Ancestral structure prediction reveals the conformational impact of the RuBisCO small subunit across time

Kaustubh Amritkar[1,2], Bruno Cuevas-Zuviría[1,3] and Betül Kaçar[1,4]

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
4. correspondence: bkacar@wisc.edu, 608-263-3622

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 variability of the large subunit early in its history, constraining 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 [1]. 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 [2]. 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) [3]. In the presence of oxygen (O_2), RuBisCO also catalyzes a competing photorespiration reaction in which RuBP combines with O_2, producing an autoinhibitory metabolite detrimental to the overall metabolic efficiency of carbon fixation [4]. While the exact age of RuBisCO is not clearly known, paleobiological inferences [5–7] 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) [8–11]. 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 [12–14].

Extant RuBisCO forms (I to IV) exhibit different dimeric and poly-dimeric assemblies. All forms have large subunits (RbcL), but Form-I RuBisCOs uniquely have an additional small subunit (RbcS). Form-I RuBisCOs are also the most abundant [15] and generally exhibit higher specificities for CO_2 [16]. These attributes make RbcS a promising candidate to study how small accessory subunits can regulate the evolution of RuBisCO, though their precise evolutionary role remains unclear. RbcS is thought to be crucial for the assembly of the RbcL octameric complex made of RbcL homodimers [17–19] and significantly impacts RuBisCO's catalytic parameters [17,20–24]. For example, incorporation of ancestral RbcS in RuBisCO was shown to increase the enzyme's specificity for CO_2, indicating that emergence of RbcS played a critical role in the

ancestor of Form-I RuBisCO [25]. Previous computational work also suggested that RbcS can act as a "CO_2-reservoir" to concentrate CO_2 molecules within the enzyme [26]. However, the extent to which the small subunit has influenced the structural dynamics and evolutionary trajectory of the Form-I RuBisCO complex remains less understood.

Here we used phylogenetic reconstructions, structural predictions as well as molecular dynamics (MD) simulations to track the evolution of the small subunit in the context of the RuBisCO complex over geologic time. Our exploration focused on the hypothesis that the molecular and structural innovations likely made Form-I RuBisCO selectively more advantageous following the planetary rise of O_2. We specifically focused on three areas: 1) the sequence and structural evolution of subunits within ancient RuBisCO complexes following integration of ancient RbcS, 2) the impact of RbcS presence or absence on the structural dynamics of ancestral and extant RuBisCO complexes, and 3) migration of CO_2 and O_2 gases in ancestral RuBisCO to assess a proposed CO_2-reservoir role for ancient small subunits [26].

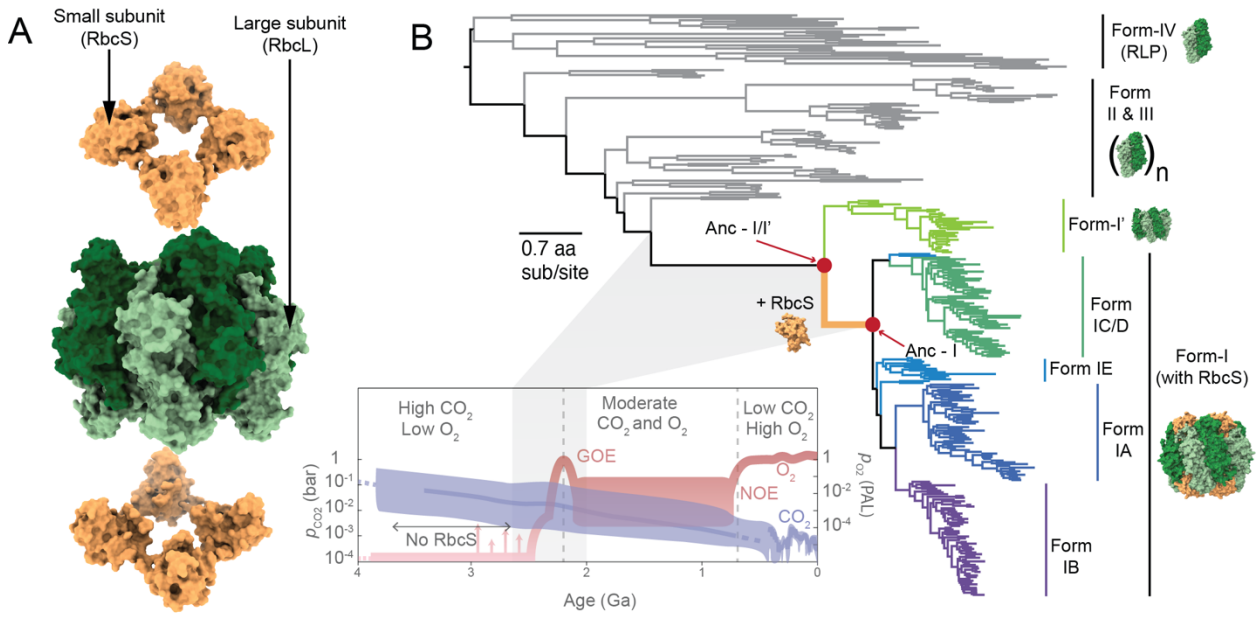


Figure 1: The evolutionary history and diversification of RuBisCO. (A) Architecture of the L_8S_8 structure of the Form-I RuBisCO complex, with RbcL shown in green and RbcS in orange (PDB:1BWV). (B) RuBisCO phylogenetic tree highlighting the emergence of RbcS coinciding with the Great Oxidation Event (GOE). Schematic on the right shows variation in the multimeric structure of the RuBisCO complex across different forms. Form-II and III RuBisCOs exhibit multiple oligomeric states of the RbcL-dimer, denoted 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.

Results and Discussion

Resurrection of ancestral RuBisCOs

We built a maximum-likelihood phylogenetic tree from a concatenated alignment of RbcL and RbcS amino acid sequences representing RuBisCO Forms I to IV, including the recently described Form-I' [27] (**Figure 2A**). The phylogeny contains 194 sets of RbcL-RbcS homologs from Form-I and 135 RbcL homologs from all other forms (including Form-I', II, II/III, III and IV). The tree contains sequences representative of known RuBisCO diversity and is rooted by Form-IV RuBisCO-like proteins in accordance with previous studies [8,28,29]. Form-I is categorized into four major subgroups: the "green-like" Form-IA and IB, prevalent in proteobacteria, cyanobacteria, green algae and plants, and the "red-like" Form-IC and ID, prevalent in proteobacteria and non-green algae [15,20]. Our phylogeny also resolves a Form-I subclade, Form-IE, that diverges before the last common ancestor of Form-IA and Form-IB [30]. Most sequences from the Form-IE clade are from metagenomic studies and belong to unclassified members of the *Chloroflexota* bacterial phylum. Form-I' RuBisCOs cluster within a monophyletic clade sister to all other Form-I sequences. Form-I' RuBisCOs are notable because their RbcL subunits are similar to those of Form-I in their multimeric assembly, but they lack the RbcS [27]. Like 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 [8,27].

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 **Figure 2A**, inferring a total of seven ancestral nodes along the phylogenetic tree.

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 (**Figure 2B,** Methods). Extant Form-I RbcS homologs show remarkably low mean pairwise sequence identity (29%) compared to Form-I RbcL sequences (66%) (**Figure S3**) as observed by others previously [23,31,32].

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 (**Figure S4A**). 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.

We further explored historical, evolutionary variation of RbcS by identifying known, functionally significant sequence and structural features [20,23] (**Figure 2C**) and mapping their presence across extant and reconstructed ancestral RbcS in our phylogenetic tree (**Figure S5**). We find that the RbcS structure contains novel features unique to different Form-I specific lineages. While all RuBisCO small subunits share a common core structure consisting of a β-sheet with four antiparallel strands (βA to βD) and two α-helices (αA & αB) [33], the Form-IB and Form-IC/D RbcS exhibit additional distinct characteristics.

Within the Form-IB clade, we find that most green-like eukaryotic RbcS homologs contain an additional N-terminal region as well as a βA-βB loop insertion. The N-terminal region is specifically known to be a signal peptide necessary for the entry of RbcS into the chloroplast, prior to assembly with RbcL [20,34]. The N-terminal extension feature is first observed in the Form-IB ancestry after the separation of cyanobacteria in the phylogeny (**Figure S5**). Absence of the N-terminal extension in Anc-IB, the common ancestor of Form-IB, as well as older ancestors indicates that their ancient hosts did not translate RbcS and RbcL in separate organelles. Rather, expression of subunits in different organelles emerged as a trait following the bifurcation of cyanobacteria and Viridiplantae in Form-IB RuBisCO.

134 We find that the βA-βB loop insertion is similarly exclusive to green-like eukaryotic RbcS
135 (**Figure 2C**). Although the significance of the βA-βB loop in RbcS is not established, it plays an
136 important part in regulating the size of the central solvent channel in RuBisCO complex [17]. Like
137 N-terminal extension, this loop first emerges after the splitting of cyanobacterial and eukaryotic
138 lineages in Form-IB (**Figure S5**). As observed with extant RbcS [20,23], the βA-βB loop insertion
139 length also varies across Form-IB ancestors (10 to 28 aa). The loop has increased by ~11 residues
140 in all eukaryotic ancestors after the divergence from cyanobacterial ancestors and by ~18 residues
141 in the ancestors from green algae lineage within Form-IB (**Figure S5**).

