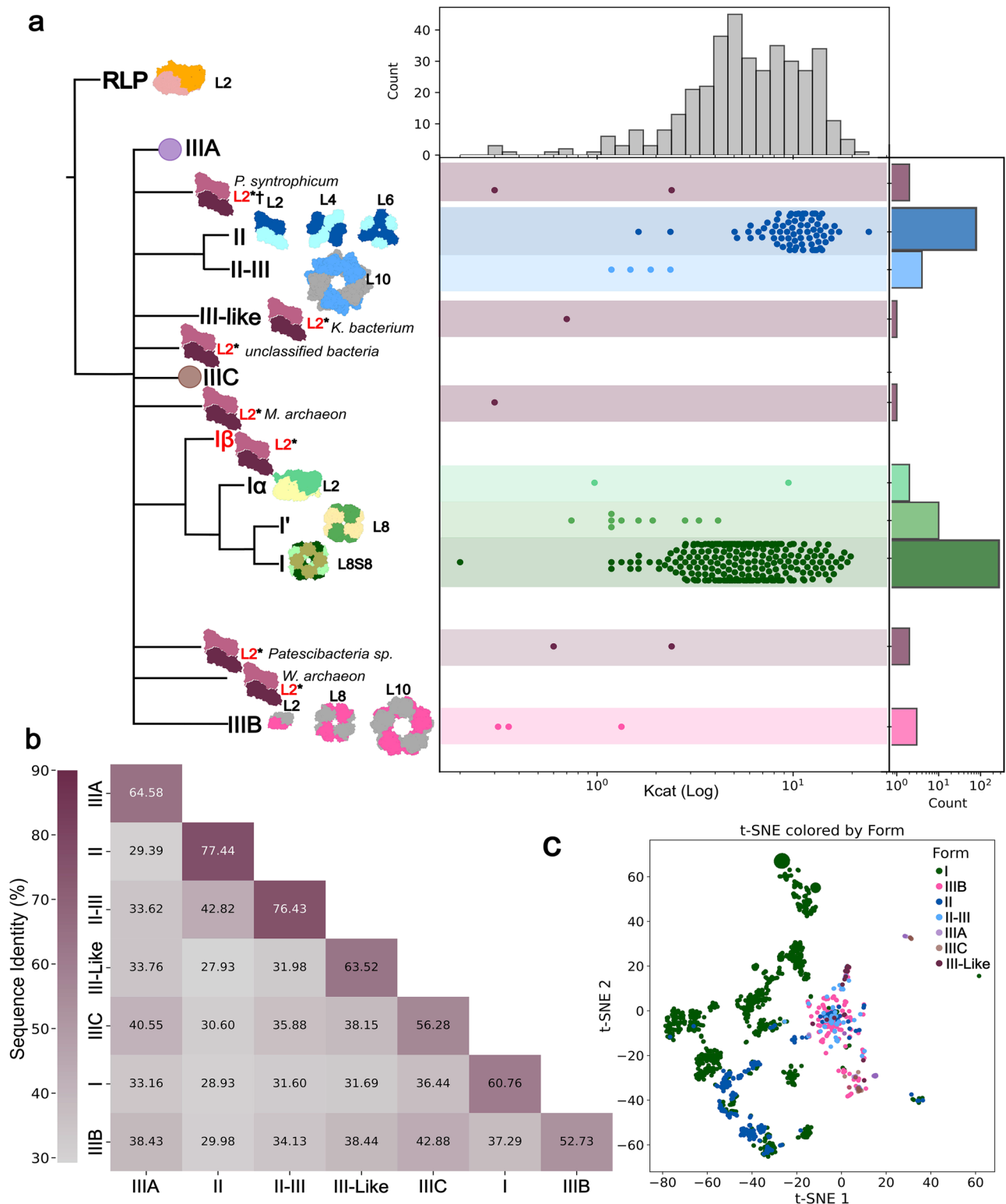


Fig. 3 | Biochemical, sequence similarity, and genomic context across the rubisco superfamily. a Oligomeric states characterized in this study are marked by a * and purple/pink dimer cartoon, † denotes crystal structure presented in this paper, red text denotes rubisco that are in previously unexplored regions of the tree. Clades or clusters with k_{cat} (Log) measurement in line with the tree. Top histogram shows the count and distribution of k_{cat} (Log) measurements by k_{cat} , the right histogram shows the count and distribution of k_{cat} (Log) measurements

by form. **b** Pairwise sequence identity heatmap. Sequence identity is shown for all forms of rubisco with numeric values and color **c**: t-distributed Stochastic Neighbor Embedding (t-SNE) of rubisco sequences colored by form. The following colors are used to depict known clades of rubisco: I: green, IIIB: pink, II: dark blue, II-III: light blue, IIIA: light purple, IIIC: brown, III-Like: dark purple. PDBs used in this figure are as follows form I: 1RBL, 6URA, form II: 5RUB, 5C2C, 7TIC, form IIIB: 1GEH, 2CWX, 8DHT, form II-III: 5MAC, RLP: 2QYG.

Table 2 | Thermal stability of phylogenetically diverse rubiscos

Accession	Organism	form	T _m °C	T _m SD
Candidatus_Lokiarchaeota_archaeon_strain_HM5_B55	Lokiarchaeota archaeon	unclassified	59.7	0.3
WP.147662334.1	Prometheoarchaeum syntrophicum	unclassified	61.6	0.4
MBU1445715.1	Patescibacteria sp. (MBU1445715.1)	unclassified	63.7	0.6
MBU2523915.1	Patescibacteria sp. (MBU2523915.1)	unclassified	57.5	0.2
HDR53286.1	Methanobacteriota archaeon	unclassified	78.8	0.1
MBI5465798.1	Kerfeldbacteria sp.	III-Like	68.7	0
WP.110780429.1	Rhodopseudomonas faecalis	II	57.6	0.2
WP_073240254.1	Rhodospirillum rubrum	II	71.8	0.1
WP.108308545.1	Thermosulfobium acidophilum	Iβ	77.7	0.4
PMB76193.1	Fervidicoccus fontis	Iβ	78.3	0.3
ACX52946.1	Ammonifex degensii	Iβ	95.2	0.2

within the form Iβ clade that exhibited high T_m values ranging from 77.5 °C to 95.2 °C, all derived from thermophilic organisms (Supplementary Fig. 6a). This observation echoes findings from Bundela et al., who reported a dimeric rubisco with the highest known T_m (114.5 °C), also from a thermophile¹⁸. In contrast, we observed a hexameric form II rubisco with a T_m of 57.6 °C—comparable to the lowest recorded values (Supplementary Fig. 6a). Taken together, these results suggest that oligomeric state alone does not correlate with thermostability, prompting consideration of alternative factors—such as sequence composition, surface chemistry, or environmental adaptation—that may underlie thermal resilience in rubisco.

