Evolutionary Dynamics of RuBisCO: Emergence of the Small Subunit and its Impact Through Time

Kaustubh Amritkar[1,2], Bruno Cuevas-Zuviría[1,3] and Betül Kaçar[1*]

[1] Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA

[2] Biophysics Graduate Degree Program, University of Wisconsin-Madison, Madison, WI, USA

[3] Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid, Madrid, Spain

* Corresponding author: E-mail: bkacar@wisc.edu

Short Title: Small subunit emergence and RuBisCO evolution

Abstract

Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) is an ancient protein critical for CO_2-fixation and global biogeochemistry. Form-I RuBisCO complexes uniquely harbor small subunits that form a hexadecameric complex together with their large subunits. The small subunit protein is thought to have significantly contributed to RuBisCO's response to the atmospheric rise of O_2 ~2.5 billion years ago, marking a pivotal point in the enzyme's evolutionary history. Here, we performed a comprehensive evolutionary analysis of extant and ancestral RuBisCO sequences and structures to explore the impact of the small subunit's earliest integration on the molecular dynamics of the overall complex. Our simulations suggest that the small subunit restricted the conformational flexibility of the large subunit early in its history, impacting the evolutionary trajectory of the Form-I RuBisCO complex. Molecular dynamics investigations of CO_2 and O_2 gas distribution around predicted ancient RuBisCO complexes suggest that a proposed "CO_2 reservoir" role for the small subunit is not conserved throughout the enzyme's evolutionary history. The evolutionary and biophysical response of RuBisCO to changing atmospheric conditions on ancient Earth showcase multi-level and trackable responses of enzymes to environmental shifts over long timescales.

Keywords:

RuBisCO, small subunit, ancestral sequence reconstruction, structural dynamics

Introduction

Life has been evolving on this planet for four billion years. Evolution allows for exploration of the vast sequence space comprised by polymers of the twenty standard amino acids – a space that exceeds the number of molecules in the universe (Wagner and Rosen 2014). From within this space, biology has discovered unique solutions to the challenges of growing and persisting in the ever-shifting diversity of Earth's environments. Undoubtedly, biogeochemically critical microbial enzymes are central to this dynamic, long-term interaction between life and the environment (Falkowski et al. 2008). One such key enzyme that evolved early in the history of life and persisted through planetary extremes is RuBisCO (ribulose-1,5-bisphosphate carboxylase/oxygenase).

RuBisCO is a globally critical, ancient enzyme with an intriguing history. It facilitates the rate-limiting step of the Calvin-Benson-Bassham cycle for carbon fixation, catalyzing the addition of atmospheric carbon dioxide (CO_2) with ribulose 1,5-bisphosphate (RuBP) (Andersson 2008). In the presence of oxygen (O_2), RuBisCO also catalyzes a competing oxygenation reaction in which RuBP combines with O_2, producing an autoinhibitory metabolite detrimental to the overall metabolic efficiency of carbon fixation (Fernie and Bauwe 2020). While the exact age of RuBisCO is not clearly known, paleobiological inferences (Ward and Shih 2019; Garcia et al. 2021; Kędzior et al. 2022) suggest that it emerged prior to the rise of atmospheric O_2 ~2.5 billion years ago, a period known as the Great Oxidation Event (GOE). Accordingly, despite being infamously sensitive to O_2, RuBisCO was maintained by organisms through this drastic atmospheric upheaval that impacted the ecosystem and its constituent biomolecules (Ashida et al. 2005; Raymond and Segrè 2006; Wang et al. 2011; Young et al. 2012; Caetano-Anollés 2017; Kacar et al. 2017; Erb and Zarzycki 2018; Garcia et al. 2021).

52 Extant RuBisCO forms (I to IV) exhibit different dimeric and poly-dimeric assemblies. All
53 forms have large subunits (RbcL), but Form-I RuBisCOs uniquely have an additional small subunit
54 (RbcS) (**Fig 1A**). Form-I RuBisCOs are also the most abundant (Tabita et al. 2008) and generally
55 exhibit higher specificities for CO_2 (Flamholz et al. 2019). These attributes make RbcS a promising
56 candidate to study how small accessory subunits can regulate the evolution of RuBisCO, though
57 their precise evolutionary role remains unclear. RbcS is thought to be crucial for the assembly of
58 the RbcL octameric complex made of RbcL homodimers (Liu et al. 2010; Esquivel et al. 2013;
59 Joshi et al. 2015) and significantly impacts RuBisCO's catalytic parameters (Gatenby 1988;
60 Spreitzer 2003; Genkov et al. 2010; Esquivel et al. 2013; Matsumura et al. 2020; Mao et al. 2022).
61 For example, incorporation of ancestral RbcS in RuBisCO was shown to increase the enzyme's
62 specificity for CO_2, indicating that emergence of RbcS played a critical role in the ancestor of
63 Form-I RuBisCO (Schulz et al. 2022). Previous computational work also suggested that RbcS can
64 act as a "CO_2-reservoir" to concentrate CO_2 molecules within the enzyme (Van Lun et al. 2014).
65 However, the extent to which the small subunit has influenced the structural dynamics and
66 evolutionary trajectory of the Form-I RuBisCO complex remains less understood.

67 Here we used phylogenetic reconstructions, structural predictions as well as molecular
68 dynamics (MD) simulations to track the evolution of the small subunit in the context of the
69 RuBisCO complex over geologic time. Our exploration focused on the hypothesis that the
70 molecular and structural innovations likely made Form-I RuBisCO selectively more advantageous
71 following the planetary rise of O_2. We specifically focused on three areas: 1) the sequence and
72 structural evolution of subunits within ancient RuBisCO complexes following integration of
73 ancient RbcS, 2) the impact of RbcS presence or absence on the structural dynamics of ancestral

74 and extant RuBisCO complexes, and 3) migration of CO_2 and O_2 gases in ancestral RuBisCO to
75 assess a proposed CO_2-reservoir role for ancient small subunits (Van Lun et al. 2014).
76

77 Results and Discussion

78 Resurrection of ancestral RuBisCOs

79 We built a maximum-likelihood phylogenetic tree from a concatenated alignment of RbcL
80 and RbcS amino acid sequences representing RuBisCO Forms I to IV, including the recently
81 described Form-I' (Banda et al. 2020) (**Fig. 1B**). The phylogeny contains 194 sets of RbcL-RbcS
82 homologs from Form-I and 135 RbcL homologs from all other forms (including Form-I', II, II/III,
83 III and IV). The tree contains sequences representative of known RuBisCO diversity and is rooted
84 by Form-IV RuBisCO-like proteins in accordance with previous studies (Tabita et al. 2007; Kacar
85 et al. 2017; Poudel et al. 2020). Form-I is categorized into four major subgroups: the "green-like"
86 Form-IA and IB, prevalent in proteobacteria, cyanobacteria, green algae and plants, and the "red-
87 like" Form-IC and ID, prevalent in proteobacteria and non-green algae (Spreitzer 2003; Tabita et
88 al. 2008). Our phylogeny also resolves a Form-I subclade, Form-IE, that diverges before the last
89 common ancestor of Form-IA and Form-IB (West-Roberts et al. 2021). Most sequences from the
90 Form-IE clade are from metagenomic studies and belong to unclassified members of the
91 *Chloroflexota* bacterial phylum. Form-I' RuBisCOs cluster within a monophyletic clade sister to
92 all other Form-I sequences. Form-I' RuBisCOs are notable because their RbcL subunits are similar
93 to those of Form-I in their multimeric assembly, but they lack the RbcS (Banda et al. 2020). Like
94 Form-IE, the Form-I' clade contains sequences from *Chloroflexota*. The topology of our

concatenated phylogeny is in agreement with previously reported RbcL-based phylogenetic trees (Kacar et al. 2017; Banda et al. 2020).

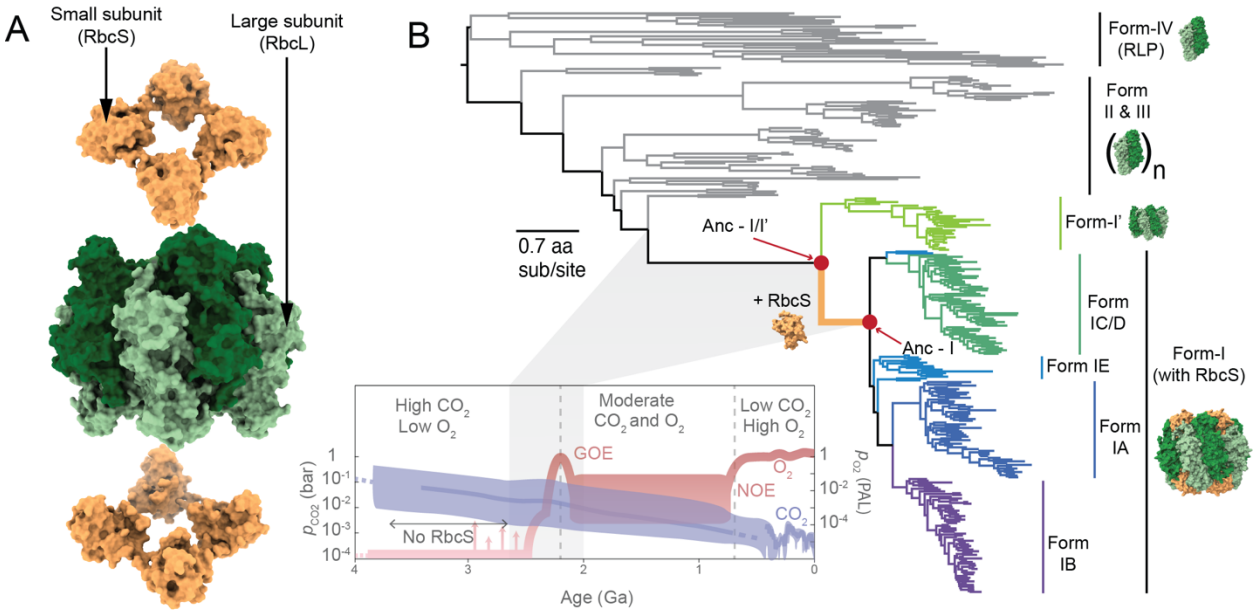

Fig. 1: The evolutionary history and diversification of RuBisCO.
(A) Architecture of the L_8S_8 structure of the Form-I RuBisCO complex, with RbcL shown in green and RbcS in orange (PDB:1BWV). (B) RuBisCO phylogenetic tree highlighting the emergence of RbcS coinciding with the Great Oxidation Event (GOE) (Kacar et al. 2017; Banda et al. 2020). Schematic on the right shows variation in the multimeric structure of the RuBisCO complex across different forms. Form-II and III RuBisCOs exhibit multiple homooligomeric states of the RbcL-dimer, such as RbcL-dimer, tetramer or hexamer (Liu et al. 2022), represented here as $(RbcL\text{-}RbcL)_n$, where n is the number of RbcL-dimers. Bottom inset schematic shows the change in atmospheric CO_2 and O_2 concentration through Earth's history (Rucker and Kaçar 2023).

