Systematic exploration of bacterial form I rubisco maximal carboxylation rates

Benoit de Pins[1], Lior Greenspoon[1], Yinon M. Bar-On[1,†], Melina Shamshoum[1], Roee Ben-Nissan[1], Eliya Milshtein[1], Dan Davidi[2,‡], Itai Sharon[3], Oliver Mueller-Cajar[4], Elad Noor[1], Ron Milo[1,*].

[1] Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot 76100, Israel
[2] Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[3] Migal Galilee Research Institute, Kiryat Shmona 11016, Israel
[4] School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore

[†] Currently: Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA
[‡] Currently: Aleph, Tel Aviv-Yafo 6688210, Israel
[*]Corresponding author: ron.milo@weizmann.ac.il

Abstract

Autotrophy is the basis for complex life on Earth. Central to this process is rubisco - the enzyme that catalyzes almost all carbon fixation on the planet. Yet, with only a small fraction of rubisco diversity kinetically characterized so far, the underlying biological factors driving the evolution of fast rubiscos in nature remain unclear. We conducted a high-throughput kinetic characterization of over 100 bacterial form I rubiscos, the most ubiquitous group of rubisco sequences in nature, to uncover the determinants of rubisco's carboxylation velocity. We show that the presence of a carboxysome CO_2 concentrating mechanism correlates with faster rubiscos with a median 5-fold higher rate. In contrast to prior studies, we find that rubiscos originating from α-cyanobacteria exhibit the highest carboxylation rates among form I enzymes (≈10 s^{-1} median versus <7 s^{-1} in other groups). Our study systematically reveals biological and environmental properties associated with kinetic variation across rubiscos from nature.

Introduction

Biological carbon fixation is the gateway for food production and energy storage in the living world. Over 99% of global carbon fixation is catalyzed by rubisco (1), probably the most abundant enzyme in the biosphere (2). Rubisco is mainly divided into four distinct forms (I, II, II/III, and III) and can be found in all domains of life, from plants to algae through autotrophic bacteria and archaea (3). Within this diversity, form I is by far the most abundant of the four forms: it is used by all plants and cyanobacteria and is responsible for almost all CO_2 fixation in nature (4).

Paradoxically, while being the most abundant, rubisco is probably the slowest (i.e. low maximum carboxylation rate - $k_{cat,C}$) central metabolic enzyme (5, 6). A systematic sampling of rubisco's genetic diversity can help grasp the boundaries of its carboxylation rate.

Our group has recently developed an approach to systematically explore the carboxylation rate of natural rubiscos. We use computational methods to select representative rubiscos from the tremendous sequence diversity space. Gene synthesis is then used to generate expression constructs encoding many rubiscos for purification and kinetic characterization. In previous work we showed the feasibility of this approach by exploring form II and II/III rubisco variants (6). We found an uncharacterized rubisco that has a $k_{cat,C}$ higher than all previously-known rubiscos - demonstrating the potential of this approach to stretch the kinetic boundaries of this pivotal enzyme.

In this work, we expand our search to form I rubiscos, which represent ≈97% of known sequences (6). Moreover, their immense ecological and sequence diversity (7–9), limited kinetic data, and higher carboxylation rates in comparison to plant rubiscos (5), make bacterial form I variants particularly interesting. By conducting a first-of-its-kind large-scale study of uncharacterized bacterial form I rubiscos, and leveraging available meta-data on their sequences, we find correlations between contextual factors (phototrophy, carboxysome association) and fast carboxylating rubiscos.

Results

Large scale survey of bacterial form I rubisco

To map the natural diversity of rubisco sequences, we performed an exhaustive search for rubisco homologs across the major genomic and metagenomic public databases. A total of 4300 unique sequences were identified as bacterial form I rubiscos (see Materials and Methods for more details).

By clustering systematically and at progressively higher sequence identity thresholds across different rubisco subgroups (as detailed in the Materials and Methods and Supplementary Fig. 1), we selected 144 rubisco variants that represent the full spectrum of bacterial form I rubisco sequence diversity. In comparison, prior work using active site quantification, a method that allows for precise measurements of turnover rates while considering the enzyme's activation state, has so far characterized only 11 bacterial form I rubiscos (Fig. 1A).

In contrast to form II, II/III and III rubiscos which are composed of a ≈50 kDa large (L) subunit organized as a homodimer, form I rubiscos comprise an additional ≈15 kDa small subunit (S) in an L_8S_8 stoichiometry. Due to this complex oligomeric structure, recombinant expression of form I rubiscos is challenging (10). To improve correct folding in *Escherichia coli*, we coexpressed rubiscos with the chaperone GroEL-GroES, which is known to help reconstitution of bacterial form I rubiscos (11–13). In the case of variants originating from β-cyanobacteria, we coexpressed *rbcL* and *rbcS* together with their cognate chaperone *rbcX* whenever that gene was present in the operon. In addition, because only about a third of these variants were soluble initially, we screened different homologs of the rubisco accumulation factor 1 (Raf1), a chaperone that mediates the assembly of β-cyanobacterial rubiscos (14–17). We identified one homolog, from the bacteria *Euhalothece natronophila*, whose co-expression nearly doubled the number of solubly-expressed β-cyanobacterial rubiscos (Supplementary Fig. 2A). We found that the average rate of β-

cyanobacterial rubiscos was not changed by adding these newly soluble rubiscos (Supplementary Fig. 2B).