142 We also tracked the evolution of the C-terminal βE and βF hairpin loop that is unique to
143 Form-ICD RbcS sequences (**Figure 2B**). The common ancestor of Form-IC/D sequences, Anc-
144 ICD, and all subsequent Form-IC/D ancestors have the βE and βF hairpin (**Figure S6**). By contrast,
145 Anc-I RbcS does not have the βE and βF hairpin, suggesting this insertion happened after the
146 divergence of Form-IC/D within the Form-I clade. The C-terminal β-hairpin is functionally
147 significant for mediating the assembly of the oligomeric complex, allowing red-like RuBisCOs to
148 assemble without specialized assembly chaperones [18]. The incorporation of the hairpin loop in
149 Form-IC/D common ancestor, Anc-IC/D, suggests that independence from assembly chaperones
150 may have first emerged in a red-like RuBisCO ancestor. Our analysis thus reveals that RbcS has
151 undergone different clade-specific structural variations since its emergence in the Form-I ancestor.

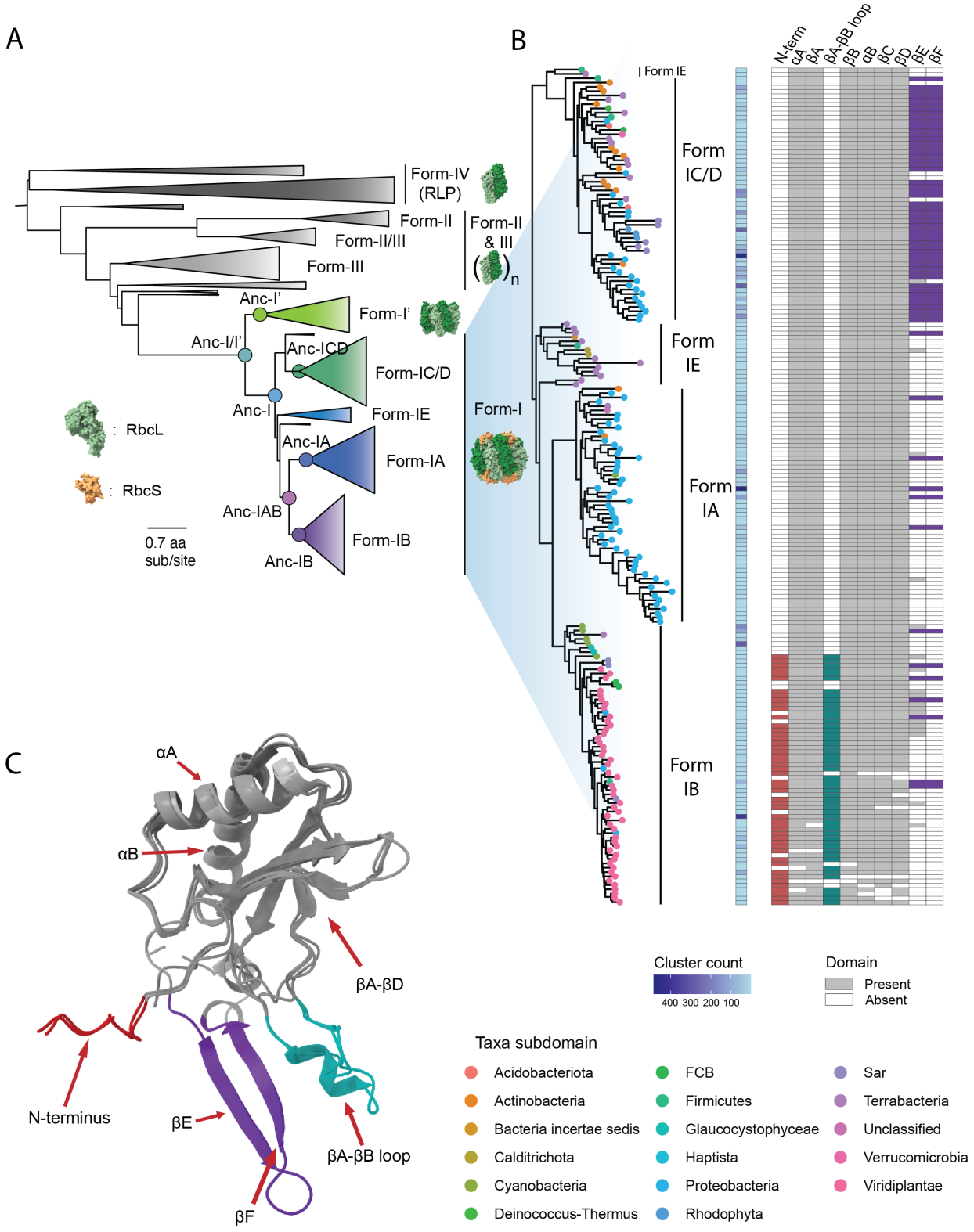


Figure 2: Phylogenetic analysis of RbcS structural diversity. (A) Collapsed maximum-likelihood phylogenetic tree of concatenated RbcL-RbcS sequences. Ancestral nodes used in this study are

155 highlighted. (B) Analysis of RbcS structural diversity, mapped to a clustered Form-I phylogeny,
156 where each tip represents a cluster of RbcS sequences with >63% sequence identity. A heatmap
157 for the number of sequences present in the RbcS cluster is represented in blue. Extant nodes are
158 colored by microbial host taxonomic diversity. The presence and absence of different RbcS
159 features in each extant sequence cluster is indicated by gray and white colors, and the color scheme
160 for the presence of uncommon features is according to (C). (C) Multiple RbcS structures across
161 Form-I subgroups aligned to highlight the different RbcS structural features.

162

163 Global distribution of residues evolutionarily linked to RbcS 164 incorporation

165 RbcL from Form-I and Form-I' RuBisCOs exhibit the same L_8 oligomeric arrangement.
166 Because the presence or absence of the small subunit is the only major structural distinction
167 between the two forms [27], we performed comparative analyses to reveal the small subunit's
168 impact on RbcL sequence and structure. We began by identifying residues that are conserved
169 within, but distinct between, RbcL of extant Form-I or Form-I' clades, which we refer to as
170 "signature positions". These residues are more likely to contribute to the functional distinctions
171 between Form-I and Form-I', including interactions (or lack thereof) with RbcS. The interface
172 region between the large and small subunit has been shown to be critical for the incorporation of
173 RbcS into the Form-I RuBisCO complex [25,33,35,36]. We therefore hypothesized that signature
174 positions would cluster within the RbcL-RbcS interface.

175 Form-I and Form-I' sequence analyses revealed eleven, discontinuous signature positions
176 (**Figure 3A**). These positions are clustered in three different regions on the RuBisCO structure.
177 Two residues (Val 70 and Asp 74; position index from *Rhodobacter sphaeroides*) are located at
178 the RbcL intradimer interface (highlighted in blue), three (Phe 201, Leu 401 and Gln 402) are
179 buried in the RbcL monomer close to the active site (highlighted in purple) and six (Lys 260, Trp
180 284, Arg 286, Asn 288, Gly 159 and Glu162) are found at the RbcL dimer-dimer interface
181 (highlighted in red) (**Figure 3A**). Out of the six residues located at the RbcL dimer-dimer interface,
182 Gly 159 and Glu 162 are the only two signature positions that are also present at the RbcL-RbcS
183 interface. A similar comparison between the Form-IA/B/E and Form-IC/D RbcS sequences did
184 not identify any signature positions (Methods).

185 We examined how these signature positions evolved through the emergence of RbcS by
186 tracking their ancestral amino acid composition through the divergence of Form-I and Form-I'
187 clades. Residues at the signature positions in Anc-I and Anc-I' most closely resemble those of their
188 respective descendants (**Figure 3B**). By contrast, the common ancestor of all Form-I and Form-I'
189 RuBisCOs, Anc-I/I', has approximately equal proportions of Form-I- and Form-I'-like signature
190 positions. Among these, signature positions at the Anc-I/I' RbcL intradimer interface (labeled
191 70,74) are Form-I'-like, whereas those near the RbcL active site (labeled 201, 401, 402) are Form-
192 I-like. Notably, the Anc-I/I' RbcL dimer-dimer interface and RbcL-RbcS interface contains a
193 mixture of Form-I- and Form-I'-like signature positions. These results identify the interface
194 residues that are likely important for oligomerization of the RuBisCO complex and underwent
195 distinct functional specialization in the presence (in Form-I) or absence (in Form-I') of RbcS.

196 In sum, the signature positions are not restricted to the RbcL-RbcS interface but are
197 scattered across the RbcL structure (**Figure 3**). We suggest that RbcS integration resulted in distant
198 mutations, which are known to influence protein function by altering intramolecular interactions
199 and the global dynamics of protein complexes [37]. We therefore submit that these positions reflect
200 how the presence or absence of RbcS affected long-range motions and conformational dynamics
201 of the RuBisCO complex, and therefore its evolutionary dynamics.