We selected ancestral nodes situated along the evolutionary trajectory immediately before and after RbcS incorporation into RuBisCO for ancestral sequence reconstruction (Methods). Specifically, we inferred the common ancestor of both Form-I' and Form-I RuBisCOs, as well as the ancestors of major Form-I clades as highlighted in **Fig. 2A**, inferring a total of seven ancestral nodes along the phylogenetic tree.

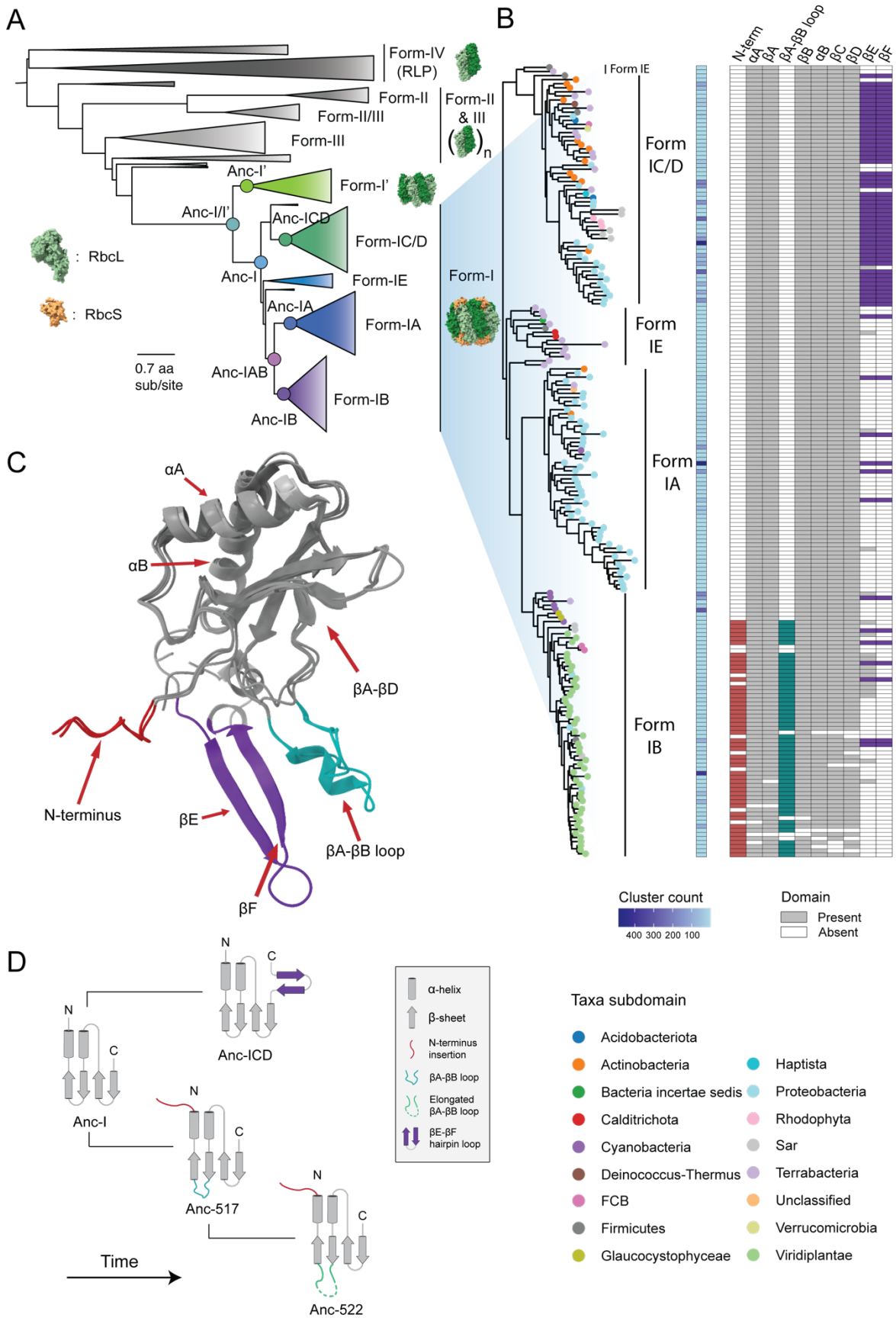
A
Form-IV (RLP)
Form-II
Form-II/III
Form-III
Anc-I'
Form-I'
Anc-I/I'
Anc-ICD
Form-IC/D
: RbcL
Anc-I
Form-IE
: RbcS
Anc-IA
Form-IA
0.7 aa sub/site
Anc-IAB
Form-IB
Anc-IB
Form-II & III
Form-I
B
Form IE
Form IC/D
Form IE
Form IA
Form IB
N-term
αA
βA
βA-βB loop
βB
αB
βC
βD
βE
βF
Cluster count
400 300 200 100
Domain
Present
Absent
C
αA
αB
βA-βD
N-terminus
βE
βA-βB loop
βF
D
N
C
Anc-ICD
Anc-I
Anc-517
Anc-522
Time
α-helix
β-sheet
N-terminus insertion
βA-βB loop
Elongated βA-βB loop
βE-βF hairpin loop
Taxa subdomain
Acidobacteriota
Actinobacteria
Bacteria incertae sedis
Calditrichota
Cyanobacteria
Deinococcus-Thermus
FCB
Firmicutes
Glaucocystophyceae
Haptista
Proteobacteria
Rhodophyta
Sar
Terrabacteria
Unclassified
Verrucomicrobia
Viridiplantae

Fig. 2: Phylogenetic analysis of RbcS structural diversity.
(A) Collapsed maximum-likelihood phylogenetic tree of concatenated RbcL-RbcS sequences. Ancestral nodes used in this study are highlighted. (B) Analysis of RbcS structural diversity, mapped to a clustered Form-I phylogeny, where each tip represents a cluster of RbcS sequences with >63% sequence identity. A heatmap for the number of sequences present in the RbcS cluster is represented in blue. Extant nodes are colored by microbial host taxonomic diversity. The presence and absence of different RbcS features in each extant sequence cluster is indicated by gray and white colors, and the color scheme for the presence of uncommon features is according to (C). (C) Multiple RbcS structures across Form-I subgroups aligned to highlight the different RbcS structural features. (D) Schematic representing the emergence of RbcS structural features across the Form-I RuBisCO phylogeny.

Sequence and structural diversity of modern and ancient RbcS proteins

We analyzed all extant Form-I RbcS sequences from our database representative of known host taxonomic diversity (**Fig. 2B**, Methods). Extant Form-I RbcS homologs show remarkably low mean pairwise sequence identity (29%) compared to Form-I RbcL sequences (66%) (**supplementary fig. 3**) as observed by others previously (Bracher et al. 2017; Mao et al. 2022; Bouvier et al. 2024).

We inferred sequences for 328 ancestral nodes within the phylogenetic tree, with 193 of these nodes belonging to the Form-I clade, including its common ancestor, Anc-I. Form-I RbcL ancestors are reconstructed with higher confidence than Form-I RbcS ancestors, with mean posterior probabilities of 0.98 and 0.90, respectively, across all positions (**supplementary fig. 4**). Anc-I subunit sequences have a mean pairwise identity of 71.5% (RbcL) and 42.2% (RbcS) compared across all extant RuBisCOs in the phylogeny, indicating a higher conservation of RbcL compared to RbcS. The mean posterior probabilities for Anc-I RbcL and RbcS are 0.95 and 0.85 respectively.

140 We further explored historical, evolutionary variation of RbcS by identifying known,
141 functionally significant sequence and structural features (Spreitzer 2003; Mao et al. 2022) (**Fig.**
142 **2C**) and mapping their presence across extant and reconstructed ancestral RbcS in our
143 phylogenetic tree (**Fig. 2B and 2D**). We find that the RbcS structure contains novel features unique
144 to different Form-I specific lineages. While all RuBisCO small subunits share a common core
145 structure consisting of a β-sheet with four antiparallel strands (βA to βD) and two α-helices (αA
146 & αB) (Knight et al. 1990), the Form-IB and Form-IC/D RbcS exhibit additional distinct
147 characteristics.

148 Within the Form-IB clade, we find that most green-like eukaryotic RbcS homologs contain
149 an additional N-terminal region as well as a βA-βB loop insertion. The N-terminal region is
150 specifically known to be a signal peptide necessary for the entry of RbcS into the chloroplast, prior
151 to assembly with RbcL (Schmidt and Mishkind 1986; Spreitzer 2003). The N-terminal extension
152 feature is first observed in the Form-IB ancestry after the separation of cyanobacteria in the
153 phylogeny (**supplementary fig. 5**). Absence of the N-terminal extension in Anc-IB, the common
154 ancestor of Form-IB, as well as older ancestors indicates that their ancient hosts did not translate
155 RbcS and RbcL in separate organelles. Rather, expression of subunits in different organelles
156 emerged as a trait following the bifurcation of cyanobacteria and Viridiplantae in Form-IB
157 RuBisCO.

158 We find that the βA-βB loop insertion is similarly exclusive to green-like eukaryotic RbcS
159 (**Fig. 2B**). Although the significance of the βA-βB loop in RbcS is not established, it plays an
160 important part in regulating the size of the central solvent channel in RuBisCO complex (Esquivel
161 et al. 2013). Like N-terminal extension, this loop first emerges after the splitting of cyanobacterial

162 and eukaryotic lineages in Form-IB (**supplementary fig. 5**). As observed with extant RbcS
163 (Spreitzer 2003; Mao et al. 2022), the βA-βB loop insertion length also varies across Form-IB
164 ancestors (10 to 28 aa). The loop has increased by ~11 residues in all eukaryotic ancestors after
165 the divergence from cyanobacterial ancestors and by ~18 residues in the ancestors from green
166 algae lineage within Form-IB (**supplementary fig. 5**).

167 We also tracked the evolution of the C-terminal βE and βF hairpin loop that is unique to
168 Form-ICD RbcS sequences (**Fig. 2B**). The common ancestor of Form-IC/D sequences, Anc-ICD,
169 and all subsequent Form-IC/D ancestors have the βE and βF hairpin (**Fig. 2D**). By contrast, Anc-
170 I RbcS does not have the βE and βF hairpin, suggesting this insertion happened after the divergence
171 of Form-IC/D within the Form-I clade. The C-terminal β-hairpin is functionally significant for
172 mediating the assembly of the oligomeric complex, allowing red-like RuBisCOs to assemble
173 without specialized assembly chaperones (Joshi et al. 2015). The incorporation of the hairpin loop
174 in Form-IC/D common ancestor, Anc-IC/D, suggests that independence from assembly
175 chaperones may have first emerged in a red-like RuBisCO ancestor. Our analysis thus reveals that
176 RbcS has undergone different clade-specific structural variations since its emergence in the Form-
177 I ancestor.

