Title
Analysis and Reconstitution of a Bacterial CO2-Concentrating Mechanism

Author
Flamholz, Abraham

Publication Date
2019

Peer reviewed|Thesis/dissertation

Analysis and Reconstitution of a Bacterial CO_2-Concentrating Mechanism

by

Abraham I. Flamholz

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor David Savage, Chair
Assistant Professor Hernan Garcia
Professor Krishna Niyogi
Professor Jeremy Thorner

Spring 2019

Analysis and Reconstitution of a Bacterial CO_2-Concentrating Mechanism

Abstract

Analysis and Reconstitution of a Bacterial CO_2-Concentrating Mechanism

by

Abraham I. Flamholz

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Associate Professor David Savage, Chair

Ribulose Bisphosphate Carboxylase/Oxygenase (Rubisco) is the central enzyme of the Calvin-Benson-Bassham (CBB) cycle and the most abundant enzyme on Earth. Nearly all carbon enters the biosphere via Rubisco carboxylation, yet no Rubisco has a maximum carboxylation rate above 15 s^{-1} and all Rubiscos catalyze a competing oxygenation of their five-carbon substrate. Evolution seems to have overcome this problem by selecting for CO_2 concentrating mechanisms (CCMs) that place Rubisco in compartments where CO_2 is highly concentrated. Distinct families of CCMs are found among bacteria, algae and plants, but they all function by elevating CO_2 to promote carboxylation and inhibit oxygenation by Rubisco.

Cyanobacteria are the ancestors of all green photosynthetic lineages, including plant chloroplasts, and have a CCM that requires a large ($>$ 200 MDa) proteinaceous organelle called the carboxysome. Carboxysomes are composed of ≈10,000 proteins and encapsulate ≈2000 Rubisco active sites. Although present-day atmosphere is rich in O_2 (21%) and CO_2-poor (0.04%), cyanobacteria use the CCM to grow robustly in air. Mutations to the CCM abrogate air growth and are only rescued by markedly elevated CO_2. By contrast, few plant species have CCMs. Land plants with CCMs, such as maize, are particularly productive, but no plant CCMs resemble those in bacteria. These considerations raise the prospect that engineering plants to express bacterial CCMs might improve their growth. Motivated by these questions, I use complementary experimental, informatic and modeling approaches to study the evolution and function of the bacterial CCM.

I first perform a meta-analysis of the kinetic properties of ≈ 300 distinct Rubiscos and show that carboxylation and oxygenation are inextricably linked: increasing carboxylation efficiency entails an equal increase to oxygenation efficiency, suggesting that oxygenation cannot be avoided by mutating Rubisco itself. Second, by mathematical modeling I show that the ≈ 20 known components of the bacterial CCM are sufficient to concentrate CO_2 and promote fast carboxylation *in silico*. As pH has wide-ranging effects on inorganic carbon chemistry, I also demonstrate that pH must be considered for CCM models to produce plausible results matching physiological measurements of cyanobacteria.

To confirm that the CCM is formed of a small number of genes, I designed a high-throughput genetic screen in the chemotrophic bacterium *H. neapolitanus*. This screen identified 20 genes in 3 operons directly involved in the *H. neapolitanus* CCM, including a novel class of inorganic carbon transporters I term "DABs." To test whether these genes are sufficient for CCM function, I developed a Rubisco-dependent *E. coli* strain, CCMB1, whose growth reports on CCM function *in vivo*. CCMB1 only grows in minimal media when Rubisco is expressed, and only under elevated CO_2 ($>$ 20x ambient). I found that expression of the *H. neapolitanus* CCM genes permits CCMB1 growth in ambient air, thereby establishing a facile system for testing the contributions of individual genes to the functioning of the bacterial CCM.

For Dad. I miss you.

Contents

List of Figures

List of Tables

Acknowledgments

The past six years have been both a privilege and a trial for me. I knew that I wanted to become an experimental scientist, but the transition from computer science, applied math and theoretical biology was not without difficulties. In the process of pursuing this degree and this work, I managed to transform myself, I believe, into a person who is competent to design and execute experiments and, more importantly, into a person whose intuition is to pursue an empirical question by experiment.

During my six years at Berkeley, I also lost my father Jack to pancreatic cancer. I am deeply grateful to my PhD advisor, David Savage for his patience and flexibility, which enabled me to pursue research fully without sacrificing family. Thanks to Dave and trans-continental air travel, I was able to spend good time with my father in his last years. This work is dedicated, first and foremost, to my father Jack Flamholz, who taught me that intellectual pursuits can also be spiritual, that awe is the appropriate orientation to nature, who taught me from early age "the difference between knowing the name of something and knowing something."

I would like to extend special thanks to all my teachers and advisers, in particular to Ron Milo and David Savage. Ron took a chance on me and gave me my first scientific home in the Milo lab at the Weizmann Institute. I spent only two years at the Weizmann, but they affected me deeply. Ron - thank you for teaching me to take pleasure in the process and people, to enjoy the small lessons that can be gleaned from every paper, every experiment and every group hike. Thanks also for the friendship and collaboration of whole WIS crew - Elad Noor, Arren Bar-Even, Niv Antonovsky, Lior Zelcbuch, Shira Amram-Weinreb, Ayelet Levin, Uri Moran, Alon Shepon and many others in the lab and department. It is a testament to Ron's perspicacity that I also want to thank several current members of the Milo lab who did not overlap with me, but have come to be great friends and excellent research collaborators: Dan Davidi, Shmuel Gleizer, and Yinon Bar-On.

To Dave: thank you for all your support over these years. Thank you for editing my NSF fellowship when it wasn't even clear that I would come to Berkeley. Thank you for giving me the freedom to pursue big projects that might not work. May it pay off for us both. Thank you for bringing together a lab environment filled with friends in science: Dana Nadler, Rachel Hood, Caleb Cassidy-Amstutz, Sean Higgins, Luke Oltrogge, Rob Nichols, Cissi Blikstad, Julia Borden, Benz Chaijarasphong, Jack Desmarais, Emeric Charles, Shin Kim and Noam Prywes in no particular order. Thanks also to our hilarious neighbors in the Coates lab, especially to Anna Engelbrektson, Maggie Stoeva, Yi Liu and Anchal Mehra for their bountiful-if-sardonic help with anaerobic experiments. Special thanks to Jack Desmarais and our most gregarious lab manager and technician, Eli Dugan, who did an enormous amount of the work described in Chapters 4 and 5. Also to our graduated but never-forgotten undergraduate researcher Sumedha Ravishankar (now pursuing a PhD at UCSD) who developed some of the *E. coli* reporter strains we now use routinely. It has been a tremendous pleasure to watch these younger scientists develop and to take small part in their training.

Thanks to Niall Mangan, my collaborator on the work presented in Chapter 3, for teaching me so much about modeling dynamical systems and for her continued support in my scientific pursuits. I don't know if there is a word for "slightly older peers," but there should be because I owe them all deep gratitude. Especially to Elad Noor, Arren Bar-Even, Niv Antonovsky, Lior Zelcbuch and Shira Amram, who taught me everything I knew about biology coming into Berkeley. Also to Eric Estrin, Dan Arlow, Antony Lee, Cissi Blikstad, Ofer Rog, Patrick Shih, and Jeremy Roop, whose friendship has made science joyful and whose advice helped me navigate graduate school in its many confusions.

I would also like to thank my thesis committee members Profs. Hernan Garcia, Jeremy Thorner, and Kris Niyogi for their feedback and support. Particular thanks to Hernan Garcia and Kris Niyogi who supplied reagents and very detailed directional advice in recent years, and to Mike Eisen, Woodward Fischer, Dan Portnoy and Dan Tawfik, my secondary thesis committee, who have provided support, encouragement and enthusiastic conversations over the years.

There are too many others to name, but some cannot be forgotten. Especially family. Thank you to my brother Bezalel and sister Eta. It has been such a privilege to come to know you as adults here in the Bay Area. I love you both. Thanks to my mother Beverly, who has carried me since before my birth. These have been hard years and I know that it has not been easy for us to be far away, Mom. May we find lighter years ahead. To my love Rachel, whose light heart and warm smile has carried me these last three years. May we find our way into the future.

Chapter 1

Introduction

1.1 Rubisco - the Primary Carboxylase of the Biosphere

After hydrogen and oxygen, which are mostly found in water, carbon is the most abundant element in living bodies [194]. Since organic molecules contain carbon-carbon bonds, organic carbon is, by definition, more reduced than CO_2. But reduced carbon is not stable in our contemporary high-oxygen atmosphere: oxidation of organic carbon to CO_2 is very favorable. Complete oxidation of glucose by O_2 to form CO_2 and H_2O, for example, is associated with a standard change in Gibbs free energy $\Delta G'^{\circ} \approx -2900$ kJ/mol [102]. This basic principle forms the basis of the biological carbon cycle. Photoautotrophic organisms like plants use light energy to reduce (fix) inorganic carbon (mostly CO_2) to make the organic molecules that form their bodies. These molecules are assimilated (eaten) by heterotrophs, which use them to produce biomass and oxidize them (partially or fully) to capture the energy required for growth and homeostasis. The net result is that CO_2, light energy, and water are converted to organic carbon and O_2 by photoautotrophs. Heterotrophs use O_2 to re-oxidize photosynthesized organic carbon, making water and CO_2 again - a futile cycle of carbon fixation and respiration whose byproduct is nearly all of life on Earth.

Nearly all biological carbon fixation ($\approx 99\%$) is due to chlorophyll-based oxygenic photosynthesis coupled to the Calvin-Benson-Bassham (CBB) cycle [27, 240]. Photosynthesis captures light energy to withdraw electrons from water and produce the energy carriers (e.g. ATP) and reducing equilvalents (e.g. NADPH) required to run the CBB cycle (Figure 1.1B). Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase (Rubisco) is the primary carbon-fixing enzyme (carboxylase) of the CBB cycle, the carbon fixation cycle responsible for growth throughout the green photosynthetic lineage and many other autotrophic taxa. So Rubisco is the ultimate source of nearly all carbon atoms entering the biosphere. We now know that Rubisco catalyzes both carboxylation and oxygenation of its five carbon substrate, ribulose 1,5-bisphophate (RuBP), followed by hydration and scission of a C-C bond (Figure 1.1A, [10]). Carboxylation and oxygenation reactions are thought to follow parallel trajectories,

though the precise details of particular microscopic steps remain under some contention [10]. Nonetheless, it is clear that oxygenation and carboxylation mechanisms must differ in their details because triplet O_2 must be activated prior to oxygenation [300].

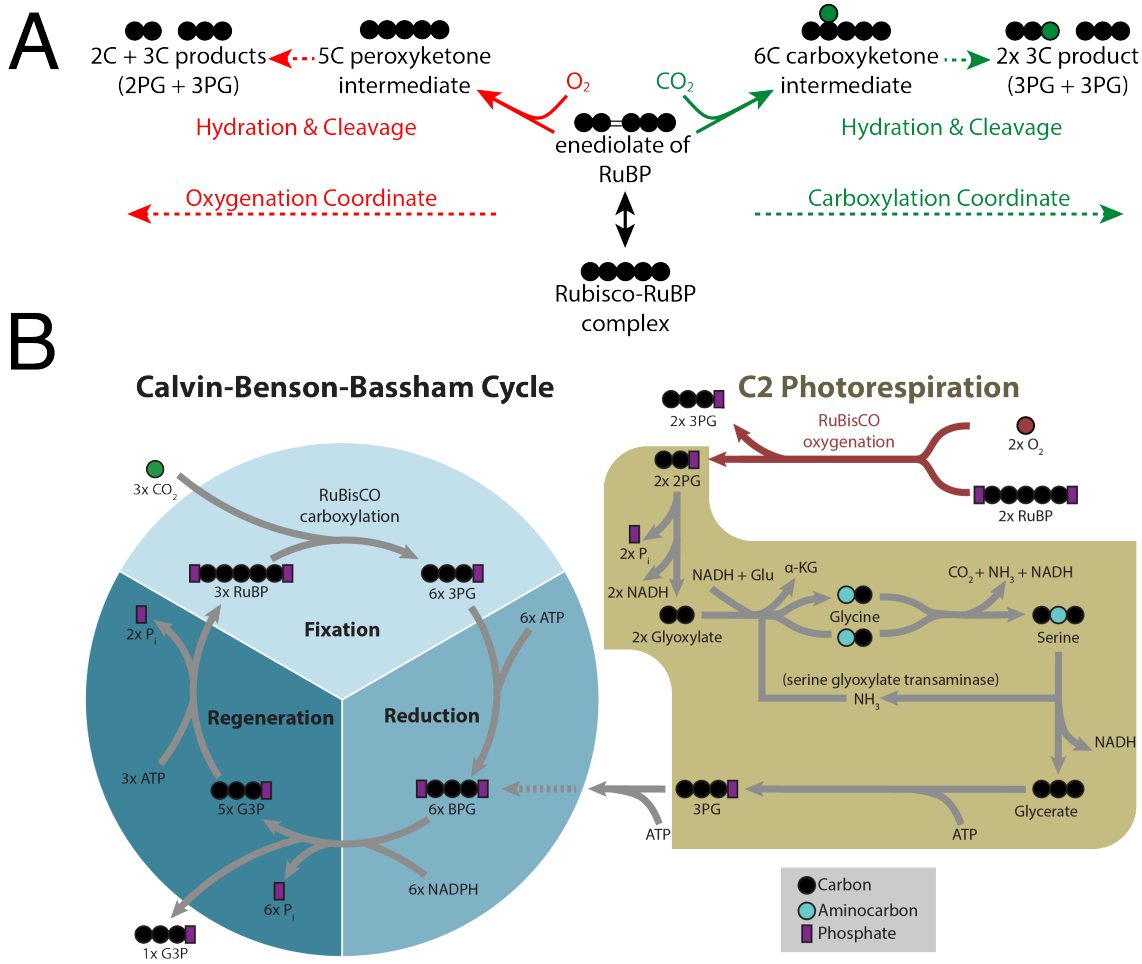


Figure 1.1: Mechanism and metabolic context of Rubisco. All known Rubiscos are capable of catalyzing both carboxylation and oxygenation of RuBP. Though the detailed mechanisms of carboxylation and oxygenation differ, these reactions are catalyzed via parallel trajectories diagrammed in panel A, where gas addition is followed by hydration and cleavage of a C-C bond, producing two new carboxylic acids. When CO_2 is added to RuBP, two molecules of 3-phosphoglycerate (3PG) are produced. Panel B shows that both 3PG molecules can proceed through the reductive phase of the CBB cycle, where they are phosphorylated and reduced to bisphosphoglycerate (BPG) to form glyceraldehyde 3-phosphate (G3P). Five molecules of G3P are condensed into three RuBP by a series of reactions in the regenerative arm of the cycle. Every three turns of the CCB cycle produces one excess G3P that can be diverted from the cycle to form biomass. However, when Rubisco oxygenates it produces one 3PG and one molecule of 2-phosphoglycolate (2PG). 2PG is not part of the CBB cycle and is not used to generate biomass, so 2PG must be recycled by a photorespiratory pathway to avoid the total loss of two reduced carbons. The canonical C_2 photorespiratory pathway condenses two 2PG to ultimately produce a 3PG which can re-enter the CBB cycle. The C_2 is metabolically expensive - consuming ATP, NADH and reduced nitrogen - and results in the loss of one carbon for every two 2PG recycled.

Following the winding path of carbon through the CBB cycle can be challenging, but the cycle is nicely summarized in three phases [30]. In the **fixation phase**, carboxylation

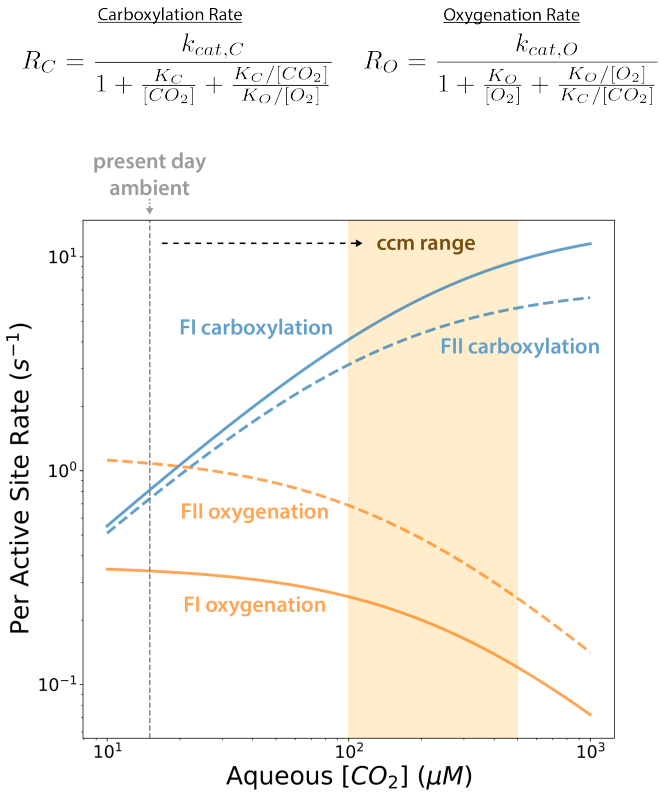


Figure 1.2: Effect of CO_2 concentration on the rates of carboxylation and oxygenation by Rubisco. Carboxylation (blue) and oxygenation rates (orange) were calculated based on an effective Michaelis-Menten model of Rubisco catalysis, which includes competitive inhibition of carboxylation by oxygenation and vice-versa. Solid lines give rates calculated for the FIB Rubisco from *Synechococcus elongatus* PCC6301 [212]. Dashed lines are for the FII Rubisco from *Rhodospirillum rubrum* [262]. The O_2 concentration was assumed to be 270 μM. A CO_2 concentrating mechanism (CCM) elevates the CO_2 concentration, which has the dual effects of saturating carboxylation and suppressing oxygenation (yellow region). On these axes FII Rubiscos appear strictly worse than FI in that they oxygenate more and carboxylate less at all CO_2 levels. Consistent with this observation, FII Rubiscos are typically found in anaerobic bacteria.

of 3 RuBP by Rubisco produces 6 molecules of 3-phosphoglycerate (3PG, Figure 1.1A-B). Phosphorylation and reduction of 3PG in the **reduction phase** allows for the production of 6 glyceraldehyde 3-phosphate (G3P). Since Rubisco carboxylated three RuBP, three of 18 carbons found in 6 G3P molecules are derived from CO_2. Only 5 molecules of G3P (15 carbons) are required to produce 3 molecules of RuBP (15 carbons), so one G3P (3 carbons) can be withdrawn from the cycle to produce biomass. The remaining 5 G3P are condensed and rearranged through a series of reactions in the **regeneration phase** to recover the 3 RuBP that we started with.

Oxygenation by Rubisco poses a problem for the continued operation of the CBB cycle. Oxygenation of RuBP produces one 3PG and one 2-phosphoglycolate (2PG, 1.1A). 2PG is not part of the CBB cycle and not used to produce biomass in most organisms, so it represents a loss of two carbons from the CBB cycle. 2PG is also considered "toxic" because knockout of chloroplast 2PG export induces bleaching and greatly inhibits growth [225, 289]. This effect is likely due to a regulatory interaction with CBB enzymes [9, 149] and not a spontaneous chemical activity of 2PG. Rather, in high O_2 or low CO_2 concentrations, oxygenation by Rubisco would cause 2PG to accumulate to levels that adversely affect net carbon fixation and downstream regulation of photosynthesis.

All plants and cyanobacteria express pathways to recycle 2PG [84, 117]. The canonical C_2 pathway ameliorates both problems - carbon loss and 2PG accumulation - by condensing

and metabolizing two 2PG to ultimately form one CO_2 and one 3PG, which can re-enter the CBB cycle. As shown in Figure 1.1B, this pathway entails the loss of one carbon for every two 2PG and carries with it substantial metabolic cost, consuming ATP, reducing equivalents and reduced nitrogen. As such, excessive photorespiration is considered "wasteful" and several labs are working to improve photorespiratory efficiency in crop plants [160, 290].

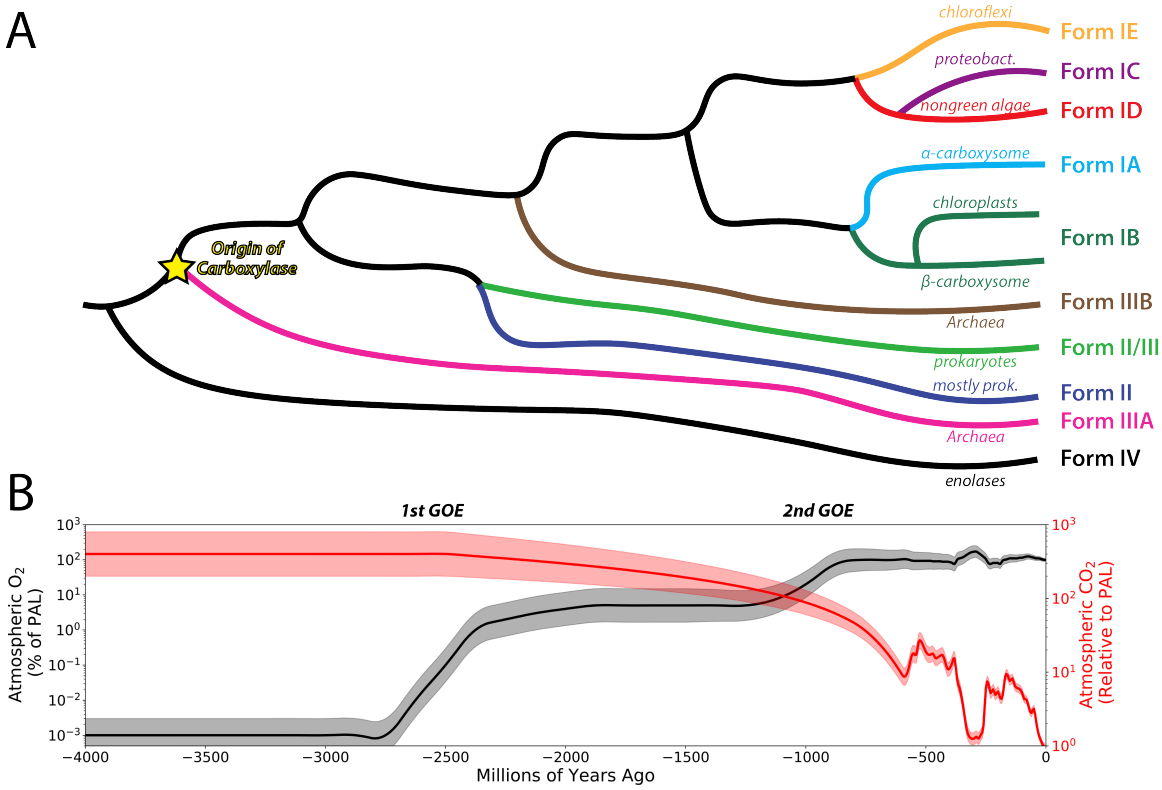


Figure 1.3: Phylogenetic relationship between Rubisco isoforms in the context of historical CO_2 and O_2 concentrations. As shown in the bottom panel (B), reproduced from [277], present day CO_2 and O_2 concentrations are very different than historical values. Geochemical proxies imply that historical O_2 concentrations were very low. Oxygenic photosynthesis arose only once, in a clade of cyanobacteria called "oxyphotobacteria" [101]. The emergence of oxygenic photosynthesis coupled to CO_2-fixation in the CBB cycle led to a sequence of so-called "Great Oxidation Events" (GOEs) where atmospheric O_2 concentrations increased greatly. Increasing O_2 was coupled to an attendant and substantial decrease atmospheric CO_2 due, in part, to fixation and burial of organic carbon. As shown in the schematic phylogenetic tree of Rubisco evolution in the top panel (A), Rubisco has undergone substantial diversification over this timescale. This schematic is derived from an unpublished collaboration with Patrick Shih. Dating of nodes in this tree is highly uncertain. Although it is unclear exactly when the carboxylase activity arose, it is clear that Rubisco is an ancient enzyme ($>$ 2.5 billion years old). Today, four broad classes of Rubisco carboxylase are known - Form I, II, III and II/III. A fifth class, Form IV, is phylogenetically related to other Rubiscos but is an enolase rather than a carboxylase [298]. The carboxylase activity likely arose when an ancestral Form IV-like enolase was selected for carboxylation activity in an anoxic environment [88]. Given their placement on the tree, Form IIIAs, which are found predominantly in Archaea with no CBB cycle, likely evolved from the earliest Rubisco carboxylases. Therefore, carboxylase activity probably arose in a salvage pathway in a heterotrophic context. Form I Rubiscos (FI for short) are found in plants, algae and cyanobacteria and are the dominant contributors to net primary productivity of the Earth biome. It is widely thought that changes in the environmental CO_2 and O_2 concentrations promoted the diversification of Rubisco and the evolution of CCMs [277]. Putative timing of important evolutionary events (e.g. emergence of the carboxylase activity) are marked on the tree.

Rubisco is an ancient enzyme, at least 2.5 billion years old [298]. Since photosynthetic organisms bearing the CBB cycle (plants and cyanobacteria) are responsible for the abundance of O_2 in our present-day atmosphere, Rubisco certainly arose when Earth's atmosphere was rich in CO_2 and contained nearly no O_2 (Figure 1.3B). In low O_2 environments, Rubisco's oxygenase activity poses no problems. Anthropogenic CO_2 emissions notwithstanding, present-day atmosphere is CO_2-poor (0.04%) and O_2-rich (21%). Although O_2 is somewhat less soluble than CO_2, water in Henry's law equilibrium with present day atmosphere (at 25 °C) nonetheless contains nearly 20-fold more O_2 ($\approx 270\mu M$) than CO_2 ($\approx 15\mu M$). Most Rubiscos are beneath CO_2-saturation in these conditions and will also oxygenate RuBP appreciably due to relatively high O_2 concentrations (grey dashed line in Figure 1.2).

As a concrete example, the fastest-carboxylating Rubisco ever observed (at 25 °C) is from the freshwater cyanobacterium *S. elongatus* PCC 7942. The PCC 7942 Rubisco has a maximum per-active site carboxylation rate ($k_{cat,C}$) reported as 14.4 s^{-1} [212]. However, because present-day atmosphere abundant O_2-rich and relatively CO_2-poor, this Rubisco carboxylates at a rate 20-fold below maximum in ambient conditions (per-active site carboxylation rate $R_C \approx 0.7$ s^{-1} per active site, Figure 1.2). Due to its relatively poor CO_2-specificity, the PCC 7942 Rubisco will also oxygenate RuBP in ambient conditions at a rate that is about half the carboxylation rate (per-active site oxygenation rate $R_O \approx 0.3$ s^{-1}), which would necessitate substantial photorespiratory flux to recycle 2PG. Since downstream processing of 2PG by the C_2 pathway leads to the loss of one carbon for every two 2PG, every two oxygenations "undoes" a carboxylation and the net rate of carboxylation by PCC 7942 Rubisco in ambient conditions is roughly $R_C - R_O/2 \approx 0.6$ carboxylations per second, roughly 4% of $k_{cat,C}$. A carboxylation rate this low would be disastrous for cyanobacteria - drastically slowing their growth. All cyanobacteria use a CO_2-concentrating mechanism (CCM) to ensure that Rubisco functions in a CO_2-rich environment. High CO_2 ensures that oxygenation is inhibited and that carboxylation proceeds at near-maximum rate. Just tenfold enrichment of CO_2 above ambient increases the carboxylation rate of PCC 7942 Rubisco to ≈ 5 s^{-1} and suppresses oxygenation to ≈ 0.2 s^{-1}, giving a net carboxylation rate of ≈ 4.6 s^{-1} per active site (Figure 1.2).

There are at least four catalytically-active Rubisco isoforms in nature (three of which are pictured in Figure 1.4) as well as several subfamilies within the Form I, II, II/III and Form III families (Figure 1.3). Form I Rubiscos (abbreviated FI) are readily distinguished from other isoforms as they assemble in to heterohexadecameric complex containing eight active sites (at the four large subunit dimer interfaces) and also eight small subunits (L_8S_8, Figure 1.4). This 500 kDa complex often relies on multiple chaperones for assembly and activity [5]. Form I Rubiscos are by far the most CO_2-specific [301, 262] and, as such, are expected to outperform other isoforms in present-day atmosphere (Figure 1.2). It is therefore not surprising that contemporary plants, algae and cyanobacteria - organisms collectively responsible for $> 99\%$ of biological carbon fixation [94] - rely exclusively on Form I Rubiscos.

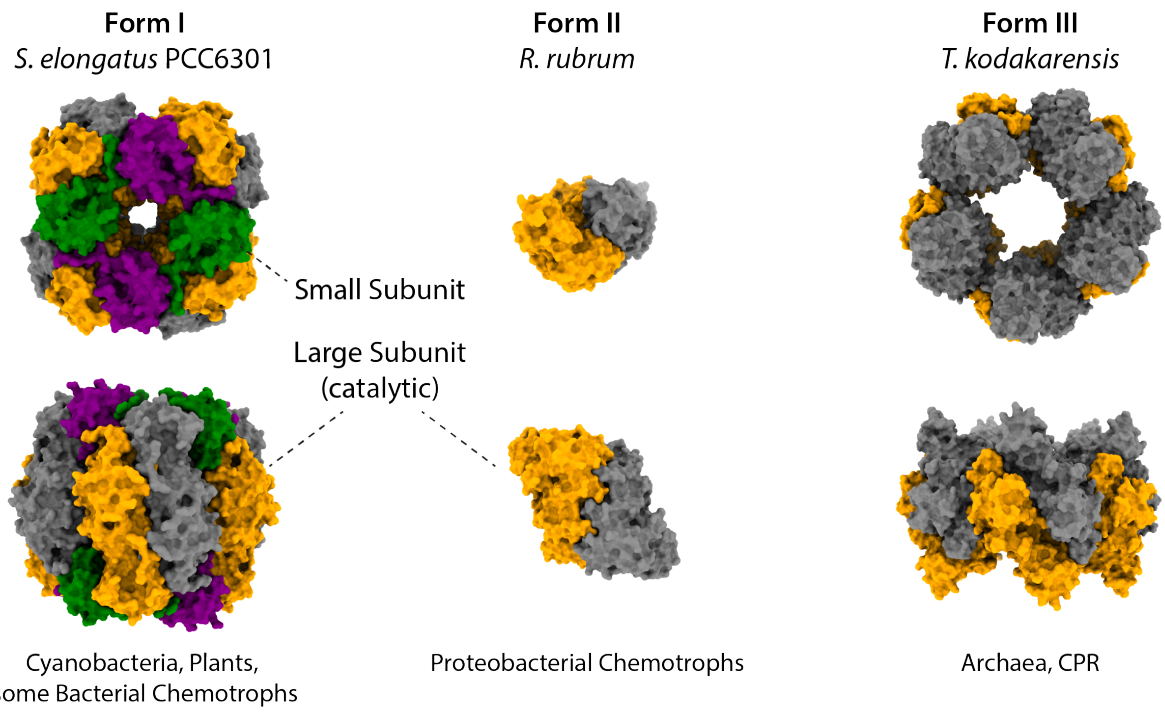


Figure 1.4: Structures of diverse Rubisco isoforms. At least four classes of catalytically active Rubisco isoforms exist in nature (Figure 1.3), three of which are depicted here. Oxygenic photosynthesis, which is found only in cyanobacteria, algae and plants, depends exclusively on Form I Rubisco, which is catalytically superior to other classes in certain respects (Figure 1.2). Some proteobacterial chemotrophs also express Form I. Form I Rubiscos assemble into a quaternary structure formed of 8 large (LSU, grey and orange) and 8 small subunits (SSU, green and purple) called L_8S_8. The LSU is the the catalytic subunit and two active sites are found at each LSU dimer interface. Form II and Form III Rubiscos lack the SSU, form different quaternary assemblies than Form I, and, so far, appear to be inferior carboxylases in terms of CO_2 specificity, maximum carboxylation rate or both. From left to right: Form IB Rubisco from the freshwater cyanobacteriuam *S. elongatus* PCC 6301, PDB 1RBL; Form II Rubisco from the facultatively chemotrophic α-proteobacterium *R. rubrum*, PDB 5RUB; Form III Rubisco from the thermophillic anaerobic Archaeon *T. kodakarensis*, PDB 5MAC.

1.2 Discovery and Characterization of Rubisco

The protein that would later be named "Rubisco" was initially identified in 1947 by Sam Wildman and colleagues at Caltech as a single species making up nearly 100% of ammonium sulfate precipitate of spinach leaf extracts - the so-called "fraction 1 protein" [323, 324]. Five years later, James Bassham, Andrew Benson and Melvin Calvin's pioneering work elucidating the Reductive Pentose Phosphate Pathway (rPP or CBB cycle) in *Chlorella pyrenoidosa* here at Berkeley demonstrated that an enzyme catalyzing the carboxylation and hydrolysis of ribulose 1,5-bisphophate (RuBP) must exist [55, 37, 30]. Arthur Weissbach, Bernard Horecker and Jerard Hurwitz of the NIH later showed that this activity could be enriched from spinach leaves and had a sedimentation coefficient matching fraction 1 protein [315]. Two years later, in 1957, Robert Dorner, Albert Kahn and Sam Wildman seized on this result and published the suggestion that fraction 1 protein was in fact the elusive RuBP

carboxylase [80].

The implication that fraction 1 protein and the CBB cycle carboxylase are one and the same was only confirmed 14 years later in 1971. Nobumaru Kawashima, P.H. Chan, Sam Wildman and colleagues showed, in a series of papers, that crystalline fraction 1 protein retains RuBP carboxylase activity [146, 145, 144, 59, 323]. In the intervening years, several labs observed a curious phenomenon called "photorespiration." Isotopic labeling experiments showed that plants produce CO_2 simultaneous with photosynthetic CO_2 fixation [205, 288]. This phenomenon was called "photorespiration" because it involves the light-dependent oxidation of organic carbon to produce CO_2. Over the course of the 1960s and early 1970s, work from many researchers, including Melvin Calvin, Ed Tolbert and Israel Zelitch, showed that glycolate was an early product of photosynthesis and that that photorespiratory CO_2 derives from metabolism of glycolate [64, 288, 336]. The source and purpose of glycolate production remained mysterious, however.

In 1971, George Bowes, Bill Ogren and Martin Hageman showed by coupled enzymatic assay that glycolate production is due to Rubisco itself [45]. Two years later, John Andrews, George Lorimer and Ed Tolbert confirmed this result by directly measuring O_2 uptake by Rubisco using a mannometer-based assay [14]. As diagrammed in Figure 1.1, oxygenase activity is apparently inherent to the Rubisco active site [300]. In a retrospective on his pioneering work studying the genetics of photorespiration [288], Chris Somerville reports that Bill Ogren and George Bowes deduced this result some five years in advance of direct evidence based on the oxygen-dependence of photosynthetic physiology. Plants grown in reduced O_2 environments (2% instead of 21%) displayed much higher rates of CO_2 fixation. Higher CO_2 levels (e.g. 4% instead of 0.04%) suppressed the effect of O_2, indicating that CO_2 and O_2 compete for the same active site (i.e. for the Rubisco). Oxygenation of RuBP by Rubisco was shown to produce one 3-phosphoglycerate and one 2-phosphoglycolate [14], the latter of which is is the source of photorespiratory glycolate. Photorespiration of some form is required to avoid 2PG accumulation and also the total loss of two carbon atoms (Figure 1.1, [84]).

The name of the enzyme was under constant flux in the 1960s and 1970s. Sam Wildman writes "it was variously called carboxydismutase, ribulose diphosphate or RuDP carboxylase, ribulose bisphosphate, or RuBP carboxylase until the final absurdity 3-phospho-D-glycerate carboxylase" [323]. By 1979 it became clear that the name would have to grow to accommodate the now firmly-established oxygenase activity. To quote Sam Wildman again:

> Lord knows where it might have ended if David Eisenberg had not called it Rubisco as a joke while delivering a talk at my retirement symposium in July 1979. He explained the acronym as Ru standing for ribulose, and the following five letters serving for *bis-c*arboxylase-*o*xygenase. If one were to substitute Na for Ru, he pointed out that it would call to mind a retiree now devoted to promoting the public acceptance of colorless, odorless, tasteless, and highly nutritious crystalline tobacco Fraction 1 protein as a food. [323]

A phenomenal example of "backronym" in action, "RuBisCO" is now taken to stand for **R**ib**u**lose-1,5-**Bis**phosphate **C**arboxylase/**O**xygenase, which I simplify here to "Rubisco."

Structural studies of Rubisco were pursued vigorously from the late 1960s through the early 1990s. By the late 1960s, it had become clear that plant Rubiscos were formed of two subunits, a large protein of mass 50-60 kDa and a smaller protein of mass 12-20 kDa [22]. These are termed the large (LSU) and small (SSU) subunits, respectively (Figure 1.4). By separating the LSU from the SSU of spinach Rubisco, Mikio Nishimura and Takashi Akazawa Nishimura showed that the large subunit carries the catalytic activity, but the presence of the SSU affects reaction kinetics [208]. Building on work in Sam Wildman's lab generating robust crystallization conditions for plant Rubiscos [59], Timothy Baker, David Eisenberg, Fredrick Eiserling and Larry Weissman reported that plant Rubisco has three orthogonal symmetry axes [22], compatible with the L_8S_8 stoichiometry now associated with all Form I Rubiscos (Figure 1.4).

The first high resolution crystal structure of a Rubisco was from a bacteria and not a plant, however. Gunter Schneider, Ylva Lindqvist, Carl-Ivar Brändén and George Lorimer reported a 2.9 Å structure of the dimeric Form II Rubisco from the α-proteobacterium *R. rubrum*, a facultative non-oxygenic phototroph [264]. This structure highlighted the position of active site lysines, demonstrated that the active site is found at the LSU dimer interface, and established that the LSU adopts an α/β barrel fold now called a TIM-barrel after triose-phosphate isomerase. Proteins with this fold are often obligate clients of folding chaperones like groEL/ES [151], as is Rubisco [86]. A later structure of the *R. rubrum* Rubisco is pictured in Figure 1.4.

The *R. rubrum* Form II structure reported by Schneider et al. [264] was not activated by carbamylation, nor was is bound to an RuBP analog. For these reasons, and because Form I Rubiscos are the dominant contributors to net primary photosynthetic productivity (NPP) [27, 94, 99], several labs sought structures of FI Rubiscos from plants and bacteria [204, 130, 60, 61, 12]. Landmark structural analysis of plant Rubisco is due to Inger Andersson, Stefan Knight, Gunter Schneider, Ylva Lindqvist, Tomas Lundquist, Carl-Ivar Brändén and George Lorimer [12]. Andersson and colleagues produced an atomic resolution structure of activated spinach Rubisco in complex with the synthetic intermediate analog 2-carboxy-D-arabinitol-1,5-bisphosphate (CABP), resolving the catalytic conformation of the active site. As expected based on earlier work, the spinach Rubisco assembles into an L_8S_8 complex, with four distinct pairs of LSU dimers (L_2, Figure 1.4).

The Rubisco active site is comprised of several residues that are not contiguous in primary sequence. Residues on both sides of the L_2 dimer interface together aid in the coordination of Mg^{2+}, explaining why a dimer of LSUs appears to be the minimal catalytic unit (as in FII enzymes). Andersson et al. were able to resolve the position of Mg^{2+} and lysine carbamate relative to substrate, suggesting that the metal plays a crucial role in coordinating both CO_2 and RuBP during catalysis. In particular, a lysine residue on loop 6 of the TIM barrel (K334 in spinach) was found to interact with the CABP carboxylate, which was taken simulate substrate CO_2. In the unactivated *R. rubrum* structure, however, loop 6 is found in an "open" conformation that permits solvent access to the active site and does not allow

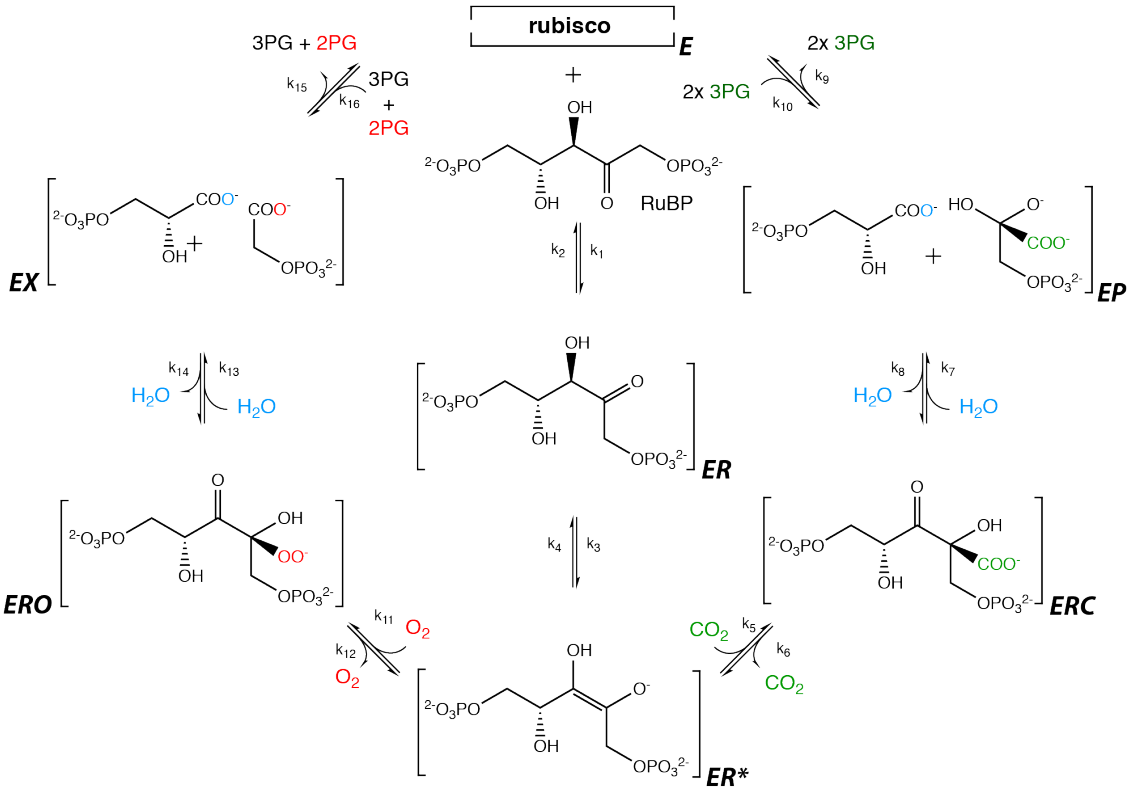


Figure 1.5: Commonly accepted mechanism of Rubisco carboxylation and oxygenation. Various nomenclature has been used to describe the microscopic kinetics of Rubisco. This description follows the nomenclature of Cummins et al. 2018 [71]. Rubisco must be carbamylated and bind Mg^{2+} before it becomes catalytically active, after which it processes three substrates - RuBP, CO_2 and O_2. RuBP binding before activation strongly inhibits activity and RuBP is removed by a catalytic chaperone called Rubisco activase in many organisms [173, 198]. The complete reactions of carboxylation and oxygenation take place through a stepwise mechanism [10, 66, 301]. RuBP binds first forming a complex (ER) with the activated holoenzyme (E), followed by enolization of RuBP (ER*) which allows binding and further processing of CO_2 or O_2. When CO_2 binds the ER* complex the ERC complex is formed while O_2 binding leads to formation of the ERO enzyme-substrate complex. Hydration and cleavage of the ERC complex leads to the formation of two enzyme-bound 3-phosphoglycerate molecules (3PG) in the EP state, each of which have 3 carbon atoms. Oxygenation proceeds through analogous steps except that the products contain 5 carbon atoms in total instead of 6 because no carbon was added. Hydration and cleavage of the ERO complex produces one enzyme-bound 3PG and one 2PG in the EX state. 2PG has two carbon atoms and is not part of the CBB cycle. As such it must be recycled through a photorespiratory pathway to avoid the accumulation of 2PG and also the loss of two carbons [304, 33]. Atoms originating from free CO_2 and O_2 are shown in green and red respectively. The oxygen atom originating from water is shown in blue.

for this lysine-carboxylate contact.

Loop 6 closure was later found to play a crucial role in coordinating carboxylation activity, likely due to stabilization of the developing carboxylate and closure of the active site to solvent [10]. K334 is absolutely conserved in all Rubiscos associated with the CBB cycle [11] and K334C mutants produce near-total abrogation of carboxylation activity in the *R. rubrum* FII Rubisco [174]. A clever chemical rescue experiment by George Lorimer and colleagues showed that K334C mutants become functional carboxylases once the cysteine is modified to resemble a lysine carbamate by aminoethylation or aminopropylation [174]. These and

subsequent structural data greatly aid in understanding the mechanisms of carboxylation and oxygenation by Rubisco [66, 11, 10], as summarized in Figure 1.5.

1.3 Discovery of CO_2 Concentrating Mechanisms

As early as 1815, Benjamin Heyne described the "daily acid taste cycle" of *Bryophyllum calycinum* [42]. Heyne found that leaves tasted acidic at night, but lost all taste during the day. He attributed this to a diurnal cycle of photosynthetic O_2 production during the day and CO_2 production by respiration at night - these cycles were already understood in the late 1700s and early 1800s based on measurements by Joseph Priestly (and others) of plant gas production over the course of a day [8, 234]. I quote from Priestly here to emphasize how much was deduced about living matter from early physiological experiments:

> ... I flatter myself that I have accidentally hit upon a method of restoring air which has been injured by the burning of candles, and that I have discovered at least one of the restoratives which nature employs for this purpose. It is vegetation. In what manner this process in nature operates, to produce so remarkable an effect, I do not pretend to have discovered; but a number of facts declare in favour of this hypothesis ...
>
> One might have imagined that, since common air is necessary to vegetable, as well as to animal life, both plants and animal had affected it in the same manner, and I own that I had that expectation, when I first put a sprig of mint into a glass-jar, standing inverted in a vessel of water; but when it had continued growing there for some months, I found that the air would neither extinguish a candle, nor was it at all inconvenient to a mouse, which I put into it.
>
> ... Accordingly, on the 17th of August 1771, I put a sprig of mint into a quantity of air, in which a wax candle had burned out, and found that, on the 27th of the same month, another candle burned perfectly well in it. This experiment I repeated, without least variation in the event, not less than eight or ten times in the remainder of the summer. [234]

Based on Priestly's work, Heyne thought that nighttime acid production was due to O_2 uptake. He was wrong.

Even before Heyne, in 1804, Theodore de Saussure measured the production and consumption of gasses by plants. de Saussure found that most plants produced O_2 exclusively during the day and CO_2 exclusively at night. The single exception was a cactus, *Opuntia*, which took up both CO_2 and O_2 in the dark [42]. In the late 1800s, Meyer and Krauss showed that CAM plants accumulate sugars during the day and acids during the night. Heyne's "daily acid taste cycle" is due to nighttime CO_2 fixation in plants possessing a metabolic adaptation called Crassulacean acid metabolism (CAM). Though we now know

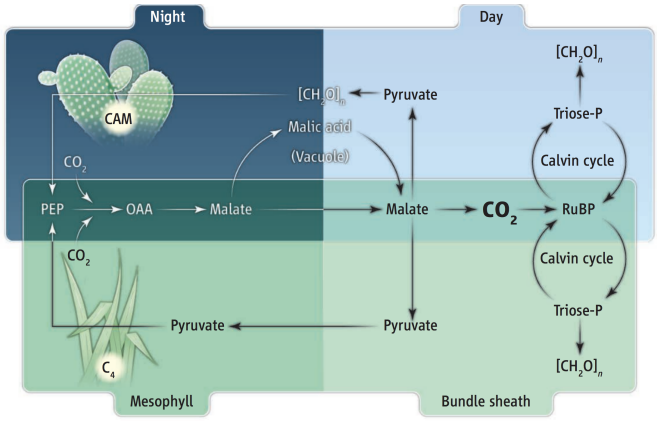


Figure 1.6: CAM and C_4 photosynthesis are analogous to each other. Both CAM and C_4 use PEP carboxylase to temporarily fix CO_2 onto PEP, forming oxaloacetate (OAA). After carboxylation, OAA is reduced to malate. CAM photosynthesis consists of a diurnal cycle of malate production at night (top left) and CBB cycle operation during the day (top right) when light is available. Malic acid is stored overnight in the vacuole. It is released and oxidatively decarboxylated during the day to produce CO_2 and reducing power for the CBB cycle. C_4 operates a similar pathway with spatial, rather than time, separation. PEP carboxylase is active in mesophyll chloroplasts where malate is produced (bottom left). Malate is transported bundle sheath chloroplasts (bottom right) where Rubisco and the CBB cycle are active. Reproduced from [317]. Reprinted with permission from AAAS.

that CAM improves carboxylation and water use efficiency, the purpose and mechanism of acid accumulation remained unclear for more than 100 years.

The details of CAM were elucidated only after another CCM, C_4 photosynthesis, was studied in detail. After Melvin Calvin and colleagues pioneered the use of isotopes to trace the path carbon through photosynthetic metabolism [30], several labs noticed that molecules outside the canonical CBB cycle were labeled at early timepoints [118]. Indeed, this is how the photorespiratory pathway was discovered [304].

Since 3PG is the sole product of Rubisco carboxylation (Figure 1.1), one would expect 3PG to be labeled first when isotopically-labeled CO_2 is provided to plants. However, in the late 1950s and early 1960s, researchers in Hawaii and Russia noticed that dicarboxylic acids like malate, aspartate and oxaloacetate were labeled even before 3PG in sugarcane and maize [118]. These plants are of great economic interest, of course. They are also of great scientific interest because they harbor a distinct leaf anatomy called the "Kranz anatomy" [256], have very high water-use efficiency and can make use of higher light levels than many other crops [42].

Early labeling of oxaloacetate and malate was explained by the discovery, by Marshall Hatch and others, that C_4 plants use an alternative primary carboxylase (not Rubisco) to fix CO_2 into oxaloacetate. Oxaloacetate is then reduced to malate and transported to a specific cell type, the bundle sheath cell, where Rubisco is localized (Figure 1.6). There, malate is oxidatively decarboxylated to produce a high CO_2 concentration that promotes fast carboxylation and suppresses oxygenation [118]. This mode of photosynthetic CO_2 fixation is called C_4 because CO_2 is temporarily fixed onto a four-carbon molecule. Since Rubisco produces a three-carbon molecule in 3PG, "regular operation" of the CBB cycle without a CCM is called C_3 by analogy. As discussed below, C_4 photosynthesis appears to be advantageous in warm climates.

The above model of C_4 photosynthesis was based on many experiments performed in the late 1960s and early 1970s [118]. These results motivated studies clarifying the mechanisms of

CAM photosynthesis. Work by numerous researchers led to the discovery that CAM is quite similar to C_4: where C_4 uses four-carbon dicarboxylic acids to transport units of inorganic carbon (C_i) in space, CAM photosynthesis uses them to store C_i units for an extended time (Figure 1.6). CAM plants temporarily fix CO_2 onto PEP to form oxaloacetate and then malate, as in C_4 photosynthesis. Malic acid is stored overnight, only to be decarboxylated during the day. Oxidative decarboxylation of malate provides CO_2 and reducing power for the CBB cycle [42]. As a result, CO_2 partial pressures inside CAM plants can exceed 2%, or 50 times ambient [67]. CAM is advantageous for plants living in hot, arid climates since primary CO_2 fixation occurs at night, when temperatures are relatively low. As such, CAM plants need only open their stomata at night, when there is less risk of water loss.

Photosynthetic bacteria, including all known cyanobacteria, also harbor CCMs. The central organelle of the bacterial CCM - the carboxysome - was first observed in the late 1960s as a "polyhedral body" on electron tomograms of cyanobacterial and proteobacterial cells [237, 201, 111]. Jessup Shively and colleagues showed in 1973 that fractions enriched in these inclusions also contained Rubisco carboxylase activity [280], which motivated the name "carboxysome." Over the subsequent decades, several groups showed that perturbing the carboxysome structure also have large effects on growth [237]. Such mutants are typically high-CO_2 requiring (HCR), implying that wild-type cells concentrate CO_2 in the carboxysome. A more detailed description of bacterial CCM is given below.

1.4 Lasting Impacts of Rubisco Studies

Studies of Rubisco have borne surprising fruits for plant, molecular and evolutionary biologists as well as geochemists. Chris and Shauna Somerville developed the now ubiquitous *Arabidopsis thaliana* as a model plant in order to investigate the genetic basis of photorespiration. Photorespiratory mutants display a high-CO_2 requiring (HCR) phenotype because they cannot recycle 2PG quickly enough to grow in ambient CO_2 and O_2 concentrations. The Somervilles ultimately mapped the entire C_2 photorespiratory pathway (diagrammed in Figure 1.1B) by screening for HCR mutants in *A. thaliana*, confirming that the mitochondrion, chloroplast and peroxisome coordinate to recycle 2PG [288].

The first-identified protein-folding chaperone, cpn60, was discovered as a single band that co-purifies with Rubisco from pea leaves [86]. cpn60 is the chloroplast homolog of the now-famous bacterial groEL/ES chaperone system, and all Rubisco large subunits are obligate clients of groEL/ES type chaperones [119]. Many Form I Rubiscos are also clients of other, Rubisco-specific folding and assembly chaperones that are required for the assembly of a functional L_8S_8 complex (Figure 1.4, [5, 119]). The discovery that proteins like Rubisco are "addicted" to folding chaperones overturned the broadest statements of Anfinsen's dogma by showing that not all proteins fold spontaneously to their most stable state [86].

Studies of Rubisco also introduced biologists to "catalytic chaperones" - proteins that maintain activity of a client enzyme. In 1976, George Lorimer, Murray Badger and John Andrews showed that Rubisco must be activated by exposure to CO_2 and Mg^{2+} [173]. They

correctly proposed that activation is due to carbamylation of an active site lysine residue. We now know that RuBP binding the pre-activated enzyme inhibits carbamylation, reducing total Rubisco activity [10]. Many plants also synthesize RuBP analogs as a means of inhibiting Rubisco in response to stress [220]. It was later discovered that dedicated chaperones - Rubisco activases - transduce the energy of ATP hydrolysis to forcibly dislodge RuBP and other inhibitors from the Rubisco active site, thereby permitting activation by spontaneous carbamylation of a crucial lysine residue [40, 198]. We now know that different lineages of Form I Rubisco are associated with specific Rubisco activases, all of which couple the mechanical work of activation to ATP hydrolysis via AAA+ ATPases [198].

Rubisco studies have had lasting impact outside of plant and molecular biology as well. Detailed understanding of the chemical mechanisms of Rubisco carboxylation and oxygenation has had a particularly profound impact on geochemistry and environmental sciences. Early kinetic models of Rubisco were directly integrated into the first models of whole plant photosynthesis. These models encode the idea of a "see-saw" between different flux limitations [98]. That is, photosynthesis can't proceed faster than Rubisco carboxylation and also can't exceed the rate of photosynthetic electron transport. The first such model, due to Graham Farquhar, Susanne von Caemmerer, and Joe Berry [95], predicts the response of plant physiology (e.g. photosynthetic rates, water uptake, and growth) to changes in important environmental variables (e.g. temperature, humidity, and CO_2 conentrations) and physiological parameters (e.g. stomatal conductance, Rubisco activation). These models have been improved and extended over time to model, for example, C_4 photosynthesis [68], but the basic form remains the same. Models of this form are now widely-used to predict the response of plant growth physiology to environmental changes [100] and engineered improvements to their photosynthetic apparatus [187, 328].

Another lasting impact of studying Rubisco has been the realization that Rubisco's isotopic fractionation —i.e. preference for $^{12}CO_2$ over $^{13}CO_2$ —can be leveraged to interrogate plant physiology as well as the biological origins of geological samples [38, 96]. Measurements of Rubisco isotope discrimination were initially used to investigate the reversibility of CO_2 addition and test which reaction steps limit the carboxylation rate [252]. C_3 plant Rubiscos have isotope effects of $\Delta \approx$ 25-30 ‰ in favor of the lighter isotope, while primary carboxylases of CAM and C_4 plants discriminate much less, with $\Delta \approx 2-3$ ‰ for PEP carboxylase [189]. As such, the carbon isotope content of plant biomass can be used to test whether the plant employs a CCM [38, 96]. C_3 plant biomass has relatively low ^{13}C content - $\approx$ 29 ‰less ^{13}C than ambient CO_2- reflecting the fractionation pattern of Rubisco. In contrast, C_4 plant biomass is more similar to atmospheric CO_2 in isotopic composition, with $\approx$ 14 ‰ less ^{13}C than ambient CO_2 [272]. These various results related to the isotopic fractionation of carbon fixation have had profound impact on the fields of plant physiolog, geochemistry and evolutionary biology. Plant biologists now routinely study plant photosynthetic physiology non-invasively by measuring isotope discrimination [272] and geochemists use similar methods to study the biotic origins of organic matter in fossils [121].

Enzyme	C_i	O_2?	k_{cat}/K_M	V_{max}	EC Class
Rubisco	CO_2	oxygenation	3×10^5	4	4.1.1.39
PEP Carboxylase	HCO_3^-	No	–	–	4.1.1.31
PEP Carboxykinase	CO_2	No	1×10^5	39	4.1.1.32
Propionyl-CoA Carboxylase	HCO_3^-	No	2×10^5	30	4.1.1.41
Pyruvate Carboxylase	HCO_3^-	No	4×10^4	30	6.4.1.1
Acetyl-CoA Carboxylase	HCO_3^-	5×10^5	18	No	6.4.1.2
Methylcrotonyl-CoA Carboxylase	HCO_3^-	3×10^4	6	No	6.4.1.4
2-oxoglutarate Carboxylase	HCO_3^-	No	6×10^4	15	6.4.1.7
Enoyl-CoA Reductase/Carboxylase (ECR)	CO_2	No	7×10^5	130	1.3.1.85
2-ketoacid Synthases	CO_2	Sensitive	–	–	1.2.7.1

Table 1.1: Comparison of natural carboxylases. The C_i column gives the inorganic substrate of the carboxylation reaction, either CO_2 or HCO_3^-. $k_{cat}K_M^{-1}$ is reported in M^{-1} s^{-1} units and denotes the slope of Michaelis-Menten relation in the limit of low C_i concentration, which is often termed the "catalytic efficiency." V_{max} is reported in U mg^{-1} and gives the average maximum catalytic rate of enzymes in this class with respect to their C_i substrate. Comparing to other natural carboxylases, Rubisco is an outlier in two respects. Most natural carboxylases have maximum rates that are substantially faster than even the fastest Rubisco (per mg protein). Some other carboxylases are sensitive to O_2 (e.g. 2-ketoacid synthases like pyruvate synthase), but Rubisco alone catalyzes oxygenation of its organic substrate. Table reproduced from Table S1 of [89] and [266]. Abbreviations: phosphoenolpyruvate (PEP), ribulose 1,5-bisphophate (RuBP).

1.5 Overview of CO_2 Concentrating Mechanisms

Typically, 20-30% of soluble protein in plant leaves is Rubisco [110]. It is thought that such high Rubisco levels are required to support C_3 plant growth precisely because Rubisco is not very fast and not very CO_2-specific [301, 262]. No known Rubisco has a maximum carboxylation rate ($k_{\mathrm{cat,C}}$) above 15 s^{-1} and all Rubiscos also catalyze a competing oxygenation RuBP [10, 301, 262]. It is widely hypothesized that the sluggishness and non-specificity of Rubisco promoted the evolution of diverse CO_2 concentrating mechanisms [301, 262, 243]. Though these various systems are functionally distinct and widely distributed among bacteria, eukaryotic algae and plants, they all serve the same basic purpose: elevating the concentration of CO_2 near RuBisCO to increase the carboxylation rate and competitively inhibit oxygenation (Figure 1.2).

Two broad classes of CCM exist [243]. Biophysical CCMs involve pumping of inorganic carbon into dedicated cellular compartments that contain a large number of Rubisco active sites. Biophysical CCMs are typically found in aquatic organisms and rely on C_i transporters and carbonic anhydrase enzymes (CAs) to function. CA catalyzes the reversible dehydration of HCO_3^- to CO_2, the true substrate of Rubisco carboxylation (Table 1.1). Biochemical CCMs like C_4 and CAM, in contrast, use a primary carboxylase other than Rubisco to fix CO_2 onto an organic carbon scaffold. As shown in Table 1.1, most other carboxylases are both faster than Rubisco and less sensitive to O_2. In biochemical CCMs, the organic product is used as a temporary store of C_i —it is ultimately transported to Rubisco, where it

is decarboxylated to produce a high local CO_2 concentration to promote fast carboxylation. Biochemical CCMs are typically found in land plants, especially those endemic to warm environments [243, 257].

All CCMs use some intermediary process to concentrate CO_2 near Rubisco in order to sidestep the well-known deficiencies of Rubisco. Irrespective of how CO_2 concentration is achieved, elevating [CO_2] suppresses oxygenation by competitive inhibition (occupying the active site with CO_2) and promotes carboxylation by mass action (Figure 1.2). A sufficiently high CO_2 concentration could produce a situation in which carboxylation is saturated (i.e. each active site carboxylates at $k_{\mathrm{cat,C}}$) and oxgenation is negligible. Despite some suggestions in the literature that CCMs block O_2 from interacting with Rubisco [150, 237], most models of CCM action requires no modulation of the O_2 concentration at all.

Biophysical CCMs

Biophysical CCMs are found in cyanobacteria, many chemotrophic proteobacteria, and most Eukaryotic algae. Aquatic phototrophs often have biophysical CCMs, but they are only rarely found in land plants [311]. Biophysical CCMs are collectively responsible for a large fraction of aquatic photosynthesis[1], which itself constitutes $\approx 50\%$ of global NPP [99, 217]. Biophysical CCMs typically operate in single cells. The carboxysome-based CCM of bacteria, for example, requires ≈ 20 genes to function, evolved convergently at least twice, and has undergone massive horizontal transfer within and between bacterial phyla, implying that the system is portable and amenable to engineering. The pyrenoid-based biophysical CCM of green algae, diatoms and other Eukaryotic algae also evolved convergently multiple times, although the genetics and molecular biology of the pyrenoid-based CCM is only now coming into view.

Carboxysome CO_2 Concentrating Mechanisms

Knockout of essential CCM components generates a high-CO_2 requiring (HCR) phenotype, meaning that CCMs are amenable to genetic characterization [229]. Work from several groups over decades has elucidated the molecular components of the bacterial CCM in various organisms [238]. These components include C_i transporters in the cell membrane and a 200+ MDa protein organelle called the carboxysome that co-encapsulates hundreds of Rubisco complexes with carbonic anhydrase (Figure 1.7). Energy-coupled C_i uptake produces an above-equilibrium bicarbonate (HCO_3^-) concentration in the cytosol (10-50 mM), which is converted into a high carboxysomal CO_2 concentration by CA activity found exclusively in the carboxysome (Figure 1.7).

This model of CCM function is dependent on several crucial features. First, the C_i uptake must be energetically activated so that intracellular C_i concentration can greatly exceed the extracelluar concentration (by 10-50 fold). C_i uptake is typically assumed to produce HCO_3^-

[1] Certainly more than 50%, likely upwards of 90% [178].

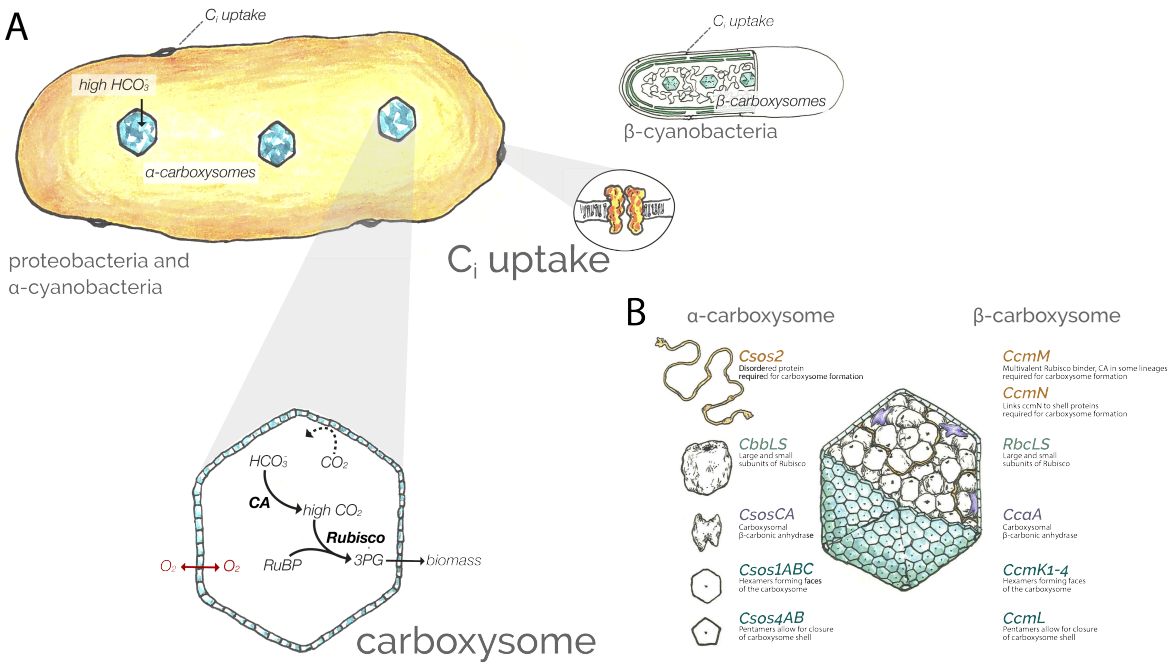


Figure 1.7: Flow of inorganic carbon through the carboxysome-based biophysical CCMs of proteobacteria and cyanobacteria. (A) Chemotrophic proteobacteria (left) and phototrophic cyanobacteria (small, right) both express carboxysome-based CCMs. There are two convergently-evolved forms of the carboxysome-CCM - α and β forms - that are believed to function in roughly the same manner. α-carboxysome CCMs are found in chemotrophic proteobacteria and oceanic cyanobacteria while β-carboxysome CCMs are predominantly found in freshwater cyanobacteria [238]. Energy-coupled C_i uptake at the cell membrane produces an above-equilibrium bicarbonate (HCO_3^-) concentration in the cytosol (10-50 mM). Two distinct classes of C_i uptake systems have been observed: (i) those that pump extracellular HCO_3^- into the cytosol and (ii) those that energize the hydration of cytosolic CO_2 (assumed to be very cell-permeable) into HCO_3^-. High cytosolic HCO_3^- is converted into a high carboxysomal CO_2 concentration by CA activity found exclusively there. Since CA activity is not coupled to any energy source, the cytosolic HCO_3^- concentration must be both high (10-50 mM) and, crucially, out-of-equilibrium with CO_2. In order for the carboxysomal CO_2 concentration to saturate Rubisco as steady-state, the carboxysome shell must slow the diffusive loss of CO_2 [248]. Panel (B) shows the molecular makeup of α- and β-carboxysomes. Despite being highly analogous structures, it is likely that the two carboxysome families evolved independently. In panel (B), structural proteins forming the carboxysome interior are highlighted in orange. Rubisco small and large subunits (light green) are found in the carboxysome interior along with carboxysomal CAs (named in purple). Shell protein names for both lineages are highlighted are in dark green. Further details about α- and β-carboxysome genes are given in Table 1.2. Illustrations by Rachel Shipps, reproduced with permission.

in the cytosol because HCO_3^- carries a net charge near neutral pH and, therefore, is much less cell-permeable than CO_2 [194]. Even though the spontaneous dehydration of CO_2 to HCO_3^- is relatively fast, cytosolic HCO_3^- must be held out-of-equilibrium with CO_2. Otherwise, the carboxysomal CA would have no effect: it can only bring HCO_3^- and CO_2 into equilibrium because it is not coupled to any energy source. Moreover, in order for CO_2 to accumulate inside the carboxysome (i.e. where Rubisco is located) the carboxysome shell must slow the diffusive loss of CO_2. The basic reason for this is that carboxysomes are very small - only 100-200 nm in diameter - and small molecule diffusion is very fast on short length scales [194]. If the carboxysome shell did not slow the diffusive loss of CO_2, CO_2 would tend

Protein Names	Function	α or β	Marker Gene?
RbcLS or CbbLS	Rubisco large and small subunits	both	Conserved in lineages without carboxysomes or CCMs
Csos1A-D	Hexameric shell proteins	α	Small (10 kDa) monomers with multiple paralogs
CcmK1-4, CcmP	Hexameric shell proteins	β	Small (10 kDa) monomers with multiple paralogs
Csos2	Biogenesis of carboxysome lumen	α	Disordered repeat protein, low conservation
CcmM	Biogenesis of carboxysome lumen, CA domain	β	Repeat protein, marker for β lineage
CcmN	Biogenesis of carboxysome lumen, CA domain	β	Marker for β lineage
CsosCA	Carbonic anhydrase	α	Marker for α lineage, unrelated to ccmM or ccaA
CcaA	Carbonic anhydrase	β	Not found in all β-carboxysomes
Csos4A-B	Pentameric shell proteins	α	Small (10 kDa) monomers with multiple paralogs
CcmL	Pentameric shell proteins	β	Small (10 kDa) monomers
DabAB	C_i transport	α	Only in proteobacterial CCM, also found in heterotrophs
SbtAB	C_i transport	both	Homologous to other Na^+ symporters, not universal
BicA	C_i transport	both	Homologous to other SulP transporters, not universal
BctA	C_i transport	β	Endemic to freshwater cyanobacteria

Table 1.2: Genes associated with α and β forms of the bacterial CCM. Bacterial CCM genes are associated either with the carboxysome or with C_i transport, as shown in Figure 1.7. While the Rubisco large subunit is highly conserved protein often found in single copy in genomes, it is also highly conserved across cyanobacteria, proteobacteria and land plants [88, 277]. With the exception of the Rubisco itself, carboxysome genes are poor markers for phylogenetic analyses. Shell proteins are small (monomers of $\approx$ 10 kDa), often have multiple paralogs, and are homologous to shells of other bacterial microcompartments. Structural proteins of the carboxysome lumen (*csos2*, *ccmM*, *ccmN*) are idiosyncratic to the α and β families and contain large unstructured regions. Similarly, lumenal CAs differ greatly between α and β lineages and so cannot serve as marker genes. C_i transporters are also poor markers since the same transporters are found in both α- or β-cyanobacteria. All C_i transporters are within families containing transporters of other substrates. As such, it is possible to annotate genomes as containing α- or β-carboxysomes, but challenging to use marker genes to study the evolutionary relationship between α and β family CCMs.

to equilibrate across the carboxysome shell and the cell membrane, negating the supposed purpose of the CCM (Figure 1.7).