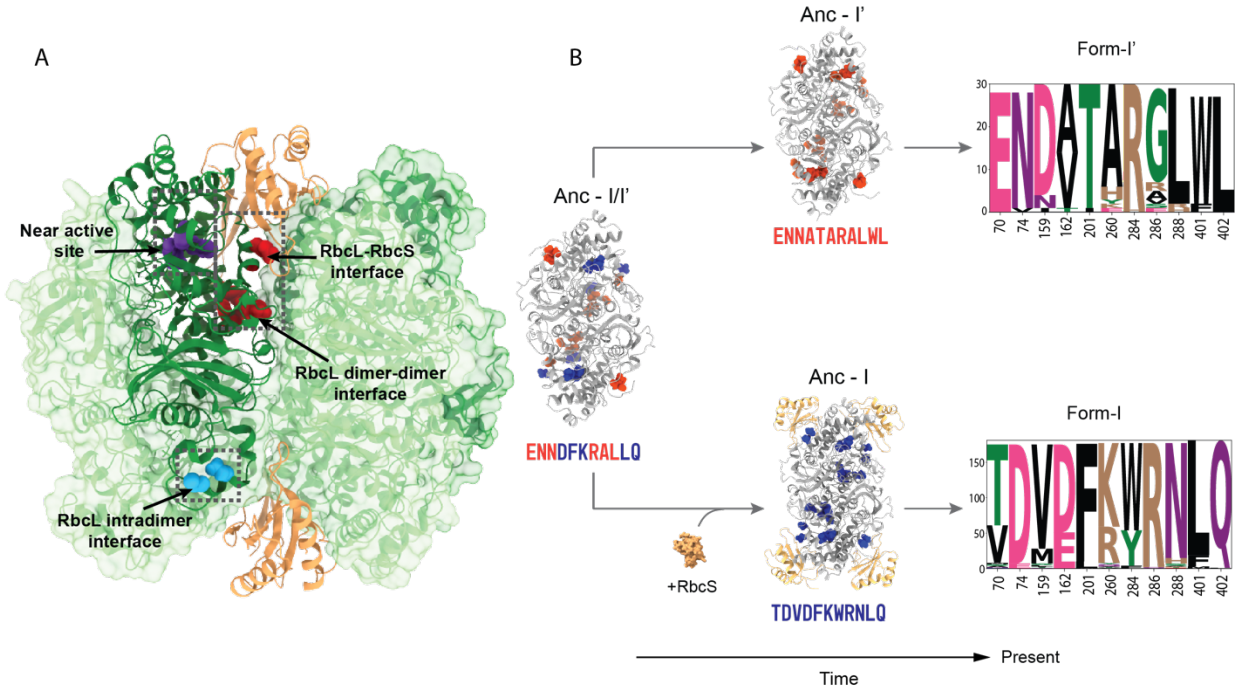


Figure 3: Schematic representation of the structural distribution and evolution of the specialized protein "signature positions" between the RuBisCO Form-I and Form-I'. (A) Signature residues are highlighted based on their structural proximity to each other along with the region it belongs to on the *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 scale with the amino acid frequency across the clade at each position. Residue numbering according to the *R. sphaeroides* sequence.

Molecular dynamics simulations of modern and ancient RuBisCOs

We studied the structural motions of RuBisCO complexes to investigate the influence of the presence or absence of RbcS. We built the ancestral RuBisCO large and small subunit structures (Methods). Modeling of the multimeric structural assembly of ancestral RuBisCO oligomers was guided by extant RuBisCO crystal structures (Methods). We simulated the

220 molecular dynamics of seven ancestral and five extant RuBisCO complexes representing a range
221 of host organisms and their associated environments (**Figure 4A, Table 1**).

222 **Table 1: Description of the extant RuBisCO structures used for analysis.**

Structure	Form	Source Organism	Environment*
1BWV [38]	I-CD	*Galdieria partita* (Red Algae)	High temperature, low pH
6FTL [39]	I-CD	*Skeletonema marinoi* (Diatom)	Low temperature, arctic waters
8RUC [40]	I-B	*Spinacia Oleracea* (Spinach)	Mesophile, surface
3ZXW	I-B	*Thermosynechococcus vestitus* (Cyanobacteria)	Mesophile, tropics
6URA [27]	I'	*Candidatus Promineofilum breve* (Chloroflexi)	Anaerobic, marine sediments

223 * Environment specifies the known natural habitat or ecological niche of the source organism.
224 All the selected structures for extant RuBisCOs are complexed with 2-Carboxyarabinitol 1,5-
225 Bisphosphate (CAP) intermediate, Magnesium ion (Mg) and the carbamylated lysine (KCX) at the
226 active site.

227

228 We used Cα root mean square deviation (RMSD) for the RbcL and RbcS subunits with
229 respect to their average trajectory structure to assess conformational differences in the complex
230 throughout the simulation. The mean RMSD across all individual large and small subunits in the
231 complexes is 0.85Å and 0.94Å respectively. RbcS, with RMSD from 0.74 to 1.26Å, exhibits
232 slightly higher fluctuations compared to RbcL, with RMSD from 0.7 to 1.03Å (**Figure 4B)**. The
233 difference in the average RMSD between the two subunits is statistically significant (RbcL: F-
234 statistic = 4.10, $p = 7.9 \times 10^{-5}$; RbcS: F-statistic = 13.46, $p = 3.8 \times 10^{-11}$; one-way ANOVA). The
235 individual outliers observed for Anc-IA and 8RUC large subunit RMSD values (**Figure 4B)**
236 correspond to the extraneous fluctuations induced by C-terminal residues of one of the eight RbcL
237 subunits. We performed pairwise comparisons of mean RMSD values among the different
238 RuBisCOs. Seven out of all possible pairs show a significant pairwise RMSD difference for RbcL
239 ($p < 0.05$, Tukey post-hoc test) (**Table S2**). All seven of these include either the Anc-I' or Anc-

240 I/I' complex, with a significantly higher RMSD in each case. A higher RMSD for Anc-I' and Anc-
241 I/I' RbcL suggests a more significant conformation change or greater flexibility compared to the
242 other complexes.

243 We used the root mean square fluctuation (RMSF) of the different residues to identify
244 region-specific differences in flexibility. Specifically, we studied the ancestral structures
245 immediately before and after divergence of RbcS-less Form-I' and RbcS-containing Form-I
246 RuBisCOs. These ancestors include Anc-I/I', Anc-I', and Anc-I. The alteration in residue-wise
247 fluctuations between the Anc-I' and Anc-I/I' complex during the MD simulation is close to zero
248 (**Figure 4C, top**). In contrast, the Anc-I RbcL exhibits less fluctuation compared to Anc-I/I' during
249 the simulations, where four distinct sections from the large subunit sequence show considerably
250 higher fluctuations Anc-I/I' compared to Anc-I (**Figure 4C, bottom**). The sections with higher
251 flexibility correspond to the top region of the large subunit that is positioned between the two small
252 subunits in the RuBisCO complex (**Figure 4C, right)**. Out of the 45 RbcL residues present at the
253 RbcS interface in Anc-I, 24 residues (~53%) show a significant difference in RMSF. The
254 localization of RbcL residues near RbcS suggests that the presence of RbcS restricts the movement
255 of these residues. This observation potentially accounts for the higher stability of the Anc-I large
256 subunit compared to the RbcS-less Anc-I/I' and Anc-I'. The differences in RMSF lend support to
257 the hypothesis that RbcS plays a role in stabilizing the RbcL dimers in the octameric complex [20].

258 The extant Form-I' RuBisCO (6URA) provides an additional basis for comparative
259 analysis. Along with Anc-I' and Anc-I/I', RbcS is also absent in 6URA. Unlike its ancestors,
260 however, the extant Form-I' complex does not display a significantly higher RMSD during the
261 pairwise comparison to the other RuBisCO systems. This indicates that in the absence of RbcS,
262 the modern Form-I' RbcL has evolved a more stable L_8 without RbcS, potentially through other
263 mutations. The L_8 assembly in the extant Form-I' complex structure is maintained by a network
264 of hydrogen bonds and salt bridges [27]. Some of the key residues Asp161, Trp165 and Tyr224
265 involved in these interactions in the extant 6URA structure are not conserved in the analyzed RbcS-
266 less ancestors, Anc-I' and Anc-I/I' (Asn140, Arg144 and Phe203 respectively). The absence of
267 these key residue interactions may explain the higher fluctuations observed in the Anc-I' and Anc-
268 I/I' RbcLs during the simulation.