178

179 Global distribution of residues evolutionarily linked to RbcS
180 incorporation

181 RbcL from Form-I and Form-I' RuBisCOs exhibit the same L_8 oligomeric arrangement.
182 Because the presence or absence of the small subunit is the only major structural distinction

183 between the two forms (Banda et al. 2020), we performed comparative analyses to reveal the small
184 subunit's impact on RbcL sequence and structure. We focused on the RbcL extant sequences from
185 the Form-I and Form-I' clades. We identified residues that are conserved within each clade, but
186 differ between the two clades, which we refer to as "signature positions". These residues are more
187 likely to contribute to the functional distinctions between Form-I and Form-I', including
188 interactions (or lack thereof) with RbcS. The interface region between the large and small subunit
189 has been shown to be critical for the incorporation of RbcS into the Form-I RuBisCO complex
190 (Knight et al. 1990; Van Lun et al. 2011; Ryan et al. 2019; Schulz et al. 2022). We therefore
191 hypothesized that signature positions would cluster within the RbcL-RbcS interface.

192 Form-I and Form-I' sequence analyses revealed eleven, discontinuous signature positions
193 (**Fig. 3A**). These positions are clustered in three different regions on the RuBisCO structure. Two
194 residues (Val 70 and Asp 74; position index from *Rhodobacter sphaeroides*) are located at the
195 RbcL intradimer interface (highlighted in blue), three (Phe 201, Leu 401 and Gln 402) are buried
196 in the RbcL monomer close to the active site (highlighted in purple) and six (Lys 260, Trp 284,
197 Arg 286, Asn 288, Gly 159 and Glu162) are found at the RbcL dimer-dimer interface (highlighted
198 in red) (**Fig. 3A**). Out of the six residues located at the RbcL dimer-dimer interface, Gly 159 and
199 Glu 162 are the only two signature positions that are also present at the RbcL-RbcS interface. A
200 similar comparison between the Form-IA/B/E and Form-IC/D RbcS sequences did not identify
201 any signature positions (Methods).

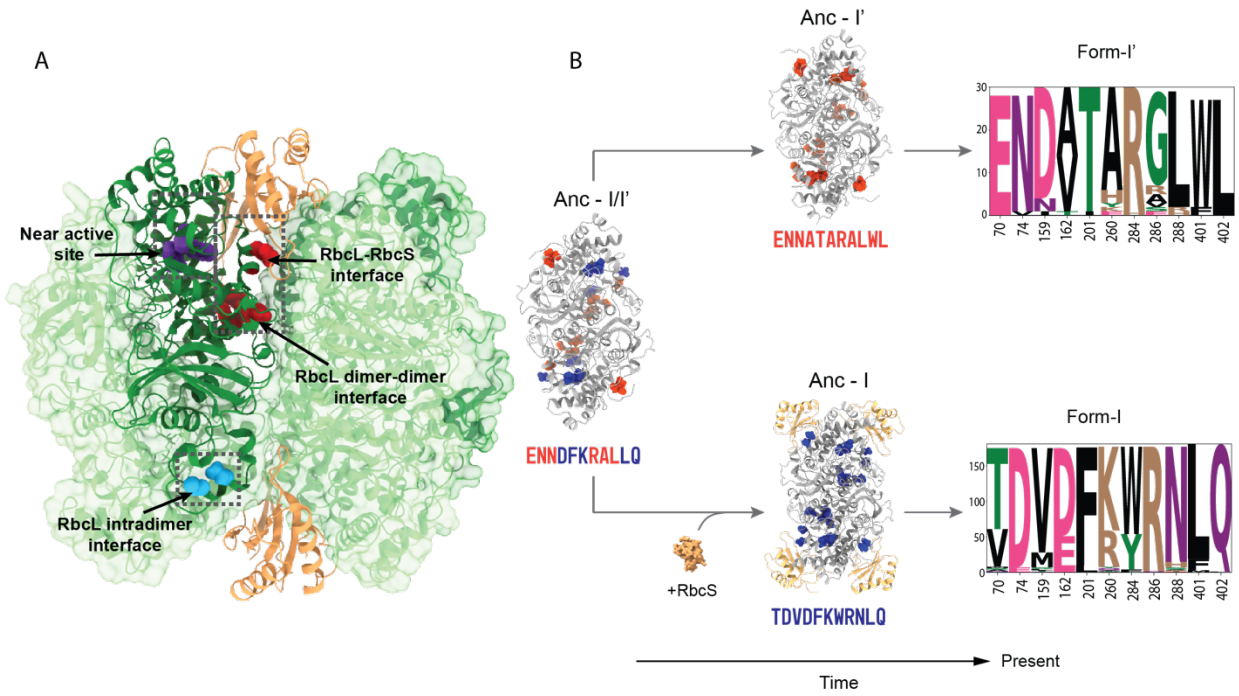

Fig. 3: Schematic representation of the structural distribution and evolution of the specialized protein "signature positions" between the RuBisCO Form-I and Form-I'. (A) Signature residues are highlighted based on their structural proximity to each other along with the region it belongs to on the *Rhodobacter sphaeroides* L_4S_2 structure (PDB: 5NV3). Residues at the signature positions located at the RbcL intradimer interface, near active site and RbcL dimer-dimer interface are highlighted in blue, purple and red respectively. (B) Evolution of the signature residues corresponding to the divergence of Form-I and Form-I' RuBisCOs. Amino acids present in the sequence motif are highlighted on the respective ancestor's L_2 or L_2S_4 structures. Sequence logos showing amino acid frequencies at each signature position across the Form-I and Form-I' clades. Residue numbering according to the *R. sphaeroides*, L denotes RbcL and S denotes RbcS.

We examined how these signature positions evolved through the emergence of RbcS by tracking their ancestral amino acid composition through the divergence of Form-I and Form-I' clades. Residues at the signature positions in Anc-I and Anc-I' most closely resemble those of their respective descendants (**Fig. 3B**). By contrast, the common ancestor of all Form-I and Form-I' RuBisCOs, Anc-I/I', has approximately equal proportions of Form-I- and Form-I'-like signature positions. Among these, signature positions at the Anc-I/I' RbcL intradimer interface (labeled

220 70,74) are Form-I'-like, whereas those near the RbcL active site (labeled 201, 401, 402) are Form-
221 I-like. Notably, the Anc-I/I' RbcL dimer-dimer interface and RbcL-RbcS interface contains a
222 mixture of Form-I- and Form-I'-like signature positions. These results identify the interface
223 residues that are likely important for oligomerization of the RuBisCO complex and underwent
224 distinct functional specialization in the presence (in Form-I) or absence (in Form-I') of RbcS.

225 In sum, the signature positions are not restricted to the RbcL-RbcS interface but are
226 scattered across the RbcL structure (**Fig. 3**). We suggest that RbcS integration resulted in distant
227 mutations, which are known to influence protein function by altering intramolecular interactions
228 and the global dynamics of protein complexes (Miton et al. 2021). We therefore submit that these
229 positions reflect how the presence or absence of RbcS affected long-range motions and
230 conformational dynamics of the RuBisCO complex, and therefore its evolutionary dynamics.

231

232 Molecular dynamics simulations of modern and ancient RuBisCOs

233 We studied the structural motions of RuBisCO complexes to investigate the influence of
234 the presence or absence of RbcS. We built the ancestral RuBisCO large and small subunit
235 structures (Methods). Modeling of the multimeric structural assembly of ancestral RuBisCO
236 oligomers was guided by extant RuBisCO crystal structures (Methods). We simulated the
237 molecular dynamics of seven ancestral and six extant RuBisCO complexes representing a range
238 of host organisms and their associated environments (**Fig. 4A, Table 1**). We performed two
239 simulation replicates for each RuBisCO variant, and data from both replicates are included in the
240 analysis.

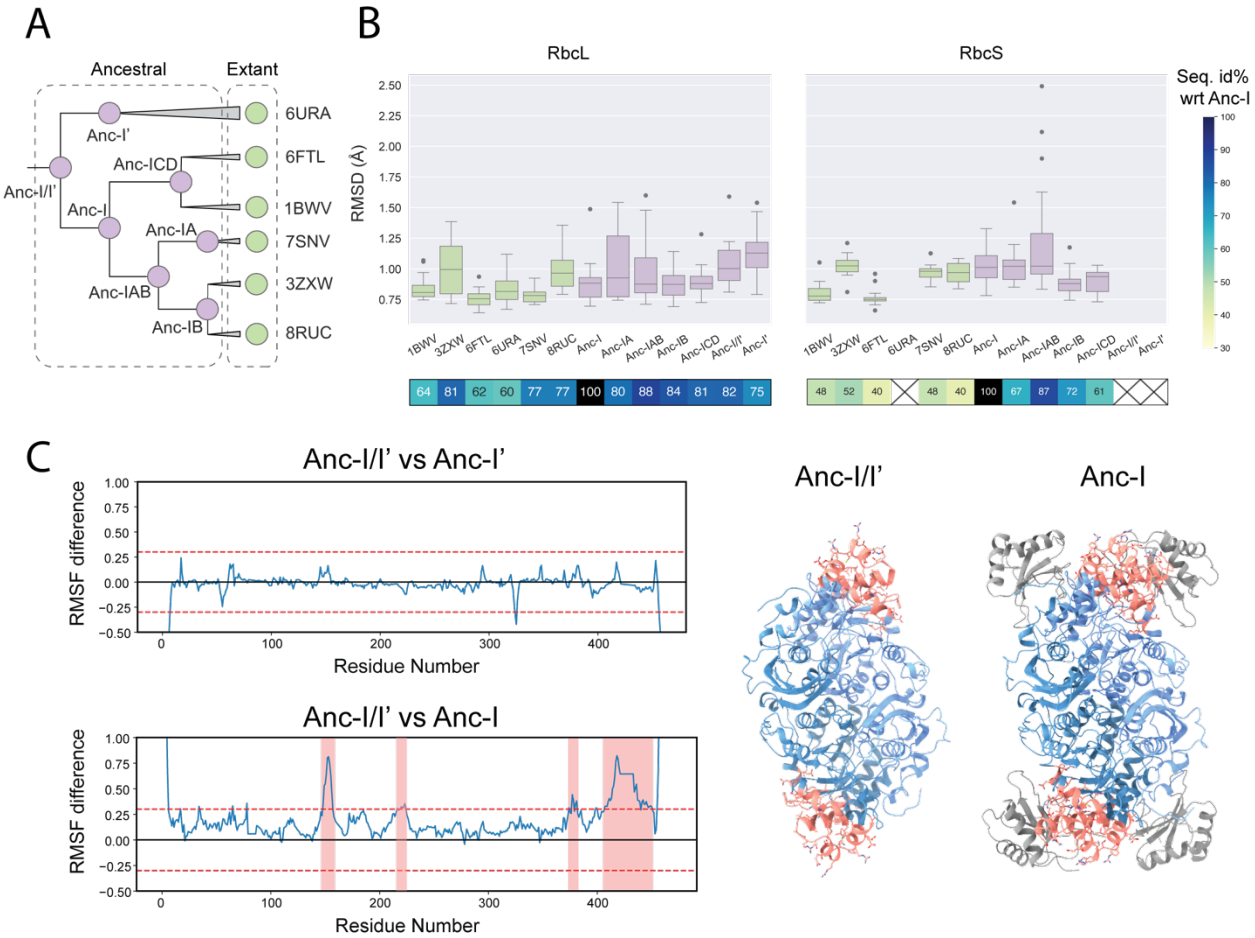

Fig. 4: Stability and fluctuation observed in the ancestral and extant RuBisCOs during MD simulations.