Pyrenoid CO_2 Concentrating Mechanisms

Like carboxysomes, pyrenoid structures are proteinaceous organelles containing a large number of Rubisco active sites (Figure 1.8A). Pyrenoids differ from carboxysomes in three important ways, however. First, pyrenoids are found almost-exclusively in Eukaryotic algae and never in bacteria. Second, pyrenoids are much larger than carboxysomes: about 1 μm in diameter in model algae like *Chlamydomonas reinhardti*. Typical pyrenoids have roughly the same volume as an *E. coli* or cyanobacterial cell [194]. Finally, pyrenoids do not have a protein shell or and any meaningful diffusional boundary so far as we know. The *C. reinhardti* pyrenoid is partially surrounded by a starch sheath, but sheath mutants grow normally [310]. Because the pyrenoid has no apparent boundary, it cannot function in the same manner as

the carboxysome CCM is thought to. Recent work has identified multiple novel proteins involved in forming the pyrenoid [178, 179, 106] and characterized several others involved in C_i pumping [137]. Nonetheless, the genetics and molecular mechanisms of pyrenoid CCMs are less well-understood than for carboxysome CCMs.

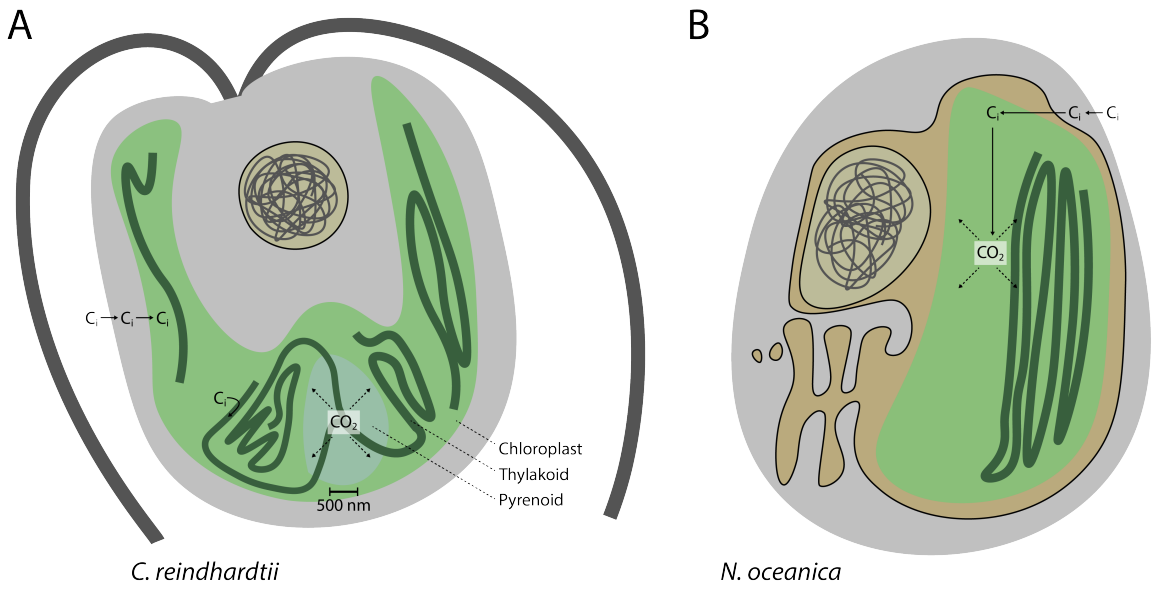


Figure 1.8: Flow of inorganic carbon through the biophysical CCMs of Eukaryotic algae. (A) A diagrammatic model of carbon flow through the pyrenoid-based CCM of the soil-dwelling green alga *Chlamydomonas reinhardti*. (B) A model of a proposed "pump-leak" CCM in the oceanic heterokont, *Nanochlorpsis oceanica*. In both of these models, C_i is pumped into the cell in the form of HCO_3^- or CO_3^{2-} and is ultimately converted to CO_2 at or near Rubisco by carbonic anhydrase activity. In pyrenoid-based CCMs (A), Rubisco is highly concentrated inside the pyrenoid structure by a polyvalent Rubisco binding protein [178]. As such, the localized production of CO_2 near the center of the pyrenoid can plausibly saturate the Rubisco. In *C. reinhardtii*, tubules emanating from the thylakoid membrane penetrate the pyrenoid and are thought to convey C_i to center of the structure [87]. For "pump-leak" type CCMs (B), there is no pyrenoid and so Rubisco is thought to be uniformly distributed throughout the chloroplast [112]. These cells display HCO_3^- uptake and CO_2 leakage. Since the chloroplasts of heterokonts and other phototrophs on the red algal lineage are surrounded by multiple lipid bilayers, it is plausible that CO_2 is retained near Rubisco long enough to improve the whole cell carboxylation rate.

Because pyrenoids are roughly ten times larger in linear dimension than carboxysomes[2], it is possible to imagine a "pump-in, diffuse-out" mechanism of CCM action, as diagrammed in Figure 1.8A. This model relies on the pyrenoid's size and the fact that fenestrations thylakoid membranes penetrate the pyrenoid [87]. Because of the larger size, CO_2 produced near the pyrenoid center is likely to encounter a Rubisco active site before diffusing out. Thylakoid tubules might convey HCO_3^- to the center of the pyrenoid where a centrally-localized CA could dehydrate HCO_3^- to CO_2. Since the thylakoid lumen is very acidic during photosynthesis (pH $\approx$ 4), dehydration of HCO_3^- pumped into the thylakoid lumen would produce a high lumenal CO_2 concentration which could be released near the center of the pyrenoid. Efforts to catalog the genetic and molecular mechanisms of pyrenoid CCMs

[2]Pyrenoid diameters are ≈ 1 μm as compared to ≈ 100 nm for the carboxysome.

are currently underway [106, 162, 153] and should fill out details missing from our current model of this environmentally significant family of CCMs.

'Pump-leak' CO_2 Concentrating Mechanisms

Some Eukaryotic algae have no apparent pyrenoid structures but nonetheless appear to have a CCM [134, 112]. CCM activity can be measured, for example, by measuring whole-cell uptake of isotopically labelled C_i [134]. Since these cells take up C_i as HCO_3^- and release CO_2 into the medium in excess of the chemical equilibrium, they are thought to employ a "pump-leak" CCM where C_i is pumped into the cell and converted into CO_2 near Rubisco. The "pump-leak" CCM is diagrammed loosely in Figure 1.8B [134, 112]. Notably, because chloroplasts from heterokonts (e.g. diatoms) are the result of a secondary endosymbiosis of a red alga (i.e. endosymbiosis of a cell already containing chloroplasts) they are separated from the cytoplasm by multiple membranes that can slow the diffusive loss of CO_2. This arrangement suggests a "pump-leak" mechanism could be used more efficiently by heterokonts than other lienages. Although recent work has shown that carbonic anhydrases are involved in pump-leak mechanisms [112], very little else is known about the molecular mechanisms underpinning this family of CCMs.

Biochemical CO_2 Concentrating Mechanisms

Biochemical CCMs are found almost exclusively in land plants, and are especially common in plants endemic to high lattitudes (i.e. warmer temperatures [256, 317]). Although most plant species are so-called C_3 plants, relying on the CBB and Rubisco cycle directly, upwards of 20% of terrestrial NPP is due to plants bearing the C_4 and CAM biochemical CCMs [243]. One core difference between terrestrial and oceanic photosynthesis is that temperature varies much more on land than it does in Earth's oceans [27]. Biochemical CCMs can broadly be considered as responding to physico-chemical problems associated with elevated temperature.

Oxygenation of RuBP is much more favorable than carboxylation, with an associated Gibbs free energy $\Delta_r G'^m \approx$ -520 kJ/mol as compared to $\approx$ -28 kJ/mol for carboxylation [102]. Despite the fact that the net reaction of RuBP oxygenation is much more favorable than carboxylation, typical maximum oxygenation rates by plant Rubisco ($k_{cat,O}$) are about fivefold below $k_{cat,C}$ [262, 301]. This disparity suggests that oxygenation rates are limited by the rate of O_2 addition to the enzyme-bound enediolate of RuBP (Figure 1.5). It stands to reason that, whatever chemical step limits the rate of O_2 addition [300], this limitation would be relaxed at higher temperature [301]. Whether or not this argument is right, increased temperature decreases CO_2-specificity and, therefore, favors oxygenation over carboxylation for all Rubiscos measured so far [47, 124].

Figure 1.9 shows the calculated temperature dependence of the ratio of oxygenation to carboxylation (R_O/R_C) of the *A. thaliana* Rubisco, which exemplifies this general trend. At low temperature and high CO_2 concentrations, carboxylation dominates. At elevated tem-

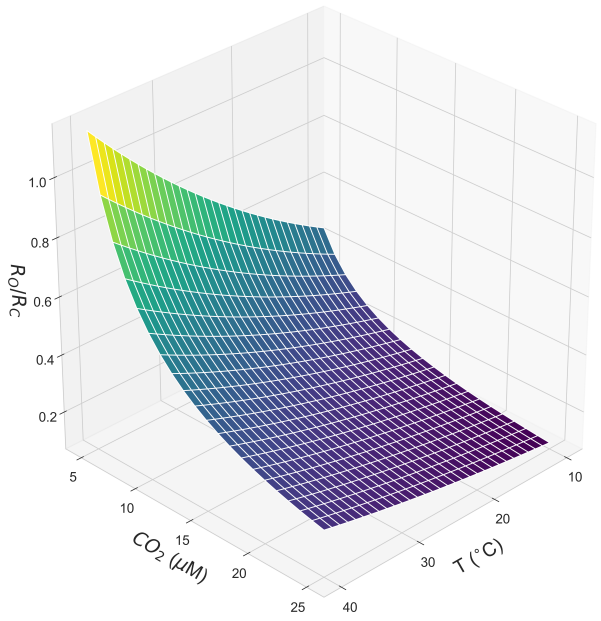


Figure 1.9: Elevated temperatures promote oxygenation by Rubisco. The temperature dependence of the *A. thaliana* Rubisco is used here to exemplify a common trend among all Rubiscos measured - that CO_2-specificity ($S_{C/O}$) decreases as temperature increases [124]. As a result of this general trend, the relative rate of oxygenation increases with temperature. This is clear from the vertical axis of the plot, which shows that the ratio of oxygenation to carboxylation rates (R_O/R_C) approaches 1.0 at elevated temperature and low CO_2 concentration. These conditions could plausibly arise in a tropical or desert plant. Arrhenius temperature dependence of the *A. thaliana* Rubisco was modeled using activation energies and reference values from [47]. Partial pressures were converted into molar concentrations of CO_2 and O_2 by calculating the Henry's law coefficients for these gases at the appropriate temperature [258].

perature and lower CO_2 concentrations, the relative rate of oxygenation grows exponentially due to an Arrhenius-type temperature dependence [47, 124].

In addition to promoting oxygenation over carboxylation, elevated temperature also increases the entropy of water vapor and gaseous CO_2, which creates two problems for plants. First, high temperature reduces the equilibrium concentration of CO_2 in water that is in Henry's law equilibrium with atmosphere [258]. Second, plants take up CO_2 through pores in the leaf surface called stomata through which water can be lost. In warm, dry environments, opening the stomata can lead to substantial water loss. C_4 and C_2 photosynthesis can be seen as responding to the high oxygenation rates associated with elevated temperatures, while CAM photosynthesis can be seen as responding to water loss associated with opening stomata in hot, arid climates.

CAM Photosynthesis

Crassulacean acid metabolism, or "CAM photosynthesis," is predominantly found in succulents and cactuses that inhabit desert environments. Major families include Agavaceae, Bromeliaceae, and Crassulaceae, after which the pathway is named [42]. The principle underlying CAM photosynthesis is time-separation of the light- and dark-reactions of photosynthesis (Figure 1.6). During the night, when temperatures are low and danger of extreme water loss is reduced, stomata are opened to permit CO_2 uptake. C_i is fixed onto phosphoenolpyruvate (PEP) by the primary carboxylase, forming a dicarboxylic acid, oxaloacetate. Oxaloacetate is reduced to malate and stored overnight in vacuoles (Figure 1.6).

Malate provides overnight storage of both C_i and reducing equivalent. During the day, when temperatures are high, stomata can remain mostly closed because the stored malate is decarboxylated to produce CO_2 and reducing equivalent for fixation through the CBB cycle [317]. CAM photosynthesis evolved independently at least 30 times, with multiple origins even within single families like bromelieads [317].

C_4 Photosynthesis

C_4 photosynthesis is very similar to CAM in pathway structure (Figure 1.6), but is based on spatial separation between Rubisco and the primary carboxylase rather than time separation. The primary carboxylase is localized to the cysotol of mesophyll cells, while Rubisco and the CBB cycle are localized to the chloroplast of bundle sheath cells [317]. In C_4 photosynthesis, C_i is fixed onto PEP by PEP carboxylase, forming a four-carbon dicarboxylic acid, oxaloacetate, from which the name derives [118]. Oxaloacetate is reduced to malate, which acts as a long-range carrier of CO_2 and reducing equivalent. Malate is transported across multiple membranes and cell walls, over a distance of $\approx 10\mu$m to the bundle sheath cell where Rubisco is localized. There, malate is decarboxylated to produce CO_2 and reducing equivalent for fixation through the CBB cycle [118].

C_4 photosynthesis it typically associated with an anatomical adaptation called the "Kranz anatonmy" [175], which has several important features:

1. Chloroplasts are dimorphic - mesophyll and bundle sheath chloroplasts differ greatly.
2. Rubisco is predominantly localized to the chloroplasts of bundle sheath cells.
3. Bundle sheath chloroplasts lack thylakoid grana and do not perform photochemistry.
4. Mesophyll cell chloroplasts do contain grana and do perform photochemistry.
5. Mesophyll cells express a primary carboxylase (e.g. PEP carboxylase).

C_4 is considered a "syndrome" rather than a pathway because there are several distinct pathways that share the architecture described above [257]. The subtypes vary in the localization of enzymes, the primary carboxylase used (PEP carboxylase, EC 4.1.1.31 or PEP carboxykinase, EC 4.1.1.32, Table 1.1) and in the mode of oxidative decarboxylation of malate (e.g. producing NADH or NADPH). Each of these variants appears to have arisen independently, and the C_4 syndrome is thought to have evolved independently > 50 times [257].

C_2 Photosynthesis

C_2 photosynthesis is an adaptation that enables re-fixation of CO_2 that would otherwise be lost to photorespiratory decarboxylation. C_2 is named for the C_2 photorespiratory pathway because the plant anatomy is structured so that photorespiratory CO_2 production occurs

near Rubisco. In a sense, C_2 is a biochemical CCM in which Rubisco is both the primary and secondary carboxylase.

C_2 is sometimes called C_3-C_4 intermediacy. C_2 it is thought to be an evolutionary intermediate between C_3 and C_4 metabolisms [257] because C_2 plants are often observed to have a proto-Kranz anatomy. In this Kranz-like anatomy, an enlarged bundle sheath has increased chloroplast content and bundle sheath chloroplasts surround bundle sheath mitochondria. Bundle sheath mitochondria produce CO_2 both by respiration and by expressing the photorespiratory glycine decarboxylase (GDC). This feature - localization of GDC to bundle sheath chloroplasts - is shared by many C_4 plants. The centripital arrangement of bundle sheath chloroplasts around mitochondria is thought to promote re-fixation of CO_2 that would otherwise be lost to oxidative decarboxylation in the mitochondria [257].

1.6 Objectives

My overarching goal has been to understand deeply how CCMs function so that we might engineer them rationally and trace the selective pressures that led to their striking multiple emergence. I am particularly motivated by the question of why certain organismal physiologies are so closely-associated with particular CCMs - i.e. why biophysical CCMs are associated with aquatic photosynthesis and, conversely, why terrestrial plants use biochemical CCMs almost exclusively. Several labs have begun efforts towards reconstituting the bacterial CCM in crop plants [165, 171, 212], so I am especially curious to understand why it is that no land plants have a CCM of this type. Indeed, the only reported land plants with any biophysical CCMs are hornworts [311], which are very small and only distantly related to useful crops. Perhaps there are physiological reasons why large terrestrial plants evolved biochemical CCMs so many times?

My interest in CCM evolution dovetails tightly with an interest in engineering photosynthesis. Given that photosynthesis is ultimately responsible for nearly all of the food calories on the planet and biological CO_2 fixation is the largest single flux in the global carbon cycle [99], it is fitting that the field of photosynthesis engineering is undergoing a renaissance. Recent work has demonstrated multiple successful rational approaches to the improvement of photosynthetic yields. These works take various approaches to optimizing the CBB cycle [195, 284], photorespiration [147, 290], and the photosynthetic light reactions [158], all of which aim to increase crop yields by improving the light interception efficiency ϵ_i or photosynthetic energy conversion ϵ_c:

$$\underbrace{w_h}_{\text{total harvested yield}} = \overbrace{S}^{\text{solar energy}} \times \underbrace{\epsilon_i}_{\text{interception eff.}} \times \overbrace{\epsilon_c}^{\text{conversion eff.}} \times \underbrace{\eta}_{\text{harvest index}}$$

Here w_h is the total harvest yield in energy units (e.g. MJ/m^2) and S is the incident photosynthetically active solar radiation (PAR) in the same units. ϵ_i, ϵ_c, and η are unitless

factors that describe the fraction of PAR intercepted by the plant, the fraction intercepted energy that is converted into chemical energy as biomass, and how much of the overall biomass is harvested, respectively [337]. Over the second half of the 20th century, humanity experienced a "Green Revolution" where selective breeding of crop plants produced staggering yield increases in excess of threefold for some staple crops like rice and wheat. This was achieved mostly by increasing η and ϵ_i [172, 337]. Introduction of dwarfed crops, which partition much more of their biomass into grains and less into stalks, helped improve η, while breeding for larger leaf sizes and rapid coverage of the ground post germination improved ϵ_i [172, 337].

Despite incredible progress in the 20th century, recent measurements demonstrate that traditional crop management strategies are ineffective or producing diminishing returns in many locales [244]. Moreover, climate change is likely to negatively impact crop yields and endanger global food supply [170]. A continuous stream of research articles [15, 158, 166, 290, 328], academic reviews [13, 24, 91, 172, 215, 236, 331, 337] and popular articles highlight new strategies to feed the world and sequester CO_2 via photosynthesis. Harvest index, η, has already increased by about twofold in the 20th century, however. Based partly on the implication that further improvements to η are implausible [172], most of the works cited above focus on increasing ϵ_i and ϵ_c. Conversion efficiency ϵ_c is a particularly attractive target because many important crop plants have measured ϵ_c threefold beneath theoretical maxima or more [337].

Another important contemporary approach to the problem of feeding humanity and reducing anthropogenic greenhouse gas emissions is to focus on the left-hand-side of the yield equation, i.e. to reduce the required w_h. Since the "feed-to-food" ratio of common livestock ranges from 5 to 40 [90], it takes 5-40 calories of plant-based calories to produce 1 calorie of meat. Meat production, especially beef, is a phenomenally wasteful use of primary productivity [90]. If we take seriously the enormous costs associated with meat production [90, 227], advocate for dietary shift [273], and work to produce viable meat substitutes [115], total demand for plant products should decrease substantially, freeing land for reforestation and other pro-environmental uses. This important work is complementary to efforts to improve photosynthetic yields by engineering plants. In my view, there is no single "silver bullet" to end climate change, abasement of Earth biomes, or human hunger. Rather, scientific, engineering and policy efforts should proceed arm-in-arm on multiple fronts to address the vital problems facing humanity in our flawed stewardship of this apparently singular planet.

I have chosen to work on these problems through the study and engineering of bacterial CCMs. Several labs are working towards improving ϵ_c by installing different CCMs in C3 crop plants [52, 187], though only partial results have been reported so far [171, 165, 212, 164]. A motivating goal of my work has been to understand the basic principles of the bacterial CCM well enough to understand whether these plant reconstitution efforts are likely to succeed. I have chosen to interrogate CCMs by a combination of mathematical modeling, functional genomics and synthetic reconstitution in pursuit of these goals. I focus on the bacterial CCM because it is well-characterized, both genetically and biochemcially, and appears to require only ≈ 20 genes to function (Figure 1.7 and Table 1.2).

1.7 Interrogating Bacterial CCMs

The subsequent chapters describe a series of studies that interrogate why CCMs arose so many times independently, how the bacterial CCM functions, and which genes and activities are necessary and sufficient for the bacterial CCM to function in cells. These efforts help contextualize the evolution of the bacterial CCM and refine current models of its function.

Chapter 2 describes a thorough meta-analysis of measured Rubisco kinetic parameters. An infinitely-fast, infinitely CO_2-selective Rubisco would obviate the need for a CCM (Figure 1.2). Yet we observe that all known Rubiscos have a maximum carboxylation rate $k_{\mathrm{cat,C}} \leq 15$ s^{-1} and limited CO_2-selectivity ($S_{\mathrm{C/O}} < 250$). In fact, selectivity is negatively correlated with $k_{\mathrm{cat,C}}$ overall, suggesting that there is some tradeoff between these two crucial properties of Rubisco. This fact might explain the prevalence of CCMs - flooding a fast-but-nonselective Rubisco with CO_2 makes its CO_2-selectivity moot. Previous efforts used small datasets to examine correlations between Rubisco kinetic parameters [301, 262] in order to reason about tradeoffs between them. I collected and analyzed a much larger dataset of Rubisco kinetic parameters and found that some existing tradeoff models must be rejected. These analyses led me to propose a simplified model of tradeoffs between Rubisco kinetic parameters, a model which can help explain why fast-and-selective Rubiscos have not been found despite abundant evidence that rate and selectivity of Rubisco carboxylation are under strong selection.

Chapter 3 describes a detailed mathematical model of the bacterial CCM aimed at understanding how CCM function is intertwined with bacterial physiology. Previous models helped elucidate the basic mechanisms of the CCM, showing that C_i must be out-of-equilibrium in the cytosol [228] and that the carboxysome must constitute a diffusional boundary to stem CO_2 leakage [248]. More recent models have considered the optimal packing of enzymes into the carboxysome [128] and attempted to map plausible trajectories for CCM evolution by considering the efficiency of intermediate forms [182]. However, all previous models produced results that were inconsistent with biochemical and physiological intuition on a number of fronts. In particular, a variety of implausible membrane permeability coefficients were used to model the membrane transit of CO_2 and HCO_3^-. These assumptions typically resulted in C_i uptake fluxes well in excess of measured values. For these fluxes to be realized, more than 50% of the cell membrane area would need to be occupied by C_i transporters alone.

The *E. coli* cell membrane is about 50% protein overall [196], including all ion channels, porins, nutrient transporters, secretory systems and membrane-bound respiratory proteins. The plant thylakoid membrane is likely the most densely protein-packed membrane in biology and is estimated to be 75% protein [155]. Filling 50% of the cell membrane with C_i transporters would likely disrupt membrane integrity and leave little space for the essential biochemistry of nutrient transport and photosynthesis. Previous models of the CCM had neglected the effects of pH on C_i pool and on enzymatic activities. As described in Chapter 3, integrating the pH-dependence of all CCM components produces a model that is consistent with bacterial physiology. This consistent model greatly aids in understanding the bacterial

CCM and implies that known CCM components are sufficient for its presumed function - ensuring Rubisco carboxylates at near maximum rate and does not oxygenate RuBP very frequently.

Though modeling suggests that known CCM components are sufficient, it remains unclear whether our current genetic "parts list" of the bacterial CCM is complete. **Chapter 4** describes the generation and high-through characterization of a whole-genome knockout library in a model CCM host, the chemoautotrophic γ-proteobacterium *H. neapolitanus*. This screen represents the most comprehensive genetic characterization of the bacterial CCM to date. As a result of this comprehensive work, we can now convincingly assert that the bacterial CCM requires no more than 25 genes to function in a native host. These data help orient the synthetic reconstitution efforts described in Chapter 5. In particular, the screen highlighted novel C_i transporters and poorly-characterized Rubisco chaperones as playing an important role in the bacterial CCM.

Chapter 5 describes a synthetic reconstitution of the bacterial CCM in an organism that does not natively express Rubisco or perform net CO_2 fixation, *E. coli*. This reconstitution was achieved by constructing a novel *E. coli* strain, CCMB1, that depends on Rubisco for growth in defined media. As expected based on the known limitations of Rubisco, Rubisco expression in CCMB1 rescues growth in elevated CO_2 but not in ambient CO_2 concentrations. Expression of 20 genes of the *H. neapolitanus* CCM rescues growth in ambient CO_2, demonstrating that a subset of the genes identified in Chapter 4 are sufficient to produce a functioning heterologous CCM. This reconstitution will enable future experiments interrogating the detailed mechanisms of the bacterial CCM, investigating, for example, the role of Rubisco chaperones and identifying the minimal gene set required for the CCM to function heterologously.

Chapter 2

Empirical Evaluation of Tradeoffs in Rubisco Kinetics

2.1 Abstract

Rubisco is the most abundant enzyme in the biosphere and one of the best-characterized. Based on correlations between Rubisco kinetic parameters, it is widely posited that tradeoffs embedded in the catalytic mechanism constrain its specificity and maximum catalytic rate. However, the reasoning that established this view was based on data from ≈20 organisms. We re-examine these tradeoff models using a dataset from ≈300 organisms. Most correlations are substantially attenuated, with the inverse relationship between carboxylation k_{cat} and specificity $S_{\mathrm{C/O}}$ being a key example. Only one tradeoff model is bolstered by our dataset. In this model, increasing catalytic efficiency ($k_{\mathrm{cat}}/K_{\mathrm{M}}$) for carboxylation requires increased catalytic efficiency for the competing oxygenation reaction, evidenced by strong power-law correlation between catalytic efficiencies. Our results imply that Rubisco evolution is constrained primarily by the physicochemical limits of O_2/CO_2 discrimination, which should reframe efforts to engineer this very central enzyme.

2.2 Introduction

Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase (Rubisco) is the primary carboxylase of the Calvin-Benson-Bassham (CBB) cycle - the carbon fixation cycle responsible for growth throughout the green lineage and many other autotrophic taxa - and the ultimate source of nearly all carbon atoms entering the biosphere [27]. Typically, 20-30% of total soluble protein in C_3 plant leaves is Rubisco [110]. As Rubisco is so highly expressed and plants are the dominant constituents of planetary biomass [28], it is often said that Rubisco is the most abundant enzyme on Earth [27]. Since Rubisco is ancient (>2.5 billion years old), abundant, and remains central to biology, one might expect it to be exceptionally fast. But Rubisco is not fast [26, 31, 277, 262, 301]. Typical central metabolic enzymes have a maximum catalytic

rate $k_{cat} \approx 80$ s^{-1} [26], but more than 95% of Rubisco carboxylation k_{cat} values are between 1-10 s^{-1}.

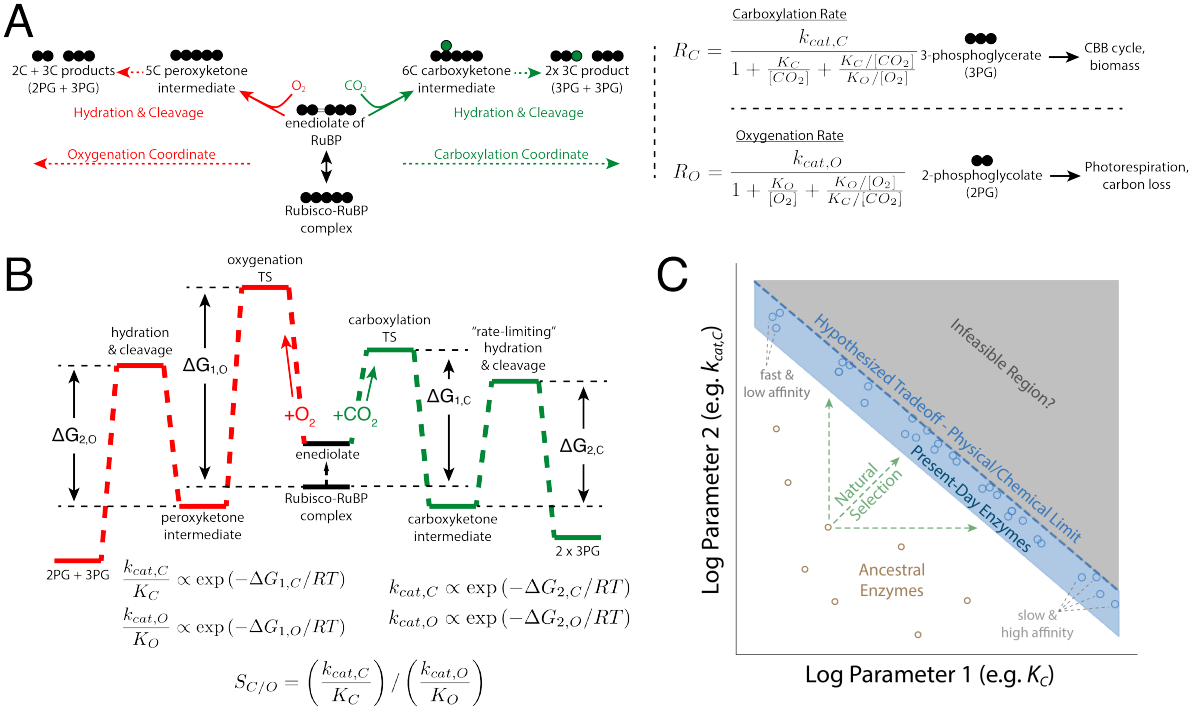


Figure 2.1: Description of the catalytic mechanism of Rubisco. The "middle-out" diagram in Panel A shows the ordered mechanisms of carboxylation and oxygenation. Circles represent carbon atoms. RuBP is first isomerized to an enediolate before carboxylation or oxygenation can occur. Addition of CO_2 or O_2 to the enediolate of RuBP are considered irreversible as are the subsequent hydration and cleavage steps of carboxylation and oxygenation arms. Carboxylation displays effective Michaelis-Menten kinetics (maximum catalytic rate $k_{cat,C}$, half-maximum CO_2 concentration $K_M = K_C$) with competitive inhibition by O_2 (assuming half-maximum inhibitory O_2 concentration $K_i = K_O$). Carboxylation results in net addition of one carbon to the five-carbon RuBP, which produces two 3PG molecules. 3PG is part of the CBB cycle and can therefore be used to continue the cycle and produce biomass. Oxygenation also displays effective Michaelis-Menten kinetics ($k_{cat,O}$, $K_M = K_O$, half-max inhibitory CO_2 concentration $K_I = K_C$). Oxygenation of RuBP produces one 3PG and one 2PG. 2PG is not part of the CBB cycle and must be recycled through photorespiration to avoid the loss of both carbons in 2PG to central metabolism. Per-active site rates of carboxylation (R_C) and oxygenation (R_O) can be calculated from kinetic parameters and the CO_2 and O_2 concentrations. The reaction coordinate diagram in panel (B) mirrors panel A and describes Rubisco carboxylation and oxygenation as a function of two "effective" barriers as in [262]. The first effective barrier describes enolization and gas addition while the second describes hydration and bond cleavage. Given standard assumptions (Appendix A), the respective catalytic efficiencies (k_{cat}/K_M) are related to the height of the first effective barrier while the k_{cat}s are related to the second. The first barrier to oxygenation is drawn higher than for carboxylation because Rubisco oxygenation is typically much slower than carboxylation. The net reactions of RuBP carboxylation and oxygenation are both quite thermodynamically favorable ($\Delta_r G'^m \approx$ -28 kJ/mol and -520 kJ/mol respectively [102]). As kinetic parameters are linearly related to the log of effective energy barriers, energetic tradeoffs should manifest as linear correlations in a log-log plot of kinetic parameters (C). As Rubisco is central to photoautotrophic growth, we expect that natural selection pushed ancestral enzymes towards the upper limits of catalytic capacity determined by any physicochemical tradeoffs (diagrammed as a dashed blue line). In the hypothetical example of a tradeoff between maximum carboxylation rate ($k_{cat,C}$) and CO_2 affinity (K_C), different environmental CO_2 and O_2 concentrations are expected to select for different combinations of rate and affinity, resulting in present-day enzymes occupying different regions in the blue shaded region.

In addition to relatively low k_{cat} values, Rubisco reacts with O_2 in a process called oxy-

genation (Figure 2.4A-B). Although both carboxylation and oxygenation of the five-carbon substrate ribulose 1,5-bisphosphate (RuBP) are energetically favorable, RuBP carboxylation is the productive reaction for incorporating carbon from CO_2 into precursors that generate biomass (Figure 2.4AB). While it might play a role in sulfur and nitrogen metabolism, oxygenation of RuBP is often considered counterproductive as it occupies Rubisco active sites and yields a product (2-phosphoglycolate, 2PG) that is not part of the CBB cycle and must be recycled through metabolically-expensive photorespiration at a partial loss of carbon [33, 50]. As such, oxygenation can substantially reduce the net rate of carboxylation by Rubisco, depending on CO_2 and O_2 concentrations and the kinetic parameters of the particular enzyme. There are at least four distinct Rubisco isoforms in nature [136, 167], but all Rubisco isoforms catalyze carboxylation and oxygenation of RuBP through the multistep mechanism described in Figure 2.4A [10, 66]. Even though many autotrophs depend on Rubisco carboxylation for growth, all known Rubiscos are relatively slow carboxylases and fail to exclude oxygenation (Figure 2.4A-B).

As a concrete example, the fastest-carboxylating Rubisco ever observed (at 25 °C) is from the freshwater cyanobacterium *S. elongatus* PCC 7942 [212]. The PCC 7942 Rubisco has a maximum per-active site carboxylation rate ($k_{\mathrm{cat,C}}$) reported as 14.4 s^{-1}. However, because present-day atmosphere contains abundant O_2 and relatively little CO_2 (≈21% O_2, ≈0.04% CO_2), this Rubisco carboxylates at a rate 20-fold below maximum in ambient conditions ($R_C \approx 0.7$ s^{-1} per active site, rate law in Figure 2.4A). Due to its relatively low specificity towards CO_2, PCC 7942 Rubisco will also oxygenate RuBP in ambient conditions at a rate that is about half the carboxylation rate ($R_O \approx 0.3$ s^{-1}), which would necessitate substantial photorespiratory flux to recycle 2PG. Since downstream processing of 2PG by the canonical C_2 photorespiratory pathway leads to the loss of one carbon for every two 2PG [33, 50], every two oxygenations "undoes" a carboxylation. So the net rate of carboxylation by PCC 7942 Rubisco in ambient conditions is roughly $R_C - R_O/2 \approx 0.55$ carboxylations per second, or ≈0.04 $k_{\mathrm{cat,C}}$. Indeed, all cyanobacteria use a CO_2-concentrating mechanism (CCM) to ensure that Rubisco functions in a CO_2-rich environment. High CO_2 ensures that oxygenation is inhibited and that carboxylation proceeds at near-maximum rate [248, 182]. Just tenfold enrichment of CO_2 above ambient increases the carboxylation rate of PCC 7942 Rubisco to ≈5 s^{-1} and suppress oxygenation to ≈0.2 s^{-1}, giving a net carboxylation rate of ≈ 4.6 s^{-1} per active site.

For comparison, the well-studied Rubisco from spinach leaves (*S. oleracea*) is characteristic of plant Rubiscos in that it has slower $k_{\mathrm{cat,C}} \approx 3$ s^{-1} and also much greater affinity towards CO_2 than the *S. elongatus* enzyme (half-maximum CO_2 concentration $K_C \approx 12$ μM for spinach as compared to $K_C \approx 170$ μM for *S. elongatus*). As a result, the spinach enzyme outperforms the cyanobacterial enzyme in ambient atmosphere, achieving a carboxylation rate of $R_C \approx 1.2$ s^{-1} and an oxygenation rate of $R_O \approx 0.4$ s^{-1}. This represents a net carboxylation rate of ≈ 1 s^{-1}, nearly double that of the cyanobacterial example above. Spinach is a C_3 plant, meaning it does not have a CO_2 concentrating mechanism, which may explain why it employs a slow-but-specific enzyme for catalysis in ambient atmosphere. Still, most enzymes catalyze far more than 1 reaction per second [26], which leads many to wonder

if Rubisco catalysis could be improved by engineering. Improved Rubisco carboxylation is expected to increase C_3 crop yields [337], but a substantially improved enzyme has evaded bioengineers for decades [291]. The repeated evolution of diverse CO_2 concentrating mechanisms, which modulate the catalytic environment rather than the enzyme itself, raises further doubts about whether Rubisco catalysis can be strictly improved [243].

Various nomenclature has been used to describe the kinetics of Rubisco carboxylation and oxygenation [262, 301, 226] since its discovery in the 1950s [323]. Here we use $k_{\mathrm{cat,C}}$ and $k_{\mathrm{cat,O}}$ to denote turnover numbers (maximum per active site catalytic rates in units of s^{-1}) for carboxylation and oxygenation respectively. K_{C} and K_{O} denote the Michaelis constants (half-saturation concentrations in μM units) for carboxylation and oxygenation. The specificity factor $S_{\mathrm{C/O}} = (k_{\mathrm{cat,C}}/K_{\mathrm{C}})\ /\ (k_{\mathrm{cat,O}}/K_{\mathrm{O}})$ is a unitless measure of the relative preference for CO_2 over O_2 (Figure 2.4A-C). Since $S_{\mathrm{C/O}}$ relates only to the ratio of kinetic parameters, it should be noted that higher $S_{\mathrm{C/O}}$ does not necessarily imply higher carboxylation rates. Rather, absolute carboxylation and oxygenation rates depend on the CO_2 and O_2 concentrations (Figure 2.4A) which can vary between organisms and environments.

As data on bacterial, archaeal and plant rubiscos has accumulated over the decades, many researchers have noted that fast-carboxylating Rubiscos are typically less CO_2-specific [139, 213, 219]. In other words, Rubiscos with high $k_{\mathrm{cat,C}}$ were observed to have lower $S_{\mathrm{C/O}}$ due either to lower CO_2-affinity (high K_{C}) or higher catalytic efficiency towards O_2 ($k_{\mathrm{cat,O}}/K_{\mathrm{O}}$). This negative correlation between $k_{\mathrm{cat,C}}$ and $S_{\mathrm{C/O}}$ is often cited to motivate the idea that the Rubisco mechanism imposes a tradeoff between carboxylation rate and specificity that constrains the evolution of this enzyme [21, 262, 301]. Indeed, if the Rubisco mechanism imposes a tradeoff between $k_{\mathrm{cat,C}}$ and $S_{\mathrm{C/O}}$ we would expect strong correlation between those parameters because Rubisco is so central to autotrophic life and has, therefore, likely experienced strong selection pressure. As diagrammed in Figure 2.4C, strong selection for Rubisco carboxylation should push the enzyme towards a point where its kinetics can be improved no further (as determined by the dashed blue tradeoff line). Since different kinetic parameters are preferable under different CO_2 and O_2 concentrations (as described above) strong selection can result in a situation where the kinetics of present-day Rubiscos extracted from various organisms trace out a curve determined by the underlying tradeoff [262, 281, 301].

Two distinct tradeoff models have been proposed to explain the observed correlations between Rubisco kinetic parameters [262, 301]. Although the proposed models are substantively different, both models imply limitations on the concurrent improvement of the maximum carboxylation rate ($k_{\mathrm{cat,C}}$) and specificity ($S_{\mathrm{C/O}}$) of natural Rubiscos. While these hypotheses appeal to physical and chemical intuition, they are based on data from only ≈20 organisms. Here we take advantage of the accumulation of new data to examine whether new data evidence the same correlations. We collected and curated literature measurements on ≈300 Rubiscos. Examining this dataset, we find that most previously-reported correlations between Rubisco kinetic parameters are substantially attenuated by the addition of new data, with the negative correlation between $k_{\mathrm{cat,C}}$ and specificity $S_{\mathrm{C/O}}$ being a key example.

Only one previously-reported correlation remains both strong and statistically significant

in the extended dataset - a power-law correlation between the catalytic efficiency for carboxylation ($k_{\mathrm{cat,C}}/K_{\mathrm{C}}$) and the catalytic efficiency for oxygenation ($k_{\mathrm{cat,O}}/K_{\mathrm{O}}$) first reported in [262]. We propose a simple physico-chemical model based on the Rubisco mechanism that can explain this very strong correlation. In this model, variation in catalytic efficiency ($k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$) derives solely from gating substrate access to the active site complex, which could help explain why Rubisco has been so recalcitrant to improvement by mutagenesis and rational engineering.

2.3 Results

An extended dataset of Rubisco kinetic parameters

To augment existing data, we collected literature data on ≈300 Rubiscos including representatives of clades and physiologies that had been poorly represented in earlier datasets e.g. diatoms, ferns, CAM plants and anaerobic bacteria (Figure 2.2A). We collected kinetic parameters associated with carboxylation and oxygenation - S, K_{C}, $k_{\mathrm{cat,C}}$, K_{O} and $k_{\mathrm{cat,O}}$- as well as measurements of the RuBP Michaelis constant (half-maximum RuBP concentration, K_{RuBP}) and experimental uncertainty for all values where available. All data considered were measured at 25 °C and near pH 8 to ensure that measured values are comparable (Methods).

The resulting dataset contains Rubisco kinetic parameters from a total of 286 distinct species including 319 $S_{\mathrm{C/O}}$ values, 275 $k_{\mathrm{cat,C}}$ values, 310 K_{C} values, 198 $k_{\mathrm{cat,O}}$ values and 256 K_{O} values (Figure 2.2B). The Michaelis constant for RuBP is not measured frequently and so only 45 values were extracted. In 198 cases there was sufficient data to calculate catalytic efficiencies for carboxylation ($k_{\mathrm{cat,C}}/K_{\mathrm{C}}$) and oxygenation ($k_{\mathrm{cat,O}}/K_{\mathrm{O}}$, Methods). Though the data include measurements of some Form II, III and II/III Rubiscos, they remain highly focused on the Form I Rubiscos found in cyanobacteria, diatoms, algae and higher plants, which make up $> 95\%$ of the dataset (Figure 2.2B). As such, we focus here on the kinetic parameters of Form I Rubiscos.

Rubisco kinetic parameters display very narrow dynamic range (Figure 2.2C). The geometric standard deviation, denoted σ^*, expresses multiplicative variability in the dataset. Geometric standard deviations are well-below one order-of-magnitude ($\sigma^* \ll 10$) for all parameters. Rubisco displays extremely low variation in $k_{\mathrm{cat,C}}$ ($\sigma^* = 1.5$) as compared to other enzymes for which 20 or more k_{cat} measurements are available (median $\sigma^* \approx 7$, Appendix A). Specificity $S_{\mathrm{C/O}}$ displays the least variation ($\sigma^* = 1.3$) of all parameters, though this may be due in part to overrepresentation of C_3 plants in the dataset, which occupy a narrow range of $S_{\mathrm{C/O}} \approx$ 80-120. Nonetheless, measurements of $S_{\mathrm{C/O}}$ for Form I and Form II enzymes are clearly distinct, with values ranging from and 7-15 for Form II and roughly 50-200 for Form I (Figure 2.2C).

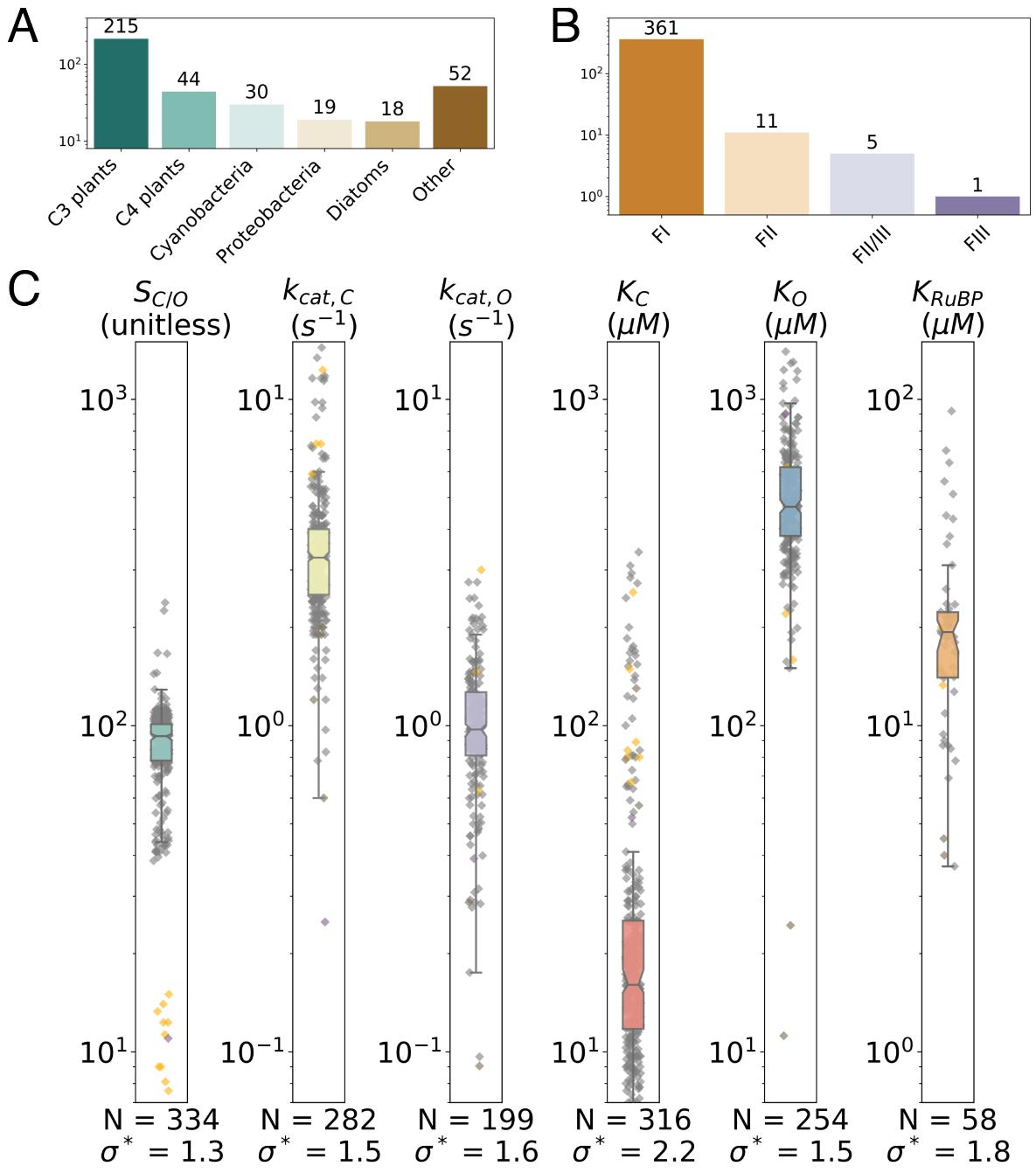


Figure 2.2: Summary of extended dataset of Rubisco kinetic parameters. We collected measurements of Rubisco kinetic parameters from a variety of organisms (A) representing four classes of Rubisco isoforms (B). The bulk of data represent Form I enzymes from green lineage organisms (A-B). As shown in panel C, the assembled kinetic parameters display narrow dynamic range. The box-plot and grey points describe the distribution of Form I Rubiscos while Form II Rubiscos are in yellow. Whiskers show the range of parameter values among Form 1 enzymes, the colored box gives the range of the central 50% of Form I values and the notch indicates the median. N is the number values and the geometric standard deviation of Form I data is reported as σ^*. $\sigma^* < 3$ for all parameters, meaning that a single standard deviation varies over less than threefold. All data presented in this figure are from wild-type Rubiscos measured at 25 °C near pH 8. More detailed histograms are given in Appendix A.

Energetic Tradeoffs Tend to Produce Power-Law Correlations

All measured kinetic parameters ($S_{C/O}$, $k_{cat,C}$, K_C, $k_{cat,O}$ and K_O) are by definition mathematically related to the microscopic rate constants of the Rubisco mechanism (Appendix A). Given common assumptions about irreversible and rate limiting steps (as elaborated below), this multi-step mechanism can be simplified so that logarithms of measured kinetic parameters are proportional to effective transition state (TS) barriers (Figure 2.4B, Appendix A). As such, tradeoffs between kinetic parameters are expected to emerge if effective TS barriers are constrained to vary together (Figure 2.4C). If, for example, lowering the effective TS barrier to CO_2 addition ($\Delta G_{1,C}$) requires an increase to the effective barrier to the subsequent hydration and cleavage steps of carboxylation ($\Delta G_{2,C}$, then we should observe a negative linear correlation between these barrier heights such that $\Delta G_{1,C} \propto -\Delta G_{2,C}$. Since, as shown in Figure 2.4B, $k_{cat,C}/K_C$ is related to the first effective carboxylation barrier ($\ln(k_{cat,C}/K_C) \propto -\Delta G_{1,C}$) and $k_{cat,C}$ to the second ($\ln(k_{cat,C}) \propto -\Delta G_{2,C}$), linear correlation between TS barrier energies should translate into log-scale correlation between kinetic parameters such that $\ln(k_{cat,C}/K_C) \propto \ln(k_{cat,C})$. These log-linear relationships are known as power laws and motivate us and others to investigate the kinetic parameters on a log-log scale.

We expect to observe strong power-law correlations between pairs of kinetic parameters when three conditions are met: (I) the associated energy barriers are subject to a tradeoff that forces them to vary together; (II) these constraints affect the net rate of carboxylation by Rubisco; and (III) the selection pressure imposed during Rubisco evolution was sufficient to reach the limits imposed by the tradeoff (as diagrammed in Figure 2.4C). As Rubisco is the central enzyme of photoautotrophic growth, we assume here that it evolved under selection pressure towards maximizing the net rate of carboxylation in each host [262]. Notably, different host physiologies and growth environments can affect the catalytic environment Rubisco experiences - Rubiscos in different organisms experience different temperature, pH and prevailing CO_2 and O_2 concentrations (e.g. due to an anaerobic host or a CO_2 concentrating mechanism enriching CO_2), which we expect to select for different combinations of kinetic parameters (Figure 2.4C).

Correlations between Kinetic Parameters of Form I Rubiscos

We performed a correlation analysis to investigate relationships between Rubisco kinetic parameters. Pairwise correlations between log-transformed Form I Rubisco kinetic parameters are given in Figure 2.3. Linear scale correlations are reported in Appendix A.

Correlations between $k_{cat,C}$ and $S_{C/O}$ as well as $k_{cat,C}$ and K_C were previously highlighted to support particular mechanistic tradeoff models [262, 301]. However, both correlations are substantially attenuated by the addition of new data ($R \approx 0.6$, Figure 2.3). Figure 2.4 inspects these two correlations in greater detail. Figure 2.4A plots $k_{cat,C}$ against $S_{C/O}$ and shows that these parameters are only modestly correlated in the extended dataset, with $R \approx 0.6$ (and extremely sensitive to outliers) as compared to $R \approx 0.9$ in previous analyses [262,

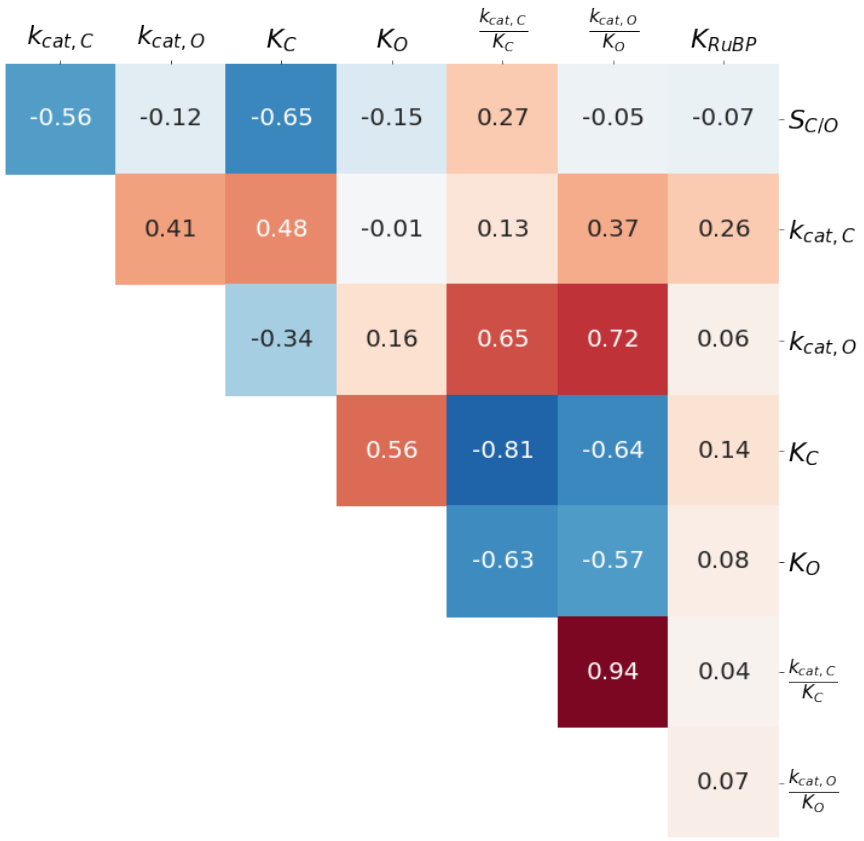


Figure 2.3: Correlations between kinetic parameters are mostly attenuated by addition of new data. The figure summarizes Pearson correlations (R) between pairs of log-transformed Form I Rubisco kinetic parameters. When multiple measurements of the same Rubisco were available, the median value was used (Methods). The $S_{C/O}$-K_C, $S_{C/O}$-$k_{cat,C}$, and K_C-$k_{cat,C}$ correlations are of particular interest because they were highlighted in previous works. None of these pairs give R > 0.7. The strongest observed correlation is between the catalytic efficiencies for carboxylation and oxygenation, $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ (R = 0.94).

301]. Similarly, Figure 2.4B plots $k_{cat,C}$ against K_C and shows that this correlation is also weakened, with R ≈ 0.5 as compared to R ≈ 0.9 previously [262]. We interpret the weakened correlations detailed in Figures 3 and 4 as evidence that previously-proposed tradeoffs should be revisited. Examining Figure 2.3 shows that the strongest log-scale correlation is between the catalytic efficiencies $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ (R = 0.93, P < 10^{-10}). We discuss possible explanations for this very strong correlation in detail below.

One might wonder why so many pairs of Rubisco kinetic parameters correlate with appreciable R values (e.g. R > 0.3 for 13 of 28 pairs in Figure 2.3). Some level of correlation is expected because the measured parameters are mathematically interrelated through the microscopic mechanism of Rubisco as it is commonly understood. For example, when we derive expressions for $k_{cat,C}$ and K_C from the Rubisco mechanism, they share common factors that should produce some level of correlation even in the absence of any tradeoff (Appendix A). Similarly, $S_{C/O}$ is defined as $(k_{cat,C}/K_C)$ / $(k_{cat,O}/K_O)$ and might correlate positively with $k_{cat,C}$ for this reason. Because modest correlation is expected, we focus on very strong correlations since these may yield insight into mechanistic constraints on Rubisco evolution.

Principal components analysis (PCA) of Rubisco kinetic parameters was previously used to interrogate constraints on Rubisco evolution. It was argued that Rubisco adaptation is constrained to a one-dimensional landscape because the first principal component (PC1)

explained > 90% of the variance in Rubisco kinetics [262]. In a one-dimensional landscape model all kinetic parameters are tightly interrelated so that changing one (e.g. $k_{\mathrm{cat,C}}$) forces all others to assume predetermined values. However, our extended dataset is not well-approximated as one-dimensional. While the orientation of PC1 is not substantially altered by the addition of tenfold more measurements, it now explains ≈70% instead of >90% of the variance in Rubisco kinetics [262]. Three principal components are required to explain >90% of the variation in the extended dataset (Appendix A), consistent with the overall reduction in pairwise correlation documented in Figure 2.3. We therefore proceed to evaluate correlations predicted by specific tradeoff models advanced in [262, 301] within the extended dataset.

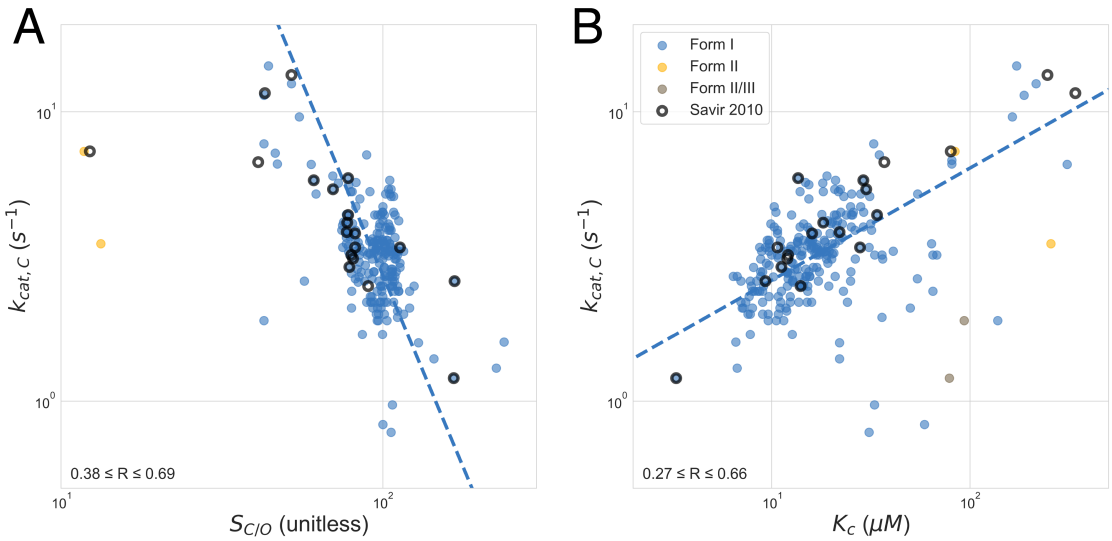


Figure 2.4: Focal correlations of previous analyses are not robust to new data. Points with black outlines are those in Savir et al. 2010 and dashed blue lines represent the best fit to all Form I Rubiscos in the extended dataset. Panel A plots the maximum carboxylation rate $k_{\mathrm{cat,C}}$ against specificity $S_{\mathrm{C/O}}$ as in [301]. Considering only Form I Rubiscos, $k_{\mathrm{cat,C}}$ and $S_{\mathrm{C/O}}$ correlate with $R \approx -0.6$. Bootstrapping gives very wide 95% confidence intervals (CIs) of (-4.0, -2.0) for the fit exponent and ($3\text{x}10^4$, $3\text{x}10^8$) for the exponential prefactor (the slope and intercept in log-log scale respectively) indicating that the form of $k_{\mathrm{cat,C}}$-$S_{\mathrm{C/O}}$ correlation is very uncertain. Panel B plots $k_{\mathrm{cat,C}}$ against the Michaelis constant for CO_2 (K_{C}) as in [262, 301]. $R \approx -0.5$ as compared to the previously reported value of 0.92. This fit is substantially more robust to outliers with bootstrapping giving 95% CIs of (0.3, 0.5) and (0.8, 1.5) for the fit exponent and prefactor respectively. More detailed plots are given in Appendix A.

Re-evaluation of Proposed Tradeoff Models

Two distinct mechanistic tradeoff models have been advanced [262, 301]. Savir et al. 2010 cast these proposals in energetic terms by relating the measured catalytic parameters to effective transition state barrier heights (Figure 2.1B, Appendix A). The first tradeoff model posits that that increased specificity towards CO_2 necessitates a slower maximum carboxylation rate, $k_{\mathrm{cat,C}}$ [262, 301]. Tcherkez et al. 2006 propose that this tradeoff is caused by

stabilization of the first carboxylation transition state. Under this model a stable Rubisco-TS complex produces high CO_2-specificity but slows the subsequent carboxylation steps and limits $k_{\mathrm{cat,C}}$ (Appendix A). This model can be construed in energetic terms as follows: lowering the effective barrier to CO_2 addition ($\Delta G_{1,C}$ in Figure 2.5A) will make Rubisco more CO_2-specific even if none of the oxygenation parameters change. This tradeoff model posits a negative coupling between CO_2 addition and the subsequent carboxylation steps of hydration and bond cleavage (effective barrier height $\Delta G_{2,C}$ diagrammed in Figure 2.5A). Therefore, the energetic interpretation of the first model predicts a negative correlation between $\Delta G_{1,C}$ and $\Delta G_{2,C}$ and, as a result, a negative power-law correlation between $k_{\mathrm{cat,C}}$ and $k_{\mathrm{cat,C}}/K_C$.

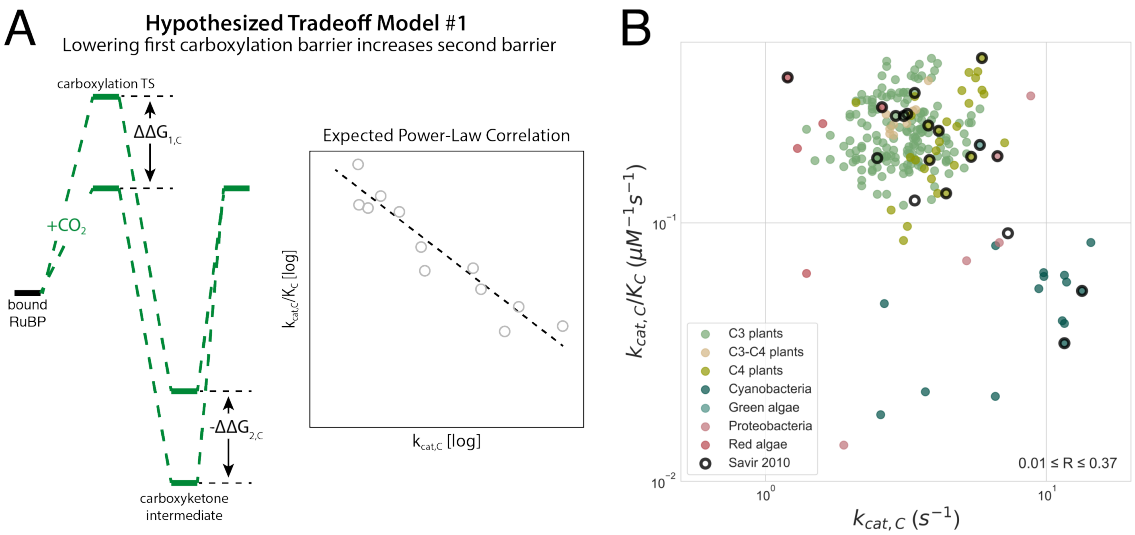


Figure 2.5: Negative power-law correlation between $k_{\mathrm{cat,C}}$ and k_{cat}/K_C is not supported by the extended dataset. Under the tradeoff model in panel (A), CO_2-specific Rubiscos have low barriers to enolization and CO_2 addition (first effective carboxylation barrier $\Delta G_{1,C}$), but lowering the first effective barrier necessarily increases the height of the effective barrier to subsequent carboxylation steps ($\Delta G_{2,C}$). This tradeoff might be due to coupling between the carboxylation transition state and the carboxyketone carboxylation intermediate [301], where stabilizing the TS also stabilizes the intermediate diagrammed in panel (A) and described in Appendix A. In this case we would expect $\Delta G_{1,C}$ and $\Delta G_{2,C}$ to be negatively correlated, which would manifest as negative linear correlation on a log-log plot of $k_{\mathrm{cat,C}}$ vs. $k_{\mathrm{cat,C}}/K_C$. (B) The extended dataset does not evidence the expected power-law correlation (R = 0.02, P = 0.8 for Form I enzymes). Fitting the entire dataset gives R = 0.13 as compared to the previously-reported R = -0.95 in [262], where two outliers were omitted. Restricting focus to particular physiologies like C_3 plants does not recover the expected correlation.

The energetic interpretation of this model was previously supported by an inverse power-law relationship between $k_{\mathrm{cat,C}}$ and $k_{\mathrm{cat,C}}/K_C$ [262]. The reported power-law correlation, however, is not strongly supported by our extended dataset (Figure 2.5B). The true barrier height to CO_2 addition depends on the CO_2 concentration, which could partially explain the apparent lack of correlation. However, correlation is not improved by restricting focus to C_3 plants for which data is abundant and for which measured leaf CO_2 concentrations vary by only 20-30% due to variation in CO_2 conductance and Rubisco activity [51, 92].

Absence of correlation does not necessarily imply the absence of a tradeoff. Rather, if the Rubisco mechanism couples $k_{cat,C}$ and $k_{cat,C}/K_C$, much decreased correlation over the extended dataset ($R < 0.4$) could result from several factors including noise in measurements, bias in organism choice leading to undersampling of faster Rubiscos (e.g. those from cyanobacteria) or, alternatively, insufficient selection pressure.

The second mechanistic tradeoff model - wherein faster CO_2 addition entails faster O_2 addition as well [262] - is extremely well-supported by the extended dataset (Figure 2.6). This model was previously supported by a power-law relationship between the catalytic efficiencies $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ with an exponent of 0.5 ($k_{cat,O}/K_O \propto (k_{cat,C}/K_C)^{0.5}$). As $k_{cat,C}/K_C$ is exponentially related to the first effective carboxylation barrier ($\ln(k_{cat,C}/K_C) \propto -\Delta G_{1,C}$) and $k_{cat,O}/K_O$ to the first effective oxygenation barrier ($\ln(k_{cat,O}/K_O) \propto -\Delta G_{1,O}$), the power-law relationship was taken to imply that decreasing the barrier to CO_2 addition will also decrease the barrier to O_2 addition ($0.5\Delta G_{1,C} - \Delta G_{1,O} = C$, Figure 2.6A). The extended dataset evidences clear power-law correlation between $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ (Figure 2.6B). Most measurements of Form I Rubiscos lie along a robust line of positive correlation in a log-log plot, while Form IIs appear to deviate systematically. The Form I enzymes display a remarkably high-confidence ($R = 0.93$, $P < 10^{-10}$) power-law relationship with and exponent of roughly 1.0: $k_{cat,O}/K_O \propto (k_{cat,C}/K_C)^{1.06}$ (Figure 2.6B).

$S_{C/O}$ is defined as the ratio of $k_{cat,C}/K_C$ to $k_{cat,O}/K_O$. A power law exponent of $\approx$1.0 implies that $S_{C/O}$ is constant (Appendix A). However $S_{C/O}$ is not strictly constant - it varies about tenfold between Form I and Form II Rubiscos and about threefold among Form I enzymes (Figure 2.2C). Subdividing the Form I enzymes by host physiology (C_3 plants, C_4 plants, cyanobacteria, etc.) reveals that all groups with sufficient data display a strong and statistically-significant power-law relationship between $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ (Figure 2.6C, Appendix A). The power-law exponent differs consistently from the previous value of 0.5 [262]. We now find a roughly 1:1 relationship of $\Delta G_{1,C}$ - $\Delta G_{1,O}$ = C, meaning that a decrease in the CO_2 addition barrier is associated with an *equal* decrease in the barrier to O_2 addition. We estimate a 95% confidence interval (CI) of 0.98-1.24 for the exponent of this power law relationship for Form I enzymes, or about double the previously-reported value.