The carboxylation rates of each expressed rubisco were determined using a modified version of the spectroscopic coupled assay reported at Davidi *et al. (6)*. Here we directly assayed the crude cell lysates without purifying the enzyme in order to best preserve the quaternary structure of form I rubiscos. The method allows determining the specific carboxylation rate even without purification thanks to the use of the rubisco inhibitor CABP (see Methods and Supplementary Note 1). Since the assay uses high CO_2 levels (4%), which is above the K_M for most rubisco variants (18), measured rates are predicted to approach the $k_{cat,C}$ values. In Supplementary Fig. 3 we compare the rates of five rubisco variants with previously published values to the measurement in our lab, showing a similar ranking of carboxylation rates in spite of the different temperatures and assay methods.

Out of 144 rubisco variants tested, 112 were successfully expressed and soluble. Of which, 98 exhibited significant catalytic activity (which we define as >0.5 reactions per active site per second). The median rate among active form I rubiscos was 5.4 s^{-1}, similar to the median $k_{cat,C}$ of plant rubisco literature values (4.7 s^{-1} when corrected to 30°C by assuming a Q10 value of 2.2 (19)). The fastest rate measured in this study was 14 s^{-1} (Fig. 1B).

Altogether, our measurements achieve a nearly tenfold increase in the number of bacterial form I rubiscos with measured carboxylation rates. As described below, our results allow for the discovery of features associated with fast form I bacterial rubiscos.

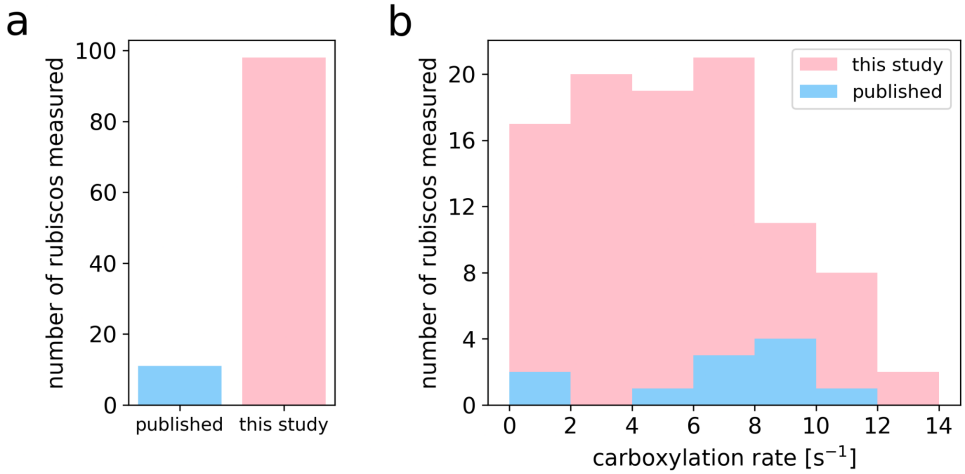

Fig. 1. Systematic exploration of the diversity of bacterial form I rubisco. *A) Number of bacterial form I rubisco variants with a carboxylation rate reported across the literature and in this study. B) Histogram of the carboxylation rates measured in this study and across the literature.*

Form I rubiscos from phototrophic bacteria are faster carboxylases than those from chemoautotrophs

Form I rubisco-expressing bacteria are autotrophs: they can convert oxidized inorganic carbon (CO_2) into the reduced organic compounds forming their biomass through the Calvin cycle. To fuel the energy-intensive reactions involved in this cycle, autotrophs can draw upon two different energy sources: light (photo-autotrophy) or chemical reactions (chemo-autotrophy). Through a literature survey (see Materials and Methods and Supplementary Data 2), we collected available information on the bacteria expressing the studied rubiscos and classified them into phototrophs and chemotrophs. We tested whether one of these trophic modes was linked to higher carboxylation rates. To refrain from selection biases, we aimed to uniformly sample rubiscos from both classes as detailed in the Materials and Methods. We find that rubiscos originating from phototrophs have a carboxylation rate of 6.5 s^{-1} [4.4-7.9 s^{-1}] (median and interquartile range), about three times faster than rubiscos derived from chemoautotrophs (2.0 s^{-1} [1.3-3.9 s^{-1}]) (Fig. 2A, Mann-Whitney U test, $p < 0.01$). The same pattern was observed when taking together all rates measured in this study, without care of uniformly covering both groups (Supplementary Fig. 4).