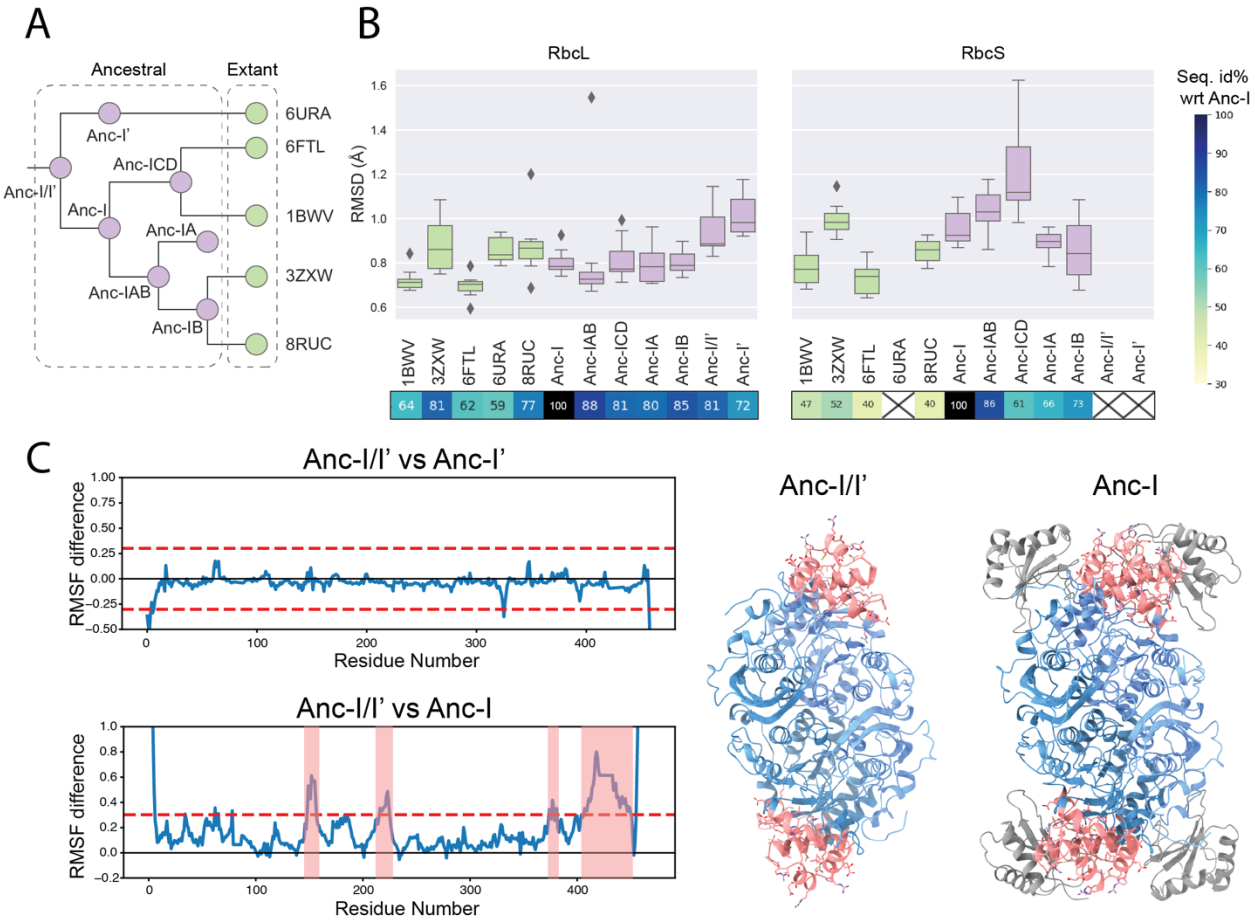


Figure 4: Stability and fluctuation observed in the ancestral and extant RuBisCOs during MD simulations. (A) Ancestral and extant RuBisCO homologs utilized in the study. Violet circles: ancestral RuBisCO, green circle: extant RuBisCO. (B) RMSD plot for the large and small subunit during the simulations. Outliers outside the quartile range are represented by diamonds. 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 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.

While RbcS is indispensable for the activity of extant Form-I RuBisCOs [41,42], its present-day importance does not necessarily imply that it played an essential role at its initial emergence. Protein subunits can become entrenched in complexes via accumulations of neutral

285 mutations that can be deleterious in monomers [43]. However, our analysis further indicates the
286 early importance of RbcS for enhanced stability of the RbcL octamer, along with prior
287 experimental work demonstrating enhanced specificity towards CO_2 [25]. These findings
288 collectively suggest an early advantage of RbcS in RuBisCO assembly at the time of its
289 acquirement.

290 Integration of RbcS restricts the conformational flexibility of Form-I 291 RuBisCO

292 The global distribution of RbcL signature positions (**Figure 3**), inferred to be functionally
293 linked to the presence of RbcS, suggests that RbcS integration likely affected the large subunit's
294 overall dynamics and motion. We utilized MD simulations for RuBisCO complexes with and
295 without RbcS to investigate the impact of RbcS on the enzyme's ability to explore conformational
296 space. A protein's conformational space can be visualized as a set of different conformational
297 states and the conformational variability of an enzyme is considered as an important factor its
298 functional diversity [44,45].

299 To characterize the motion and variation in conformations for the different RuBisCO
300 simulations, we performed a principal component analysis (PCA) of the trajectory of the Cα atoms
301 independently [46,47]. Dimensionality reduction shows that changes along the first two principal
302 components (PC1 and PC2) of the RbcL MD trajectories reveals a topology that is conserved
303 across most of the RuBisCO complexes (**Figure 5A**). The distribution spans the same regions in
304 all simulations (a high PC1-high PC2 region; an intermediate PC1-low PC2 region; and a low PC1-
305 high PC2 region) with most data points placed within the last two regions. Exceptions to this were
306 the Anc-I' and 6URA systems, which both belong to the Form-I' clade of RuBisCO and exhibit
307 two different distributions (**Figure 5A**). Further analyses across five independent MD simulations
308 reveal a consistent topology for PC1 versus PC2 for Anc-I (**Figure S10**). In contrast, the
309 distributions for the RuBisCO ancestors lacking RbcS (Anc-I/I', and Anc-I') exhibit lower
310 similarity to Anc-I (**Figure S10**). Inclusion or exclusion of the RbcS and the N- and C-terminal
311 residues of RbcL for the analysis did not change the pattern of PCA across the different RuBisCOs
312 (**Figure S11, S12)**. A joint projection with identical essential modes across all RbcL trajectories
313 (**Figure S13**) captures approximately 66% of all motions during the simulations along the first

314 three principal components (**Figure 5B**). We observe a distinction between the ancestral and extant
315 RuBisCOs along the PC1. Another separation of the Form-I' clade Anc-I' and 6URA RuBisCOs
316 is found along the PC2 (**Figure 5B**), parallel with the pattern observed in the PCA for individual
317 simulations.

318 We find that the conformational changes for both ancestral and extant Form-I RbcLs are
319 more restricted compared to Form-I' and their common ancestor Anc-I/I'. We thus suggest that
320 RbcL explores a narrower range of conformational states in the presence of RbcS. The unaltered
321 conformational variability of Form-I RbcL prompts us to infer that the introduction of RbcS
322 limited RbcL's ability to explore new conformational states, constraining the conformational
323 landscape available to RbcL in the enzyme complex. While initial incorporation of RbcS is
324 beneficial in conferring the complex's conformational stability (**Figure 4C**), we suggest that
325 addition of RbcS in Form-I resulted in higher RbcL conformational rigidity relative to Form-I'
326 RbcL. This finding indicates that RbcS regulates the enzyme's structural dynamics by acting akin
327 to a structural stability anchor.

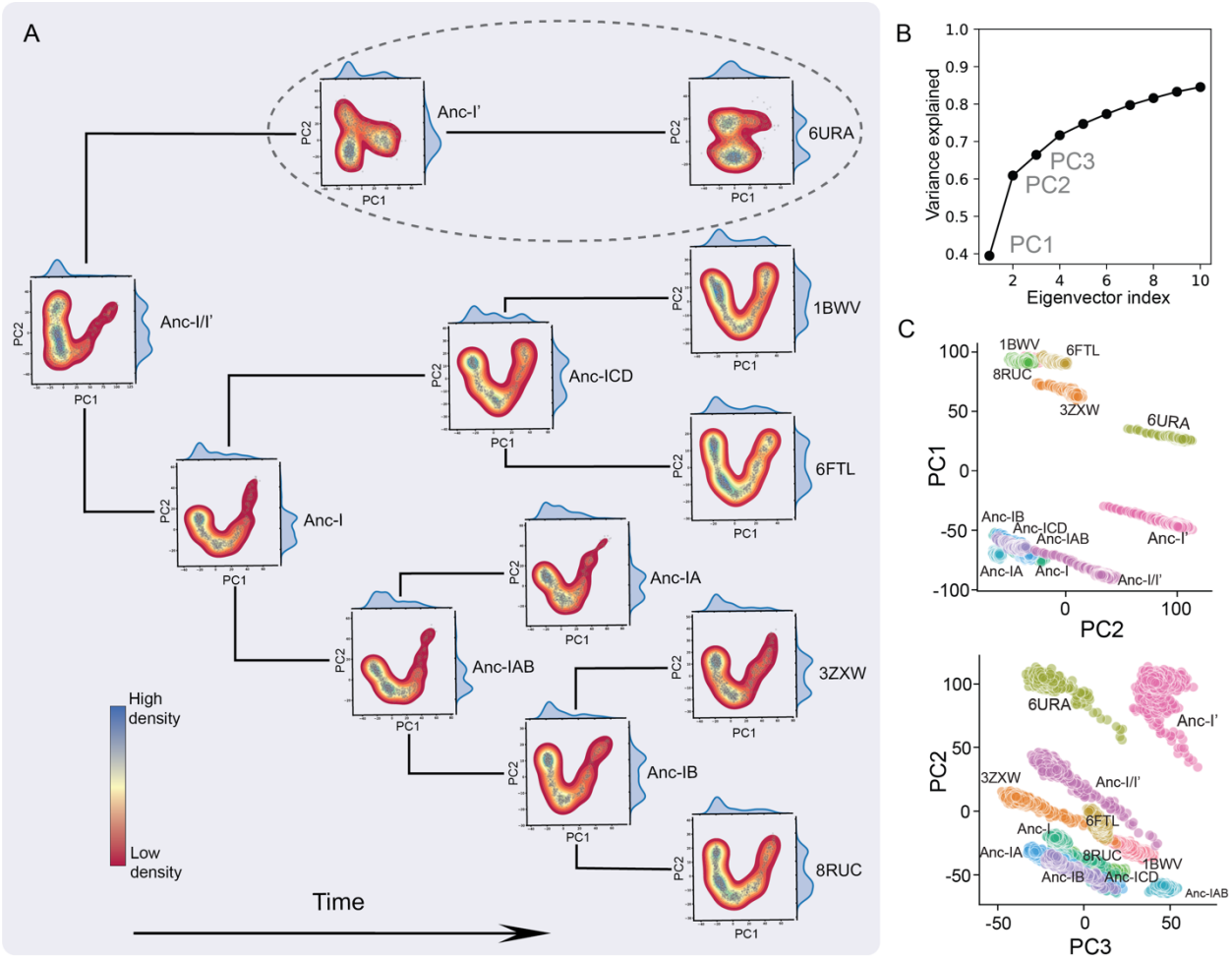


Figure 5: Representation of the MD-trajectories projected along its principal components based on Cα atoms of RbcLs from different RuBisCO complexes during the simulations. (A) Individual PCA along the two most important principal components (PC1 and PC2) for different RuBisCOs. Gray dashed circle: Form-I' RuBisCOs. (B) Explained variance ratio for the joint PCA with identical principal components across all simulations. (C) Pairwise PCA of the first three principal components (PC1-PC2 and PC2-PC3) after aligning the trajectories of all RuBisCO simulations.