Implications for the mechanism of CO_2/O_2 discrimination by Rubisco

Figure 2.6 shows that effective barriers to CO_2 and O_2 addition and oxygenation vary in proportion with each other ($\Delta G_{1,C} - \Delta G_{1,O} \approx$ constant). A roughly 1:1 correlation between effective barriers to CO_2 and O_2 addition suggests that a single factor controls both. We offer a model based on the known catalytic mechanism of Rubisco that would produce a 1:1 relationship between barriers. In this model, the RuBP-bound Rubisco active site fluctuates between reactive and unreactive states (Figure 2.7A). We deonte the fraction of enzyme in the reactive state as ϕ. In the unreactive state neither oxygenation or carboxylation is possible. In the reactive state, either gas can react its intrinsic rate, which does not

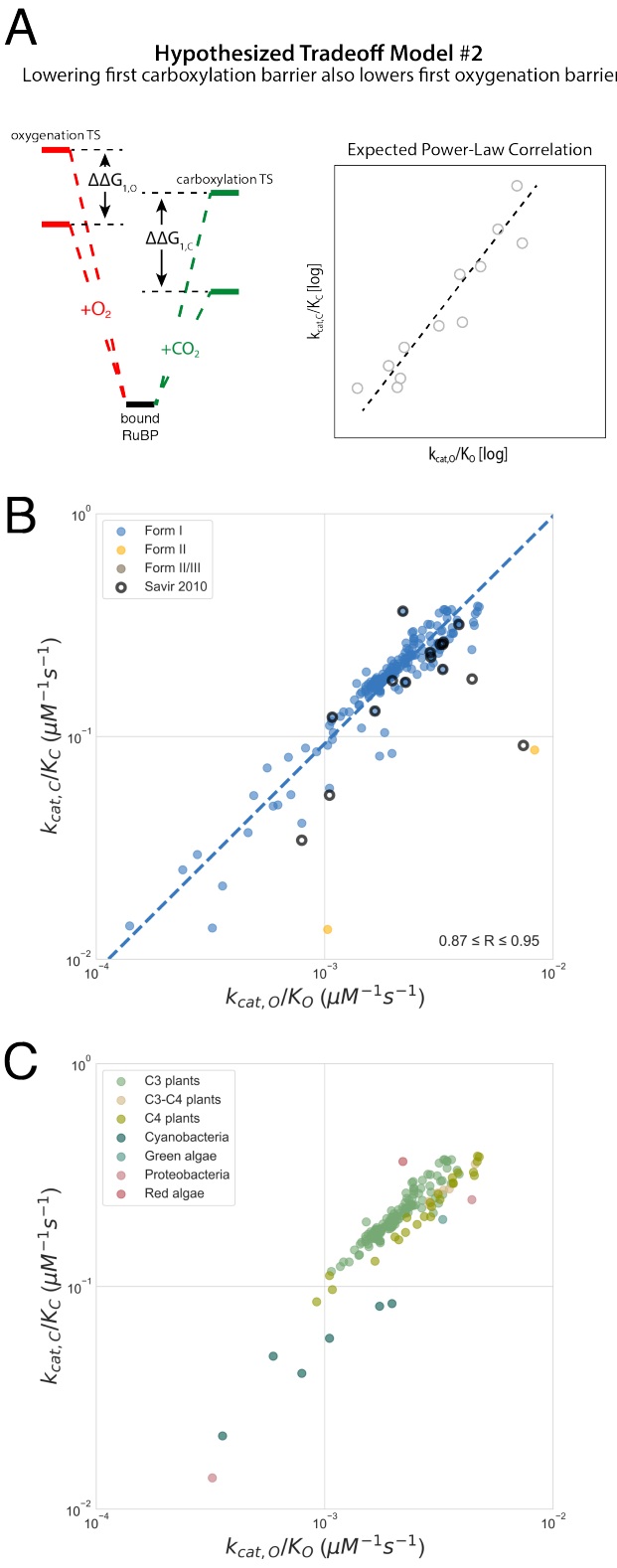


Figure 2.6: The second mechanistic proposal is remarkably well-supported by the extended dataset, but with a different power-law exponent. (A) In this proposal, CO_2 and O_2 addition rates are coupled, with faster CO_2 addition necessitating faster O_2 addition. This can be framed in energetic terms, where lowering the effective barrier to enolization and CO_2 addition ($\Delta G_{1,C}$) lowers the first effective barrier to O_2 addition ($\Delta G_{1,O}$) as well. Given this model, we would expect the barrier heights to be positively correlated, which would manifest as a positive linear correlation on a log-log plot of $k_{cat,C}/K_C$ vs $k_{cat,O}/K_O$. (B) most measurements cluster along a power-law (linear in log-log) front in the $k_{cat,C}/K_C$ vs $k_{cat,O}/K_O$ plot (dashed blue line, R = 0.94). While some Form I Rubiscos appear to lie beneath this front, Form II and Form II/III enzymes deviate most profoundly. A total least squares fit to the Form I enzymes produces a very strong power-law correlation ($P < 10^{-10}$, blue dashed line). 95% CIs for the exponent and prefactor are (0.93, 1.1) and (63, 199), respectively. The best fit power law is $(k_{cat,O}/K_O \propto (k_{cat,C}/K_C)^{1.04})$, but forcing an exponent of 1.0 gives a fit of nearly identical quality. (C) Restricting focus to particular physiologies - e.g. C_3 and C_4 plants, cyanobacteria - reveals that each grouping obeys a distinct power law. These power laws differ primarily in the exponential prefactor, which causes variation in the Y-intercept but not the slope on a log-log plot. 95% CIs on the power-law exponent are (0.87, 1.01) for C_3 plants, (0.82, 1.01) for C_4 plants, (0.38, 1.31) for C_3-C_4 plants and (0.38, 1.06) for cyanobacteria.

vary across Rubiscos of the same class. Since RuBP must undergo enolization in order for carboxylation or oxygenation to occur, ϕ may be determined by the degree of enolization of RuBP (Appendix A).

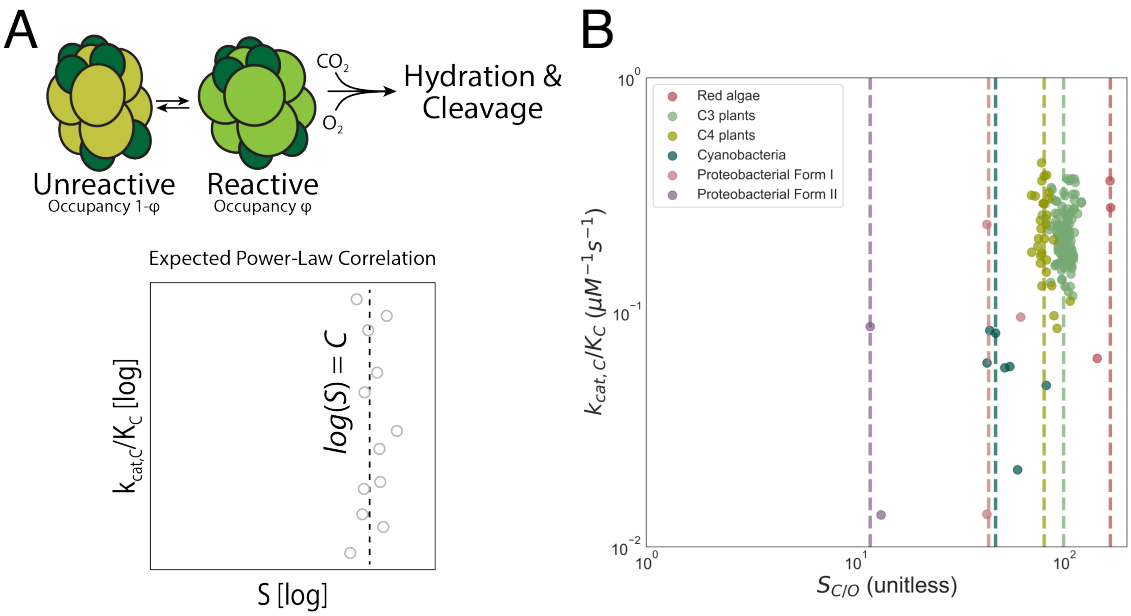


Figure 2.7: A power-law relationship between $k_{cat,C}/K_C$ vs $k_{cat,O}/K_O$ with an exponent of roughly 1.0 can be explained by an active site that fluctuates between "reactive" and "unreactive" states. (A) In this model CO_2 and O_2 can react with the bound RuBP only if the enzyme is in the reactive state. If the difference in intrinsic reactivities of the active site complex ($\Delta G^{*}_{1,O} - \Delta G^{*}_{1,C}$) is organism-independent, we derive a power-law relationship between $k_{cat,C}/K_C$ vs $k_{cat,O}/K_O$ that has an exponent of 1.0 (Appendix A). This model predicts that $S_{C/O}$ is constant. However, $S_{C/O}$ varies 3-4 fold across Form I Rubiscos. (B) Rubiscos within the same physiological grouping - e.g. C_3 or C_4 plants - have roughly constant $S_{C/O}$ independent of $k_{cat,C}/K_C$ (log scale standard deviations $\sigma^* \leq 1.2$ in all cases). The dashed vertical line is drawn at the median $S_{C/O}$ value for each group, with red algal Rubiscos having the highest measured $S_{C/O}$ values ($S_{C/O} \approx 150$) at 25 °C.

This model can be phrased quantitatively as

$$\frac{k_{cat,C}}{K_C} \propto \phi \exp(-\Delta G^{*}_{1,C}/RT)$$
$$\frac{k_{cat,O}}{K_O} \propto \phi \exp(-\Delta G^{*}_{1,O}/RT)$$

where $\Delta G^{*}_{1,C}$ and $\Delta G^{*}_{1,O}$ are the intrinsic reactivities of the enediolate of RuBP to CO_2 and O_2 respectively (Appendix A). Given this model, we expect to observe a power-law relationship with exponent 1.0 between $\frac{k_{cat,C}}{K_C}$ and $\frac{k_{cat,O}}{K_O}$ (Appendix A). Since $S_{C/O} = (k_{cat,C}/K_C)$ / $(k_{cat,O}/K_O)$, specificity should be roughly constant under this model (Appendix A). This model implies that ϕ can vary between related Rubiscos, perhaps by evolutionary tuning of the equilibrium constant for RuBP enolization. $S_{C/O}$ is independent of the equilibrium

fraction of on-enzyme RuBP enolization (K_E), so variation in K_E would affect $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ without altering $S_{C/O}$ (Appendix A). Though individual groups of Rubiscos have roughly constant $S_{C/O}$, specificity clearly varies between C_3 plants and cyanobacteria, for example (Figure 2.7B). Variation in specificity could be achieved by adjusting the difference between intrinsic reactivities $\Delta G^*_{1,O} - \Delta G^*_{1,O}$ through changes to the conformation of the enediolate of RuBP. This would produce roughly constant $S_{C/O}$ among C_3 plants while permitting variation in $S_{C/O}$ between C_3 plants, cyanobacteria and proteobacterial Form I Rubiscos. See Appendix A for a full derivation of this model and discussion of its potential implications.

2.4 Discussion

We collected and analyzed roughly 300 literature measurements of Rubisco kinetic parameters (Figure 2.2A). The collection is quite biased, with the readily-purified Rubiscos of land plants making up ≈80% of the data (Figure 2.2B). Better sampling of Rubisco diversity including more algal, bacterial and archaeal Rubiscos would greatly improve our understanding of the evolution and capacity of this enzyme [167]. Despite incomplete coverage, some trends are clear. All Rubisco kinetic parameters display limited dynamic range, with standard deviations in log-scale being less than one order-of-magnitude in all cases (Figure 2.2C). Rubisco $k_{cat,C}$ appears particularly constrained, displaying much less variability in k_{cat} than any other enzyme for which sufficient data is available (Appendix A). 97% of Rubisco $k_{cat,C}$ values are between 1 and 10 s^{-1}. The highest Rubisco $k_{cat,C}$ measured at 25 °C is 14.4 s^{-1} (*S. elongatus* PCC 7942, [212]), only 18 times the lowest measured Form I $k_{cat,C}$ (0.83 s^{-1} from the diatom *Cylindrotheca* N1, [246]). Altogether, these data suggest that there is some limitation on the maximum carboxylation rate by Rubisco in the presence of O_2.

Other extremal Rubiscos of note include an enzyme from the thermophilic red alga *G. sulphuraria*, which has the lowest $K_C \approx 3$ M. Red algal Rubiscos are generally the most CO_2-specific ($S_{C/O} \approx$ 160-200) measured at 25 °C [306, 321]. Many Rubiscos are quite slow oxygenators with more than half of measurements having $k_{cat,O} < 1$ s^{-1} (Figure 2.2A). Similarly, O_2 affinity is quite low in general: the median K_O is ≈ 470 M, nearly double the Henry's law equilibrium of water with a 21% O_2 atmosphere (≈ 270 μM at 25 °C). Rubisco from the diatom *Thalassiosira weissflogii*, for example, has a $K_O \approx 2$ mM [333], more than five times the ambient O_2 concentration [258].

Specificity $S_{C/O}$ varies the least of all Rubisco parameters (Figure 2.2C). Nonetheless, Form I Rubiscos are much more CO_2-specific than their Form II, III and II/III counterparts (Figure 2.7B, Appendix A). This might be explained by the prevalence of Form II, III and II/III enzymes in bacteria and archaea that fix CO_2 in anaerobic conditions where oxygenation should be negligible. We note, however, that there is substantial variation among measurements of the model Form II Rubisco from *R. rubrum*. This and the general paucity of data on non-Form I Rubiscos indicates that more measurements are required to evaluate stereotyped differences within and between Form II, III and II/III Rubiscos. As

such, we chose to focus here on Form I Rubiscos for which data is abundant (Figure 2.2).

Overall, we observed that Rubisco appears less constrained than previously supposed. Rubisco kinetics were previously argued to vary in a one-dimensional landscape [262] and hypothesized be "nearly perfectly optimized" [301]. If these arguments were true, we would expect very limited variation in Rubisco kinetics because all enzymes should attain near-optimal parameters (as diagrammed in Figure 2.4C). However, the extended dataset is not strictly one-dimensional (Appendix A. Consistent with this analysis, Figure 2.3 and Appendix A document an an overall reduction in correlation between Form I Rubisco kinetic parameters. There are some stereotyped differences between Rubiscos from different kinds of organisms - cyanobacterial Rubiscos are among the fastest (highest $k_{cat,C}$) and red algal Rubiscos are the most CO_2-specific (highest $S_{C/O}$). However, when we plot the assembled data in Figure 2.4A, there is only modest correlation between $k_{cat,C}$ and $S_{C/O}$. Similarly, Figure 2.4B shows relatively modest correlation between $k_{cat,C}$ and K_C. Overall weakened correlations led us to reject the notion that Rubisco kinetics vary in a one-dimensional landscape and to investigate evidence for previously-suggested mechanistic tradeoff models within our extended dataset.

The mechanistic tradeoff models described in Figures 2.5 and 2.6 are based on a simple chemical intuition: that the intrinsic difficulty of binding and discriminating between CO_2 and O_2 requires the enzyme to differentiate between carboxylation and oxygenation transition states. The requirement of TS discrimination is a direct consequence of two common assumptions that are supported by experimental evidence [226]. Briefly, it is assumed that addition of either gas is irreversible and that there is no binding site for CO_2 or O_2 and, thus, no "Michaelis complex" for either gas [302]. If CO_2 bound a specific site on Rubisco before reacting, K_C could be modulated by mutation without substantially affecting the kinetics of subsequent reaction steps. In the unlikely case that gas addition is substantially reversible [71, 302], we would expect to find Rubiscos that evolved enhanced selectivity by energy-coupled kinetic proofreading. Energy coupling would enable amplification of selectivity determined by differential CO_2 and O_2 off-rates [131, 207]. The fact that no such Rubiscos have been found suggests that gas addition is irreversible or that the off-rates of CO_2 and O_2 are incompatible with kinetic proofreading in some other way [261, 262].

As Rubisco likely does not bind CO_2 directly, Tcherkez et al. 2006 hypothesize that high specificity towards CO_2 (i.e. large $S_{C/O}$) is realized by discriminating between the first carboxylation and oxygenation transition states (i.e. between the carboxyketone and the peroxyketone, Appendix A). A late carboxylation TS would be maximally discriminable from the oxygenation TS because the developing carboxylic acid is distinguishable from the peroxyl group of the oxygenation intermediate [301]. Since a late TS resembles the carboxyketone carboxylation intermediate, Tcherkez et al. further argue that CO_2-specific Rubiscos must tightly bind the carboxyketone, which could slow the subsequent hydration and cleavage steps and restrict $k_{cat,C}$. Though this model is motivated by the need for discrimination between CO_2 and O_2, it implies coupling between the kinetics of carboxylation steps alone. That is: specificity requires tighter binding of the carboxylation intermediate, which slows downstream processing of that same intermediate irrespective of the oxygenation

steps. The extraordinarily tight binding of the carboxyketone analog CABP to plant Rubisco provides strong support for the idea of TS discrimination. Negative correlation between $k_{\mathrm{cat,C}}$ and $S_{\mathrm{C/O}}$ was previously taken to support the idea of tighter TS binding slowing carboxylation [301].

Savir et al. 2010 articulate a related model, noting that $k_{\mathrm{cat,C}}$ and $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ are inversely correlated in their dataset [262]. Since $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ is related to the effective barrier to enolization and CO_2 addition and $k_{\mathrm{cat,C}}$ is related to the effective barrier to hydration and cleavage (Figure 2.4B), Savir et al. argue that lowering the effective barrier to CO_2 addition entails a higher barrier for the subsequent steps (i.e. a lower $k_{\mathrm{cat,C}}$, Figure 2.5A). In both of these descriptions, the initial steps of carboxylation are negatively coupled to the subsequent steps in a manner that produces the observed correlations. However, those correlations - between $S_{\mathrm{C/O}}$ and $k_{\mathrm{cat,C}}$, K_{C} and $k_{\mathrm{cat,C}}$ and $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,C}}$- are attenuated by the addition of new measurements (Figures 2.3 and 2.4) which calls these proposals into question. Importantly, we do not argue that the chemical logic advanced by Tcherkez et al. 2006 is incorrect, but rather that the assembled data do not support such a tradeoff being optimized over the evolution of Form I Rubiscos.

The second tradeoff model posited by Savir et al. 2010 is that faster CO_2 addition to the Rubisco-RuBP complex necessarily allows faster O_2 addition. This model can be motivated by the catalytic mechanism of Rubisco as well. Since Rubisco likely does not bind CO_2 or O_2 directly, the concentrations of CO_2 and O_2 in the Rubisco active site should be determined by their solution concentrations alone (e.g. in the chloroplast stroma). Rubisco might limit the active site concentration of O_2 by closing the active site to diffusion, but this would slow CO_2 entry and carboxylation as well. This model was previously supported by a positive power-law correlation between the catalytic efficiencies for carboxylation and oxygenation ($k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ respectively), which can be understood as a positive coupling of the effective barriers to enolization and gas addition for both CO_2 and O_2 (Figure 2.6A, Appendix A). We showed that the extended dataset strongly supports this power-law relation and suggests lowering the effective CO_2 addition barrier (enabling faster carboxylation) requires a *roughly equal* reduction in the effective barrier to O_2 addition (i.e. enabling faster oxygenation as well). Though several research groups have attempted to isolate improved Rubisco mutants, none of the mutants examined so far exceed the wild-type enzymes on these axes (Appendix A).

A power law relation with an exponent of 1.0 can be seen as resulting from an active site that fluctuates between a reactive and unreactive state (Figure 2.7A). In this model, the average occupancy of the reactive state dictates the rate of CO_2 and O_2 addition and throttles the subsequent steps of carboxylation and oxygenation equally (Figure 2.7). This model can be mapped onto the Rubisco mechanism by noting that RuBP must be enolized before CO_2 or O_2 can react, suggesting that the occupancy of the reactive state (ϕ) is related to the degree of enolization of RuBP (Appendix A). One implication of this model is that $S_{\mathrm{C/O}}$ is roughly constant (Appendix A). While $S_{\mathrm{C/O}}$ does vary over roughly tenfold across the entire dataset and 3-4 fold across Form I enzymes (Figure 2.4B), Rubiscos from the same physiological groupings display roughly constant $S_{\mathrm{C/O}}$ values independent of $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ (Figure 2.7B).

More measurements of bacterial Form I, II and III Rubiscos as well as the notably high-specificity Form ID enzymes of red algae will be crucial to evaluate the generality of this observation.

In previous work, where Rubisco kinetics were thought to vary in a one-dimensional landscape, setting $k_{\mathrm{cat,C}}$ determined all other kinetic parameters [262]. In this setting it was argued that Rubisco kinetic parameters were wholly determined by the prevailing CO_2 and O_2 concentrations since there was a unique choice of parameters on the one-dimensional curve that maximize the net rate of carboxylation [262]. Since the data is no longer clearly one-dimensional, we cannot argue that Rubisco is "perfectly optimized" to match prevailing concentrations. Moreover, the single surviving tradeoff model does not, on its own, explain why we have not found faster-carboxylating Rubiscos. The model presented in Figures 2.6 and 2.7 describes a tradeoff between CO_2 and O_2 addition, but sets no upper limit on $k_{\mathrm{cat,C}}$, suggesting that selection for increased carboxylation in the absence of O_2 could produce Rubiscos with superlative $k_{\mathrm{cat,C}}$ values (i.e. $k_{\mathrm{cat,C}} \gg 15\ \mathrm{s}^{-1}$).

The prospect of engineering an improved Rubiscos is tantalizing not only because it could plausibly increase crop yields substantially [337], but also because the task tests our understanding of proteins and enzymes on a very basic level. It is clear from the data presented here that there is some evolutionary constraint on Rubisco catalysis. Indeed, no known Rubisco has a $k_{\mathrm{cat,C}}$ greater than 15 s^{-1} at 25 °C and no measured $S_{\mathrm{C/O}}$ exceeds 250. Surely a superlative Rubisco would have arisen if it was mutationally accessible from existing enzymes. However, the Rubisco large subunit displays extremely limited sequence variation [143]. Perhaps exploring a wider swath of sequence space via protein engineering techniques [63, 105, 283] would enable strict improvements to Rubisco kinetics? In order to better-resolve the evolutionary constraints imposed on Rubisco kinetics and evaluate the prospects of future Rubisco engineering, we suggest several avenues of research.

First, the kinetics of non-plant Rubiscos should be characterized more thoroughly. These should include the Form II, III and II/III enzymes of bacteria and archaea as well as Form I enzymes of cyanobacteria and diverse eukaryotic autotrophs [167]. Ideally these enzymes would be sampled from accumulated genomic data in a manner that maximizes sequence and phylogenetic diversity [6] and characterized for their binding (e.g. of RuBP and CABP) and catalytic activity (measuring $k_{\mathrm{cat,C}}$, K_{C}, $\mathrm{k_{cat,O}}$, K_{O} and $S_{\mathrm{C/O}}$) as a function of temperature and pH [214, 272]. These data would likely resolve whether Rubisco isoforms display characteristic differences in catalytic potential. It is possible, for example, that Form II, III or II/III enzymes are subject to different constraints than Form I Rubiscos and might serve as useful chassis for bioengineering.

Furthermore, it is important to revisit the classic experiments undergirding our understanding of the Rubisco catalytic mechanism, especially those supporting the central assumptions that (a) there is no Michaelis complex for CO_2 or O_2 and (b) that gas addition is irreversible [71, 226, 300]. As mentioned above, these assumptions imply substantial limitation on CO_2 specificity by, for example, disallowing a kinetic proofreading based mechanism for the amplification of specificity. If we were to find Rubiscos for which these assumptions are relaxed, they might be used as a basis for future engineering of a fast-and-selective

carboxylase. On the other hand, it may be the case that all Rubiscos share these same limitations and are constrained by the same tradeoffs. Since tradeoffs in Rubisco catalysis are likely described by couplings between transition state barriers (e.g. as in Figure 2.6) it would be very useful to measure TS barrier heights for many variants. One avenue for further investigation would be measurement of carbon and oxygen kinetic isotope effects (KIEs) for a wide variety of Rubiscos. Kinetic isotope effects report indirectly on TS barrier heights [120, 188] and KIEs could plausibly be measured in relatively high throughput via mass spectrometry. Investigating the relationship between kinetic isotope effects and kinetic parameters will hopefully refine our understanding of the Rubisco mechanism and help clarify whether different families of Rubisco enzymes are subject to the same constraints [301].

There remains some disagreement about the precise ordering of Rubisco carboxylation steps [10, 66, 301] and the mechanism of oxygenation is not well understood [300]. Chemical reasoning about the mechanisms of Rubisco carboxylation and oxygenation would benefit from progress in structural biology - intermediates and transition state analogs should be used to capture the active site at various points along the reaction trajectory [11]. If experiments and structural analyses confirm that the above assumptions hold for all Rubiscos, it would greatly limit our capacity to engineer Rubisco and strongly suggest that alternative strategies for improving carbon fixation should be pursued [24, 187, 290]. If, however, these assumptions are invalidated, many enzyme engineering strategies would be viable. Such data and analyses will be instrumental in guiding the engineering of carbon fixation for the next decade.

2.5 Methods

Data Collection and Curation

We reviewed the literature to find Rubisco kinetic data measured at 25 °C and near pH 8. Ultimately 61 primary literature studies were included, yielding 334 $S_{\mathrm{C/O}}$, 282 $k_{\mathrm{cat,C}}$, 316 K_{C}, and 254 K_{O} values for Rubiscos from 304 distinct organisms (Datasets S1 and S2). We also recorded 52 measurements of the Michaelis constant for RuBP (K_{RuBP}). Experimental error was recorded for all of these values (when reported) along with the pH, temperature and other metadata. Data was filtered as described in the Appendix A. $k_{\mathrm{cat,O}}$ is usually not measured directly, but is rather inferred as $k_{\mathrm{cat,O}} = (k_{\mathrm{cat,C}}/K_{\mathrm{C}})\ /\ (S_{\mathrm{C/O}}/K_{\mathrm{O}})$. We assumed that experimental error is normally distributed and used 10^4-fold bootstrapping to estimate 198 $k_{\mathrm{cat,O}}$ values and 95% confidence intervals thereof. We used an identical procedure to estimate $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ and confidence intervals thereof (Appendix A). Altogether, we were able to calculate 274 $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and 199 $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ values. Datasets S1 and S2 provide all source and inferred data respectively.

Fitting Power Laws

Certain model Rubiscos are measured frequently. For example, we found 12 independent measurements of the Rubisco from spinach. We used the median measured value in correlation and regression analyses to avoid bias. In contrast to textbook examples with one independent and one dependent variable, there is experimental error associated with both variables in all scatter plots shown here (e.g. plotting $k_{\mathrm{cat,C}}$ against K_{C} in Figure 2.4B). As such we used total least squares linear regression in log scale to fit relationships between Rubisco parameters. Because R^2 values of total least squares fits do not convey the explained fraction of Y axis variance, they are challenging to interpret. We instead report the degree of correlation as Pearson R values of log-transformed values. Bootstrapping was used to determine 95% confidence intervals for the Pearson correlation coefficient, power-law exponents and prefactors (i.e. the slopes and intercepts of linear fits in log-log scale). In each iteration of the bootstrap, data were subsampled to 90% with replacement. Total least squares regression was applied to each subsample to determine a point estimate of R, the power-law exponent and prefactor. This procedure was repeated 10^4 times to determine a 95% confidence intervals on the above parameters. Python source code is available at github.com/flamholz/rubisco.

Acknowledgements

We would like to thank Uri Alon, Kapil Amarnath, Doug Banda, Arren Bar-Even, Cissi Blikstad, Jack Desmarais, Woodward Fischer, Vahe Galstyan, Laura Helen Gunn, Itai Halevy, Robert Nichols, Elad Noor, Jeremy Roop, Yonatan Savir, Patrick Shih, Daniel Stolper, Dan Tawfik, Guillaume Tcherkez, Tsvi Tlusty and Renee Wang for helpful conversations and comments on the manuscript.

Chapter 3

A Physiologically-Plausible Model of the Bacterial CCM

Adapted from *Proceedings of the National Academy of Sciences of the United States of America*, Volume 113, Mangan NM, Flamholz A, Hood RD, Milo RM, and DF Savage, Pages E5354-62, Copyright 2016, with permission from *PNAS*.

3.1 Abstract

Many carbon-fixing bacteria rely on a CO_2 concentrating mechanism (CCM) to elevate the CO_2 concentration around the carboxylating enzyme Rubisco. The CCM is postulated to simultaneously enhance the rate of carboxylation and minimize oxygenation, a competitive reaction with O_2 also catalyzed by Rubisco. To achieve this effect, the CCM combines two features: active transport of inorganic carbon into the cell and co-localization of carbonic anhydrase and Rubisco inside proteinaceous microcompartments called carboxysomes. Understanding the significance of the various CCM components requires reconciling biochemical intuition with a quantitative description of the system. To this end, we have developed a mathematical model of the CCM in order to analyze its energetic costs and the inherent intertwining of physiology and pH. We find that intracellular pH greatly affects the cost of inorganic carbon accumulation. At low pH the inorganic carbon pool contains more of the highly cell-permeable H_2CO_3, necessitating a substantial expenditure of energy on transport to maintain internal inorganic carbon levels. An intracellular pH ≈ 8 reduces leakage, making the CCM orders of magnitude more energetically efficient. This pH prediction coincides well with our measurement of intracellular pH in a model cyanobacterium. We also demonstrate that CO_2 retention in the carboxysome is necessary, while selective uptake of HCO_3^- into the carboxysome would not appreciably enhance energetic efficiency. Altogether, integration of pH produces a model that is quantitatively consistent with cyanobacterial physiology, emphasizing that pH cannot be neglected when describing biological systems interacting with inorganic carbon pools.

3.2 Significance

Cyanobacteria are responsible for roughly 10% of global photosynthetic primary production of reduced carbon. Though cyanobacteria are incredibly diverse, all known species contain a complex protein system called the CO_2 concentrating mechanism (CCM), which enables rapid growth even in environments with extremely limited CO_2. The CCM enables cyanobacteria to accumulate HCO_3^- and convert this inorganic carbon pool to utilizable CO_2. We demonstrate here that a quantitative description of the CCM must include the effect of pH on the abundance of HCO_3^- and H_2CO_3. This pH-dependent description is consistent with cyanobacterial physiology. Furthermore, the model predicts that alkaline cytosolic pH reduces the energetic cost of the CCM, consistent with pH measurements photosynthesizing cyanobacteria.

3.3 Introduction

Cyanobacteria and many other autotrophs employ a CO_2 concentrating mechanism (CCM) to increase the cellular pool of inorganic carbon and facilitate the Calvin-Benson-Bassham (CBB) cycle [243]. Specifically, the CCM functions to supply CO_2 to ribulose bisphosphate carboxylase/oxygenase (Rubisco), the primary carboxylating enzyme of the CBB cycle. High levels of CO_2 are essential to cyanobacterial metabolism because Rubisco has relatively slow carboxylation kinetics and is promiscuous, catalyzing an off-pathway reaction with O_2 called oxygenation [66, 262, 301].

Rubisco oxygenation produces 2-phosphoglycolate (2PG), which is not part of the CBB cycle and must be recycled. Recycling 2PG through photorespiratory pathways is costly, consuming reduced carbon and energy resources [33, 213]. The problem of Rubisco's limited specificity is all the more pronounced because there is $\approx$ 20 times more O_2 than CO_2 in aqueous solutions equilibrated with present day atmosphere (Appendix B,. CCMs overcome these problems by concentrating CO_2 near Rubisco, favorably increasing the ratio of CO_2 to O_2. High concentrations of CO_2 maximize the rate of carboxylation and competitively inhibit oxygenation. Indeed, it is widely thought that the evolution of CCMs served to ameliorate energetic costs associated with large photorespiratory fluxes in the present day atmosphere [243].

Based on diverse experimental studies, a convincing model of the bacterial CCM has emerged (Figure 3.1) wherein the CCM has two primary components: active accumulation of inorganic carbon (C_i) in the cytosol and organization of Rubisco with carbonic anhydrase (CA) inside proteinaceous organelles called carboxysomes [238]. Perturbations to either component, uptake systems or carboxysomes, disrupt the CCM and produce mutants that require elevated CO_2 for growth [81, 228]. Two energetically activated transport mechanisms —HCO_3^- transport and facilitated uptake of CO_2 —enable accumulation of bicarbonate (HCO_3^-) in the cytosol [238]. HCO_3^- transport is generally coupled to Na^+ gradients or ATP hydrolysis and facilitated uptake is hypothesized to couple oxidation of NAD(P)H

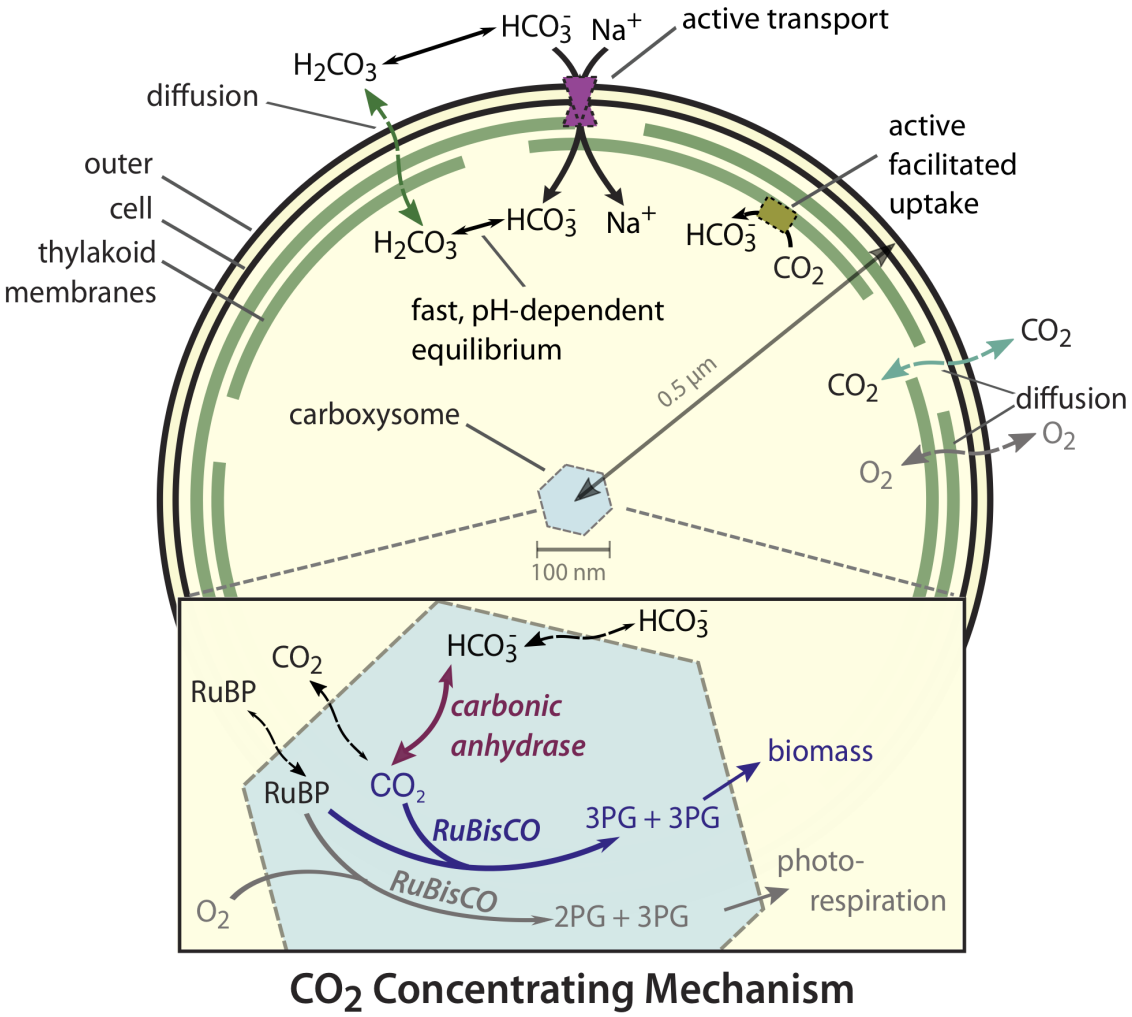


Figure 3.1: Schematic of the generally accepted model of the cyanobacterial CO_2 concentrating mechanism. The cyanobacterial CCM concentrates HCO_3^- in the cytosol via two classes of uptake systems: transporters and facilitated uptake systems. We focus on Na^+-activated transporters, which transport one Na^+ with each HCO_3^-. Facilitated uptake of CO_2 is catalyzed by proteins on the thylakoid membrane and is thought to couple NAD(P)H oxidation to the vectorial conversion of CO_2 into HCO_3^-. Roughly 250 Rubiscos (2000 active sites) and 100 carbonic anhydrases are localized to the carboxysome. HCO_3^- concentrated in the cytosol enters the carboxysome where it is converted to a high CO_2 concentration by carbonic anhydrase activity. This elevated CO_2 concentration increases the rate of Rubisco carboxylation and competitively inhibits oxygenation, increasing the overall efficiency of carbon fixation. For parameter values and model details see Appendix B. References for each model component are given in the main text. Abbreviations: RuBP (ribulose-1,5-bisphosphate); 3PG (3-phosphoglycerate); 2PG (2-phosphoglycolate).

to unidirectional hydration of CO_2 (Figure 3.1) [238]. Accumulation of charged HCO_3^- is preferable to accumulation of CO_2 because HCO_3^- escapes much less readily through the cell membrane, as we discuss below. Together, these uptake systems generate a cytosolic

HCO_3^- concentration > 10 mM, which is ≈ 30 times the equilibrium concentration of HCO_3^- in water at neutral pH at 25 °C (Appendix B) [320].

Carboxysomes are icosahedral compartments approximately 100 *nm* in diameter and composed of a protein shell surrounding an enzyme-filled lumen [330]. The carboxysome lumen is densely packed (> 400 mg protein/ml) with about 2000 Rubisco and 100 CA active sites [330]. As CA activity is absent from the cytosol and the spontaneous dehydration of HCO_3^- to CO_2 is relatively slow [181], HCO_3^- does not equilibrate with CO_2 in the cytosol [228]. Rather, HCO_3^- enters the carboxysome where carbonic anhydrase activity readily equilibrates it with CO_2 [228]. Crucially, the carboxysome shell must slow the diffusive loss of CO_2 so that CA activity produces a locally elevated carboxysomal CO_2 concentration [81, 132]. At sufficiently high CO_2 concentrations, Rubisco is saturated, oxygenation is inhibited, and carboxylation proceeds at the maximum rate [243].

In order to define the regimes wherein the cyanobacterial CCM leads to efficient carbon fixation, we previously developed a reaction-diffusion model of the CCM [182]. Even when presented with very low extracellular CO_2 concentrations (i.e. very high relative O_2 concentrations), the modeled CCM can saturate Rubisco and drastically reduce oxygenation [182]. Surprisingly, the model suggested that the CCM does not require selective uptake of HCO_3^- into the carboxysome, selective retention of CO_2 inside the compartment, or exclusion of O_2. Rather, a non-specific permeability barrier at the carboxysome shell could give rise to CO_2-concentrating activity. These results built upon previous models, [107, 249, 248] and further explained how CCM activity, which makes cyanobacterial growth largely insensitive to the environmental concentration of inorganic carbon, could arise from its known protein components.

An unexplained observation of our previous work was substantial leakage of HCO_3^- across the cell membrane, with > 99% of carbon import leaking out of the cell as HCO_3^- [182]. As HCO_3^- accumulation is energetically driven, large leakage fluxes consume substantial energy and require an implausible fraction (> 50%) of the membrane surface for transporters (Appendix B). Here, we identify the cell membrane permeability to HCO_3^- as a key parameter determining leakage. All previous models of the CCM tracked only HCO_3^-, implicitly ignoring other species of hydrated inorganic carbon (H_2CO_3, HCO_3^- and CO_3^{2-}, which we collectively term H_{total}) and their contribution to cell permeability [107, 132, 182, 249, 248]. This implicit assumption is at odds with biochemical intuition that charge is a major determinant of membrane permeability [251]. Indeed, the often-cited permeability coefficient of $\approx 3 \times 10^{-4}$ cm/s is representative of the uncharged H_2CO_3 but 3-4 orders of magnitude too high for HCO_3^- and CO_3^{2-} (Appendix B, [116, 251]). A recent model of the minimal CCM of *Prochlorococcus* MED4 made the inverse assumption: that the cell membrane is negligibly permeable to HCO_3^-, implicitly ignoring the rapid interconversion of HCO_3^- with the uncharged and highly-permeable H_2CO_3 [132].

In order to reconcile these varying treatments and conduct a careful accounting of inorganic carbon species, we take advantage of pH as a key physiological parameter governing the composition of the inorganic carbon pool. The relative concentration of species in H_{total} (H_2CO_3, HCO_3^- and CO_3^{2-}) depends strongly on pH (Appendix B). As these species differ

in their net charge, pH will influence the rate of inorganic carbon leakage from the cell and the energetic costs associated with carbon accumulation. Since Rubisco and CA activities are pH-dependent, the pH will also affect enzymatic rates inside the carboxysome. We have integrated these effects into our previous analytical and numerical models of the CCM [182] to produce a 'pH-aware' model of the CCM, which is described in mathematical detail in the Appendix B.

The pH-aware model is now consistent with cyanobacterial physiology in a number of important ways. In optimal conditions, the model produces absolute fluxes that are similar to measured values and consistent with cyanobacterial growth rates. Unlike in previous models, these fluxes can be supported by transporters occupying a very small fraction of membrane surface area ($< 1\%$) leaving space for the essential biochemistry of transport, photosynthesis and chemiosmosis (Appendix B). Finally, characteristic differences in the inorganic carbon transport modalities used by oceanic and freshwater cyanobacteria can be explained through the pH-aware model by the characteristic pH and salinity differences between ocean and freshwater (Appendix B). Given this broad consistency with cyanobacterial physiology and genetics, the pH-aware model can also be used to examine open questions related to the CCM.

Our updated, pH-aware model also enables us to test the hypothesis that the CCM reduces the energetic costs associated with fixing carbon. We calculate the effect of pH on the energetic cost of carbon fixation, carbon concentration, and photorespiration. According to our calculation, the CCM requires considerably less energy than implied by our previous model (Appendix B). We further demonstrate that a selective carboxysome that prefers to take up HCO_3^- and retain CO_2 is not required to produce an energy-efficient CCM. Rather, a low absolute permeability to CO_2 is the crucial characteristic of an energy efficient carboxysome. Moreover, the pH-aware model predicts an intracellular pH range that optimizes CCM efficiency in cyanobacteria actively fixing carbon. We validate this prediction by measuring the intracellular pH in live cyanobacterial cells (*Synechococcus elongatus* PCC 7942). Thus, the CCM offers another example where the properties of complex systems central to bacterial growth are well-explained by the principle of energetic cost minimization [29, 103, 269], in this case explaining the remodeling of cytosolic pH and transport modalities to minimize the cost of inorganic carbon accumulation for the CCM.

3.4 Results

The Effect of pH on the Permeability of Inorganic Carbon to the Cell Membrane

All cells regulate their cytosolic pH due to the inherent pH-dependence of biochemical reactions [7]. However, cytosolic pH varies substantially between organisms and growth conditions [186, 308]. Here, we use a mathematical model to examine how the cyanobacterial CCM functions over a range of cytosolic pH values. We include the effect of pH on the

permeability of the cell membrane to constituents of the inorganic carbon pool and on the enzymatic activities located within the carboxysome.

The equilibrium composition of the C_i pool (CO_2, H_2CO_3, HCO_3^- and CO_3^{2-}) is highly pH dependent (Appendix B). Whether these species reach equilibrium with each other, however, depends on how the rate of uncatalyzed interconversion compares to the rates of other processes —e.g. transport and enzymatic catalysis —that produce and consume specific species. The spontaneous dehydration of H_{total} to CO_2 ($T_{1/2} > 10$ s, [181]) is much slower than diffusion and transport, and so the CCM can maintain H_{total} out of equilibrium with CO_2 in the cytosol [228]. Equilibration between HCO_3^-, H_2CO_3 and CO_3^{2-} within the H_{total} pool is, on the other hand, extremely fast ($T_{1/2} < 1\ \mu s$, [2]). As such, the dominant species among HCO_3^-, H_2CO_3 and CO_3^{2-} will be determined by the pH, and so the contribution of each species to the cell permeability of H_{total} must be examined individually.

Due to the energetic penalty associated with the passage of charge into the membrane, small, charged molecules typically have membrane permeability coefficients 10^4-10^5 times smaller than uncharged molecules of comparable size [251]. Consequently, HCO_3^- and CO_3^{2-} are dramatically less cell-permeable than the uncharged H_2CO_3. In addition to having a low membrane-permeability, CO_3^{2-} contributes negligibly ($< 5\%$) to the H_{total} pool below pH 9 and can be neglected (Appendix B). As such, H_2CO_3 can be treated as a monoprotic acid equilibrating quickly with HCO_3^-. The primary literature on the permeability of small molecules gives permeability coefficients of $\approx 10^{-3}\frac{cm}{s}$ for H_2CO_3 and $\approx 10^{-7}\frac{cm}{s}$ for HCO_3^- [116, 329].

Given these values, we assume $k_m^{H_2CO_3} \ll k_m^{HCO_3^-}$, and derive that the rate of H_{total} diffusion across the membrane is dominated by the diffusion of uncharged H_2CO_3,

diffusive H_{total} flux $= k_m^{H_2CO_3}\Delta H_2CO_3 + k_m^{HCO_3^-}\Delta HCO_3^- \approx k_m^{H_2CO_3}\Delta H_2CO_3$.

Here, ΔH_2CO_3 and ΔHCO_3^- are the concentration differences of H_2CO_3 and HCO_3^- across the cell membrane (see Appendix B for derivation). To a first approximation, H_2CO_3 is the only species of H_{total} that will cross the cell membrane diffusively at an appreciable rate, regardless of the relative concentrations of H_2CO_3 and HCO_3^-.

The above equation describes the diffusion of H_{total} into the cell as a function of the H_2CO_3 concentration gradient. However, the CCM model tracks HCO_3^- and not H_2CO_3 because HCO_3^- is the substrate of CA. In order to integrate this equation into the CCM model, therefore, we assume fast equilibrium of the H_{total} pool and calculate the equilibrium ratio $\frac{[HCO_3^-]}{[H_2CO_3]} = 10^{pH-pK_1}$ in order to determine ΔH_2CO_3 across the cell membrane. As the concentration and composition of H_{total} may differ markedly across the cell membrane (due to differences in pH, ionic strength and the action of C_i transporters), this substitution yields separate terms for the inward (first term) and outward (second term) diffusional velocities of H_{total}

$$\text{diffusive } HCO_3^- \text{ flux} = \left[HCO_3^-\right]_{\text{out}} \times k_m^{H_2CO_3} \times 10^{(pK_1 - pH_{\text{out}})} - \left[HCO_3^-\right]_{\text{cytosol}} \times k_m^{H_2CO_3} \times 10^{\left(pK_1 - pH_{\text{cytosol}}\right)}$$

Here, $k_m^{H_2CO_3} = 3 \times 10^{-3}$ cm/s is the velocity of H_2CO_3 permeation and $10^{(pK_1 - pH)}$ is the ratio of H_2CO_3 to HCO_3^- as a function of the pH and the first pK_a of H_2CO_3 ($pK_1 \approx 3.2$) [313, 329]. This equation holds across the entire pH range considered here ($6.5 \leq pH \leq 9$, see Appendix B for full derivation). If the cytosolic and extracellular pH are equal, the pH-dependent velocity of H_{total} membrane permeation $k_m^{H_{\text{total}}} \approx k_m^{H_2CO_3}\ 10^{(pK_1 - pH)}$, recovering a functional form equivalent to the previous model. At pH 7, this equation yields a velocity of $\approx 5 \times 10^{-7}$ cm/s, 1000-fold smaller than the commonly-used value of 3×10^{-4} cm/s [232, 230]. Indeed, this higher effective velocity implies an implausible cytosolic pH of about 4 (Appendix B).

Functional Form of the pH-aware CCM model

The pH-aware CCM model is a system of coupled reaction-diffusion differential equations in spherical coordinates. The equations describe the entry of C_i into the cell and the carboxysome as well as diffusion and the chemical reactions within the carboxysome. This system can be solved both numerically and analytically at steady state [182]. We write the fluxes for CO_2, C, and HCO_3^-, H, at the cell membrane, $r = R_b$, as:

$$\begin{aligned} D\frac{\partial C}{\partial r} &= -\ \alpha C_{\text{cytosol}} + k_m^C \left(C_{\text{out}} - C_{\text{cytosol}}\right) \\ D\frac{\partial H}{\partial r} &= j_c H_{\text{out}} +\ \alpha C_{\text{cytosol}} + \\ &\quad \left(\left[HCO_3^-\right]_{\text{out}}\ k_m^{H_2CO_3}\ 10^{(pK_1 - pH_{\text{out}})} - \left[HCO_3^-\right]_{\text{cytosol}}\ k_m^{H_2CO_3}\ 10^{\left(pK_1 - pH_{\text{cytosol}}\right)}\right) \end{aligned}$$

Here D is the diffusion constant for small molecules in water (10^{-5} cm^2/s); j_c is the velocity of active HCO_3^- transport; α is the velocity of CO_2 to HCO_3^- conversion; and k_m^C and $k_m^{H_2CO_3}$ are the permeability of the cell membrane to CO_2 and H_2CO_3 respectively. k_m^C is set to 0.3 cm/s while $k_m^{H_2CO_3} = 3 \times 10^{-3}$cm/s as described above.

As discussed above, the spontaneous dehydration of HCO_3^- is negligibly slow (Appendix B) and so our model cell has no relevant chemical reactions in its cytosol (Figure 3.1). Therefore, diffusion sets the steady-state concentrations of CO_2 and HCO_3^- in the cytosol via $\nabla^2 C = 0$, $\nabla^2 H = 0$, where ∇^2 is the second derivative in spherical coordinates. Assuming the same carboxysome permeability, k_c, for both CO_2 and HCO_3^- at the carboxysome shell, diffusive leakage at the carboxysome shell is expressed as:

$$D\frac{\partial C}{\partial r} = k_c \left(C_{\text{cytosol}} - C_{\text{carboxysome}}\right)$$

$$D\frac{\partial H}{\partial r} = k_c(H_{\text{cytosol}} - H_{\text{carboxysome}})$$

at r = R_c , the boundary of the carboxysome. The pH-aware model can also be used to consider differential carboxysome permeability to CO_2 and HCO_3^-, as described in the Discussion and Appendix B. Inside the carboxysome, the model considers diffusion of substrates as well as carbonic anhydrase (R_{CA}) and Rubisco (R_{Rub}) activities. The balance of enzymatic and diffusive rates set the steady-state concentrations of CO_2 and HCO_3^- in the carboxysome as

$$D\nabla^2 C + R_{\text{CA}} - R_{\text{Rub}} = 0$$

$$D\nabla^2 H - R_{\text{CA}} = 0$$

where the enzymatic rates R_{CA} and R_{Rub} are depend on pH in the manner described below.

The Effect of pH on Enzymatic Activity in the Carboxysome

The enzymatic mechanisms of Rubisco and CA are inherently pH dependent [18], and so it is critical to include the pH dependence of these enzymes in describing the pH dependence of the CCM. The carboxysomal CA is an efficient enzyme, well described by reversible Michaelis-Menten kinetics, with a dehydration $k_{\text{cat}} \approx 5\times10^5\ s^{-1}$ [122]. When the intra-shellular CA is saturated, increased HCO_3^- transport has no effect on the carboxysomal CO_2 concentration. As a result, saturated CA implies that HCO_3^- uptake could be lower without affecting the carboxylation rate [182]. Therefore, efficient CCM function occurs, in part, when CA is not saturated. As CA is a fast enzyme with ≈ 100 active sites in the carboxysome, CO_2 and HCO_3^- are held in equilibrium within the carboxysome when CA is not saturated (Appendix B). In this pseudo-equilibrium condition, the ratio of CO_2 to HCO_3^- in the carboxysome is determined by the pH dependence of the reaction equilibrium constant $K'_{\text{eq}}(\text{pH}) = \frac{[HCO_3^-]}{[CO_2]}$. We calculate the pH dependence of K'_{eq} from the Gibbs formation energies of the relevant species [7, 211]. For reversible reactions like the dehydration of bicarbonate, a modest change in the pH can have a large effect on K'_{eq} —about tenfold between pH 7 and 8 (Appendix B).

Rubisco activity also varies with pH, with the carboxylation k_{cat} reaching maximum near pH 7.5 [18]. We describe the Rubisco reaction with irreversible Michaelis-Menten kinetics and assume Rubisco kinetic parameters measured at pH 7.8 for the *S. elongatus* PCC 6301 Rubisco. The pH dependence of Rubisco kinetics was modeled by rescaling the Michaelis-Menten kinetic constants by the pH dependence observed for the cyanobacterial Rubisco from *Anabaena variabilis* (Appendix B).

A pH-aware CCM Model

With the pH-dependence of permeability and enzymatic activity integrated into the model, we asked a simple question: what is the role of cytosolic pH in shaping the flow of carbon through the CCM? To answer this question, we calculated the HCO_3^- transport flux required to achieve the measured ≈ 30 mM cytosolic HCO_3^- concentration [295, 320, 327] across a range of cytosolic pH values (Figure 3.2). For this calculation, the carboxysome permeability was set to 3x10^{-5} cm/s (optimum for cytosolic pH 8, Appendix B) and facilitated CO_2 uptake was set to zero for simplicity. The carboxysome permeability represents the velocity at which CO_2 and HCO_3^- traverse the carboxysome shell, including both transit of pores found in shell proteins and any "leakiness" of the shell itself. Carbon is conserved and so fluxes importing inorganic carbon into the cell (active HCO_3^- transport) must equal the sum of fluxes consuming inorganic carbon within the cell: H_{total} leakage, CO_2 leakage, and carboxylation. This conservation law is captured by the equation shown in Figure 3.2.

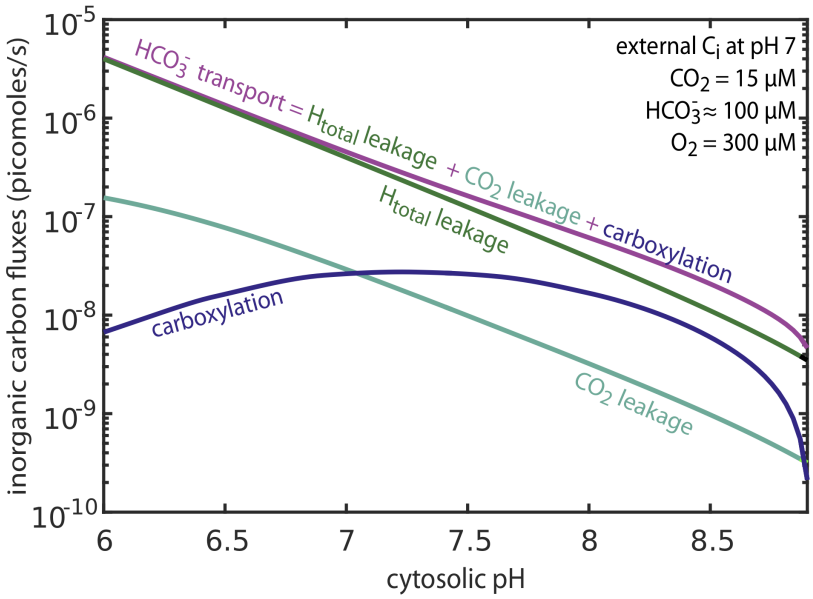


Figure 3.2: Alkaline cytosolic pH reduces C_i fluxes required to achieve efficient CO_2 fixation. Inorganic carbon fluxes are plotted as a function of pH. As carbon is conserved, the HCO_3^- transport flux (purple) equals the sum of H_{total} leakage (green), CO_2 leakage (teal), and carboxylation (blue). Rubisco achieves a maximum carboxylation rate near pH 7.5. Carboxysome permeability was set to 3×10^{-5} cm/s and HCO_3^- transport rate was set to yield 30 mM cytosolic HCO_3^- at each pH. The extracellular C_i pool was assumed to be in equilibrium with the external pH, which was fixed at 7 for this analysis.

Analyzing the pH dependence of carbon fluxes shows that CCM performance is expected to improve dramatically with increasing cytosolic pH (Figure 3.2). Indeed, the amount of active transport required to maintain 30 mM cytosolic HCO_3^- decreases exponentially with increasing pH (i.e. linearly on a log scale). At pH 7 and below, more than 85% of the HCO_3^- influx leaks out of the cell. Above pH 7, CO_2 leakage and carboxylation fluxes contribute more prominently to the flux balance with leakage representing ≈ 65% at pH 8. The log-linear relationship between pH and H_{total} leakage arises from the exponential (power-law) dependence of H_{total} permeability on pH described above. HCO_3^- transport fluxes also display a log-linear relationship with cytosolic pH (Figure 3.2) and absolute transport rates at pH

8 are on the order of $10^{-7}\frac{\text{pmol}}{cell\times s}$, approaching the $\approx 10^{-8}\frac{\text{pmol}}{cell\times s}$ uptake flux measured in the marine cyanobacterium *Prochlorococcus* MED4 [132].

In Figure 3.2, the carboxylation rate declines above pH 7.5 because Rubisco achieves maximal catalytic rate at that pH [18]. At a cytosolic pH of 8 $\approx$ 30% of C_i uptake is ultimately fixed by Rubisco, as compared to $\approx$ 1% in the previous model. Economic operation of the CCM is strongly dependent on the cytosolic pH, with pH $\cong$ 8 curtailing diffusive loss of H_{total}. Although greatly reduced, loss of HCO_3^- (through H_2CO_3 leakage) remains a substantial flux and cannot be neglected even above pH 8. A full description of the model, including all kinetic parameters and pH dependencies, is given in the Appendix B.

The Energetic Cost of CO_2 Fixation

The pH-aware model can also be used to calculate the energetic cost of fixing carbon through the CCM. This cost is composed of three components: the cost of transport, the cost of the CBB cycle and the cost of photorespiratory pathways recycling 2PG (Figure 3.3). The per-fixation cost of the CBB cycle is approximately constant, but the costs of transport and photorespiration depend on the pH and the efficacy of the CCM (i.e. how much carboxylation and oxygenation take place). Here, we express the cost of fixation in units of H^+ gradient dissipation —the number of H^+ transported along the concentration gradient —i.e. the fundamental currency of the electrochemical potential. Cyanobacteria convert a proton gradient into ATP by means of the F_1-F_o ATP synthase, which synthesizes one ATP for every 4 H+ translocated [309].

Active transport of C_i is the chief energetic cost associated with the CCM. The biochemical mechanism of facilitated CO_2 uptake is not known and so we focus on the primary cyanobacterial HCO_3^- transporters, Na^+ symporters, which import one Na^+ with each HCO_3^-. In model cyanobacteria like *S. elongatus*, antiporters exchange Na^+ and H^+, and so sodium and proton gradients are interchangeable [85, 142]. This allows HCO_3^- transport cost to be expressed in units of H^+ transported across the membrane per CO_2 fixed. We assume a cost of 4 H^+ per HCO_3^- transported (Appendix B) in order to calculate the per-fixation cost of HCO_3^- transport shown in Figure 3.2.

The cost of the CBB cycle and the C_2 photorespiratory pathway can be estimated from their known stoichiometry. We calculate the cost of the CBB cycle as $\approx$ 37 H^+ per CO_2 fixed and the cost of the C_2 pathway $\approx$ 26 H^+ per 2PG recycled (Appendix B). The total cost of CO_2 fixation and 2PG recovery will depend on the relative amount of oxygenation according to the formula

$$\text{fixation and recovery cost} = \frac{\left(37 + 26\ \frac{V_o}{V_c}\right)}{1 - \frac{1}{2}} \times \frac{V_o}{V_c}$$

where (V_o/V_c) is the ratio of oxygenation to carboxylation rates, 37 H^+ is the per-carboxylation cost of the CBB cycle and 26 H^+ is the per-oxygenation cost of C_2 photorespiration. The cost has units of H^+ per carboxylation. This formula accounts for the fact that

each turn of the C_2 cycle releases one CO_2, which must be fixed again to maintain carbon balance (Appendix B). In the limit of low photorespiration, (V_o/V_c) approaches zero and the above formula converges to ≈ 37 H^+ per CO_2, the calculated cost of fixation through the CCB cycle. As the modeled CCM strongly limits oxygenation, the "fixation and recovery cost" is dominated by the cost of the CBB cycle, which is roughly constant (blue line in Figure 3.3). The rate and cost of photorespiration, in contrast, is tied to the ratio of CO_2 to O_2 in the carboxysome (i.e. CCM efficacy) and so depends on pH (grey line Figure 3.3).

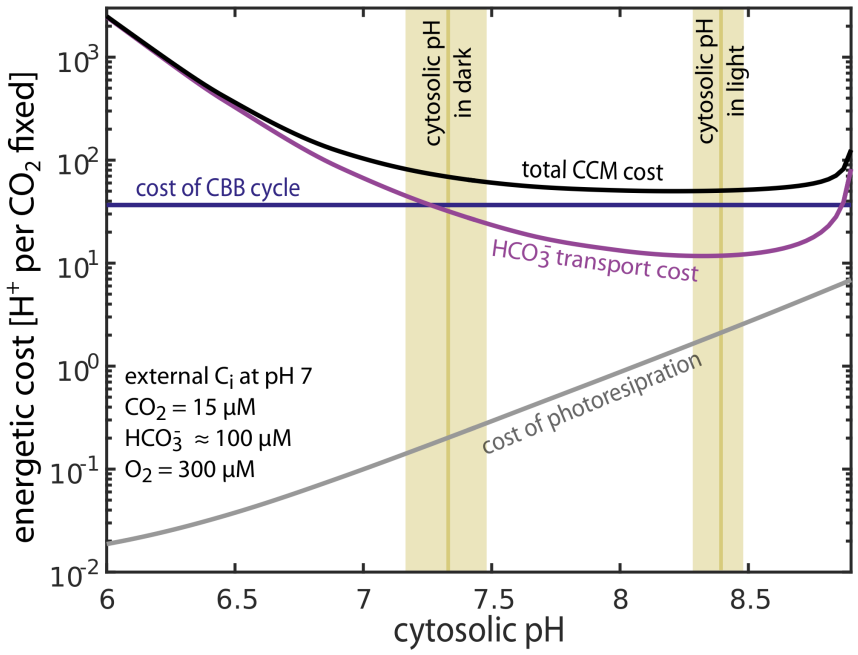


Figure 3.3: Energetic costs associated with the CCM depend strongly on the cytosolic pH. The total energetic cost of concentrating CO_2 per CO_2 fixation (black) is plotted for cytosolic pH ranging from 6-9. The purple line denotes the cost of HCO_3^- transport while the blue line gives the per-fixation cost of the CBB cycle, which is assumed to be independent of pH. The grey line gives the cost of recycling 2PG through the C_2 photorespiratory pathway. The measured cytosolic pH of *S. elongatus* PCC 7942 is shown as gold bars with bar width denoting the standard error (Methods). The extracellular C_i pool was assumed to be in equilibrium with the external pH, which was fixed at 7 for this analysis.

The projected cost of transport (purple line) falls beneath that of the CBB cycle (Figure 3.3) as the cytosolic pH surpasses 7. The modeled cyanobacterial CCM asymptotically approaches a total cost of ≈ 50 H^+ per fixation near pH 8. At pH $\approx$ 8, active transport costs about tenfold less than the CBB cycle and, we calculate, requires $< 1\%$ of cell surface area (Appendix B). Notably, photorespiratory costs increase at elevated pH (Figure 3.3) and oxygenation could account for as many as 2% of Rubisco turnovers near pH 8. This might explain the absolute requirement for photorespiratory pathways in model cyanobacteria [84]. Increased photorespiration can be understood as follows: CA equilibrates CO_2 with HCO_3^- in the carboxysome and increased pH favors HCO_3^- over CO_2 (Appendix B). As such, the carboxysomal CO_2 concentration decreases as the pH increases, increasing the O_2:CO_2 ratio and, consequently, photorespiration. So long as the pH remains beneath 8.5, the cost of photorespiration never exceeds 3% of total. Taken together, these cost calculations suggest that a cytosolic $7.5 <$ pH < 8.5 minimizes total energetic costs without requiring substantial photorespiratory flux. At a cytosolic pH < 6.2 transport is prohibitively costly, at least an order of magnitude more so that carbon fixation itself.

S. elongatus Cyotosolic pH is within the Optimal Range for CCM Operation

Increased cytosolic pH decreases the permeability of the bicarbonate pool to the cell membrane, reduces the amount of HCO_3^- transport required per CO_2 fixation and decreases the total cost of CO_2 fixation through the CCM. Our model predicts an optimal cytosolic pH range of 7.5-8.5 to minimize the total energetic cost of fixing carbon through the CCM. To test the prediction of an optimal cytosolic pH range, we measured the cytosolic pH of light- and dark-acclimated cyanobacteria (*S. elongatus* PCC 7942) using the ratiometric pH-sensitive dye BCECF-AM (Methods). As shown in Figure 3.3, cytosolic pH differs between dark-acclimated cells —which do not grow or fix CO_2 appreciably —and light-acclimated cells —which fix carbon and grow. Strikingly, the cytosolic pH of *S. elongatus* increases from 7.3 ± 0.2 in the dark to 8.4 ± 0.1 in the light (Figure 3.3, Appendix B). This is expected as the photosynthetic light reactions pump H^+ from the cytosol into the thylakoid lumen [217] and the pH values are consistent with previous measurements of cytosolic pH in cyanobacteria [35, 93, 224]. The pH in the light coincides well with our prediction of an optimal cytosolic pH range for carbon-fixing cyanobacteria.

3.5 Discussion

Here, we demonstrate that CCM function is strongly dependent on pH. Most importantly, we show that C_i leakage is dominated by H_2CO_3 within the physiological pH range (Appendix B). Elevated cytosolic pH increases the relative abundance of charged HCO_3^- and so reduces leakage and the energetic cost of maintaining a high cytosolic concentration of H_{total}. This mirrors heterotrophic transport strategies like the phosphorylation of glucose after uptake, which serves to "trap" glucose in the cell. We predict an optimal cytosolic pH range $7.5 - 8.5$ for carbon-fixing cyanobacteria. At pH 7.7 the total energetic cost of fixing carbon through the CCM is minimized, with higher pH offering diminishing returns (Figure 3.3). We experimentally verified that cyanobacteria (*S. elongatus* 7942) achieve a cytosolic pH in this range while fixing carbon (Figure 3.3). Moreover, *S. elongatus* cytosolic pH in the light (during carbon fixation) is ≈ 8.4 and differs markedly from the pH ≈ 7.3 measured in dark-acclimated cells (Figure 3.3).

Notably, efficient operation of the CCM depends on two crucial unknown parameters —the velocity of active HCO_3^- transport and the permeability of the carboxysome to CO_2 and HCO_3^- [182]. These parameters are interdependent: changing the carboxysome permeability alters the HCO_3^- transport required to concentrate CO_2 in the carboxysome, as shown in Appendix B. Integrating pH into the model does not qualitatively change the interdependence of carboxysome permeability and active HCO_3^- transport. However, at pH 8, the pH-aware CCM model demands 10^2-10^4 times less active HCO_3^- transport than our previous model to achieve efficient CO_2 fixation (Figure 3.2). We defined the optimum permeability as the value that minimizes the active HCO_3^- transport flux required to saturate Rubisco. The

optimal carboxysome permeability at pH 8 is also thirty-fold lower than previously reported (3×10^{-5} instead of 10^{-3} cm/s), but carboxysome permeabilities as high as 10^{-2} cm/s can support CO_2 concentrating activity (Appendices B and C).

This updated, pH-aware model now agrees with cyanobacterial physiology in several ways. In optimal pH conditions the predicted fluxes for HCO_3^- transport, leakage, and carboxylation are all within an order of magnitude of measured values [132]. In contrast to earlier models, which required $> 50\%$ of the cell membrane surface for C_i transport, these fluxes demand less than 1% of cell membrane surface area (Appendix B). The pH-aware model also helps rationalize the distribution of C_i transport systems among cyanobacteria. Freshwater cyanobacteria, which live near neutral pH, typically have genes coding for both facilitated CO_2 uptake and energetically activated HCO_3^- transport systems. Oceanic cyanobacteria, which live at pH ≈ 8, typically encode only HCO_3^- transporters [132, 230]. Based on our analysis of the pH-aware model, a sizable fraction of C_i can be taken up as CO_2 in a near-neutral pH environment like freshwater, while in an environment at pH 8 CO_2 uptake would contribute negligibly to the overall C_i uptake rate (Appendix B).

Selectivity at the Carboxysome Shell

Throughout this text, we assumed that the carboxysome is equally permeable to CO_2 and HCO_3^-. However, recent structures of carboxysome shell proteins offer a potential mechanism for differential permeability of CO_2 and HCO_3^-: the pores of shell proteins typically carry positive charge, which might increase the rate of HCO_3^- transit relative to CO_2 [154, 330]. Indeed, recent experimental evidence suggests that protein compartments can be selectively permeable [65, 114]. Intuitively, it seems that a very high HCO_3^- permeability and a very low CO_2 permeability would be best for the efficient operation of the CCM. Such permeabilities would maximize HCO_3^- uptake and minimize loss of CO_2, ensuring that every carbon entering the carboxysome is ultimately fixed.

We used the pH-aware model to examine this intuition. As shown in Figure 3.3, selectivity does not substantially improve the performance of the CCM. Rather, we find that a low permeability for CO_2 ($\approx 10^{-4}$ cm/s) is a critical requirement for CCM function. To explain this non-intuitive result, note that the best possible case for the CCM is that the carboxysomal CA brings the cytosolic HCO_3^- pool into perfect equilibrium with CO_2 inside the carboxysome (i.e. with negligible leakage of carboxysomal CO_2, dashed black line in Figure 3.3A). As no energetic coupling is known or hypothesized to exist inside the carboxysome, equilibration of HCO_3^- with CO_2 would maximize the carboxysomal CO_2 concentration. Indeed, the optimal single carboxysome permeability $k_c = 3 \times 10^{-5}$ cm/s very nearly achieves this equilibrium CO_2 concentration (within 5%, grey dashed line in Figure 3.3A).