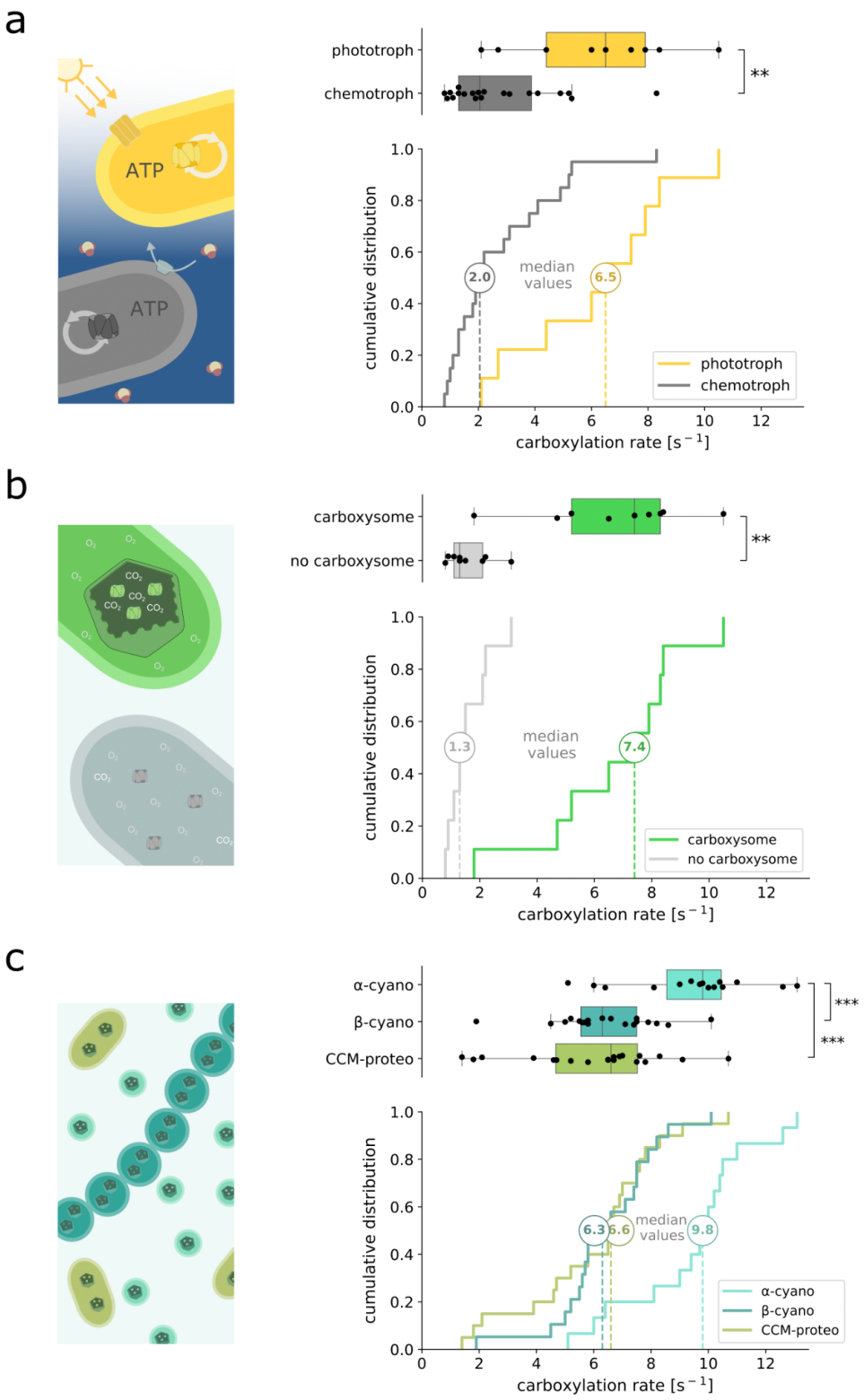

Fig. 2. Large-scale analysis of the biological parameters associated with fast carboxylating rubiscos. *Box and cumulative distribution plots of rubisco carboxylation rates from different clusters: chemo- and phototrophic bacterial rubiscos (A), carboxysome-associated rubiscos and their counterparts (B), α- and β- cyanobacterial, and carboxysome-associated proteobacterial rubiscos (C). To ensure*

*unbiased study of every group, we selected 20 and 9 class-representative chemo- and phototroph-associated rubiscos, 9 and 9 class-representative carboxysome-associated and non-associated rubiscos, and 15, 19 and 20 class-representative α- and β- cyanobacterial, and carboxysome-associated proteobacterial rubiscos respectively (see Materials and Methods). Mann-Whitney U test (A and B) or Kruskal-Wallis followed by Dunn multiple comparison tests (C) were applied. **p < 0.01, ***p < 0.001. Legend abbreviations are as follows: α-cyano, α-cyanobacterial rubisco; β-cyano, β-cyanobacterial rubisco; CCM-proteo, carboxysome-associated proteobacterial rubisco.*

Carboxysome-associated form I rubiscos are significantly faster

A biological parameter that could have influenced rubisco evolution is the presence of a carboxysome-based CO_2 concentrating mechanism (CCM). This cellular mechanism combines the active transport of inorganic carbon into the cell and the colocalization of carbonic anhydrase and rubisco inside subcellular proteinaceous microcompartments called carboxysomes, locally increasing CO_2 concentration around rubisco (20). High CO_2 levels inhibit oxygenation by competitive inhibition, which can permit the use of less CO_2-affine but faster rubisco variants, following the observation of a kinetic tradeoff between these two parameters (21, 22).

To compare carboxysomal and non-carboxysomal form I rubiscos, we uniformly sampled rubiscos from each class and compared their measured carboxylation kinetics (see Materials and Methods). We found that with a median catalytic rate of 7.4 s^{-1} [5.2-8.3 s^{-1}], carboxysome-associated rubiscos are more than 5 times faster than their non-carboxysomal counterparts (1.3 s^{-1} [1.1-2.1 s^{-1}]) (Fig. 2B, Mann–Whitney U test, $p < 0.01$). We found consistent results when taking together all rates measured in this study, i.e. ignoring uniform coverage (Supplementary Fig. 4). An exemplary case within our dataset is the gammaproteobacteria *Hydrogenovibrio kuenenii*, which expresses two form I rubiscos. The carboxysomal rubisco has a rate of 8.3 s^{-1}, which is twice as fast as its non-carboxysomal counterpart (4.2 s^{-1}, Supplementary Data 1), thus exemplifying the impact of carboxysome association on rubisco carboxylation.

Alpha-cyanobacteria express the fastest form I rubiscos across the tree of life

Carboxysomes are known to be expressed by all cyanobacteria and some chemotrophic proteobacteria (our current analysis also showed bioinformatically that it can be found in phototrophic proteobacteria such as the purple sulfur bacteria *Thiorhodococcus drewsii)*. Cyanobacteria can be divided into two sub-clades, α- and β-cyanobacteria (23, 24). It has been posited that α- and β-cyanobacteria rubiscos had identical catalytic rates (25) or that β-cyanobacteria rubiscos are faster (22, 26), as they included the fastest form I rubisco characterized to date (from *Synechococcus elongatus* PCC 6301) (18). However, such statements were made based on scarce measurements, especially among α-cyanobacterial enzymes where only two $k_{cat,C}$ values are currently available (27, 28). We now reevaluate this hypothesis using a wider and systematic kinetic sampling of these different rubisco subforms. We have class representative subsets of 15, 19 and 23 variants uniformly covering α-, β-cyanobacteria, and proteobacteria carboxysome-associated rubiscos diversity from our dataset (see Materials and Methods).