The functional diversity (e.g., promiscuity) of a protein depends on its ability to explore more conformational states and new folds [48–51]. Therefore, a reduction in the conformational variability of RuBisCO following the integration of RbcS may directly relate to an increase in CO_2-specificity, as well as a decrease in the "evolvability" of the enzyme [48,52]. Our analysis illustrates that RbcL cannot explore new conformations or folds when RbcS is present, locking

341 RbcS in the RuBisCO complex. Further investigations utilizing enhanced sampling techniques
342 would also help explore larger portions of the conformational space accessible to the RuBisCO
343 complex [53,54].

344 Heterogeneity of CO_2/O_2 diffusion patterns across ancestral and extant 345 RuBisCOs

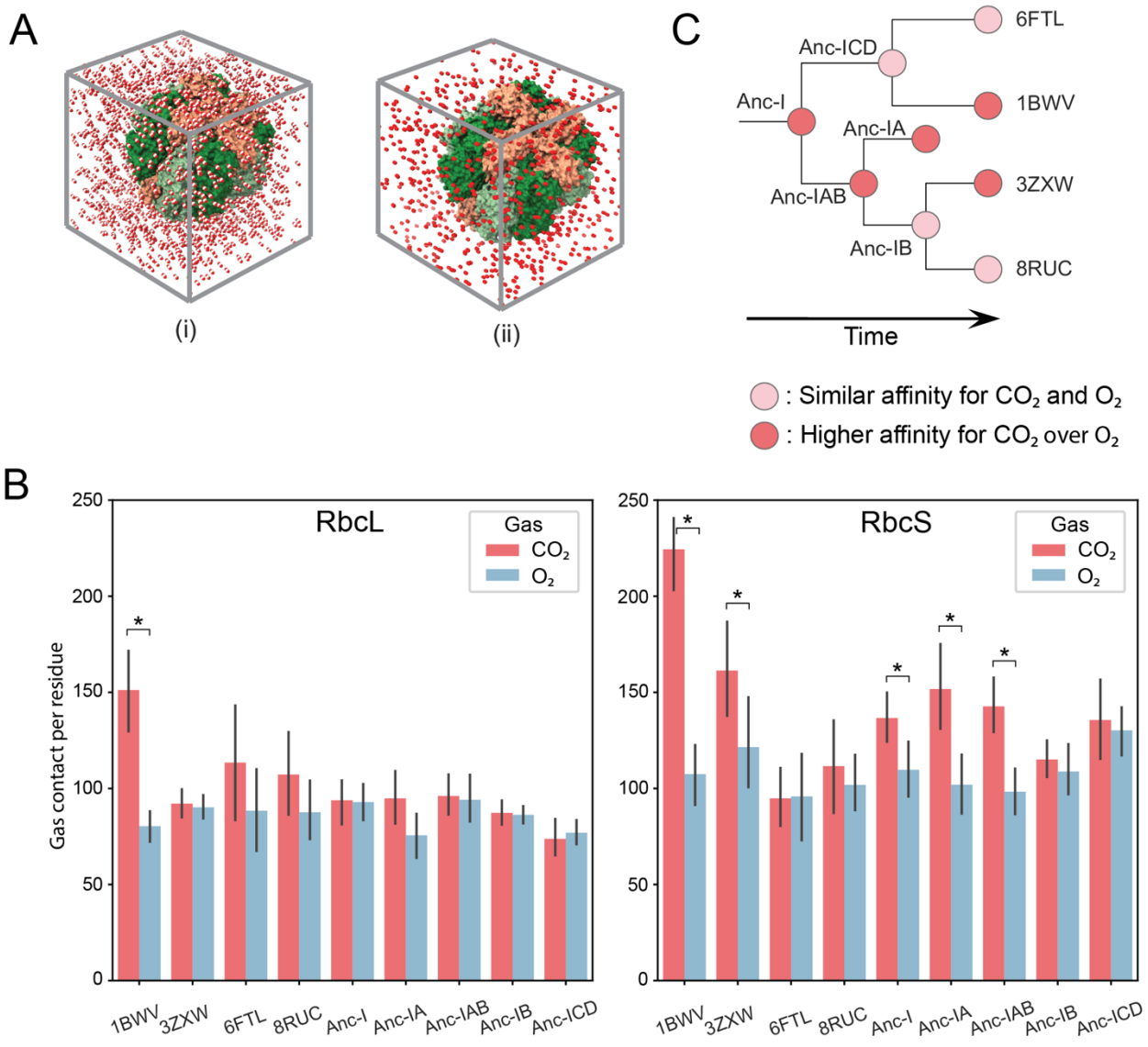
346 Previous work suggested that RbcS acts as a CO_2-reservoir [26], increasing the
347 concentration of CO_2 molecules near the enzyme and making CO_2 more accessible to the active
348 site. This hypothesis also provides an additional functional justification for the integration of RbcS,
349 given the increase in Earth's atmospheric O_2 concentration ~2.5 billion years ago, around the time
350 of RbcS integration into the RuBisCO complex.

351 We performed MD simulations for the different RuBisCO complexes with CO_2 and O_2 gas
352 molecules present in the medium (**Figure 6A**) to assess the affinity of RbcS residues towards both
353 gases. We estimated the relative affinities of the gas molecules by counting the number of
354 interactions between the gases and the protein throughout the simulation within a distance
355 threshold of 6Å. The MD simulations for O_2 and CO_2 gas molecules were performed separately
356 (**Figure 6A**). Five independent replicates for MD simulations of one extant (8RUC) and one
357 ancestral (Anc-I) enzyme were conducted to assess the robustness and replicability of the gas
358 dynamics modeling approach. To assess the impact of gas concentrations, we performed
359 simulations with varying CO_2 and O_2 gas concentrations. The trends observed for affinity of the
360 subunits across different gas concentrations remains relatively consistent between the two subunits
361 (**Figure S14**).

362 An independent two sample t-test was performed to compare the mean CO_2 and O_2 gas
363 contacts across all simulations. For RbcL, we observe that only 1BWV has a significantly higher
364 affinity for CO_2 over O_2 ($t = 5.92$; $p = 3.7x10^{-5}$), whereas other complexes do not show a
365 statistically significant ($p < 0.05$) difference (**Figure 6B**). Alternatively, five RbcS (1BWV,
366 3ZXW, Anc-I, Anc-IA, Anc-IAB) out of the nine exhibit a significantly ($p < 0.05$) higher affinity
367 for CO_2 than for O_2 molecules in the surrounding medium (**Figure 6B**) (**Table S3**). RbcS that
368 exhibit a higher affinity for CO_2 over O_2 are not confined to a specific clade in the phylogenetic
369 tree (**Figure 6C**). We find that these results do not vary significantly over five independent

370 simulations (**Figure S15**). Our finding that the CO_2 reservoir property is not conserved in both
371 ancestral and extant RbcS suggests that this role was not evolutionarily relevant in the early
372 incorporation of RbcS into RuBisCO. Rather, an adaptive property to concentrate CO_2 may have
373 evolved in select ancestors and lineages due to other intercellular or environmental pressures.

374



375

376 **Figure 6**: CO_2 and O_2 gas diffusion molecular dynamics across different RuBisCO systems. (A)
377 A snapshot of the MD simulation box with CO_2 (i) and O_2 (ii) gas molecules in the medium at a
378 concentration of 400 mM with the protein complex. (B) Barplot representing the number of gas
379 contacts per residue for RbcL and RbcS for each of RuBisCO for the two CO_2 and O_2 gas

380 molecules. Error bars represent the standard deviation across the 8 subunits in the RuBisCO
381 complex. Proteins with a significant difference between the CO_2 and O_2 affinity are marked with
382 an asterisk (*). (C) Trend for the difference in affinity for CO_2 and O_2 gas molecules for RbcS as
383 observed in the barplot (B). Time is displayed along the y-axis.

384 Conclusions

385 The co-evolution of the large and small subunits of RuBisCO presents an opportunity to
386 analyze molecular evolutionary events within key macroevolutionary and geochemical transitions,
387 independently documented by Earth's geologic record. We examined the historical impact of
388 RuBisCO small subunit emergence on the structural motions and evolutionary trajectory of the
389 enzyme, with specific focus on a possible adaptive role in concentrating CO_2. Both extant and
390 ancestral Form-I RuBisCO complexes that contain RbcS show higher CO_2-specificity than other
391 forms of RuBisCO lacking RbcS [16,25]. These observations have led to the hypothesis that the
392 improved CO_2-specificity of RuBisCO after RbcS integration would have been advantageous after
393 the GOE, given the subsequent rise in atmospheric O_2 levels and decrease in CO_2 concentrations.
394 We illustrate that not all ancestral and extant RbcS proteins act as a reservoir to concentrate CO_2
395 around the RuBisCO complex. It is possible that certain RuBisCOs with RbcS might have adopted
396 a CO_2-reservoir strategy based on other intercellular or possibly environmental factors as we find
397 that this adaptation appears multiple times in the enzyme's history. Moreover, our analysis
398 suggests additional consequences of RbcS integration: increased Form-I RbcL stability and lower
399 flexibility relative to that of Form I' RbcL. This addition limited the enzyme's ability to explore
400 novel conformations and was led to a "locked in", albeit beneficial, state in the face of a major
401 environmental shift, particularly with regards to environmental oxygen levels.