While a selective system could reduce the amount of active HCO_3^- transport required per fixation (Appendix B), transport costs represent ˜20% of the total cost at a cytosolic pH of 8 (Figure 3.2) and so selectivity would have only a very small effect on the total cost of fixation in these conditions ($< 2\%$, grey dashed line in Figure 3.3B). Increasing the permeability of the carboxysome to CO_2 beyond 10^{-4} cm/s, on the other hand, exponentially increases the cost

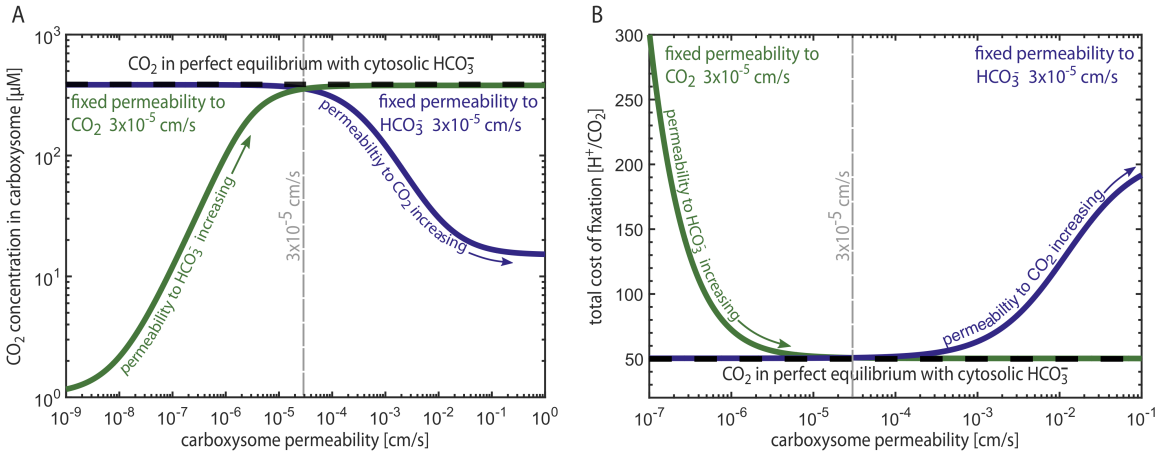


Figure 3.4: A selective carboxysome would not substantially improve CCM efficiency. The CO_2 concentration in the carboxysome is maximized when the cytosolic HCO_3^- pool is brought into equilibrium inside the carboxysome, shown as the black dashed line in panel (A). A CCM using the single optimal permeability computed through the pH-aware model (dashed grey line) achieves a carboxysomal CO_2 concentration within 5% of this maximum. Selectivity at the carboxyome shell is thought to increase HCO_3^- permeability relative to CO_2 by means of charge interactions in the pores of the carboxysome shell. Selectivity might intuitively result in greater trapping of CO_2 (tracing the purple curve towards lower permeabilities) or faster uptake of HCO_3^- (tracing the green curve towards higher permeabilities) but neither of these strategies can increase the carboxysomal CO_2 concentration above equilibrium. As shown in panel (B), selectivity would not substantially reduce the total cost of fixing carbon through the CCM, which is already nearly minimized at the single optimal permeability of 3×10^{-5} cm/s. Increasing the CO_2 permeability beyond 10^{-4} cm/s, however, exponentially increases the cost of fixation due to leakage of CO_2 from the carboxysome to the cytosol.

of fixation irrespective of the HCO_3^- permeability (Figure 3.3B, Appendix A). As a result, we find that selectivity is not necessary to produce a functional and energetically efficient CCM and suggest that future research should focus on understanding how the carboxysome maintains a permeability barrier to CO_2 (Appendices B and C).

A Relatively Acidic Carboxysome would Benefit the CCM

It has been suggested that the carboxysome might maintain a pH gradient, with the cytosolic and carboxysomal pHs differing [190, 320]. Altering the pH in the carboxysome would influence CCM efficiency in two ways: pH affects the equilibrium composition of C_i and the kinetics of the carboxysomal enzymes. A more acidic pH < 8 would increase the equilibrium CO_2 concentration relative to the cytosol (Appendix B, and increase Rubisco's maximum carboxylation rate (Appendix B). Indeed, as the combined action of CA and Rubisco produces a net H^+, it may be possible for the cell to maintain a steady state where the carboxysomal pH differs from the cytosol (Appendix B).

Figure 3.5 shows that, within limits, a relatively acidic carboxysome would result in a higher degree of Rubisco saturation at the same rate of cellular HCO_3^- uptake, with a carboxysomal pH ≈ 7 minimizing the amount of HCO_3^- uptake required to saturate Rubisco (blue region). If the carboxysome is too acidic, however, it can deleteriously affect Rubisco

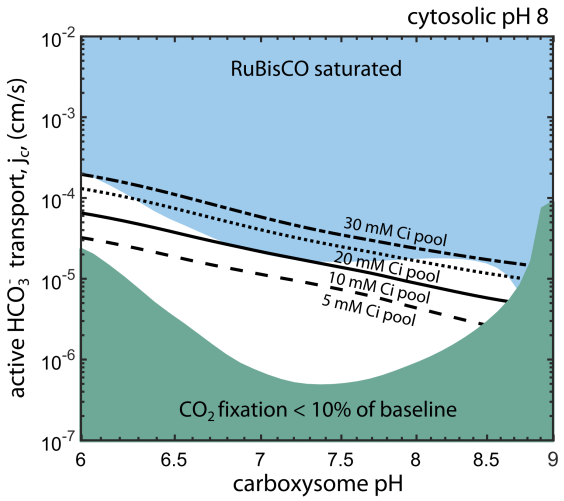


Figure 3.5: A relatively acidic carboxysome would improve CCM performance. The phase space depicts the effect of varying carboxysomal pH and HCO_3^- transport on the degree of Rubisco saturation in the carboxysome. The cytosolic pH was set to 8. The blue region denotes the portion of phase space wherein Rubisco is saturated with CO_2. When the carboxysomal pH is relatively acidic (roughly pH 7) less HCO_3^- uptake is required to saturate Rubisco because the equilibrium between CO_2 and HCO_3^- favors CO_2 more than at pH 8 (Appendix B). Moreover, Rubisco has an increased carboxylation k_cat at this pH. The minimum cytosolic HCO_3^- concentration (black lines) that allows for saturation of the carboxysomal Rubisco is 10-20 mM, depending on Rubiscos CO_2 affinity, which also varies with pH (Appendix B). The teal region denotes the portion of phase space where the total CO_2 fixation flux is less than 10% of the flux in reference conditions (pH 8 in the cytosol and carboxysome). This region is labeled to emphasize that the fixation flux collapses at basic pH because the carboxylation k_cat is projected to vanish in this regime (Appendix B).

kinetics (Appendix B). An alkaline carboxysomal pH > 8.5 would encumber the CCM on two fronts: as the pH increases, the equilibrium CO_2 concentration decreases exponentially and Rubisco's maximum carboxylation rate vanishes [18]. Furthermore, we find that a cytosolic HCO_3^- concentration beneath ≈ 10 mM is insufficient to saturate the carboxysomal Rubisco (Figure 3.5, Appendix B). This can be seen by calculating the CO_2 concentration that would be achieved by equilibrating 5 mM HCO_3^-, $[CO_2] = \left[HCO_3^-\right] 10^{(pK_{\text{eff}} - pH)} \approx 5\text{ mM} \times 10^{6.1-7} \approx 600\ \mu M$, where $pK_{eff} = 6.1$ is the effective pK_a, between aqueous CO_2 and HCO_3^- (Appendix B). The modeled Rubisco, by contrast, has a carboxylation K_{M} that increases with decreasing pH, reaching approximately 1 mM at pH 7 (Appendix B), and so cannot be saturated by 5 mM cytosolic HCO_3^-. Though the absolute carboxylation K_{M} varies between Rubisco forms [262, 301], the plant enzyme displays a similar pH dependence, with pH 7 resulting in at least twofold reduced CO_2 affinity [18, 46, 270].

Model Assumptions and Future Directions

The major insight of the pH-aware model is that the effective permeability of the membrane to H_{total} depends on the relative abundance of the species comprising H_{total} (HCO_3^-, H_2CO_3 and CO_3^{2-}) and hence on the pH and ionic strength (Appendix B). The model ignores the effect of H_2CO_3 and CO_3^{2-} on all other CCM fluxes. HCO_3^- is the true substrate of carbonic anhydrase, so it is reasonable to ignore the enzyme's interaction with other species. However, it is not known whether HCO_3^-, H_2CO_3 and CO_3^{2-} are equally permeable to the carboxysome. As mentioned above, pores on the carboxysome shell often carry positive charge, so it is plausible that charge would affect the rate substrates enter the carboxysome [154, 330]. Direct measurement of the permeability characteristics of the carboxysome would

be of great help in understanding the CCM.

We further assume throughout this work that CO_2 is the limiting substrate for Rubisco carboxylation, i.e. that the carboxylation rate does not depend on the concentration of the five-carbon carboxylation substrate ribulose-1,5-bisphosphate (RuBP). This assumption is supported by measurements of millimolar RuBP in *S. elongatus* [320], but it remains unclear how RuBP enters the carboxysome [53, 154, 250]. Similarly, we assume that O_2 enters the carboxysome quickly enough to equalize concentrations across the carboxysome shell (Appendix B), i.e. that the CCM is not an "oxygen blocking mechanism."

In our analysis of the membrane permeability of H_{total}, we assumed that the permeability of H_2CO_3 is well-approximated by those of formic and acetic acids (H_2CO_2 and $H_4C_2O_2$ respectively, Appendix B). Model results are particularly sensitive to our assumptions about membrane and carboxysome permeabilities (Appendix B), emphasizing that quantitation of these permeability values is vital to understanding the cyanobacterial CCM. We further assume the cytosolic and carboxysomal pH are equal based on the observation of fast pH equilibration across the α-carboxysome shell [190]. However, it may be possible for the CCM to maintain a ΔpH across the carboxysome shell (Appendix B). Although the pH-aware model suggests that a relatively acidic carboxysomal pH ≈ 7 would enhance CCM efficiency (Figure 3.5), it is difficult to imagine how the carboxysomal pH might be measured *in vivo*. Finally, we assume a 30 mM concentration of cytosolic HCO_3^- based on a number of measurements [295, 320, 327]. As discussed above, ≈ 10 mM is the lowest cytosolic HCO_3^- concentration that would saturate the modeled carboxysomal Rubisco with CO_2 (Figure 3.5, Appendix B).

A long history of research into photosynthetic physiology has shown that diverse phototrophs maintain a basic pH near 8 around Rubisco while fixing carbon. Spinach chloroplasts, which contain no CCM, maintain a pH near 7 in the dark and shift to pH ≈ 8 in the light [316]. As such, it is unclear whether the prevailing pH is an adaptation to the CCM or the CCM adapted to the pH. Indeed, active photosynthesis requires pumping of protons into the thylakoid and so it is sensible that the pH should increase in the stroma/cytoplasm [93, 123]. We simply note that these explanations are not mutually exclusive, i.e. that it is possible that a pH ≈ 8 results from photosynthetic proton pumping and also optimizes CCM efficiency. Many eukaryotic algae have a CCM based on a structure called the pyrenoid that is evolutionarily distinct from, but physiologically similar to the cyanobacterial CCM [142, 243]. Similar coordinated pH shifts are also known to occur in algal chloroplasts [49]. Perhaps a shift from pH ≈ 7 in the dark to pH ≈ 8 in the light reflects a coordination of chloroplast pH with the algal CCM? We hope that future investigations into pH homeostasis and intracellular fluxes of carbon-fixing organisms help refine our understanding of the inner workings of CCMs.

3.6 Methods

Modeling

The analytic equations of our previous model [182] were updated to account for the effect of pH on the composition of H_{total}, H_{total} membrane permeability and Rubisco and carbonic anhydrase activity. The pH-dependent composition of H_{total} was calculated using thermodynamic potentials derived by [211]. Permeability coefficients for CO_2, H_2CO_3 and HCO_3^- and were derived from literature values for those and similar molecules. The pH dependence of Rubisco and CA kinetics were extracted from biochemical studies and scaled to match the carboxysomal enzymes. We further developed a mathematical framework to integrate the pH dependence of H_{total} permeability and selectivity at the carboxysome shell while preserving our capacity to solve the model analytically. In addition to the integration of detailed pH dependence, we described a carbon flux balance relation and model the energetic cost of all fluxes in the model except for facilitated CO_2 uptake. To verify the accuracy of our analytic solutions, we compared them to results produced by simulating a numerical model. The updated model is described in detail in the Appendix B, implemented in MATLAB and freely available at github.com/SavageLab/ccm/.

S. elongatus pH Measurement

The intracellular pH of *S. elongatus* PCC 7942 was measured using the ratiometric pH dye 2,7-Bis(2-carboxyethyl)-5(6)-carboxyfluorescein acetoxymethyl ester (BCECF-AM). The calibration curve was generated as follows. *S. elongatus* 7942 was grown to mid-log phase, washed and resuspended in BG11 media with 20 μM BCECF-AM. Cells were incubated with BCECF-AM for 30 minutes in the light at 30 °C and then resuspended in BG11 of defined pH containing 20 μM of the ionophore nigericin. After a 10-minute incubation, four replicates of each pH condition were quickly loaded onto a 96-well plate and fluorescence was measured in a Tecan M1000 plate reader with excitation/emission pairs 440/535 and 490/535. The 490:440 emission ratio was fit to a Boltzmann sigmoid to generate the calibration curve in Appendix B.

S. elongatus PCC 7942 in mid-log phase were pre-incubated in light or in the dark for 9 hours. These cultures were washed and resuspended in their spent growth media with 20 μM BCECF-AM and incubated for 30 minutes at 30 °C in the light or dark as appropriate. For dark-treated cultures, all pipetting steps were carried out in a dark room with a low-intensity green LED light. Fluorescence was measured as above. Fluorescence ratios were calculated for each replicated and converted to pH values by inverting the calibration curve.

Acknowledgements

We are grateful to Arren Bar-Even, Elad Noor, Christopher Jakobson, Patrick Shih, Yinon Bar-On, Martin Jonikas, Dan Arlow, Dan Tawfik, John Raven, Grant Rotskoff, Anna

Chen, Luke Oltrogge, Yoni Savir, Rosalie Lawrence, Niv Antonovsky, Sumedha Ravishankar, Jeremy Roop, and Antony Lee for fruitful discussions and helpful insights.

Chapter 4

Comprehensive Characterization of Bacterial CCM Genes by Transposon Mutagenesis

4.1 Abstract

Bacterial autotrophs often rely on CO_2 concentrating mechanisms (CCMs) to assimilate carbon. Although many CCM proteins have been identified, a systematic screen of CCM components has not been carried out. Here, we performed a genome-wide barcoded transposon screen to identify essential and CCM-related genes in the γ-proteobacterium *H. neapolitanus*. Screening revealed that the CCM comprises about 25 genes mostly encoded in 3 operons. Two of these operons contain a two-gene locus encoding a domain of unknown function (PFAM:PF10070) and a putative cation transporter (PFAM:PF00361). Physiological and biochemical assays demonstrate that these proteins, which we name DabA and DabB for "DABs accumulate bicarbonate," assemble into a heterodimeric complex, contain a putative β-carbonic anhydrase-like active site, and function as an energy-coupled inorganic carbon (C_i) pump. Surprisingly, DabAB operons are found in diverse bacteria and archaea. We demonstrate that functional DABs are present in the human pathogens *B. anthracis* and *V. cholerae*. Based on these results, we propose that DABs constitute a new class of energized C_i pump and play a critical role in C_i metabolism throughout prokaryotic phyla.

4.2 Introduction

Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase (Rubisco) is the primary carboxylase of the Calvin-Benson-Bassham (CBB) cycle and the major entry point of C_i into the biosphere. Rubisco activity is critical to agriculture and a major flux removing anthropogenic CO_2 from the atmosphere. Despite its centrality and abundance, Rubisco is not a fast enzyme [26, 31]. Nor is Rubisco specific - all known Rubiscos admit molecular oxygen (O_2) as a substrate

in place of CO_2 [300]. Oxygenation does not fix carbon and produces a product that must be recycled through metabolically-expensive photorespiratory pathways [33]. Many studies support the hypothesis that improvements to Rubisco could improve crop yields, but Rubisco has proven recalcitrant to protein engineering. It remains unclear whether or how Rubisco can be improved [262, 301].

Organisms that depend on Rubisco for growth often employ CO_2 concentrating mechanisms (CCMs) that concentrate CO_2 near Rubisco so that carboxylation proceeds at its maximum rate and oxygenation is competitively inhibited [243]. All cyanobacteria and many chemolithoautotrophic proteobacteria have a CCM [237]. The bacterial CCM has garnered particular interest among bioengineers because it is well-understood, it is thought to consist of relatively few genes and it operates inside single cells. Detailed modeling suggests that transplantation of the bacterial CCM into crops might improve yields [187] and efforts towards transplantation are already underway [165, 171].

Diverse experimental studies make it clear that the bacterial CCM requires two components to function: active C_i transport driving accumulation of HCO_3^- in the cytosol, and organization of Rubisco with carbonic anhydrase (CA) in the lumen of a protein organelle called the carboxysome [228, 132]. Energy-coupled C_i pumps keep the cytosolic HCO_3^- concentration high (> 10 mM) and, crucially, out-of-equilibrium with CO_2 [129, 228, 320]. CA activity interconverts $HCO_3^- + H^+$ with $CO_2 + H_2O$. As such, the carboxysomal CA converts a high cytosolic HCO_3^- concentration into a high carboxysomal CO_2 concentration, promoting faster carboxylation by Rubisco and competitively inhibiting (Figure 1.2). Genetic lesions to either component - C_i uptake systems or carboxysomes - disrupt the CCM and mutants have growth defects unless CO_2 is supplemented [184, 229]. The high-CO_2 requiring (HCR) mutant phenotype is commonly-used to identify CCM components in screens.

Despite early screens, a comprehensive list of bacterial CCM components remains unknown, leaving the possibility that additional activities are required for CCM function. Although well-assembled carboxysome structures can be produced in bacteria and plants [44, 171], the functionality of these carboxysomes in a heterologous CCM has not been demonstrated. Bioinformatic studies show that several non-carboxysomal genes are associated with carboxysome operons [16, 138]. Further, experimental [229, 276] and modeling studies [182, 132] make it clear that energy-coupled C_i uptake systems are required for CCM function. Several different C_i pump families, including transporters and facilitated uptake systems are known for cyanobacterial lineages, but mechanistic understanding of C_i uptake is limited [230].

Here we use a genome-wide barcoded transposon mutagenesis screen (RB-TnSeq) to interrogate the CCM of *H. neapolitanus* (henceforth *Hnea*). *Hnea* is a sulfur oxidizing γ-proteobacterial chemolithoautotroph and a model system for studying α-carboxysomes [122]. *Hnea* possesses a C_i uptake system that uses CO_2 as its substrate and is powered by membrane potential, but the components of this system are unknown [129]. In addition to producing the first essential gene set for a bacterial chemolithoautotroph, we leverage our pooled mutant library to comprehensively screen for knockouts that produce an HCR phenotype. This screen identified all known CCM components and confirmed that a two-

gene operon containing a large, conserved, poorly-characterized protein (PFAM:PF10070, hereafter DabA) and a member of a large family of cation transporters (PFAM:PF00361, hereafter DabB) is required for CCM function. Scott and colleagues recently identified and validated homologs of these genes as a C_i import system in hydrothermal vent chemolithoautotrophs [183, 268, 267]. Based on this work and the results described below, we propose naming this locus the DAB operon for "**D**ABs **A**ccumulate **B**icarbonate."

Here we show that the products of the DAB operon form a protein complex that is capable of energetically-coupled C_i uptake. Both proteins are necessary for activity and treatment with an ionophore abrogates DAB-mediated C_i uptake. Structural homology modeling suggests that DabA contains a domain distantly homologous to a type II β-CA. Indeed, DabA binds zinc, likely in a manner similar to β-CAs. These results are consistent with a model of activity dependant on unidirectional hydration of CO_2 to HCO_3^- in the cytosol via a CA-like mechanism and energized by coupling to a cation gradient. Phylogenomic analysis demonstrates that DAB operons are widespread throughout prokaryotes including carbon-fixing bacteria and archaea. Surprisingly, DAB operons are also found in many heterotrophic bacteria. We demonstrate that functional operons are present in the notable pathogens *V. cholera* and *B. anthracis*. We therefore propose that DABs constitute a novel class of C_i uptake pump whose biochemical tractability facilitates mechanistic analyses and whose widespread occurrence merits further investigation.

4.3 Results

Transposon Mutagenesis and Gene Essentiality

We constructed a randomly-barcoded genome-wide pooled knockout library of *Hnea* by conjugation (Figure 4.1A). The conjugated vector contained a barcoded Tn5-derived transposon encoding a kanamycin resistance marker. The library was produced in 5% CO_2 enabling isolation of CCM gene knockouts.

Transposon barcodes simplify the use of the library for pooled screens using the 'barseq' approach (Methods, [318]). Transposon insertion sites and cognate barcodes were mapped using standard TnSeq methods (Methods). The library was found to contain $\approx 10^5$ insertions, or one insertion for every ≈25 base pairs in the *Hnea* genome. Since the average gene contains ≈35 insertions, genes with no insertions are very likely essential for growth. A simple statistical model identified 551 essential genes and 1787 nonessential genes out of 2408 genes in the *Hnea* genome (Methods, Figure 4.1A-B). The remaining 70 genes were classified as "ambiguous" due either to their short length or because replicate mapping experiments were discordant (Methods). Genes associated with known essential functions including central carbon metabolism, ribosome production, and DNA replication were found to be essential (Figures 1C and S1). Importantly, known CCM genes, including carboxysome components, were not essential for growth at 5% CO_2 (Figure 4.2).

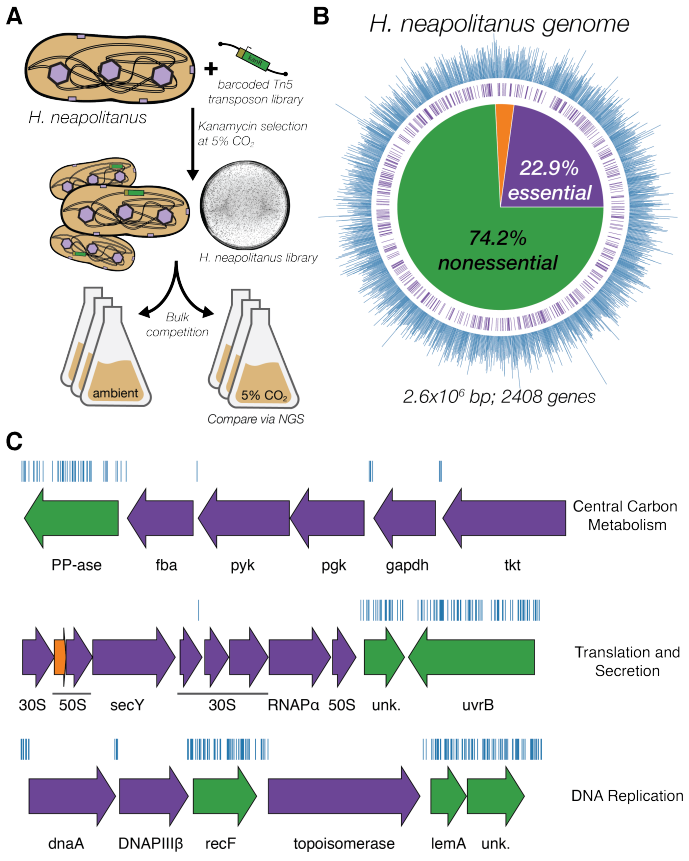


Figure 4.1: Enumerating *H. neapolitanus* essential genes by transposon mutagenesis. (A) Schematic depicting the generation and screening of the RB-TnSeq library. Transposons were inserted into the *Hnea* genome by conjugation with an *E. coli* donor strain. The transposon contains a random 20 base pair barcode (yellow) and a kanamycin selection marker (green). Selection for colonies containing insertions was performed in the presence of kanamycin at 5% CO_2 and insertions were mapped by sequencing as described in the Methods. Subsequent screens were carried out as bulk competition assays and quantified by BarSeq. (B) Insertions and essential genes are well-distributed throughout the *Hnea* genome. The outer track (blue) is a histogram of the number of barcodes that were mapped to a 1 kb window. The inner track annotates essential genes in purple. The pie chart shows the percentages of the genome called essential (purple), ambiguous (orange), and nonessential (green). (C) Representative essential genes and nonessential genes in the *Hnea* genome. The blue track indicates the presence of an insertion. Genes in purple were called essential and genes in green are nonessential. Genes labeled "unk." are hypothetical proteins. The first genomic locus contains 5 essential genes involved in glycolysis or the CBB cycle including pyruvate kinase (pyk) and transketolase (tkt). The 8 essential genes in the second locus encoding 30S and 50S subunits of the ribosome, the secY secretory channel, and an RNA polymerase subunit. Essential genes in the third example locus include topoisomerase and DNA polymerase III β.

Comprehensive Screen for *Hnea* CCM Components

Based on the current model of the bacterial CCM (Figure 4.2A), knockouts of CCM genes are expected to have reduced fitness in atmospheric CO_2 conditions [184, 229]. As our pooled library contains ≈70,000 barcodes that map to exactly one position in the *Hnea* genome, we were able to use barseq to quantify the fitness effects of single gene knockouts for all nonessential *Hnea* genes in a pooled competition experiment (Methods, Figure 4.2B). Since the library contains about 20 uniquely-mapped knockouts per gene, this screen contains multiple internal biological replicates testing the effect of gene knockouts. We assigned a gene knockout an HCR phenotype if the average effect of all knockout mutants was a twofold (or greater) growth defect in ambient CO_2 as compared to 5% in two replicate experiments.

As expected, knockouts of carboxysome genes consistently produced large and specific fitness defects in ambient CO_2 (Figures 2B-C). These genes include *cbbLS* - the large and small subunits of the α-carboxysomal Rubisco; *csoS2* - an intrinsically disordered protein required for α-carboxysome assembly [58]; *csoSCA* - the carboxysomal carbonic anhydrase; *csoS4AB* - the pentameric proteins thought to form vertices of the α-carboxysome; and

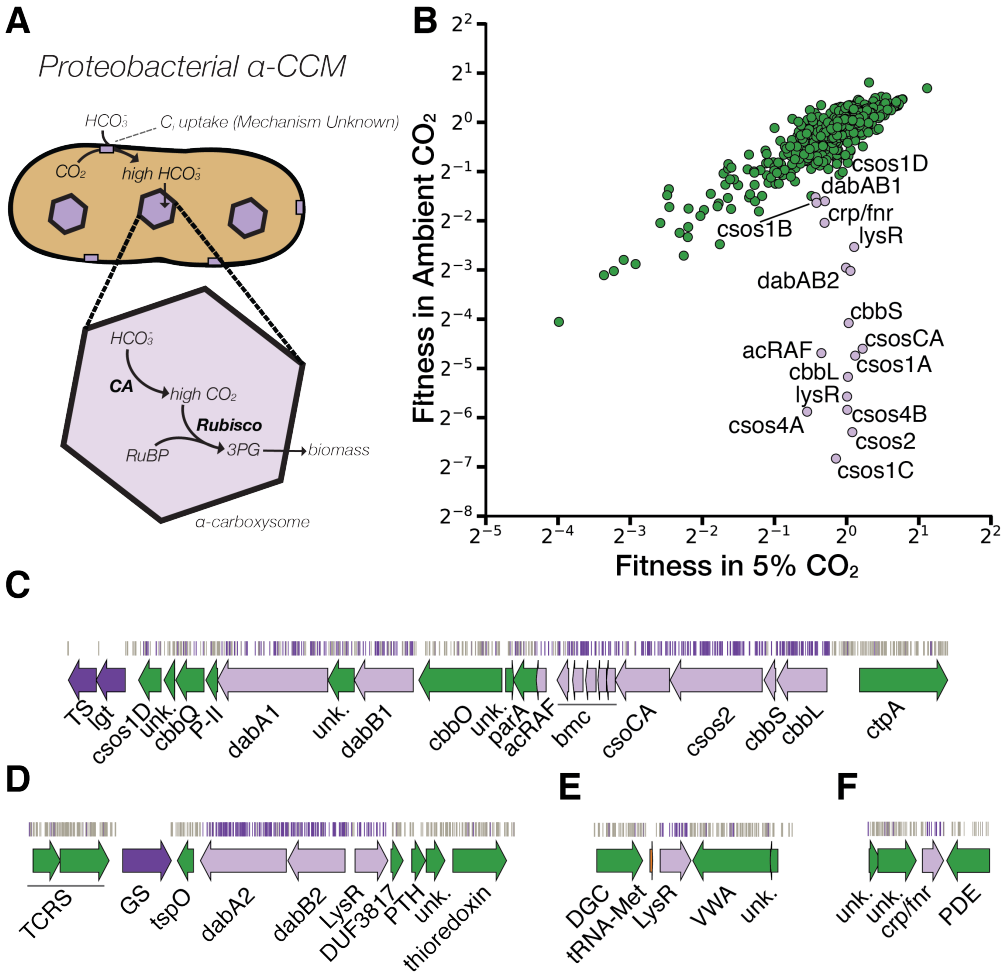


Figure 4.2: A systematic screen for high CO_2-requiring mutants identifies genes putatively associated with the CCM. (A) Simplified model of the α-CCM of chemolithoautotrophic proteobacteria. Inorganic carbon is concentrated via an unknown mechanism, producing a high cytosolic HCO_3^- concentration. High cytosolic HCO_3^- is converted into high carboxysomal CO_2 by CA, which is localized only to the carboxysome. (B) Fitness effects of gene knockouts in 5% CO_2 as compared to ambient CO_2. Data is from one of two replicates of BarSeq - the second replicate gives consistent results. The effects of single transposon insertions into a gene are averaged to produce the gene-level fitness value plotted. We define HCR mutants as those displaying a twofold fitness defect in ambient CO_2 relative to 5% CO_2 in both replicates. HCR genes are colored light purple. Data from both replicates and the associated error bars are shown in the Appendix. Panels C-F show regions of the *Hnea* genome containing genes annotated as HCR in panel A. Essential genes are in dark purple, HCR genes are in light purple, and other genes are in green. The top tracks show the presence of an insertion in that location. Insertions are colored grey unless they display a twofold or greater fitness defect in ambient CO_2, in which case they are colored light purple. (C) The gene cluster containing the carboxysome operon and a second CCM-associated operon. This second operon contains acRAF, a Form IC associated cbbOQ-type Rubisco activase and *dabAB1*. (D) The DAB2 operon and surrounding genomic context. (E) The genomic context of a lysR-type transcriptional regulator that shows an HCR phenotype. (F) Genomic context of a crp/fnr-type transcriptional regulator that displays an HCR phenotype. Genes labeled "unk." are hypothetical proteins.

csoS1CAB - the hexamers that form the faces of the α-carboxysome shell [238]. Knockouts of *csoS1D*, a shell hexamer with a large central pore [250], had too weak a phenotype to be

considered HCR (Figures 2B-C). The *Hnea* genome also contains a non-carboxysomal Form II Rubisco that is likely not involved in CCM activity as its disruption confers no fitness defect. A number of genes that are not associated with the carboxysome structure also exhibited HCR phenotypes. These include two LysR transcriptional regulators, a Crp/Fnr type transcriptional regulator, a protein called acRAF that is involved in Rubisco assembly [5, 319], and two paralogous loci encoding DAB genes (hereafter DAB1 and DAB2, Figure 4.2B-F).

DAB Operon Composition

DAB1 is a cluster of 2 genes found in an operon directly downstream of the carboxysome operon (Figure 4.2C). Though DAB1 is part of an 11-gene operon containing several genes associated with Rubisco proteostasis, including acRAF [5, 319] and a cbbOQ-type Rubisco activase [198], we refer to DAB1 as an "operon" for simplicity. DAB2 is a true operon and is not proximal to the carboxysome operon in the *Hnea* genome. These "operons" are unified in that they both display HCR phenotypes and possess similar genes (Figures 2B-D) and are commonly found in association with proteobacterial α-carboxysomes (Figure 4.3).

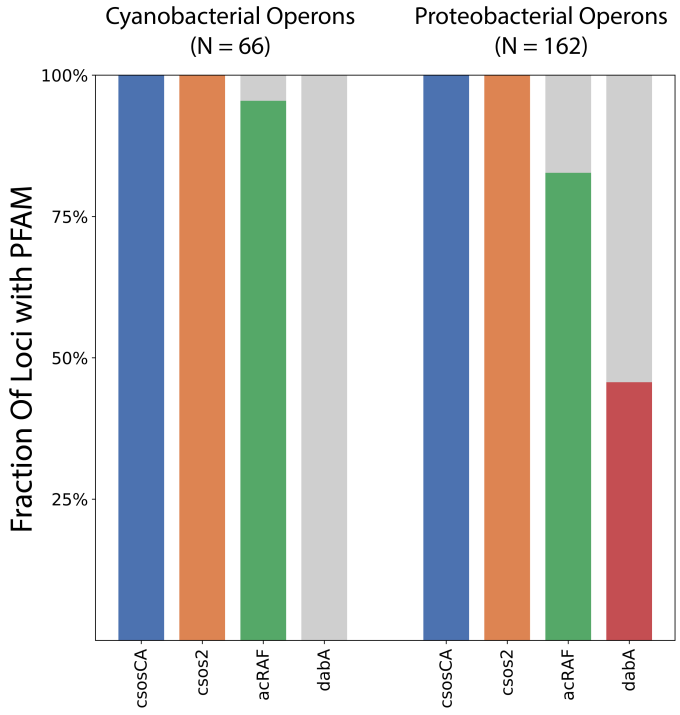


Figure 4.3: DAB operons are specifically associated with the proteobacterial CCM. α-carboxysome operons were culled from the IMG database by searching for loci containing a large and small subunit of Rubisco, an α-carboxysomal CA gene, and a bacterial microcompartments (BMC) protein (PFAMs PF00016, PF00101, PFAM08936, PF00936, respectively). We extracted the genes within 100 kilobases of these operons to determine the genetic locus. *dabA* was used as a marker to investigate the dispersion of DAB operons. Nearly all loci contain the positive control α-carboxysome gene, *csos2*, which was not searched for. The putative Rubisco chaperone *acRAF* is found in $>$ 75% of proteobacterial and cyanobacterial loci. In contrast, *dabA* was found only proteobacterial loci and only in $\approx$ 45% of them, suggesting that some proteobacteria use other C_i uptake mechanisms.

Both operons contain a conserved protein of unknown function (PFAM:PF10070) that we term DabA. DabAs have no predicted transmembrane helices or signal peptides and appear to be large (DabA1: 118.5 kDa, DabA2: 91.7 kDa), soluble, cytoplasmic proteins (Methods, Figure 4.4A). Both DAB operons also contain a member of the cation transporter

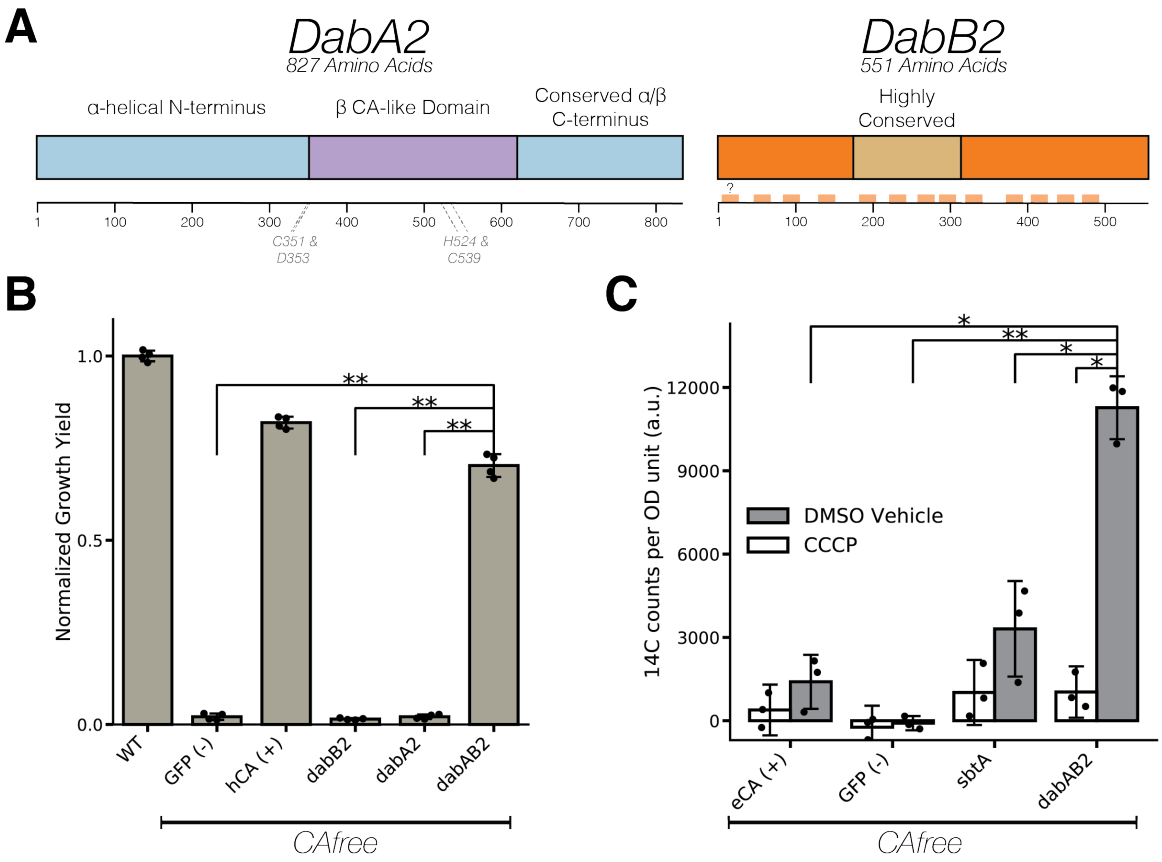


Figure 4.4: DABs catalyze active transport of C_i and are energized by a cation gradient. (A) Diagrammatic representation of DabA2 and DabB2 based on bioinformatic annotation. The four predicted active site residues (C351, D353, H524, C539) are marked on the primary amino acid sequence. Amino acid numbers are marked below each gene and predicted transmembrane helices are marked in light orange. (B) DAB2 was tested for ability to rescue growth of CAfree *E. coli* in ambient CO_2 conditions. Expression of the full operon (DabAB2) rescues growth, as does the positive control, and human carbonic anhydrase II (hCA). Error bars represent standard deviations of 4 replicate cultures. (C) CAfree *E. coli* were tested for C_i uptake using the silicone-oil centrifugation method. Expression of DabAB2 produced a large and statistically significant increase in ^{14}C uptake as compared to all controls. Moreover, treatment with the ionophore CCCP greatly reduces DabAB2-mediated ^{14}C uptake, suggesting that DabAB2 is coupled to a cation gradient. *E. coli CA* (eCA) was used as a control for a non-vectorial CA. *Synechococcus elongatus* PCC 7942 *sbtA* was used as a known C_i transporter. GFP was used as a vector control. Error bars represent standard deviations of 3 technical replicates. In (B) and (C) "*" denotes that the means are significantly different with Bonferroni corrected $p < 0.05$ according to a two-tailed t-test. "**" denotes $p < 5 \times 10^{-4}$. In panel B, dabAB2 has a larger rescue than GFP (t=42.6, corrected $p < 10^{-7}$), dabA2 (t=43.4, corrected $p < 10^{-7}$), and dabB2 (t=44.5, corrected $p < 10^{-7}$). In panel C, dabAB2 expressing cells treated with DMSO have greater uptake than dabAB2 expressing cells treated with CCCP (t=13.6, corrected $p = 6 \times 10^{-4}$), sbtA expressing cells treated with DMSO (t = 6.7, corrected $p = 10^{-2}$), GFP expressing cells treated with DMSO (t=17.1, corrected $p = 2 \times 10^{-4}$), or eCA expressing cells treated with DMSO (t=11.5, corrected $p = \times 10^{-3}$).

family (PFAM:PF00361) that includes H^+-pumping subunits of respiratory complex I and Mrp Na^+:H^+ antiporters. This protein, which we call DabB (DabB1: 62.2 kDa, DabB2: 59.3 kDa), is predicted to have 12-13 transmembrane helices (Figure 4.4A). The complex I subunits in PF00361 are H^+-pumping proteins and do not contain redox active groups, e.g. iron-sulfur clusters or quinone binding sites. Phylogenetic analysis suggests DabB proteins

form a clade among PF00361 members (Appendix) distinct from complex I subunits. Therefore, homology between DabB and complex I subunits (e.g. NuoL) suggests cation transport but does not imply redox activity [159]. Importantly, operons of this type were recently demonstrated to be capable of C_i uptake in the hydrothermal vent chemolithoautotroph *Hydrogenovibrio crunogenus* [183, 268, 267].

dabA2 and *dabB2* are Necessary and Sufficient for Energy-Coupled C_i Accumulation in *E. coli*

In order to facilitate testing for C_i transport activity, we generated an *E. coli* strain, CAfree, that has knockouts of both CA genes (Methods). It was previously shown that deletion of the constitutive CA, *can*, gene produces an HCR phenotype in *E. coli* [191] that is complemented by expression of cyanobacterial bicarbonate transporters [82]. Deleting both CA genes replicates this phenotype and greatly reduces the likelihood of escape mutants. Since DAB2 disruption is associated with a larger fitness defect than DAB1 (Figure 4.2B), we used CAfree to test DAB2 for C_i uptake activity. DAB2 expression enables growth of CAfree in ambient CO_2 while expression of either gene alone is not sufficient (Figures 3B and S5). ^{14}C uptake assays demonstrate that DAB2 facilitates import of extracellular C_i to levels significantly above that of the appropriate control (Figure 4.4C). Moreover, DAB2-associated C_i uptake is strongly inhibited by the ionophore CCCP (white bars in Figure 4.4C), indicating that DAB2 is energetically-coupled, either directly or indirectly, to a cation gradient (e.g. H^+ or Na^+). This is consistent with previous observations that *Hnea* C_i uptake is powered by a membrane gradient [129].

DabA2 and DabB2 Interact to Form a Complex

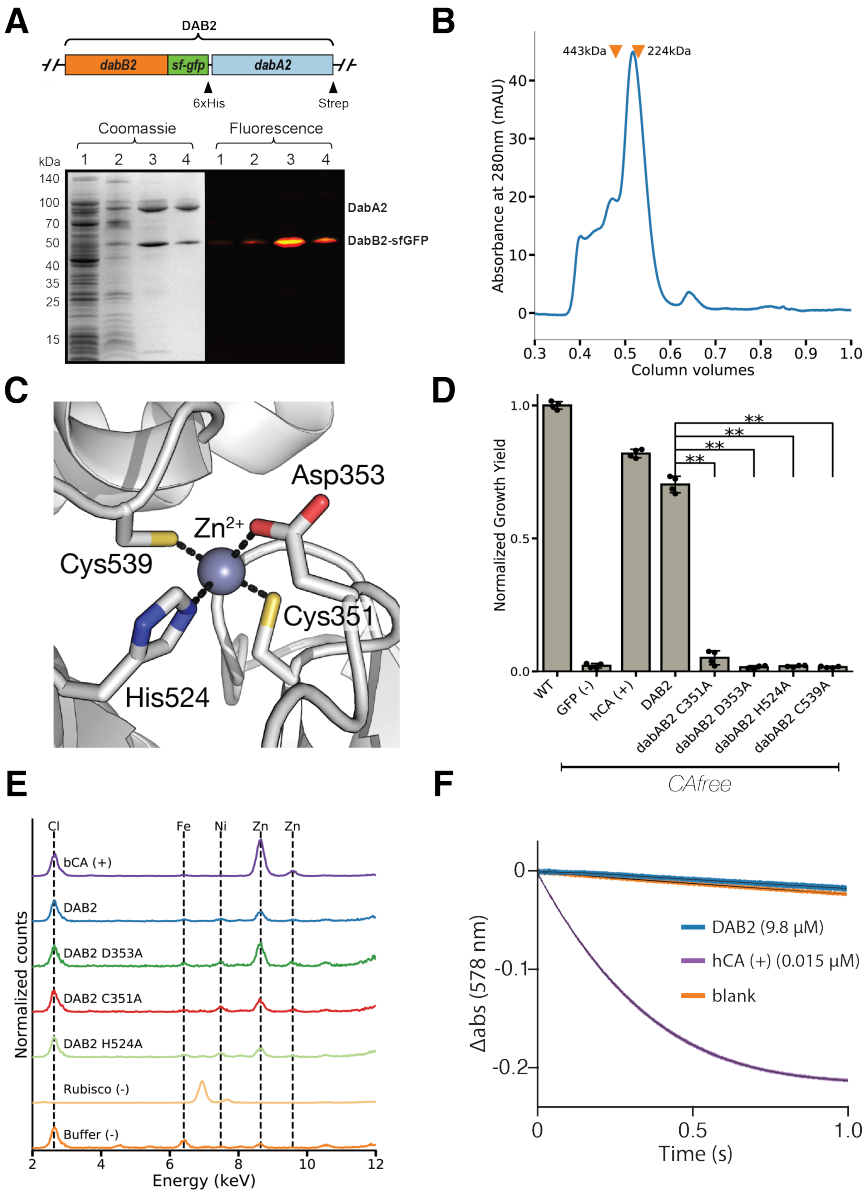


Figure 4.5: DabA function depends on on residues making up a predicted β-CA-like active site. (A) Purification of DabAB2 complex from *E. coli*. DabA2 was C-terminally tagged to a Strep-tag and DabB2 was C-terminally tagged with sf-GFP and a 6xHis-tag. Purification was monitored using SDS-PAGE imaged with fluorescence (right view) before coomassie staining (left view). Lane 1: clarified lysate; 2: solubilized membranes; 3: Ni-NTA resin eluent; 4: strep-tactin resin eluent. DabA2 and DabB2 co-purify as a single complex without any obvious interactors. (B) Size-exclusion chromatogram of His/Strep purified DabAB2 with retention volumes (orange arrows) and molecular weights (kDa) indicated for standard samples (apoferritin, 443 kDa; β-amylase, 224 kDa). DabAB2 runs with a mass of ≈270 kDa, which is likely an oligomer of DabA and DabB. (C) Structural model of the DabA2 active site based on a β-CA of *E. coli* (PDB 1I6P). Typical β-CAs rely on two cysteine and one histidine residues to bind Zn^{2+}. Asp coordinates Zn^{2+} but is likely displaced during catalysis [70]. (D) Alanine mutants of the putative DabA2 active site residues (C351A, t=54.3, $p < 10^{-7}$; D353A, t=144, $p < 10^{-10}$; H524A, t=44, $p < 10^{-7}$; C539A, t=44.3, $p < 10^{-7}$) abrogate rescue of CAfree *E. coli* compared to wild-type dabAB2. Error bars give standard deviations of four replicates. "*" denotes that means differ with Bonferroni corrected $p < 0.05$ by a two-tailed t-test, and "**" denotes $p < 5 \times 10^{-4}$. (E) X-ray fluoresence data indicate that DabAB2 binds zinc like all known β-CAs. Single mutations to the active site do not abrogate zinc binding. (F) Purified DabAB2 does not display any obvious CA activity despite being present in 650-fold excess over the positive control (Human carbonic anhydrase II, hCA) in our assays.

In order to determine if the genetic interaction between *dabA2* and *dabB2* reflects a physical interaction, we attempted to co-purify the two proteins. DabA2 was fused to a C-terminal Strep-tag, DabB2 was fused to a C-terminal GFP with 6xHis-tag, and the genes were co-expressed in *E. coli* (Methods). Tandem-affinity purification following detergent solubilization revealed that DabA2 and DabB2 form a complex in *E. coli* (Figure 4.5A). The complex runs as a single major peak on size exclusion chromatography and has a retention volume consistent with a heterodimer of DabA2 and DabB2 (Figure 4.5B). We did not observe co-purification of any *E. coli* proteins, suggesting that DAB2 operates as an independent complex within the membrane (Figure 4.5A). Moreover, *DAB2* expression rescues CAfree growth even when complex I is knocked out ($\Delta nuoA$-N), providing further evidence that DAB function is independent of complex I.

pH Independence of Rescue Suggests CO_2 is the DAB2 Substrate

Aqueous CO_2 spontaneously interconverts with the gas phase as well as hydrated C_i species (H_2CO_3, HCO_3^-, CO_3^{-2}). The equilibrium of $CO_2^{(aq)}$ and $CO_2^{(gas)}$ is not affected by pH, but the conversion from CO_2 to hydrated C_i is pH dependent. Thus, the equilibrium concentration of HCO_3^- increases 100 fold between pH 5 and 7 without an accompanying change in CO_2 concentration [193]. SbtA, a known HCO_3^- transporter, rescues CAfree growth at pH 7 but not at pH 5, while DabAB2 rescues growth at both pHs (Appendix). Since DabAB2 rescue is pH-independent in this range, its substrate is likely CO_2 and not H_2CO_3, HCO_3^-, or CO_3^{2-}. This is consistent with previous observations that CO_2 is the substrate of *Hnea* C_i uptake [129].

Requirement of Putative Zn^{2+}-binding Residues for DAB function

Structural homology modeling software predicts that the middle of DabA2 has sequence elements related to a β-CA (Figure 4.4A). Phyre2 predictions identify C539 and H524 as part of a potential Zn^{2+} binding site distantly homologous to a bacterial type II β-CA (10% coverage of DabA, 90.8% confidence). I-TASSER predicts a Zn^{2+} binding site including the same residues along with an additional cysteine (C351), and aspartic acid (D353). As shown in Figure 4.5C, these residues could make up the active site of a type II β-CA [69]. We generated individual alanine mutants for each of these putative active site residues (C351A, D353A, H524A and C539A) and tested them in CAfree. All mutants failed to rescue CAfree growth in ambient CO_2 (Figure 4.5D). We proceeded to assay zinc binding of purified DabAB2 complex using X-ray fluorescence spectroscopy and found that wild-type DabAB2 and three of the single mutants (C351A, D353A, and H524A) bind zinc (Figure 4.5E). Single mutants retain three of four zinc-coordinating residues[69], which could explain why the mutants bind zinc. Indeed, mutational studies of the human CA II show that single mutations to Zn^{2+}-binding residues reduce but do not abrogate Zn^{2+} binding [157].

Purified DabAB2 Complex does not have Conspicuous CA Activity

We tested whether detergent solubilized, purified DabAB2 displays carbonic anhydrase activity (Figure 4.5F). CA activity was not detected. DabAB2 was assayed at high protein concentrations (> 650-fold more protein than the positive control) and under CO_2 concentrations that are typically saturating for CAs, but displayed no activity (Figure 4.5F). Absence of activity *in vitro* argues either that DabAB2 has extremely low CA activity or, perhaps, that DabAB2 must reside in a cell membrane holding a cation gradient to function as an energetically-activated carbonic anhydrase.

Genomes of Important Human Pathogens Contain Functional DABs

Searching Uniprot for the DabA pfam (PF10070) yielded 878 putative DabA sequences. DabAs were found in a wide variety of prokaryotes including bacteria and archaea (Figure 4.6A and S9). Represented clades include not only Proteobacteria, but also Euryarchaeota, Firmicutes, Planctomycetes, and Bacterioides. Many *dabA* sequences were found in genomes of organisms that cannot fix CO_2 including the heterotrophic human pathogens *V. cholera*, *B. anthracis*, and *L. pneumophila* (Figure 4.6A). Notably, 843 (96%) of the identified *dabA* sequences were either within three genes of, or fused to, a *dabB*.

We assayed whether the DAB homologs from heterotrophic pathogens are functional C_i pumps. Operons from *V. cholera* E7946 El Tor Ogawa and *B. anthracis* Sterne were cloned and expressed in CAfree. Both DAB operons rescued CAfree growth in ambient CO_2 (Fig. 5B and S10). Thus, DAB operons from heterotrophic, human pathogens are functional.

4.4 Discussion

Here we generated a knockout library containing ≈35 individual knockouts for every gene in the genome of the proteobacterial chemolithoautotroph *H. neapolitanus*. Using these data, we compiled the first essential gene set for a chemolithoautotroph (Figure 4.1). We were able to confidently identify 551 essential genes and 1787 nonessential genes. Mapping essential genes will provide insight into the metabolism and growth physiology of sulfur-oxidizing chemolithoautotrophs.

In addition to mapping essential genes, transposon mutagenesis produced a library of ≈70,000 mutants with unique barcodes. These mutants were isolated in high CO_2 and so we were able to disrupt all known components of the bacterial CCM (Figure 4.2). We used the resulting genome-wide knockout library to perform the first comprehensive screen for novel bacterial CCM genes. This screen highlighted a small number of genes (about 17) as having the HCR phenotype associated with the CCM (Figure 4.2B-F), nearly all of which are known to be associated with the α-carboxysome. Though it is possible that genetic redundancy,

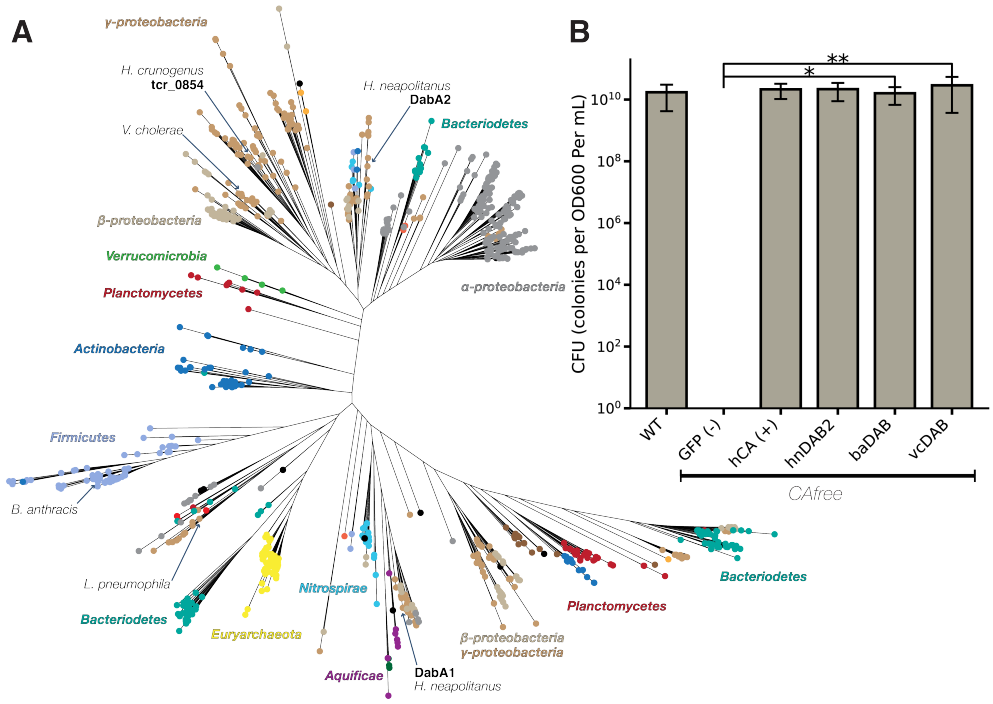


Figure 4.6: DAB operons are widespread among prokaryotes. (A) Approximate maximum likelihood phylogenetic tree of DabA homologs associated with PF10070.9 (Methods). DabA homologs are found in > 15 prokaryotic clades, including archaea. *Hnea* DabA1 and DabA2 represent two different groupings that are commonly found in proteobacteria. Inspecting the tree reveals several likely incidents of horizontal transfer, e.g. between Proteobacteria and Firmicutes, Nitrospirae and Actinobacteria. Moreover, the genomes of several known pathogens contain a high-confidence DabA homolog, including *B. anthracis*, *V. cholerae*, and *L. pneumophila*. Detailed annotations are given in the Appendix. Scale bar indicates one substitution per site. (B) Functional DABs are found in human pathogens. Colony forming units per OD600 per ml were measured on LB plates with induction in air. DAB operons from *B. anthracis* (baDAB, t=6.0, $p = 4 \times 2 \times 10^{-4}$) and *V. cholerae* (vcDAB, t=4.0, $p = 4 \times 10^{-3}$) rescued growth of CAfree cells. The *Hnea* operon DAB2 is abbreviated as hnDAB2. Error bars represent the standard deviation of 6 replicate platings for WT, GFP (-), hCA (+), and hnDAB2. Error bars represent standard deviations of 12 replicate platings for baDAB and vcDAB. "*" denotes that means differ with Bonferroni corrected $p < 0.05$ by a two-tailed t-test, and "**" denotes $p < 5 \times 10^{-4}$.

conditional phenotypes, or impairment only at sub-ambient CO_2 permit some genes to escape notice, we can nonetheless confidently say that the proteobacterial CCM is composed of < 30 distinct functional units. Moreover, none of the genes identified have unexpected functions, suggesting that current models of bacterial CCMs incorporate all necessary functions.

Our screen identified 3 transcriptional regulators as well as 3 distinct CCM operons (Figures 2B-F). Identification of transcriptional regulators with HCR phenotypes (Figures 2D-F) may inform the study of CCM regulation. The first operon contains nearly all known components of the α-carboxysome, all of which confer HCR phenotypes on knockout (Figure 4.2C). The second operon is adjacent to the carboxysome operon and contains 11 genes of which only 3 - the Rubisco chaperone *acRAF*, and the putative C_i transporter *dabAB1* - had HCR phenotypes (Figure 4.2C). The remaining 8 genes had no associated phenotype but might nonetheless have roles in the CCM. These genes include *cbbOQ*, *csos1D*, *p-II*, and a

parA homolog (Figure 4.2C). The third operon contains two genes, *dabAB2,* both with HCR phenotypes (Figure 4.2D).

Prior to this study, the *Hnea* C_i transporter was a unknown. *Hnea* C_i uptake was known to act on CO_2 in a manner coupled to the electrochemical potential across the cell membrane [129]. DAB1 and DAB2 were promising candidate C_i importers because they are homologous to C_i transporters in hydrothermal vent chemolithoautotrophs [183, 268, 267]. However, the mechanism of DAB C_i uptake systems has not been studied. We characterized the DAB2 operon to determine if DABs match *Hnea* transport characteristics.

We showed that the DAB2 operon encodes a two-component protein complex that has C_i uptake activity when heterologously expressed in *E. coli* (Figures 3B-C & 4A). This complex is likely a heterodimer as suggested by size-exclusion chromatography (Figure 4.5B). As C_i uptake is inhibited by the ionophore CCCP (Figure 4.4C), we suspect that DAB2 activity is energetically-coupled to a cation gradient (Figure 4.6A). Since DabAB2 shows pH-independent rescue of CAfree *E. coli* (pH 5-7) CO_2 is likely the transported substrate (Figure 4.5C). This is further supported by the fact that DabA has some homology to a type II β-CA and binds a zinc (Figures 3-4), suggesting that a CA active site hydrates transported CO_2. Mutations to the putative zinc-binding residues (C351A, D353A, H524A, and C539A) ablate function *in vivo* (Figure 4.5D). For these reasons, we propose a model of DAB activity wherein CO_2 is passively taken into the cell and then unidirectionally hydrated to HCO_3^- by energy-coupled CA activity of DabA. This model is consistent with previous observations of the *Hnea* C_i uptake system [129].

Model carbonic anhydrases are not coupled to any energy source (e.g., ATP, cation gradient). Rather, they equilibrate CO_2 and HCO_3^- [157, 296]. However, energy coupled CA activity could favor CO_2 hydration, allowing the DAB system to actively accumulate HCO_3^- in the cytosol and power the CCM (Figure 4.2A). Given the similarity of DabB to H^+-pumping proteins, we propose that DABs use the H^+ gradient, though our results are equally consistent with other cation gradients, e.g. Na^+. This mechanism would require tight coupling of cation flow to CA activity by DabA, consistent with our observation that purified DabAB2 displays no measurable CA activity. Interestingly, type II β-CAs are the only CAs known to display allosteric regulation [70]. Allosteric control is thought to be mediated by Zn^{2+} capping and uncapping by the active site aspartic acid (D353 in DabA2). A similar mechanism might couple cation flow through DabB to the active site of DabA.

Cyanobacteria possess vectorial CAs called CUPs, which may provide clues to the DAB mechanism [180, 230, 275]. Indeed, both DAB and CUP systems contain subunits in the Mrp protein family (DabB and NdhD/F are in PF00361) that also contains the H^+-pumping subunits of complex I. This commonality might suggest similar mechanisms. CO_2 hydration by CUPs is thought to be coupled to energetically-favorable electron flow because CUPs associate with complex I [32]. However, the Mrp protein family (PF00361) is very diverse and contains many cation transporters that do not associate with complex I or any other redox-coupled complex [159]. Moreover, DabB sequences are only distantly related to complex I and CUP subunits (Appendix), DAB2 subunits do not co-purify with *E. coli* complex I (Figure 4.5A) and DAB2 rescues CAfree growth in a complex I knockout (Appendix). We

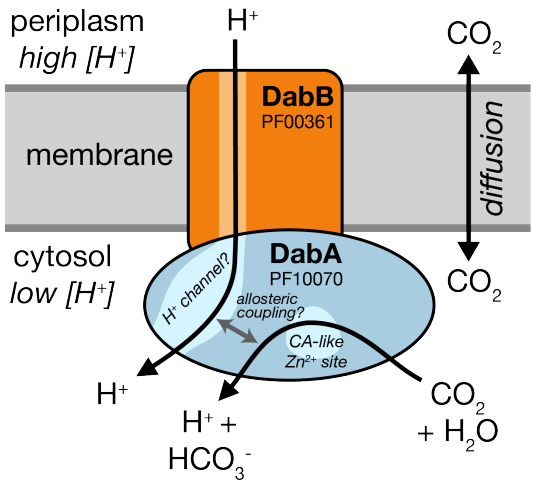


Figure 4.7: Speculative model of energetically-coupled DAB-mediated unidirectional CA activity. We propose that DabAB complexes couple CA activity of DabA to a cation gradient across the cell membrane, producing unidirectional hydration of CO_2 to HCO_3^-. The cation gradient could be H^+ or Na^+. Energy-coupled CA activity is required for the DABs role as a C_i uptake system in the proteobacterial CCM, as discussed in the text. Because it appears that DabAB2 is not active as a purified complex outside of the membrane, it is assumed protein tightly couples the inflow of cations with CO_2 hydration so that there is no "slippage." Indeed, slippage - i.e., uncoupled CA activity - would be counterproductive for CCM function [228]. Notably, Zn^{2+} binding by the active site aspartic acid of type II β-CAs (D353 in DabA2) is thought to allosterically regulate [70]. This Asp-mediated activity switch could, therefore, provide a means for allosteric coupling of a β-CA active site to distal ion transport.

therefore propose that DAB activity is not coupled to electron flow through complex I but, rather, to a cation gradient across the membrane (Figure 4.7).

DabAB2 functions robustly, as demonstrated by complementation of CAfree (Figure 4.4B) and ^{14}C uptake measurements (Figure 4.4C). Indeed, we observed that DabAB2 functions substantially better in *E. coli* than SbtA, a C_i transporter from cyanobacteria (Figure 4.4C). As *E. coli* and *Hnea* are proteobacteria, this observation could result from greater "compatibility" of proteobacterial proteins with *E. coli* expression. It may also be the case, though, that the α-CCM of proteobacteria is more "portable" than the β-CCM of freshwater cyanobacteria. Indeed, α-CCM genes are typically found in a single gene cluster in chemolithoautotrophs throughout α- β- and γ-proteobacteria and the α-CCM was clearly horizontally transferred from proteobacteria to marine cyanobacteria [238]. DabA homologs are widespread in prokaryotes and were likely horizontally transferred multiple times (Figure 4.6A). Since DAB complexes are prevalent among prokaryotes and have superlative activity, DAB-family transporters are an attractive target for protein engineering and heterologous expression in plants and industrial microbes, where elevated intracellular C_i could be useful [15].

Finally, DABs are present in a wide variety of bacteria [183, 267] as well as some archaea. We found high-confidence DabA homologs not only in large numbers of autotrophs but also in heterotrophs (Figure 4.6A & S9). Moreover, homologs are present in the notable heterotrophic pathogens *V. cholerae*, *B. anthracis*, and *L. pneumophila* (Figure 4.6A). We showed that DABs from *V. cholerae* and *B. anthracis* are active in *E. coli* (Figure 4.6B). This leads us to wonder: what do heterotrophic pathogens use C_i uptake systems for? Carbonic anhydrase activity is essential for growth of the heterotrophs *E. coli* and *S. cerevisiae* in ambient CO_2 [4, 191]. In the heterotrophic context, CA activity is thought to supply bicarbonate for biotin-dependent carboxylases in central metabolism, for which HCO_3^- is the substrate [4, 191]. Additionally, bicarbonate levels have been linked to virulence in both

V. cholera and *B. anthracis* [1, 285]. Perhaps DAB-family C_i uptake systems play roles in the growth or virulence of these important pathogens? We hope that future research will delineate the role of energetically-activated C_i uptake in heterotrophic and pathogenic organisms.

4.5 Materials and Methods

Bacterial Strains and Growth Conditions

E. coli strain APA766 was used as the conjugation donor to transfer the Tn5 transposon to *H. neapolitanus* C2 (*Hnea*) via conjugation [318]. The *E. coli* double CA deletion strain "CAfree" (BW25113 Δ*can* Δ*cynT*) was generated by curing the KEIO collection *cynT* knockout (BW25113 Δ*cynT*, KEIO strain JW0330) of kanamycin resistance via pCP20-mediated FLP recombination and subsequent P1 transduction (and curing) of kanamycin resistance from the *can* knockout strain EDCM636 (MG1655 Δ*can*, Yale Coli Genomic Stock Center, [17, 191]). Complex I knockout strains (Δ*nuoA-N*) were produced in both BW25113 and CAfree backgrounds. These strains were generated by lambda red mediated recombination of a Kan^R resistance cassette flanked by FRT sites into the nuo locus such that the entire operon was removed. The pSIM5 plasmid carrying the lambda red recombinase was heat cured at 42 °C. Lysogeny broth (LB) and LB agar were used as *E. coli* growth media unless otherwise specified. *E. coli* strains were grown at 37 °C in the presence of 0.1 mg/ml carbenicillin, 0.06 mg/ml kanamycin, or 0.025 mg/ml chloramphenicol as appropriate. *Hnea* was grown in DSMZ-68 media at 30 °C and in the presence of 0.03 mg/ml kanamycin when appropriate.

Transposon Mutagenesis and RB-TnSeq Library Generation

A barcoded library of *Hnea* transposon mutants was generated by adapting the methods of Wetmore *et al.[318]*. Conjugations were performed as follows. *Hnea* and APA766 were cultured and harvested by centrifugation. Both cultures were washed once in 10 mL antibiotic-free growth media per conjugation reaction and resuspended in 100 l. 5 OD600 units of *Hnea* were mixed with 20 OD600 units of APA766 on a 0.45 μM Millipore MCE membrane filter and cultured overnight at 30 °C in 5% CO_2 on an antibiotic-free LB agar plate containing 0.06 mg/ml diaminopimelic acid. Cells were scraped from the filter into 2 mL DSMZ-68 and collected in a 2 mL microcentrifuge tube. Recovered cells were pelleted by centrifugation at 16000 x g for 1 minute, washed in 2 mL DSMZ-68, pelleted again at 9000 x g for 1 minute, and resuspended in 2 ml DSMZ-68 before 200 μl was plated onto 10 separate DSMZ-68 kanamycin plates (per conjugation). Plates were incubated at 30 °C under 5% CO_2 until colonies formed (≈ 7 days). Colonies were counted and scraped into 55 mL DSMZ-68. Two 1.4 OD600 unit samples were taken and used to prepare genomic DNA (Qiagen DNeasy blood and tissue kit). Transposon insertions were amplified from gDNA and trans-

posons were mapped after Illumina sequencing using protocols and software from Wetmore *et al.[318]* 1.6 OD600 unit aliquots were then flash frozen in 50% glycerol for subsequent BarSeq experiments.

Essential gene assignment

Following the logic of Wetmore *et al.* [318] and Rubin *et al.* [255], we categorized genes as essential if we observed significantly fewer transposon insertions than would be expected by chance. If insertion occurred uniformly at random, the number of insertions per gene would be expected to follow a binomial distribution. The probability of observing at most k insertions into a gene of length n is therefore expressed as:

$$P(k; n, p) = \sum_{i=0}^{k} \frac{n!}{k!(n-k)!} p^i (1-p)^{n-i}$$

Here, p is the average rate of transposon insertion per base pair genome-wide. Genes were determined to be essential if they received a lower-than-expected number of insertions in both replicates of the library mapping, i.e. if the probability of observing k or fewer insertions was beneath 0.05 after Bonferroni correction. Genes were called "ambiguously essential" in two cases: (i) replicates were discordant or (ii) zero insertions were observed but the gene was short enough that the formula could not yield a Bonferroni-corrected p-value beneath 0.05 even in the case of zero insertions.

Gene Fitness Experiments

Fitness experiments were performed according to a modification of the protocol in Wetmore *et al.[318]*. This method allows pooled library fitness experiments to be performed comparing different growth conditions by comparing barcode abundance changes in order to track changes in the abundance of the transposon mutants. In short, a library aliquot was thawed and used to inoculate three 33 mL cultures. Cultures were grown to OD600 ≈0.08 in 5% CO_2. At this point, 20 mL were removed and harvested by centrifugation as two t_0 (input) samples. Cultures were back-diluted 1:64 into 128 mL and incubated for 6.5-7.5 doublings under 5% CO_2 or ambient conditions. 50 mL of culture was harvested by centrifugation. gDNA was prepared and barcodes were amplified for fitness determination via Illumina sequencing as described in Wetmore *et al.[318]*. Fitness values were calculated using existing software[318]. Genes were assigned an HCR phenotype if they had a fitness defect of two fold or greater in ambient CO_2 compared to 5% CO_2 in two replicate experiments.