We find that, in contrast to previous statements, α-cyanobacterial rubiscos show the highest carboxylation rates (9.8 s^{-1} [8.6-10.5 s^{-1}]) among all bacterial form I rubiscos, ≈50% higher than

β-cyanobacterial rubiscos and their proteobacterial counterparts (6.3 s^{-1} [5.6-7.5 s^{-1}] and 6.6 s^{-1} [4.7-7.5 s^{-1}] respectively, Fig. 2C, Kruskal-Wallis test followed by Dunn multiple comparison test, $p < 0.001$). We observed the same result when taking together all rates measured in this study, regardless of achieving uniform coverage across groups (Supplementary Fig. 4). Future work can validate this result with direct assays and tests on the impact of other CO_2 concentrations, different temperatures etc.

We further analyzed the correlation between rubisco carboxylation rate and various biological and environmental parameters such as bacterial environmental source, rubisco subtype, bacterial halotolerance, pH, oxygen-sensitivity, or optimal growth temperature. As presented in Supplementary Fig. 5-10, these showed much weaker or no correlations. For example, while pH is known to be crucial for carboxysome-efficiency (29, 30), the optimal growth pH of a bacteria does not show any correlation with its rubisco carboxylation rate (Supplementary Fig. 8), likely as it does not directly affect the tightly controlled intracellular pH. Additionally, the slightly lower carboxylation rate of rubiscos originating from thermophiles (Supplementary Fig. 10) aligns with expectations, considering that these rubiscos naturally work at higher temperatures than in our in vitro assay (30°C). Investigating these specific rubiscos under varying temperatures in subsequent studies could further explore this effect.

To evaluate possible dependencies among the features showing correlations, we performed a joint analysis to see the contribution of each feature while accounting for the other features. In order to achieve this, we trained a random forest regressor model using our dataset to predict rubisco's carboxylation rate based on the main factors explored in this work (see Materials and Methods and Supplementary Fig. 11). The derived Shapley additive explanations (SHAP) values (31, 32) quantify the influence of each feature on the predicted rate (Fig. 3). Among all the features considered in this study, carboxysome-association is by far the most important one for determining the carboxylation rate of form I rubisco. This is followed by belonging to alpha-cyanobacteria. The influence of phototrophy, while present, is only marginal when correcting for the presence or absence of a carboxysome.

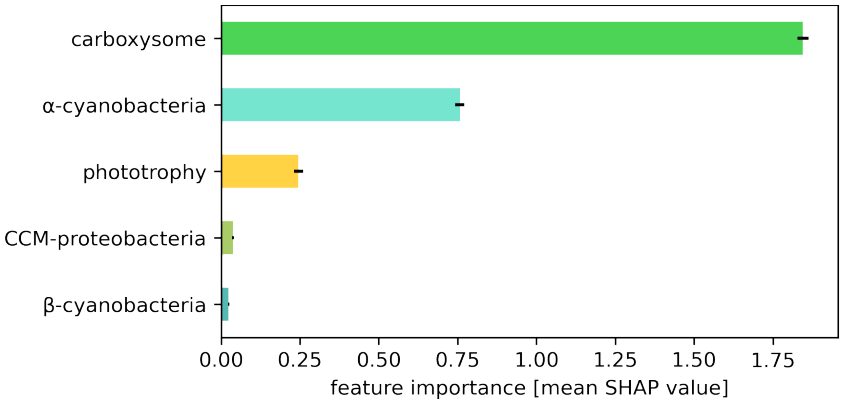

Fig. 3. The presence of a carboxysome is the primary factor influencing form I rubisco carboxylation rate. *Feature importance was determined using absolute SHAP (Shapley additive explanations) values from a random forest regressor model. The model assessed the rubisco carboxylation rate based on bacterial trophic mode, carboxysome-association, and belonging to specific carboxysome-expressing bacterial group: alpha-cyanobacteria, beta-cyanobacteria, or carboxysome-associated proteobacteria (CCM-proteobacteria). Error bars are the standard deviations across 100 different train-test splits.*

Discussion

Rubisco is among the enzymes that have helped shape Earth's biosphere and geosphere the most. Its kinetic parameters result from billions of years of evolution, following (and causing) changes in the atmosphere and climate. We present here a systematic large-scale survey of ≈140 rubisco variants covering the genetic diversity of bacterial form I rubiscos.

We note that the present study is limited by the risk of underestimating some rates if, for instance, enzymes are partially denatured in the expression conditions used. Yet, a denatured enzyme probably could not interact with RuBP or CABP, and more generally, we have no reason to presume *a priori* a systematic bias affecting some of the studied groups over the others.

We find that carboxysome-associated rubiscos are, on average, more than 5 times faster than counterparts which are not associated with a carboxysome. Moreover, among the main parameters examined in this study, it exhibits the strongest association with the occurrence of fast carboxylating rubiscos. Carboxysomes likely evolved during the Proterozoic eon - in the context of the rise of oxygen and the decrease of carbon dioxide in our atmosphere (20) - to maintain carboxylase activity in this changing atmosphere. One evolutionary strategy is the emergence of carbon-concentrating mechanisms (CCMs), including carboxysomes, that locally concentrate CO_2 around rubisco and therefore maintain local gas concentrations favorable to carboxylation. Another strategy consists of the evolution of rubisco towards stronger affinity for CO_2. It has long been postulated that rubisco is a constrained enzyme, limited by catalytic tradeoffs, notably between its carboxylation rate and affinity for CO_2 (illustrated by the positive correlation between $k_{cat,C}$ and $K_{M,CO2}$) (21, 22). Rubiscos that evolved in the context of a CCM might have faced less selective pressure towards stronger CO_2 affinity and have been postulated to present higher carboxylation rates as observed in C_4 versus C_3 plants (33). Our findings with carboxysomal rubiscos support this conjecture using the most comprehensive kinetic sampling to date.