(A) Ancestral and extant RuBisCO homologs utilized in the study. Violet circles: ancestral RuBisCO, green circle: extant RuBisCO. (B) Average root mean square deviation (RMSD) of all amino acids in the large and small subunits throughout the simulations. Outliers outside the quartile range are represented by circles. Heatmap on the x-axis represents the pairwise sequence identity for each sequence relative to the oldest RbcS ancestor, Anc-I. (C) Difference in residue-wise root mean square fluctuation (RMSF) for RbcL between Anc-I/I' vs Anc-I' (top), and Anc-I/I' vs Anc-I (bottom). Horizontal dashed red line represents the RMSF difference of 0.3 Å, used as threshold to classify the residue fluctuation as considerable, and the horizontal black line represents the baseline when there is a zero difference in RMSF between the residues from the two proteins. Residues with RMSF over 0.3 Å are highlighted in red on the Anc-I/I' and Anc-I L_2 and L_2S_4 complex structures respectively on the right.

Table 1: Overview of extant RuBisCO structures selected for analysis.

Structure	Form	Source Organism	Environment*
1BWV (Sugawara et al. 1999)	I-CD	*Galdieria partita* (Red Algae)	High temperature, low pH
6FTL (Valegård et al. 2018)	I-CD	*Skeletonema marinoi* (Diatom)	Low temperature, arctic waters
7SNV (Blikstad et al. 2023)	I-A	*Halothiobacillus neapolitanus* (Gammaproteobacteria)	Aerobic, with alpha-carboxysome
8RUC (Andersson 1996)	I-B	*Spinacia Oleracea* (Spinach)	Mesophile, surface
3ZXW	I-B	*Thermosynechococcus vestitus* (Cyanobacteria)	Mesophile, tropics, with beta-carboxysome
6URA (Banda et al. 2020)	I'	*Candidatus Promineofilum breve* (Chloroflexi)	Anaerobic, marine sediments

258 * Environment specifies the known natural habitat or ecological niche and presence of CO_2-
259 concentrating mechanism of the source organism. All the selected structures for extant RuBisCOs
260 (except 7SNV) are complexed with 2-Carboxyarabinitol 1,5-Bisphosphate (CAP) intermediate,
261 Magnesium ion (Mg^{2+}) and the carbamylated lysine (KCX) at the active site (Methods). The 8RUC
262 structure does not include the N-terminal extension of the RbcS subunit.

263

264 We calculated the root mean square deviation (RMSD) of backbone atoms for all residues
265 in the RbcL and RbcS subunits relative to their average trajectory structure to evaluate
266 conformational differences in the complex during the simulation. The mean RMSD of all the
267 individual large and small subunits across all complexes is 0.928Å and 0.964Å, respectively. The
268 average RbcL RMSD ranges from 0.78Å to 1.13Å, while the RbcS RMSD ranges from 0.76 to
269 1.26Å (**Fig. 4B)**. The outliers in Anc-IAB small subunit RMSD values (**Fig. 4B)** reflect the N-
270 terminal residues residue protrusion and unfolding observed across the simulation replicates. We

271 performed pairwise comparisons of mean RMSD values among the different RuBisCOs. Ten out
272 of all possible pairs show a significant pairwise RMSD difference for RbcL ($p < 0.01$, Tukey post-
273 hoc test). And eight of these include the Anc-I' or Anc-I/I' complex, with a significantly higher
274 RMSD in each case (**supplementary table 3**). A higher RMSD for the Anc-I/I' and Anc-I' RbcL
275 suggests a greater structural flexibility compared to the other complexes during the simulations.

276 We used the root mean square fluctuation (RMSF) of the different residues to identify
277 region-specific differences in flexibility. Specifically, we studied the ancestral structures
278 immediately before and after divergence of RbcS-less Form-I' and RbcS-containing Form-I
279 RuBisCOs. These ancestors include Anc-I/I', Anc-I', and Anc-I. The alteration in residue-wise
280 fluctuations between the Anc-I' and Anc-I/I' complex during the MD simulation is close to zero
281 (**Fig. 4C, top**). In contrast, the Anc-I RbcL exhibits less fluctuation compared to Anc-I/I' during
282 the simulations, where four distinct sections from the large subunit sequence show considerably
283 higher fluctuations Anc-I/I' compared to Anc-I (**Fig. 4C, bottom**). The sections with higher
284 flexibility correspond to the top region of the large subunit that is positioned between the two small
285 subunits in the RuBisCO complex (**Fig. 4C, right)**. Out of the 45 RbcL residues present at the
286 RbcS interface in Anc-I, 23 residues (~51%) show a significant difference in RMSF. The
287 localization of RbcL residues near RbcS suggests that the presence of RbcS restricts the movement
288 of these residues. This observation potentially accounts for the higher stability of the Anc-I large
289 subunit compared to the RbcS-less Anc-I/I' and Anc-I'. The differences in RMSF lend support to
290 the hypothesis that RbcS plays a role in stabilizing the RbcL dimers in the octameric complex
291 (Spreitzer 2003).

292 The extant Form-I' RuBisCO (6URA) provides an additional basis for comparative
293 analysis. Along with Anc-I' and Anc-I/I', RbcS is also absent in 6URA. Unlike its ancestor,
294 however, the extant Form-I' complex does not display a significantly higher RMSD during the
295 pairwise comparison to the other RuBisCO systems (**supplementary table 3**). Similarly, 6URA
296 RbcL displays lower fluctuations per residue compared to the Anc-I/I' and Anc-I' RbcL
297 (**supplementary fig. 9**). This indicates that in the absence of RbcS, the modern Form-I' RbcL has
298 evolved a more stable L_8 without RbcS, potentially through other mutations. The L_8 assembly in
299 the extant Form-I' complex structure is maintained by a network of hydrogen bonds and salt
300 bridges (Banda et al. 2020). Some of the key residues Asp161, Trp165 and Tyr224 involved in
301 these interactions in the extant 6URA structure are not conserved in the analyzed RbcS-less
302 ancestors, Anc-I' and Anc-I/I' (Asn140, Arg144 and Phe203 respectively) (**supplementary fig.**
303 **7**). The absence of these key residue interactions may explain the higher fluctuations observed in
304 the Anc-I' and Anc-I/I' RbcLs during the simulation.

305 While RbcS is indispensable for the activity of extant Form-I RuBisCOs (Andrews 1988;
306 Lee and Tabita 1990), its present-day importance does not necessarily imply that it played an
307 essential role at its initial emergence. Protein subunits can become entrenched in complexes via
308 accumulations of neutral mutations that can be deleterious in monomers (Hochberg et al. 2020).
309 However, our analysis further indicates the early importance of RbcS for enhanced stability of the
310 RbcL octamer, along with prior experimental work demonstrating enhanced specificity towards
311 CO_2 (Schulz et al. 2022). These findings collectively suggest an early advantage of RbcS in
312 RuBisCO assembly at the time of its acquirement.

Impact of RbcS integration on the conformational variation of Form-I RbcL

The global distribution of RbcL signature positions (**Fig. 3**), inferred to be functionally linked to the presence of RbcS, suggests that RbcS integration likely affected the large subunit's overall dynamics and motion. We utilized MD simulations for RuBisCO complexes with and without RbcS to investigate the impact of RbcS on the enzyme's ability to explore conformational space. A protein's conformational space can be visualized as a set of different conformational states and the conformational variability of an enzyme is considered as an important factor for its functional diversity (Nobeli et al. 2009; Babtie et al. 2010).

To characterize the variation in conformations for the different RuBisCO simulations, we performed a principal component analysis (PCA) of the trajectory of the RbcL Cα atoms (Hayward and de Groot 2008; David and Jacobs 2014). This analysis allows us to reduce the dimensionality of the simulation trajectories into a few relevant dimensions, referred to as principal components (PC). To compare different RbcLs, we extracted the coordinates of conserved residues across all RbcLs and projected the trajectories onto a shared set of PCs (Methods). The dimensionality reduction shows the changes along the top two PCs (PC1 and PC2) for the RbcL's MD trajectories, capturing ~27% of all RbcL motions during the simulations (**Fig. 5A**). We observe that RbcL from each RuBisCO variant clusters closely together (**Fig. 5B**) suggesting that the dynamics of each individual RbcL are similar to those of its counterparts within the RuBisCO complex.

We observe that Form-I RuBisCOs (with RbcS) and non-Form-I RuBisCOs (without RbcS) occupy distinct regions along PC1 (**Fig. 5C**) with statistical significance (U-statistic = 2.09×10^7; $p = 0.0$; Mann-Whitney U test). In contrast, PC2, does not show a specific trend between

RuBisCO variants. This suggests that the primary conformational variation (~19% of all RbcL motions) corresponds to the distinction between Form-I and non-Form-I RuBisCOs.