CAfree Rescue Experiments

Electrocompetent CAfree cells were prepared using standard protocols and transformed with pFE plasmids expressing genes of interest by electroporation. CAfree pre-cultures were

grown overnight in 10% CO_2 and diluted into 96 well plates (3 μl cells in 250 μl media). Growth curves were measured by culturing cells in a Tecan M1000 microplate reader under ambient conditions with continuous shaking, and measuring OD600 every 15 minutes. When samples are marked "induced," 200 nM anhydrotetracycline (aTc) was added to the media. Growth yields are calculated as the maximum OD600 achieved after 24 hours of growth and normalized to the yield of a wild type control. CFU experiments were performed by back diluting cultures to OD600 0.2 before performing 10X serial dilutions. 3 μl of the OD600 0.2 sample and each of the serial dilutions were then spotted on plates with 200 nM aTc and grown overnight in ambient conditions (400 ppm CO_2). The spot with the highest dilution that yielded more than one colony was counted and a minimum of six replicates were averaged for each strain.

Silicone Oil Centrifugation Assay for Measurement of C_i Uptake

The silicone oil filtration method was modified from Dobrinski [78] and used to measure uptake of radiolabeled inorganic carbon. Assay tubes were generated using 0.6 ml microcentrifuge tubes containing 20 μl of dense kill solution (66.7% v/v 1 M glycine pH 10, 33.3% v/v triton X-100) covered by 260 μl of silicone oil (4 parts AR20:3.5 parts AR200). Electrocompetent CAfree cells were prepared using standard protocols and transformed with pFA-based plasmids containing genes of interest by electroporation. CAfree cultures were grown overnight in 10% CO_2, back diluted to an OD600 of 0.1 and allowed to grow to mid-log phase in 10% CO_2 in the presence of 200 nM aTc for induction. Cells were then harvested by centrifugation, washed once in PBS (pH 7.55) and resuspended to OD600 0.6 in PBS + 0.4% glucose. ^{14}C-labeled sodium bicarbonate (PerkinElmer) was added to a final concentration of 4.1 nM and an activity of 0.23 μC_i. Cells were incubated with ^{14}C for 4 minutes before centrifugation at 17,000 x g for 4 minutes to separate cells from buffer. Pellets were clipped into scintillation vials containing 5 ml Ultima Gold scintillation fluid and 300 μl 3M NaOH using microcentrifuge tube clippers or medium dog toenail clippers. Counts were measured on a PerkinElmer scintillation counter. ^{14}C counts are normalized to 1 OD600 unit of cells added. During inhibition assays, cells were incubated in PBS pH 7.55 with 0.4% glucose + 0.4% DMSO and the inhibitor (100 μM CCCP) for 10 minutes before assay.

Generation of DabA Phylogenetic Tree

We searched the Uniprot reference proteome database using the Pfam Hidden Markov Model PF10070.9 with a cutoff e-value of 10^{-4}. Our search recovered 941 candidate DabA proteins. These sequences were aligned using MAFFT and manually pruned to remove fragments and poorly aligning sequences. The remaining 878 candidate DabA sequences were re-aligned with MAFFT and an approximate maximum likelihood phylogenetic tree was constructed using FastTree. Taxonomy was assigned to nodes in the tree based on NCBI taxonomy information for the genomes harboring each sequence. Genomic neighborhoods for each gene in the tree were determined using the EFIGNT online server [334] and genomes with a

dabB gene within 3 genes of *dabA* and oriented in the same direction were considered to have full DAB operons. *dabAB* fusions were found by visual inspection of genomic neighborhoods from those genomes that did not have separate *dabB* genes located close to *dabA*.

Generation of DabB Phylogenetic Tree

DabB homologs were collected manually by searching MicrobesOnline for close homologs of four PF00361 members in the Hnea genome (*dabB1*, *dabB2*, *Hneap_1953*, *Hneap_1130*) and other characterized PF00361 members including Synechococcus *elongatus ndhF1*, Synechococcus *elongatus ndhF3*, and Synechococcus *elongatus ndhF4*. Genes were clustered to 95% similarity and genes with divergent operon structure were removed manually using MicrobesOnline treeview [76]. *nuoL* from *Escherichia coli*, *nqo12* from *Thermus thermophilus*, and *ndhF1/3/4* from *Thermosynechococcus elongatus* BP-1 were added as markers. ClustalOmega was used to construct a multiple sequence alignment and an approximate maximum likelihood phylogenetic tree was constructed using FastTree [233]. The tree was visualized using the Interactive Tree of Life [282].

Protein Annotation and Structural Homology Modeling

Secondary structural annotations for DabA and DabB were generated using XtalPred [286]. Structural Homology modeling of DabA was performed using Phyre2 and I-TASSER web servers with default parameters [148, 254]. A list of close DabB homologs was assembled by searching MicrobesOnline for PF00361 members with similar operon structure. A ClustalOmega alignment was used to calculate residue-level conservation of DabB proteins while the MAFFT alignment generated during the creation of the DabA tree was used to calculate residue level conservation of DabA proteins (Appendix).

Purification of DAB2

Chemically competent BL21-AI *E. coli* were transformed with a pET14b-based vector containing the *dabAB* genes. 1 liter of 2xYT media was inoculated with 20 ml of an overnight culture of BL21-AI E. coli in LB+CARB and allowed to grow to mid log at 37 °C. When midlog was reached, cells were induced with 20 ml of 50 mg/ml arabinose and transitioned to 20 °C for overnight growth. Cultures were pelleted and resuspended in 10 ml TBS (50 mM Tris, 150 mM NaCl, pH 7.5) supplemented with 1.2 mM phenylmethylsulfonyl fluoride, 0.075 mg/ml lysozyme and 0.8 ug/ml DNAse I per liter of starting culture and then incubated at room temperature on a rocker for 20 minutes. Cells were lysed with four passes through a homogenizer (Avestin). Lysate was clarified at 15,000 x g for 30 minutes. Membranes were pelleted at 140,000 x g for 90 minutes. Membrane pellets were resuspended overnight in 25 ml TBS supplemented with 1 mM phenylmethylsulfonyl fluoride and 1% β-dodecyl-maltoside (DDM, Anatrace) per liter of culture following [206]. Membranes were then re-pelleted at 140,000 - 200,000 x g for 60 minutes and the supernatant was incubated

with Ni-NTA beads (Thermo Fisher) for 90 min at 4 °C. The resin was washed with "Ni buffer" (20 mM Tris + 300 mM NaCl + 0.03% DDM, pH 7.5) supplemented with 30 mM imidazole and eluted with Ni buffer supplemented with 300 mM imidazole. Eluent was then incubated with Strep-Tactin (Millipore) resin for 90 min at 4 °C. Resin was washed with "strep buffer" (TBS + 0.03% DDM) and eluted with strep buffer supplemented with 2.5 mM desthiobiotin. Eluent was concentrated using Vivaspin 6 100 kDa spin concentrators and buffer exchanged into strep buffer by either spin concentration or using Econo-Pac 10DG (Biorad) desalting columns. For analytical purposes, 300 g of strep-purified protein was injected onto a Superdex 200 Increase 3.2/300 size-exclusion column pre-equilibrated in strep buffer and eluted isocratically in the same buffer.

Carbonic Anhydrase Assays

CA-catalyzed CO_2 hydration of purified DAB2 complex and human carbonic anhydrase (hCA) was measured using the buffer/indicator assay of Khalifah [152] on a KinTek AutoSF-120 stopped-flow spectrophotometer at 25 °C. The buffer/indicator pair used was TAPS/*m*-cresol purple measured at a wavelength of 578 nm using a pathlength of 0.5 cm. Final buffer concentration after mixing was 50 mM TAPS, pH 8.0 with the ionic strength adjusted to 50 mM with Na_2SO_4, and 50 μM of pH-indicator. Final protein concentration used was: 9.8 μM DAB2 (His-elution) and 0.015 μM hCA (positive control; Sigma Aldrich C6624). Saturated solution of CO_2 (32.9 mM) was prepared by bubbling CO_2 gas into milli-Q water at 25 °C. The saturated solution was injected into the stopped-flow using a gas-tight Hamilton syringe, and measurements were performed in a final CO_2 concentration of 16.5 mM. Progression curves were measured in 7 replicates.

X-ray Fluorescence Spectroscopy for Metal Analysis

50-100 g of protein in 20-200 μl of TBS + 0.03% DDM was precipitated by addition of 4 volumes of acetone and incubation at -20 °C for 1 hour. Samples were centrifuged at 21,130 x g for 15 minutes in a benchtop centrifuge and the supernatant was removed. Pellets were stored at 4 °C until analysis. Fluorescence analysis was performed by breaking up the pellet into 5 μl of TBS + 0.03% DDM with a pipette tip. Small pieces of the pellet were looped with a nylon loop and flash frozen at the beamline under a nitrogen stream. The sample was excited with a 14 keV X-ray beam and a fluorescence spectrum was collected. Sample emission spectra were then used to identify metals. Metal analysis was performed on wild-type DAB2, Zn-binding mutants C351A, D353A, and H524A, bovine CA (positive control; Sigma Aldrich C7025), and a buffer blank was used as a negative control. A Rubisco crystal containing cobalt salts was also used as a zinc free control. Displayed traces are averages of at least two experiments. Experiments were performed at the Lawrence Berkeley National Laboratory Advanced Light Source Beamline 8.3.1.

Acknowledgements

Thanks to Adam Deutschbauer, Morgan Price and Kelly Wetmore for assistance with RB-TnSeq experiments and analysis. Genomic DNA samples were kindly provided by Zoe Netter and Kimberly Seed (*V. cholera)* and Dan Portnoy and Richard Calendar (*B. anthracis* Sterne). Thanks to Spencer Diamond for assistance in producing phylogenetic trees and to Andreas Martin and Jared Bard for assistance with stopped flow experiments. Thanks to Emeric Charles, Woodward Fischer, Britta Forster, Ben Long, Luke Oltrogge, Robert Nichols, Dean Price and Patrick Shih for useful conversations and comments on the manuscript. X-ray-based experiments were performed at the Lawrence Berkeley National Laboratory Advanced Light Source Beamline 8.3.1.

Chapter 5

Synthetic Reconstitution of the Bacterial CCM in *E. coli*

5.1 Abstract

Ribulose Bisphosphate Carboxylase/Oxygenase (Rubisco) is the central carboxylase of the Calvin-Benson-Bassham (CBB) cycle and a notoriously frustrated catalyst. No known Rubisco has a maximum carboxylation rate above 15 s^{-1} and all Rubiscos also catalyze a competing oxygenation of their five-carbon substrate ribulose 1,5-bisphosphate (RuBP). It is widely hypothesized that the sluggishness and non-specificity of Rubisco promoted the evolution of diverse CO_2 concentrating mechanisms (CCMs). Though these various systems are functionally distinct and widely distributed among bacteria, eukaryotic algae and plants, they all serve the same basic purpose: elevating the concentration of CO_2 near Rubisco to increase the carboxylation rate and competitively inhibit oxygenation. Here we demonstrate the first transplantation of a functional CCM into a non-native host by expressing the α-carboxysome-based CCM of Halothiobacillus neapolitanus in an engineered *E. coli* strain that depends on Rubisco carboxylation for growth.

5.2 Introduction

Virtually all inorganic carbon (C_i) in the biosphere enters via Rubisco-based CO_2 fixation in the CBB cycle. Rubisco is an ancient enzyme that arose more than 2.5 billion years ago (Gya, ref) when Earths atmosphere contained virtually no O_2 and 20% CO_2. RuBisCO variants are often classified by maximum carboxylation rate ($k_{cat,C}$) and specificity $S_{C/O}$, a measure of the degree to which they prefer CO_2 as a substrate over O2. However, as discussed in Chapter 2, there is surprisingly little variation in these parameters, suggesting that the catalytic mechanism strongly constrains the evolution of this pivotal enzyme [301, 262]. CCMs accelerate Rubisco catalysis by simulating the atmosphere of early Earth in the

vicinity of the enzyme - that is, by elevating the CO_2 concentration near Rubisco so that oxygenation is competitively inhibited and carboxylation rate is maximized.

There are at least four distinct types of natural CCMs. These systems have evolved convergently multiple times and are found in diverse photosynthetic lineages ranging from cyanobacteria and aquatic algae to land plants (Chapter 1). Photosynthetic organisms with CCMs are responsible for more than half of the primary productivity on Earth [243] and some of the most productive human crops have CCMs (e.g. maize and sugarcane are C4 plants). Since CCMs accelerate Rubisco, there is a great interest in transplanting then into C_3 crop plants, which do not have a CCM. The carboxysome-based CCMs of cyanobacteria and proteobacteria are especially attractive from a biotechnological perspective because they are well-studied, function in single cells and require 20 genes for activity [171, 187].

Knockout of essential CCM components generates a high-CO_2 requiring (HCR) phenotype, meaning that the bacterial CCM is amenable to genetic characterization (refs). Work from several groups over decades has elucidated the molecular components of the bacterial CCM. These components include inorganic carbon (C_i) transporters and a 200+ MDa protein organelle called the carboxysome that co-encapsulates hundreds of Rubisco complexes with carbonic anhydrase (CA, Figure 1.7A). A genome-wide screen in our lab highlighted two adjacent operons in *H. neapolitanus* that are required for CCM function (Chapter 4). We sought to determine whether these operons are also sufficient for CCM function by expressing them in an *E. coli* strain (CCMB1) engineered to depend on RuBisCO carboxylation for growth.

5.3 Results

Construction of a Rubisco-Coupled *E. coli*, CCMB1

We used the OptSlope algorithm [15] to search the *E. coli* genome for central metabolic knockouts that are complemented by Rubisco during growth on specific carbon sources. As described in Figure 5.1A, OptSlope searched for knockouts wherein increasing biomass production requires carboxylation flux through Rubisco. We term these knockouts "Rubisco-coupled" mutants. Rubisco-coupled strains are distinct from "Rubisco-dependent" mutants wherein Rubisco expression is required for growth, but a very low Rubisco flux will suffice independe. This property is attractive because it implies that laboratory evolution experiments and suppressor screens will select for improvements to Rubisco expression and activity in such strains. A previously published Rubisco-dependent *E. coli* strain, $\Delta gapA$ [199], does not appear to be Rubisco-coupled in this manner (Figure 5.1B-C).

OptSlope identified many Rubisco-coupled knockout mutants (Figure 5.1B). We focused on ribose-5-phosphate isomerase (EC 5.3.1.6, *rpiAB* genes). We constructed the $\Delta rpiAB$ Δedd, which is complemented by Rubisco during growth on glycerol, xylose and gluconate minimal media. This can be explained simply for xylose: ribose-5-phosphate (Ri5P) is essential for production of nucleotides and an rpiAB deletion prevents Ri5P production

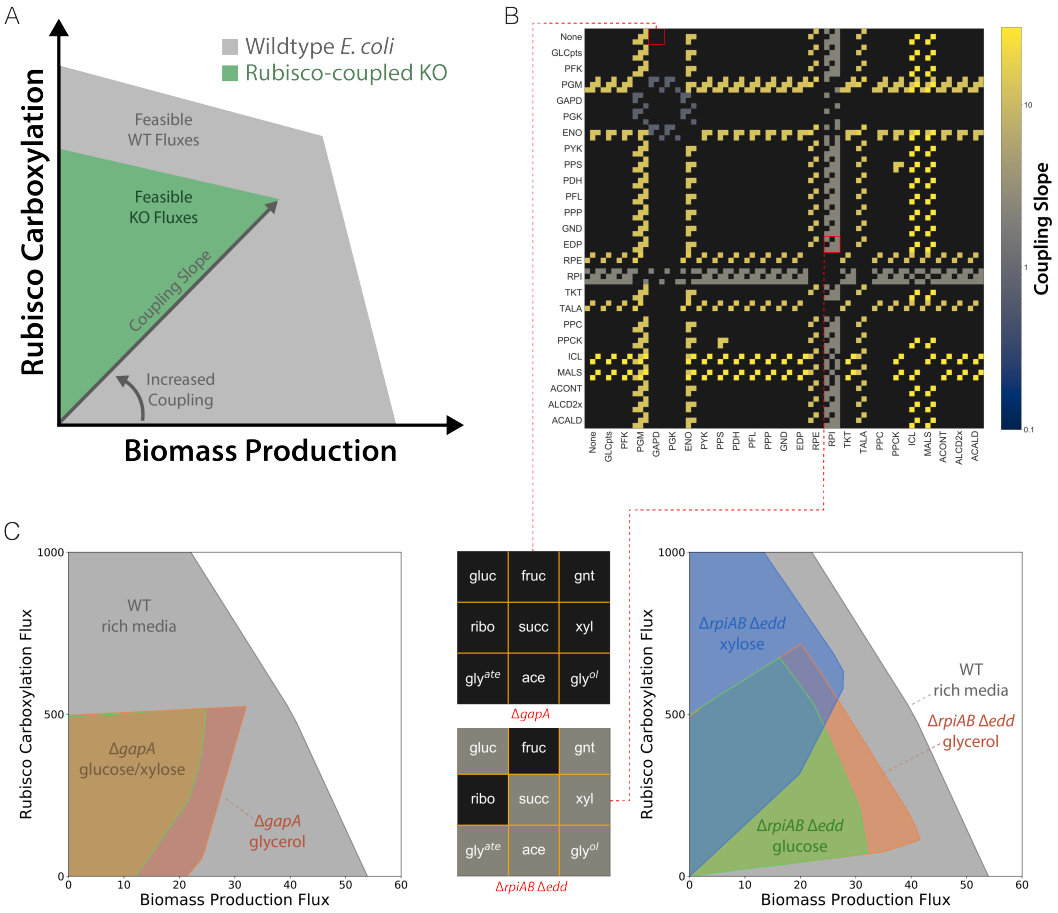


Figure 5.1: Use of the OptSlope algorithm for designing Rubisco-coupled *E. coli* strains. The OptSlope algorithm searches for metabolic knockout mutants in which the production of biomass is coupled to flux through a reaction of choice (e.g. Rubisco) at all growth rates. This is diagrammed in (A), which shows the phase space of feasible biomass production and Rubisco fluxes for wildtype (WT, grey) and a hypothetical knockout mutant (green). In WT, biomass production flux and, therefore, growth rate, is independent of Rubisco at all feasible growth rates (i.e. within the grey polygon). The mutant is "Rubisco-coupled" because maximal biomass production requires nonzero Rubisco flux and increasing biomass production demands increased carboxylation. The slope of this relationship is the "coupling slope." (B) We ran OptSlope on pairs of *E. coli* central metabolic knockouts and calculated the coupling slope for growth on 9 metabolically-distinct carbon sources: glucose, fructose, gluconate, ribose, succinate, xylose, glycerate, acetate and glycerol. Each double knockout is summarized in (B) with a 3×3 matrix of coupling slopes as shown in (C). Black denotes a mutant that is not Rubisco-coupled. (C) The published mutant $\Delta gapA$ [199] has a coupling slope of 0, meaning that the rate of biomass production is independent of the Rubisco carboxylation flux (left). Our chosen strain, $\Delta rpiAB\ \Delta edd$, is Rubisco-coupled on seven of the nine carbon sources (right). Coupling is predicted to be stronger in xylose than glycerol media, suggesting that modulation of the growth media can alter the stringency of selection for Rubisco carboxylation in this strain.

from pentoses like xylose. Expression of phosphoribulokinase (prk) and Rubisco opens a new route of Ri5P production via Rubisco carboxylation and gluconeogenesis (Figure 5.2A). Rubisco-dependent growth on glycerol can be explained by noting that rpiAB deletion forces 1:2 stoichiometry of Ri5P and xylulose 5-phosphate (Xu5P). Expression of Rubisco and prk permits arbitrary Ri5P:Xu5P stoichiometry (Figure 5.2B).

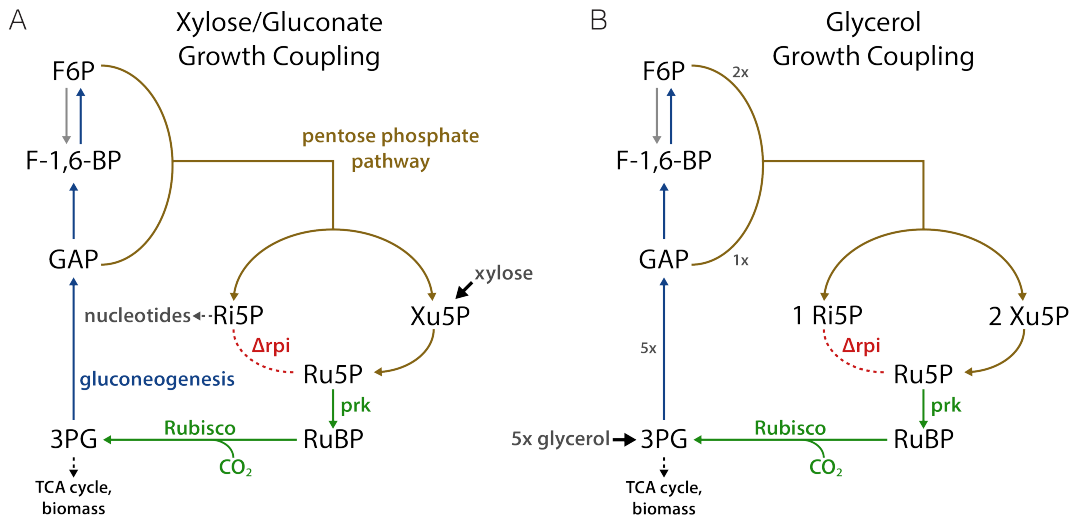


Figure 5.2: Model for coupling of growth to Rubisco carboxylation flux during growth on xylose, gluconate and glycerol. (A) Coupling of biomass production to carboxylation by Rubisco is easiest to understand in the context of growth on ribose as the sole carbon source. Ribose is a pentose that is metabolized via the pentose phosphate (PP) pathway. Δ*rpiAB* cannot synthesize ribose-5-phosphate (Ri5P) from xylose and Ri5P is required to produce nucleic acids for transcription and genome replication. Expression of Rubisco and *prk* enables growth by permitting Ri5P synthesis through gluconeogenesis the upper portion of the PP pathway. Since CCMB1 is also an *edd* knockout, it must metabolize gluconate via decarboxylation to ribulose-5-phosphate (Ru5P). Therefore, gluconate growth requires Rubisco and prk activity for the same reasons as xylose does. (B) Growth on glycerol also requires Rubisco. The *rpiAB* knockout mutant must produce Ri5P, but, due to the stoichiometry of the PP pathway, Ri5P production requires Ru5P production as well. The mutant cannot process Ru5P, however. Expression of prk and Rubisco enables Ru5P metabolism and unblocks biomass production.

We constructed BW25113 Δ*rpiAB* Δ*edd* by repeated transduction of kanamycin resistance markers from the KEIO collection of *E. coli* knockouts (Methods). Since cytosolic CA activity is incompatible with the CCM [228] we also knocked out both endogenous CAs in Δ*rpiAB*to generate the CCMB1 strain (BW25113 Δ*rpiAB* Δ*edd* Δ*cynT* Δ*can*). CCMB1 grows in a Rubisco-independent manner in rich media at elevated CO_2, but requires Rubisco and prk activity for growth on minimal media supplemented with glycerol as predicted (Figure 5.3C). CCMB1 does not grow in ambient CO_2 concentrations, i.e. it displays an HCR phenotype (Figure 5.3C), likely due to the low rate of Rubisco carboxylation in ambient CO_2. As evidenced by Rubisco-dependent growth under an anoxic mix of 10:90 CO_2:N2, carboxylation alone is sufficient for growth of CCMB1 and oxygenation is not required (Figure 5.3C).

Rubisco carboxylation also produces biomass precursors in complemented CCMB1. CCMB1 cells were transformed with pFE1A and grown in M9 xylose media supplemented with 100 mM $H^{13}CO_3^-$ in airtight growth vessels. The ^{13}C content of amino acids in a total metabolite extract was measured via high-performance liquid chromatography coupled to mass-spectrometry (HPLC-MS, Methods). Complemented CCMB1 was found to accumulate ≈ 50% more 13C into proteinogenic amino acids than the negative control, indicating that Rubisco actively fixes CO_2 in these cells. Altogether, these results imply that CCMB1 depends on Rubisco carboxylation for growth and biomass production in minimal media and that

oxygenation by Rubisco is not required for growth rescue.

Expression of CCM Components from *H. neapolitanus*

We generated two plasmids, pFECB and pFACCM, that together express all known components of the *Hnea* CCM under an anhydrotetracycline (aTc) inducible promoter $P_{LtetO-1}$ (Methods). pFECB expresses all 10 *Hnea* α-carboxysome genes [44], including large and small subunits of Rubisco, along with a prk, which generates the Rubisco substrate RuBP (Figure 5.4A). In *Hnea*, a second operon adjacent to the carboxysome operon contains several CCM genes, including Rubisco chaperones and a C_i transport complex (Chapter 4). We cloned this 12 kb operon into a compatible plasmid backbone to generate pFACCM (Figure 5.4A, Methods).

Since pFECB harbors both Rubisco and prk, it was expected to complement CCMB1 growth in elevated CO_2. Indeed, *Hnea* Rubisco complements CCMB1 when expressed separately from the carboxysome operon (Figure 5.3A). We did not observe growth initially, however. Plating on minimal media supplemented with glycerol and, subsequently, with gluconate (both under 10% CO_2) was used to select for Rubisco-dependent growth. Gluconate suppressor colonies were picked, plasmid was extracted and transformed into naive CCMB1 to verify plasmid-linkage of growth. Whole plasmid sequencing identified point mutations in regulatory regions of pFECB plasmids isolated from suppressor colonies. We chose to proceed with the plasmid isolated from gluconate colony 9 (termed pFECB.9).

Selection for Rubisco-Dependent Growth in Ambient Air

CCMB1 was co-transformed with pFECB.9 and pFACCM and plated on M9 glycerol media under ambient air (Methods). As a negative control, CCMB1 was co-transformed with pFE1A and pFACCM and plated in the same conditions. Both strains grew extremely poorly under ambient air, with a light patina of cells visible on titer plates. After $\approx$ 20 days, suppressor colonies appeared on the experimental plates but not for the negative control (Figure 5.4B-C). Suppressor colonies were picked, genotyped to ensure that they were not contaminants (Methods), and plated again in ambient CO_2. Of the 40 colonies tested, 12 regrew in ambient air to some degree (six exemplars are shown in Figure 5.4D.

6 of the 12 colonies were chosen based on a qualitative assessment of regrowth. These strains were grown in rich media under elevated CO_2 to purify plasmid DNA, which was then transformed into naive CCMB1 cells. Retransformation of plasmid DNA from suppressor colony 4 (henceforth s.4) conferred reproducible plasmid-linked growth (Figure 5.4E) and so we proceeded with this sample. Retransformants were plated for single colonies under 10% CO_2, picked into liquid M9 glycerol medium and incubated under ambient lab air. Colony 13 (s.4.r.13 denoting retransformant 13 of suppressor 4) was found to grow best in liquid media (Figure 5.4F). Carboxysomes purified from air-grown s.4.r.13 had normal protein content and morphology (Figure 5.4G-H).

Illumina and Sanger sequencing of plasmid DNA purified from s.4.r.13 identified two point mutations and one large structural rearrangement in the two plasmids. Both plasmids contain a nonsynonymous mutation (E37A) to a glutamate residue of the Tet Repressor (TetR) also observed in the parent plasmid pFECB.9. Glu37 is involved in DNA-binding [216] and E37A mutants have been shown to affect the specificity of Tet operator binding [39], suggesting that this mutation relaxes TetR-mediated repression of the CCM operons. The pFECB plasmid from s.4.r.13 also carries a nucleotide substitution to the 3' half of the second Tet operator (TetO2 +8G> T) previously found in pFECB.9. This mutation produces an asymmetric second operator, O2, but it is unclear how this would affect repression because G8 is not directly contacted by the repressor in co-crystal structures [216]. Finally, pFACCM from s.4.r.13 acquired the high-copy colE1 origin from pFECB in place of its original medium-copy P15A origin [176]. Altogether, mutations observed appear to reflect selection for higher expression of CCM genes through increased plasmid copy number and reduced repression of gene expression.

Plasmid Linked Growth in Ambient Air

Subsequent to the selection experiments described above, we found that reduced antibiotic concentrations dramatically improve growth of complemented CCMB1 strains in all conditions tested. It is intuitive that excess concentrations of two anti-ribosomal antibiotics (kanamycin and chloramphenicol) would inhibit growth when expressing 22 genes on medium and high-copy plasmids [29, 196, 269]. Based on these data (not shown), we chose to omit kanamycin from growth media. The pFE family plasmids carrying kanamycin resistance also uniformly carry Rubisco in this work, and Rubisco is required for CCMB1 growth on M9 media (Figure 5.3). We also reduced chloramphenicol to half of our lab standard concentration (12.5 mg/L instead of 25, Methods).

Reduced antibiotic was sufficient to prevent contamination while greatly improving growth in ambient CO_2. We found that co-transformation of parental pFECB.9 and pFACCM permitted growth of naive CCMB1 in ambient air. Cultures grew in ambient air within 2 days and without any need for suppressor screening (Figure 5.5). By generating targeted mutations to CCM genes, we showed that air growth depends on the known essential CCM features. An inactivating Rubisco mutation (pFECB.9 *cbbL*:K194M) abrogates growth, as does an inactivating mutation to the carboxysomal CA (pFECB.9 csosCA:C173S). Deletion of the Rubisco-binding N-terminal domain of *csos2* (pFECB.9 *csos2*:ΔNTD), a disordered protein that is involved in biogenesis of the carboxysome interior [58]. pFACCM is also required for growth: co-transformation of pFECB.9 with a pFA vector control does not permit growth of CCMB1 in ambient lab air (Figure 5.5B).

5.4 Discussion

Here we demonstrate that expression of 20 genes from *H. neapolitanus* produces the hallmark structures and phenotypes of the CCM in an *E. coli* strain, CCMB1, with an artificial dependence on carboxylation by Rubisco. The reconstituted CCM behaves similarly to native CCMs: Rubisco and CA activity are required for growth, as are carboxysome structures and an inorganic carbon transporter. As such, this work represents a *bona fide* reconstitution of the bacterial CCM. CCMB1 *E. coli* enables several lines of future research that are worth pursuing.

As shown in Chapter 4, only three fourths of the 20 genes expressed here have knockout phenotypes in *H. neapolitanus*. It is unlikely that all 20 are necessary for CCM function. CCMB1 should be used to clarify which genes are required for CCM function. *acRAF* and *cbbOQ* are involved in maintenance of Rubisco activity, but *cbbOQ* knockouts have no apparent phenotype in *H. neapolitanus* (Chapter 2). DAB gene knockouts do have an HCR phenotype in *H. neapolitanus*, but the mechanism by which they transport C_i remains mysterious. Detailed genetic characterization of the Rubisco chaperones *acRAF* [319] and *cbbOQ* [198] and the DAB-family C_i transporters (Chapter 4) in CCMB1 would be of great interest.

The bacterial CCM is found in two convergently-evolved forms featuring the α- and β-carboxysome, respectively. α and β-carboxysomes are morphologically similar but formed of distinct proteins that are not closely related [238]. *Hnea* expresses α-carboxysomes, but many freshwater cyanobacteria express β-carboxysomes. CCMB1 can be used to produce a similar reconstitution of the β-carboxysome CCM to investigate whether the two systems operate on the same principles and if components are interchangeable. These studies would be particularly useful for interrogating C_i transporters, which can be challenging to express heterologously.

Since it is challenging and time-consuming to integrate large DNA fragments into plant genomes, bacterial reconstitution will be crucial in evaluating CCM components and minimal CCM constructs prior to plant expression [82, 165, 171]. CCMB1 will provide a means of testing CCM expression constructs prior to expression in model plants. Importantly, bacteria and plants have substantially different physiology, which raises the question of whether heterologous expression of the bacterial CCM would in fact improve plant growth [187]. Physiological experiments using CCMB1 can be used to refine existing models of the bacterial CCM [187] and evaluate whether CCM expression can be expected improve plant growth.

5.5 Materials and Methods

Computational Design of RuBisCO-Dependent Strains

We used a variant of flux balance analysis (FBA) called OptSlope [15] to design central metabolic mutants in which cell growth is strongly coupled to Rubisco. We considered combinations of 1-3 central metabolic knockouts and removed any that permit growth without Rubisco and phosphoribulokinase activity. For the remaining putative Rubisco-dependent knockout mutants, we evaluated the degree of coupling between Rubisco flux and biomass production during growth under various carbon sources. This approach highlighted several candidate mutants, including ΔrpiAB. Candidates were evaluated in the context of a genome-scale model of *E. coli* metabolism to determine if growth coupling is due specifically to Rubisco carboxylation and not oxygenation. *E. coli* can, in principle, metabolize the oxygenation product 2-phosphoglycolate (2PG) by dephosphorylation to glycolate and oxidation to glyxolate, but this pathway is not constitutively expressed [221, 222]. Deletion of 2PG phophatase (PGPase) has no noticeable growth effects [222] but greatly reduced electroporation efficiency in our hands (data not shown). We reasoned that a 2PG salvage pathway would ultimately be required, as it is in cyanobacteria [84], and so we chose to proceed with 2PG metabolism intact. Source code is available at `https://gitlab.com/elad.noor/optslope`.

Strains and Genomic Modifications

Unless otherwise noted, all knockout strains are derived from *E. coli* BW25113, the parent strain of the KEIO knockout collection [17]. The strain we term "Δ*rpiAB*" has the genotype BW25113 Δ*rpiAB* Δ*edd* and was constructed by repeated rounds of P1 transduction from the KEIO collection followed by pCP20 curing of the kanamaycin selection marker [17, 72]. CCMB1 has the genotype BW25113 Δ*rpiAB* Δ*edd* Δ*cynT* Δ*can* and was constructed from Δ*rpiAB* by deleting both native carbonic anhydrases (*can* and *cynT*) using the same methods, first transducing the KEIO Δ*cynT* and then Δ*can* from EDCM636 [191]. CCMB1 was expected to depend on Rubisco for growth in xylose and glycerol media for the reasons described in Figure 5.2. Because of the *edd* deletion, gluconate metabolism must proceed entirely through the pentose phosphate pathway and so gluconate metabolism in CCMB1 is nearly equivalent to xylose metabolism. Transformations were performed by electroporation and electrocompetent stocks were prepared using standard protocols.

Heterologous Expression of CCM Components

pFE21 and pFA31 are compatible vectors derived from pZE21 and pFA31 [176]. pF vectors were modified to express the Tet Repressor (TetR) under a constitutive promoter so that expression is repressed in the absence of aTc. Phosphoribulokinase expression is deleterious to *E. coli* because it does not natively metabolize RuBP [326]. We found that repressed

plasmids pFE21 are pFA31 aid in cloning genes like *prk*. Cloning was performed by Gibson or Golden-Gate approaches as appropriate. Large plasmids like pFECB and pFACCM were verified by Illumina resequencing using BreSeq software [75]. When necessary, Sanger sequencing of PCR amplicons was used to verify regions of shared sequence (e.g. TetR) in mixed pools of pFECB and pFACCM plasmids purified from suppressor colonies, for example.

Growth Conditions

Unless otherwise noted, cells were grown on M9 minimal media supplemented with 0.4% v/v glycerol and 0.5 mg/L thiamin. 100 nM anhydrotetracycline (aTc) was used in induced cultures. Chloramphenicol and kanamycin antibiotics were supplied at 25 mg/L and 60 mg/L respectively, unless otherwise noted. Where "reduced antibiotic conditions" are invoked, it denotes 12.5 mg/L chloramphenicol and 0 mg/L kanamycin. Agar plates were incubated at 37 °C in a humidified gas-controlled incubator (Eppendorf New Brunswick s41i) under 10% CO_2, unless otherwise noted. Growth curves were obtained by culturing cells 96-well plates incubated at 37 °C in a gas-controlled shaking incubator with a refillable humidity cassette (Tecan Spark). In experiments in which frozen bacterial stock was used to inoculate the culture, cells were first streaked on M9 agar plates and incubated at 10% CO_2 to facilitate fast growth. Precultures derived from single colonies were grown in 5 mL liquid M9 under 10% CO_2 with a matching 1 mL control under ambient CO_2. Negative control strains unable to grow in M9 (e.g. K194M active site mutant of Rubisco) were streaked on and precultured in rich media (LB) under 10% CO_2. Precultures grown in rich media were pelleted and washed once in double the culture volume of M9 medium with no carbon source. Precultures were diluted to an optical density (600 nm) of 0.2 prior to inoculation or serial dilution for titer plates. Anaerobic growth was achieved by incubating agar plates in a BD BBL GasPak anaerobic system. The GasPak was flushed 5-6 times, replacing atmosphere with a 10:90 mix of CO_2:N_2.

Carboxysome Purification

Cells were grown in 1 L M9 cultures under ambient CO_2. Negative controls were grown in the same media under 10% CO_2. Cells were harvested by centrifugation at 5000g for 20 min. The cell pellet was resuspended in 50 ml buffer TEMB [10 mM Tris, 10 mM $MgCl_2$, 1 mM ethylenediaminetetraacetic acid, and 20 mM Na^+ HCO_3^- (pH 8.4)] with 1 mM phenylmethylsulfonyl fluoride, 0.1 mg/ml lysozyme (Sigma-Aldrich), and 25 U/μl benzonase (EMD Millipore). Cell lysis was performed by passing the cell suspension through an Avestin EmulsiFlex-C3 homogenizer three times. The crude cell extract was centrifuged at 12,000g for 30 min to remove cell debris. The supernatant was subsequently centrifuged at 40,000g for 30 min to pellet the carboxysomes. The pellet was then resuspended in 20 ml of 33% cellLytic B (Sigma-Aldrich) in TEMB buffer in order to remove lipid contaminants and centrifuged again at 40,000g for 30 min. The pellet was resuspended in 3 ml TEMB and

centrifuged at 3000g. The supernatant was then applied to 25 ml sucrose gradient made from 10%, 20%, 30%, 40%, and 50% sucrose. The gradient was centrifuged at 105,000g for 30 min. Fractions of 1 ml were collected and analyzed for the presence of carboxysomes via SDS-PAGE gel. The fractions that contained carboxysomes were pooled and centrifuged at 105,000g for 90 min to pellet the carboxysomes. Carboxysomes were then resuspended in 1 ml TEMB and stored at 4 °C.

Metabolite Extraction for Isotope Enrichment Measurements

Cells were grown to stationary phase in sealed airtight culture tubes containing 10 mL M9 minimal media supplemented with 4 g/L xylose, 100 nM aTc and 100 mM HCO_3^-, $H^{12}CO_3^-$ or $H^{13}CO_3^-$ as appropriate. The negative control strain in Figure 5.3D is a strain we term "CAfree," which has the genotype BW25113 Δcan $\Delta cynT$. This strain, which is described more fully in Chapter 4, was chosen because it lacks all native CA enzymes but does not depend on Rubisco for growth in minimal media at elevated CO_2. 1.5 mL of culture was pelleted by centrifugation at 13,000 × g for 1 min at room temperature. Cell pellets were quenched with 250 μl methanol, mixed by pipetting and stored at -20 °C. After thawing on ice, 250 μl ultrapure water was added and the mixture was centrifuged at 13,000 × g for 10 min at 4 °C. 450 μl of the supernatant was then transfered to a 3 kDa cutoff filter (Amicon Ultra) and filtered by centrifugation at 13,000 × g for 90 minutes. 400 μl of filtrate was mixed with 600 μl water in a 2 ml microcentrifuge tube and lyophilized for 24-48 hours to concentrate the sample. Samples were reconstituted in 100 μl 1:1 methanol:water before analysis.

HPLC-MS Isotope Enrichment Measurements

Amino acids were separated by the hydrophilic interaction liquid chromatography method described by Dong et al. [79] and were detected via electrospray ionization time-of-flight mass spectrometry as described by Bokinsky et al. [43]. Data processing were performed using the El-Maven software package [34]. Amino acids presented in Figure 5.3 were selected by manual inspection of peaks and comparison to both standards and blanks. If peaks aligned well with standards and were sufficiently above background, the mass isotopomer distribution of the amino acid was obtained without fragmentation. Labeling patterns for the following intracellular amino acids were considered: alanine (Ala), glycine (Gly), glutamate (Glu), lysine (Lys), phenylalanine (Phe), and proline (Pro).

Acknowledgements

We would like to thank Matt Davis for P1 transduction materials and advice, Hernan Garcia and Han Lim for supplying pZ plasmids, Maggie Stoeva, Anna Engelbrektson, Anchal Mehra, Sophia Ewens and Tyler Barnum for help with anaerobic growth conditions, and Rob Egbert and Adam Arkin for KEIO strains. Thanks to Dan Arlow, Arren Bar-Even, Dan Davidi,

Eric Estrin, Shmuel Gleizer, Luke Oltrogge, Noam Prywes, Patrick Shih, Elad Noor, Jeremy Roop, Rachel Shipps, Dan Tawfik, Morgan Price for support, advice and helpful discussions throughout this work.

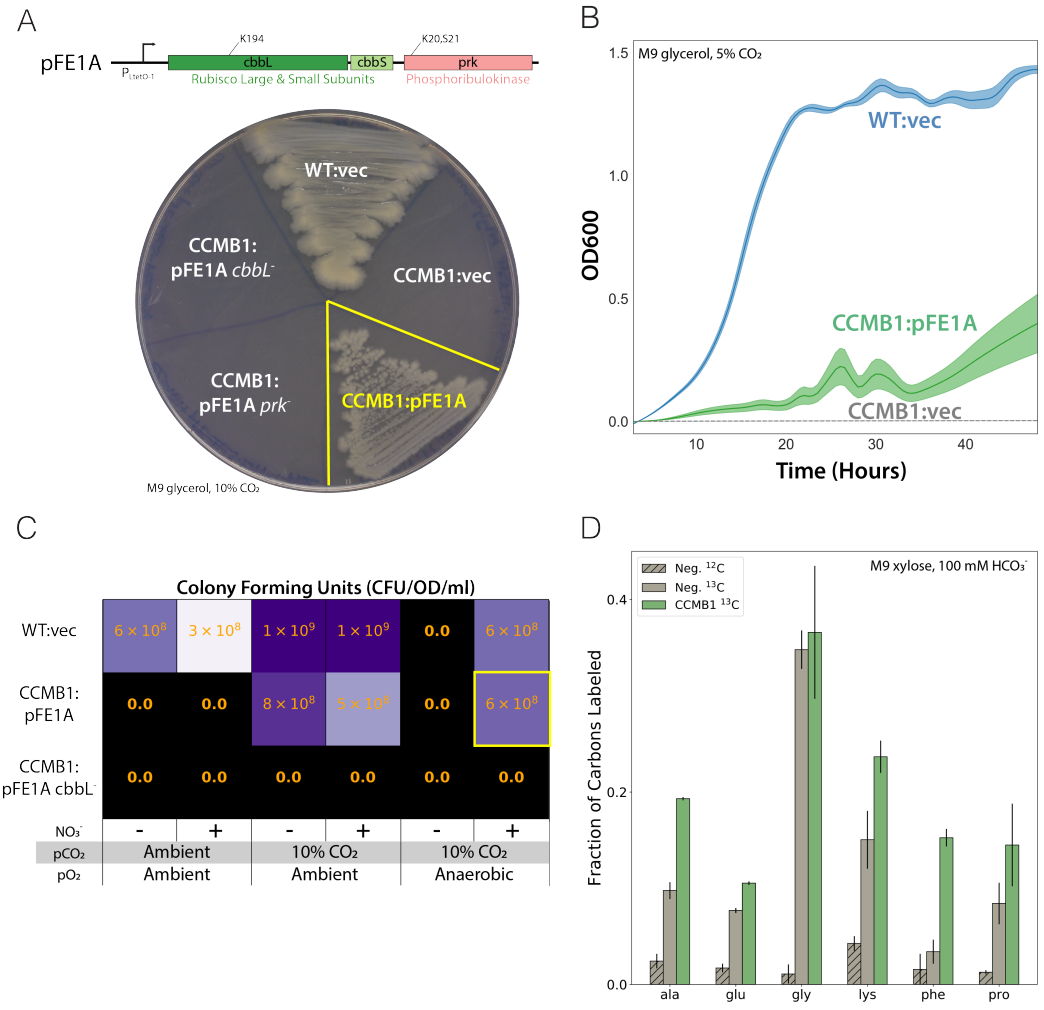


Figure 5.3: CCMB1 depends on Rubisco carboxylation for growth. (A) The plasmid pFE1A harbors the large and small subunits of the *H. neapolitanus* Rubisco along with a *prk* gene. Expression of these genes in CCMB1 rescues growth on solid M9 glycerol media in 10% CO_2. A catalytically inactive Rubisco mutant (pFE1A *cbbL*$^-$ harboring *cbbL* K194M, [66, 54]) fails to complement CCMB1, as does a nucleotide-binding deficient *prk* mutant (pFE1A *prk*$^-$ harboring *prk* K20M S21A, [54]). pFE1A also enables growth in liquid media under elevated CO_2 (B) and in anaerobic environments (C). Panel (C) quantifies growth of WT *E. coli*, complemented CCMB1:pFE1A, and the negative control CCMB1:pFE1A *cbbL*$^-$ as colony forming units (CFU) per unit optical density (at 600 nm) per ml (CFU/OD/ml). WT cannot ferment glycerol [293, 294], but can grow anaerobically on M9 glycerol media supplemented with 20 mM NO_3^- as a terminal electron acceptor [307]. CCMB1:pFE1A grows anaerobically on glycerol in the presence of NO_3^- (C), indicating that O_2 is not required. Since CCMB1:pFE1A fails to grow in ambient CO_2 and O_2, O_2 is in fact inhibitory. (D) We confirmed that CCMB1:pFE1A performs net CO_2 fixation into biomass precursors during growth on xylose by examining the incorporation of ^{13}C, provided as $H^{13}CO_3^-$, into intracellular amino acids. CCMB1:pFE1A incorporated an average of 52% more ^{13}C than the negative control (Methods).

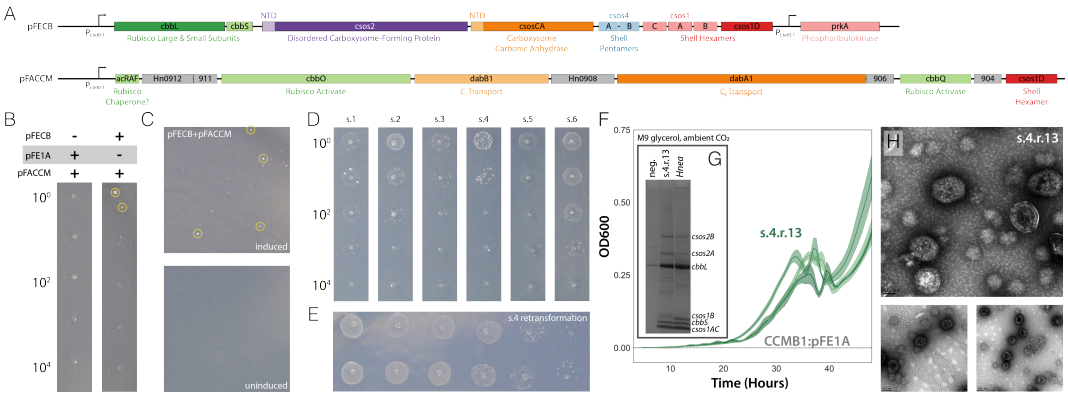


Figure 5.4: Expression of *H. neapolitanus* CCM components in CCMB1 *E. coli* produces suppressor mutants capable of growing in ambient CO_2. (A) pFECB and pFACCM plasmids encode all 20 distinct activities known to be required for CCM function in *H. neapolitanus*. pFECB expresses all 10 carboxysome genes, including the carboxysomal Rubisco (*cbbLS*) and carbonic anhydrase (*csosCA*). pFACCM expresses a secondary CCM operon that is directly adjacent to the carboxysome operon in *Hnea*. This operon contains putative (*acRAF*) and known Rubisco chaperones (*cbbOQ*) as well as a recently-characterized C_i transporter (*dabAB*). We co-transformed these plasmids into CCMB1 and plated titers on M9 glycerol media under ambient atmosphere. Neither strain grew well, suppressors emerged after 20 days of incubation, as shown in (B). Suppressors developed only from cultures expressing all 20 genes and not from negative control cultures lacking carboxysome genes (but expressing Rubisco and prk from pFE1A). (C) Appearance of suppressor mutants was also dependent on aTc induction of CCM genes. (D) Several individual suppressor colonies regrew in air. Pooled plasmid isolated from suppressor 4 (s.4) displayed plasmid-linked growth in ambient CO_2 when retransformed into naive CCMB1 on plates (E). Individual retransformants were picked and tested for growth in liquid M9 glycerol media. (F) Colony 13 (s.4.r.13) grew best in liquid media. The green lines in (F) represent three biological replicates and plot mean and standard error of measured optical density for 4 technical replicates. Growth was induction-dependent in liquid and solid media (data not shown). Carboxysomes isolated from s.4.r.13 are similar to native *H. neapolitanus* carboxysomes in protein content (G) and have normal morphology in electron micrographs of pooled sucrose gradient fractions (H, Methods). All growth assays in this figure were performed in ambient lab air.

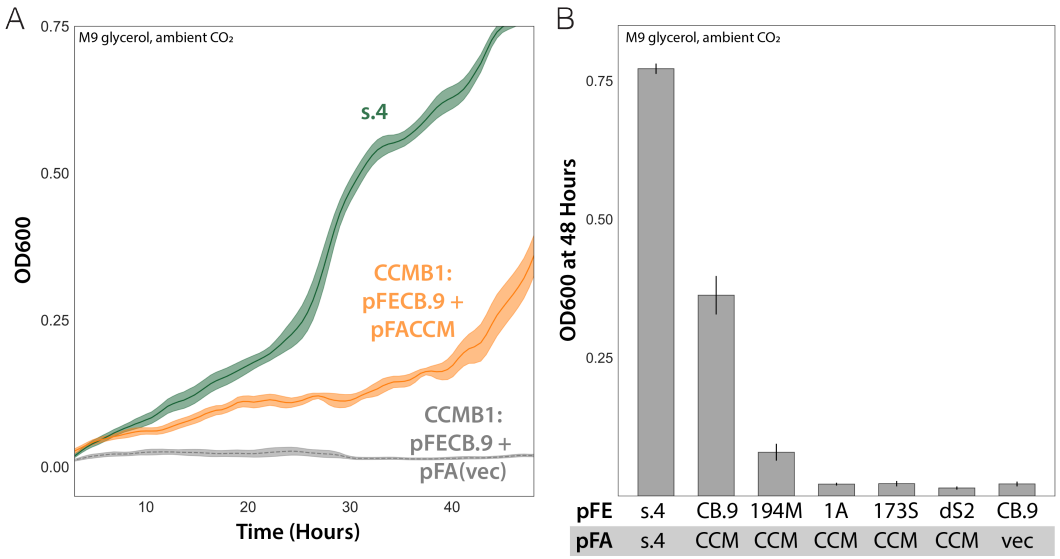


Figure 5.5: Reduced antibiotic permits growth of naive CCMB1 in ambient CO_2 without suppressor screening. Naive CCMB1 were co-transformed with pFECB.9 and pFACCM. (A) Co-transformation (orange) enables growth in ambient CO_2 that is comparable to the suppressor mutant s.4 (green). A negative control CCB1 strain (grey) carrying pFECB.9 and a pFA vector control is shown in grey for comparison. (B) Targeted mutations to carboxysome genes ablate growth. Growth is quantified in (B) as the mean optical density (600 nm) measured 48 hours post inoculation. Abbreviated names for pFE plasmids are: pFECB.9 (CB.9) pFECB.9 *cbbL*:K194M (194M), pFE1A (1A), pFECB.9 *csosCA*:C173S (173S), *csos2*:ΔNTD (dS2). All cultures in this figure were grown in ambient lab air in reduced antiobiotic conditions and 100 nM aTc (Methods). Positive control titers of the same cultures were plated on the same media under 10% CO_2, all of which grew except CCMB1: pFECB.9 *cbbL*:K194M + pFACCM (Appendix).

Chapter 6

Conclusions

6.1 Discussion

The work I presented here is framed by two core questions. First, I am curious why it is that CCMs evolved independently so many times throughout photosynthetic lineages. If we include C_2, C_4, CAM, carboxysome and pyrenoid CCMs, it appears that CCMs evolved independently more than 100 times [257, 317, 238, 311]. This striking case of multiple emergence begs for explanation. Considering the distribution of CCMs highlights that biophysical CCMs are nearly universally expressed by single-celled aquatic phototrophs, while biochemical CCMs are mostly endemic to land plants. This observation provokes a second question: is the distribution of CCMs are historical accident, or are the various CCM types uniquely suited to particular organismal physiologies. The fact that all chloroplasts descend from an ancestral cyanobacterium [185, 241, 278] makes this question even sharper: if all chloroplasts descend from cyanobacteria and all cyanobacteria that we know of have a carboxysome CCM, why don't land plants have carboxysomes?

In Chapter 2, I presented a suite of literature data on ≈ 300 natural Rubiscos. Based on these data, it appears that Rubisco carboxylation is subject to certain inherent limitations in present-day CO_2 and O_2 concentrations. In particular, it appears that Rubiscos that are more efficient at carboxylation are also more efficient at oxygenation. These data provide an answer to the first question, suggesting that CCMs represent common response to the inherent deficiencies of Rubisco. CCMs are thought to flood the Rubisco active site with CO_2, which would suppress the oxygenation reaction by competitive inhibition, making moot any tradeoff between carboxylation and oxygenation.

Thorough mathematical modeling of the bacterial CCM (Chapter 3) shows that this idea is plausible. That is, the known components of the bacterial CCM are sufficient, in principle, to saturate Rubisco with CO_2 to the point where the oxygenation reaction is negligible ($\leq 1\%$ of total). By accounting for the the inherent intertwining of physiology and pH, I was able to produce a model that is consistent with measured cyanobacterial carbon fluxes as well as basic physiological and biochemical intuition. This vetted model helps refine our intuition

about the bacterial CCM, showing that there is no need for carboxysomes to block O_2 entry [150, 237], nor is it necessary for carboxysomes to selectively take up HCO_3^- or selectively retain CO_2 in order for the CCM to function efficiently.

Based on the results I articulate in Chapter 3, it is clear that the CCM works in theory. But how does it work in practice? How many genes are required for the bacterial CCM to function in a native host? In Chapter 4 I present the results of a genome-wide screen in a native CCM host, *H. neapolitanus*. In addition to enumerating the first essential gene set for a bacterial chemolithoautotroph, this screen shows conclusively that ≤ 25 distinct genetic units are requires for the CCM to function *in vivo*. These results are very informative because no *in vivo* CCM reconstitution has been reported (prior to this work) and so it was unclear how many bacterial CCM genes remained unidentified. The screen presented in Chapter 4 confirms that CCM components are encoded in 3 distinct operons and highlights the importance of Rubisco folding chaperones [319, 5] and inorganic carbon transporters [82] to the functioning of the CCM.

Chapter 5 presents the first *in vivo* reconstitution of any CCM. This work was enabled by two innovations: a complete CCM "parts list" based on the results of Chapter 4, and a computational method that enables design of Rubisco-dependent *E. coli* mutants [15]. Our designed Rubisco-dependent *E. coli* strain, CCMB1, can grow in minimal media in elevated CO_2 only if Rubisco and phosphoribulokinase are expressed. Complemented CCMB1 *E. coli* mimick the phenotype of cyanobacterial and proteobacterial CCM knockouts - they are high-CO_2 requiring (HCR) presumably because Rubisco carboxylation is too slow to support growth in low CO_2. Heterologous expression of 20 *H. neapolitanus* CCM genes enables CCMB1 to grow in ambient lab air. Moreover, targeted mutagenesis demonstrates that air growth depends on known essential CCM features - Rubisco and CA activity are required, Rubisco must be localized to carboxysomes, and inorganic carbon transporters must be present to permit growth in ambient air.

These results delineate a set of 20 genes that are sufficient for CCM function *in vivo* in bacteria. However, the motivating questions I presented above remain unresolved. I believe that future work on reconstituted CCMs can help explain why CCMs evolved so many times independently and why particular families of CCMs are so strongly-associated with particular organismal physiologies. In the forgoing text I put forward speculative answers to these questions and highlight future avenues of research that may help bring clarity to the field.

Why Don't Plants have Carboxysomes?

All known cyanobacteria have a functioning carboxysome CCM [19, 238]. Chloroplast genome sequences make it abundantly clear that plant and algal chloroplasts derive from a free-swimming cyanobacterium by endosymbiosis [185, 241, 278]. Despite their sisterhood with cyanobacteria, no plants or algae produce carboxysomes. It is often suggested that transplanting the carboxysome CCM into plants would improve crop yields [187] and transplantation efforts are already underway [171]. Yet, why is it that plants do not already have

a carboxysome CCM if their chloroplasts were once cyanobacteria? There are, broadly, two plausible answers:

1. The carboxysome CCM evolved after the primary endosymbiosis event, or
2. Biophysical CCMs evolved prior to endosymbiosis but were lost because they did not aid in plant growth.

Environmentally-significant cyanobacteria - oceanic *Prochlorococcus* and freshwater *Synechococcus* species - express distinct and likely unrelated CCMs (i.e. α and β-carboxysomes, respectively). Moreover, the α-carboxysome appears to have been horizontally transferred from proteobacteria to cyanobacteria [16, 150]. As such, it is plausible that the carboxysome CCM evolved after the primary endosymbiosis and then came to dominate the cyanobacteria by a mixture of horizontal transfer and selection. This is my preferred explanation, though it is challenging to test by phylogenetic methods, as discussed below. Phylogenetic methods, moreover, will not resolve the core physiological question: could carboxysome CCMs improve plant growth?

We might similarly ask: why is it that so few land plants have pyrenoid CCMs? So far, the hornworts are the only terrestrial plants shown to have pyrenoids [311]. The fact the pyrenoid CCM evolved multiple times in extremely evolutionary distinct algae (e.g. in both green algae [106] and diatoms [153, 333]) suggests, in my view, that the pyrenoid is not found in most plants because biophysical CCMs are less appropriate for land plants than they are for aquatic microbes. I sketch this argument in two parts below. First, I argue from quantitative physiological principles that the presence of a biophysical CCM will greatly increase the the growth rates of small, aquatic phototrophs. Second, I argue that there is a fundamental difference between the action of biophysical and biochemical CCMs and, consistent with their observed dispersion, biochemical CCMs are more appropriate for large-bodied terrestrial plants. Before describing these arguments, I discuss why phylogenetic methods are unlikely to be helpful in resolving these questions.

Prospects of Ordinating CCM evolution by Phylogenetic Methods

Plant chloroplasts, which have no carboxysomes, are sister to a clade of cyanobacteria that uniformly express the β-carboxysome (Figure 1.3). Rubsicos from plant chloroplasts and β-cyanobacteria are collectively termed Form IB Rubiscos and are easily separated from other Form I Rubiscos by sequence alignment [11]. This fact is compatible with two models: either (a) the ancestor of chloroplasts had a β-carboxysome or (b) the β-carboxysome CCM evolved after the primary endosymbiosis. This former model is bolstered by the fact that the Form IB Rubiscos are sister to a clade of Form IA Rubiscos that is associated with the α-carboxysome CCM. If α- and β-carboxysomes are related, e.g. if the α-carboxysome was derived from the β, then the tree structure would strongly imply that the ancestral chloroplast had and lost a carboxysome [278].

However, it seems likely that α- and β-carboxysomes evolved convergently and not by a combination of descent and divergence. The shell proteins of α and β lineages are clearly distinct and they encapsulate carbonic anhydrases from different families that adopt very different tertiary structures [238]. Moreover, the proteins responsible for forming the carboxysome lumen in each family, csos2 and ccmM, are wholly unrelated in sequence and bind Rubisco by different mechanisms [314, 58].[1] As such, it remains uncertain whether the ancestor of all chloroplasts had a carboxysome.

It is unlikely that phylogenetic methods will resolve the historical ordering of carboxysome CCM evolution and primary plastid endosymbiosis because carboxysome CCM genes are not well-suited to detailed phylogenetic analysis (Table 1.2). The carboxysome CCM has two components: C_i transporters and the carboxysome itself. Multiple distinct families of C_i transporters are associated with the bacterial CCM [238], so the C_i transporters cannot reasonably be used as markers for the presence of the CCM. Carboxysome genes are often very short, have multiple paralogs, and are sometimes unstructured, highly variable repeat proteins (e.g. csos2, Figure 1.7B and Table 1.2). It is very challenging to produce meaningful phylogenies from such genes.

Phylogenetic analysis of pyrenoid CCM genes is similarly unlikely to be helpful in ordinating the evolution of pyrenoid CCMs. Pyrenoids arose by multiple evolution in distinct clades of eukaryotic phototrophs [113, 311], though it is unclear when those events occurred along the evolutionary trajectory between free-swimming algae and multicellular land plants. Moreover, the model pyrenoid of *C. reinhardtii* is formed by liquid-liquid phase separation [106] mediated by a highly disordered repeat protein, EPYC1. EPYC1 is not well-conserved in other pyrenoid-bearing algae [178] and is therefore unsuitable as a marker gene. Carbonic anhydrases (CAs) are another common component of algal CCMs [137] but CAs evolved convergently several times and are ubiquitous across the tree of life [287, 297]. The genetic determinants of model pyrenoid CCMs are not yet fully understood [162], but it seems unlikely that useful marker genes will be found to help trace their emergence.

Fast cyanobacterial growth depends on the CCM

Many experiments over multiple decades have shown that cyanobacteria require the CCM for growth in ambient CO_2 partial pressures [238]. Indeed, the carboxysome was discovered by screening for genes producing high-CO_2 requiring (HCR) phenotypes [330]. Chemotrophic proteobacteria bearing the α-carboxysome CCM show similar phenotypes when the carboxysome is knocked out - they grow robustly under elevated CO_2 but are not viable in ambient air [238]. In this section I argue that the purpose of the biophysical CCM is to enhance the per active site catalytic rate of Rubisco and, thereby, enable bacteria to express 5-10 times less Rubisco than would be needed in the absence of a CCM (Figure 6.1). Excess protein expression substantially attenuates bacterial growth, so it is advantageous for small, fast-growing organisms to express protein economically [29, 269, 335]. This argument does

[1] Also unpublished results from Luke Oltrogge, Allen Chen, and Cecilia Blikstad in the Savage lab.

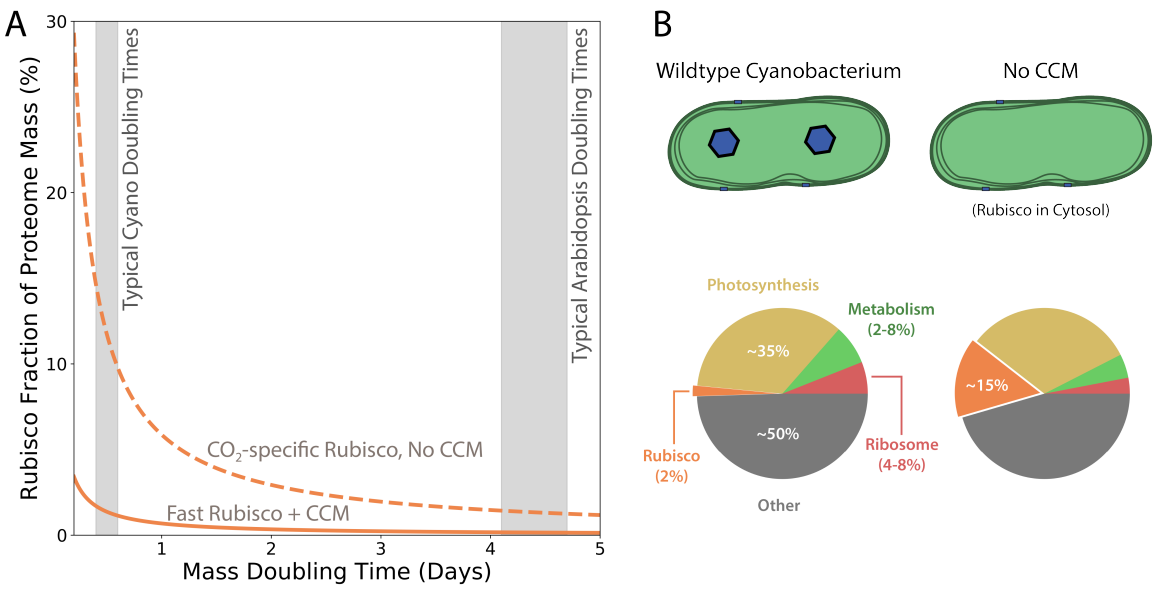


Figure 6.1: Presence of CCM greatly increases expected cyanobacterial growth rates. Knowing the CO_2 and O_2 concentrations, we can calculate the net carboxylation rate of a particular Rubisco. Assuming all cellular carbon is produced by Rubisco fixation, the whole cell rate of net fixation sets an upper bound on the rate of growth. Conversely, if we know the growth rate (e.g. model cyanobacteria double their mass every 10-15 hours) then we can calculate the amount of Rubisco mass required to produce this rate. Here we imagine a wildtype cyanobacterium that employs a carboxysome CCM and a relatively fast Rubisco with relatively low CO_2-specificity. Due to the action of the CCM, the Rubisco is found in a high CO_2 environment that promotes carboxylation and inhibits oxygenation (Figure 1.2). For comparison, we also consider a cyanobacterium that has no CCM but instead expresses the slower but more CO_2-specific Rubisco from spinach. Panel (A) shows that in order to achieve typical cyanobacterial growth rates, the spinach Rubisco would need make up about 10-15% of the cyanobacterial proteome by mass. When a CCM is present, by contrast, Rubisco makes up only $\approx$ 2% of the proteome mass [163, 335]. As shown in (B), a fivefold increase in Rubisco mass would necessarily decrease the proteome fractions of other cellular machinery like ribosomes and photosynthetic complexes. Since ribosomes, photosynthesis and central metabolism are intimately linked with growth, displacing these complexes would, in turn, decrease the growth rate (i.e. increase the doubling time) [269, 335]. This tradeoff between Rubisco expression and growth rate likely does not apply to large-bodied land plants because their typical mass doubling times are much longer than cyanobacterial doubling times, as shown for *Arabidopsis thaliana* in (A).

not apply as stringently to terrestrial plants because they grow much more slowly on a mass basis than cyanobacteria do (Figure 6.1).

Typical cyanobacteria weigh 1-10 pg/cell and double every $\approx$ 12 hours. Dry cyanobacterial cell mass is about 50% carbon by mass (BNID 105530) and 40-50% protein [335]. Since all the carbon in a phototroph derives ultimately from fixation by Rubisco, we can use these physiological parameters to calculate the amount of Rubisco required to achieve a particular growth rate (expressed as doubling time in Figure 6.1A). As a gedankenexperiment, imagine a cyanobacterium that has no CCM but instead expresses the relatively slow, but CO_2-specific Rubisco from spinach chloroplasts. In order to match the net carboxylation rate of a CCM and produce growth rates typical of cyanobacteria, the spinach Rubisco would need to occupy 10-15% of proteome mass (Figure 6.1A). Such high expression levels will inevitably have secondary effects, displacing the expression of crucial growth factors like the ribosome

and photosynthetic complexes (Figure 6.1B). Notably, this tradeoff between Rubisco expression and growth rate likely does not apply to large-bodied land plants because their typical growth rates are much slower than cyanobacteria on a mass basis, as shown for *A. thaliana* in (Figure 6.1A).

Biochemical and Biophysical CCMs are Functionally Distinct

Most biophysical and biochemical CCMs can be described as "converting" Rubisco from a CO_2-fixing enzyme into an HCO_3^--fixing enzyme. Bicarbonate is advantageous because it is more easily distinguished from O_2 (Table 1.1), is dramatically less cell-permeable [116], and has 10-100 times higher equilibrium concentration than CO_2 at pH 7-8.[2] In the case of biochemical CCMs, this "conversion" is achieved by employing an alternate primary carboxylase whose inorganic substrate is usually HCO_3^- (Table 1.1). In biophysical CCMs, conversion is accomplished by pumping C_i into the cell, usually as HCO_3^-. When CO_2 is the species pumped, it is thought to be converted to HCO_3^- in the cell [238] for the reasons listed above.

One important distinction between biophysical and biochemical CCMs, however, is that the pumps associated with biophysical CCMs increase the total C_i concentration in the cell. Measurements of algae and cyanobacteria typically find intracellular C_i concentrations well in excess of ambient, often more than tenfold above equilibrium [320, 133, 132]. Plants, in contrast, typically operate at C_i deficits: leaf tissue CO_2 concentrations are typically beneath the surrounding air due to the dual effects of insufficient stomatal conductance (i.e. low CO_2 uptake) and also fixation by Rubisco reducing the CO_2 concentration in the plant [51, 97].

This difference is due to the fact that terrestrial plants have internalized the air-water interface, which lies entirely outside the bodies of aquatic phototrophs. In oceans, where the vast majority of aquatic photosynthesis occurs, CO_2 equilibrates across the air-water interface at the surface of the ocean. For plants, however, CO_2 must enter the aqueous phase at stomatal pores, which are small and sometimes closed. The longest axis of stomata is determined by the size of "guard cells," which have typical lengths of 10-80 μm (BNID 101757). The fraction of leaf area covered by stomata ranges from 0.5-5% [325], which sets an upper bound on the area available for CO_2 conductance. Since there is no known means of pumping CO_2 gas into the liquid phase, the CO_2 concentration in the plant must be beneath ambient. The ambient CO_2 concentration $\approx$ 400 ppm therefore sets an upper bound on the CO_2 concentration *in planta*. Models and experiments confirm that the CO_2 concentration inside plants is substantially beneath ambient.