Eventually, among cyanobacteria, α-cyanobacterial rubiscos were found to be 50% faster than their β-cyanobacterial counterparts. One possible difference between α- and β-cyanobacteria could be their cell size. α-cyanobacteria are known to encompass many members of the so-called picocyanobacteria, the smallest cyanobacteria on Earth (25, 34). We collected from the literature the size dimensions of cyanobacteria, when available. With a median cell volume of 0.5 μm^3 (0.3-1.0 μm^3), α-cyanobacteria are more than 25 times smaller in volume than their β counterparts (13.5 μm^3 [3.9-25 μm^3]) (Supplementary Fig. 12, Mann–Whitney U test, $p < 0.0001$). Smaller cells offer higher surface-to-volume ratio and increased exchange with the medium (35), which could support a higher supply of nutrients, and contribute to the evolution of faster rubiscos in these specific cyanobacteria (see Supplementary Note 2 for more details). We also note that α-carboxysomes are smaller than β-carboxysomes, and that their rubiscos are less densely packed than in β-carboxysomes (36). The concentration of CO_2 molecules at rubisco active site may therefore be even higher in α-cyanobacteria.

This study provides a systematic exploration of bacterial form I rubisco maximal rates and its relationship with various contextual factors that could have shaped the evolution of this most abundant enzyme on Earth. It will hopefully help our ability to select and engineer it for human needs.

Acknowledgments

We thank Yoav Peleg, Ron Sender, Noam Prywes, Ralf Steuer, Avi Flamholz and David Savage for important conversations and productive feedback on this manuscript. We thank Michelle Gehring for additional information about unpublished data from their laboratory. This research was supported by the Mary and Tom Beck Canadian Center for Alternative Energy Research, Miel de Botton, the Schwartz Reisman Collaborative Science Program, and the Charles and Louise Gartner Professorial Chair.

Materials and Methods

Rubisco sequence collection

Rubisco large subunit sequences were collected from (i) the NCBI's nr database (37) downloaded in December 2020 and searched following the method described in Davidi *et al. (6)*; (ii) in-house assemblies of the 244 samples from the Tara Oceans expedition (38); (iii) assemblies and rubisco sequences published by (39–43). Sequences outside the length range of 300-700 amino acids were removed. The remaining sequences were then clustered at an 80% sequence identity threshold using USEARCH algorithm (44). Cluster representatives were aligned using MAFFT (v7.475, default parameters) (45), and columns with more than 95% gaps were removed using trimAl (v1.4.rev15, -gt 0.05) (46). A phylogenetic tree was constructed using FastTree (v2.1.10, default parameters) (47). To identify the different rubisco forms, we relied on annotated sequences from NCBI, Tabita *et al.* (48), and Banda *et al.* (43). This process resulted in a total of 72,395 sequences, including 56,161 form I rubiscos, and most notably for this study, 4,302 non-eukaryotic form I rubiscos. The latter were further re-clustered at 90% identity using USEARCH algorithm (custom python script; see below), and a phylogenetic tree was constructed using RAxML (49).

Rubisco variants selection for characterization

Form I rubiscos are divided into 5 separate groups (8, 42, 50–52) (see Supplementary Fig. 6A): forms IA and IB (forming the “green” type, found in cyanobacteria and some proteobacteria); forms ICD and IE (the “red” type, found in proteobacteria); and the recently discovered form I “Anaero” (found in bacteria related to anaerobic, thermophilic Chloroflexaeota and Firmicutes) (53). To comprehensively sample the sequence space of form I rubisco diversity, an iterative clustering approach of the large subunit gene was employed using USEARCH algorithm. The resulting representative sequences from each cluster were selected for characterization. Thresholds were chosen in line with the number of variants we could afford to synthesize and measure in the span of this study. Initially, 32 rubiscos were chosen to cover the entire diversity of form I rubisco at a threshold of 75% identity. Subsequently, further clustering was performed on smaller groups of rubiscos of particular interest, with increasing threshold percentages. Throughout the study, this successively included 38 rubiscos representing the diversity of form IA and B rubisco at 85% identity, 13 rubiscos representing the diversity of cyanobacterial rubisco at 88% identity, 31 rubiscos representing the diversity of IB rubisco at 91% identity, 20 rubiscos representing the diversity of cyanobacterial IA rubisco at 97.5% identity, and 23 rubiscos representing the diversity of proteobacteria carboxysome-associated rubisco at 90% identity (See Supplementary Fig. 1). These representative sequences could sometimes overlap, resulting in a total of 129 different rubisco sequences tested in this study. Additionally, 15 rubiscos were arbitrarily selected for setting-up the experimental pipeline. In total, 144 different rubisco variants were selected for characterization in this study.

Gene synthesis

For each chosen rubisco variant, the complete rubisco operon, encompassing rubisco large and small subunit genes, as well as the chaperone *rbcX* gene for IB rubiscos, was retrieved. The operons were then codon-optimized for expression in *E. coli* (Twist Codon Optimization tool) and synthesized by Twist Bioscience. Following synthesis, these operons were cloned into a pET-29b(+) overexpression vector (NdeI_XhoI insertion sites). Validation of gene synthesis and cloning was conducted through next-generation sequencing as part of the Twist bioscience service.