To contextualize rubisco's catalytic constraints, we compared turnover rates across variants characterized in this study, integrating our measurements with recent efforts to expand kinetic profiling beyond historically overrepresented clades (Fig. 3a). Previous studies have highlighted potential sampling bias toward slower, predominantly eukaryotic rubiscos, suggesting that true median activity may be underestimated^{12,28,29}. A recent expansion of kinetic measurements across forms I, Iα, II-III, and IIIB supports an updated global median of $\approx 6 \text{ s}^{-1}$, approaching that of typical enzymes measured at $\approx 10 \text{ s}^{-1}$ ^{28,30,31}. Yet rubisco remains catalytically constrained relative to central metabolic enzymes, whose k_{catC} values often range between 20–80 s^{-1} ^{30,31}. Our data reinforce this trend while adding kinetic measurements from underrepresented rubisco clades, contributing to a more complete view of rubisco's catalytic landscape (Supplementary Table 3, Supplementary Fig. 7). Notably, two of the variants we characterized—*Prometheoarchaeum syntrophicum* and *Patescibacteria* sp. (MBU2523915.1)—exhibit relatively fast turnover rates ($k_{\text{catC}} = 2.4 \text{ s}^{-1}$), compared to nearly all deep-branching, non-CBB related rubiscos.

Large, deep-branching rubisco highlights structural diversity within dimers

Beyond phylogenetic diversity, the expanding catalog of rubisco sequences reveals variants with striking structural features. Two proteins—*L. archaeon* and *P. syntrophicum*—identified in a sparsely sampled deep-branching clade basal to form III-Like, stood out due to their exceptional sequence lengths: 557 and 559 amino acids, respectively. These are approximately 13% longer than other characterized rubisco catalytic subunits in the current literature and 20% larger than the average rubisco length of 466 amino acids (Fig. 4). Their extended sequences include expanded N- and C-termini and multiple insertions, representing primary structure deviations not observed in other characterized rubiscos (Fig. 4a, 4c).

The first insertion, spanning residues 79–84, modestly elongates the loop connecting β-strand 2 to α-helix 5. A second insertion, located at positions 108–124, introduces a short α-helix (helix 6) that bridges α-helix 5 and β-strand 3. Most strikingly, a third insertion between

residues 340–354 forms a β-sheet through interaction with the terminal region of the N-terminus—an architecture not observed in canonical rubiscos. To determine the phylogenetic distribution of these motifs, we analyzed the multiple sequence alignment of the 509 rubisco clusters used for our phylogenetic reconstruction (excluding RLPs). We found that the insertions at residues 79–84 and 108–124 are highly restricted, identified in only ~1.6% of sequences and exclusively localized to the *P. syntrophicum* clade. However, the larger insertion at residues 340–354—which forms the β-sheet—is slightly more widespread (~3.9%), appearing in both the *P. syntrophicum* clade and the Form III-like lineage. This suggests that while the β-sheet motif is a conserved structural strategy shared across these deep-branching groups, the additional helical and loop insertions are evolutionary innovations specific to the *P. syntrophicum* lineage. In addition to these insertions, we identified a major deletion (Fig. 4c): the absence of a β-hairpin motif typically conserved in bona fide rubiscos². This hairpin, which extends prominently in forms II-III and III-Like, is notably absent in RLPs and missing in these deep-branching variants.

These changes in primary and secondary structure cascade into the tertiary structure, resulting in visibly distinct monomeric folds compared to the consensus model or any previously characterized rubisco (Fig. 4b, 4d). Together, these structural differences suggest functional divergence and raise intriguing questions about the evolutionary pressures shaping rubisco architecture in these lineages—a hypothesis further supported by the low carboxylation turnover rates observed for this variant.

To explore structure–function relationships in deep-branching rubiscos, we determined the crystal structure of *P. syntrophicum*, a representative of this unclassified clade with a k_{catC} of 2.4 s^{-1} (Supplementary Table 4, Supplementary Fig. 8). The structure confirms multiple secondary-structure motifs observed in the sequence analysis. Most notably, the canonical β-hairpin motif following loop 6—typically present in bona fide carboxylating rubiscos and essential for active-site stabilization—is absent. Strikingly, *P. syntrophicum* appears to compensate for this deletion through a previously unobserved interaction: residues 16–26 form an extended loop that engages in a 2.2 Å side-chain–side-chain hydrogen bond between T20 and E395 (Fig. 4f). This interaction stabilizes the active-site region and may substitute mechanically for the missing β-hairpin. In addition, residues 16–22 contribute hydrophobic packing that mimics the β-hairpin's structural support. A further stabilizing element is provided by residues 342–353, which form a β-sheet that interacts with N-terminal residues 3–6 via an extended hydrogen-bonding network (Fig. 4f, Supplementary Fig. 9). This linkage could cascade to stabilize the 16–26 loop, suggesting a convergent strategy for maintaining carboxylation competency.

Relative to *R. rubrum* (Fig. 4e, 4f), a well-characterized form II rubisco, *P. syntrophicum* additionally exhibits an expanded C-terminal