The above argument can be illustrated with a simple calculation. During steady-state photosynthesis, net carboxylation (A) should equal net CO_2 conductance at the stomata ($g_c(c_a - c_i)$) due to conservation of carbon atoms.

[2]This pH range is typical of cytosol, chloroplast stroma as well as both freshwater and oceanic environments.

$$A = g_c(c_a - c_i)$$
$$c_i = c_a - \frac{A}{g_c} \approx c_a - 10^{-4}$$

Here A and g_c are given in mol m^{-2} s^{-1} units, c_a is the volume fraction of CO_2 in air and c_i is the volume fraction of CO_2 in the intercellular space of the plant [97, 23]. Typical net photosynthesis values are of the order $A \approx 10-20$ μmol m^{-2} s^{-1} and typical CO_2 conductance values $g_c \approx 0.1-0.5$ mol m^{-2} s^{-1} [97], so $\frac{A}{g_c} \approx 10^{-4}$. Since $c_a \approx 400$ ppm 4×10^{-4}, the combined effect of CO_2 conductance and Rubisco fixation is of the same order as the concentration of CO_2 in air, and so we expect these phenomena to appreciably affect the amount of CO_2 in the plant. Measurements and more detailed models of plant carbon fixation confirm this intuition [68, 95].

As such, it unclear whether introduction of a biophysical CCMs can, on its own, improve C_3 plant growth [187, 328]. There may be insufficient C_i in the plant to effectively operate biophysical CCM. Moreover, biophysical CCMs invest energy in transporting C_i as HCO_3^-. Since HCO_3^- undergoes spontaneous dehyration to CO_2 quite quickly (seconds to minutes depending on the pH) [181], energy invested in pumping HCO_3^- would be wasted if C_i lingers too long before fixation. Plant cells are 10-50 times larger than cyanobacteria (e.g. BNID 100825) and so spontaneous dehydration of HCO_3^- is likely not negligible in a large, compartmentalized plant cell. It is possible that operating a functional biophysical CCM in a macroscopic land plant will require additional modifications to plant physiology, including engineering of stomates and photosynthetic electron transport [328]. For these reasons, I endeavored to interrogate the intertwining CCM function and host physiology in a simpler context - by describing it mathematically and reconstituting it in a simple, genetically tractable host - *E. coli* (Chapter 5).

6.2 Future Research Directions

High-Throughput Approaches to Understanding Rubisco

Rubisco is perhaps the best-characterized enzyme in the history of science (Chapter 2). Nonetheless, since Rubisco is ancient and highly diverged in the present-day, current literature on Rubisco folding, assembly and kinetics is quite far from covering the impressive diversity of Rubisco isoforms (Forms I, II, II/III, III and IV, Figure 1.3) and Rubisco-dependent organismal physiologies (e.g. C3, C4, CAM plants, cyanobacteria, archaea, red algae, diatoms, etc.). Since Rubisco is central to life on Earth and a major target of international bioengineering efforts, high-throughput means of studying this enzyme would be very useful.

High-Throughput Rubisco Kinetic Measurements

Rubisco kinetic assays are challenging to perform and are nearly-always accomplished in a labor-intensive, low-throughput fashion. The active site must be pre-activated by carbamylation, which is achieved by incubation in elevated CO_2 or HCO_3^-. Since activation can be variable, it is now standard to quantify the number of active sites by binding to a ^{14}C-labeled stoichiometric inhibitor, CABP, which must be synthesized. The carboxylation rate is then measured as the rate of $^{14}CO_2$ incorporation into acid-stable organic matter [57].

Notably, oxygenation parameters cannot be measured directly in the standard assay protocol because O_2 is not labeled. The Michaelis constant for O_2 (K_O) is typically inferred from measurements of carboxylation under varying O_2 concentration. This approach assumes that K_O is the same as the constant for O_2-inhibition of carboxylation, but this assumption has been vetted for only a few Rubiscos [47]. The maximum oxygenation rate ($k_{cat,O}$) is then inferred by separate measurements of the specificity factor $S_{C/O} = \frac{k_{cat,C}}{K_C} \times \frac{K_O}{k_{cat,O}}$. $S_{C/O}$ measurements are themselves fairly laborious, involving radiolabeled carbon or liquid chromatography [141, 305]. Now that I have a full understanding of these assays, I am all the more impressed with recent works reporting data for many (up to 75) Rubiscos [109, 124, 214, 235, 326].

Development of a reliable coupled enzymatic assay that could be run in small volumes in a gas-controlled incubator would greatly simplify kinetic measurements. Moreover, Rubisco kinetic parameters are typically measured at 25 °C, but photosynthesis does not take place at a single temperature [108]. Oceanic photosynthesis typically occurs closer to 15 °C, and many plants live in tropical or desert environments where temperatures can exceed 35 °C [27]. Increased oxygenation by Rubisco at higher temperatures is known to be problematic for tropical and desert plants (as discussed in Chapter 1). A high throughput assay could be used to measure the temperature-dependence of kinetic parameters for a large number (> 100) of ecologically and environmentally significant enzymes. This would be simplest for bacterial Form I and Form II enzymes, which are reliably purified from *E. coli*. A larger dataset of bacterial Rubiscos, especially Form II Rubiscos, would help elucidate whether different isoforms are subject to the same constraints as observed for Form I enzymes from plants and cyanobacteria (Chapter 2).

Another empirical approach to studying the Rubisco mechanism is to measure kinetic isotope effects (KIEs). KIEs report directly on transition state barrier heights [120] and could plausibly be measured in modest throughput using membrane-inlet mass-spectrometry (MIMS) [189]. Mass spectrometry can also be used to measure Rubisco kinetics and, notably, is capable of measuring both carboxylation and oxygenation directly in a single assay [47]. A medium- or high-throughput MIMS Rubisco assay would be transformative since it could measure all carboxylation and oxygenation parameters directly, avoid separate measurement of $S_{C/O}$, and might also be used to measure KIEs.

In Vivo Selection Experiments

In Chapter 5, I described an *E. coli* strain, CCMB1, for which growth in minimal media is coupled to Rubisco carboxylation. In addition to CCMB1, the OptSlope search algorithm highlighted at least four other knockout strains for which growth is expected to be coupled to Rubisco in defined media (Figure 5.1). Another strategy for producing a Rubisco-coupled *E. coli* involves using Rubisco expression to alleviate the toxicity associated with *prk* expression [326]. If growth is indeed coupled to the rate of carboxylation by Rubisco in some of these strains, faster growth implies faster carboxylation. Therefore, once fully vetted, strains like CCMB1 offer the possibility of testing Rubisco mutants in high-throughput by barcoded bulk competition assays of the kind described for *H. neapolitanus* in Chapter 4.

This approach is attractive for a number of reasons. First, no comprehensive mutagenesis of any Rubisco isoform has been reported. Several selection experiments have been published [54, 199, 200, 326], but the most comprehensive mutagenesis reported so far is an alanine scan of five conserved residues on the large subunit [292]. There is likely much to be learned about Rubisco catalysis, folding and assembly by testing the full complement of single and double amino acid substitutions to the large subunit (LSU, ≈ 500 amino acids). Furthermore, the Rubisco LSU displays extremely limited sequence variation [143]. Perhaps exploring a wider swath of sequence space by mutagenesis and protein engineering techniques [105, 283, 63, 127] would enable strict improvements to Rubisco kinetics?

The general strategy would be to assemble a large library of mutant Rubiscos. Since there are $\approx 10^9$ *E. coli* cells in 1 ml of media with optical density 1.0 [194], an *E. coli*-based strategy can be used to test very large libraries of Rubisco mutants [127]. In practice, sequencing is the limiting reagent in these screens. Using current sequencing platforms, perhaps 10^8 mutants could be evaluated in a quantitative fashion [127]. Fortunately, there are about 5×10^7 possible double amino acid substitution mutants to the Rubisco LSU.[3] A qualitative selection experiment - one in which only the best variants are sequenced, rather than the entire library - could plausibly evaluate upwards of 10^{10} mutants. These mutant libraries could be assembled by a number of different strategies, including targeted mutagenesis [105, 127], "protein breeding" [283], and topological rearrangement [126, 202] of natural sequences.

Approaches to Studying CCM Components in *E. coli*

One overarching theme of this work has been the use of designed *E. coli* strains, CAfree (Chapter 4) and CCMB1 (Chapter 5), to enable study of CCM components by genetic and biochemical means. These techniques can be applied to some of the central questions in CCM research.

[3] That is, the number of mutants to a 500 amino acid sequence with two distinct amino acid substitutions, which is $\approx 20^2 \times \binom{500}{2} \approx 5 \times 10^7$ unique mutant sequences

Delineating the Minimal Complement of Bacterial CCM Genes

In addition to enabling selection for carboxylation by Rubisco, the CCMB1 strain described in Chapter 5 also enables further study of the bacterial CCM. I have shown that 20 *H. neapolitanus* genes are sufficient to reconstitute the bacterial CCM in *E. coli*, but it is unlikely that all 20 genes are necessary. "Leave-one-out" experiments [125] should be used identify genes that can be omitted from future CCM expression constructs. This approach can be used to generate a "minimal CCM" construct that would enable controlled investigation of particular CCM genes. The Rubisco chaperones *acRAF* and *cbbOQ* are of particular interest. *acRAF* is related to plant chaperones involved in Rubisco folding [5] and the *acRAF* knockout has a strong HCR phenotype in *H. neapolitanus* (Chapter 4), but it's role in the CCM is not known. In contrast, *cbbOQ* are known to be Rubisco activases, which might be required to active the carboxysomal Rubisco, but the *cbbOQ* knockouts have no noticeable phenotype in *H. neapolitanus*. Experiments in CCMB1 will help us understand which of these genes is involved in producing a functional CCM, which will lay the groundwork for more detailed studies on CCM biogenesis.

Reconstitution of a Functional β-Carboxysome CCM

All the experimental work described here involves the α-carboxysome CCM of *H. neapolitanus*. Most of the cyanobacterial productivity on Earth occurs in the oceans [99] and is due to α-cyanobacteria like *P. marinus* [104], which bears an α-carboxysome CCM [237]. α-cyanobacteria likely acquired their CCM by horizontal transfer from proteobacteria [16, 150]. A second form of bacterial CCM prevails in freshwater cyanobacteria and appears to have evolved convergently [237]. The β-carboxysome CCM relies on a carboxysome organelle that is quite similar to the α-carboxysome in morphology and likewise co-encapsulates Rubisco with a CA enzyme. However, the structural proteins making up these two carboxysomes are entirely unrelated and they appear to form by different mechanisms [314, 238].

CCMB1 *E. coli* can be used to reconstitute a functional β-carboxysome CCM and interrogate whether these two convergently evolved CCMs indeed function by analogous principles. This work would help trace the evolution of these strikingly similar physiological adaptations and help clarify whether they in fact converged on the same mechanism.

Purifying Carboxysomes from CCMB1

H. neapolitanus, the typical α-carboxysome model, grows slowly and does not reach very high cell densities. Typical stationary phase cultures have an OD600 of ≈ 0.3, meaning that very large volumes of cell culture are required for carboxysome purifications [122]. Additionally, poor growth is a major limitation in study of carboxysomes because genetic manipulations of *H. neapolitanus* can take several weeks. Cyanobacterial α-carboxysome models like *Cyanobium marinum* grow more robustly, but genetic systems have not yet been developed for these models [171].

CCMB1 offers an alternative path forward. Since the CCM is expressed from two plasmids, genetic manipulation is simple. CCMB1 strains can be grown robustly to high titer in a Rubisco-independent fashion in rich media. By replacing rich media with minimal media, a dense culture can be made to grow in a Rubisco-dependent fashion for 1-2 generations before carboxysome purification. Simplified preparation of purified, functional carboxysomes will enable researchers to study many more mutants and address core questions about carboxysome function. As discussed in Chapter 2, the permeability of the carboxysome shell to important metabolites - CO_2, O_2, HCO_3^-, RuBP, 3PG, etc. - is a crucial unknown variable in models of the CCM [248]. These parameters could plausibly be measured by stopped-flow assays of whole and broken carboxysomes [122], which would be greatly aided by simplified carboxysome purification.

Studying Carbonic Anhydrases in CAfree

Carbonic anhydrases (CAs) are ubiquitous enzymes that interconvert CO_2 and HCO_3^-. These enzymes are present throughout the tree of life and evolved convergently at least 5 times [296]. Because of their multiple evolution, CAs may not be correctly identified by bioinformatic tools [223]. CAs are also particularly common in photosynthetic organisms where they appear to be involved in C_i flow through both biophysical and biochemical CCMs [77].

In Chapter 4, I presented an improved CA-reporter strain, CAfree [82, 191]. CAfree is a double knockout mutant that lacks all of *E. coli*'s native CA genes and fails to grow in ambient CO_2. Expression of a carbonic anhydrase or C_i transporter complements growth in ambient CO_2. I have not yet observed any suppressor mutants in the absence of heterologous CA expression (suppressor frequency $< 10^{-10}$, data not shown). Since CAs and C_i transporters are core components of biophysical CCMs, CAfree is a natural and facile means of testing candidate CCM genes without using radioisotopes or purifying protein, which are much more laborious and necessary for C_i uptake assays [129] and CA assays [152] respectively.

Modeling of CCM function and Interaction with Photosynthetic Physiology

Several labs are already pursuing a reconstitution of the bacterial CCM in model plants [171, 165] and others have begun heterologous expression of algal CCM components in crops [192, 210, 239]. Despite modeling results arguing that these CCMs would improve plant photosynthesis [187], I am skeptical that this notion has been fully evaluated. One cause for skepticism is a recent modeling effort from the lab of Graham Farquhar [328] suggesting that improvements to Rubisco would not, on their own, improve photosynthetic performance in C_3 plants. In this model, other manipulations (e.g. increased rate of photosynthetic electron transport) are required in order for faster carboxylation to translate into improved photosynthesis and increased biomass yield.

Most current models of plant photosynthesis do not fully account for the effects of pH on CO_2 fixation. In plants, pH can vary between different compartments and across the diurnal cycle. As discussed in Chapter 3, pH has outsized effects on the physical chemistry of the C_i pool and will therefore greatly affect the performance of biophysical CCMs in plants. The pH-dependence of the spontaneous dehydration of HCO_3^- is particularly important to consider because dehydration of HCO_3^- wastes energy invested in C_i pumping in biophysical CCMs.

A landmark mathematical model in this field is due to McGrath & Long [187], which argues that introduction of the bacterial CCM would increase maximum photosynthetic rates in C_3 plants. In my view, however, McGrath et al. did not adequately test the effects of two crucial variables on CCM performance: pH and stomatal conductance.[4] Moreover, their model neglects the spontaneous dehydration of HCO_3^-. Though I argued in Chapter 3 that spontaneous dehydration should be negligible in cyanobacteria, it not clear that this would be the case in plants: plant cells are 10-50 times larger than bacteria, are divided by many more membranes, and draw their CO_2 through stomata with limited surface area (as discussed above).

Labs and consortia undertaking efforts to introduce biophysical CCMs into plants should use the model of McGrath & Long as a basis for modeling the growth of effects of CCM expression. They should also extend this model to include the effects of pH on the membrane-permeability and spontaneous dehydration of the C_i pool. More broadly, *E. coli* reconstitution systems like the ones I describe here should be used to measure and manipulate the important parameters governing CCM function, including membrane and carboxysome permeabilities and pH in various compartments. Such experiments would help refine our models of biophysical CCM function and, thereby, greatly inform efforts to reconstitute these CCMs in land plants.

[4] McGrath & Long did perform a sensitivity analysis to test their model subject to variation in both pH and stomatal conductance. However, they only tested 10% variation to key parameters. Crucially, they vary the H^+ concentration by 10%, not the pH. A 10% change in [H^+] corresponds to a ΔpH of 0.04 units, which is within the typical range of measurement error for *in vivo* pH measurements.

Appendices

Appendix A

Empirical Evaluation of Tradeoffs in Rubisco Kinetics

A.1 Review of Previous Literature

Overview of Rubisco Folding, Activation and Catalysis

Rubisco is likely the most abundant enzyme in nature [27], being the central enzyme of the Calvin-Benson-Bassham cycle responsible for nearly all annual carbon fixation [243]. Rubisco is a notoriously complex enzyme that depends on multiple chaperones for folding, assembly and catalysis [5, 66, 83, 167, 198], requires a post-translational covalent modification to become active, and is inhibited by multiple metabolites , including its own five-carbon substrate [40, 198, 220]. Activation consists of carbamylation of an active-site lysine residue and binding of a catalytic Mg^{2+} ion. If the five-carbon substrate ribulose 1,5,-bisphosphate (RuBP) binds prior to activation, the enzyme cannot become activated. Diverse ATP-coupled chaperones, collectively termed Rubisco activases, have evolved to eject RuBP from the active site so that Rubisco can become activated [40, 198].

Once activated, all known Rubiscos catalyze both carboxylation and oxygenation of RuBP through a multistep mechanism (Figure A.1). Both carboxylation and oxygenation of RuBP are energetically favorable, but only carboxylation is considered productive because it incorporates carbon from CO_2 into precursors that can generate biomass. Oxygenation is often portrayed as counterproductive as it occupies Rubisco active sites and yields a product (2-phosphoglycolate, 2PG) that is not part of the CBB cycle and must be recycled through metabolically-expensive photorespiration at a loss of carbon [33, 50]. As photorespiration is known to play a role in signalling and nitrogen metabolism in plants [33, 50, 117], there is disagreement about this view of oxygenation and photorespiration as "wasteful." However, many phototrophs have evolved CO_2 concentrating mechanisms (CCMs) that are thought to compensate for oxygenation by elevating the CO_2 concentration near the Rubisco active site [243]. Moreover, efforts to improve photorespiration through pathway engineering have proved fairly successful in cyanobacteria and plants [147, 279, 290, 303], implying that the

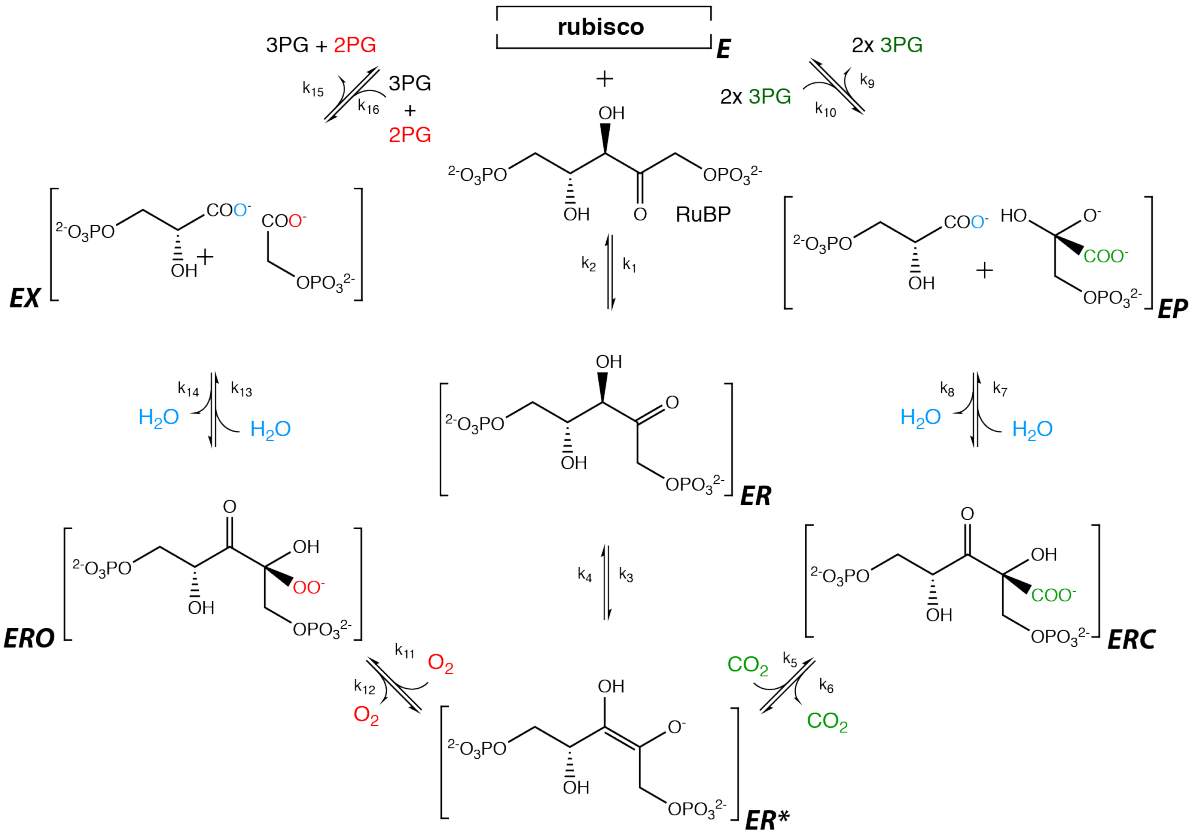


Figure A.1: The mechanism of RuBisCO following the nomenclature of Cummins et al. 2018 [71]. RuBisCO must be carbamylated and then bind Mg^{2+} before it becomes catalytically active, after which it processes three substrates - ribulose bisphosphate (RuBP), CO_2 and O_2. Rubisco can bind the five-carbon RuBP before or after activation. RuBP binding before activation strongly inhibits activity and RuBP is removed by a catalytic chaperone called Rubisco activase in many organisms [198]. The complete reactions of carboxylation and oxygenation take place through a stepwise mechanism [10, 66, 301]. RuBP binds first forming a complex (ER) with the activated form of the enzyme (E), followed by enolization of RuBP (ER*) which allows binding and further processing of CO_2 or O2. When CO_2 binds the ER* complex the ERC complex is formed while O_2 binding leads to formation of the ERO enzyme-substrate complex. Hydration and cleavage of the ERC complex leads to the formation of two enzyme-bound 3-phosphoglycerate molecules (3PG) in the EP state, each of which have 3 carbon atoms. Oxygenation proceeds through analogous steps except that the products contain 5 carbon atoms in total instead of 6 because no carbon was added. Hydration and cleavage of the ERO complex produces one enzyme-bound 3PG and one 2-phosphoglycolate (2PG) in the EX state. 2PG has two carbon atoms and is not part of the CBB cycle. As such it must be recycled through a photorespiratory pathway to avoid the accumulation of 2PG and also the loss of two carbons [33]. Atoms originating from free CO_2 and O_2 are shown in green and red respectively. The oxygen atom originating from water is shown in blue.

products of oxygenation by Rubisco do in fact impinge on autotrophic growth, albeit in a complex manner. Despite the fact that many autotrophs depend on Rubisco carboxylation for growth, all known Rubiscos are relatively slow carboxylases that fail to exclude O_2 (Figure 2.1A-B, Figure A.1).

While it is often claimed that Rubisco is "slow" and "non-specific" it is actually the case that Rubisco is a slightly below-average enzyme in terms of carboxylation $k_{\mathrm{cat,C}}$ and roughly average in terms of $k_{\mathrm{cat}}/K_{\mathrm{M}}$ [26]. Moreover, comparing the rate of the Rubisco

carboxylation to the uncatalyzed reaction we find that Rubisco confers a rate enhancement of 10^{16} fold [31]. A rate enhancement of 10^{16} fold is quite impressive, especially in comparison to canonical "perfect" enzymes like carbonic anhydrase, superoxide dismutase and triose-phosphate isomerase that display rate enhancements on the order of 10^7-10^9 10^7-10^9 fold [31].

Rather than claiming that Rubisco is absolutely "slow" we instead are surprised to find that Rubisco is not faster given its centrality to life and its abundance [26, 31]. Indeed, Rubisco is routinely measured to comprise upwards of 20% of total leaf protein in C_3 plants [110]. Moreover, oxygenation by Rubisco competitively inhibits carboxylation and lowers the effective rate of carboxylation [262, 300], which leads us to expect selection for faster carboxylation $k_{\mathrm{cat,C}}$ to compensate for inhibition by O_2. CO_2 concentrating mechanisms are thought to elide this problem by elevating the CO_2 concentration near the Rubisco active site and competitively inhibiting oxygenation [243]. The curious slowness and non-specificity of such a central enzyme, combined with the fact that CCMs evolved multiple times independently in diverse lineages [243], suggests that perhaps Rubisco itself cannot be strictly improved - i.e. that some physicochemical limitation imposes a tradeoff on the enzyme such that it cannot be both fast and specific [21, 262, 277, 301].

Definition of Rubisco Kinetic Parameters

Over the decades since its discovery, various nomenclature has been used to describe the kinetics of Rubisco carboxylation and oxygenation. Here we use $k_{\mathrm{cat,C}}$ and $k_{\mathrm{cat,O}}$ to denote the maximum catalytic rates (measured in units of s^{-1}) for carboxylation and oxygenation respectively. K_{C} and K_{O} denote the Michaelis constants (half-saturation concentrations in μM) for carboxylation and oxygenation. The specificity factor $S_{\mathrm{C/O}} = (k_{\mathrm{cat,C}}/K_{\mathrm{C}})$ / $(k_{\mathrm{cat,O}}/K_{\mathrm{O}})$ is a unitless measure of the relative preference for CO_2 over O_2 (Figure 2.1A).

Using these definitions, the per-active site rates of carboxylation (R_{C}) and oxygenation (R_{O})

$$R_C = k_{\mathrm{cat,C}}(1 + \frac{K_{\mathrm{C}}}{[CO_2]} + \frac{K_{\mathrm{C}}/[CO_2]}{K_{\mathrm{O}}/[O_2]})^{-1}$$
$$R_O = k_{\mathrm{cat,O}}(1 + \frac{K_{\mathrm{O}}}{[O_2]} + \frac{K_{\mathrm{O}}/[O_2]}{K_{\mathrm{C}}/[CO_2]})^{-1}$$

These equations take the form of an irreversible Michaelis-Menten type relationship with mutual competitive inhibition [262]. That is, both carboxylation and oxygenation are irreversible (both reactions are associated with strongly negative $\Delta G'_m$values) and both reactions mutually inhibit each other competitively by occupying the same active site [10, 66, 102, 300]. These per-enzyme rates can be scaled up to the total rates of carboxylation V_{C} and V_{O} by multiplying by the concentration of active Rubisco denoted [E].

Previous Literature on Tradeoffs in Rubisco Catalysis

The long-standing observation of correlations between the specificity factor $S_{C/O}$ and other Rubisco kinetic parameters [21, 140, 219] is often cited to motivate the notion that tradeoffs mediated by the catalytic mechanism strictly constrain Rubisco's catalytic potential [21, 262, 277, 301]. Tcherkez et al. 2006 [301] focuses primarily on the correlation between $S_{C/O}$ and $k_{cat,C}$ and argues that this relationship is best explained by the intrinsic difficulty of discriminating between CO_2 and O_2, both of which are small, volatile and similarly nonpolar molecules. As a result, Tcherkez et al. claim [301], discrimination must occur between the carboxylation and oxygenation transition states (TS). An "advanced, product-like" carboxylation TS allows for maximum discrimination because the developing carboxylate intermediate is more readily distinguished from the peroxyacid of the oxygenation intermediate. However, Tcherkez et al. 2006 argue that tight-binding of a product-like intermediate will necessarily limit the throughput of the subsequent, rate-limiting, reaction steps by slowing the release of the carboxyketone carboxylation intermediate (Figure A.2).

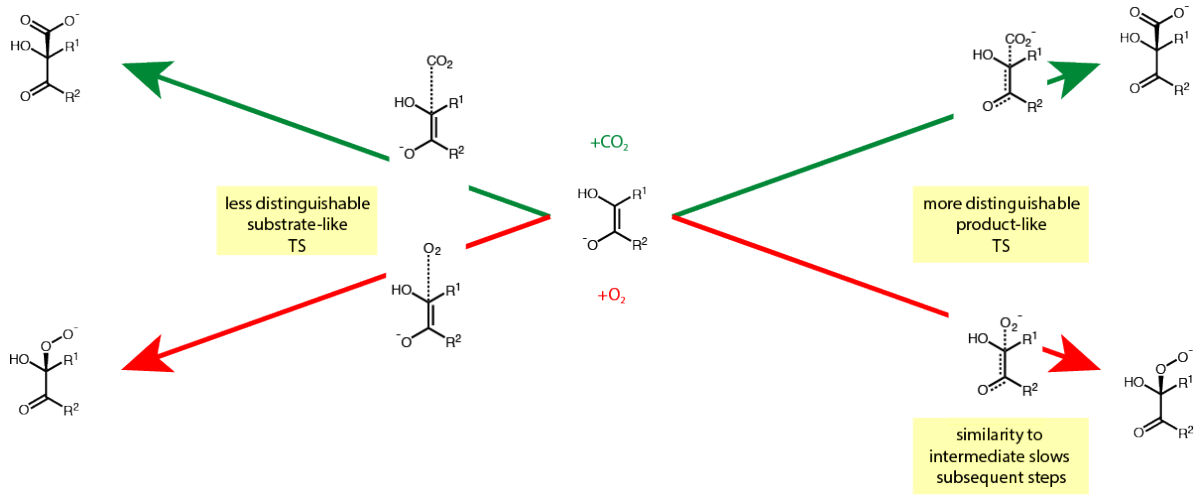


Figure A.2: Visual explanation of coupling between carboxylation rate and affinity as proposed by Tcherkez et al. In this model, it is assumed that Rubisco discriminates between CO_2 and O_2 at the level of the first transition state [301]. The need for selectivity in the presence of high O_2 or low CO_2 concentrations promotes a late carboxylation transition state (green trajectory in the right half of the figure) which more resembles the carboxylate of the carboxyketone intermediate. The developing carboxylate of a late TS is maximally discriminable from the peroxyacid of the first oxygenation TS (red trajectory). However, in this model, stabilization of a late carboxylation TS forces tight binding of the carboxyketone carboxylation intermediate (following Hammonds postulate) and slows the downstream rate-limiting hydration and cleavage of the intermediate. Though this model is motivated by the need for selectivity, note that the tradeoff implied by the model is between initial binding of CO_2 ($k_{cat,C}/\ K_C$) and subsequent processing of the carboxylation intermediate ($k_{cat,C}$). That is, the tradeoff is strictly independent of the O_2 concentration and affects only those kinetic parameters related to carboxylation.

This explanation is consistent with a number of lines of evidence given in [301], including correlation between specificity ($S_{C/O}$) and maximum carboxylation rate ($k_{cat,C}$), trends in the carbon kinetic isotope effect and the extremely tight binding of the carboxyketone intermediate analog CABP. However, data is limited for all of these claims and the linear

correlation plots in [301] have notable outliers in the cyanobacterial Rubiscos, which are faster than the overall trend would predict. Moreover, if the tradeoffs are energetic - as implied by the description advanced in [301] - we expect correlations on a log scale and not necessarily on a linear scale. Accordingly, when the plots in Tcherkez et al. 2006 are drawn on a log scale cyanobacterial Rubiscos are no longer outliers, but fall along the trend line [262].

Before describing [262], we note that this requirement of TS discrimination is a direct consequence of two assumptions made by both papers and supported by experimental evidence. Briefly, it is assumed that addition of either gas is irreversible and that there is no binding site for CO_2 or O_2 and, thus, no Michaelis complex. If CO_2 bound a specific site on Rubisco before reacting with the enediolate of RuBP, the affinity for CO_2 could be modulated (e.g. by mutating that binding site) without affecting the kinetics of subsequent reaction steps. If gas addition were reversible, discrimination could also be implemented via differential CO_2 and O_2 off-rates as in Hopfield's model of kinetic proofreading [131].

Savir et al. 2010 [262] can be understood as starting from a simple motivating point: the specificity factor S is a function of the other catalytic parameters ($S_{C/O} = (k_{cat,C}/K_C)$ / $(k_{cat,O}/K_O)$). It is therefore not surprising that we observe correlations between $S_{C/O}$ and other kinetic parameters. But which parameters drive the observed correlations? If all kinetic parameters ($S_{C/O}$, K_C, $k_{cat,C}$, K_O and $k_{cat,O}$) were free variables, then they would vary in a 4-dimensional space[1]. Savir et al. 2010 uses PCA in the space of log-transformed kinetic parameters to show that the kinetic parameters of ≈20 diverse Rubiscos are roughly one dimensional, meaning that $S_{C/O}$ is well-approximated as a function of $k_{cat,C}$ alone. Indeed, according to this analysis all of the carboxylation and oxygenation parameters can be calculated from a single free variable (e.g. $k_{cat,C}$). Savir et al. 2010 also argues that the maximum oxygenation rate $k_{cat,O}$ is uncorrelated with the other parameters, implying either that $k_{cat,O}$ is unconstrained or that it has not been selected for strongly in the evolution of these ≈20 enzymes.

Based on the established kinetic model of Rubisco activity and some assumptions, Savir et al. [262] note that the catalytic efficiencies for carboxylation and oxygenation are related to "effective barriers" to CO_2 and O_2 addition respectively. That is, $\ln(k_{cat,C}/K_C)$ $-(\Delta G_{1,C}/RT)$ and $\ln(k_{cat,C})$ $-(\Delta G_{2,C}/RT)$ as diagrammed in Figure 2.1 and explained in detail below. The catalytic efficiencies $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ can be calculated straightforwardly from measured kinetic parameters. Similarly, given some assumptions (see below) $k_{cat,C}$ can be related to the height of a second "effective barrier" to carboxylation ($\ln(k_{cat,C})$ $-\Delta G_{2,C}/RT$, Figure 2.1). This effective barrier height ($\Delta G_{2,C}$) represents hydration and cleavage of the bound carboxylation intermediate (the state labeled ERC in Figure A.1).

Savir et al. 2010 demonstrates that $k_{cat,C}$ correlates negatively with $k_{cat,C}/K_C$ on a log scale (in their dataset) and propose the following explanation [262]. $k_{cat,C}/K_C$ and $k_{cat,C}$ are related (exponentially) to the height of the first and second effective carboxylation barriers respectively (Figure 2.1, below). Their correlation implies that these barrier heights sum

[1] not 5 because S is wholly determined by the other parameters

to a constant, or, as interpreted by Savir et al., that a maximum "deformation energy" that can be applied to RuBP and this energy must be partitioned between the two reaction steps. This proposed tradeoff between CO_2 addition and cleavage rates is quite similar to the proposal of Tcherkez et al. [301] and could produce the observed correlation between $S_{C/O}$ and $k_{cat,C}$ without reference to the kinetics of oxygenation (as diagrammed in Figure A.2). Savir et al. 2010 also shows that $k_{cat,C}/K_C$ is positively correlated with $k_{cat,O}/K_O$ on a log scale, implying that there is some coupling between the effective barriers to CO_2 and O_2 addition (as diagrammed in Figures 1B and 6A). Given this correlation they propose that mutations that make CO_2 addition to the ERC complex (Figure A.1) also increase the rate of O_2 addition to that same complex.

While the use of log-scale plots in Savir et al. [262] is sensible (linear energetic tradeoffs will produce log-scale correlations between kinetic parameters), neither of their mechanistic claims is straightforward in light of the microscopic mechanism of Rubisco (see below) and both warrant exploration in the light of new data. Indeed more than 200 Rubisco variants have been characterized since 2010. Here we examine whether new data supports or contradicts the arguments of [262] and [301].

A.2 Description and Analysis of the Extended Dataset

Collection and Curation of the Extended Dataset

The raw extended dataset including kinetic parameter values from the primary literature as well as temperature and pH of measurements and manual annotations is given as Dataset S1. Experimental error was recorded when reported. We attempted to exhaustively extract kinetic parameters for wild-type Rubiscos. Several values for mutant and hybrid enzymes were also extracted in the process. We reviewed the primary literature and manually annotated cases where the underlying data did not appear trustworthy, for example measurements taken before the active site stoichiometry of Form I Rubiscos was well-understood [20]. Equivocal measurements, mutant enzymes and some values measured at temperatures other than 25 °C were removed before further analysis.

In cases where experimental error was not reported, we assumed that experimental error scales with the measured value. We calculated the mean coefficient of variation (CV, the ratio of the standard deviation to the mean) for each commonly measured kinetic parameter (S, $k_{cat,C}$, K_C and K_O) and assumed this CV to assign a plausible standard deviation in cases of unreported error.

In some cases multiple values for the same parameter were found in the same reference, for example multiple similar S values (and standard deviations) for *R. rubrum* and *S. oleracea* Rubiscos in [139]. When these values were similar, they were combined into a single mean and standard deviation by bootstrapping. Briefly, it was assumed that measurements were normally distributed with the reported standard deviation and means. 10^4 values were

randomly sampled from the each implied distribution (i.e. one per measurement being combined) and a posterior mean and standard deviation were calculated from the samples.

$S_{C/O}$ is often measured at pH 8.3 while kinetic parameters $k_{cat,C}$, K_C and K_O are usually measured near pH 8.0 [124, 214]. It is commonly assumed that $S_{C/O}$ is pH-independent after Jordan et al. [139]. We made the same assumption and combined kinetic measurements and $S_{C/O}$ values from the same reference. This assumption is used to calculate $k_{cat,O}$ from the other four parameters as $k_{cat,O} = (k_{cat,C}/K_C)\ /\ (S_{C/O}/K_O)$ from the definition of $S_{C/O}$. $k_{cat,O}$ is often calculated in this manner because K_O is usually high enough that saturating the oxygenation rate is technically challenging. Given the combined data, we calculated a 95% confidence interval on $k_{cat,O}$ by 10^4-fold bootstrapping (described in the Methods section) in cases where $S_{C/O}$, $k_{cat,C}$, K_C and K_O were measured in the same reference. When sufficient data was available we also calculated 95% confidence intervals for $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ in a similar manner. The median of bootstrapped values was used to represent $k_{cat,O}$, $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ in plots (e.g. in Figures 4-6). Inferred values were checked for internal consistency and consistency with literature values before performing any analyses. We refer to the resulting dataset - where data from individual references are merged and unmeasured values are inferred - as the extended dataset.

The complete extended dataset is given in Dataset S2 and source code used for dataset normalization and error inference is available at https://github.com/flamholz/rubisco. Dataset S2 is filtered such that it contains only measurements of native (i.e. wild-type, non-mutant) enzymeos taken near pH 8 and at 25 °C. An unfiltered version of Dataset S2 is given as Dataset S4. This repository contains data about mutant and hybdrid Rubiscos as well as some values measured away from 25 °C. We did not attempt to comprehensively determine search for mutant measurements or measurements at non-standard temperature, so Dataset S4 is likely far from complete in this respect. Supplementary datasets are described in Tables S1 and S2.

The extended dataset displays little variation among all carboxylation and oxygenation parameters with central values varying over less than fourfold in all cases (Figure A.3). This is clear from the $\log_{10}$-scale standard deviation of Form I Rubisco kinetics (reported as σ^*in Figure A.3), which is less than 0.5 for all parameters. Rubisco differs notably from other enzymes in this regard: for most reaction classes kinetic parameters vary over 2-3 orders of magnitude [26, 73]. In particular, Rubisco displays extremely low variation in $k_{cat,C}$ (σ^*= 0.2, Figure A.4) as compared to other enzymes for which multiple k_{cat} measurements are available from the BRENDA database [265]. Form I and Form II enzymes differ most in their $S_{C/O}$, $k_{cat,C}$ and K_C values, which can be seen by comparing the median of blue and yellow distributions in Figure A.3. In addition to analyzing variation between Rubisco isoforms, the extended dataset can be used to see variation in kinetic parameters between Rubiscos from different hosts. Figure A.5 shows that different host physiologies are associated with characteristic carboxylation kinetics, with C_3 plant Rubiscos being more CO_2-specific than cyanobacterial Rubiscos, for example.

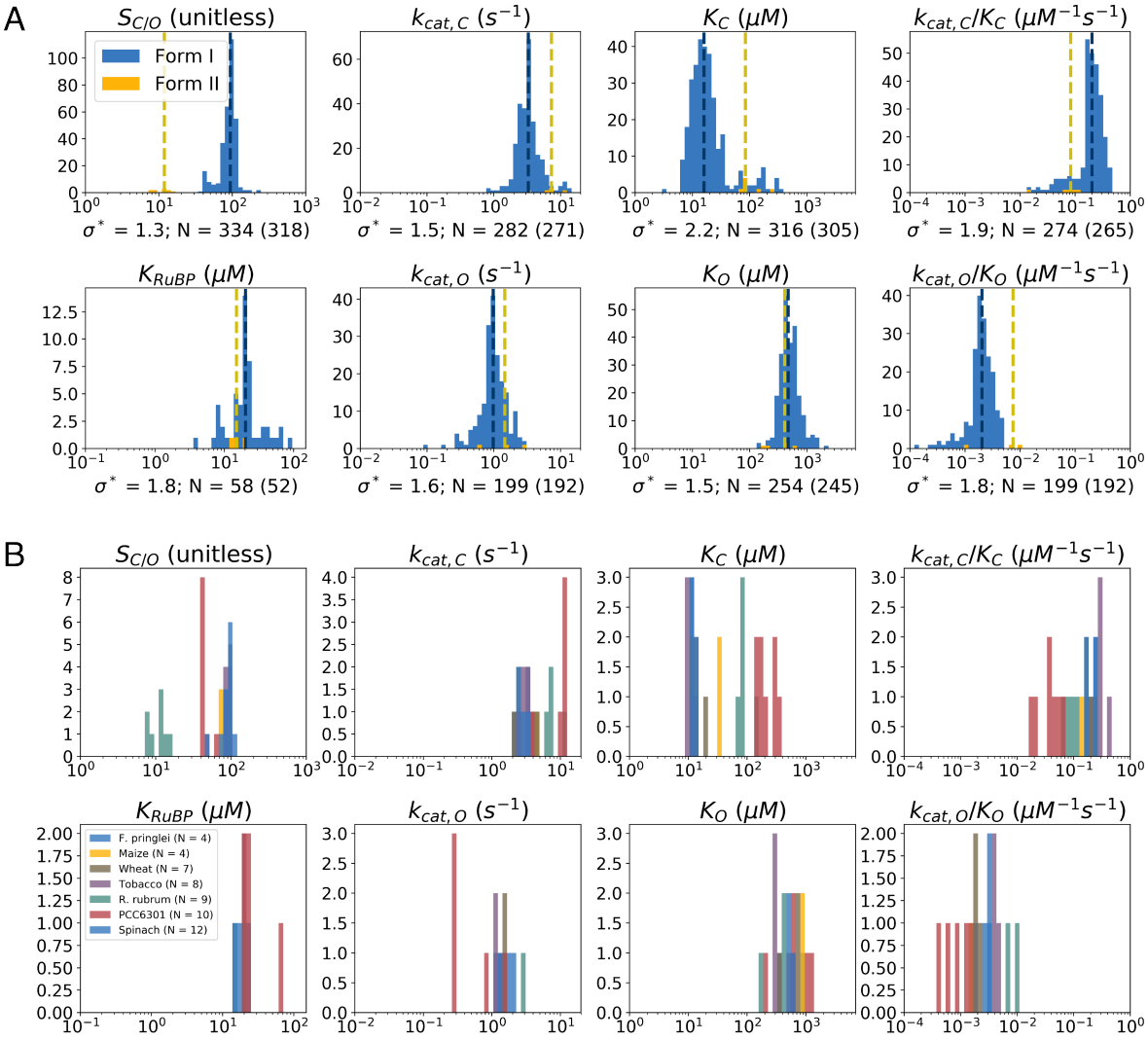


Figure A.3: Histograms of Rubisco kinetic parameters from the extended dataset. In all histograms, kinetic parameter values are on the X-axis and number of data points in on the Y-axis. (A) Parameter distributions for all Form I and Form II Rubiscos in the extended dataset. Form I distributions are in blue and Form II distributions are in yellow. Other Rubisco isoforms are omitted because of scant data. The dashed lines give the medians of each distribution. The number of values in the dataset is given as N and the number of values for Form I Rubiscos is in parentheses. The geometric standard deviation of Form I values is given as σ^*. Notice that $\sigma^* < 3$ in all cases, meaning that the central values of each parameter vary over less than threefold. (B) Parameter distributions for the seven Rubiscos with ¿ 3 measurements. In the figures in Chapter 2 used median values to represent species with multiple measurements. Panel B shows that multiple measurements of the same Rubisco are broadly consistent, though there is clear variation e.g. early $k_{\mathrm{cat,C}}$ measurements of the *Synechococcus* PCC 6301 Rubisco are roughly fivefold lower than recent measurements, likely due to improved quantitation of the number of active sites.

Principal Component Analysis of The Extended Dataset

We repeated the PCA analysis of Savir et al. in order to see if the principal axis of variation is altered substantially by a tenfold increase in the dataset size [262]. PCA was performed

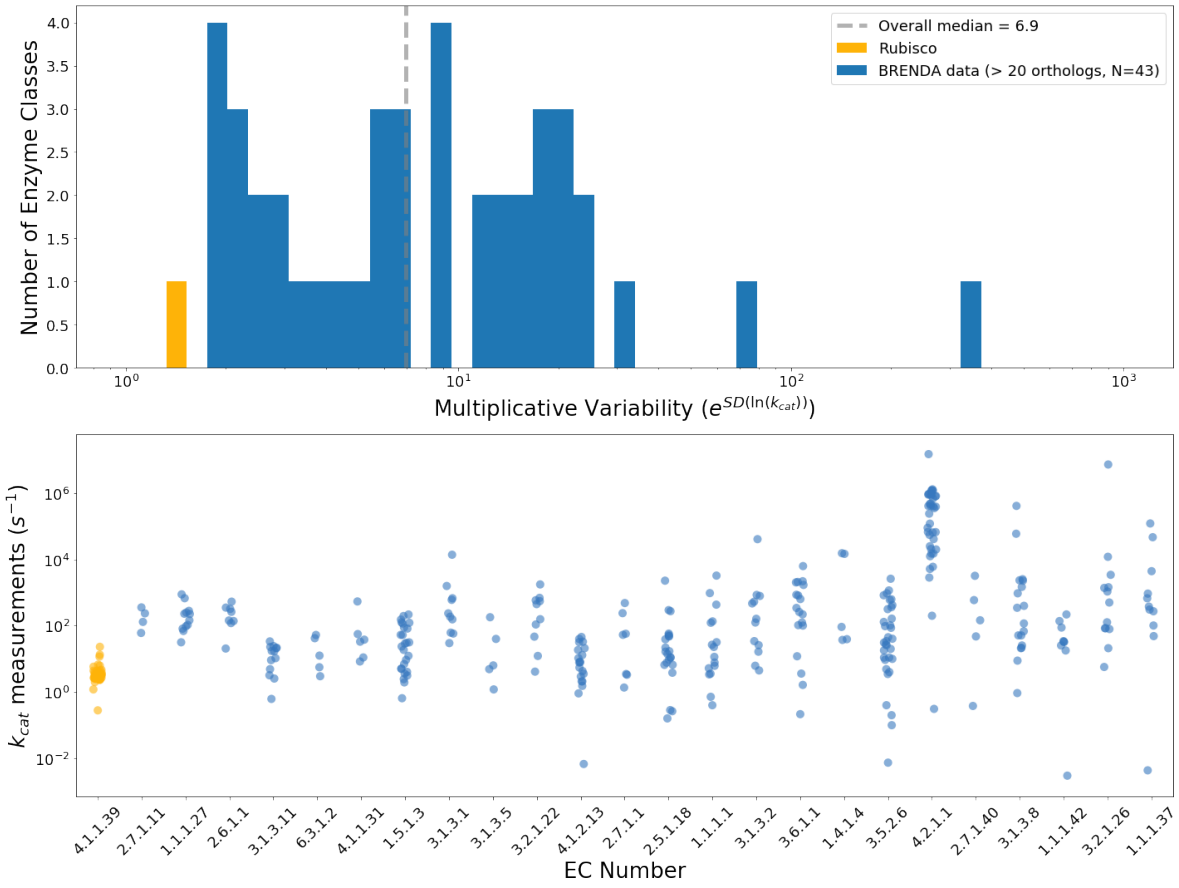


Figure A.4: Rubisco displays very little variation in $k_{cat,C}$ values as compared to other enzymes for which > 20 measurements are available. k_{cat} data for wild-type enzymes other than Rubisco was drawn from the BRENDA database [74]. In cases where multiple measurements of the same enzyme from the same organism were found, the median value was used as representative. This procedure produced 240 $k_{cat,C}$ values for Rubisco with a geometric standard deviation (or multiplicative variability) $\sigma^* = 1.48$, defined as the exponent of standard deviation in logarithmic scale $\exp(SD(\ln k_{cat}))$. This quantity reflects the multiplicative spread of central k_{cat} values around the mean. When the underlying distribution is log-normal, a value of $\sigma^* = 2.0$ connotes that the central 68% of values are within two-fold of the mean. Panel (A) shows that Rubisco is an outlier among enzymes for which > 20 k_{cat} measurements were available, displaying fivefold less variation in kcat than the median enzyme in this grouping (median $\sigma^* = 6.9$). Panel (B) shows k_{cat} data for the 25 enzyme classes with the most available data, ordered from least-to-most multiplicative variability. Roughly fivefold more data is available for Rubisco than any other enzyme in this dataset, but it displays the least variation in k_{cat} of any enzyme in the dataset.

on a 4 dimensional space of parameters that uniquely determine net carbon fixation: [K_C, $k_{cat,C}$, $S_{C/O}$, K_C/K_O]. Analysis was restricted to those Rubiscos for which $S_{C/O}$, K_C, $k_{cat,C}$, and K_O measurements were available from the same reference. This data set contains > 150 more Rubiscos points than used in [262]. The data were log-transformed and Z-score normalized before PCA. Normalization and PCA was performed using the Python sklearn package.

Focusing on Form I Rubiscos as in [262] shows that the principal axis remains mostly unchanged by the addition of new data. PC1 = [0.47 0.52 -0.48 0.53] over the Form I subset

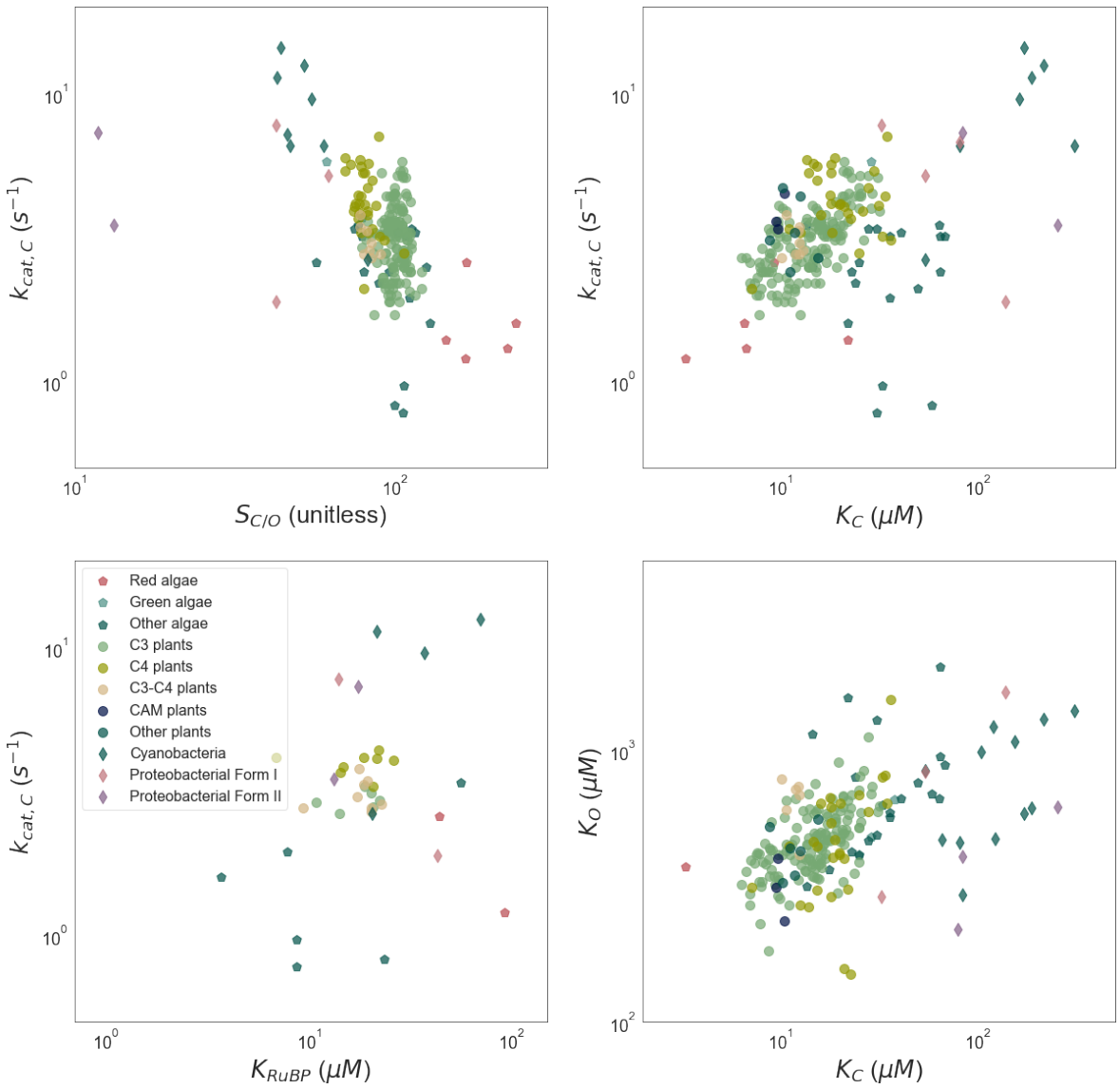


Figure A.5: Rubisco kinetic parameters by host organism physiology. There are characteristic differences in kinetic parameters depending on host physiology. Cyanobacteria generally possess the fastest Rubiscos and red algal Rubiscos have characteristically high CO_2 specificity. The top left panel shows that different host physiologies generally occupy distinct regions of the $k_{\mathrm{cat,C}}$ vs. $S_{\mathrm{C/O}}$ plot, consistent with the analysis of [301]. However, as documented in Chapter 2, these broad trends do not produce a very strong correlation, which is what we would expect if Rubisco was perfectly optimized with respect to a tradeoff between $k_{\mathrm{cat,C}}$ and $S_{\mathrm{C/O}}$.

of the extended dataset as compared to PC1 = [-0.52 -0.51 0.44 -0.52] over the 15 data points in the Form I subset of Savir et al. [262]. The proportions associated with the principal components (the percent of variance explained) are [70% 12% 11% 5%] of the variance as compared to [91%, 6%, 2%, 0.6%], indicating that there is substantially more variance in the extended dataset even as the principal axis is quite similar. Indeed, this variation is apparent

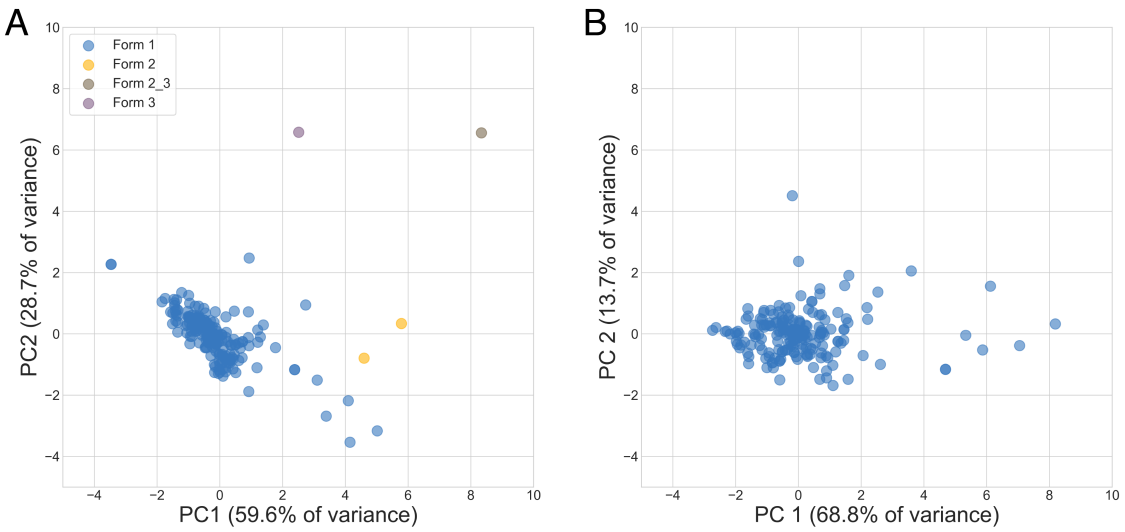


Figure A.6: Principal components analysis of the extended dataset. Data were normalized before PCA in order to ensure that the very different absolute values of Rubisco kinetic parameters did not bias analysis. Panel (A) gives the projection onto the first two principal components (PCs) determined by the subset of the extended dataset for which sufficient data was available - i.e. for which the vector [$k_{cat,C}$, K_C, $S_{C/O}$, K_C/ K_O] could be calculated. It is clear from the plot that the first two PCs separate the Rubisco isoforms from each other. When PCA is restricted only to Form I RuBisCOs, as in panel (B), there is still substantially more variation in Rubisco kinetic parameters (i.e. lower percent of variance explained by the first principal component, PC1) than described in previous work, where PC1 accounted for roughly 90% of the variation in the data and Rubisco was described as evolving within a roughly 1-dimensional space [262]

from a plot of Form I data projected onto the first two principal components (Figure A.6).

Correlation and Regression Analyses

Certain model Rubiscos are characterized frequently. For example we found 12 independent measurements of the model Rubisco from spinach and 10 for the model cyanobacterial Rubisco from Synechococcus sp. PCC 6301. We used the median measured value in correlation and regression analyses to avoid bias towards frequently-characterized Rubisco variants. Once multiple measurements were merged, we used performed correlation and regression analyses on a log scale to investigate coupling between Rubisco parameters. The reasoning for log-scale analyses is discussed in detail below, but is based on our expectation (rooted in transition state theory) that relationships between kinetic parameters will have exponential form. An exponential relationship of the form $Y = a\ X^k$ is called a power law. Power laws produce a linear relationship (e.g. $\log_{10}(Y) = \log_{10}(a) + m\ \log_{10}(X)$) on a log-log plot. The slope of the linear fit corresponds to the power law exponent and the Y-intercept corresponds to the exponential prefactor.

Regressions were performed using total least squares (TLS) regression, which is sometimes also called orthogonal distance regression (ODR). TLS regression is applicable to cases

where there is experimental error associated with both X and Y variables, which is the case here because both variables result from fits to multiple measurements of Rubisco activity. However, R^2 values of TLS fits do not convey the explained fraction of Y axis variance and are therefore not easily interpreted. As such, we report the degree of correlation as Pearson R values and 95%. 10^4-fold bootstrapping was used to estimate 95% confidence intervals for R, power-law exponents and prefactors (i.e. the slopes and intercepts of linear fits in log-log scale). In each iteration of the bootstrap, data were subsampled to 90% with replacement. TLS regression was applied to each subsample to determine a point estimate of R, the power-law exponent and prefactor. The 95% confidence interval on R gives an indication of the robustness of the underlying power-law correlation, which the confidence intervals on power-law exponent and prefactor quantify the degree to which the power law is determined by the data. Results of most correlation and regression analyses are discussed in detail in Chapter 2.

Correlations with $k_{cat,O}$

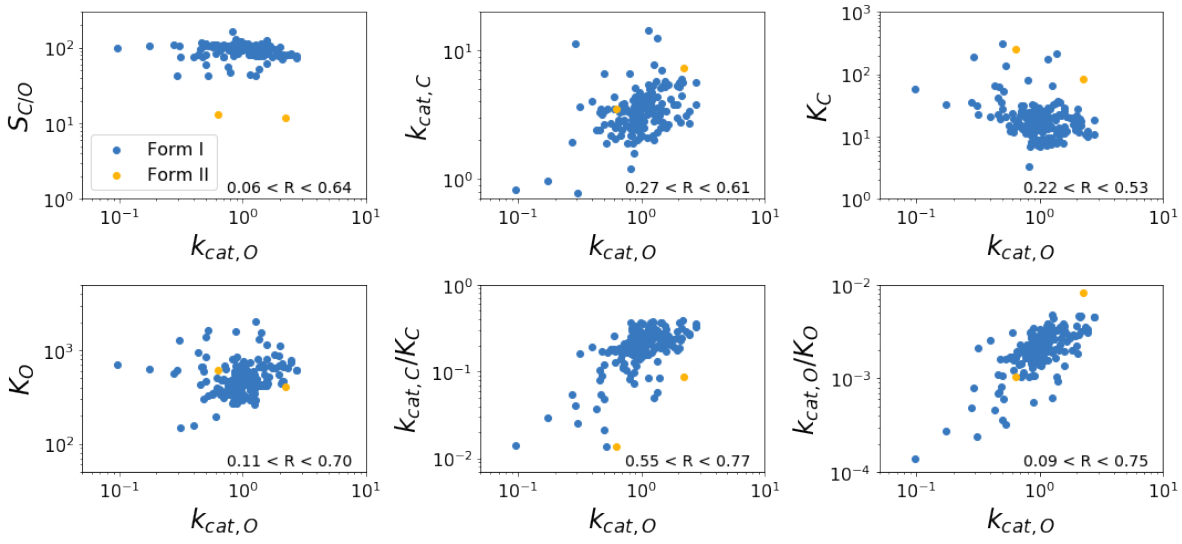


Figure A.7: Correlations between $k_{cat,O}$ and other kinetic parameters. $S_{C/O}$ is unitless, k_{cat} values have s^{-1} units and K_M values have μM units. Form I Rubiscos are in blue and Form IIs are in yellow. A 95% confidence interval on the Pearson correlation R among Form I enzymes was calculated by 1000-fold bootstrapping and indicates that the $k_{cat,O}$ may be correlated with $k_{cat,C}$ and $k_{cat,C}/K_C$ but is only weakly correlated with other parameters.

An intriguing suggestion of [262] is that $k_{cat,O}$ might vary independently of other kinetic parameters. This might imply that a very slowly oxygenating Rubisco could be designed or evolved in a laboratory. Alternatively, it might imply that typical O_2 concentrations are low enough that $k_{cat,O}$ is not strongly selected for. Figure A.7 evaluates correlations between $k_{cat,O}$ and other parameters in the extended dataset. $k_{cat,O}$ is now somewhat correlated with

$k_{\mathrm{cat,C}}$ and $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$, suggesting that earlier thinking about oxygenation rates should be re-evaluated.

Notes on Use of Log-Scale Correlations

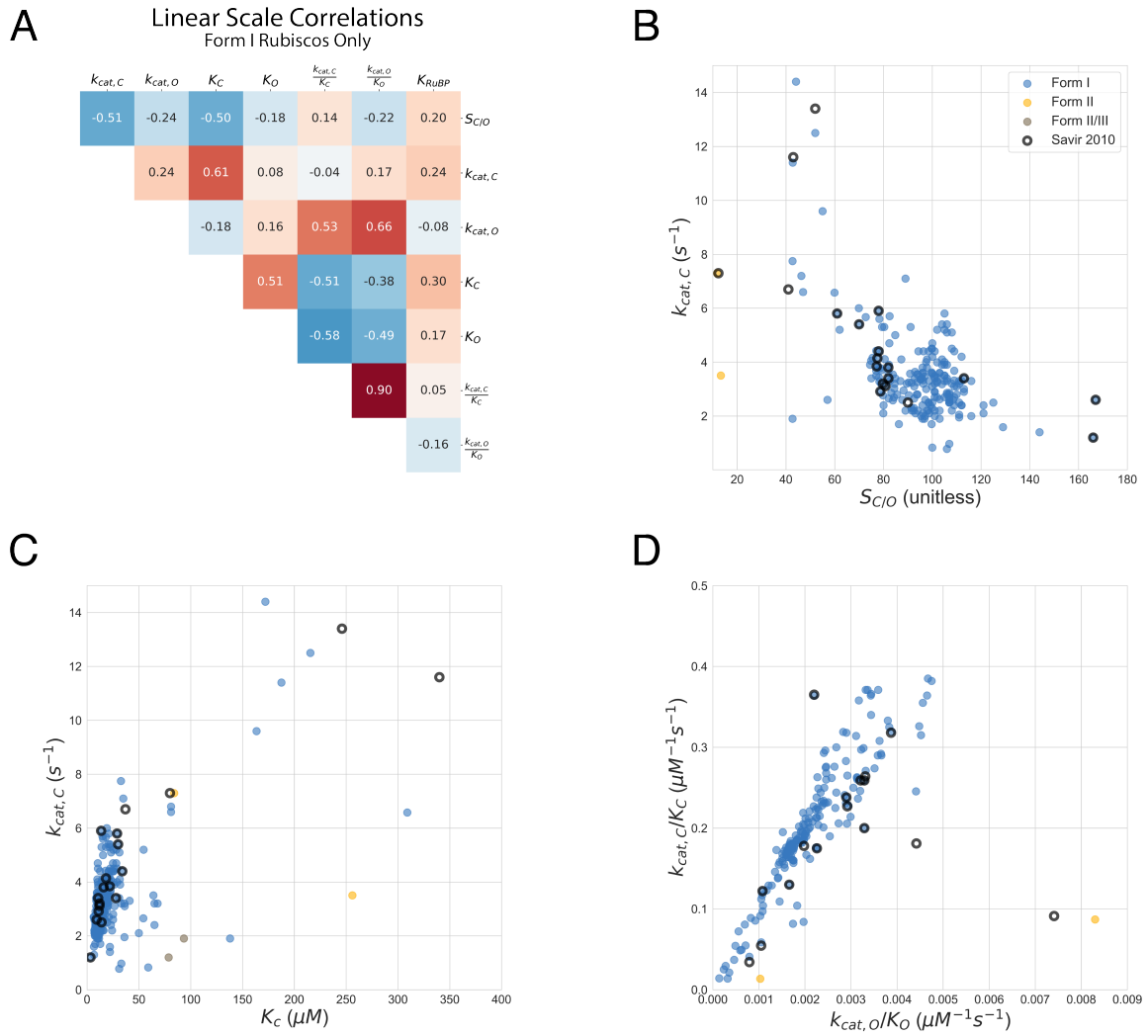


Figure A.8: Linear scale correlations between key pairs of Rubisco kinetic parameters. (A) A heatmap of pairwise linear-scale correlations is qualitatively similar to Figure 2.3. The $S_{\mathrm{C/O}}$-K_{C}, $S_{\mathrm{C/O}}$-$k_{\mathrm{cat,C}}$, and K_{C}-$k_{\mathrm{cat,C}}$ correlations are of particular interest because they were highlighted in previous works. None of these pairs give R ¿ 0.65. As in the log-scale Figure 2.3, the strongest linear scale correlation is between the catalytic efficiencies for carboxylation and oxygenation, $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ (R = 0.90). (B) A linear-scale scatter plot of $k_{\mathrm{cat,C}}$ against $S_{\mathrm{C/O}}$ shows that the data has very limited dynamic range and correlations are determined by a small number of outliers. (C) A similar scatter plot of $k_{\mathrm{cat,C}}$ against K_{C} shows limited dynamic range and small number of extreme examples. (D) Plotting the catalytic efficiencies $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ against $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ shows that robust log-scale correlation is recapitulated on a linear scale, as would be expected for a power-law exponent of roughly 1.0.

Throughout Chapter 2 we presented log-scale correlations and regressions to investigate tradeoffs between Rubisco kinetic parameters. For reasons described here we strongly prefer to investigate these questions on a logarithmic scale. For those interested we nevertheless give linear scale correlations and scatter plots in Figure A.8. Figure S8A shows that pairwise linear-scale correlations are qualitatively similar to the log scale correlations in main-text Figure 2.3. Correlation between $k_{\mathrm{cat,C}}$ and $S_{\mathrm{C/O}}$ is modest on both scales (R ≈ -0.5) and clearly driven by outlying measurements (Figures S8B). The correlation between $k_{\mathrm{cat,C}}$ and K_{C} is somewhat higher in a linear scale (R = 0.61) than in log scale (R = 0.48) and is similarly driven by outlying measurements (Figure A.8C). More data on very fast (e.g. cyanobacterial and proteobacterial) and very specific (e.g. red algal) Rubiscos is needed to better-resolve these correlations. As in log-scale, the strongest linear scale correlation is between the catalytic efficiencies for carboxylation and oxygenation, $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ (R = 0.90). Plotting the catalytic efficiencies $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ against $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ (Figure A.8D) shows that robust log-scale correlation is recapitulated on a linear scale, as would be expected for a power-law exponent of roughly 1.0.

We reasoned that log-scale correlations are appropriate for two reasons. First, the proposed mechanisms all involve tradeoffs between effective transition state (TS) energy barriers. If these tradeoffs are additive in energy units then they should manifest as power laws as functions of rate constants. The TS theory based logic is explained in the next paragraphs. Second, linear regressions on a logarithmic scale are more robust to multiplicative error, which is common in biochemical measurements and likely present in this case (as discussed below).

If two TS barrier heights $\Delta \mathrm{G}_1$ and $\Delta \mathrm{G}_2$ trade off with each other, the tradeoff will manifest as some coupling between the barrier heights. If the coupling is negative, one barrier height must decrease when the other increases, which would produce negative correlation between the two barrier heights. If that correlation is linear then

$$\begin{aligned} \Delta G_1 + m\Delta G_2 &= b \\ \exp(\Delta G_1) &= \exp(-\Delta G_2)^m \times \exp(b) \end{aligned}$$

Where m and b are positive constants. Transition state theory posits that first order kinetic constants k_1 and k_2 are related to their respective barrier heights as follows:

$$\begin{aligned} ln(k_1) &\propto -\Delta G_1 \\ ln(k_2) &\propto -\Delta G_2 \end{aligned}$$

Where the ΔG values are in units of RT so that we can omit the factor of RT here and above. It follows from these equations

$$\frac{1}{c_1 k_1} = \exp(b) \times (c_2 k_2)^m k_1 = c' {k_2}^{-m}$$

Here c_1 and c_2 are constants of proportionality, and $c' = (e^{-b}\ c_1^{-m})\ /\ c_2$ is a constant as well. In summary: we expect to find a negative power law correlation with exponent $-m$ between two kinetic constants when there is a negative linear tradeoff between the associated transition state barrier heights (with slope $-m$).