402 Our observations help explain the integration of RbcS in all Form-I RuBisCOs.
403 Nevertheless, the origin of RbcS still remains unresolved. Certain cyanobacteria have two other
404 proteins that are part of the carbon fixation machinery: β-carboxysome structural protein (CcmM)
405 and a RuBisCO activase-like protein (ALC). Both these proteins have one or multiple domains
406 that are "RbcS-like" and are considered homologs to RbcS [35,55,56]. The presence of these
407 protein domains is specific to cyanobacteria. Further studies should explore whether the RbcS-like
408 protein domains could have emerged alongside or after the RbcS in response to the ancient shifts
409 in atmospheric CO_2 and O_2 levels. Future empirical work that tracks the evolutionary history of

410 Form-I RuBisCO in light of these revelations will deepen our knowledge of the ancestry, structure,
411 and the historical role of RbcS and RbcS-like proteins. In summary, we reveal that ancestral
412 RuBisCO had a remarkable evolutionary flexibility responding to a global environment shift,
413 though this adaptation resulted in increased rigidity imparted by a novel subunit. Thus, the rigidity
414 that makes modern RuBisCO notoriously resistant to artificial improvements in CO_2 specificity
415 may have ironically emerged from nature's own solution to this challenge.

Methods

Phylogenetic Reconstruction of RuBisCO

Homologous sequences for RbcS protein from the NCBI non-redundant database were identified using PSI-BLAST [57,58]. The RbcS sequence from *Thermosynechococcus elongatus* (PDB id: 2YBV) [59] 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 [60] and aligned by MAFFT [61] (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 [62] with the LG+R6 evolutionary model (**Figure S1**). ModelFinder Plus [63] 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 RbcL sequences were combined with RbcL sequences for Form-I', II, III and IV from Banda et al. [27] to build an RbcL dataset. An RbcL phylogenetic tree was constructed using IQ-Tree with the LG+R9 evolutionary model (all other parameters were the same as for the RbcS tree) (**Figure S2**).

The sequences in the RbcL dataset were aligned by MAFFT, with the same parameters as those mentioned above, and were concatenated with the RbcS alignment. A maximum-likelihood phylogenetic tree was constructed for the concatenated alignment using IQ-Tree, with a partition model [64] and ModelFinder plus [63] to search and implement the best-fit evolutionary models corresponding to the RbcL (LG + R9) and RbcS (LG + R5) section of the concatenated alignment. In line with prior research, Form-IV "RuBisCO-like" protein sequences were used to root the

444 phylogenetic tree [8,25,27,29,65]. Ancestral sequence inference for the concatenated RbcL-RbcS
445 tree was performed using PamL4.9 (LG model) [66]. Reconstruction of the gaps in the ancestral
446 sequences was performed using a binary likelihood model as described by Aadland et al. [67].

447 RbcS sequence and structural diversity analysis

448 Candidate RbcS structures (listed in **Table 1**) from Form-I subclades were used to identify
449 the different structural features in RbcS. The amino acid regions corresponding to these features
450 were identified in the multiple sequence alignment of extant RbcS. The binary heatmap indicating
451 the presence or absence of these features across all extant RbcS sequences in the phylogeny was
452 generated based on the presence of at least half of the residues in the respective region of the RbcS
453 alignment. Host taxonomy for the extant sequences was assigned using the REST API provided
454 by Ensembl (https://rest.ensembl.org/).

455 RbcL specialized signature position analysis

456 Sequence specialization between the Form-I and Form-I' extant RbcL sequences was
457 analyzed using TwinCons [68], with blosum62 as substitution matrix for score calculation, and
458 Zebra2 [69], using the webserver default parameters. Specialized residues identified by Zebra2
459 and TwinCons score were then filtered based on conservation in the multiple sequence alignment
460 for Form-I and Form-I' sequences. Same steps as above were followed to identify separately
461 conserved amino acids between 131 RbcS sequences from Form-IA/B/E and 59 from Form-IC/D.
462 None of the RbcS positions had a TwinCons score < -1, resulting in no signature positions.

463 Structural Modeling of RuBisCO

464 We predicted the structures of ancestral RuBisCOs using the deep-learning based
465 Colabfold software [70]. We built our models by combining predictions and structure alignments.
466 We built the model for the RbcL-dimer with RbcS-dimer for each ancestral protein with templates
467 from PDB, structure refinement using Amber and 3 prediction recycles. Low-confidence modeled
468 terminal regions for the predicted structures were removed using a pLDDT threshold of 50. The
469 extant RuBisCO structure from *Thermosynechococcus elongatus* (PDB: 2YBV) was used as a

470 template to obtain a predicted hexadecameric complex using UCSF Chimera [71]. Anc-I' and Anc-
471 I/I' did not have an RbcS sequence resulting in an octameric complex.

472 RuBisCO MD simulations

473 We simulated the molecular dynamics of both extant and ancient RuBisCO complexes. We
474 selected seven ancestral structures (nodes highlighted in **Figure 2A**) corresponding to the last
475 common ancestors of different Form-I and Form-I' clades and five experimentally determined
476 extant RuBisCO structures, representing the Form-IA, IB (two from this clade), IC/D, and I'
477 RuBisCO clades. PDB IDs for these entries are: 1BWV (red algae), 3ZXW (green algae), 6FTL
478 (diatom), 6URA (chloroflexi), and 8RUC (plantae) (more details in **Table 1**). The 6URA entry
479 represents the first case discovered of a Form-I-like RuBisCO without a small subunit. This
480 selection represents the biological diversity of the RuBisCO enzyme while preserving the
481 following attributes: high resolution (under 2.0 Å), carbamylation of the lysine at the active site
482 [72], and CABP and magnesium at the active site.

483 We obtained parameters for the CABP molecule through the standard Amber protocol for
484 ligands [73]. RuBisCO proteins include different post-translational modifications that required
485 additional steps. We computed the CABP and the carbamylated lysine charges at HF/6-31*G level,
486 as specified in the Amber protocol. The post-translational modifications found in 6FTL were
487 modeled with semi-empirical charges. In the case of the modeled ancient structures, we manually
488 added both ligands (CABP), ions (Mg), and the post-translation modification at the carbamylated
489 lysine.

490 We built the topology and parameters of the molecular dynamics systems using tleap [74]
491 and Amber14 [75]. Periodic solvation boxes were constructed with 10 Å spacing and water
492 molecules according to the TIP3P model [76]. Sodium and chloride ions counterbalanced the
493 charge of the system. The particle-mesh Ewald summation method was used for long-range
494 electrostatics and a 10 Å cutoff was set for short-range non-bonded interactions. Initial geometries
495 in all systems were minimized at 5,000 conjugate-gradient steps after which water was equilibrated
496 at 298 K and 1 atm for 100 ps at 2 fs time steps. Production runs were then performed for 75 ns in
497 the NPT ensemble at P = 1 atm and T = 298 K. We used the hydrogen mass partitioning method
498 to ensure the stability of the simulation under large integration steps, thereby increasing the

499 simulation speed. Langevin dynamics for T control and Nosé-Hoover Langevin piston method for
500 P control were used. We carried out the MD simulation on OpenMM 7.7 [77] running in the Nvidia
501 Tesla A100 nodes of the Center for High Throughput Computation at the University of Wisconsin-
502 Madison.

503 Gas diffusion MD simulations

504 We obtained CO_2 and O_2 parameters through the standard Amber protocol. We computed
505 the charges of these molecules at the HF/6-31*G level through the RESP protocol [78]. The
506 embedding of proteins in solutions with gas molecules in dissolution consisted of replacing water
507 molecules from a previously solvated system to avoid clashes among the gas molecules. We
508 computed the number of molecules that match a given concentration following the Amber
509 recommended method to compute ion concentrations in a solvent. The other aspects of the
510 simulation were equivalent to the previous epigraph.

511 Analysis of MD simulations

512 We employed MDAnalysis [79] and ProDy [80] to analyze and process the outcome of the
513 MD simulations. For the principal component analysis, we used the scikit-learn implementation
514 instead of the native MDAnalysis tool for performance reasons. Each of the 75 ns simulations was
515 performed at the same temperature, pH, and solvent conditions. The initial 25 ns of all simulations
516 were used for equilibration. The overall dynamics combined add more than 1μs of simulations.
517 The average sizes of the systems are ~300k atoms including the solvent molecules.

518 Data accessibility

519 All information related to sequence datasets, phylogeny, and structures is available at
520 https://github.com/kacarlab/RuBisCO_evolution.

521 Competing interests

522 The authors declare no competing interests.

Acknowledgement

We would like to thank Alessandro Senes, Phil Romero, Tina Wang and Srivatsan Raman for their feedback on the study, Aya Klos, 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.

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.

Funding

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).