High-throughput rubisco expression

Chemocompetent BL21(DE3) cells, previously transformed with a pESL plasmid coding for the chaperone GroEL-GroES (12), were transformed with the rubisco library and incubated at 37°C, 250 rpm in 8 ml of LB media supplemented with 30 µg/ml chloramphenicol and 50 µg/ml kanamycin. Growth was performed in 24-deep-well plates. When cells reached an OD600 of 0.6, GroEL-GroES expression was induced by adding arabinose (0.2% final) and incubating at 23°C for 45 min. Rubisco expression was then induced by adding 0.2 mM IPTG (isopropyl β-d-thiogalactoside) and incubating at 23°C for 21 h. For protein extraction, cells were harvested by centrifugation (15 min; 4,000 g; 4°C) and pellets were lysed with BugBuster® ready mix (Millipore) for 25 min at room temperature; 70 µl BugBuster master mix (EMD Millipore) was added to each sample. Crude extracts were then centrifuged for 30 min at 4,000 g, 4°C to remove the insoluble fraction. For quality control, all samples were run on an SDS–PAGE gel (Supplementary Fig. 13).

Raf1-IB rubisco co-expression

For IB rubiscos (originating from β-cyanobacteria), a pilot study was performed to find an homologue of the chaperone Rubisco accumulation factor 1 (Raf1), that could help solubly-express these rubisco variants. 3 different Raf1 were first tested with their cognate rubisco, from 3 β-cyanobacteria (namely, *Trichormus variabilis*, *Pseudanabaena sp. PCC 6802*, and *Euhalothece natronophila*). Codon-optimized *raf1* genes were synthesized and cloned by Twist Bioscience in polycistron with their cognate rubisco genes into pET-29b(+) vectors. Proteins were expressed and tested *in vitro*, as described above, and Raf1 from *Euhalothece natronophila* was chosen, based on ability to solubilize its cognate rubisco. The gene was therefore cloned into a pESL plasmid, in the same operon as GroEL-ES genes, and transformed into chemocompetent BL21(DE3) cells. IB rubiscos were then transformed and expressed in these cells as described previously.

High-throughput determination of rubisco carboxylation rates

To determine rubisco carboxylation rates, we performed kinetic assays directly from the soluble fraction of prepared lysates since purifying both large and small rubisco subunits together was not feasible in a high-throughput manner. Soluble fractions were incubated with 4% CO_2 and 0.4% O_2 for rubisco activation (15 min; 4°C; plate shaker at 250 rpm). Following activation, rubisco carboxylase activities were tested as described in Davidi *et al. (6)*. Briefly, 10 µl of the activated rubisco sample was added to six aliquots of 80 µl of assay mix (a detailed list of all assay components and their sources is provided in Supplementary Table 1) containing different concentrations of CABP (0, 0, 10, 20, 30, and 90 nM). The mix was incubated for 15 minutes at 30°C in a plate reader (Infinite® 200 PRO; TECAN) connected to a gas control module (TECAN pro200) ensuring atmospheric conditions of 4% CO_2 and 0.4% O_2. Rubisco carboxylation activity was initiated by adding 10 µl RuBP to each sample (final concentration of 1 mM and a total volume of 100 µl). The carboxylation rate was determined through a coupled reaction (54, 55). In brief, 3-phosphoglycerate, the product of ribulose 1,5-bisphosphate carboxylation, was phosphorylated and subsequently reduced into glyceraldehyde 3-phosphate, involving NADH oxidation. The latter could be monitored through 340 nm absorbance for 15 minutes at 2-second intervals. To convert from NADH to rubisco reactions per second, we assumed a stoichiometric ratio of 2:1 between NADH and carboxylation reactions. The active-site concentration was determined by fitting a linear regression model (custom python script; see below) to the measured reaction rates as a function of the CABP inhibitor concentrations (See Supplementary Note 1 and Supplementary Fig. 14 for more details). Due to the use of soluble fractions of total cell lysates, the initial concentration of rubisco could not be determined beforehand, which often led to saturation of the first assay with rubisco. To overcome this issue, we adjusted the concentration of lysates by dilution, to obtain a rubisco concentration that enabled measurable inhibition by CABP, allowing for an accurate quantification of the active-site concentration. We finally obtained the rate per active site by dividing the activity with no CABP by the concentration of active sites. As was done in Davidi *et al. (6)*, and since not all variants were tested on the same day, the form-II rubisco from *R. rubrum*, commonly used as a standard in rubisco kinetic assays, was consistently included as an internal reference in every measurement.

Bacterial, ecological and rubisco data collection

Data were collected on the bacteria expressing rubiscos studied in this work. A literature survey was performed to gather as much available information on each bacteria. Additionally, sample data associated with these bacteria were collected on NCBI (BioSample database). Furthermore, the identification of carboxysome and non-carboxysome-associated rubiscos was carried out by systematically examining the presence of carboxysome genes following the small subunit gene of rubisco, *rbcS*. All collected information and references are presented in Supplementary Data 2.

Rubisco variants selection for comparative analysis

For comparing rubisco carboxylation rates between groups without sampling bias, we aimed to select representative sets of variants uniformly covering the different groups studied. For the comparison of rubiscos originating from phototrophic versus chemotrophic organisms, we selected a set of 32 rubiscos, covering the entire diversity of form I rubiscos at a 75% similarity threshold, enriched with a set of 36 rubiscos, covering the entire diversity of form IA and B rubiscos at an 85% similarity threshold. Among these 68 variants, phototrophic rubiscos were distinguished from chemotrophic ones based on their co-occurrence with photosystem genes in the respective genome. This resulted in 53 chemotropic and 15 phototrophic rubiscos, of which 20 and 9 respectively were soluble and active *in vitro*.