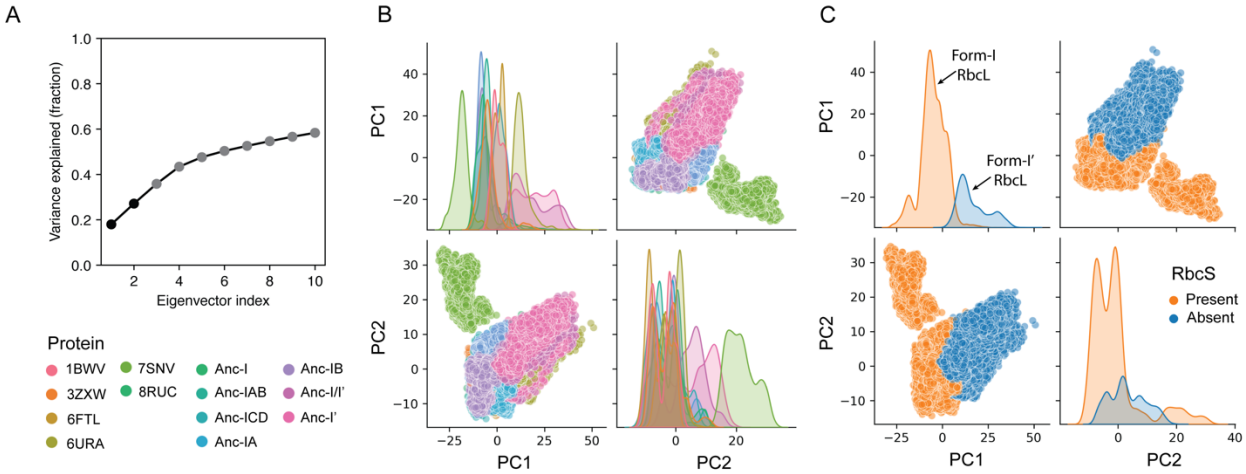

Fig. 5: Principal Component Analysis for the conserved RbcL residues of MD-simulation trajectories across ancestral and extant Form-I and Form-I' RuBisCO variants. (A) Variance explained by each eigenvector in the PCA, showing the contribution of the top 10 principal components to overall fluctuation across all simulations. (B) Pairwise representation of PC1 and PC2, highlighting the different Form-I and Form-I' RuBisCO variants. (C) Pairwise representation of PC1 and PC2, highlighting the RbcS presence and absence in the RuBisCO complex.

We simulated the dynamics of Anc-I, the common ancestor of Form-I RuBisCO, without the RbcS subunit and projected its trajectory onto the previously defined PCs for Form-I and Form-I'. The results show an intermediate positioning between Form-I and Form-I' RuBisCOs (**supplementary figure 10A**). Removing RbcS from Form-I RuBisCO, shifts it dynamics closer to Form-I', but does not fully replicate Form I' behavior. This indicates that RbcS plays a critical, though not exclusive, role in driving the distinct conformational variations between Form-I and Form-I'.

353 We assessed the impact of octameric oligomerization of RbcL on enzyme dynamics by
354 including MD simulations of a Form-II RuBisCO (9RUB) in our PCA decomposition.
355 Incorporating Form-II RbcL generates a distinct set of PCs (*PC'*). The first PC (PC1') separates
356 Form-II from Form-I and Form-I'; while the second PC (PC2') distinguishes Form-I and Form-I'
357 (**supplementary figure 10B**). This separation suggests that the octamerization in Form-I/Form-I'
358 induces significant conformational changes compared to the non-octameric Form-II, with the
359 Form-I and Form-I' distinction becoming apparent only in PC2'.

360 Our results suggest that the presence or absence of RbcS in the RuBisCO complex impacts
361 the major dynamics of RbcL. We hypothesize that the incorporation of RbcS induces a
362 conformational shift in RbcL, a modification that has remained consistent across diverse Form-I
363 and Form-I' RuBisCO variants over evolutionary time. The functional diversity (e.g., promiscuity)
364 of a protein is closely tied to its ability to explore diverse conformational states (Tokuriki and
365 Tawfik 2009; Zou et al. 2015; Petrović et al. 2018; Jackson et al. 2022). In this context, our analysis
366 shows that the RbcL explores distinct conformations in the presence of RbcS, suggesting that the
367 shift in RuBisCO's conformational variability may be directly associated with the enzyme's
368 increased CO_2 specificity following the emergence of RbcS (Schulz et al. 2022).

369 Heterogeneity of CO_2/O_2 diffusion patterns across ancestral and extant
370 RuBisCOs

371 Previous work suggested that RbcS acts as a CO_2-reservoir (Van Lun et al. 2014),
372 increasing the concentration of CO_2 molecules near the enzyme and making CO_2 more accessible
373 to the active site. This hypothesis also provides an additional functional justification for the

374 integration of RbcS, given the increase in Earth's atmospheric O_2 concentration ~2.5 billion years
375 ago, around the time of RbcS integration into the RuBisCO complex.

376 We performed MD simulations for the different RuBisCO complexes with CO_2 and O_2 gas
377 molecules present in the medium (**Fig. 6A**) to assess the interaction of RbcS residues towards both
378 gases. We estimated the relative contact-scores of the gas molecules by counting the number of
379 interactions between the gases and the protein throughout the simulation within a distance
380 threshold of 6Å. The MD simulations for O_2 and CO_2 gas molecules were performed separately
381 (**Fig. 6A**). Five independent replicates for MD simulations of one extant (8RUC) and one ancestral
382 (Anc-I) enzyme were conducted to assess the robustness and replicability of the gas dynamics
383 modeling approach. To assess the impact of gas concentrations, we performed simulations with
384 varying CO_2 and O_2 gas concentrations. The trends observed for affinity of the subunits across
385 different gas concentrations remains relatively consistent between the two subunits
386 (**supplementary fig. 11)**.

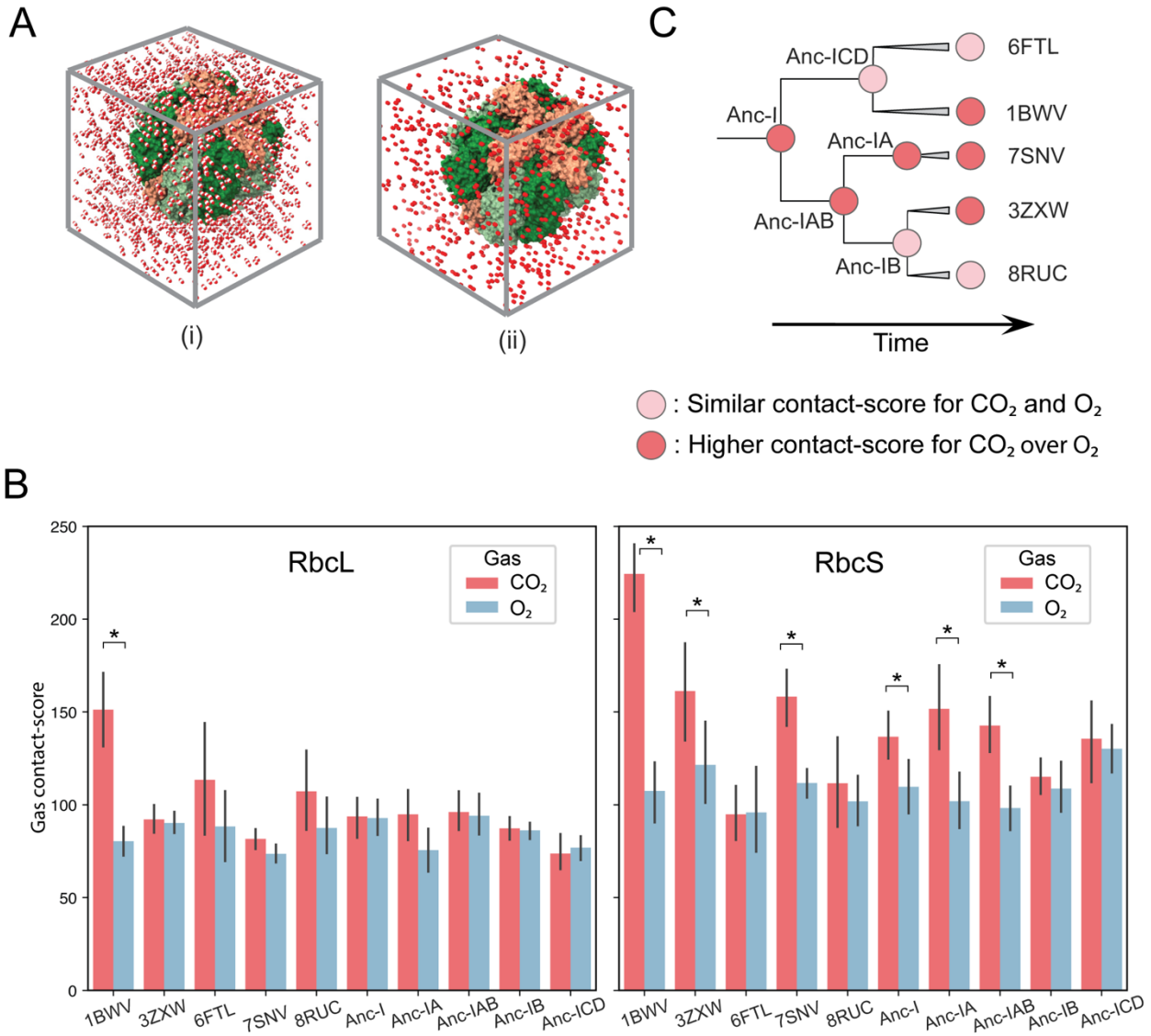

Fig. 6: CO_2 and O_2 gas diffusion molecular dynamics across different RuBisCO systems. (A) A snapshot of the MD simulation box with CO_2 (i) and O_2 (ii) gas molecules in the medium at a concentration of 400 mM with the protein complex. (B) Barplot representing the contact-score (number of gas contacts per residue) for RbcL and RbcS for each of RuBisCO for the two CO_2 and O_2 gas molecules. Error bars represent the standard deviation across the 8 subunits in the RuBisCO complex. Proteins with a significant difference between the CO_2 and O_2 affinity are marked with an asterisk (*). (C) Trend for the difference in contact-score for CO_2 and O_2 gas molecules for RbcS as observed in the barplot (B). Time is displayed along the y-axis.

397 An independent two sample t-test was performed to compare the mean CO_2 and O_2 gas
398 contacts across all simulations. For RbcL, we observe that only 1BWV has a significantly higher
399 contact-score for CO_2 over O_2 ($t = 5.92$; $p = 3.7x10^{-5}$, two-sided), whereas other complexes do not
400 show a statistically significant ($p > 0.05$) difference (**Fig. 6B**). Alternatively, six RbcS (1BWV,
401 3ZXW, 7SNV, Anc-I, Anc-IA, Anc-IAB) out of the ten exhibit a significantly ($p < 0.05$) higher
402 contact-score for CO_2 than for O_2 molecules in the surrounding medium (**Fig. 6B**, **supplementary**
403 **table 4**). RbcS that exhibit a higher affinity for CO_2 over O_2 are not confined to a specific clade in
404 the phylogenetic tree (**Fig. 6C**). We find that these results do not vary significantly over five
405 independent simulations (**supplementary fig. 12)**.