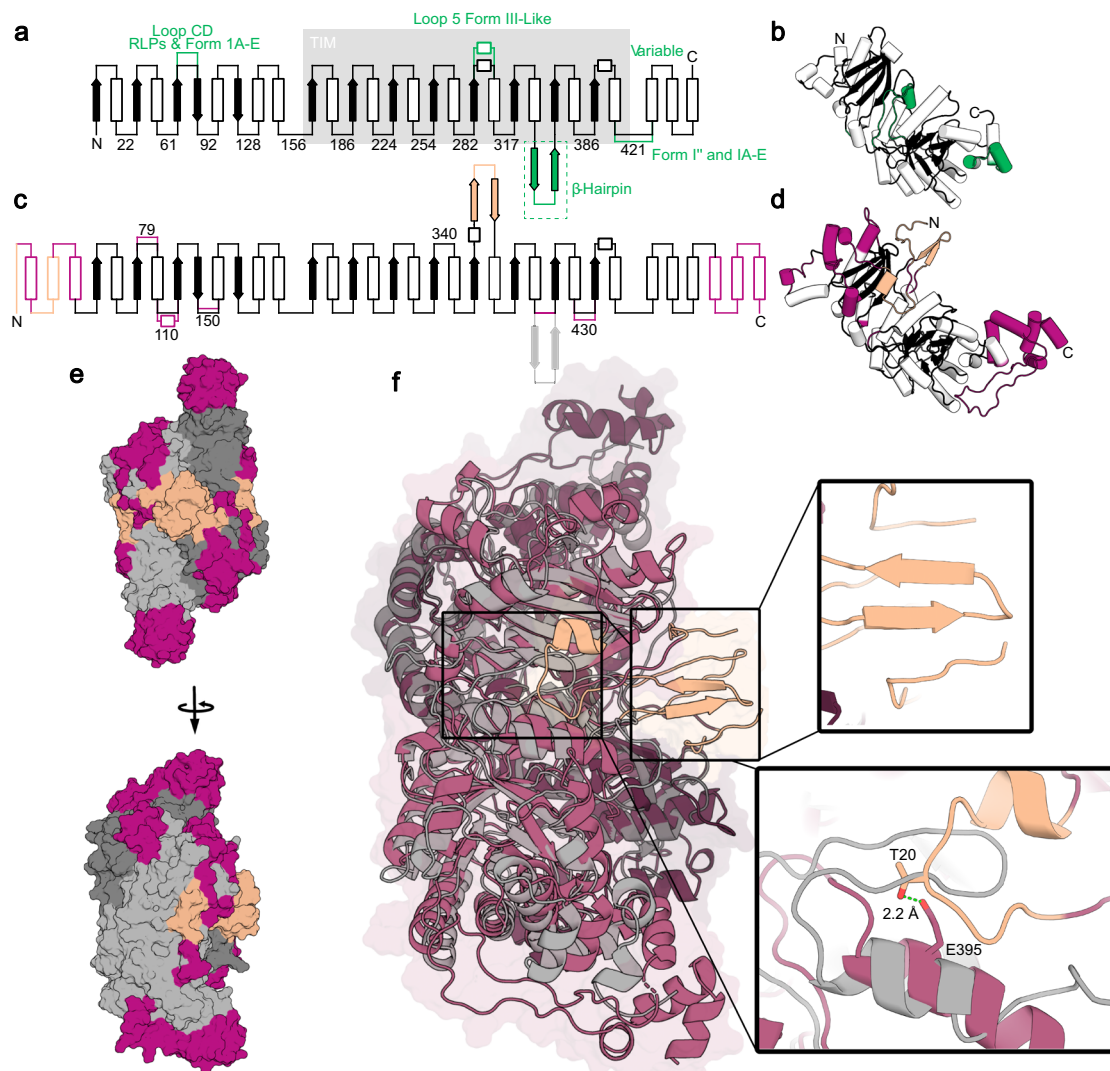


Fig. 4 | Deep-branching rubisco reveals unprecedented structural diversity of the core rubisco dimer. **a** 2d topology of rubisco consensus, **b** tertiary structure of consensus rubisco. **c** 2d topology of *Prometheoarchaeum syntrophicum* rubisco. Insertions shown in purple, with β -sheet and N-terminus shown in orange. **d** Tertiary structure of *Prometheoarchaeum syntrophicum* rubisco, Insertions shown in purple, with β -sheet and N-terminus shown in orange. **e** Surface representation of the crystal structure of *Rhodospirillum rubrum* (9RUB) and apo

Prometheoarchaeum syntrophicum rubisco (9Z01) superimposed shown in grey, and purple and pink respectively with 90° rotation. Orange as in **d**. **f** Transparent surface and cartoon representation of apo *Prometheoarchaeum syntrophicum* rubisco structure shown in purple and pink, 9RUB shown superimposed in gray, zoom in of distinct β -sheet motif shown in orange, and missing β -hairpin and functional equivalent loop light gray and orange, respectively. 2.2 Å sidechain-sidechain hydrogen bond shown with green dashed line.

region comprising a short terminal α -helix bundle followed by a large unstructured loop spanning residues 502–533. This extension, in combination with the insertions described above, may influence the rigidity of the quaternary structure and, despite being distal from the active sites, could ultimately affect carboxylation efficiency. Phylogenetically, *P. syntrophicum* occupies a deep-branching position within the rubisco clade, and its atypical structural features may reflect early divergence from canonical carboxylation architectures, offering insight into the range of structural plasticity tolerated within the rubisco lineage.

Together, these findings underscore rubisco's structural plasticity and reveal an alternative architectural solution for preserving carboxylation function in the absence of canonical motifs. The discovery of these compensatory mechanisms shows that rubisco can evolve alternative structural solutions to maintain carboxylation function when canonical motifs are lost and reinforces the need for deeper structural and functional exploration across rubisco clades. Future work to validate the functional role of these structural features will be

critical for deepening our understanding of the rubisco structure–function relationship. These motifs may represent an alternative architectural solution and offer an intriguing avenue for future studies into rubisco evolution and engineering.

Discussion

This study expands the functional and structural characterization of rubisco by exploring under-sampled phylogenetic clades and revising their evolutionary relationships. Rubisco's catalytic efficiency has long been studied, largely in the context of its importance in photosynthesis and carbon fixation; however, kinetic parameters remain sparsely documented outside form I and form II clades for this reason. To broaden this dataset, we quantified carboxylation turnover ($k_{\text{cat,C}}$) for seven deep-branching rubiscos spanning underexplored phylogenetic diversity. Our findings confirm that elevated catalytic rates have evolved only in Form I and II clades, while most rubiscos exhibit slower carboxylation kinetics. A previous study (de Pins et al.) identified one fast ($k_{\text{cat,C}} > 2 \text{ s}^{-1}$) rubisco outside of the form I and form II clades, but

upon further analysis we place that enzyme at the base of Form I²⁸. This high rate is therefore consistent with an evolution towards faster rates as rubisco became involved in the CBB cycle^{32,33}.

This evolutionary drive toward increased efficiency coincided with a fundamental reorganization of the form I clade, which we now show originates from a taxonomically diverse set of basal bacterial lineages. The earliest branching clade within form I, designated I β (Fig. 1b, Supplementary Fig. 10), includes genera such as *Ammonifex* (*Bacillota*), *Thermodesulfitomonas* (*Bacillota*), *Thermodulovibrio* (*Aquificota*), and *Thermodesulfobium* (*Thermodesulfobiota*)³². The latter of these was shown to harbor a modified, and perhaps ancestral, form of the CBB cycle^{1,32}. This clade bridges the transition from deep-rooted form I sequences with L₂ dimer oligomerization to those associated with higher-order oligomers L₈ in form I' and SSU-containing L₈S₈ *Chloroflexi*, and the L₈S₈ geometry continues through form IE (predominantly composed of sequences from the *Actinobacteria*) into the IC and IA groups, which are largely composed of *Proteobacteria*. This degree of taxonomic diversity suggests an extensive history of horizontal gene transfer (HGT) involving rubisco, and possibly other components of the CBB pathway, even among a set of relatively late-evolving and structurally constrained rubiscos.