For the comparison of carboxysome and non-carboxysome-associated rubiscos, we selected a set of 14 rubiscos covering the entire diversity of carboxysome-associated rubiscos at an 85% similarity threshold, and of 30 rubiscos covering the entire diversity of non-carboxysome-associated rubiscos at a 75% similarity threshold. 9 and 9 of them were soluble and active *in vitro*.

For the comparison of α- and β-cyanobacterial, and carboxysome-associated proteobacterial rubiscos, we selected 19 rubiscos covering the entire diversity of α-cyanobacteria rubiscos at a 97.5% similarity threshold, 29 rubiscos covering the entire diversity of β-cyanobacteria rubiscos at a 91% similarity threshold, and 23 rubiscos covering the entire diversity of carboxysome-associated proteobacterial rubiscos at a 90% similarity threshold. 15, 19 and 20 of them respectively were soluble and active *in vitro*.

Random forest regressor model and feature importance analysis

A random forest regressor model was used to predict rubisco's carboxylation rate as a function of the main features showing correlation in the study: the trophic mode of the bacteria (photo or chemotrophy), the association of the rubisco with a carboxysome, and, among those harboring a carboxysome, those belonging to proteobacteria, α- or β-cyanobacteria. One hundred individual decision trees were trained with a maximum depth of 3 and a fixed random seed of 42. In each tree generation, the dataset was randomly split into training (75% of the data) and testing (25%) sets. For every generated tree, Shapley additive explanations (SHAP) values were computed for each estimator. SHAP values represent the average contribution of each feature to the difference between the model's prediction and the expected prediction. The SHAP values of each parameter were eventually averaged and plotted for feature importance comparison.

418 **Data and Code Availability**

419 All the data supporting the findings of this study as well as the codes used for generating our list
420 of rubiscos and for analyzing the results is open source and can be found in the following link:
421 https://gitlab.com/milo-lab-public/rubisco-F1.
422
423

424 ## References

425 1. J. A. Raven, Contributions of anoxygenic and oxygenic phototrophy and chemolithotrophy
426 to carbon and oxygen fluxes in aquatic environments. *Aquat. Microb. Ecol.* **56**, 177–192
427 (2009).

428 2. Y. M. Bar-On, R. Milo, The global mass and average rate of rubisco. *Proc. Natl. Acad. Sci.*
429 *U. S. A.* **116**, 4738–4743 (2019).

430 3. N. Prywes, N. R. Phillips, O. T. Tuck, L. E. Valentin-Alvarado, D. F. Savage, Rubisco
431 Function, Evolution, and Engineering. *Annu. Rev. Biochem.* (2023)
432 https:/doi.org/10.1146/annurev-biochem-040320-101244.

433 4. S. G. Wildman, Along the trail from Fraction I protein to Rubisco (ribulose bisphosphate
434 carboxylase-oxygenase). *Photosynth. Res.* **73**, 243–250 (2002).

435 5. A. I. Flamholz, *et al.*, Revisiting Trade-offs between Rubisco Kinetic Parameters.
436 *Biochemistry* **58**, 3365–3376 (2019).

437 6. D. Davidi, *et al.*, Highly active rubiscos discovered by systematic interrogation of natural
438 sequence diversity. *EMBO J.* **39**, e104081 (2020).

439 7. F. R. Tabita, Microbial ribulose 1, 5-bisphosphate carboxylase/oxygenase: a different
440 perspective. *Photosynth. Res.* **60**, 1–28 (1999).

441 8. M. R. Badger, E. J. Bek, Multiple Rubisco forms in proteobacteria: their functional
442 significance in relation to CO2 acquisition by the CBB cycle. *Journal of Experimental*
443 *Botany* **59**, 1525–1541 (2008).

444 9. B. Witte, *et al.*, Functional prokaryotic RubisCO from an oceanic metagenomic library. *Appl.*
445 *Environ. Microbiol.* **76**, 2997–3003 (2010).

446 10. H. Aigner, *et al.*, Plant RuBisCo assembly in E. coli with five chloroplast chaperones
447 including BSD2. *Science* **358**, 1272–1278 (2017).

448 11. C. Liu, *et al.*, Coupled chaperone action in folding and assembly of hexadecameric
449 Rubisco. *Nature* **463**, 197–202 (2010).

450 12. P. Goloubinoff, J. T. Christeller, A. A. Gatenby, G. H. Lorimer, Reconstitution of active
451 dimeric ribulose bisphosphate carboxylase from an unfolded state depends on two
452 chaperonin proteins and Mg-ATP. *Nature* **342**, 884–889 (1989).

453 13. D. N. Greene, S. M. Whitney, I. Matsumura, Artificially evolved Synechococcus PCC6301
454 Rubisco variants exhibit improvements in folding and catalytic efficiency. *Biochem. J* **404**,
455 517–524 (2007).

14. F. Huang, *et al.*, Rubisco accumulation factor 1 (Raf1) plays essential roles in mediating Rubisco assembly and carboxysome biogenesis. *Proceedings of the National Academy of Sciences* **117**, 17418–17428 (2020).

15. L.-Y. Xia, *et al.*, Molecular basis for the assembly of RuBisCO assisted by the chaperone Raf1. *Nature Plants* **6**, 708–717 (2020).

16. T. Hauser, *et al.*, Structure and mechanism of the Rubisco-assembly chaperone Raf1. *Nat. Struct. Mol. Biol.* **22**, 720–728 (2015).

17. P. Kolesinski, I. Belusiak, M. Czarnocki-Cieciura, A. Szczepaniak, Rubisco Accumulation Factor 1 from Thermosynechococcus elongatus participates in the final stages of ribulose-1,5-bisphosphate carboxylase/oxygenase assembly in Escherichia coli cells and in vitro. *FEBS J.* **281**, 3920–3932 (2014).