References

1. Wagner A, Rosen W. 2014 Spaces of the possible: universal Darwinism and the wall between technological and biological innovation. *J. R. Soc. Interface* **11**, 20131190. (doi:10.1098/rsif.2013.1190)

2. Falkowski PG, Fenchel T, Delong EF. 2008 The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science* **320**, 1034–1039. (doi:10.1126/science.1153213)

3. Andersson I. 2008 Catalysis and regulation in Rubisco. *J. Exp. Bot.* **59**, 1555–1568. (doi:10.1093/jxb/ern091)

4. Fernie AR, Bauwe H. 2020 Wasteful, essential, evolutionary stepping stone? The multiple personalities of the photorespiratory pathway. *Plant J.* **102**, 666–677. (doi:10.1111/tpj.14669)

5. Garcia AK, Cavanaugh CM, Kacar B. 2021 The curious consistency of carbon biosignatures over billions of years of Earth-life coevolution. *ISME J.* **15**, 2183–2194. (doi:10.1038/s41396-021-00971-5)

6. Ward LM, Shih PM. 2019 The evolution and productivity of carbon fixation pathways in response to changes in oxygen concentration over geological time. *Free Radic. Biol. Med.* **140**, 188–199. (doi:10.1016/J.FREERADBIOMED.2019.01.049)

7. Kędzior M, Garcia AK, Li M, Taton A, Adam ZR, Young JN, Kaçar B. 2022 Resurrected Rubisco suggests uniform carbon isotope signatures over geologic time. *Cell Rep.* **39**, 110726. (doi:10.1016/j.celrep.2022.110726)

8. Kacar B, Hanson-Smith V, Adam ZR, Boekelheide N. 2017 Constraining the timing of the Great Oxidation Event within the Rubisco phylogenetic tree. *Geobiology* **15**, 628–640. (doi:10.1111/gbi.12243)

9. 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. (doi:10.1016/J.RESMIC.2005.01.014)

10. Caetano-Anollés G. 2017 RubisCO and the Search for Biomolecular Culprits of Planetary Change. *BioEssays* **39**, 1700174. (doi:10.1002/bies.201700174)

11. Erb TJ, Zarzycki J. 2018 A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme. *Curr. Opin. Biotechnol.* **49**, 100–107. (doi:10.1016/j.copbio.2017.07.017)

12. Raymond J, Segrè D. 2006 The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**, 1764–1767. (doi:10.1126/SCIENCE.1118439/SUPPL_FILE/RAYMOND.SOM.PDF)

13. Wang M, Jiang YY, Kim KM, Qu G, Ji HF, Mittenthal JE, Zhang HY, Caetano-Anollés G. 2011 A universal molecular clock of protein folds and its power in tracing the early history of

572 aerobic metabolism and planet oxygenation. *Mol. Biol. Evol.* **28**, 567–582.
573 (doi:10.1093/MOLBEV/MSQ232)

574 14. Young JN, Rickaby REM, Kapralov MV, Filatov DA. 2012 Adaptive signals in algal Rubisco
575 reveal a history of ancient atmospheric carbon dioxide. *Philos. Trans. R. Soc. B Biol. Sci.* **367**,
576 483–492. (doi:10.1098/rstb.2011.0145)

577 15. Tabita FR, Satagopan S, Hanson TE, Kreel NE, Scott SS. 2008 Distinct form I, II, III, and IV
578 Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and
579 structure/function relationships. *J. Exp. Bot.* **59**, 1515–1524. (doi:10.1093/JXB/ERM361)

580 16. Flamholz AI, Prywes N, Moran U, Davidi D, Bar-On YM, Oltrogge LM, Alves R, Savage D,
581 Milo R. 2019 Revisiting Trade-offs between Rubisco Kinetic Parameters. *Biochemistry* **58**,
582 3365–3376.
583 (doi:10.1021/ACS.BIOCHEM.9B00237/SUPPL_FILE/BI9B00237_SI_005.XLSX)

584 17. Esquivel MG, Genkov T, Nogueira AS, Salvucci ME, Spreitzer RJ. 2013 Substitutions at the
585 opening of the Rubisco central solvent channel affect holoenzyme stability and CO2/O2
586 specificity but not activation by Rubisco activase. *Photosynth. Res.* **118**, 209–218.
587 (doi:10.1007/S11120-013-9916-0/FIGURES/7)

588 18. Joshi J, Mueller-Cajar O, Tsai Y-CC, Hartl FU, Hayer-Hartl M. 2015 Role of Small Subunit
589 in Mediating Assembly of Red-type Form I Rubisco. *J. Biol. Chem.* **290**, 1066–1074.
590 (doi:10.1074/jbc.M114.613091)

591 19. Liu C *et al.* 2010 Coupled chaperone action in folding and assembly of hexadecameric
592 Rubisco. *Nature* **463**, 197–202. (doi:10.1038/nature08651)

593 20. Spreitzer RJ. 2003 Role of the small subunit in ribulose-1,5-bisphosphate
594 carboxylase/oxygenase. *Arch. Biochem. Biophys.* **414**, 141–149. (doi:10.1016/S0003-
595 9861(03)00171-1)

596 21. Gatenby AA. 1988 Synthesis and assembly of bacterial and higher plant Rubisco subunits in
597 Escherichia coli. *Photosynth. Res.* **17**, 145–157. (doi:10.1007/BF00047686/METRICS)

598 22. Genkov T, Meyer M, Griffiths H, Spreitzer RJ. 2010 Functional Hybrid Rubisco Enzymes
599 with Plant Small Subunits and Algal Large Subunits: ENGINEERED rbcS cDNA FOR
600 EXPRESSION IN CHLAMYDOMONAS. *J. Biol. Chem.* **285**, 19833–19841.
601 (doi:10.1074/JBC.M110.124230)

602 23. Mao Y, Catherall E, Díaz-Ramos A, Greiff GRL, Azinas S, Gunn L, McCormick AJ. 2022
603 The small subunit of Rubisco and its potential as an engineering target. *J. Exp. Bot.*
604 (doi:10.1093/JXB/ERAC309)

605 24. Matsumura H *et al.* 2020 Hybrid Rubisco with Complete Replacement of Rice Rubisco Small
606 Subunits by Sorghum Counterparts Confers C4 Plant-like High Catalytic Activity. *Mol. Plant*
607 **13**, 1570–1581. (doi:10.1016/J.MOLP.2020.08.012)

608 25. Schulz L *et al.* 2022 Evolution of increased complexity and specificity at the dawn of form I
609 Rubiscos. *Science* **378**, 155–160. (doi:10.1126/SCIENCE.ABQ1416)

610 26. Van Lun M, Hub JS, Van Der Spoel D, Andersson I. 2014 CO2 and O2 distribution in rubisco
611 suggests the small subunit functions as a CO2 reservoir. *J. Am. Chem. Soc.* **136**, 3165–3171.
612 (doi:10.1021/ja411579b)

613 27. Banda DM *et al.* 2020 Novel bacterial clade reveals origin of form I Rubisco. *Nat. Plants* **6**,
614 1158–1166. (doi:10.1038/s41477-020-00762-4)

615 28. Tabita FR, Hanson TE, Li H, Satagopan S, Singh J, Chan S. 2007 Function, Structure, and
616 Evolution of the RubisCO-Like Proteins and Their RubisCO Homologs. *Microbiol. Mol. Biol.*
617 *Rev.* **71**, 576–599. (doi:10.1128/MMBR.00015-07)

618 29. Poudel S, Pike DH, Raanan H, Mancini JA, Nanda V, Rickaby C REM, Falkowski PG. 2020
619 Biophysical analysis of the structural evolution of substrate specificity in RuBisCO. *PNAS*
620 (doi:10.1073/pnas.2018939117)

621 30. West-Roberts JA *et al.* 2021 The Chloroflexi supergroup is metabolically diverse and
622 representatives have novel genes for non-photosynthesis based CO2 fixation. *bioRxiv* ,
623 2021.08.23.457424-2021.08.23.457424. (doi:10.1101/2021.08.23.457424)

624 31. Bracher A, Whitney SM, Hartl FU, Hayer-Hartl M. 2017 Biogenesis and Metabolic
625 Maintenance of Rubisco*. *Httpsdoiorg101146annurev-Arplant-043015-111633* **68**, 29–60.
626 (doi:10.1146/ANNUREV-ARPLANT-043015-111633)

627 32. Bouvier JW, Emms DM, Kelly S. 2024 Rubisco is evolving for improved catalytic efficiency
628 and CO2 assimilation in plants. *Proc. Natl. Acad. Sci.* **121**, e2321050121.
629 (doi:10.1073/pnas.2321050121)

630 33. Knight S, Andersson I, Br~ind~n C-I. 1990 Crystallographic Analysis of Ribulose 1,5-
631 Bisphosphate Carboxylase from Spinach at 2.4 A Resolution Subunit Interactions and Active
632 Site. **215**, 113–160.

633 34. Schmidt GW, Mishkind ML. 1986 THE TRANSPORT OF PROTEINS INTO
634 CHLOROPLASTS.