In addition to their phylogenetic significance, the environmental origin of the deep-branching form I groups offer further insights into rubisco diversification. Members of the candidate division WOR-3 inhabit anaerobic sediments, hydrothermal vents, and hot springs, typically under acidic to neutral conditions (pH 2–7) and moderate to high temperatures^{17,34}. These organisms are generally facultative anaerobes, engaging in hydrogenotrophic and sulfur-cycling metabolisms^{17,34}. In contrast, species within the I β clade—*Ammonifex*, *Thermodesulfitomonas*, *Thermodulovibrio*, and *Thermodesulfobium*—are strict anaerobes that thrive in thermophilic, often geothermal or subsurface environments^{35–38}. *Ammonifex* and *Thermodesulfobium* are chemolithoautotrophs, utilizing hydrogen and reducing sulfur compounds, while *Thermodesulfitomonas* and *Thermodulovibrio* display organotrophic sulfur-reducing metabolisms^{39,40}. Despite broad similarities in temperature and redox conditions, WOR-3's broader pH tolerance and facultative anaerobiosis distinguish it ecologically from its counterparts with I β rubiscos.

These distinct ecological pressures may be drivers of the structural innovations and varied oligomeric states observed across the rubisco superfamily. By mapping these results onto our refined phylogenetic framework alongside our oligomeric state profiling, we demonstrate that the basic dimeric oligomeric state is very widespread, consistent with a dimeric origin of the rubisco superfamily. The observed structural and thermostability diversity, particularly among archaeal sequences, underscores the adaptive breadth of rubisco. In this context, higher-order assemblies (L₄, L₆, L₈, L₁₀) outside of form I appear to have arisen secondarily, perhaps as adaptive responses to stability demands, with oligomerization emerging as a consequence rather than a primary functional requirement. Such oligomeric complexity may have been driven by the same selective pressures that also shaped the evolution of the small subunit and the recruitment of protein chaperones. Strong selection for catalytic efficiency can destabilize protein structure, fostering alternative folding pathways and promoting interactions with stabilizing partners^{41–43}.

The prevalence of dimers among deep-branching lineages suggests that the ancestral rubisco functioned without the reinforcement of additional subunits. However, once higher-order oligomerization is established, mutations that would be tolerated in dimers become deleterious, preventing a return to the dimeric state and locking in complex architectures^{42,44}. Reversion remains possible so long as the interface is stabilized primarily by hydrogen-bonding networks rather than hydrophobic packing⁸. As Ng et al. point out, the more critical the protein, the more susceptible the organism is to purifying selection of mutations that disrupt the chaperone relationship or, in this case, the

higher order oligomerization. Our phylogenetic analysis is consistent with an extension of this hypothesis, since rubiscos involved in the CBB have frequently developed higher order oligomers, it may be the case that the initial establishment of higher order oligomerization or, indeed, chaperone dependence, is more likely when selection is strong and even slight, near-neutral improvements in protein stability come with a strong fitness advantage.

Beyond oligomeric state, deeper structural analysis of deep-branching rubiscos can broaden our understanding of the diversity of secondary structural elements and motifs that have evolved within this enzyme family billions of years ago. Notably, deep-branching forms such as *P. syntrophicum* exhibit atypical secondary structural elements and insertions that expand the known structural repertoire of the enzyme, highlighting a broader range of architectural plasticity and structural diversity than previously recognized. The absence of the β -hairpin in this structure is only seen in forms IIIA and IIIC, as well as all RLPs. The compensation of the β -hairpin in this structure (Fig. 4f) by an extension of Loop 5 raises the question of whether such compensation is necessary for rubisco activity. RLP structures do not compensate in this location (i.e. PDB structures 3QFW or 3FK4). Further structural studies on form IIIA and IIIC may or may not reveal additional compensations testifying to the importance of this structural feature.

Despite being one of the most abundant and consequential enzymes on Earth, rubisco remains unevenly characterized across its vast phylogenetic diversity, with only a fraction of lineages represented by purified and experimentally studied enzymes. This study underscores the difficulty of studying rubiscos from deep branching lineages and archaea; to date, only eleven rubiscos have been purified from an archaeal lineage, seven of which being described in this work. More comprehensive sampling of these clades, which represent the majority of rubisco sequence diversity, may yet reveal surprises. Further sampling at the base of the form I and II branches is also necessary to follow the evolution of rubisco as it became the primary carboxylase on our planet.

Methods

Phylogenetic tree inference for the rubisco superfamily

We began by building a comprehensive set of protein sequences from the rubisco superfamily, prioritizing large sequence databases and a subset of publications amalgamating rubisco kinetic data. Specifically, we queried the National Center for Biotechnology Information (NCBI) non-redundant protein database (*nr*) as well as the *enu_nr* (November 2024) database using a previously published set of form-specific Hidden Markov Models (HMMs)⁴⁵ using *hmmsearch* from the HMMER suite⁴⁶. Alignments were filtered to those in which target sequences obtained 90% or greater of overall model coverage. Resulting target sequences were appended to a separate set selected for biochemical characterization in this study, as well as those reported in previous publications^{11,12,28,29,47}. To remove partial or composite sequences, the total set was then secondarily filtered to those falling in between 358 and 716 amino acids in length.

Size-filtered sequences were next clustered using USEARCH (*-cluster_fast*) at a variety of identity thresholds. For each cluster at a 65% sequence identity threshold, we selected a representative, prioritizing those which we had attempted biochemical characterization, or, secondarily, those with the largest sequence lengths. Sequences from the Protein Data Bank (PDB) were not selected as representatives due to the frequent appearance of N- and C-terminal expression tags. Similarly, we removed a small number ($n=6$) of RLP with unstable phylogenetic positions in preliminary reconstructions. We used the same process to more deeply sample the phylogenetic branch including form I rubiscos, this time employing a sequence identity threshold of 90%. Form IIIA rubiscos were used as an outgroup in this case.

For both the full tree and form I subtree, representative sequences were aligned using MAFFT (--auto --reorder)⁴⁸, and alignments were subsequently trimmed using TRIMAL (-gt 0.1)⁴⁹ to remove phylogenetically uninformative positions. Sequence alignments were manually inspected in Geneious. Maximum-likelihood trees were inferred using IQTREE2 with a model selection step and 1000 ultra-fast bootstraps (-bnni, -m TEST -st AA -bb 1000)⁵⁰. Trees were similarly inferred for the four next best evolutionary models, and differences among the five topologies were summarized by calculating the pairwise normalized Robinson-Foulds (nRF) distances (mean nRF = 0.32, σ = 0.01)⁵¹. Clades were defined through a combination of automated and manual approaches, adjusting clade membership determined in a previous study¹ in consideration of 1) sequence diversity 2) the degree of bootstrap support 3) degree of taxonomic conservation and 4) prior literature. Characterization status for each representative sequence was assigned based on all members of its corresponding cluster; if any kinetic measurement had been previously made for any cluster member; that cluster was marked as successfully characterized. Similarly, if kinetic measurements had been attempted but failed, we marked clusters accordingly.