Following this same logic and as shown in detail below, Rubisco kinetic parameters can be treated as exponentially related to effective TS barriers given particular assumptions. Therefore we perform correlation analysis on a log scale in order to interpret correlations between kinetic parameters are related to the microscopic kinetics of Rubsico. It is not possible to interpret linear scale correlations as directly related to TS theory and so performing correlation analysis on a log scale greatly aids in interpretation.

The second reason that we prefer logarithmic scale is that regressions on a log scale are appropriate when measurement errors are multiplicative. That is, when the error scales with the measured quantity as implied by statements like "within 10% error" or similar. Multiplicative error is common in many experimental settings and error in Rubisco kinetic measurements appears to scale with measured values as well (Figure A.9). Multiplicative error on a linear scale that errors are normally distributed on a log scale. Least-squares regression assumes that error is normally distributed, so it is sensible to perform regression on a log scale when errors are multiplicative.

A.3 Relationship Between Tradeoff Models and Microscopic Kinetics of Rubisco

Derivation of Rubisco Kinetic Equations

Figure A.1 diagrams the microscopic kinetic scheme for Rubisco carboxylation and oxygenation following the nomenclature of [71]. A complete derivation of the Michaelis-Menten type rate law for Rubisco carboxylation and oxygenation is given in the supplement of Cummins et al. 2018 [71]. The only assumption made in this derivation is that product release (of carboxylation and oxygenation products both) is irreversible, i.e. that k_{10}, $k_{15} = 0$ (Figure A.1). We will use their notation and results in this supplement and encourage readers to refer to that paper for full detail.

Catalytic Efficiencies (k_{cat}/K_M) are Related to the First Effective Barrier to Carboxylation and Oxygenation

One central result from the derivation of [71] is that the catalytic efficiencies for carboxylation and oxygenation can be expressed as:

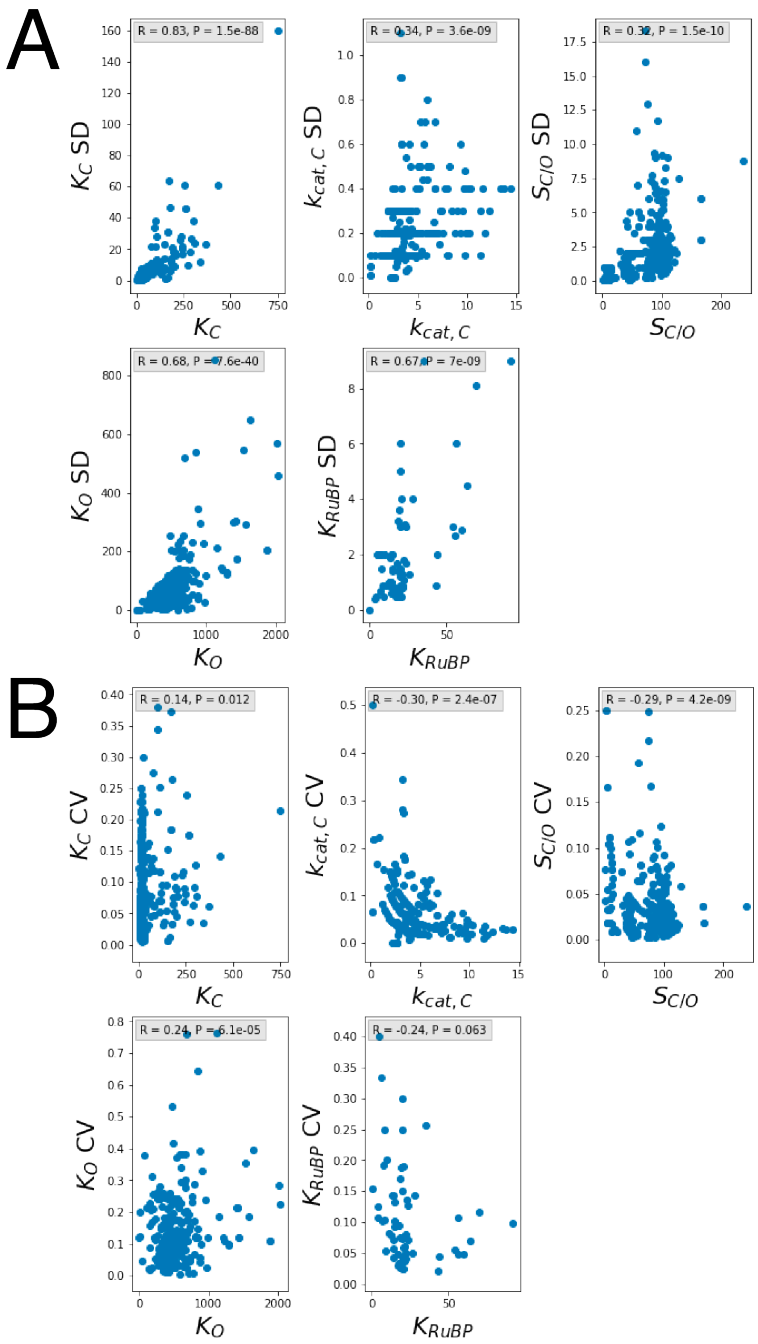


Figure A.9: Scaling of kinetic measurement error with measured value justifies use of regression on a logarithmic scale. Panel (A) shows that reported error for all kinetic measurements scales with the reported mean value, consistent with multiplicative error. $k_{\mathrm{cat,O}}$ is omitted because it is rarely measured directly and experimental error is not usually reported. Panel (B) shows that the coefficient of variation (CV, standard deviation normalized to mean value) does not depend as strongly on the measured value. This is expected if error depends multiplicatively on the measured value.

$$\frac{k_{\text{cat,C}}}{K_{\text{C}}} = \frac{k_{\text{cat,C}} K_R k_5}{k_{\text{cat,C}} + \gamma_C k_6}$$
$$\frac{k_{\text{cat,O}}}{K_{\text{O}}} = \frac{k_{\text{cat,O}} K_R k_{11}}{k_{\text{cat,O}} + \gamma_O k_{12}}$$

Where γ_C and γ_O are defined as

$$\gamma_C = \frac{k_3 k_8 + k_3 k_9}{k_3 k_7 + k_3 k_8 + k_3 k_9 + k_7 k_9}$$
$$\gamma_O = \frac{k_3 k_{14} + k_3 k_{15}}{k_3 k_{13} + k_3 k_{14} + k_3 k_{15} + k_{13} k_{15}}$$

As diagrammed in Figure A.1, k_5 is the rate constant for CO_2 association with the Rubisco-enediolate complex (ER*) and k_6 is the rate constant for CO_2 dissociation from the ERC complex of Rubisco with the carboxylation intermediate. Similarly, k_{11} is the rate of O_2 association ER* complex and k_{12} is the rate constant for O_2 dissociation from the ERO complex of Rubisco with the oxygenation intermediate. $K_R = \frac{k_3}{k_3 + k_4}$ is related to the equilibrium fraction of the Rubisco-RuBP complex found in the enediolate form ($K_E = \frac{ER^*}{\text{ER}} = \frac{k_3}{k_4}$; $K_R = \frac{K_E}{1+K_E}$). In a commentary on Cummins et al. 2018 [71], Tcherkez et al. 2018 [302] summarize experimental evidence that decarboxylation and deoxygenation rates are small in comparison to carboxylation and oxygenation, respectively. As a result of this evidence, it is usually assumed decarboxylation and deoxygenation are small in comparison to the forward reactions (i.e. $k_{\text{cat,C}} \gg \gamma_C k_6$ and $k_{\text{cat,O}} \gg \gamma_O k_{12}$) [262, 300, 302, 301] to derive that

$$\frac{k_{\text{cat,C}}}{K_{\text{C}}} \approx K_R k_5$$
$$\frac{k_{\text{cat,O}}}{K_{\text{O}}} \approx K_R k_{11}$$

There remains disagreement about the validity of this assumption [71, 302]. Nonetheless, under these assumptions $k_{\text{cat,C}}/K_{\text{C}}$ and $k_{\text{cat,O}}/K_{\text{O}}$ are determined by enolization and gas addition rates alone.

Savir et al. [262] apply transition state theory to obtain *effective* barriers to enolization and gas addition for carboxylation and oxygenation from the above expressions.

$$\frac{k_{\text{cat,C}}}{K_{\text{C}}} \propto \exp(-\Delta G_{1,\text{C}})$$
$$\frac{k_{\text{cat,O}}}{K_{\text{O}}} \propto \exp(-\Delta G_{1,\text{O}})$$

Where the effective barriers to enolization and gas addition -$\Delta G_{1,C}$ and -$\Delta G_{1,O}$ are expressed in units of RT in [262]. Note however that the right hand side of these expressions have units of of s^{-1} μM^{-1} meaning that the true barrier heights will depend on the CO_2 and O_2 concentrations [301]. The CO_2 concentration near Rubisco can vary between organisms, especially between those that have a CO_2-concentrating mechanism and those that do not. C3 plants lack a CO_2-concentrating mechanism and so their Rubiscos experience roughly the same CO_2 concentration. Measurements of the CO_2 partial pressure in C_3 plant leaves typically lie between 65-85% of ambient [51, 92], meaning that the soluble CO_2 concentration likely varies by at most 30% within this group. For this reason we highlighted the C_3 plants in describing Figure 2.5B.

$k_{cat,C}$ is Related to the Second Effective Barrier for Carboxylation

Following the notation of [71]

$$k_{\mathrm{cat,C}} = \frac{k_3 k_7 k_9}{k_3 k_7 + k_3 k_8 + k_3 k_9 + k_7 k_9}$$

k_3 is the only term in that expression that relates to enolization. However, if hydration and cleavage are rate limiting, as is often assumed [226, 262], then the k_7 k_9 term should be small compared to the other terms in the denominator as k_7 and k_9 are the rate constants for those two steps (Figure S1). In that limit, k_3 cancels and

$$k_{\mathrm{cat,C}} \approx \frac{k_7 k_9}{k_7 + k_8 + k_9}$$

$k_{\mathrm{cat,C}}$ is therefore roughly independent of enolization (k_3) if hydration and cleavage are rate-limiting for carboxylation. Savir et al. apply transition state theory to obtain an *effective* barriers to hydration and cleavage of the 6-carbon carboxyketone intermediate from the above expression [262].

$$k_{\mathrm{cat,C}} \propto \exp(-\Delta G_{2,C})$$

Where -$\Delta G_{2,C}$ is expressed in units of RT in [262] as above. A similar result can be obtained for $k_{\mathrm{cat,O}}$ but the oxygenation mechanism is poorly understood [300] and such an expression is not required to interpret our main-text figures.

Interpretation of Power-Law Correlation Between $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$

As shown in Chapter 2, we find very strong power-law correlation between $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ (Figures 2.6 and 2.7). These correlations are simplest to understand if we assume

that decarboxylation and deoxygenation rates are effectively 0, i.e. k_6, $k_{12} \approx 0$. In this limit, the catalytic efficiencies can be expressed as functions of enolization and gas addition alone, as described above. A power law relation implies that

$$\frac{k_{\mathrm{cat,O}}}{K_{\mathrm{O}}} \approx \alpha \left(\frac{k_{\mathrm{cat,C}}}{K_{\mathrm{C}}}\right)^{\beta}$$
$$\ln\left(\frac{k_{\mathrm{cat,O}}}{K_{\mathrm{O}}}\right) \approx \ln(\alpha) + \beta \ln\left(\frac{k_{\mathrm{cat,C}}}{K_{\mathrm{C}}}\right)$$

where α and β are constants. Substituting in our above derivation for the catalytic efficiencies:

$$\ln(K_R) + \ln(k_{11}) \approx \ln(\alpha) + \beta \ln(K_R) + \beta \ln(k_5)$$

Empirically, we found $0.93 < \beta < 1.1$ when considering the whole Form I dataset (Figure 2.6). If $\beta = 1.0$ then

$$\ln(K_R) + \ln(k_{11}) \approx \ln(\alpha) + \ln(K_R) + \ln(k_5)$$

Notably, the enolization constant $\mathrm{K_R}$ is the same for both the carboxylation and oxygenation branches of each Rubisco. If k_5 and k_{11} varied independently of each other and their variation was large relative to the variation in $\mathrm{K_R}$, we should not observe power-law correlation between $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$. Therefore, the observation of strong power-law correlation with an exponent of roughly 1.0 suggests that variation in $\mathrm{K_R}$ is the primary driver of variation in $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ (with the caveats described above).

If this description of Rubisco catalysis is correct, then, as shown in Figure 2.7, the specificity factor $S_{\mathrm{C/O}}$ must be roughly constant. This can be seen by noting that

$$\begin{aligned} S_{\mathrm{C/O}} &= \frac{k_{\mathrm{cat,C}} K_{\mathrm{O}}}{K_{\mathrm{C}} k_{\mathrm{cat,O}}} \\ &= \frac{k_5 k_{\mathrm{cat,C}} (k_{\mathrm{cat,O}} + \gamma_O k_{12})}{k_{11} k_{\mathrm{cat,O}} (k_{\mathrm{cat,C}} + \gamma_C k_6)} \end{aligned}$$

Again, if we assume that k_6, $k_{12} \approx 0$ then we can simplify to $S_{\mathrm{C/O}} = \frac{k_5}{k_{11}}$. Taking the natural logarithm

$$\ln(S) = \ln\left(\frac{k_5}{k_{11}}\right) \approx \ln(\alpha)$$

So under this model $S_{C/O}$ should be roughly constant. However, $S_{C/O}$ is not constant: there is 3-4 fold variation in $S_{C/O}$ values among Form I Rubiscos and tenfold variation between Form I and Form II Rubiscos (Figure 2.6B, Figure A.3).

This inconsistency is partially resolved in Figure 2.6C where we plot $k_{cat,C}/K_C$ vs $k_{cat,O}/K_O$ separately for Rubiscos purified from organisms with different physiologies, e.g. C_3 and C_4 plants. This plot shows that each subcategory with sufficient data (e.g. C_3 plants and cyanobacteria) displays a strong and statistically significant a power-law correlation of its own. These power laws appear to differ primarily in prefactor (Y-intercept on a log-log plot) and not in exponent (slope in a log-log plot). 95% confidence intervals on the exponents for C_3 and C_4 plants - the categories with the most data and highest-confidence fits - place them between 0.8 and 1.0.

Derivation of the Active Site Gating Model

The model diagrammed in Figure 2.7A envisions that the Rubisco-RuBP complex can be found in either a "reactive" or "unreactive" state. Note that these states can be directly mapped onto the catalytic mechanism drawn in Figure A.1 as all evidence indicates that the Rubisco-RuBP complex is only reactive to CO_2 and O_2 once RuBP is isomerized to its enediolate form [10, 66]. As such, the "reactive" state is likely related to the Rubisco-enediolate (ER*) complex and the "unreactive" state might be related to the Rubisco-RuBP complex (ER). These states need not to map directly onto one another, however. Conformational rearrangements in Rubisco might also be required to convert between the reactive and unreactive states, for example [48].

In the "reactive" state both CO_2 and O_2 can enter and react with their intrinsic reactivities, denoted by the barrier heights $\Delta G^*_{1,C}$ and $\Delta G^*_{1,O}$. The fractional occupancy of the reactive state is denoted ϕ. As discussed above, this factor ϕ could arise solely from enolization of RuBP but could also be due to other factors - e.g. electrostatics, hydration and protonation of active site residues, as well as protein motions - so long as these factors affect both CO_2 and O_2 equally. Whatever factors determine the reactivity of the active site, we assume that they are fast-equilibrating.

Given these assumptions,

$$k_{cat,C}/K_C \propto \phi \exp(-\Delta G^*_{1,C}/RT)$$
$$k_{cat,O}/K_O \propto \phi \exp(-\Delta G^*_{1,C}/RT)$$

Where the constants of proportionality are assumed to be the same in both cases. If enolization is the only factor determining reactivity, then $\phi = K_R = \frac{K_E}{1+K_E}$ and $\phi^{-1} = e^{\Delta G_E/RT} + 1$. Since

$$\begin{aligned} S_{C/O} &= (k_{cat,C}/K_C)/(k_{cat,O}/K_O) \\ &= \exp((\Delta G^*_{1,O} - \Delta G^*_{1,C})/RT) \end{aligned}$$

and we assumed that the intrinsic reactivities $\Delta G^{*}_{1,C}$ and $\Delta G^{*}_{1,O}$ are organism independent, $S_{C/O}$ is constant under this model. By taking the log of both sides we can also derive the expected power-law correlations.

$$\begin{aligned}\ln(S_{C/O}) &= \ln(k_{cat,C}/K_C) - \ln(k_{cat,O}/K_O) \\ &= (\Delta G^{*}_{1,O} - \Delta G^{*}_{1,C})/RT = C \\ \ln(k_{cat,C}/K_C) &= C + \ln(k_{cat,O}/K_O)\end{aligned}$$

where C is a constant.

At first glance it would seem that $\phi = 1$ would be the best possible outcome for all Rubiscos because this would yield the same constant $S_{C/O}$ but maximize $k_{cat,C}/K_C$. However, this intuition stems from a misunderstanding what $S_{C/O}$ connotes. $S_{C/O}$ is not trivially related to the net rate of carboxylation. Rather $S_{C/O}$ is the slope of the ratio of carboxylation rate to oxygenation rate (R_C/R_O) in the limit of very low CO_2 and O_2 concentrations. This can be seen by inspecting the equations for R_C and R_O given in Figure 2.1.

$$\begin{aligned}R_C &= \frac{k_{cat,C}}{1 + K_C/[CO_2] + K_C[O_2]/(K_O[CO_2])} \\ &= k_{cat,C}[CO_2]([CO_2] + K_C + K_C[O_2]/K_O)^{-1} \\ R_O &= \frac{k_{cat,O}}{1 + K_O/[O_2] + K_O[CO_2]/(K_C[O_2])} \\ &= k_{cat,O}[O_2]([O_2] + K_O + K_O[CO_2]/K_C)^{-1}\end{aligned}$$

In the limit where $[CO_2] \ll K_C$ and $[O_2] \ll K_O$ we get that

$$R_C/R_O = \frac{k_{cat,C}[CO_2]}{k_{cat,O}[O_2]} \frac{(K_O + K_O[CO_2]/K_C)}{(K_C + K_C[O_2]/K_O)}$$

If CO_2 and O_2 concentrations are sufficiently low, we can also neglect the competitive inhibition terms $K_C[O_2]/K_O$ and $K_O[CO_2]/K_C$. In this limit we have that $R_C/R_O = S_{C/O}$. When the CO_2 and O_2 concentrations are not negligible, inspection of R_C and R_O, indicates the relationship between $S_{C/O}$ and the actual ratio of carboxylation and oxygenation rates is neither linear nor simple. As such, we caution readers against the simplistic logic of higher $S_{C/O}$ being better for carboxylation.

Implications of the Active Site Gating Model

The implication of the data presented in Figures 2.6 and 2.7 and the model described above is that ϕ varies within the various Rubisco groups (Figure A.10), perhaps by varying the barrier to RuBP enolization (ΔG_E), which does not appear in the expression for $S_{C/O} = \frac{k_5}{k_{11}}$

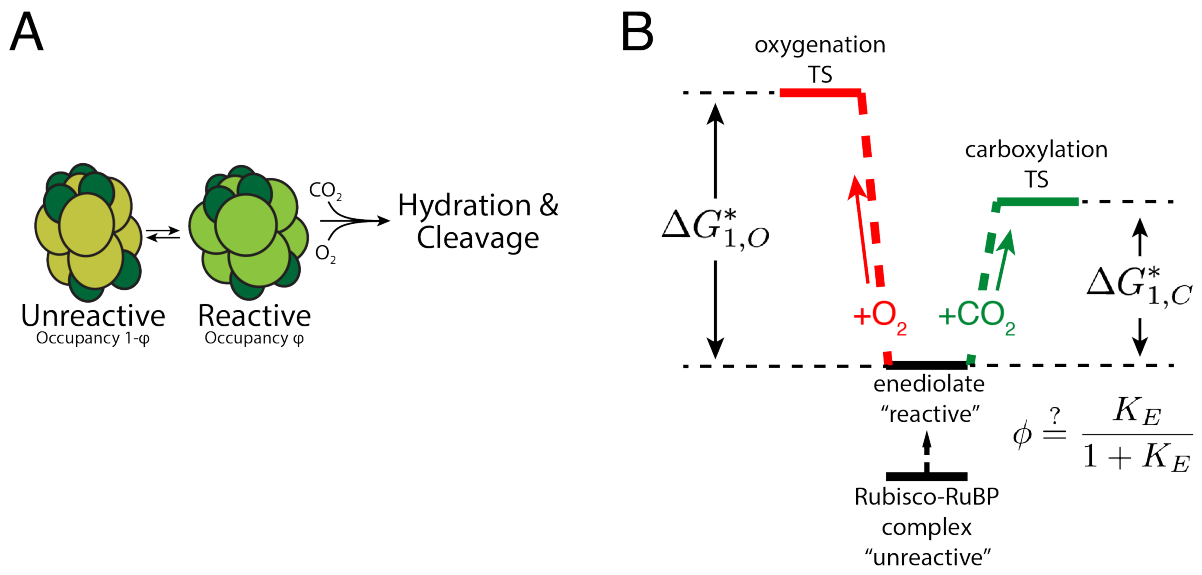


Figure A.10: Description of the relationship between the active site gating model and the accepted mechanism of Rubisco. Panel (A) describes the overall model, in which Rubisco fluctuates between a reactive and unreactive state. Only in the reactive state can gasses react with RuBP. Panel (B) relates this model to the mechanism of Rubisco catalysis. In this model, the first effective barrier to carboxylation and oxygenation are decomposed into their two elementary chemical steps - enolization of RuBP and gas addition. We hypothesize that the fractional occupancy of the reactive state (ϕ) is related to the equilibrium constant for on-enzyme enolization of RuBP (K_E). This hypothesis assumes that enolization is in quasi-equilibrium, i.e. that the forward and reverse rates are fast relative to gas addition. We further hypothesize that the intrinsic gas addition rates are determined by intrinsic barrier heights $\Delta G^*_{1,O}$ and $\Delta G^*_{1,C}$ where the difference between these barrier heights (G*1,O-G*1,C) does not vary between Rubiscos within the same physiological grouping. Presuming this is true, variation in K_E (and hence ϕ) would produce the strong power-law correlations between $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ shown in Figure 2.6 and roughly constant specificity $S_{C/O}$ within the group shown in Figure 2.7. We hypothesize that changes in the conformation of the Rubisco-bound enediolate cause G*1,O-G*1,C to vary across groupings of Rubiscos - i.e. comparing C_3 and C_4 plant enzymes - which would produce the characteristic differences in $S_{C/O}$ between groupings of Rubisco enzymes. See text for derivation of these relationships.

derived above. Varying ϕ will cause a proportional increase in both $k_{cat,C}/K_C$ and $k_{cat,O}/K_O$ which should produce the power law correlation shown in Figure 2.6B-C.

According to this power law, $S_{C/O}$ should be approximately constant. As noted in Chapter 2 and Figures 2, 6 and S3, however, $S_{C/O}$ varies about tenfold between Form I and Form II Rubiscos and 3-4 fold among the Form I enzymes. When we examine Rubiscos isolated from hosts belonging to the same physiological group - e.g. C_3 or C_4 plants - they do display a characteristic and roughly constant $S_{C/O}$ independent of $k_{cat,C}/K_C$ (Figure 2.7B).

One implication of characteristic variation in $S_{C/O}$ between these groups of Form I Rubiscos (C3 plants, C_4 plants, cyanobacteria, etc.) might be that the difference between intrinsic reactivities $\Delta G^*_{1,O} - \Delta G^*_{1,C}$ does vary between the groupings (Figure A.10). As described above, our model implies that $S_{C/O}$ is determined by the difference in intrinsic barriers $\Delta G^*_{1,O} - \Delta G^*_{1,C}$. This mode of variation would produce roughly constant $S_{C/O}$ among C_3 plants while allowing sizable variation in $S_{C/O}$ between C_3 plants, cyanobacteria and proteobacterial Form I Rubiscos. Characteristic variation in $S_{C/O}$ between these

groups of Form I Rubiscos might be understood via the conformational proofreading model [261]. Conformational proofreading offers an alternative means of amplifying specificity in conditions when kinetic proofreading cannot be applied because, for example, the off-rates associated with on- and off-target substrates are negligible [131, 207]. In the conformational proofreading model, intentionally reducing complementarity between enzyme and substrate can lead to increased specificity if the change affects off-target substrates (e.g. O_2) even more than it affects on-target ones (e.g. CO_2).

Characterized Rubisco Mutants Do Not Exceed Wild-Type Tradeoffs

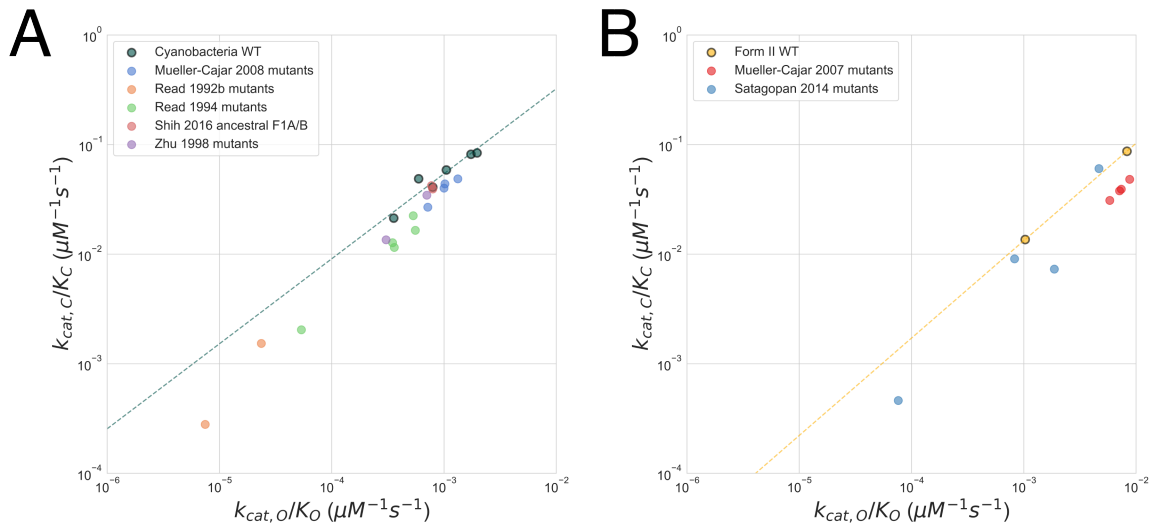


Figure A.11: Mutant cyanobacterial Form I and proteobacterial Form II Rubiscos do not exceed tradeoffs calculated from counterpart wild-type enzymes. Panel (A) plots $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ against $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ for WT and mutant cyanobacterial Form I enzymes. The dashed green line gives the best-fit regression line determined by the WT cyanobacterial enzymes. Note that several of the mutants from the Read & Tabita references [245, 246] have tenfold lower catalytic efficiency towards both CO_2 and O_2 than all WT cyanobacterial enzymes. Panel (B) plots $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ against $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ for WT and mutant Form II Rubiscos from proteobacteria. The dashed yellow line gives the best-fit regression line determined by 3 WT Form II enzymes. In both panels the best fit is determined by ¡ 10 data points and should not be considered high-confidence. Rather the data here are given to suggest that mutant Rubiscos investigated so far do not surpass the catalytic efficiency tradeoff described above and in Chapter 2.

Various schemes have been used to choose Rubisco mutants for characterization. Overall, we extracted data for ≈40 mutant Rubiscos from which 15 were sufficiently well characterized to calculate both catalytic efficiencies $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$. These mutants are predominantly 1-3 amino acid substitutions to wild-type enzymes, but there is also kinetic data on chimeric rubiscos with large and small subunits from different hosts [245] - and reconstructed ancestral rubisco sequences [277]. Cyanobacterial Form I [21, 197, 200, 247, 246, 277, 326]

and proteobacterial Form II [199, 259] Rubiscos are the most commonly mutagenized in our dataset. Figure A.11 plots the $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ against $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ for these mutant enzymes along with the appropriate wild-types (WT) for comparison. Broadly, it appears that the mutants are "worse" than the WT enzymes in that the catalytic efficiency for carboxylation is worse than would be expected based on the trend among WT Rubiscos of the same class. Moreover, several of the cyanobacterial Rubisco mutants are strictly less efficient than the WT enzymes with both $k_{\mathrm{cat,C}}/K_{\mathrm{C}}$ and $k_{\mathrm{cat,O}}/K_{\mathrm{O}}$ being more than tenfold lower than WT values (Figure A.11).

Appendix B

Mathematical Model of the Bacterial CCM

Adapted from the supplementary text of *Proceedings of the National Academy of Sciences of the United States of America*, Volume 113, Mangan NM, Flamholz A, Hood RD, Milo RM, and DF Savage, Pages E5354-62, Copyright 2016, with permission from *PNAS*.

B.1 Relationship to Previous Models of the Cyanobacterial CCM

In a recent paper Mangan and Brenner [182] used both numerical and analytical methods to solve a set of reaction diffusion equations at steady state. In the numerical model the concentration in the carboxysome was allowed to vary spatially. The numerical solutions showed that for a very large range of physically relevant parameters, CO_2 and bicarbonate concentrations had negligible gradients across the carboxysome. As there is negligible carbonic anhydrase and Rubisco activity in the cytosol and the spontaneous dehydration of bicarbonate is relatively slow (see below), cytosolic concentrations have a C = A/r + B dependence (i.e. the solution to the Laplacian). Concentration gradients in the cytosol are therefore incorporated into the analytical model using this dependence. Having noted that there are negligible gradients in the carboxysome and finding that the most physically relevant case is when carbonic anhydrase is unsaturated, Mangan and Brenner were able to solve the entire system analytically. These analytic solutions very closely match the numerical solutions for all relevant parameter regimes [182].

In order to arrive at analytical solutions, Mangan and Brenner assumed fixed pH in all compartments and equal permeability to CO_2 and HCO_3^- at the carboxysome shell [182]. Here we build upon these analytic solutions, incorporating two new features into the earlier model

1. Modeling the pH dependence of membrane permeability and enzymatic activities;

2. Modeling selective uptake of HCO_3^- and selective retention of CO_2 at the carboxysome shell;

These modifications allow us to consider two highly salient questions about the CCM: how the pH in each compartment affects its efficiency and, similarly, how selective uptake at the carboxysome shell might improve CCM performance. Throughout this work, we are careful to stay within the parameter regimes for which the analytic solutions are valid. Complete source code for the updated, pH-aware model is available at github.com/SavageLab/ccm/.

The Mangan and Brenner model is by no means the only previous model of the CCM. An elegant and inspiring early paper by Reinhold et al. employed an analytical model of a standalone carboxysome to investigate the capacity of a proteinaceous compartment to concentrate CO_2 and activate carboxylation [248]. Reinhold et al. proposed that Rubiscos are packed around a central CA in order to guarantee elevated CO_2 concentrations near the Rubisco active site. This result depended on a number of assumptions that now, in light of 25 years of subsequent investigations, appear outmoded: (i) that the number of Rubisco active sites is near the limit of spherical packing and (ii) that Rubisco is arranged around CA active sites. The first assumption would yield approximately 600 Rubisco hexadecamers (4,800 active sites) in a 100 nm carboxysome, 2-3 times more than observed in electron micrographs of α-carboxysomes [135, 263]. Biochemical evidence now puts the second assumption in question, suggesting that carboxysomal carbonic anhydrases are linked to the shell [238].

Later models of the CCM generally employed numerical methods to produce steady-state solutions for the entire CCM system, i.e. including the carboxysome as well as transport and permeability to the cell membrane (sometimes also including the thylakoid membranes). Fridlyand et al. produced a numerical reaction-diffusion model of the CCM that included the permeability of the thylakoid membranes as well as the activity of a CO_2 scavenging system [107]. Hopkinson et al. used measurements of uptake fluxes and photosynthetic rates to fit a numerical model of the Prochlorococcus MED4 CCM [132]. McGrath and Long used a numerical model of the cyanobacterial CCM to investigate how a C_3 plant might benefit from the heterologous expression of CCM components [187]. These models make varying assumptions about the permeability of H_{total} to the cell and thylakoid membranes, but none of these previous models (including that of Mangan and Brenner) accounts for the pH dependence of the H_{total} permeability. Moreover, previous numerical models typically make very charitable assumptions about the permeability of the carboxysome shell.

Long et al., for example, assumed that the carboxysome shell represents no barrier to the passage of HCO_3^- but greatly inhibits the passage of CO_2 such that the permeability values differ by 1000 fold [187]. Hopkinson et al. begin with the unlikely assumption that the cell membrane is entirely impermeable to HCO_3^- (and, by proxy, H_2CO_3) and the carboxysome presents no barrier to HCO_3^- [132]. They proceed to fit the carboxysome permeability to CO_2 based on measured carbon and photosynthetic fluxes, finding a best-fit CO_2 permeability value of roughly 10^{-6} cm/s (i.e. 10^5 fold selectivity between CO_2 and HCO_3^- at the carboxysome shell). Fridlyand et al., in contrast, assume that differences in the carboxysome permeability to CO_2 and HCO_3^- stem from a presumed 10% difference in their diffusion

coefficients within the densely-packed carboxysome lumen [107]. This very small degree of selectivity is compatible with our parsimonious null assumption of equal permeability of CO_2 and HCO_3^- to the carboxysome shell.

B.2 Functional Form of the pH-aware CCM Model

As described in the text, we update the model to include the effects of cellular and external pH on the function of the CO_2 concentrating mechanism (CCM). We incorporate the effect of pH at a number of points in the model. Previously we wrote the fluxes for CO_2, C, and HCO_3^-, H, at the cell membrane, $r = R_b$, as:

$$D\frac{\partial C}{\partial r} = -\ \alpha C_{\text{cytosol}} + k_m^C\,(C_{\text{out}} - C_{\text{cytosol}})$$

$$D\frac{\partial H}{\partial r} = j_c H_{\text{out}} +\ \alpha C_{\text{cytosol}} + k_m^H(H_{\text{out}} - H_{\text{cytosol}})$$

Here D is the diffusion constant for small molecules in water; α, is the rate of CO_2 to HCO_3^- conversion assuming it is unsaturated; and k_m^C and k_m^H are the permeability of the cell membrane to CO_2 and HCO_3^- respectively. The CO_2 to HCO_3^- conversion was previously written using Michaelis-Menten kinetics, but in the code we assume the unsaturated limit as the equations cannot be solved analytically with the full Michaelis-Menten form. As noted in the supplement of our previous paper, the saturated case is less interesting because the rate is a constant, independent of the CO_2 concentration. Therefore $\alpha = \ \alpha/K_\alpha$ from the previous text.

As described in the text, the leakage of HCO_3^- in our original model was too high, motivating us to re-examine our treatment of $k_m^H(H_{\text{out}} - H_{\text{cytosol}})$, the diffusive transport of HCO_3^- across the cell membrane. Here we incorporate the effects of the other hydrated inorganic carbon species, most importantly H_2CO_3, on diffusive transport of inorganic carbon. This treatment ends up requiring separate treatment of the permeability of hydrated inorganic carbon species into the cell and permeability out of the cell. Additionally, the external HCO_3^- concentration will depend strongly on the external pH. We express both the external HCO_3^- concentration and diffusive transport of HCO_3^- such that they depend on pH and the pK_as of the various species of inorganic carbon.

In the cytosol of the model cell there are no chemical reactions. Therefore, when the system is at steady state, diffusion sets the CO_2 and HCO_3^- concentration in the cytosol: $\nabla^2 C = 0$, $\nabla^2 H = 0$. Here ∇^2 is the second derivative in the radial direction, r, in spherical coordinates.

Assuming the same carboxysome permeability, k_c, for both CO_2 and HCO_3^- at the carboxysome shell, diffusive leakage at the carboxysome shell is expressed as:

$$D\frac{\partial C}{\partial r} = k_c\,(C_{\text{cytosol}} - C_{\text{carboxysome}})$$

$$D\frac{\partial H}{\partial r} = k_c(H_{\text{cytosol}} - H_{\text{carboxysome}})$$

Where r = R_c at the boundary of the carboxysome. Inside the carboxysome there is diffusion, carbonic anhydrase (R_{CA}) and Rubisco (R_{Rub}):

$$D\nabla^2 C + R_{\text{CA}} - R_{\text{Rub}} = 0 D\nabla^2 H - R_{\text{CA}} = 0$$

The rate of the reversible carbonic anhydrase reaction takes the form:

$$R_{\text{CA}} = \frac{V_{\text{ba}}K_{\text{ca}}H - V_{\text{ca}}K_{\text{ba}}\text{C}}{K_{\text{ba}}K_{\text{ca}} + K_{\text{ca}}H + K_{\text{ba}}\text{C}}$$

where V_{ba} and V_{ca} are the maximum rates of dehydration and hydration for carbonic anhydrase, respectively. K_{ba} and K_{ca} are the half-maximum concentrations for dehydration and hydration by carbonic anhydrase. R_{CA} is not to be confused with R_C the radius of the carboxysome.

Previously we found that carbonic anhydrase should be unsaturated for the CCM to operate efficiently, past carbonic anhydrase saturation further HCO_3^- transported cannot be utilized by the cell. We also found that diffusion of inorganic carbon within the small carboxysome is fast enough to produce radially uniform concentrations when carbonic anhydrase is unsaturated (presuming reasonable CA reaction rates). In those conditions, CA equilibrates CO_2 and HCO_3^- within the carboxysome. As a result

$$\frac{H_{\text{carboxysome}}}{C_{\text{carboxysome}}} = \frac{V_{\text{ca}}K_{\text{ba}}}{V_{\text{ba}}K_{\text{ca}}}$$

inside the carboxysome. Here $\frac{H_{\text{carboxysome}}}{C_{\text{carboxysome}}} = K'_{\text{eq}}(pH)$ is the pH-dependent equilibrium constant of the CO_2 hydration reaction. This relationship between enzyme kinetic parameters and reaction equilibrium is known as the Haldane relation and is derived by assuming $R_{CA} = 0$ at steady state. We here express the equilibrium ratio as a function of pH and effective pK_a between CO_2 and HCO_3^- (derivations below).

In contrast to CO_2 hydration, Rubisco's carboxylation reaction has $-20\frac{\text{kJ}}{\text{mol}} \leq \Delta_r G'^{\circ} \leq -40\frac{\text{kJ}}{\text{mol}}$ and so likely operates far from equilibrium [102]. Indeed, Rubisco must operate away from equilibrium in order for there to be net CO_2 fixation. Therefore, the pH dependence of the Rubisco reaction cannot be simplified using a Haldane relation. Instead we model Rubisco carboxylation kinetics explicitly and include competitive binding of oxygen:

$$R_{\text{Rub}} = \frac{V_{\text{max}}\text{C}}{K_{\text{M}}\left(1 + \frac{O}{K_{\text{O}}}\right) + C}$$

where C is the CO_2 concentration and O is the O_2 concentration in the carboxysome. Here, $V_{\max}$ is the maximum rate of CO_2 fixation by Rubisco and $K_{\mathrm{M}}(1 + \frac{O}{K_{\mathrm{O}}})$, is the half max concentration of CO_2 including the effect of competitive binding with oxygen. $V_{\max}$ is related to the k_{cat} of the reaction by $V_{\max} = \frac{k_{\mathrm{cat}} \times N_{\mathrm{Rub}}}{V_{\mathrm{csome}} \times N_A}$, where $N_{\mathrm{Rub}}$is the number of Rubisco active sites, V_{csome} is the volume of the carboxysome and N_A is Avagadro's number. The maximum reaction rate, $V_{\max}$, and half-max concentration, K_{M}, are both pH dependent. We incorporate the empirically determined pH dependence of Rubisco carboxylation in a manner discussed below. All kinetic parameters for the carboxysomal Rubisco are given in Table B.2.

The rate of oxygenation assumes the same form, but with the maximum rate of O_2, V_O, and the concentration of O_2 and CO_2, and half max concentration values to O_2, K_{O}, and CO_2, K_m, reversed:

$$R_{\mathrm{Rub,O}} = \frac{V_O O}{K_{\mathrm{O}}\left(1 + \frac{C}{K_{\mathrm{M}}}\right) + O}$$

We use a literature value for K_{O}, but V_O is calculated from the specificity, S, using the relationship $V_O = \frac{V_{\max} K_{\mathrm{O}}}{K_m S}$ [262]. K_{O} is written as K_i in our previous paper [182] and in the literature.

In the following sections we detail how we model the pH dependence of (i) diffusive HCO_3^- transport across the cellular membrane, (ii) the external HCO_3^- concentration, (iii) the carbonic anhydrase induced equilibrium between HCO_3^- and CO_2 in the carboxysome, and (iv) the empirical pH dependence of Rubisco reaction rate. The pH range explored is set by the limits of empirically measured Rubisco values.

B.3 The Effect of pH on C_i Composition and Permeability

Effect of pH on the Equilibrium Composition of C_i

We formulate the equilibrium composition of inorganic carbon concentrations in terms of the pH and pK_as as calculated from chemical formation energies. These pK_a values are used to model the composition of the inorganic carbon pool inside and outside the cell as a function of pH, which has consequences for transport as well as catalytic rates in the carboxysome. For a given inorganic carbon species j, the formation energy, $\Delta_f G'^{\circ}(j)$, can be expressed as a function of the pH [7]:

$$\Delta_f G'^{\circ}(j) = \Delta_f G^{\circ,I}(j) + N_H \mathrm{RT} \log(10) pH$$

As we will be calculating the behavior as a function of internal and external pH, it is useful to separate the formation energy into a pH independent and pH dependent term. The second term on the right hand side represents the direct effect of the H^+ potential (or pH)

on the formation energy. Here N_H is the number of hydrogen atoms in the molecule, R is the gas constant, and T is the temperature.

The pH independent term, $\Delta_f G^{\circ,I}(j)$, is the standard Gibbs free energy corrected for the ionic strength of the solution;

$$\Delta_f G^{\circ,I}(j) = \Delta_f G^{\circ}(j) - \frac{2.91482\,(z^2 - N_H)\,\sqrt{I}}{1 + 1.6\sqrt{I}}$$

The $\Delta_f G^{\circ}(j)$ is the *standard Gibbs free energy of formation* for the species. The second term is the extended Debye-Huckel correction for the non-ideality of solutions with modest ionic strength (0.05 to 0.25 M range). In this term, z is the total charge and I is the ionic strength of the solution in Molar units [7, 211]. Notably, this correction is not accurate at the high ionic strength of ocean water $\approx 0.75\ M$ (BNID 106144), characteristic of oceanic environments. For these environments we use empirical values for the pH independent term, $\Delta_f G^{\circ,I}(j)$.

Alberty provides standard formation energies from equilibrium measurements [7, 211], $\Delta_f G^{\circ}(j)$, for the relevant species: aqueous CO_2, H_2CO_3, HCO_3^- and CO_3^{2-} (Table B.3). When these species equilibrate at a particular pH and ionic strength, their proportions follow the Boltzmann distribution determined by their standard transformed Gibbs energies of formation $\Delta_f G'^{\circ}(j)$ such that species j has concentration

$$[j] \propto exp\left(-\frac{\Delta_f G'^{\circ}(j)}{\mathrm{RT}}\right)$$

One can calculate the pH dependence of the mole fraction a species j by normalizing as follows

$$f_j = \frac{exp\left(-\frac{\Delta_f G'^{\circ}(j)}{\mathrm{RT}}\right)}{\sum_i exp\left(-\frac{\Delta_f G'^{\circ}(i)}{\mathrm{RT}}\right)}.$$

The fractional abundances of H_2CO_3, HCO_3^- and CO_3^{2-} are calculated and shown in Figure B.1. Using the framework of Gibbs energies of formation we can mathematically describe the pH-dependent composition of the equilibrium C_i pools described above and in Chapter 3.

Calculating pK_as for C_i Equilibria

The pK_as calculated here are used to calculate the relative concentrations of the different species of inorganic carbon. Here we demonstrate the calculation of the first pK_a of H_2CO_3, i.e. the pH at which H_2CO_3 and HCO_3^- have equal concentration. This value can also be derived by considering the difference in $\Delta_f G^{\circ}(j)$ energies between the two species:

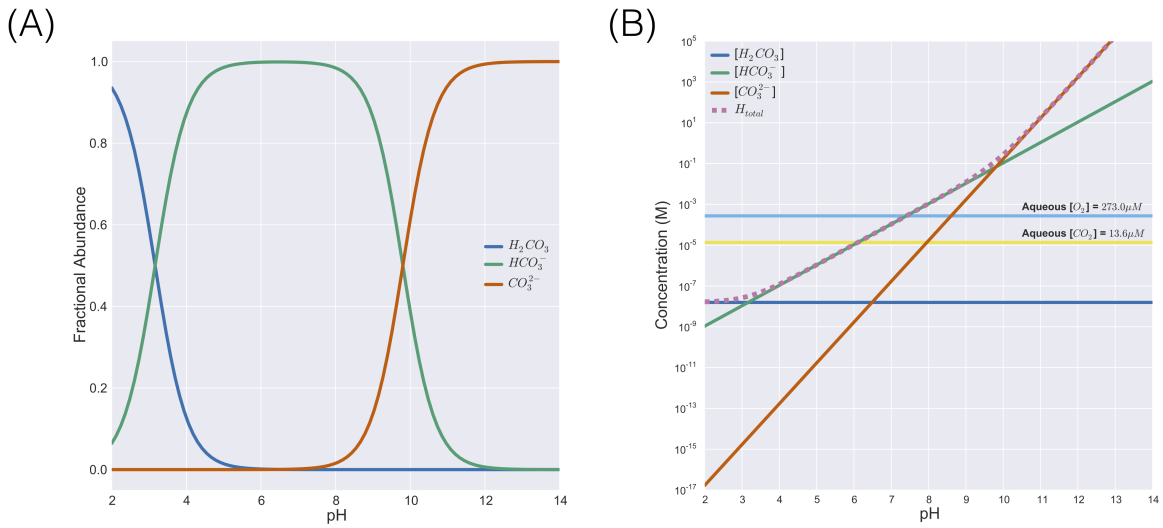


Figure B.1: The effect of pH on the composition and concentration of H_{total} at an ionic strength of 0.2 M. In both panels, the concentration of H_{total} species was calculated from their formation energies. (A) HCO_3^- is the dominant species of Htotal within the pH range considered in this work (pH 6 8.5). As the deprotonation of HCO_3^- to CO_3^{2-} has a pK_a near 10, it is reasonable to ignore the cell-permeability of the doubly-charged CO_3^{2-} in this regime. While H_2CO_3 is similarly a minor constituent of the pool at pH 6-8.5, it is uncharged and so its cell-permeability cannot be neglected. (B) The concentrations of aqueous CO_2 and O_2 were calculated assuming 400 ppm and 21% partial pressures respectively by assuming Henrys law equilibrium between the gas and aqueous phase using solubility coefficients for those species at 25 °C. Notably, O_2 is roughly 20 times as abundant as CO_2 in aqueous solution in equilibrium with present day atmosphere. The concentrations of H_2CO_3, HCO_3^-, and CO_3^{2-} were then calculated from the CO_2 concentration and the pH as in panel (A). Near pH 7.5, the concentration of HCO_3^- surpasses the aqueous concentration of O_2, which provides one intuitive reason for the cyanobacterial CCM to concentrate HCO_3^- in the cytosol rather than CO_2.

$$
\begin{aligned}
K_{a,1} &= \frac{\left[\mathrm{HCO_3^-}\right]\left[\mathrm{H^+}\right]}{\left[\mathrm{H_2CO_3}\right]} \\
&= \exp\left(-\frac{\Delta_f G^{\circ,I}\left(\mathrm{HCO_3^-}\right) + \Delta_f G^{\circ,I}\left(H^+\right) - \Delta_f G^{\circ,I}\left(\mathrm{H_2CO_3}\right)}{\mathrm{RT}}\right) \\
pK_1 &= -\log_{10}\left(K_a\right) = \frac{-\Delta_f G^{\circ,I}\left(\mathrm{H_2CO_3}\right) + \Delta_f G^{\circ,I}\left(\mathrm{HCO_3^-}\right) + \Delta_f G^{\circ,I}\left(\mathrm{H^+}\right)}{\ln\left(10\right)\mathrm{RT}}
\end{aligned}
$$

In biochemical thermodynamic treatments, the formation energy of free protons H^+ is set to 0 and the potential of protons is calculated from the pH [7, 211]. Therefore, using the values in Table B.3 and an ionic strength representative of the cytosol, I = 0.2M, we get a first pK_a for H_2CO_3 of

$$
pK_1 = \frac{\Delta_f G^{\circ,I}\left(\mathrm{HCO_3^-}\right) - \Delta_f G^{\circ,I}\left(\mathrm{H_2CO_3}\right)}{\ln(10)RT} \cong 3.2
$$

The second pK_a– the pH at which HCO_3^- and CO_3^{2-} have equal concentration – can be calculated similarly:

$$pK_2 = \frac{\Delta_f G^{\circ,I}\left(CO_3^{2-}\right) - \Delta_f G^{\circ,I}\left(\mathrm{HCO_3^-}\right)}{\ln(10)RT} \cong 9.8$$

Textbooks sometimes quote a first pK_a of ≈ 6.3 for carbonic acid [94]. This value results from "lumping together" aqueous CO_2 and H_2CO_3, which have the same elemental composition and charge, and calculating the equilibrium of that group with HCO_3^-. We term this value the effective pK_a, which gives the pH at which HCO_3^- and CO_2 have equal concentration.

$$pK_{\mathrm{eff}} = \frac{\Delta_f G^{\circ,I}\left(CO_2\right) - \Delta_f G^{\circ,I}\left(\mathrm{HCO_3^-}\right)}{\ln(10)RT} \cong 6.1$$

These calculations can be corrected for different ionic strengths representative of the cytosol and external freshwater (Table B.6). The ionic strength of seawater is very high ($I \approx 0.75\ M$) and the extended Debye-Huckel approximation is not appropriate. In that regime, we use experimentally determined pK_as which are described in [193] and shown Figure B.2.

As we treat CO_2 separately from the species of H_{total} throughout the model, tracking their concentrations and permeability separately, pK_{eff} is not directly relevant to our modeling of membrane permeability. Rather, we use pK_1 to calculate the ratio of uncharged H_2CO_3 to charged HCO_3^- in the derivations below. Notably, this value is consistent with recent direct spectroscopic measurements of the first pK_a of H_2CO_3 in solution as 3.45 ± 0.15 [2]. Figure B.1A shows the results of these calculations for the case that only H_{total}– H_2CO_3, HCO_3^- and CO_3^{2-}– is in equilibrium, as is assumed to be true in the cyanobacterial cytosol. Python source code for generating Figures B.1 and B.2 is available online at `https://github.com/SavageLab/ccm/`.

Effect of pH on the Permeability and Leakage of H_{total}

Here we derive the effective permeability for HCO_3^- assuming that it is in equilibrium with H_2CO_3. We can neglect leakage of CO_3^{2-}, as it is both very impermeable to the cell membrane due to its double negative charge and also in very low concentration below pH 9 (Figure B.1). Therefore, we can treat H_2CO_3 as a monoprotic acid in order to model its permeability to the cell membrane.

Consider the passive diffusion of an uncharged monoprotic acid HA (e.g. H_2CO_3) and its conjugate base A- (e.g. HCO_3^-) across the cell membrane and into the cell (i.e. positive flux is inward directed). We are interested in the passage of the pool into the cell and the total concentration of the pool $[A_{\mathrm{tot}}] = [HA]_{\mathrm{int}} + [A^-]_{\mathrm{int}}$ inside the cell. Let HA have a permeability P and A^- have permeability coefficient Q, both having units of cm/s and with $Q << P$ because HA is uncharged. Importantly, the pH may differ inside and outside the

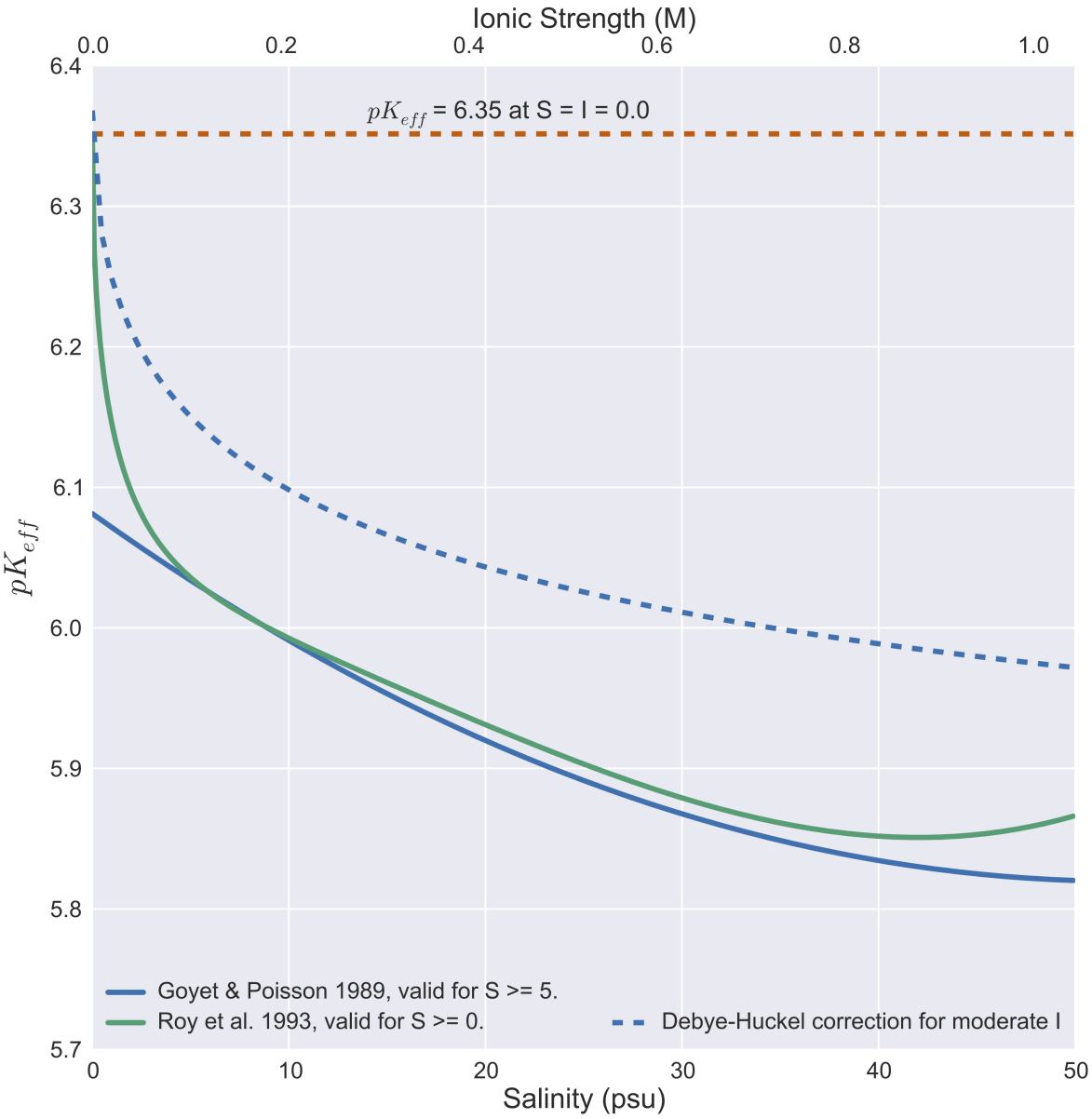


Figure B.2: Measurements of C_i pK_as as a function of ionic strength. Measurements of the dependence of C_i equilibria on salinity are reviewed in [193]. The figure shows the dependence of pK_{eff} on salinity (as defined in the text). Two different polynomial regressions of measurements against ionic strength agree at high salinity, though they diverge quickly from the extended Debye-Huckel treatment as I increases. For this reason, we use measured pK_as when calculating equilibria in seawater.

cell, causing the equilibrium concentrations of HA and A^- to differ across the cell membrane. We'll define $[E] = [HA]_{ext} + [A^-]_{ext}$ as the total exterior pool and C as the total interior pool with [E] and [C] denoting the total concentrations of those pools.

Define $f = [A^-]_{ext} / [E]$ as the fraction of the exterior pool that is deprotonated (A^-) and

let g be the same fraction but for the cytosol (g = $[A^-]_{in}$ / [C]). Note that

$$pK_a = -\log_{10}\frac{[A-]}{[HA]} + pHf = \frac{[A-]_{ext}}{[A-]_{ext} + [HA]_{ext}} = \frac{1}{10^{(pK_a - pH_{ext})} + 1}$$

when the extracellular pool is in equilibrium with the pH. The cytosolic fraction, g, can be calculated in the same way with 'ext' replaced with 'int.' Since protonation is typically very fast (faster than nanosecond timescales) we will assume that both the interior and exterior pools are in equilibrium with the pH. As such, the fractions f and g can be calculated directly from the acid dissociation constant and the pH.

We can calculate the total flux of HA + A^- into the cell according to Fick's law

$$D\frac{d\left[A_{tot}\right]_{in}}{dr} = P\Delta\left[HA\right] + Q\Delta[A^-]$$

Where $\Delta\left[HA\right]$ and $\Delta[A^-]$ are the concentration differences of HA and A^- across the cell membrane. Consider just the right hand side term

$$\begin{aligned} P\Delta\left[HA\right] + Q\Delta\left[A^-\right] &= P\left[E\right](1-f) - P\left[H_{tot}\right](1-g) + Q\left[E\right]f - Q\left[A_{tot}\right]g \\ &= P\left[E\right] - P\left[A_{tot}\right] + g\left[A_{tot}\right](P-Q) - f[E](P-Q) \end{aligned}$$

Assuming $Q \ll P$, we get that $P - Q \approx P$ and the above equation simplifies to

$$\begin{aligned} &\approx P\left[E\right] - P\left[A_{tot}\right] + Pg\left[A_{tot}\right] - Pf\left[E\right] \\ &= P\left[E\right](1-f) - P\left[A_{tot}\right](1-g) \\ &= P\left[HA\right]_{out} - P\left[HA\right]_{cytoplasm} \\ &= P\Delta\left[HA\right] \end{aligned}$$

So long as we appropriately correct for pH, therefore, we can assume that the $Q\Delta[A-]$ contributes negligibly to diffusive transport.

Using only the assumption that Q<<P we have derived that $D\frac{d[A_{tot}]_{in}}{dr} = P\Delta\left[HA\right]$. Now we calculate the rate of A^- leakage from the cell. This is the value of interest in our case because $A^- = HCO_3^-$ is the true substrate of carbonic anhydrase and, therefore, the concentration of interest inside the cell. Using the first equation $[HA] = [A^-] \times 10^{(pK_a - pH)}$ and $[A_{tot}] = [A^-]_{in} + [A^-]_{in} \times 10^{(pK_a - pH)}$. Substituting both of these into our equation for the total flux at the membrane we get

$$D\frac{d\left[A^-\right]_{in}}{dr}\left(1 + 10^{(pK_a - pH_{in})}\right) = P\left(\left[A^-\right]_{out} \times 10^{(pK_a - pH_{out})} - \left[A^-\right]_{in} \times 10^{(pK_a - pH_{in})}\right)$$

Given the first $pK_a = 3.4$ of H_2CO_3, we can neglect the $10^{(pKa - pH_{in})} \ll 1$ for pH ≥ 5, giving:

$$D\frac{d\left[A^{-}\right]_{\text{in}}}{\text{dr}} = \left(\left[A^{-}\right]_{\text{out}} \times P \times 10^{(pK_a - pH_{\text{out}})} - \left[A^{-}\right]_{\text{in}} \times P \times 10^{(pK_a - pH_{\text{in}})}\right)$$

So the effective permeability coefficient for A- can be written as $P \times 10^{(pK_a - pH)}$. Note that since the permeability is now pH dependent and pH may differ inside and outside the cell, you will have different permeability coefficients for the diffusive flux in and out. For the case of HCO_3^- and H_2CO_3, this means that we can write the diffusive flux of HCO_3^- across the membrane as

$$\text{diffusive } HCO_3^- \text{ flux } = \left(\left[HCO_3^-\right]_{\text{out}} \; k_m^{H_2CO_3} \; 10^{(pK_1 - pH_{\text{out}})} - \left[HCO_3^-\right]_{\text{in}} \; k_m^{H_2CO_3} \; 10^{(pK_1 - pH_{\text{in}})}\right)$$

where, pK_1 is the first pK_a of H_2CO_3 as derived in the previous section. In our previous work we assumed that the membrane permeability assumed same value of $k_m^{\text{eff}} = 3 \times 10^{-4}$ cm/s both inside and outside the cell. As $pK_1 \approx 3.2$ and $k_m^{H_2CO_3} \approx 3 \times 10^{-3}$ cm/s, the implied intra- and extracellular pH of our previous work was an implausible $pH = pK_1 - \log_{10}\left[\frac{k_m^{\text{eff}}}{k_m^{H_2CO_3}}\right] = 3.2 - \log_{10}\left[\frac{3\times10^{-4}}{3\times10^{-3}}\right] = 4.2$.

Absolute Permeability of C_i Species

Gutknecht, Bisson and Tosteson measured the membrane permeability of ^{14}C-labeled CO_2 *in vitro* using synthetic lipid bilayers composed of egg-lecithin and cholesterol [116]. These experiments were performed at various pH, with or without added carbonic anhydrase to rapidly equilibrate CO_2 and H_{total} on one side of the membrane. In the limiting case of pH 9-10, the CO_2 flux across the membrane was found to saturate. The saturating flux was used to calculate a CO_2 permeability coefficient of 0.35 cm/s.

The same study measured the membrane conductance of HCO_3^-. The conductance was converted into a one-way flux based on the assumption that all of the ionic flux through the membrane was due to HCO_3^- (producing an upper bound on the flux). This upper bound was then divided by the HCO_3^- concentration to produce an upper bound on the permeability of HCO_3^- of 2×10^{-7} cm/s. Gutknecht et al. also estimate a permeability coefficient of 10^{-3} cm/s for H_2CO_3 on the basis of a measurement of acetic acid ($C_2H_4O_2$). Charge and size are the primary determinants of the membrane permeability of small molecules [251, 329] and acetic acid has the same charge and roughly the same mass as H_2CO_3 ($\approx$ 60 Da). Moreover, the permeability $10^{-3} cm/s$ is consistent with measurements of formic acid (CH_2O_2; $\approx$ 45 Da), which is 25% less massive than acetic and carbonic acid and has a measured permeability coefficient $k_m^{CH_2O_2} \approx 3\times10^{-3}$ cm/s [329]. The per-species permeability values we assume throughout the text are given in Table B.5. On the basis of these primary literature measurements and estimates we are comfortable assuming that HCO_3^- is negligibly permeable compared to H_2CO_3.

Calculating the Extracellular HCO_3^- Concentration

The concentration of HCO_3^- outside the cell is needed to model the loss of HCO_3^- across the membrane in the form of H_2CO_3 leakage. Assuming Henry's law equilibrium with an atmosphere containing $\approx$ 400 PPM CO_2, the aqueous concentration of CO_2 is 13.6 μM, which we 'round' to 15 μM in the model (Table B.4). From the CO_2 concentration we can calculate the size of the total inorganic carbon pool at a particular pH by assuming that the extracellular H_{total} pool is in equilibrium. The total inorganic carbon pool ($H_{\text{total}} + CO_2$) $= \frac{CO_2}{f_{CO_2}}$, where f_{CO_2} is the equilibrium fraction of CO_2 calculated from the Boltzmann distribution:

$$f_{\text{CO}_2} = \frac{\exp\left(-\frac{\Delta_f G'^{\circ}(CO_2)}{\text{RT}}\right)}{\sum_i \exp\left(-\frac{\Delta_f G'^{\circ}(i)}{\text{RT}}\right)}$$

To calculate the HCO_3^- concentration, we use the formula for fractional abundance times the total inorganic carbon, $[HCO_3^-] = f_{\text{HCO}_3^-} \times ([H_{\text{total}}] + [CO_2]) = f_{\text{HCO}_3^-} \times \frac{[CO_2]}{f_{CO_2}}$.

$$f_{\text{HCO}_3^-} = \frac{\exp\left(-\frac{\Delta_f G'^{\circ}\left(\text{HCO}_3^-\right)}{\text{RT}}\right)}{\sum_i \exp\left(-\frac{\Delta_f G'^{\circ}(i)}{\text{RT}}\right)}$$

Notice, that the denominator in both is the same, so

$$[HCO_3^-] = f_{\text{HCO}_3^-} \times \frac{[CO_2]}{f_{CO_2}}$$

$$= [CO_2] \times \exp\left(-\frac{\Delta_f G'^{\circ}\left(\text{HCO}_3^-\right)}{\text{RT}} + \frac{\Delta_f G'^{\circ}\left(CO_2\right)}{\text{RT}}\right)$$

$$= [CO_2] \times 10^{-pKa_{\text{eff}}+pH}$$

We therefore use the pK_{eff} at the appropriate ionic strength in order to calculate the external HCO_3^- concentration as a function of the extracellular pH (as in Figure B.1B).

B.4 pH Dependence of Carboxysomal Enzymes

Carbonic Anhydrase Equilibrates CO_2 and HCO_3^- in the Carboxysomes

As carbonic anhydrase catalyzes a reversible reaction, it equilibrates CO_2 and HCO_3^- in the carboxysome when it is not saturated. To incorporate the pH dependence of this process

into the model we need the pH dependent equilibrium coefficient between CO_2 and HCO_3^-. Similar to above

$$CO_2 + H_2O \longleftrightarrow HCO_3^- + H^+$$

$$\Delta_r G^{\circ,I}\left(\mathrm{CO_2\ to\ HCO_3^-}\right) = \Delta_f G^{\circ,I}\left(HCO_3^-\right) + \Delta_f G^{\circ,I}\left(H^+\right) - \Delta_f G^{\circ,I}\left(CO_2\right) - \Delta_f G^{\circ,I}\left(H_2O\right)$$

$$\frac{\left[HCO_3^-\right]\left[H^+\right]}{\left[CO_2 + H_2O\right]} = \ K_{\mathrm{eq}}^{\mathrm{eff}} = e^{-\frac{\Delta_r G^{\circ,I}\left(\mathrm{CO_2\ to\ HCO_3^-}\right)}{\mathrm{RT}}}$$

where again $\Delta_f G^{\circ,I}\left(H^+\right) = 0$. To get the ratio of $\frac{\left[HCO_3^-\right]}{[CO_2]}$ we can assume water is abundant and obtain the concentration of H^+ from the pH.

$K'_{\mathrm{eq}} = \frac{\left[HCO_3^-\right]}{[CO_2]} = \frac{K_{\mathrm{eq}}^{\mathrm{eff}}}{[H^+]} = K_{\mathrm{eq}}^{\mathrm{eff}} \times 10^{\mathrm{pH}} = 10^{-\mathrm{pK_{eff}}+pH}$

Carbonic anhydrase rapidly drives the reaction to equilibrium, but does not alter the equilibrium ratio at a particular pH. Therefore the relation from the model derivation, $\frac{H_{\mathrm{carboxysome}}}{C_{\mathrm{carboxysome}}} = \frac{V_{\mathrm{ca}}K_{\mathrm{ba}}}{V_{\mathrm{ba}}K_{\mathrm{ca}}}$, must follow the pH dependent ratio $K'_{\mathrm{eq}} = \frac{\left[HCO_3^-\right]}{[CO_2]} = 10^{-\mathrm{pK_{eff}}+pH}$. This pH-dependence relation will hold in the carboxysome as long as the carbonic anhydrase is unsaturated with HCO_3^-. We check that carbonic anhydrase remains unsaturated in our calculations by comparing that the carboxysomal HCO_3^- concentration to the value determined by analytical solutions generated under the assumption that CA is saturated with HCO_3^-. The two solutions match at the transition to CA saturation. So long as the carboxysomal HCO_3^- concentration remains beneath that limit, CA is unsaturated [182].

The pH dependence of the HCO_3^- to CO_2 at equilibrium in the carboxysome means that HCO_3^- will be favored over CO_2 in the carboxysome when the carboxysomal pH exceeds $\mathrm{pK_{eff}}$. The pH dependence of the concentrations of HCO_3^- to CO_2 in the carboxysome with full CCM operation is shown in Figure B.3.