406

407 Conclusions

408 The co-evolution of the large and small subunits of RuBisCO presents an opportunity to
409 analyze molecular evolutionary events within key macroevolutionary and geochemical transitions,
410 independently documented by Earth's geologic record. We examined the historical impact of
411 RuBisCO small subunit emergence on the structural motions and evolutionary trajectory of the
412 enzyme, with specific focus on a possible adaptive role in concentrating CO_2. Both extant and
413 ancestral Form-I RuBisCO complexes that contain RbcS show higher CO_2-specificity than other
414 forms of RuBisCO lacking RbcS (Flamholz et al. 2019; Schulz et al. 2022). These observations
415 have led to the hypothesis that the improved CO_2-specificity of RuBisCO after RbcS integration
416 would have been advantageous after the GOE, given the subsequent rise in atmospheric O_2 levels
417 and decrease in CO_2 concentrations. We illustrate that not all ancestral and extant RbcS proteins

418 act as a reservoir to concentrate CO_2 around the RuBisCO complex. It is possible that certain
419 RuBisCOs with RbcS might have adopted a CO_2-reservoir strategy based on other intercellular or
420 possibly environmental factors as we find that this feature appears multiple times in the enzyme's
421 history. Moreover, our analysis highlights additional consequences of RbcS integration: increased
422 stability of Form-I RbcL and decreased flexibility compared to Form I' RbcL. The small subunit
423 integration may have shaped the enzyme's major conformational and functional variations in
424 response to significant shifts in atmospheric composition.

425 The origin of the small subunit still remains unresolved, but these observations shed light
426 on the circumstances that surrounded and likely facilitated the emergence of RbcS. Certain
427 cyanobacteria have two other proteins that are part of the carbon fixation machinery: β-
428 carboxysome structural protein (CcmM) and a RuBisCO activase-like protein (ALC). Both these
429 proteins have one or multiple domains that are "RbcS-like" and are considered homologs to RbcS
430 (Ryan et al. 2019; Wang et al. 2019; Lechno-Yossef et al. 2020). The presence of these protein
431 domains is specific to cyanobacteria. Further studies should explore whether the RbcS-like protein
432 domains could have emerged alongside or following the RbcS in response to the ancient shifts in
433 atmospheric CO_2 and O_2 levels. In summary, we show that ancestral RuBisCO dynamically
434 responded to a global environmental shift. The integration of the small subunit resulted in
435 increased rigidity, allowing it to maintain substrate specificity despite decreased substrate
436 availability. Ironically, this natural solution to an ancient challenge may have led to modern
437 RuBisCO's notorious resistance to artificial improvements in CO_2 specificity.

438

Methods

Phylogenetic Reconstruction of RuBisCO

Homologous sequences for RbcS protein from the NCBI non-redundant database were identified using PSI-BLAST (Altschul et al. 1997; Altschul and Koonin 1998). The RbcS sequence from *Thermosynechococcus elongatus* (PDB id: 2YBV) (Gubernator et al. 2008) was used as the query sequence for 5 PSI-BLAST iterations with 50,000 sequences per iteration and an E-value cutoff of 0.005. The dataset was curated to remove partial sequences and sequences with incomplete annotations. The sequences in the dataset were dereplicated at 63% amino acid identity with CD-Hit (Li and Godzik 2006) and aligned by MAFFT (Katoh et al. 2002) (default parameters) with iteration refinement over 1000 cycles. Any poorly aligned sequences were subsequently replaced with a different representative of their CD-HIT cluster.

A maximum-likelihood phylogenetic tree was constructed using IQ-Tree (Nguyen et al. 2015) with the LG+R6 evolutionary model (**supplementary fig. 1**). ModelFinder Plus (Kalyaanamoorthy et al. 2017) was used to select the best-fit evolutionary model. Branch support values were calculated using the Shimodaira–Hasegawa–like approximate likelihood-ratio test (SH-aLRT) with 1000 bootstrap replicates and 1000 ultrafast bootstrap (UFBoot) replicates optimized by nearest neighbor interchange (NNI).

RbcL sequences from the same taxa represented in our RbcS sequence dataset were identified using the NCBI Identical Protein Groups database (IPG) (https://www.ncbi.nlm.nih.gov/ipg). For RbcS entries with missing IPG RbcL sequences, Taxonomy-restricted BLASTp was used to search for the corresponding RbcL sequence. Form-I

460 RbcL sequences were combined with RbcL sequences for Form-I', II, III and IV from Banda et al.
461 (Banda et al. 2020) to build an RbcL dataset. An RbcL phylogenetic tree was constructed using
462 IQ-Tree with the LG+R9 evolutionary model (all other parameters were the same as for the RbcS
463 tree) (**supplementary fig. 2**).

464 The sequences in the RbcL dataset were aligned by MAFFT, with the same parameters as
465 those mentioned above, and were concatenated with the RbcS alignment. A maximum-likelihood
466 phylogenetic tree was constructed for the concatenated alignment using IQ-Tree, with a partition
467 model (Chernomor et al. 2016) and ModelFinder plus (Kalyaanamoorthy et al. 2017) to search and
468 implement the best-fit evolutionary models corresponding to the RbcL (LG + R9) and RbcS (LG
469 + R5) section of the concatenated alignment. In line with prior research, Form-IV "RuBisCO-like"
470 protein sequences were used to root the phylogenetic tree (Kacar et al. 2017; Banda et al. 2020;
471 Poudel et al. 2020; Camel and Zolla 2021; Schulz et al. 2022). Ancestral sequence inference for
472 the concatenated RbcL-RbcS tree was performed using PamL4.9 (LG model) (Yang 2007).
473 Reconstruction of the gaps in the ancestral sequences was performed using a binary likelihood
474 model as described by Aadland et al. (Aadland et al. 2019).

475 RbcS sequence and structural diversity analysis

476 Candidate RbcS structures (listed in **Table 1**) from Form-I subclades were used to identify
477 the different structural features in RbcS. The amino acid regions corresponding to these features
478 were identified in the multiple sequence alignment of extant RbcS. The binary heatmap indicating
479 the presence or absence of these features across all extant RbcS sequences in the phylogeny was
480 generated based on the presence of at least half of the residues in the respective region of the RbcS

481 alignment. Host taxonomy for the extant sequences was assigned using the REST API provided
482 by Ensembl (https://rest.ensembl.org/).

483 RbcL specialized signature position analysis

484 Sequence specialization between the Form-I and Form-I' extant RbcL sequences was
485 analyzed using TwinCons (Penev et al. 2021), with blosum62 as substitution matrix for score
486 calculation, and Zebra2 (Suplatov et al. 2020), using the webserver default parameters. Specialized
487 residues identified by Zebra2 and TwinCons score were then filtered based on conservation in the
488 multiple sequence alignment for Form-I and Form-I' sequences. Same steps as above were
489 followed to identify separately conserved amino acids between 131 RbcS sequences from Form-
490 IA/B/E and 59 from Form-IC/D. None of the RbcS positions had a TwinCons score < -1, resulting
491 in no signature positions.

492 Structural Modeling of RuBisCO

493 We predicted the structures of ancestral RuBisCOs using the deep-learning based
494 Colabfold software (Mirdita et al. 2022). We built our models by combining predictions and
495 structure alignments. We built the model for the RbcL-dimer with RbcS-dimer for each ancestral
496 protein with templates from PDB, structure refinement using Amber and 3 prediction recycles.
497 Low-confidence modeled terminal regions for the predicted structures were removed using a
498 pLDDT threshold of 50. The extant RuBisCO structure from *Thermosynechococcus elongatus*
499 (PDB: 2YBV) was used as a template to obtain a predicted hexadecameric complex using UCSF
500 Chimera (Pettersen et al. 2004). Anc-I' and Anc-I/I' did not have an RbcS sequence resulting in
501 an octameric complex.

502 ## RuBisCO MD simulations

503 We simulated the molecular dynamics of both extant and ancient RuBisCO complexes. We
504 selected seven ancestral structures (nodes highlighted in **Fig. 2A**) corresponding to the last
505 common ancestors of different Form-I and Form-I' clades and six experimentally determined
506 extant RuBisCO structures, representing the Form-IA, IB (two from this clade), IC/D (two from
507 this clade), and I' RuBisCO clades. PDB IDs for these entries are: 1BWV (red algae), 3ZXW
508 (green algae), 6FTL (diatom), 6URA (chloroflexi), 7SNV (gammaproteobacteria) and 8RUC
509 (plantae) (more details in **Table 1**). The 6URA entry represents the first case discovered of a Form-
510 I-like RuBisCO without a small subunit. This selection represents the biological diversity of the
511 RuBisCO enzyme while preserving the following attributes: high resolution (under 3.0 Å),
512 carbamylation of the lysine at the active site (except for 7SNV) (Stec 2012), and CABP and
513 magnesium at the active site. Sequence identity and alignments for ancestral and extant RuBisCOs
514 analyzed in the study are presented in **supplementary fig. 6 and 7**, respectively.

515

516 We obtained parameters for the CABP molecule through the standard Amber protocol for
517 ligands (Wang et al. 2004). RuBisCO proteins include different post-translational modifications
518 that required additional steps. We computed the CABP and the carbamylated lysine charges at
519 HF/6-31*G level, as specified in the Amber protocol. The post-translational modifications found
520 in 6FTL were modeled with semi-empirical charges. In the case of the modeled ancient structures,
521 we manually added both ligands (CABP), ions (Mg^{2+}), and the post-translation modification at the
522 carbamylated lysine (also for 7SNV).

523 We built the topology and parameters of the molecular dynamics systems using tleap
524 (Maier et al. 2015) and Amber14 (Salomon-Ferrer et al. 2013). Periodic solvation boxes were
525 constructed with 10 Å spacing and water molecules according to the TIP3P model (Jorgensen et
526 al. 1983). Sodium and chloride ions counterbalanced the charge of the system. The particle-mesh
527 Ewald summation method was used for long-range electrostatics and a 10 Å cutoff was set for
528 short-range non-bonded interactions. Initial geometries in all systems were minimized at 5,000
529 conjugate-gradient steps after which water was equilibrated at 298 K and 1 atm for 100 ps at 2 fs
530 time steps. Production runs were then performed for 250 ns in the NPT ensemble at P = 1 atm and
531 T = 298 K. We used the hydrogen mass partitioning method to ensure the stability of the simulation
532 under large integration steps, thereby increasing the simulation speed. Langevin dynamics for T
533 control and Nosé-Hoover Langevin piston method for P control were used. We carried out the MD
534 simulation on OpenMM 7.7 (Eastman et al. 2017) running in the Nvidia Tesla A100, L40 and
535 H100 GPU nodes of the Center for High Throughput Computation at the University of Wisconsin-
536 Madison. We performed two replicates of MD simulations. The equilibration of simulations across
537 different complexes was evaluated by tracking the RMSD of the complex structure over time for
538 both replicates (**supplementary fig. 8**). The RMSD across all RuBisCO complexes stabilizes at a
539 plateau, indicating that the complexes reached an equilibrated state.