Analysis of genomic context for diverse rubiscos

To reconstruct the genomic loci for diverse rubiscos, we created a more deeply sampled sequence set from the clustering analysis performed above (95% sequence identity). For each of these rubiscos, we queried NCBI using a custom Python script that retrieves all other upstream and downstream protein sequences on the DNA fragment from which the rubisco was originally reported (code available at <https://github.com/alexanderjaffe/rubisco-diversity>). Taxonomic affiliations for each locus were also obtained through this method. Only loci with at least 10 open reading frames (ORFs) upstream and downstream of the rubisco sequence were retained for further analysis.

Next, all 'neighbor' protein sequences were clustered into protein families using MMSeqs (--cov-mode 0)⁵². Protein families were filtered to those with 5 or more members across all rubisco loci, and loci encoding RLPs were removed. Next, we computed the similarity of rubisco loci based on presence/absence patterns of their neighboring genes. This was accomplished by adapting a previous bioinformatic workflow²⁴. Briefly, a presence/absence matrix of protein families by locus was created and subjected to Principal Component Analysis (PCA) using Python's sklearn package. The first two PCA components were then passed to t-distributed stochastic neighbor embedding (t-SNE, also from sklearn) for additional dimensionality reduction. Finally, we visualized loci along the first two t-SNE axes as a function of their rubisco form, taxonomic affiliation, and presence/absence of selected KEGG orthologies (as determined by kofamscan)⁵³.

Plasmids, cloning

Representative *rbcl* genes were synthesized by JGI (sequences available as supplementary data) and cloned into a pET28 vector with an N-terminal His14-bdSUMO tag Plasmids pSF1389⁵⁴.

Expression and purification of recombinant proteins

Rubiscos were prepared by cotransforming plasmids containing His14-bdSUMO-tagged rubisco *RbcL* into chemically competent BL21 DE3 Star *E. coli* with pBADES/EL plasmid. Cells were grown to mid-log phase at 37 °C (OD600 $\approx$ 0.7–0.8) and overexpression of GroEL/ES was induced by the addition of 0.2% w/v arabinose and further incubation for 2 h. Cells were resuspended in fresh TB media (without arabinose) with 100 mM IPTG and shaken for 16 h at 16 °C. Pelleted cells were resuspended in pH 8.0 lysis buffer (20 mM sodium phosphate, 300 mM NaCl, 10 mM imidazole, 5% glycerol, 2 mM MgCl₂) with 5 mM PMSF and subject to a freeze-thaw cycle at -80 °C before lysis by use of a sonicator. The soluble fraction was collected by centrifugation

(15,000x g, 20 min). Clarified cell lysate was used to resuspend pre-equilibrated Ni-NTA resin, Ni-NTA resin was allowed to settle before the column was allowed to flow. Columns were washed thoroughly before resuspension in the bdSENPI reaction buffer (20 mM sodium phosphate pH 8.0, 300 mM NaCl, 10% glycerol). Purified bdSENPI was added to resuspended columns and rocked gently overnight at 4 °C to facilitate cleavage of the His14-bdSUMO tag from the target protein. Flow-through from the bdSENPI reaction was collected. Samples were stored in 20 mM sodium phosphate pH 8.0, 150 mM NaCl, 10 mM MgCl₂, 10 mM NaHCO₃ at -80 °C.

PAGE analyses

SDS-PAGE samples were prepared according to standard procedures in Laemmli Sample Buffer (Bio-rad) with 2-mercaptoethanol and heated at 98 °C for 5 min. Samples were resolved on 4-12% NuPAGE™ Bis-Tris Mini Protein Gels (Invitrogen) in 1 × Tris/Glycine/SDS buffer (Bio-Rad) and stained in InstantBlue® Coomassie protein stain (Abcam).

Small-angle X-ray-scattering (SAXS) data collection and analysis

Small-angle X-ray-scattering (SAXS) coupled in line with size-exclusion chromatography (SEC) experiments were performed with 100 μ l samples containing 1.45 mg ml⁻¹ of rubisco incubated with or without CABP prepared in 20 mM HEPES-OH (pH 8.0), 300 mM NaCl, 10 mM MgCl₂, 10 mM NaHCO₃. SEC-SAXS data were collected at the ALS beamline 12.3.1 at the Lawrence Berkeley National Laboratory⁵⁵. The X-ray wavelength was set at λ = 1.227 Å and the sample-to-detector distance was 2,100 mm, resulting in scattering vectors (q) ranging from 0.01 to 0.4 Å⁻¹. The scattering vector is defined as $q = 4\pi\sin\theta/\lambda$, where 2θ is the scattering angle. All experiments were performed at 20 °C, and the data were processed as described previously⁵⁶. Briefly, a SAXS flow cell was directly coupled with an inline 1260 Infinity HPLC system (Agilent) using a Shodex KW804 column (Showa Denko). The column was equilibrated with running buffer (20 mM HEPES-OH pH 8.0, 300 mM NaCl, 10 mM MgCl₂, 10 mM NaHCO₃) with a flow rate of 0.5 ml min⁻¹. A total 90 μ l of sample was separated by SEC and 3-s X-ray exposures were collected continuously during a 30-min elution. The SAXS frames recorded before sample analysis were subtracted from all other frames. The subtracted frames were investigated by R_g derived by the Guinier approximation, $I(q) = I(0) \exp(-q^2 R_g^2/3)$ with the limits $q R_g < 1.6$. The elution peak was mapped by comparing the integral of ratios to background and R_g relative to the recorded frame using the program RAW⁵⁷. Uniform R_g values across an elution peak represent a homogenous assembly. Final merged SAXS profiles, derived by integrating multiple frames across the elution peak, were used for further analysis, including Guinier plot, which determined aggregation-free state. The molecular weight from SAXS was determined by Volume of Correlation (V_c)⁵⁸.

All Size-exclusion chromatography-small-angle X-ray-scattering (SEC-SAXS) data, including processed data, are deposited in the Simple Scattering database (<https://simplescattering.com/>) with the ID codes listed in the Supplementary Table 1 and Supplementary Table 2.

SAXS modelling

The atomistic model of rubiscos in the open conformation was prepared based on the crystal structure of the closed conformation presented in this study, with missing N- and C-terminal residues added using the program MODELLER⁵⁹. Different extensions and compactions of the unfolded tails were built to screen conformational variability. The experimental SAXS profiles were then compared to theoretical scattering curves generated from these atomistic models using FoXS^{60,61}. Theoretical scattering profiles were fitted to the experimental SAXS data. To validate solution-state conformations of rubiscos. Models of rubisco were obtained through ALPHAFOLD⁶².