635 35. Ryan P, Forrester TJB, Wroblewski C, Kenney TMG, Kitova EN, Klassen JS, Kimber MS.
636 2019 The small RbcS-like domains of the β-carboxysome structural protein CcmM bind
637 RubisCO at a site distinct from that binding the RbcS subunit. *J. Biol. Chem.* **294**, 2593–5195.
638 (doi:10.1074/JBC.RA118.006330)

639 36. Van Lun M, Van Der Spoel D, Andersson I. 2011 Subunit interface dynamics in
640 hexadecameric Rubisco. *J. Mol. Biol.* **411**, 1083–1098. (doi:10.1016/j.jmb.2011.06.052)

641 37. Miton CM, Buda K, Tokuriki N. 2021 Epistasis and intramolecular networks in protein
642 evolution. *Curr. Opin. Struct. Biol.* **69**, 160–168. (doi:10.1016/J.SBI.2021.04.007)

38. Sugawara H, Yamamoto H, Shibata N, Inoue T, Okada S, Miyake C, Yokota A, Kai Y. 1999 Crystal Structure of Carboxylase Reaction-oriented Ribulose 1,5-Bisphosphate Carboxylase/Oxygenase from a Thermophilic Red Alga, *Galdieria partita* *. *J. Biol. Chem.* **274**, 15655–15661. (doi:10.1074/jbc.274.22.15655)

39. Valegård K *et al.* 2018 Structural and functional analyses of Rubisco from arctic diatom species reveal unusual posttranslational modifications. *J. Biol. Chem.* **293**, 13033–13043. (doi:10.1074/jbc.RA118.003518)

40. 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. (doi:10.1006/jmbi.1996.0310)

41. Lee B, Tabita FR. 1990 Purification of recombinant ribulose-1,5-bisphosphate carboxylase/oxygenase large subunits suitable for reconstitution and assembly of active L8S8 enzyme. *Biochemistry* **29**, 9352–9357. (doi:10.1021/bi00492a007)

42. Andrews TJ. 1988 Catalysis by cyanobacterial ribulose-bisphosphate carboxylase large subunits in the complete absence of small subunits. *J. Biol. Chem.* **263**, 12213–12219. (doi:10.1016/S0021-9258(18)37741-X)

43. Hochberg GKA, Liu Y, Marklund EG, Metzger BPH, Laganowsky A, Thornton JW. 2020 A hydrophobic ratchet entrenches molecular complexes. *Nature* **588**, 503–508. (doi:10.1038/s41586-020-3021-2)

44. Babtie A, Tokuriki N, Hollfelder F. 2010 What makes an enzyme promiscuous? *Curr. Opin. Chem. Biol.* **14**, 200–207. (doi:10.1016/j.cbpa.2009.11.028)

45. Nobeli I, Favia AD, Thornton JM. 2009 Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157–167. (doi:10.1038/nbt1519)

46. David CC, Jacobs DJ. 2014 Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. *Methods Mol. Biol. Clifton NJ* **1084**, 193. (doi:10.1007/978-1-62703-658-0_11)

47. Hayward S, de Groot BL. 2008 Normal Modes and Essential Dynamics. In *Molecular Modeling of Proteins* (ed A Kukol), pp. 89–106. Totowa, NJ: Humana Press. (doi:10.1007/978-1-59745-177-2_5)

48. Tokuriki N, Tawfik DS. 2009 Protein Dynamism and Evolvability. *Science* **324**, 203–207. (doi:10.1126/science.1169375)

49. Jackson C, Toth-Petroczy A, Kolodny R, Hollfelder F, Fuxreiter M, Kamerlin SCL, Tokuriki N. 2022 Adventures on the Routes of Protein Evolution—In Memoriam Dan Salah Tawfik (1955–2021). *J. Mol. Biol.* **434**, 167462–167462. (doi:10.1016/J.JMB.2022.167462)

50. Petrović D, Risso VA, Kamerlin SCL, Sanchez-Ruiz JM. 2018 Conformational dynamics and enzyme evolution. *J. R. Soc. Interface* **15**, 20180330. (doi:10.1098/rsif.2018.0330)

679 51. Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB. 2015 Evolution of Conformational
680 Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme.
681 *Mol. Biol. Evol.* **32**, 132–143. (doi:10.1093/molbev/msu281)

682 52. James LC, Tawfik DS. 2003 Conformational diversity and protein evolution – a 60-year-old
683 hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368. (doi:10.1016/S0968-0004(03)00135-
684 X)

685 53. Bernardi RC, Melo MCR, Schulten K. 2015 Enhanced Sampling Techniques in Molecular
686 Dynamics Simulations of Biological Systems. *Biochim. Biophys. Acta* **1850**, 872–877.
687 (doi:10.1016/j.bbagen.2014.10.019)

688 54. Pravin MA, Singh SK. 2023 Enhanced Sampling and Free Energy Methods to Study Protein
689 Folding and Dynamics. In *Protein Folding Dynamics and Stability: Experimental and*
690 *Computational Methods* (eds P Saudagar, T Tripathi), pp. 165–179. Singapore: Springer
691 Nature. (doi:10.1007/978-981-99-2079-2_9)

692 55. Lechno-Yossef S, Rohnke BA, Belza ACO, Melnicki MR, Montgomery BL, Kerfeld CA. 2020
693 Cyanobacterial carboxysomes contain an unique rubisco-activase-like protein. *New Phytol.*
694 **225**, 793–806. (doi:10.1111/nph.16195)

695 56. Wang H *et al.* 2019 Rubisco condensate formation by CcmM in β-carboxysome biogenesis.
696 *Nat. 2019 5667742* **566**, 131–135. (doi:10.1038/s41586-019-0880-5)

697 57. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped
698 BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids*
699 *Res.* **25**, 3389–3402. (doi:10.1093/NAR/25.17.3389)

700 58. Altschul SF, Koonin EV. 1998 Iterated profile searches with PSI-BLAST—a tool for
701 discovery in protein databases. *Trends Biochem. Sci.* **23**, 444–447. (doi:10.1016/S0968-
702 0004(98)01298-5)

703 59. Gubernator B, Bartoszewski R, Kroliczewski J, Wildner G, Szczepaniak A. 2008 Ribulose-
704 1,5-bisphosphate carboxylase/oxygenase from thermophilic cyanobacterium
705 Thermosynechococcus elongatus. *Photosynth. Res.* **95**, 101–109. (doi:10.1007/S11120-007-
706 9240-7/FIGURES/3)

707 60. Li W, Godzik A. 2006 Cd-hit: a fast program for clustering and comparing large sets of protein
708 or nucleotide sequences. *Bioinformatics* **22**, 1658–1659.
709 (doi:10.1093/BIOINFORMATICS/BTL158)

710 61. Katoh K, Misawa K, Kuma KI, Miyata T. 2002 MAFFT: a novel method for rapid multiple
711 sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066.
712 (doi:10.1093/NAR/GKF436)

713 62. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015 IQ-TREE: A Fast and Effective
714 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**,
715 268–274. (doi:10.1093/MOLBEV/MSU300)

716 63. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017 ModelFinder:
717 fast model selection for accurate phylogenetic estimates. *Nat. Methods 2017 146* **14**, 587–589.
718 (doi:10.1038/nmeth.4285)

719 64. Chernomor O, Von Haeseler A, Minh BQ. 2016 Terrace Aware Data Structure for
720 Phylogenomic Inference from Supermatrices. *Syst. Biol.* **65**, 997–1008.
721 (doi:10.1093/SYSBIO/SYW037)

722 65. Camel V, Zolla G. 2021 An Insight of RuBisCO Evolution through a Multilevel Approach.
723 *Biomolecules* **11**, 1761–1761. (doi:10.3390/biom11121761)

724 66. Yang Z. 2007 PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**,
725 1586–1591. (doi:10.1093/MOLBEV/MSM088)

726 67. Aadland K, Pugh C, Kolaczkowski B. 2019 High-Throughput Reconstruction of Ancestral
727 Protein Sequence, Structure, and Molecular Function. In *Computational Methods in Protein*
728 *Evolution* (ed T Sikosek), pp. 135–170. New York, NY: Springer. (doi:10.1007/978-1-4939-
729 8736-8_8)

730 68. Penev PI, Alvarez-Carreño C, Smith E, Petrov AS, Williams LD. 2021 TwinCons:
731 Conservation score for uncovering deep sequence similarity and divergence. *PLOS Comput.*
732 *Biol.* **17**, e1009541–e1009541. (doi:10.1371/JOURNAL.PCBI.1009541)

733 69. Suplatov D, Sharapova Y, Geraseva E, Švedas V. 2020 Zebra2: advanced and easy-to-use
734 web-server for bioinformatic analysis of subfamily-specific and conserved positions in diverse
735 protein superfamilies. *Nucleic Acids Res.* **48**, W65–W71. (doi:10.1093/nar/gkaa276)

736 70. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022 ColabFold:
737 making protein folding accessible to all. *Nat. Methods* **19**, 679–682. (doi:10.1038/s41592-022-
738 01488-1)

739 71. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004
740 UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput.*
741 *Chem.* **25**, 1605–1612. (doi:10.1002/JCC.20084)

742 72. Stec B. 2012 Structural mechanism of RuBisCO activation by carbamylation of the active site
743 lysine. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18785–18790.
744 (doi:10.1073/PNAS.1210754109/SUPPL_FILE/PNAS.201210754SI.PDF)

745 73. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. 2004 Development and testing of a
746 general amber force field. *J. Comput. Chem.* **25**, 1157–1174. (doi:10.1002/JCC.20035)

747 74. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. 2015 ff14SB:
748 Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J.*

Chem. Theory Comput. **11**, 3696–3713. (doi:10.1021/ACS.JCTC.5B00255/SUPPL_FILE/CT5B00255_SI_001.PDF)

75. Salomon-Ferrer R, Case DA, Walker RC. 2013 An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3**, 198–210. (doi:10.1002/WCMS.1121)

76. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. 1983 Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935. (doi:10.1063/1.445869)

77. Eastman P *et al.* 2017 OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13**, e1005659–e1005659. (doi:10.1371/JOURNAL.PCBI.1005659)

78. Woods RJ, Chappelle R. 2000 Restrained electrostatic potential atomic partial charges for condensed-phase simulations of carbohydrates. *J. Mol. Struct. THEOCHEM* **527**, 149–156. (doi:10.1016/S0166-1280(00)00487-5)

79. 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. (doi:10.1002/JCC.21787)

80. Bakan A, Meireles LM, Bahar I. 2011 ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **27**, 1575–1577. (doi:10.1093/BIOINFORMATICS/BTR168)