540 Gas diffusion MD simulations

541 We obtained CO_2 and O_2 parameters through the standard Amber protocol. We computed
542 the charges of these molecules at the HF/6-31*G level through the RESP protocol (Woods and
543 Chappelle 2000). The embedding of proteins in solutions with gas molecules in dissolution
544 consisted of replacing water molecules from a previously solvated system to avoid clashes among

545 the gas molecules. We calculated the number of molecules corresponding to a given concentration
546 using the Amber-recommended method for determining ion concentrations in a solvent. All other
547 simulation aspects were consistent with those described in the previous section. These simulations
548 were run for 75 ns, with the final 50 ns used for analysis following a 25 ns equilibration period.

549 Analysis of MD simulations

550 We employed MDAnalysis (Michaud-Agrawal et al. 2011) and ProDy (Bakan et al. 2011)
551 to analyze and process the outcome of the MD simulations. All simulations were conducted under
552 identical temperature, pH, and solvent conditions. The initial 50 ns of the 250 ns simulations were
553 designated for equilibration, and only the final 200 ns were used for analysis. Data from both MD
554 simulation replicates were included, resulting in a combined simulation time exceeding 7 μs. The
555 average system size was approximately 300,000 atoms, including solvent molecules.

556 RMSD and RMSF calculations

557 We aligned all the simulation trajectories to their average structure and calculated the
558 RMSD across all Cα backbone atoms for each RbcL and RbcS subunit using MDAnalysis. The
559 boxplot in Figure 4B presents the average RMSD value for each large and small subunit across the
560 two simulation replicates. RMSF was calculated by aligning the MD trajectories to their reference
561 structure first. The Cα atomic positions of each residue were mapped onto the respective sequence,
562 and fluctuations were calculated as the average positional deviation from the mean structure over
563 the simulation period across the eight RbcL subunits using MDAnalysis. Simulation data from
564 replicates were included in the calculations of the RMSD and RMSF.

565 ## Principal component analysis

566 The principal component analysis helps reduce the complexity of MD simulation data by
567 generating orthogonal eigenvectors (principal components, PCs) that represent the primary axes
568 of motion. We conducted PC analyses using the scikit-learn implementation (Pedregosa et al.
569 2011). Specifically, we first extracted all conserved residues across select RbcL sequences shown
570 in **Fig. 4A**. All the individual trajectories for each RbcLs were then aligned onto the reference
571 extant RbcL structure (1BWV). The covariance matrix of positional fluctuations was constructed
572 exclusively for the conserved RbcL Cα atoms. Eigenvalues and eigenvectors were derived from
573 this covariance matrix to identify the PCs shared across all RbcL trajectories, including each
574 simulation replicate. PCA, including the Anc-I without RbcS (**supplementary fig. 10A**), was
575 conducted by extracting trajectories for conserved residues and projecting them onto the PC-space
576 defined by Form-I and Form-I' RuBisCOs in **Fig. 5**. Additionally, a separate PCA was performed
577 following the same procedure, incorporating RbcL simulation trajectories from Form-II alongside
578 Form-I and Form-I', as shown in **supplementary fig. 10B**.

579

580 # Data and resource availability

581 The data underlying this article along with the code for analysis can be found at
582 https://github.com/kacarlab/RuBisCO_evolution. Molecular dynamics simulation files for the
583 extant and ancestral RuBisCOs are available at Zenodo (https://zenodo.org/records/14187581).

584

Acknowledgements

We would like to thank Alessandro Senes, Phil Romero, Tina Wang and Srivatsan Raman for their feedback on the study, Aya Klos, Evrim Fer, Josh MacCready, Zach Adam and Amanda Garcia for critical reading and useful comments and suggestions on the manuscript. The Center for High Throughput Computing (CHTC) at the University of Wisconsin-Madison provided computing resources for this work. This work was supported by the Human Frontier Science Program (HFSP) [*RGY0072/2021*] with additional support from the NASA Exobiology [*NNH23ZDA001N*] and ICAR Programs [*80NSSC17K0296*]. B.C.Z acknowledges the Margarita Salas Postdoctoral Fellowship, founded by the Unión Europea - Next Generation EU (B.C.Z.; UP2021-035).

Author Contribution

Conceptualization: KA, BCZ and BK; Methodology: KA and BCZ; Data curation: KA and BCZ; Data analysis: KA, BCZ and BK; Figures: KA and BCZ; First draft: KA; Resources and supervision: BK. All authors edited and approved the final draft.

Conflict of Interest

The authors declare no competing interests.

References

Aadland K, Pugh C, Kolaczkowski B. 2019. High-Throughput Reconstruction of Ancestral Protein Sequence, Structure, and Molecular Function. In: Sikosek T, editor. Computational Methods in Protein Evolution. New York, NY: Springer. p. 135–170. Available from: https://doi.org/10.1007/978-1-4939-8736-8_8

Altschul SF, Koonin EV. 1998. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.* 23:444–447.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Andersson I. 1996. Large Structures at High Resolution: The 1.6 Å Crystal Structure of Spinach Ribulose-1,5- Bisphosphate Carboxylase/Oxygenase Complexed with 2-Carboxyarabinitol Bisphosphate. *J. Mol. Biol.* 259:160–174.

Andersson I. 2008. Catalysis and regulation in Rubisco. *J. Exp. Bot.* 59:1555–1568.

Andrews TJ. 1988. Catalysis by cyanobacterial ribulose-bisphosphate carboxylase large subunits in the complete absence of small subunits. *J. Biol. Chem.* 263:12213–12219.

Ashida H, Danchin A, Yokota A. 2005. Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism? *Res. Microbiol.* 156:611–618.

Babtie A, Tokuriki N, Hollfelder F. 2010. What makes an enzyme promiscuous? *Curr. Opin. Chem. Biol.* 14:200–207.

Bakan A, Meireles LM, Bahar I. 2011. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* 27:1575–1577.

Banda DM, Pereira JH, Liu AK, Orr DJ, Hammel M, He C, Parry MAJ, Carmo-Silva E, Adams PD, Banfield JF, et al. 2020. Novel bacterial clade reveals origin of form I Rubisco. *Nat. Plants* 6:1158–1166.

Blikstad C, Dugan EJ, Laughlin TG, Turnšek JB, Liu MD, Shoemaker SR, Vogiatzi N, Remis JP, Savage DF. 2023. Identification of a carbonic anhydrase–Rubisco complex within the alpha-carboxysome. *Proc. Natl. Acad. Sci.* 120:e2308600120.

Bouvier JW, Emms DM, Kelly S. 2024. Rubisco is evolving for improved catalytic efficiency and CO2 assimilation in plants. *Proc. Natl. Acad. Sci.* 121:e2321050121.

Bracher A, Whitney SM, Hartl FU, Hayer-Hartl M. 2017. Biogenesis and Metabolic Maintenance of Rubisco*. *Httpsdoiorg101146annurev-Arplant-043015-111633* 68:29–60.

636 Caetano-Anollés G. 2017. RubisCO and the Search for Biomolecular Culprits of Planetary
637 Change. *BioEssays* 39:1700174.

638 Camel V, Zolla G. 2021. An Insight of RuBisCO Evolution through a Multilevel Approach.
639 *Biomolecules* 11:1761–1761.

640 Chernomor O, Von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for
641 Phylogenomic Inference from Supermatrices. *Syst. Biol.* 65:997–1008.

642 David CC, Jacobs DJ. 2014. Principal Component Analysis: A Method for Determining the
643 Essential Dynamics of Proteins. *Methods Mol. Biol. Clifton NJ* 1084:193.

644 Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP,
645 Simmonett AC, Harrigan MP, Stern CD, et al. 2017. OpenMM 7: Rapid development of
646 high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* 13:e1005659–
647 e1005659.

648 Erb TJ, Zarzycki J. 2018. A short history of RubisCO: the rise and fall (?) of Nature's
649 predominant CO2 fixing enzyme. *Curr. Opin. Biotechnol.* 49:100–107.

650 Esquivel MG, Genkov T, Nogueira AS, Salvucci ME, Spreitzer RJ. 2013. Substitutions at the
651 opening of the Rubisco central solvent channel affect holoenzyme stability and CO2/O2
652 specificity but not activation by Rubisco activase. *Photosynth. Res.* 118:209–218.

653 Falkowski PG, Fenchel T, Delong EF. 2008. The Microbial Engines That Drive Earth's
654 Biogeochemical Cycles. *Science* 320:1034–1039.

655 Fernie AR, Bauwe H. 2020. Wasteful, essential, evolutionary stepping stone? The multiple
656 personalities of the photorespiratory pathway. *Plant J.* 102:666–677.

657 Flamholz AI, Prywes N, Moran U, Davidi D, Bar-On YM, Oltrogge LM, Alves R, Savage D,
658 Milo R. 2019. Revisiting Trade-offs between Rubisco Kinetic Parameters. *Biochemistry*
659 58:3365–3376.

660 Garcia AK, Cavanaugh CM, Kacar B. 2021. The curious consistency of carbon biosignatures
661 over billions of years of Earth-life coevolution. *ISME J.* 15:2183–2194.

662 Gatenby AA. 1988. Synthesis and assembly of bacterial and higher plant Rubisco subunits in
663 Escherichia coli. *Photosynth. Res.* 17:145–157.

664 Genkov T, Meyer M, Griffiths H, Spreitzer RJ. 2010. Functional Hybrid Rubisco Enzymes with
665 Plant Small Subunits and Algal Large Subunits: ENGINEERED rbcS cDNA FOR
666 EXPRESSION IN CHLAMYDOMONAS. *J. Biol. Chem.* 285:19833–19841.

667 Gubernator B, Bartoszewski R, Kroliczewski J, Wildner G, Szczepaniak A. 2008. Ribulose-1,5-
668 bisphosphate carboxylase/oxygenase from thermophilic cyanobacterium
669 Thermosynechococcus elongatus. *Photosynth. Res.* 95:101–109.

670 Hayward S, de Groot BL. 2008. Normal Modes and Essential Dynamics. In: Kukol A, editor.
671 Molecular Modeling of Proteins. Methods Molecular Biology™. Totowa, NJ: Humana
672 Press. p. 89–106. Available from: https://doi.org/10.1007/978-1-59745-177-2_5

673 Hochberg GKA, Liu Y, Marklund EG, Metzger BPH, Laganowsky A, Thornton JW. 2020. A
674 hydrophobic ratchet entrenches molecular complexes. *Nature* 588:503–508.