Crystallization and structural determination of rubisco

Ni-NTA-purified rubisco were further subject SEC on a Superose 6 Increase 10/300 GL, in a final buffer containing 100 mM Hepes (pH 8), 100 mM NaCl, 25 mM MgCl₂, 5 mM NaHCO₃, and 1 mM DTT. The dimeric rubisco *Prometheoarchaeum syntrophicum* (PsRuB) was screened against the following crystallization screens: MCSG-1 (Ana-trace); Crystal Screen, SaltRx, PEG/Ion, Index, and PEGRx (Hampton Research); and Berkeley Screen⁶³. Crystals of the *Prometheoarchaeum syntrophicum* rubisco were found in 100 mM Bis-Tris-HCl pH 6.5, 100 mM Ammonium phosphate dibasic, 20% (w/v) PEG 3,350 and 5% (v/v) 2-Propanol. Crystals from the *P. syntrophicum* rubisco were then placed in a reservoir solution containing 20% (v/v) glycerol and flash-cooled in liquid nitrogen.

The X-ray dataset for the PsRuB was collected at the Berkeley Center for Structural Biology beamline 8.2.1 at the Advanced Light Source at Lawrence Berkeley National Laboratory. The diffraction data were processed using the program Xia2⁶⁴. The crystal structure of PsRuB was solved by molecular replacement with the program PHASER⁶⁵ using the search model generated by ALPHAFOLD⁶². The atomic positions obtained from the molecular replacement were used to initiate model building using phenix.autobuild within the Phenix suite^{66,67}. Structure refinement was performed using the phenix.refine program⁶⁸. Manual rebuilding was done using COOT⁶⁹. Root mean square deviation differences from ideal geometries for bond lengths, angles, and dihedrals were calculated with Phenix⁶⁷. The stereochemical quality of the final models of *Prometheoarchaeum syntrophicum* were assessed by the program MOLPROBITY⁷⁰. A summary of crystal parameters, data collection, and refinement statistics can be found in Supplementary table 4. Structures and coordinates for PsRuB can be found in the PDB under accession ID [9Z01](#).

Rubisco activity assays

The rubisco carboxylation rate was measured using an NADH spectroscopic assay coupled to 3-phosphoglycerate (3-PGA) production, as previously described^{12,71,72}. Rubisco was activated in the presence of all assay components minus ribulose-1,5-bisphosphate (RuBP), the substrate. RuBP was synthesised in-house as previously described⁷³. Assay activating conditions include; pH 8, 25 °C, 4% CO₂ and 0.5% O₂ with orbital shaking at 1440 rpm for ~20 minutes in a 96 well flat-bottom transparent plate (Corning, Costar). The reaction was initiated with the addition of RuBP. NADH oxidation at 340 nm was used to estimate the RuBP turnover rate, accounting for the production of 2 3-PGA molecules per molecule of RuBP. Rubisco active site concentration was estimated using the known rubisco inhibitor, CABP, at $n=2$, using linear regression (Supplementary Fig. 3). Rates of carboxylation were calculated using methods as previously defined²¹. Two batches of CABP were synthesized from RuBP using both cold and ¹⁴C-labeled KCN^{74,75}. An aliquot of activated *R. rubrum* was used for determination of rubisco active sites via ¹⁴C-CABP binding and liquid scintillation, as previously described⁷⁶. The ¹⁴C-CABP quantified *R. rubrum* rubisco was then used to quantify the cold CABP for use in the NADH assay.

Protein thermal shift (PTS) assay

The PTS assay was conducted using a Protein Thermal Shift kit (Thermo Fisher). Samples were prepared with 1 mg ml⁻¹ protein in 1 × PTS phosphate buffer and 4 × PTS dye in Thermo Fisher MicroAmp Optical 8-Tube Strips. The assay was conducted on an Applied Biosciences QuantStudio 3 machine for quantitative PCR with reverse transcription. The assay consisted of initial cooling and hold at 16 °C for 1 min, followed by an 0.05 °C s⁻¹ increase to 100 °C and a final hold at 100 °C for 1 min. Data were analyzed in Protein Thermal Shift Software.

Reporting summary

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

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Protein sequences used in this study are available in Supplementary Data 7. Data for Phylogenetic trees are provided in Supplementary Data 1 and Supplementary Data 2 respectively. All data for kinetic distribution of the rubisco super family is provided in Supplementary Data 3. All data for pairwise sequence identity plots and TSNE analyses are provided in Supplementary Data 4 and Supplementary Data 5 respectively. All data for sequences represented in this study are provided in Supplementary Data 6. All data for TM analysis are provided in Supplementary Data 8. All Size-exclusion chromatography-small-angle X-ray-scattering (SEC-SAXS) data, including processed data, are deposited in the Simple Scattering database (<https://simplescattering.com/>) with the ID codes listed in the Supplementary Table 1 and Supplementary Table 2. The structure of PsRuB has been deposited in the PDB under accession ID [9Z01](#). The structures for [3QFW](#), [1RBL](#), [6URA](#), [5RUB](#), [5C2C](#), [7TIC](#), [1GEH](#), [2CWX](#), [8DHT](#), [5MAC](#), [2QYG](#), [9RUB](#) and [3FK4](#) are available in the PDB.

Code availability

Code produced for this study is available at the following link: <https://zenodo.org/records/19614364>.

References

1. Prywes, N., Phillips, N. R., Tuck, O. T., Valentin-Alvarado, L. E. & Savage, D. F. Rubisco function, evolution, and engineering. *Annu. Rev. Biochem.* **92**, 385–410 (2023).
2. Erb, T. J. & Zarzycki, J. A short history of RubisCO: the rise and fall (?) of Nature's predominant CO₂ fixing enzyme. *Curr. Opin. Biotechnol.* **49**, 100–107 (2018).
3. Bar-On, Y. M. & Milo, R. The global mass and average rate of rubisco. *Proc. Natl. Acad. Sci. USA.* **116**, 4738–4743 (2019).
4. Tabita, F. R. et al. Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. *Microbiol. Mol. Biol. Rev.* **71**, 576–599 (2007).
5. Hanson, T. E. & Tabita, F. R. A ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO)-like protein from *Chlorobium tepidum* that is involved with sulfur metabolism and the response to oxidative stress. *Proc. Natl. Acad. Sci. USA.* **98**, 4397–4402 (2001).
6. Erb, T. J. et al. A RubisCO-like protein links SAM metabolism with isoprenoid biosynthesis. *Nat. Chem. Biol.* **8**, 926–932 (2012).
7. Liu, D., Ramya, R. C. S. & Mueller-Cajar, O. Surveying the expanding prokaryotic Rubisco multiverse. *FEMS Microbiol. Lett.* **364**, (2017).
8. Liu, A. K. et al. Structural plasticity enables evolution and innovation of RuBisCO assemblies. *Sci. Adv.* **8**, eadc9440 (2022).
9. Gunn, L. H., Valegård, K. & Andersson, I. A unique structural domain in *Methanococcoides burtonii* ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) acts as a small subunit mimic. *J. Biol. Chem.* **292**, 6838–6850 (2017).
10. Schulz, L. et al. Evolution of increased complexity and specificity at the dawn of form I Rubiscos. *Science* **378**, 155–160 (2022).
11. de Pins, B. et al. A systematic exploration of bacterial form I rubisco maximal carboxylation rates. *EMBO J.* **43**, 3072–3083 (2024).
12. Davidi, D. et al. Highly active rubiscos discovered by systematic interrogation of natural sequence diversity. *EMBO J.* **39**, e104081 (2020).
13. Jaffe, A. L., Castelle, C. J., Dupont, C. L. & Banfield, J. F. Lateral gene transfer shapes the distribution of RuBisCO among Candidate Phyla Radiation bacteria and DPANN Archaea. *Mol. Biol. Evol.* **36**, 435–446 (2019).