18. Y. Savir, E. Noor, R. Milo, T. Tlusty, Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3475–3480 (2010).

19. Y.-P. Cen, R. F. Sage, The regulation of Rubisco activity in response to variation in temperature and atmospheric CO2 partial pressure in sweet potato. *Plant Physiol.* **139**, 979–990 (2005).

20. A. Flamholz, P. M. Shih, Cell biology of photosynthesis over geologic time. *Curr. Biol.* **30**, R490–R494 (2020).

21. D. B. Jordan, W. L. Ogren, Species variation in the specificity of ribulose biphosphate carboxylase/oxygenase. *Nature* **291**, 513–515 (1981).

22. C. Iñiguez, *et al.*, Evolutionary trends in RuBisCO kinetics and their co-evolution with CO2 concentrating mechanisms. *Plant J.* **101**, 897–918 (2020).

23. M. R. Badger, G. D. Price, CO2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. *J. Exp. Bot.* **54**, 609–622 (2003).

24. L. Whitehead, B. M. Long, G. D. Price, M. R. Badger, Comparing the in vivo function of α-carboxysomes and β-carboxysomes in two model cyanobacteria. *Plant Physiol.* **165**, 398–411 (2014).

25. P. J. Cabello-Yeves, *et al.*, α-cyanobacteria possessing form IA RuBisCO globally dominate aquatic habitats. *ISME J.* (2022) https:/doi.org/10.1038/s41396-022-01282-z.

26. N. D. Nguyen, *et al.*, Towards engineering a hybrid carboxysome. *Photosynth. Res.* **156**, 265–277 (2023).

27. P. M. Shih, *et al.*, Biochemical characterization of predicted Precambrian RuBisCO. *Nat. Commun.* **7**, 10382 (2016).

28. B. M. Long, *et al.*, Carboxysome encapsulation of the CO2-fixing enzyme Rubisco in tobacco chloroplasts. *Nat. Commun.* **9**, 3570 (2018).

29. N. M. Mangan, A. Flamholz, R. D. Hood, R. Milo, D. F. Savage, pH determines the energetic efficiency of the cyanobacterial CO_2 concentrating mechanism. *Proceedings of*

the National Academy of Sciences **113**, E5354–E5362 (2016).

30. B. M. Long, B. Förster, S. B. Pulsford, G. D. Price, M. R. Badger, Rubisco proton production can drive the elevation of CO2 within condensates and carboxysomes. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021).

31. S. M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions in *Advances in Neural Information Processing Systems*, I. Guyon, *et al.*, Eds. (Curran Associates, Inc., 2017).

32. S. M. Lundberg, *et al.*, From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56–67 (2020).

33. J. R. Seemann, M. R. Badger, J. A. Berry, Variations in the specific activity of ribulose-1,5-bisphosphate carboxylase between species utilizing differing photosynthetic pathways. *Plant Physiol.* **74**, 791–794 (1984).

34. I. Jasser, C. Callieri, "Picocyanobacteria" in *Handbook of Cyanobacterial Monitoring and Cyanotoxin Analysis*, (John Wiley & Sons, Ltd, 2017), pp. 19–27.

35. Young Kevin D., The Selective Value of Bacterial Shape. *Microbiol. Mol. Biol. Rev.* **70**, 660–703 (2006).

36. B. D. Rae, B. M. Long, M. R. Badger, G. D. Price, Functions, compositions, and evolution of the two types of carboxysomes: polyhedral microcompartments that facilitate CO2 fixation in cyanobacteria and some proteobacteria. *Microbiol. Mol. Biol. Rev.* **77**, 357–379 (2013).

37. D. A. Benson, *et al.*, GenBank. *Nucleic Acids Res.* **41**, D36–42 (2013).

38. S. Sunagawa, *et al.*, Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

39. K. C. Wrighton, *et al.*, Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**, 1661–1665 (2012).

40. C. T. Brown, *et al.*, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

41. K. Anantharaman, *et al.*, Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 1–11 (2016).

42. F. R. Tabita, S. Satagopan, T. E. Hanson, N. E. Kreel, S. S. Scott, Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *J. Exp. Bot.* **59**, 1515–1524 (2008).

43. D. M. Banda, *et al.*, Novel bacterial clade reveals origin of form I Rubisco. *Nat Plants* **6**, 1158–1166 (2020).

44. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

45. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

46. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

47. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

48. F. R. Tabita, T. E. Hanson, S. Satagopan, B. H. Witte, N. E. Kreel, Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 2629–2640 (2008).

49. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

50. R. J. Spreitzer, M. E. Salvucci, Rubisco: structure, regulatory interactions, and possibilities for a better enzyme. *Annu. Rev. Plant Biol.* **53**, 449–475 (2002).

51. S. W. Park, *et al.*, Presence of duplicate genes encoding a phylogenetically new subgroup of form I ribulose 1,5-bisphosphate carboxylase/oxygenase in Mycobacterium sp. strain JC1 DSM 3803. *Res. Microbiol.* **160**, 159–165 (2009).

52. A. Grostern, L. Alvarez-Cohen, RubisCO-based CO2 fixation and C1 metabolism in the actinobacterium Pseudonocardia dioxanivorans CB1190. *Environ. Microbiol.* **15**, 3040–3053 (2013).

53. L. Schulz, *et al.*, Evolution of increased complexity and specificity at the dawn of form I Rubiscos. *Science* **378**, 155–160 (2022).

54. R. M. Lilley, D. A. Walker, The reduction of 3-phosphoglycerate by reconstituted chloroplasts and by chloroplast extracts. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **368**, 269–278 (1974).

55. D. S. Kubien, C. M. Brown, H. J. Kane, Quantifying the amount and activity of Rubisco in leaves. *Methods Mol. Biol.* **684**, 349–362 (2011).