675 Jackson C, Toth-Petroczy A, Kolodny R, Hollfelder F, Fuxreiter M, Kamerlin SCL, Tokuriki N.
676 2022. Adventures on the Routes of Protein Evolution—In Memoriam Dan Salah Tawfik
677 (1955–2021). *J. Mol. Biol.* 434:167462–167462.

678 Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. 1983. Comparison of
679 simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

680 Joshi J, Mueller-Cajar O, Tsai Y-CC, Hartl FU, Hayer-Hartl M. 2015. Role of Small Subunit in
681 Mediating Assembly of Red-type Form I Rubisco. *J. Biol. Chem.* 290:1066–1074.

682 Kacar B, Hanson-Smith V, Adam ZR, Boekelheide N. 2017. Constraining the timing of the
683 Great Oxidation Event within the Rubisco phylogenetic tree. *Geobiology* 15:628–640.

684 Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder:
685 fast model selection for accurate phylogenetic estimates. *Nat. Methods 2017 146* 14:587–
686 589.

687 Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: a novel method for rapid multiple
688 sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.

689 Kędzior M, Garcia AK, Li M, Taton A, Adam ZR, Young JN, Kaçar B. 2022. Resurrected
690 Rubisco suggests uniform carbon isotope signatures over geologic time. *Cell Rep.*
691 39:110726.

692 Knight S, Andersson I, Br~ind~n C-I. 1990. Crystallographic Analysis of Ribulose 1,5-
693 Bisphosphate Carboxylase from Spinach at 2.4 A Resolution Subunit Interactions and
694 Active Site.

695 Lechno-Yossef S, Rohnke BA, Belza ACO, Melnicki MR, Montgomery BL, Kerfeld CA. 2020.
696 Cyanobacterial carboxysomes contain an unique rubisco-activase-like protein. *New Phytol.*
697 225:793–806.

698 Lee B, Tabita FR. 1990. Purification of recombinant ribulose-1,5-bisphosphate
699 carboxylase/oxygenase large subunits suitable for reconstitution and assembly of active
700 L8S8 enzyme. *Biochemistry* 29:9352–9357.

701 Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein
702 or nucleotide sequences. *Bioinformatics* 22:1658–1659.

703 Liu AK, Pereira JH, Kehl AJ, Rosenberg DJ, Orr DJ, Chu SKS, Banda DM, Hammel M, Adams
704 PD, Siegel JB, et al. 2022. Structural plasticity enables evolution and innovation of

RuBisCO assemblies. *Sci. Adv.* [Internet] 8. Available from: https://www.science.org/doi/10.1126/sciadv.adc9440

Liu C, Young AL, Starling-Windhof A, Bracher A, Saschenbrecker S, Rao BV, Rao KV, Berninghausen O, Mielke T, Hartl FU, et al. 2010. Coupled chaperone action in folding and assembly of hexadecameric Rubisco. *Nature* 463:197–202.

Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* 11:3696–3713.

Mao Y, Catherall E, Díaz-Ramos A, Greiff GRL, Azinas S, Gunn L, McCormick AJ. 2022. The small subunit of Rubisco and its potential as an engineering target.Sharwood R, editor. *J. Exp. Bot.* [Internet]. Available from: https://academic.oup.com/jxb/advance-article/doi/10.1093/jxb/erac309/6645828

Matsumura H, Shiomi K, Yamamoto A, Taketani Y, Kobayashi N, Yoshizawa T, Tanaka S ichi, Yoshikawa H, Endo M, Fukayama H. 2020. Hybrid Rubisco with Complete Replacement of Rice Rubisco Small Subunits by Sorghum Counterparts Confers C4 Plant-like High Catalytic Activity. *Mol. Plant* 13:1570–1581.

Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. 2011. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* 32:2319–2327.

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nat. Methods* 19:679–682.

Miton CM, Buda K, Tokuriki N. 2021. Epistasis and intramolecular networks in protein evolution. *Curr. Opin. Struct. Biol.* 69:160–168.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* 32:268–274.

Nobeli I, Favia AD, Thornton JM. 2009. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* 27:157–167.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12:2825–2830.

Penev PI, Alvarez-Carreño C, Smith E, Petrov AS, Williams LD. 2021. TwinCons: Conservation score for uncovering deep sequence similarity and divergence. *PLOS Comput. Biol.* 17:e1009541–e1009541.

Petrović D, Risso VA, Kamerlin SCL, Sanchez-Ruiz JM. 2018. Conformational dynamics and enzyme evolution. *J. R. Soc. Interface* 15:20180330.

740 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004.
741 UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput.*
742 *Chem.* 25:1605–1612.

743 Poudel S, Pike DH, Raanan H, Mancini JA, Nanda V, Rickaby C REM, Falkowski PG. 2020.
744 Biophysical analysis of the structural evolution of substrate specificity in RuBisCO. *PNAS.*

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

747 Rucker HR, Kaçar B. 2023. Enigmatic evolution of microbial nitrogen fixation: insights from
748 Earth's past. *Trends Microbiol.* [Internet] 0. Available from:
749 http://www.cell.com/article/S0966842X23000914/fulltext

750 Ryan P, Forrester TJB, Wroblewski C, Kenney TMG, Kitova EN, Klassen JS, Kimber MS. 2019.
751 The small RbcS-like domains of the β-carboxysome structural protein CcmM bind
752 RubisCO at a site distinct from that binding the RbcS subunit. *J. Biol. Chem.* 294:2593–
753 5195.

754 Salomon-Ferrer R, Case DA, Walker RC. 2013. An overview of the Amber biomolecular
755 simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3:198–210.

756 Schmidt GW, Mishkind ML. 1986. THE TRANSPORT OF PROTEINS INTO
757 CHLOROPLASTS. Available from: www.annualreviews.org

758 Schulz L, Guo Z, Zarzycki J, Steinchen W, Schuller JM, Heimerl T, Prinz S, Mueller-Cajar O,
759 Erb TJ, Hochberg GKA. 2022. Evolution of increased complexity and specificity at the
760 dawn of form I Rubiscos. *Science* 378:155–160.

761 Spreitzer RJ. 2003. Role of the small subunit in ribulose-1,5-bisphosphate
762 carboxylase/oxygenase. *Arch. Biochem. Biophys.* 414:141–149.

763 Stec B. 2012. Structural mechanism of RuBisCO activation by carbamylation of the active site
764 lysine. *Proc. Natl. Acad. Sci. U. S. A.* 109:18785–18790.

765 Sugawara H, Yamamoto H, Shibata N, Inoue T, Okada S, Miyake C, Yokota A, Kai Y. 1999.
766 Crystal Structure of Carboxylase Reaction-oriented Ribulose 1,5-Bisphosphate
767 Carboxylase/Oxygenase from a Thermophilic Red Alga, *Galdieria partita* *. *J. Biol. Chem.*
768 274:15655–15661.

769 Suplatov D, Sharapova Y, Geraseva E, Švedas V. 2020. Zebra2: advanced and easy-to-use web-
770 server for bioinformatic analysis of subfamily-specific and conserved positions in diverse
771 protein superfamilies. *Nucleic Acids Res.* 48:W65–W71.

772 Tabita FR, Hanson TE, Li H, Satagopan S, Singh J, Chan S. 2007. Function, Structure, and
773 Evolution of the RubisCO-Like Proteins and Their RubisCO Homologs. *Microbiol. Mol.*
774 *Biol. Rev.* 71:576–599.

775 Tabita FR, Satagopan S, Hanson TE, Kreel NE, Scott SS. 2008. Distinct form I, II, III, and IV
776 Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and
777 structure/function relationships. *J. Exp. Bot.* 59:1515–1524.

778 Tokuriki N, Tawfik DS. 2009. Protein Dynamism and Evolvability. *Science* 324:203–207.

779 Valegård K, Andralojc PJ, Haslam RP, Pearce FG, Eriksen GK, Madgwick PJ, Kristoffersen AK,
780 van Lun M, Klein U, Eilertsen HC, et al. 2018. Structural and functional analyses of
781 Rubisco from arctic diatom species reveal unusual posttranslational modifications. *J. Biol.*
782 *Chem.* 293:13033–13043.

783 Van Lun M, Hub JS, Van Der Spoel D, Andersson I. 2014. CO2 and O2 distribution in rubisco
784 suggests the small subunit functions as a CO2 reservoir. *J. Am. Chem. Soc.* 136:3165–3171.

785 Van Lun M, Van Der Spoel D, Andersson I. 2011. Subunit interface dynamics in hexadecameric
786 Rubisco. *J. Mol. Biol.* 411:1083–1098.

787 Wagner A, Rosen W. 2014. Spaces of the possible: universal Darwinism and the wall between
788 technological and biological innovation. *J. R. Soc. Interface* 11:20131190.

789 Wang H, Yan X, Aigner H, Bracher A, Nguyen ND, Hee WY, Long BM, Price GD, Hartl FU,
790 Hayer-Hartl M. 2019. Rubisco condensate formation by CcmM in β-carboxysome
791 biogenesis. *Nat. 2019 5667742* 566:131–135.

792 Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. 2004. Development and testing of a
793 general amber force field. *J. Comput. Chem.* 25:1157–1174.

794 Wang M, Jiang YY, Kim KM, Qu G, Ji HF, Mittenthal JE, Zhang HY, Caetano-Anollés G. 2011.
795 A universal molecular clock of protein folds and its power in tracing the early history of
796 aerobic metabolism and planet oxygenation. *Mol. Biol. Evol.* 28:567–582.

797 Ward LM, Shih PM. 2019. The evolution and productivity of carbon fixation pathways in
798 response to changes in oxygen concentration over geological time. *Free Radic. Biol. Med.*
799 140:188–199.

800 West-Roberts JA, Matheus-Carnevali PB, Schoelmerich MC, Al-Shayeb B, Thomas AD, Sharrar
801 A, He C, Chen L-X, Lavy A, Keren R, et al. 2021. The Chloroflexi supergroup is
802 metabolically diverse and representatives have novel genes for non-photosynthesis based
803 CO2 fixation. *bioRxiv*:2021.08.23.457424-2021.08.23.457424.

804 Woods RJ, Chappelle R. 2000. Restrained electrostatic potential atomic partial charges for
805 condensed-phase simulations of carbohydrates. *J. Mol. Struct. THEOCHEM* 527:149–156.

806 Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.*
807 24:1586–1591.

808 Young JN, Rickaby REM, Kapralov MV, Filatov DA. 2012. Adaptive signals in algal Rubisco
809 reveal a history of ancient atmospheric carbon dioxide. *Philos. Trans. R. Soc. B Biol. Sci.*
810 367:483–492.

811 Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB. 2015. Evolution of Conformational
812 Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist
813 Enzyme. *Mol. Biol. Evol.* 32:132–143.

814