14. Tabita, F. R., Satagopan, S., Hanson, T. E., Kreel, N. E. & Scott, S. S. Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *J. Exp. Bot.* **59**, 1515–1524 (2008).
15. Park, S. W. et al. Presence of duplicate genes encoding a phylogenetically new subgroup of form I ribulose 1,5-bisphosphate carboxylase/oxygenase in *Mycobacterium* sp. strain JC1 DSM 3803. *Res. Microbiol.* **160**, 159–165 (2009).
16. Ji, M. et al. Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* **552**, 400–403 (2017).
17. Zeng, J. et al. The anaerobic WOR-3 Bacteria: Hydrogenotrophic metabolism and unique carbon fixation via Archaeal form III RuBisCO. *Microbiology*, (2025).
18. Bundela, R., Keown, J., Watkin, S. & Pearce, F. G. Structure of a hyperthermostable dimeric archaeal Rubisco from *Hyperthermus butylicus*. *Acta Crystallogr. D Struct. Biol.* **75**, 536–544 (2019).
19. Huang, Q. & Szebenyi, D. M. E. Crystal structure of a type III Rubisco in complex with its product 3-phosphoglycerate. *Proteins* **91**, 330–337 (2023).
20. Nishitani, Y. et al. Structure-based catalytic optimization of a type III Rubisco from a hyperthermophile. *J. Biol. Chem.* **285**, 39339–39347 (2010).
21. Liu, A. K. et al. Deep-branching evolutionary intermediates reveal structural origins of form I rubisco. *Curr. Biol.* **33**, 5316–5325.e3 (2023).
22. Hayer-Hartl, M. & Hartl, F. U. Chaperone machineries of Rubisco - the most abundant enzyme. *Trends Biochem. Sci.* **45**, 748–763 (2020).
23. Badger, M. R. & Bek, E. J. Multiple Rubisco forms in proteobacteria: their functional significance in relation to CO₂ acquisition by the CBB cycle. *J. Exp. Bot.* **59**, 1525–1541 (2008).
24. Barnum, T. P. & Coates, J. D. Chlorine redox chemistry is widespread in microbiology. *ISME J.* **17**, 70–83 (2023).
25. Perica, T. et al. Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science* **346**, 1254346 (2014).
26. Fraser, N. J. et al. Evolution of protein Quaternary structure in response to selective pressure for increased thermostability. *J. Mol. Biol.* **428**, 2359–2371 (2016).
27. Maeda, N., Kanai, T., Atomi, H. & Imanaka, T. The unique pentagonal structure of an archaeal Rubisco is essential for its high thermostability. *J. Biol. Chem.* **277**, 31656–31662 (2002).
28. de Pins, B. et al. Rubisco is slow across the tree of life. *bioRxiv* <https://doi.org/10.1101/2025.01.19.633714> (2025).
29. Flamholz, A. I. et al. Revisiting trade-offs between Rubisco kinetic parameters. *Biochemistry* **58**, 3365–3376 (2019).
30. Bar-Even, A. et al. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410 (2011).
31. Davidi, D., Longo, L. M., Jabłońska, J., Milo, R. & Tawfik, D. S. A bird's-eye view of enzyme evolution: Chemical, physico-chemical, and physiological considerations. *Chem. Rev.* **118**, 8786–8797 (2018).
32. Frolov, E. N. et al. Form III RubisCO-mediated transaldolase variant of the Calvin cycle in a chemolithoautotrophic bacterium. *Proc. Natl. Acad. Sci. USA*. **116**, 18638–18646 (2019).
33. Bouvier, J. W., Emms, D. M. & Kelly, S. Rubisco is evolving for improved catalytic efficiency and CO₂ assimilation in plants. *Proc. Natl. Acad. Sci. USA*. **121**, e2321050121 (2024).
34. Mori, K. et al. Isolation and characterization of a bacterium affiliated with the hitherto uncultured candidate phylum WOR-3 from a deep-sea hydrothermal fluid. *Appl. Environ. Microbiol.* **91**, e0018825 (2025).
35. Miroshnichenko, M. L. et al. *Ammonifex thiophilus* sp. nov., a hyperthermophilic anaerobic bacterium from a Kamchatka hot spring. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY* **58**, 2935–2938 (2008).
36. Mori, K., Kim, H., Kakegawa, T. & Hanada, S. A novel lineage of sulfate-reducing microorganisms: *Thermodesulfobiaceae* fam. nov., *Thermodesulfobium narugense*, gen. nov., sp. nov., a new thermophilic isolate from a hot spring. *Extremophiles* **7**, 283–290 (2003).
37. Slobodkina, G. B., Baslerov, R. V., Novikov, A. A., Bonch-Osmolovskaya, E. A. & Slobodkin, A. I. *Thermodesulfitomonas autotrophica* gen. nov., sp. nov. a thermophilic, obligate sulfite-reducing bacterium isolated from a terrestrial hot spring. *Int. J. Syst. Evol. Microbiol.* **67**, 301–305 (2016).
38. Henry, E. A. et al. Characterization of a new thermophilic sulfate-reducing bacterium: *Thermodesulfovibrio yellowstonii*, gen. nov. and sp. nov.: its phylogenetic relationship to *Thermodesulfobacterium commune* and their origins deep within the bacterial domain. *Arch. Microbiol.* **161**, 62–69 (1994).
39. BenHania, W., Ollivier, B., Fardeau, M.-L. The family *thermodesulfobiaceae*. The Prokaryotes 421–426. Springer: Berlin Heidelberg, Berlin, Heidelberg, 2014.
40. Podstawka, A. *Ammonifex thiophilus* SR.
41. Pereira-Leal, J. B., Levy, E. D. & Teichmann, S. A. The origins and evolution of functional modules: lessons from protein complexes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 507–517 (2006).
42. Levy, E. D., Boeri Erba, E., Robinson, C. V. & Teichmann, S. A. Assembly reflects evolution of protein complexes. *Nature* **453**, 1262–1265 (2008).
43. Ng, J. et al. Origin of chaperone dependence and assembly complexity in Rubisco biogenesis. *Evol. Biol.* (2025).
44. Hochberg, G. K. A. et al. A hydrophobic ratchet entrenches molecular complexes. *Nature* **588**, 503–508 (2020).
45. Jaffe, A. L., Salcedo, R. S. R. & Dekas, A. E. Abundant and metabolically flexible bacterial lineages underlie a vast potential for rubisco-mediated carbon fixation in the dark ocean. *Genome Biol* **26**, 167 (2025).
46. HMMER. <http://hmmer.org>.
47. Iñiguez, C. et al. Evolutionary trends in RuBisCO kinetics and their co-evolution with CO₂ concentrating mechanisms. *Plant J.* **101**, 897–918 (2020).
48. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
49. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
50. Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
51. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
52. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
53. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
54. Frey, S. & Görllich, D. A new set of highly efficient, tag-cleaving proteases for purifying recombinant proteins. *J. Chromatogr. A*. **1337**, 95–105 (2014).
55. Dyer, K. N. et al. High-throughput SAXS for the characterization of biomolecules in solution: a practical approach. *Methods Mol. Biol.* **1091**, 245–258 (2014).

56. Hura, G. L. et al. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat. Methods* **6**, 606–612 (2009).
57. Hopkins, J. B. BioXTAS RAW 2: new developments for a free open-source program for small-angle scattering data reduction and analysis. *J. Appl. Crystallogr.* **57**, 194–208 (2024).
58. Rambo, R. P. & Tainer, J. A. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* **496**, 477–481 (2013).
59. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
60. Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* **38**, W540–W544 (2010).
61. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105**, 962–974 (2013).
62. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
63. Pereira, J. H., McAndrew, R. P., Tomaleri, G. P. & Adams, P. D. Berkeley Screen: a set of 96 solutions for general macromolecular crystallization. *J. Appl. Crystallogr.* **50**, 1352–1358 (2017).
64. Winter, G., Lobley, C. M. C. & Prince, S. M. Decision making in xia2. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1260–1273 (2013).
65. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
66. Terwilliger, T. C. et al. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).
67. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **75**, 861–877 (2019).
68. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).
69. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
70. Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **27**, 293–315 (2018).
71. Lilley, R. M. & Walker, D. A. The reduction of 3-phosphoglycerate by reconstituted chloroplasts and by chloroplast extracts. *Biochim. Biophys. Acta* **368**, 269–278 (1974).
72. Kubien, D. S., Brown, C. M. & Kane, H. J. Quantifying the amount and activity of Rubisco in leaves. *Methods Mol. Biol.* **684**, 349–362 (2011).
73. Horecker, B. L., Hurwitz, J. & Weissbach, A. The enzymatic synthesis and properties of ribulose 1,5-diphosphate. *J. Biol. Chem.* **218**, 785–794 (1956).
74. Pierce, J., Carlson, T. J. & Williams, J. G. A cyanobacterial mutant requiring the expression of ribulose biphosphate carboxylase from a photosynthetic anaerobe. *Proc. Natl. Acad. Sci. USA.* **86**, 5753–5757 (1989).
75. Kane, H. J., Wilkin, J. M., Portis, A. R. & John Andrews, T. Potent inhibition of ribulose-bisphosphate carboxylase by an oxidized impurity in ribulose-1,5-bisphosphate. *Plant Physiol* **117**, 1059–1069 (1998).
76. Whitney, S. M. & Sharwood, R. E. Rubisco engineering by Plastid transformation and protocols for assessing expression. *Methods Mol. Biol.* **2317**, 195–214 (2021).

Acknowledgements

The authors wish to extend their sincere gratitude to their families for the steadfast support and encouragement provided throughout the course of this work. We are deeply indebted to our partners, parents, and loved ones, whose patience and understanding sustained us during the demands of this research. Their contributions, though not visible in the data presented herein, are inseparable from its completion.

Author contributions

A.J.K., J.B.S., P.M.S., and N.P. carried out overall experimental design. A.L.J., L.V.A., J.F.B. and N.P. carried out generation of a comprehensive rubisco diversity library. A.L.J., L.M.W., and A.J.K. conducted the genomic context analysis. All protein samples were prepared by A.J.K. *E. coli* transformations, growth and expression were carried out by J.L. and C.Y. A.J.K. and M.H. conducted all SEC-SAXS-MALS analyses. J.L. conducted all protein thermal shift experiments. L.T.K. conducted all kinetic experiments. J.H.P., P.D.A. performed x-ray crystallography data acquisition, image processing, and structure determination. All authors contributed to writing and manuscript preparation.

Funding

This work was part of the DOE Joint BioEnergy Institute (<https://www.jbei.org>) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. The work (proposal:509657 DOI:10.46936/10.25585/60008804) conducted by the U.S. Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. SAXS data was collected at the Advanced Light Source (ALS) at the SIBYLS beamline, through the Integrated Diffraction Analysis Technologies (IDAT) program, supported by DOE Office of Biological and Environmental Research. Additional support comes from the National Institute of Health project ALS-ENABLE (P30 GM124169). A.L.J. was supported by the Stanford Science Fellows program. L.M.W. was supported by the National Science Foundation Graduate Research Fellowship Program.

Competing interests

P.M.S. has financial interests in BasidioBio and Totality Biosciences. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-73982-5>.

Correspondence and requests for materials should be addressed to Noam Prywes or Patrick M. Shih.

Peer review information *Nature Communications* thanks Robert H. Wilson, who co-reviewed with Yunlong Zhao and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

¹Biophysics Graduate Group, University of California, Davis, Davis, CA, USA. ²Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. ³Feedstocks Division, Joint BioEnergy Institute, Emeryville, CA, USA. ⁴Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA 94720, USA. ⁵Department of Earth System Science, Stanford University, Stanford, CA, USA. ⁶Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ⁷Genome Center, University of California Davis, Davis, CA, USA. ⁸Department of Bioengineering, University of California, Berkeley, California, USA. ⁹Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. ¹⁰Department of Electrical Engineering and Computer Science, University of California, Berkeley California, USA. ¹¹Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Clayton, Australia. ¹²Department of Bioengineering, University of California, Berkeley, CA 94720, USA. ¹³Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA, USA. ¹⁴Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA. ¹⁵Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, USA. ¹⁶Department of Chemistry, University of California-Davis, Davis, California 95616, USA. ¹⁷Department of Biochemistry and Molecular Medicine, University of California-Davis, Davis, California 95616, USA. ¹⁸Department of Biochemistry, University of Cambridge, Cambridge, UK. ¹⁹These authors contributed equally: Alexander J. Kehl, Leah Taylor-Kearney.

✉ e-mail: np648@cam.ac.uk; pmshih@berkeley.edu