Integrating the Empirical pH-Dependence of Cyanobacterial Rubisco

To model the pH dependence of Rubisco we incorporate the empirically determined pH dependence of the Michaelis-Menten kinetics into our model [18]. This dependence has been measured for the Rubisco from the filamentous cyanobacterium *Anabaena variabilis*. We interpolate the data and rescale the dependence to the Rubisco parameters for *Synechoccocus* PCC 6301 at pH 7.8 [262]. The pH dependence of *A. variabilis* Rubisco was only measured between pH 6 and 8.3, which restricts our analysis of the pH dependence of the CCM. To give a slightly broader range we extrapolate the Rubisco kinetic parameters to pH 8.9. Figure B.4 shows the interpolated and rescaled maximum per-active site catalytic rate, k_{cat} (Figure B.4A), half-maximum CO_2 concentration for carboxylation, K_{M} (Figure B.4B) and

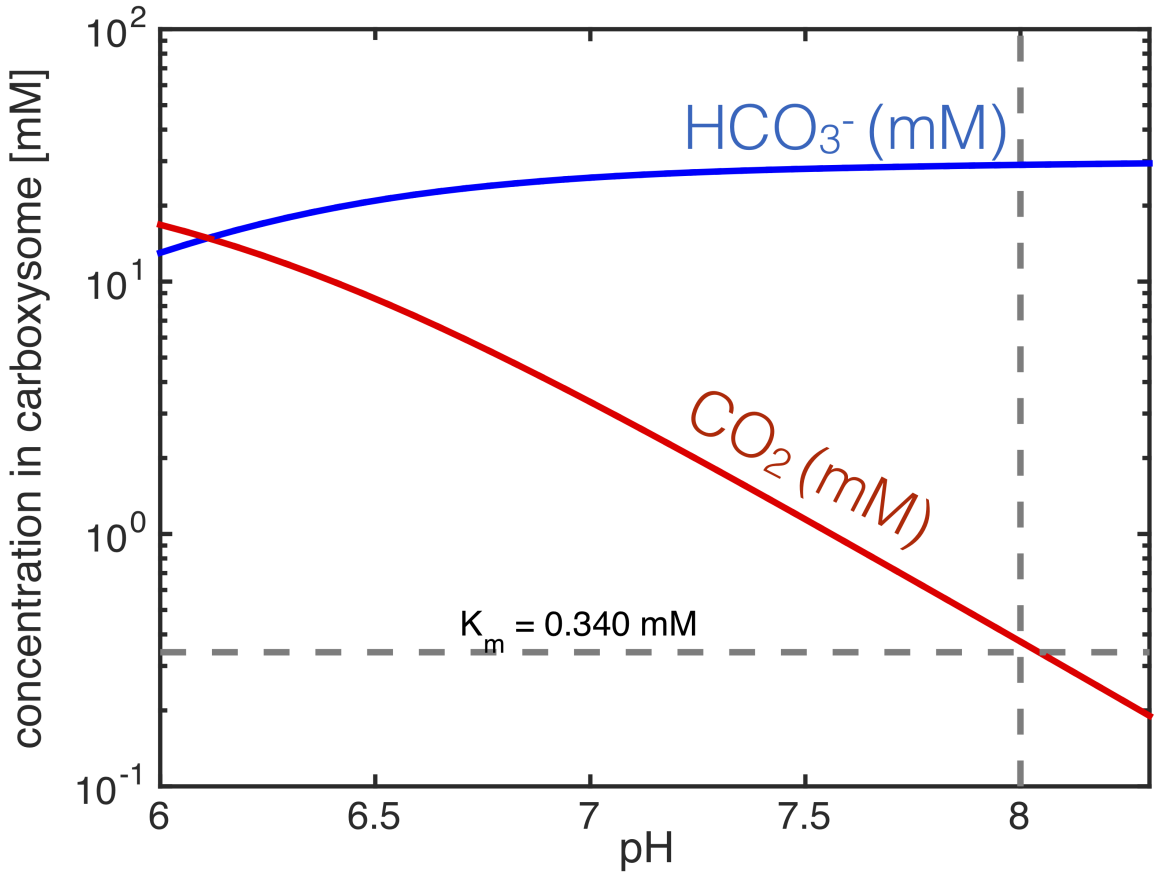


Figure B.3: Carboxysomal concentrations of CO_2 and HCO_3^- as a function of cytosolic pH. Assumes full CCM operation. The HCO_3^- concentration is roughly constant around 10-30 mM, which is sensible because the model enforces a 30 mM cytosolic HCO_3^- concentration. The carboxysomal CO_2 concentration drops from 20 mM to 370 μM from pH 6 to 8 due to the pH-dependence of the CO_2:HCO_3^- equilibrium described in the text.

half-maximum O_2 concentration for oxygenation, K_O (Figure B.4C). We use this rescaled and interpolated data as a lookup table in code.

Justification for Modeling pH-Dependence

To incorporate the pH-dependence into our model, we rely on the analytic solutions derived in our previous work [182]. We previously showed that the analytic solutions match the numerical solution exactly as long as the gradient in the concentration of CO_2 and HCO_3^- are negligible across the carboxysome. As the time required for CO_2 and HCO_3^- to diffuse across the carboxysome is small and the reaction rate of carbonic anhydrase is very fast, the gradients are generally negligibly shallow.

We can rigorously define the regimes in which this condition holds. The equations for HCO_3^- and CO_2 concentration in the carboxysome at steady state are $D\nabla^2 H - R_{CA} =$

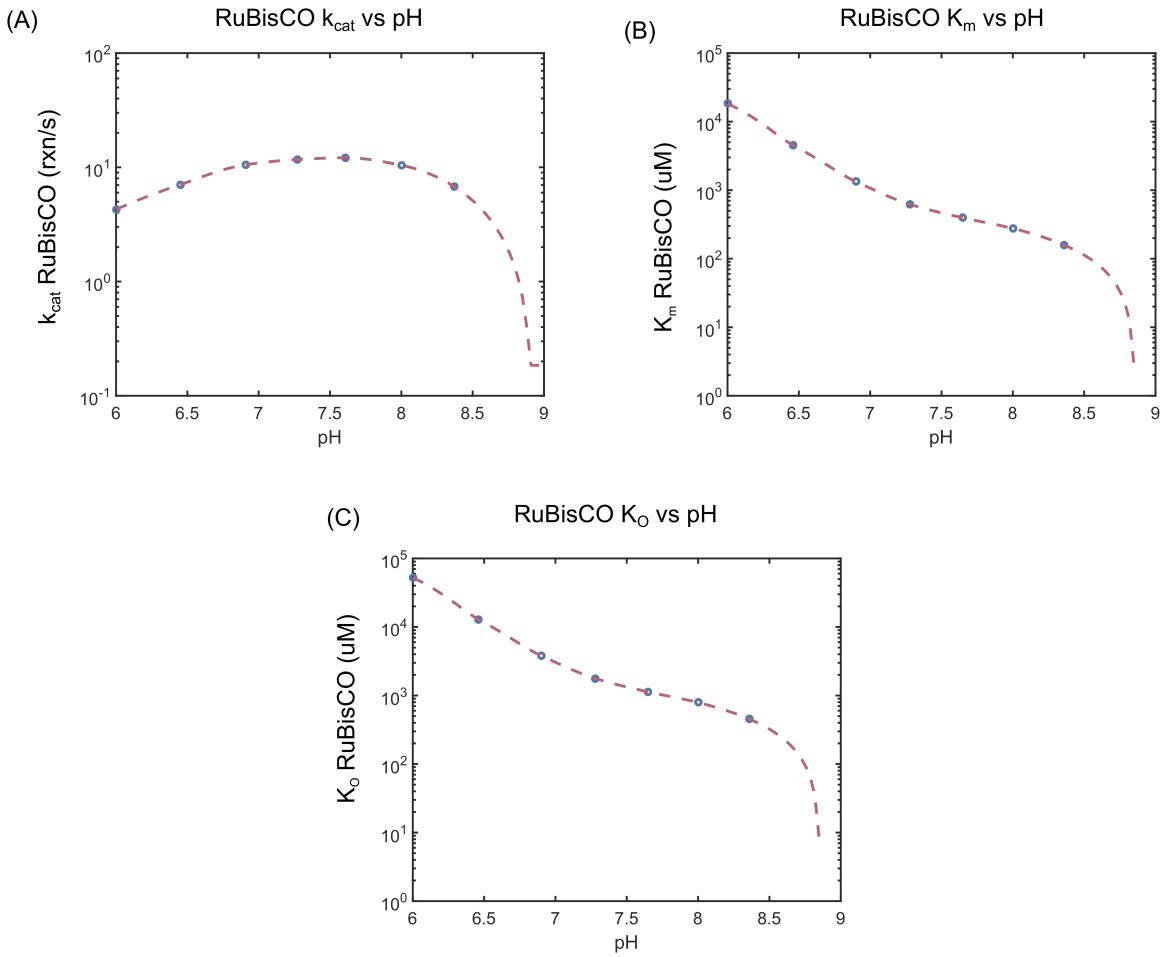


Figure B.4: pH dependence of the Rubisco kinetic parameters. Data for the *Anabena variabilis* enzyme was extracted was rescaled to have a carboxylation $k_{\rm cat}$ of 11.6 s^{-1}, $K_{\rm M}$ of 340 μM and $K_{\rm O}$ of 972 μM at pH 7.8, as measured for the *S. elongatus* PCC 7942 enzyme. Panel (A) shows the interpolated and rescaled carboxylation $k_{\rm cat}$, panel (B) shows the carboxylation $K_{\rm M}$ and panel (C) shows the oxygenation half-maximum concentration $K_{\rm O}$. In all panels, red circles show pH values at which the *Anabena variabilis* enzyme kinetic parameter was measured and the blue interpolated curve was used to calculate the same parameter at intermediate pH values. As shown in panel (A), the maximum kcat of 12 s^{-1} is achieved at pH 7.6 while the half-maximum concentrations $K_{\rm M}$ and $K_{\rm O}$ decrease monotonically between pH 6 and 8.3, as seen in panels (B) and (C).

0 and $D\nabla^2 C + R_{\rm CA} - R_{\rm Rub} = 0$. Here the reaction rates are for carbonic anhydrase, $R_{\rm CA} = \frac{V_{\rm ba}K_{\rm ca}H - V_{\rm ca}K_{\rm ba}{\rm C}}{K_{\rm ba}K_{\rm ca} + K_{\rm ca}H + K_{\rm ba}{\rm C}}$, and Rubisco, $R_{\rm Rub} = \frac{V_{\rm max}{\rm C}}{K'_m(1+\frac{O}{K_i})+C}$, as described above. To determine when the gradients are flat, we compare the magnitude of the diffusion and reaction terms. A rigorous way to make this comparison is to employ non-dimensionalization. To do this we normalize the partial derivative using the relevant length scale, the radius of the carboxysome, and the HCO_3^- and CO_2 concentrations by the appropriate half max reaction rates. This means that the diffusion term $D\nabla^2 H = \frac{{\rm D}\,K_{\rm ba}}{R_c^2}\widetilde{\nabla}^2 h$, and $D\nabla^2 C = \frac{{\rm D}\,K_{\rm ca}}{R_c^2}\widetilde{\nabla}^2 c$. Here the non-dimensional parameters are $h = \frac{H}{K_{\rm ba}}$, $c = \frac{C}{K_{\rm ca}}$, and $\widetilde{\nabla}^2 = R_c^2\nabla^2$. The reaction

terms are then, $R_{\mathrm{CA}} = V_{\mathrm{ba}} \frac{h - \frac{V_{\mathrm{ca}}}{V_{\mathrm{ba}}} c}{1+h+c}$, and $R_{\mathrm{Rub}} = \frac{V_{\max} \mathrm{c}}{\frac{K_m'}{K_{\mathrm{ca}}}(1+\ \frac{O}{K_i})+c}$.

As $\mathrm{C_i}$ is supplied to the system in the form of HCO_3^-, the first term of the carbonic anhydrase rate equation will be larger than the second (also $\frac{V_{\mathrm{ca}}}{V_{\mathrm{ba}}} < 1$) and, therefore, the rate $R_{\mathrm{CA}} \sim \mathrm{V_{ba}}$. Similarly, for reasonable carbonic anhydrase and Rubisco kinetic parameters, $\frac{K_m'}{K_{\mathrm{ca}}}\left(1 + \ \frac{O}{K_i}\right) < 1$, so the Rubisco rate $R_{\mathrm{Rub}} \sim V_{\max}$. The diffusion term scales like $\frac{\mathrm{D}\ K_{\mathrm{ba}}}{R_c^2}$. Now we can compare the relative sizes of these terms. For the HCO_3^- equation diffusion will be negligible when $\frac{\mathrm{D}\ K_{\mathrm{ba}}}{R_c^2} < V_{\mathrm{ba}}$ or, equivalently, $\frac{K_{\mathrm{ba}}}{V_{\mathrm{ba}}} < \frac{R_c^2}{D}$. Diffusion of CO_2 within the carboxysome will be negligible when $\frac{\mathrm{D}}{R_c^2} < \frac{V_{\max}}{K_{\mathrm{ca}}},\ \frac{V_{\mathrm{ba}}}{K_{\mathrm{ca}}}$ or, equivalently, $\frac{K_{\mathrm{ca}}}{V_{\mathrm{ba}}}, \frac{K_{\mathrm{ca}}}{V_{\max}} < \frac{R_c^2}{D}$.

Each of these terms represents time scales. These inequalities indicate that the diffusion gradient will be negligible when the time constant for diffusion across the carboxysome, $\frac{R_c^2}{D}$, is larger than the time scale for the reaction rates $\frac{K_{\mathrm{ca}}}{V_{\mathrm{ba}}}, \frac{K_{\mathrm{ca}}}{V_{\max}}$, and $\frac{K_{\mathrm{ba}}}{V_{\mathrm{ba}}}$. This relation and assumption holds for our use of the pH-aware model. It also shows us that other parameters will not change or break the assumption that the concentrations of CO_2 and HCO_3^- are constant across the carboxysome.

Therefore, we can utilize the analytic solutions derived in the supplement of our previous paper. We insert our pH-dependent derivations for the external HCO_3^- concentration, cell membrane permeability to HCO_3^-, and carbonic anhydrase equilibrated CO_2 to HCO_3^- concentration ratio within the carboxysome.

B.5 Derivation of an pH-Aware Analytical Model

In this section we provide justification for various features of the CCM model and explain the derivation of simplified equations used to generate figures presented here and in Chapter 3. For a more complete mathematical treatment, see below.

Justification for Ignoring Spontaneous Interconversion of CO_2 and HCO_3^-

HCO_3^- is spontaneously dehydrated to CO_2 even in the absence of carbonic anhydrase [181]. This spontaneous equilibration process has a timescale of 10-15 seconds at room temperature (Figure B.5). In order to produce an analytically solvable model, we ignored spontaneous dehydration of HCO_3^-. This assumption is well justified: the spontaneous dehydration of HCO_3^- is 4-6 orders of magnitude slower than diffusion in the cytosol, diffusion across the carboxysome shell and the aggregate catalytic activity of the carboxysomal Rubisco.

The cytosolic small molecule diffusion coefficient $D = 1 \times 10^{-5} \frac{cm^2}{s}$ used in the model gives a timescale of $t = \frac{\mathrm{distance}^2}{6 \times D} = \frac{\left(10^{-4}\ \mathrm{cm}\right)^2}{6 \times 10^{-5} cm^2/s} = 1.7 \times 10^{-4}$ s to traverse a 1 $\mu m = 10^{-4}$cm length scale (roughly the size of a cyanobacterial cell). So diffusion of CO_2 and HCO_3^- in the cytosol diffusion is much faster than the spontaneous dehydration reaction.

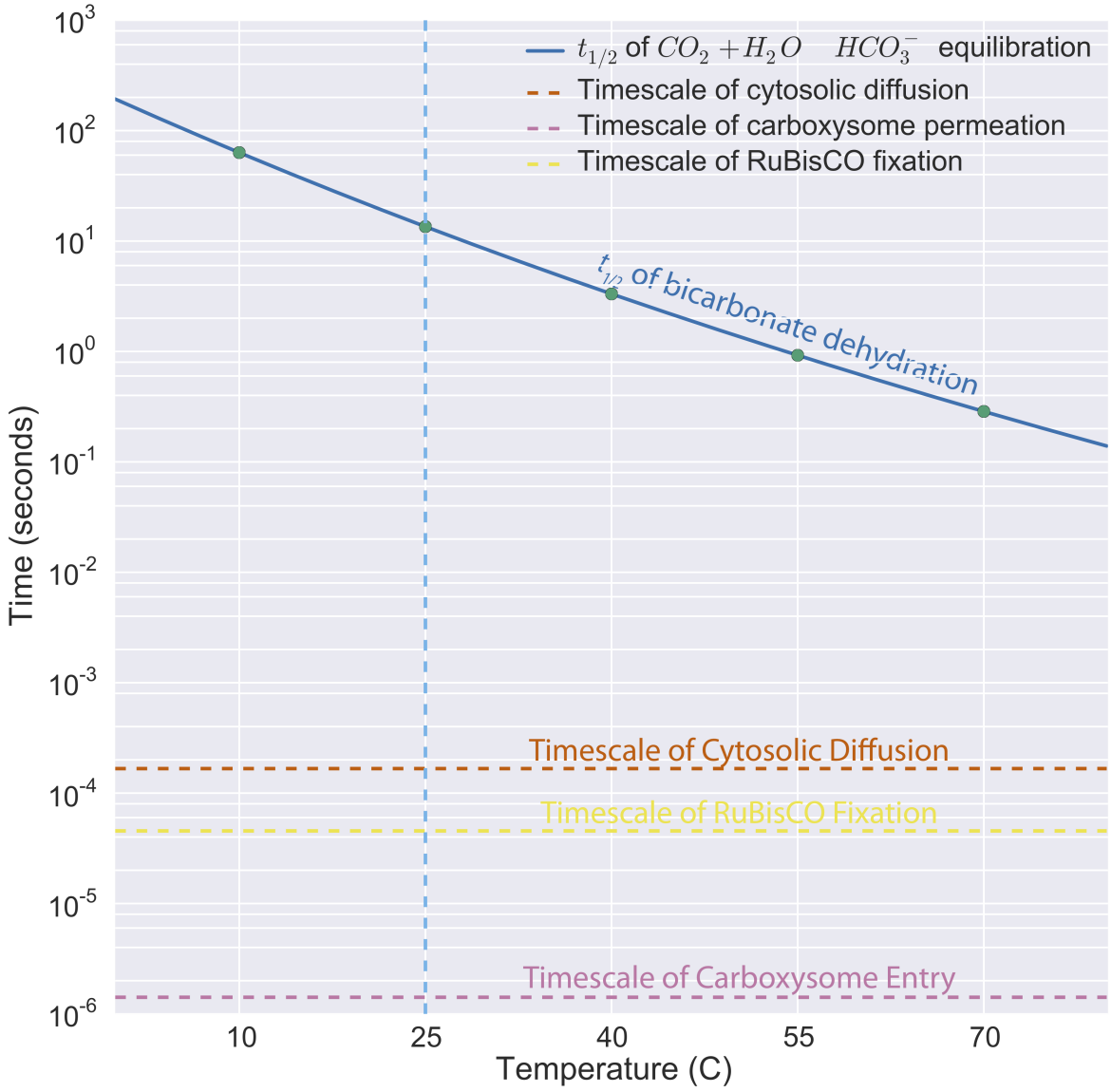


Figure B.5: Spontaneous dehydration of HCO_3^- is negligibly slow. The temperature-dependent timescale of spontaneous HCO_3^- dehydration was calculated using the measured Arrhenius dependence of the relevant rate constants from [181]. The timescales associated with diffusion in the cytosol, entry into the carboxysome and Rubisco fixation were calculated using model parameters (with no temperature dependence) as described. These processes are all calculated to be at least 10,000 times faster than the spontaneous dehydration of HCO_3^- at 25 °C.

If the carboxysome lumen is at pH 8 and equilibrates a 30 mM pool of cytosolic HCO_3^-, it will produce a carboxysomal CO_2 concentration of roughly 500 μM, yielding a 500 μM HCO_3^- concentration gradient across the carboxysome (i.e. $\left[HCO_3^-\right]_{\text{cytosol}} - \left[HCO_3^-\right]_{\text{carboxysome}} = 500\ \mu M$). Assuming a carboxysome permeability in the optimal regime determined by the model ($\sim 10^{-5}\ cm/s$) and a carboxysome surface area around $2 \times 10^{-10}\ cm^2$ then about

6×10^5 HCO_3^- molecules traverse the shell per second, giving a timescale of 1.7×10^{-6} seconds for traversal of the shell. If the carboxysome was more acidic, the gradient would be larger and the timescale even faster. See sections below for the derivation of the optimal carboxysome permeability, calculation of the carboxysome surface area and analysis of pH gradients across the carboxysome shell.

In order to calculate a rough timescale for CO_2 fixation in the carboxysome, we note that there are ≈2000 Rubisco active sites in the carboxysome [135, 263] and each site has a k_{cat} for CO_2 around 10 s^{-1} (Table B.2). Assuming the CCM is effective, all Rubisco active sites are nearly saturated and the carboxysome can catalyze 2×10^4 fixations per second, giving a fixation timescale around 5×10^{-5} seconds. The results of this calculation are shown diagrammatically in Figure B.5.

Derivation of a Carbon Flux Balance Equation

In order to determine the HCO_3^- transport rate needed to support a particular cytosolic HCO_3^- concentration, we write a conservation law that includes the fluxes across the cell membrane as well as CO_2 fixation in the carboxysome. When Rubisco is saturated, the CO_2 fixation rate will equal the total fixation rate integrated over the volume of the carboxysome, $\int V_{\max} dV_{\text{carboxysome}} = \frac{4}{3}\pi R_c^3 V_{\max}$. This rate of fixation must be balanced by the total flux of all inorganic carbon at the cell membrane, $\int D\frac{\partial(C+H)}{\partial r} d(SA_{\text{cell}})$. Here the fluxes are defined by the boundary conditions at the cell membrane:

$$
\begin{aligned}
D\frac{\partial C}{\partial r} &= -\alpha C_{\text{cytosol}} + k_m^C \left(C_{\text{out}} - C_{\text{cytosol}}\right) \\
D\frac{\partial H}{\partial r} &= j_c H_{\text{out}} + \alpha C_{\text{cytosol}} + \\
&\quad \left(H_{\text{out}}\ k_m^{HCO_3^-}\ 10^{(pK_1 - pH_{\text{out}})} - H_{\text{cytosol}}\ k_m^{HCO_3^-}\ 10^{\left(pK_1 - pH_{\text{cytosol}}\right)}\right)
\end{aligned}
$$

Where α is the first-order rate constant associated with facilitated uptake at the thylakoid membrane (diagrammed in Figure 3.1). All of these concentrations and fluxes are evaluated at the cell membrane. This integral gives

$$
\int D\frac{\partial(C+H)}{\partial r} d(SA_{\text{cell}}) = 4\pi R_b^2 \times
\begin{bmatrix}
j_c H_{\text{out}} + \ \alpha C_{\text{cytosol}} \\
+ \left(H_{\text{out}}\ k_m^{H_2CO_3}\ 10^{(pK_1 - pH_{\text{out}})} - H_{\text{cytosol}}\ k_m^{H_2CO_3}\ 10^{\left(pK_1 - pH_{\text{cytosol}}\right)}\right) \\
- \ \alpha C_{\text{cytosol}} + k_m^C \left(C_{\text{out}} - C_{\text{cytosol}}\right)
\end{bmatrix}
$$

Notice that the conversion from CO_2 to HCO_3^- ($\alpha C_{\text{cytosol}}$) cancels exactly between the two equations. Assuming that Rubisco is saturated with CO_2 (or nearly so), the total fixation

flux can be calculated as $\frac{4}{3}\pi R_c^3 V_{\max}$. Setting the flux across cell surface area integral equal to the total CO_2 fixation gives

$$\begin{aligned}\frac{4}{3}\pi R_c^3 V_{\max} &= 4\pi R_b^2 \times \\ &\quad \left[j_c H_{\text{out}} + \left(H_{\text{out}} k_m^{\text{H}_2\text{CO}_3} 10^{(pK_1 - pH_{\text{out}})} - H_{\text{cytosol}}\ k_m^{\text{H}_2\text{CO}_3} 10^{\left(pK_1 - pH_{\text{cytosol}}\right)}\right)\right] \\ &\quad + k_m^C \left(C_{\text{out}} - C_{\text{cytosol}}\right) \\ j_c H_{\text{out}} &= \frac{R_b^2}{3R_c^3} V_{\max} - \left(H_{\text{out}} k_m^{\text{H}_2\text{CO}_3} 10^{(pK_1 - pH_{\text{out}})} - H_{\text{cytosol}} k_m^{\text{H}_2\text{CO}_3} 10^{\left(pK_1 - pH_{\text{cytosol}}\right)}\right) \\ &\quad - k_m^C \left(C_{\text{out}} - C_{\text{cytosol}}\right)\end{aligned}$$

If we are using this equation to calculate the flux needed to maintain a given internal inorganic carbon concentration, we can proscribe the internal HCO_3^- concentration, H_{cytosol}. This internal pool of hydrated C_i has been measured at roughly 30 mM [320, 327, 295], and HCO_3^- is the dominant contributor to this pool at pH > 5. The CO_2 concentration in the cytosol, C_{cytosol}, depends on all the other parameters including the carboxysome permeability. We note that, as derived, this conservation law applies only when Rubisco is near saturation. The full model must be used to derive the transport rate when the carboxysomal CO_2 concentration is beneath Rubisco's K_M for carboxylation.

Phase Space of CCM Function Under Varying HCO_3^- Transport and Carboxysome Permeability

Figure B.6 shows the regions of effective CCM function for varying HCO_3^- transport and carboxysome permeability. For a given carboxysome permeability, we calculate the minimum HCO_3^- transport velocity sufficient to saturate Rubisco. These values define the line between Rubisco saturation and drawn in Figure B.6. To the left of this line is a dark blue region wherein the carboxysomal CO_2 concentration too low to saturate Rubisco. Similarly we can solve for the HCO_3^- transport velocity and carboxysome permeability values where carbonic anhydrase is saturated, defining the maroon region in Figure B.6 where energy is wasted importing excess HCO_3^- that cannot be converted to CO_2. In the green region (between the maroon and dark blue) conditions are "just right" and produce "efficient fixation," where inorganic carbon transport saturates Rubisco with CO_2, activating carboxylation and suppressing oxygenation without expending excess energy on transport. The curves dividing each region are lines of equal carboxysomal CO_2 concentration.

The shape of the region of efficient fixation is largely unchanged from previous work. However, the updated model requires between 10^2-10^4 times less HCO_3^- transport to achieve efficient CO_2 fixation. The optimal carboxysome permeability (black dashed line in Figure B.6) is also about thirty-fold lower than previously reported ($3 \times 10^{-5} \frac{\text{cm}}{s}$ instead of $10^{-3} \frac{\text{cm}}{s}$), although permeabilities as high as 10^{-2} cm/s could support efficient carboxylation. The

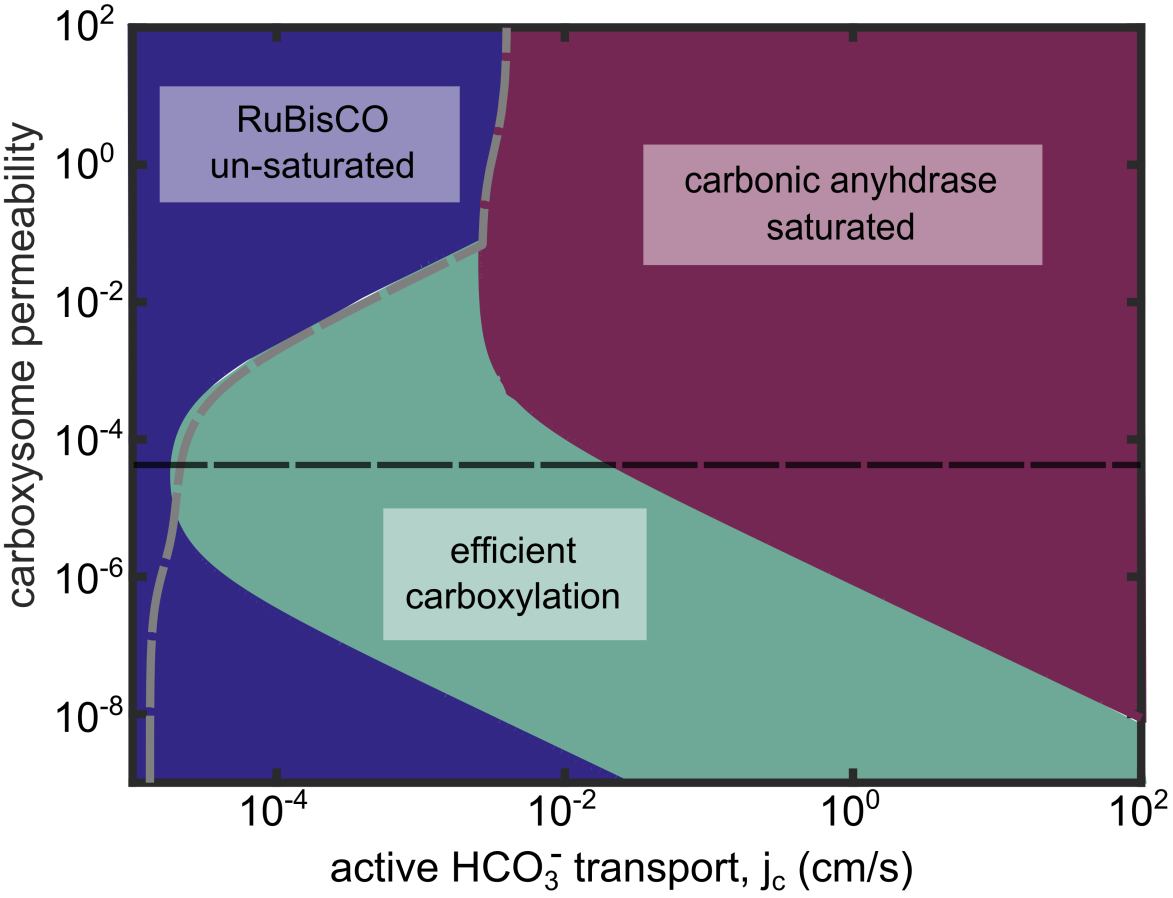


Figure B.6: Phase space defining regimes of efficient carbon fixation. The CCM model depends on two crucial unknown parameters the amount of energetically activated C_i uptake and the permeability of the carboxysome to CO_2 and HCO_3^-. The phase space examines how the CCMs capacity to produce efficient CO_2 fixation (defined as saturating Rubisco for carboxylation without saturating the carboxysomal carbonic anhydrase) depends on the value of these two unknown parameters, where jc is the velocity of energetically activated HCO_3^- transport. In the left-most blue region Rubisco is unsaturated and increasing the carboxysomal CO_2 concentration would elevate the carboxylation rate. Carbonic anhydrase is saturated in the right-most maroon region and so energy is wasted transporting excess HCO_3^- that cannot be converted to CO_2 by CA. A carboxysome permeability $\approx 3 \times 10^{-5}$ cm/s allows for efficient fixation at a minimum transport velocity $j_c \approx 10^{-5}$ cm/s. Carboxysome permeabilities as high as $\approx 10^{-2}$ cm/s cm/s can produce efficient fixation at the expense of 100-1000 fold higher transport rates. These calculations were performed assuming cytosolic pH = 8 and external pH = 7, incorporating the pH dependence of carbonic anhydrase, Rubisco, and the membrane permeability as described in the text. The grey dashed line indicates the values of carboxysome permeability and HCO_3^- transport where the CCM generates a 30mM HCO_3^- concentration in the cytosol.

optimal single carboxysome permeability is set by two conflicting mechanisms: the benefit of better trapping CO_2 near Rubisco versus the detriment of slowed HCO_3^- entry into the carboxysome. The optimal carboxysome permeability is now lower because carboxysomal CA produces CO_2 inside the cell and CO_2 leakage from the cell is, at pH 8, a much larger

contributor to overall carbon loss due to the CCM. In Chapter 3 and below we demonstrate that allowing the carboxysome to be selective, i.e. have different permeabilities to CO_2 and HCO_3^-, would not drastically improve its performance. A carboxysome with a more acidic pH than the cytosol would, however, enable efficient carboxylation with lower HCO_3^- influx, as discussed below.

The updated phase space also displays interdependence between the rate of transport required to produce 30 mM HCO_3^- in the cytosol (dashed grey line) and the carboxysome permeability. This dependence can be understood as follows: at pH 8, where HCO_3^- permeability to the membrane is greatly reduced, a much larger fraction of HCO_3^- influx is ultimately fixed (Figure 3.2). The rate of CO_2 fixation is now a substantial contributor to the carbon flux balance derived above. We showed previously that carboxysome permeability strongly influences the availability of CO_2 to Rubisco and thus, the rate of CO_2 fixation [182]. As HCO_3^- influx and CO_2 fixation are both substantial contributors to the flux balance at pH 8 (Figure 3.2), it follows the rate of HCO_3^- transport should also depend on the carboxysome permeability.

B.6 Enegetic Cost Calculations

The Cost of C_i Transport

Model cyanobacteria like *S. elongatus* PCC7942 have a number of energetically activated HCO_3^- transporters. These transporters generally fall into two classes based on the mode of energetic activation: Na^+/ HCO_3^- symporters and ATP-binding cassette (ABC) type transporters [238]. In *S. elongatus*, the ABC-type transporter is encoded by the *cmpABCD* operon [209]. The *cmp* genes are induced by limiting C_i conditions but are not required for growth in those conditions: a mutant deficient in the periplasmic binding protein *cmpA* was shown to have wild-type like growth in low C_i [209]. As such, it seems likely that the Na^+/ HCO_3^- symporters are the primary transport mechanism in freshwater cyanobacteria, dissipating one unit of Na^+ gradient per HCO_3^- translocated.

Cyanobacterial genomes encode Na^+:H^+ antiporters that are located on the cell membrane [41]. Cyanobacterial Na^+:H^+ antiporters are homologous to the *E. coli* NhaA transporter, whose structure has been solved and appears to display a 2:1 H^+:Na^+ stoichiometry [161]. Overexpression of these genes in freshwater cyanobacteria was shown to confer high salt tolerance [312]. However, knockout of these antiporters induces no defect in salt tolerance [85], implying some other functional role for these transporters. Here we assume that these antiporters exchange Na^+ for H^+ at steady state, allowing us write the energetic cost of HCO_3^- influx as 2 H^+ per HCO_3^- translocated as a first approximation.

This value is likely an underestimate of the transport cost for two reasons. First, a fraction of HCO_3^- influx is likely due to the ABC transport system, which hydrolyzes one ATP per transport reaction. According to the stoichiometry of the cyanobacterial and chloroplast ATP synthase, synthesis of one ATP dissipates 4 units of H+ gradient [309]. Second, the

molecular biology of Na^+ gradient maintenance in cyanobacteria is not well characterized. Indeed, the fact that antiporter mutants are viable [85] suggests that some other factors are involved in Na^+ homeostasis. It may well be the case that other energetic currencies are dissipated in the maintenance of the Na^+ gradient. As such, we proceed on the assumption of a 4:1 H^+:Na^+ stoichiometry. This assumption is parsimonious as it assigns the same cost to ATP and Na^+ activated bicarbonate transport. Below, we perform a detailed sensitivity of the model results to this and several other assumptions.

We do not attempt to calculate the energetic cost associated with facilitated CO_2 uptake. It is hypothesized that vectorial CO_2 hydration is energetically activated via coupling to NAD(P)H oxidation. This hypothesis is based on the homology of facilitated uptake systems to the *E. coli* respiratory complex NAD(P)H:Quinone oxidoreductase [238, 32]. However, NAD(P)H binding domains are absent from the cyanobacterial operons [32] and so it remains unclear whether (i) the uptake system is coupled to electron flow at all, (ii) if NAD(P)H is the electron donor and (iii) if H^+ are pumped across the thylakoid membrane in tandem with CO_2 hydration. These factors prevent us from attempting to estimate the energetic cost associated with facilitated CO_2 uptake.

The Cost of Carbon Fixation

The energetic cost of carbon fixation was approximated according to the per-CO_2 ATP and NADPH consumption of the Calvin-Benson-Bassham cycle. The CBB cycle stoichiometry is diagrammed in Figure B.7. In order to convert ATP and NADPH into units of H^+ gradient dissipation, two numbers are required: the H^+:ATP stoichiometry of the cyanobacterial ATP synthase and the prevailing H^+:NADPH ratio in photosynthesizing cyanobacteria.

As mentioned above, the H^+:ATP ratio of the cyanobacterial ATP synthase is 4:1 [309]. However, the cyanobacterial ATP synthase is resident on both the cell and thylakoid membranes [274]. When the resident on the thylakoid membrane, the ATP synthase dissipates the H^+ gradient between the thylakoid lumen and cytosol through ATP synthesis. When resident on the cell membrane the synthase dissipates gradient across that membrane. In approximating the cost of transport we considered the H^+ gradient across the cell membrane as it is unclear how to convert thylakoid gradient dissipation into these units. For lack of a better approach, we proceed by ignoring the distinction between these two separate electrochemical gradients.

Either a respiratory or a photosynthetic calculation can be used to estimate the H^+:NADPH ratio of growing cyanobacteria. Based on the presence of the b_6f complex in the cyanobacterial respiratory chain, which is analogous to the mitochondrial bc_1 complex [203], we would assume a mitochondrial-like respiratory H^+:NADPH ratio of 10:1. However, cyanobacteria primarily respire in the dark and so it is unclear whether this number is relevant to cyanobacteria in the light. We can also calculate the H^+:NADPH ratio by considering cyclic and linear electron flow through the photosystems as follows. The Z scheme (linear electron flow) produces 2 NADPH for every 9 photons and pumps 12 H^+ across the thylakoid mem-

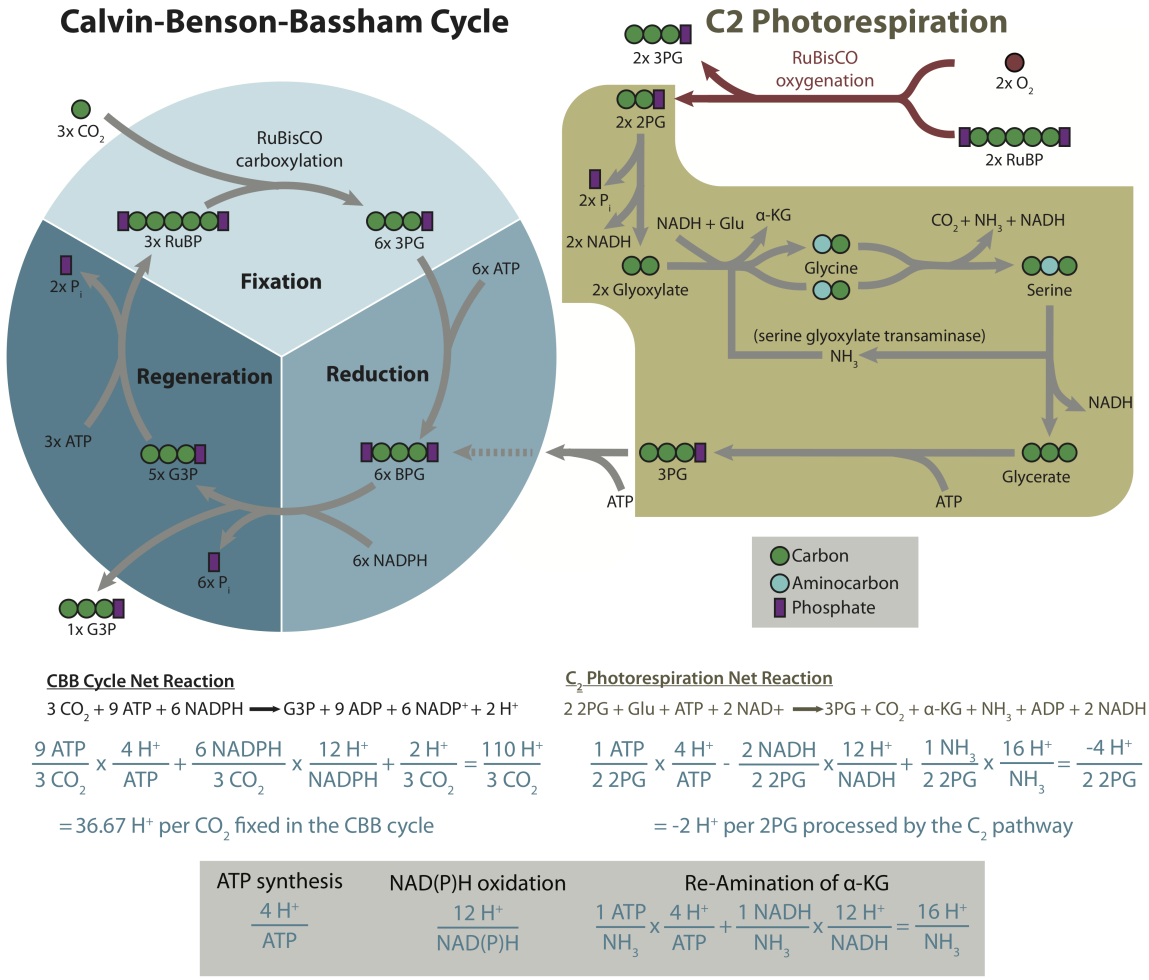


Figure B.7: Calculation of the energetic cost of the CBB cycle and C_2 photorespiratory pathway. A simplified diagram of the Calvin-Benson-Bassham cycle (left) and C_2 photorespiratory pathway (right) are given so that the path of carbon can be followed. The energetic costs of each pathway are then calculated based on the stoichiometry of ATP synthase and the presumed stoichiometry of the photosynthetic apparatus. We assumed α-ketoglutarate is converted to glutamate through the GOGAT pathway, providing a cost estimate for the photorespiratory de-amination. The full energetic cost of fixing carbon and recycling oxygenation products depends on the ratio of the rates through these two pathways. See text for explanation of how we integrate these two separate estimates to calculate the total cost of carbon fixation and photorespiration in a given condition. Abbreviations: P_i (orthophosphate), G3P (glyceraldehyde 3-phosphate), BPG (bisphophoglycerate), Glu (glutamate), α-KG (α-ketoglutarate), 3PG (3-phosphoglycerate), 2PG (2-phosphoglycolate).

brane [242]. Cyclic electron flow would use those same 9 photons to pump 36 H+ across the thylakoid membrane and make no NADPH [242]. As such,

$$9 \text{ photons } = 36\ H^+ = 2\ NADPH + 12\ H^+ \ \Rightarrow\ NADPH = 12\ H^+$$

As these two numbers (10 or 12 H^+ per NADPH) are similar, we proceed with the number based on photosynthetic stoichiometry. Neither number is perfectly accurate as both include

H^+ pumped across the thylakoid membrane, while we have assumed that H^+ interconverts with Na^+ at the cell membrane to activate bicarbonate transport. Regardless, we treat these as reasonable estimates for the H^+ gradient dissipation required for ATP and NADPH synthesis in cyanobacteria.

Per CO_2, the CBB cycle consumes 3 ATP and 2 NADPH. The carboxylation reaction itself produces two new carboxylic acids that are deprotonated near physiological pH. As a result, two additional H^+ are produced for every 3 carboxylation events. Therefore, the energetic cost associated with CBB fixation is

$$\frac{3\ ATP}{CO_2} \times \frac{4\ H^+}{\text{ATP}} + \frac{2\ NADPH}{CO_2} \times \frac{12\ H^+}{\text{ATP}} + \frac{2\ H^+}{3\ CO_2} = 36.67 \frac{H^+}{CO_2}$$

We write the net reaction of the CBB cycle per-CO_2 as

$$CO_2 + 36.67\ H^+ \rightarrow \frac{1}{3} G3P$$

The Cost of Photorespiration

Unlike higher plants, model cyanobacteria have multiple photorespiratory pathways [84]. We assumed that the plant-like C_2 pathway is the primary 2PG recovery pathway in cyanobacteria as well. This assumption is based on the broad conservation of the C_2 cycle and the observation that it decarboxylates less than the other cyanobacterial pathways (once for every two 2PG processed). Further research is required to determine the relative fluxes through the photorespiratory pathways of model cyanobacteria. We note that this far from a trivial goal: flux measurement in phototrophs is complicated by the fact that C1 compounds like CO_2 and HCO_3^- do not admit multiple carbon labeling patterns [177, 332]. Time-varying labeling patterns and high CO_2 concentrations are generally used to measure fluxes in phototrophs [332], making these measurement conditions inappropriate for investigations of the CCM.

The energetic cost of the C_2 pathway is calculated following the pathway diagram in Figure B.7. So that we can balance photorespiration with the CBB calculation, we treat the C_2 cycle as consuming G3P and producing CO_2 (via decarboxylation in the C_2 pathway). We begin with the condensation of G3P into RuBP (in the CBB cycle) and its subsequent oxygentation and then trace the path of carbon through the C_2 pathway and back to G3P through the CBB cycle. The cost of this process can be calculated in steps as follows:

1. Condensation in the CBB Cycle: $\frac{5}{3}\ G3P \rightarrow RuBP$
2. Oxygenation: $RuBP + O_2 \rightarrow 3PG + 2PG$
3. Photorespiration: $3PG + 2PG \rightarrow 3PG + \frac{1}{2}3PG + \frac{1}{2}CO_2$
4. Reduction in the CBB cycle: $1.5\ 3PG + \frac{1}{2}CO_2 \rightarrow 1.5\ G3P + \frac{1}{2}CO_2$

Condensation of 5 G3P into 3 RuBP in the CBB cycle consumes 3 ATP, or one ATP for every 5/3 G3P (Figure B.7). Oxygenation is not energetically activated and so costs nothing. Based on the stoichiometry of the C_2 pathway (Figure B.7), photorespiration is calculated to cost

$$\frac{1}{2}\frac{\text{ATP}}{2PG} \times \frac{4\ H^+}{\text{ATP}} - \frac{2}{2}\frac{\text{NADPH}}{2PG} \times \frac{12\ \text{H}^+}{\text{NADPH}} + \frac{1}{2}\frac{NH_3}{2PG} \times \frac{16\ \text{H}^+}{NH_3} = -2\frac{H^+}{2PG}$$

Giving a cost of -2 H+ for every 2PG molecule entering the photorespiratory pathway. Reduction of 1.5 3PG into 1.5 G3P through the CBB cycle costs 1.5 ATP and 1.5 NADPH. The net cost of the whole process is therefore

$$1\ ATP\ \times \frac{4\ H^+}{\text{ATP}} - 2\ H^+ + 1.5\ ATP\ \times \frac{4\ H^+}{\text{ATP}} + 1.5\ NADPH \times \frac{12H^+}{\text{NADPH}} = 26\ H^+$$

and we can write the net reaction per-G3P as

$$\frac{1}{6}G3P + 26H^+ \rightarrow \frac{1}{2}\ CO_2$$

This net reaction may seem pointless as it simply converts reduced carbon in the form of G3P into inorganic CO_2. Remember, however, that photorespiration represents a response to oxygenation, which would have wasted 2 reduced carbons (as 2PG) instead of $\frac{1}{2}$ a carbon if not for photorespiratory carbon recovery.

The Total Cost of CO_2 Fixation and Photorespiration

A simple approach to calculating the total cost of carbon fixation is to take the direct sum of the costs due to fixation and the costs due to photorespiration (i.e. multiplying the per-pathway costs by their respective fluxes and summing). However, the C_2 pathway results in one decarboxylation for every two 2PG processed (Figure B.7). So our analysis must also include the cost associated with "re-carboxylating." The simplest way to do this is to "blame" the C_2 pathway for the cost of re-carboxylation, which increases the cost of photorespiration by $\frac{1}{2}$ the cost of one turn of the CBB cycle, or $\frac{1}{2} \times 36.67\ \approx 18.3\ H^+$. Under this assumption, the total cost of photorespiration equals $26\ + 18.3\ = 44.3\frac{H^+}{2PG}$.

The above assumption oversimplifies, however, because "re-carboxylating" through the CBB cycle will result in oxygenation some fraction of the time (that fraction depending on the CO_2 and O_2 concentrations). In conditions where the oxygenation flux is high, this "second-order" oxygenation may contribute non-negligibly to the total cost of fixing CO_2. The following calculation implicitly includes the cost of secondary oxygenation by ensuring that photorespiration and fixation are mass-balanced regardless of the carboxylation and oxygenation fluxes.

Given a particular ratio of oxygenation to carboxylation ($R_o = V_o/V_c$), we can balance the net reactions above to calculate the total cost of fixing carbon and recovering the products of oxygenation.

$$\left(CO_2 + 36.67\ H^+ \rightarrow \frac{1}{3}G3P\right) + R_O \times \left(\frac{1}{6}G3P + 26H^+ \rightarrow \frac{1}{2}\ CO_2\right)$$

$$(36.67 + 26R_O)\, H^+ + \frac{2 - R_O}{2}CO_2 \rightarrow \frac{2 - R_O}{6}G3P$$

$$\frac{(36.67 + 26R_O)}{1 - \frac{R_O}{2}}\frac{H^+}{CO_2} + CO_2 \rightarrow \frac{1}{3}G3P$$

To sanity check this calculation we first note that it is carbon-balanced: there is one carbon atom in CO_2 and one carbon atom in 1/3 of a G3P. Second, the calculation clearly produces negative costs when $R_O \geq 2$. This is desirable because $R_O \geq 2$ means that the oxygenation rate is at least twice the carboxylation rate and so there is as much (or more) photorespiratory decarboxylation as carbon fixation. In this regime net carbon fixation is impossible, and so it is appropriate that the cost function diverges at that point. Here and in Chapter 3 we to restrict analysis to regimes where $R_O \leq 2$.

In C_3 plants $R_O \approx 0.3$ [271], giving a per-CO_2 cost of $\frac{(36.67+26\times0.3)}{1-0.15} = 52.3\frac{H^+}{CO_2}$. As the oxygenation flux is high, the cost of secondary oxygenation is not negligible in this case. This can be seen by calculating the total cost using the "first-order" approximation above, which gives $\frac{V_c\times36.67+V_c\times44.3}{V_c} = 36.67 + 0.3 \times 44.3 = 50\ \frac{H^+}{CO_2}$.

Notably, this cost calculation assumes that CO_2 released by photorespiratory decarboxylation doesn't drastically alter the CO_2 concentration, which would in turn change R_O. This is especially concerning when we model the energetic cost of a cytosolic Rubisco with no CCM (see below) because R_O might be as high as 0.3 in atmospheric CO_2 and even higher when CO_2 is limiting. Here we justify the use of this calculation even in those extreme cases. The frame of the argument is as follows. When there is no CCM, only two fluxes produce CO_2 in the cytosol: diffusion across the membrane and photorespiration. If the diffusional timescale is much faster than the photorespiratory one, then the photorespiratory contribution to the cytosolic CO_2 concentration will be negligible.

If the absolute rate of carboxylation is on the order of $10^{-20}\frac{\text{mol}}{cell\times s}$ (as produced by the model here and measured in [132]) then the absolute rate of photorespiration is roughly $0.5 \times 0.3 \times 10^{-20}\frac{\text{mol}}{cell\times s} = 1.5 \times 10^{-21}\frac{\text{mol}}{cell\times s}$ as there is one decarboxylation for every two 2PG entering the C_2 pathway (Figure B.7). If that CO_2 production was trapped in a model cell with volume ≈0.5 fL, it would produce an additional 3 $\mu M\frac{CO_2}{cell\times s}$. Since the membrane permeability to CO_2 is here taken to be 0.3 cm/s, the accumulated 3 μM of cytosolic CO_2 would cross the membrane diffusively at a rate $0.3\frac{\text{cm}}{s} \times 3\mu M = 9\times10^{-7}\frac{\text{mol}}{cm^2\times s}$. Multiplying by the model cell's surface area ($3.14 \times 10^{-8}\ cm^2$) we get $2.8 \times 10^{-14}\frac{\text{mol}}{cell\times s}$. So even in the limiting case of high photorespiration, the rate at which CO_2 leaks through the membrane is 6-7 orders of magnitude faster than the rate at which it is produced through photorespiration. Even when R_O approaches 2, at the very limits of feasibility of CBB cycle-based autotrophy, the rate of photorespiration would have a negligible effect on the equilibrium of CO_2 across

the membrane. As such, it is reasonable to neglect the effects of photorespiratory CO_2 production on the cytosolic CO_2 concentration.

H^+ Production and Consumption by Carbonic Anhydrase

Carbonic anhydrase catalyzes the dehydration of HCO_3^- to $CO_2 + H_2O$. Based on diverse structural and enzymatic studies, HCO_3^- is the true substrate of CA [296]. However, as HCO_3^- equilibrates spontaneously with H_2CO_3 and CO_3^{2-}, we can think of the CA reaction as equilibrating the H_{total} pool with $CO_2 + H_2O$. Across the pH range we consider (pH 6-9), a large fraction of H_{total} pool is deprotonated (Figure B.1). CA activity will, therefore, tend to consume protons by converting a partially deprotonated pool of H_{total} into a completely protonated pool of H_2O, thereby counteracting cytosolic acidification due to transport.

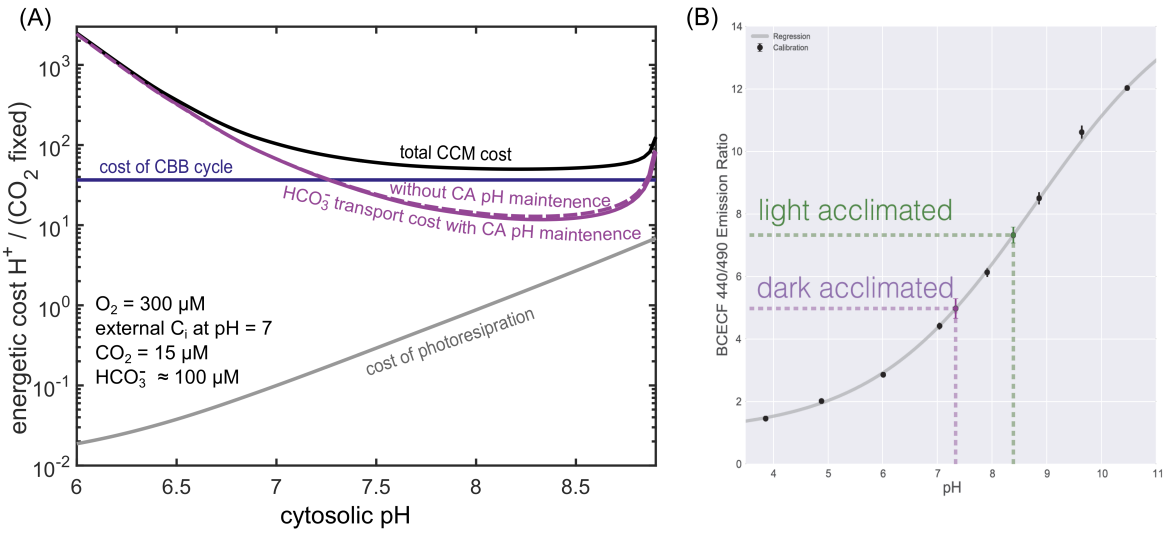


Figure B.8: The pH dependent cost of fixing CO_2 using a CCM. Panel (A) is the same as Figure 3.3, but explicitly includes the cost-compensation due to the activity of CA. In the CCM, CA primarily dehydrates HCO_3^- to $CO_2 + H_2O$, thereby absorbing protons transported into the cell and partially ameliorating the cost of transporting C_i. The black curve shows the total energetic cost of operating the CCM on a per-fixation basis. The blue horizontal line denotes the cost of the CBB cycle. The solid maroon solid line shows the cost of HCO_3^- transport including the correction for pH maintenance due to carbonic anhydrases dehydration activity. The dashed maroon line represents the cost of HCO_3^- transport without the correction. The grey curve gives the cost of photorespiration. Panel (B) shows the calibration curve and measurement of the cytosolic pH of *S. elongatus* PCC 7942. The intracellular pH of *S. elongatus* was measured using the ratiometric pH dye BCECF-AM. The calibration curve was generated by measuring the 490:440 emission ratio in the presence of the ionophore nigericin in BG11 media with defined pH and was then fit to a Boltzmann sigmoid (Methods). The intracellular pH measurement was carried out by incubating light- and dark-acclimated *S. elongatus* cultures with BCECF-AM for 30 min in the light or dark as appropriate (Methods). The dark-acclimated culture was found to have a cytosolic pH $= 7.3 \pm 0.2$, which differs markedly from the pH $= 8.4 \pm 0.1$ measured for light-acclimated cells (error given is the standard deviation of the measured 490:440 emission ratio converted into pH units via the calibration curve).

We account for cost compensation due to CA by subtracting the flux through carbonic anhydrase from the H^+ cost of bicarbonate uptake. The corrected and uncorrected cost

of transport is plotted in Figure B.8 and accounts for the pH-dependence of CA kinetics discussed above. We included CA-based cost compensation in the total cost and transport costs plotted in Figure 3.3 as well. As expected, cost compensation due to CA is greater at increased pH where H_{total} leakage is diminished. Notably, H^+ produced by Rubisco catalysis are accounted for in our analysis of the energetic cost of carbon fixation and photorespiration, as described above.

B.7 Quantitative Evaluation of CCM Variants and Alternatives

Effectiveness of C_i Uptake Mechanisms Depends on Extracellular pH

Two distinct energy-coupled transport mechanisms are associated with the CCM: facilitated CO_2 uptake and HCO_3^- transport. Note that facilitated CO_2 uptake expends energy to convert cytosolic CO_2 into HCO_3^-, while the active HCO_3^- transport imports HCO_3^- from outside the cell. As the mechanism of facilitated CO_2 uptake is not known, however, we cannot compare these mechanisms in energetic terms. We compare the effectiveness of the systems at different external pH by assuming that they have the same velocity, i.e. that the cell's capacity for CO_2 and HCO_3^- uptake are equal, and using the model to calculate the fraction of C_i uptake due to each system. Based on this analysis, facilitated CO_2 uptake is projected to contribute equally to inorganic carbon influx at an external pH of 6.2, i.e. when the external inorganic carbon pool is roughly 50% CO_2 (Figure B.9). This contrasts with our previous work using a single permeability coefficient for H_{total}, which found that facilitated uptake only contributes substantially to carbon influx when the extracellular C_i pool is at least 80% CO_2, i.e. beneath pH 5.8 [182].

At an extracellular pH of 7 and ionic strength of 50 mM (approximating freshwater), the two transport fluxes differ by only fivefold (Figure B.9). It is possible to close this gap by assuming that the cell has fivefold more capacity for facilitated uptake or by assuming HCO_3^- transport consumes more energy. At an oceanic pH of 8 and ionic strength of 750 mM, however, there is a $\approx$ 100-fold difference between the fluxes through the two transport systems, suggesting that facilitated CO_2 uptake is not a useful strategy in the ocean. This is consistent with the phylogenetic dispersion of the two transport systems, where freshwater cyanobacteria (pH $\approx$ 7) have genes coding for both transport systems while the oceanic cyanobacteria generally have only genes coding for HCO_3^- transport [132, 231].

Energetic Cost of the CCM in Varying CO_2 Concentrations

C_3 land plants grow well in present day atmosphere but have no CCM. Instead, they employ a relatively high-specificity Rubisco to fix carbon in the stroma of the chloroplast (topologically

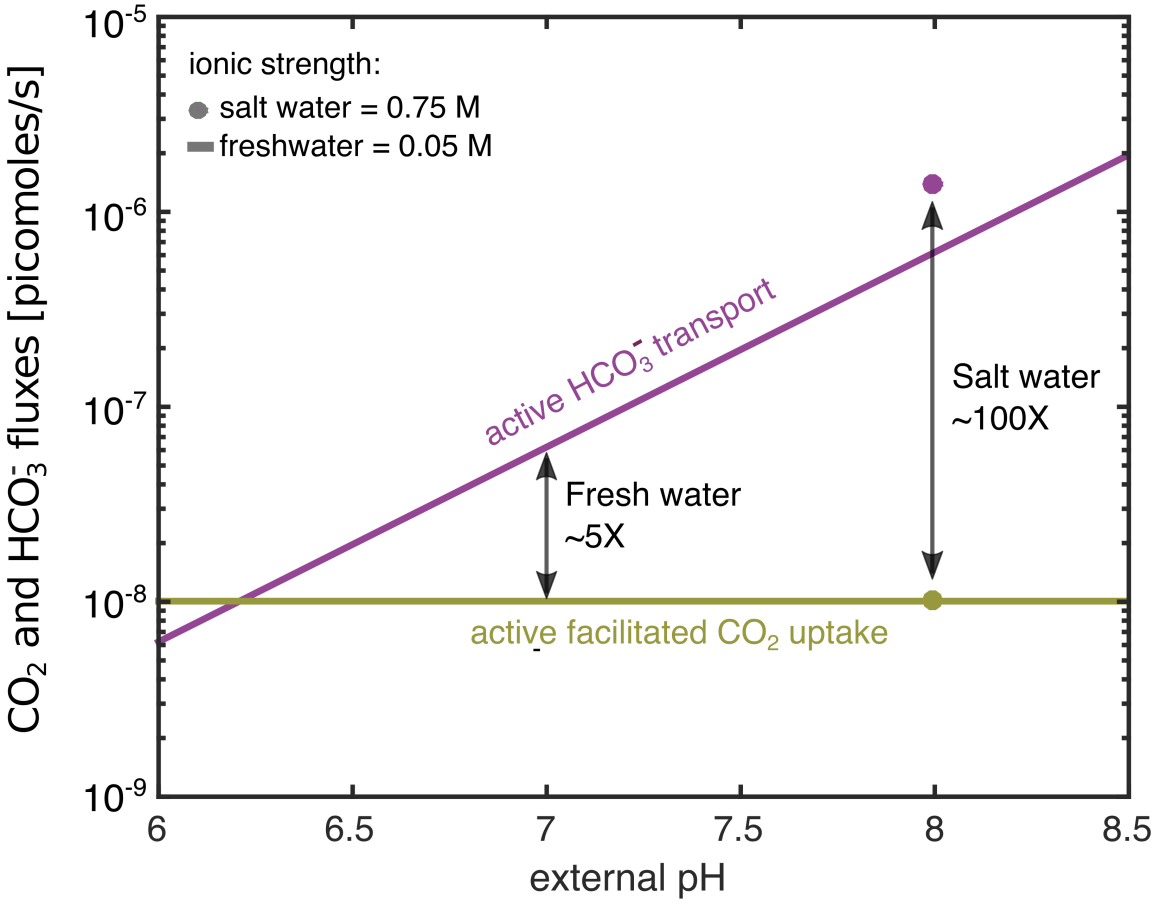


Figure B.9: Effectiveness of C_i uptake mechanisms depends on extracellular pH. Comparison of optimal C_i fluxes into the cell from the CO_2 facilitated uptake system (gold) and active HCO_3^- transport (purple) at external pH 6 to 8.5, for ionic strength 0.05 M (representative of freshwater). The CO_2 to HCO_3^- conversion velocity and HCO_3^- active transport velocity were both set to 2×10^{-5} cm/s, and the carboxysome permeability was set to 3×10^{-5}. At pH 7 (typical of freshwater) active HCO_3^- transport produces only fivefold more flux than active conversion of CO_2 to HCO_3^-. At pH 8 and ionic strength 0.75 M, representative of seawater, there is 100 times greater through active HCO_3^- uptake. Altogether these results suggest that CO_2 transport will be relatively more advantageous in acid pH, as shown for *H. neapolitanus* in Chapter 4.

equivalent to the cyanobacterial cytosol). In C_3 plants, Rubisco oxygenation proceeds at 15-30% of the carboxylation rate [33, 271] incurring an additional cost as high as $\approx 15\ H^+$ per CO_2 fixation due to photorespiration. Figure B.10 compares the energetic cost of the modeled CCM against the natural alternative of using a high-specificity Rubisco in the bacterial cytosol (i.e. employing the C_3 plant approach in a cyanobacterium). Here the cost of the CCM is plotted as a function of the external CO_2 concentration with the external H_{total} concentration calculated assuming the extracellular pH is 7.

The grey band in Figure B.10 denotes the plausible range of the total energetic cost of fixation through the CCM. This range was calculated by allowing the cytosolic pH and

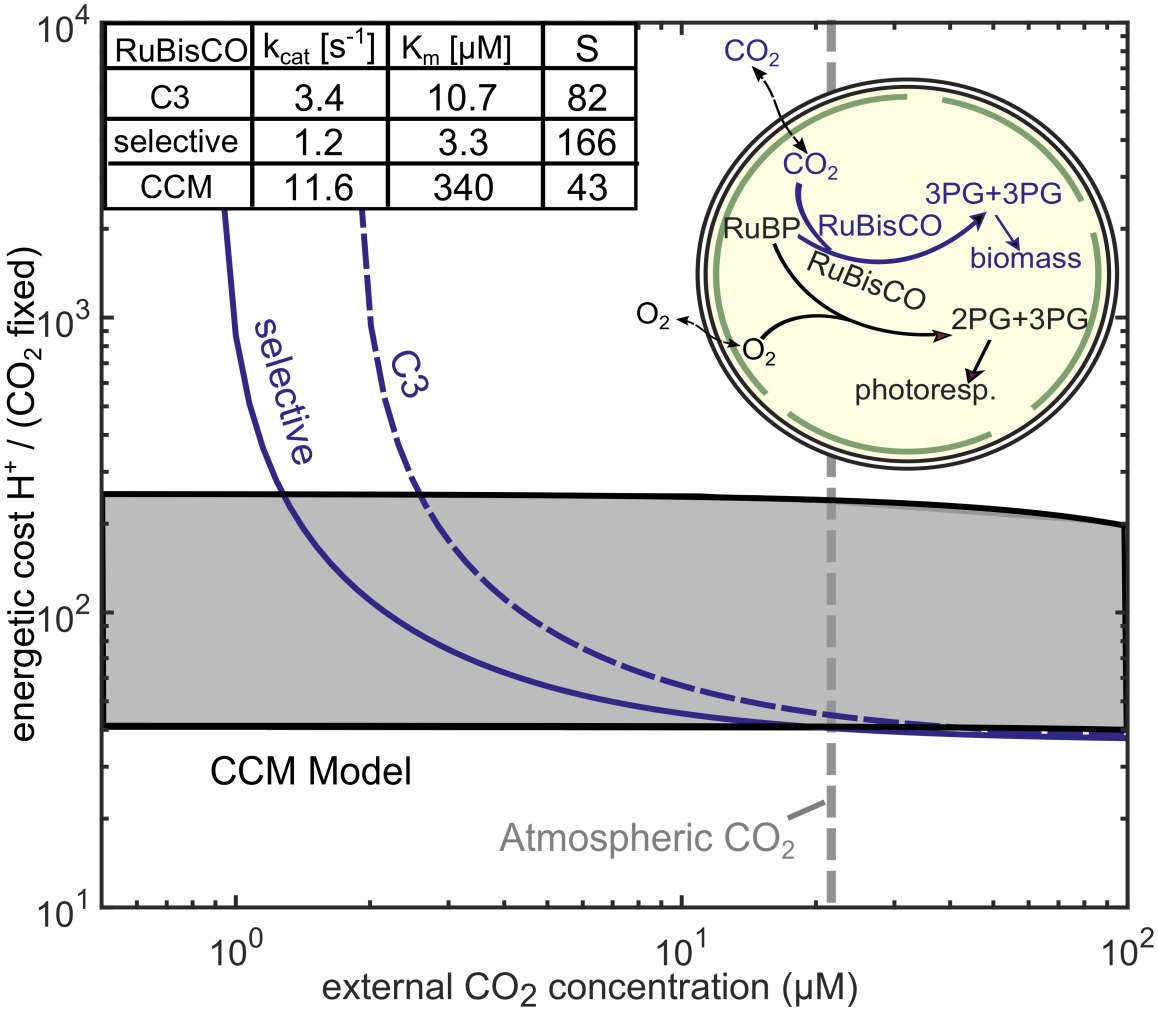


Figure B.10: Comparison of CCM energetics against a high-specificity cytosolic Rubisco. This figure plots the total cost of carbon fixation schemes as a function of the external CO_2 concentration. Here the cytosolic pH is set to 8 and the extracellular pH is set to 7, approximating freshwater. For simulations of the full CCM, the active HCO_3^- flux was set to achieve a cytosolic HCO_3^- concentration of 30 mM. The range of CCM costs was calculated by considering the optimal parameter regime and also regimes where the cytosolic pH and carboxysome permeability assume feasible but suboptimal values. For the simulations with Rubisco alone (inset), there is no carbonic anhydrase or active transport and two sets of Rubisco kinetic parameters were tested: those typical of a C_3 plant Rubisco (*N. tabacum*) and parameters measured for the highly-specific enzyme from *Galdiera sulfuraria*. These enzymes have twofold and fourfold higher CO_2 specificity ($S_{C/O}$) than the cyanobacterial Rubisco, respectively (Table B.2).

the carboxysome permeability to assume feasible but suboptimal values (see below). For comparison, we calculated the energetic cost of fixation using a cytosolic Rubisco in the absence of a CCM. The CCM and a high-specificity cytosolic Rubisco appear approximately equally energetically efficient near atmospheric CO_2 concentrations (Figure B.10). A highly selective Rubisco like that of the red alga *Goldiera sulfuraria* will be near saturation at atmospheric CO_2 concentrations ($\approx$ 10 μM) and will oxygenate roughly once for every

10 carboxylations, yielding a total cost energetic cost of $\approx 42\frac{H^+}{CO_2}$. In optimal conditions (cytosolic pH 8, carboxysome permeability $\approx 3 \times 10^{-5}$) and in atmospheric CO_2 the CCM model expends ≈ 4 H^+ per fixation on transport, yielding a total cost of $\approx 51\frac{H^+}{CO_2}$. If the cost of transport were lower (2 H^+ per fixation, see above) then the total cost of fixation with or without the CCM would be comparable at atmospheric CO_2 concentrations (41 and 42 $\frac{H^+}{CO_2}$ respectively).

In Figure B.10, the total energetic cost of the CCM is displayed as a shaded range. The cost range of the CCM is calculated by evaluating the CCM model under five parameter regimes and drawing the range between the minimum and maximum calculated cost at each extracellular CO_2 concentration. In the first regime, optimal CCM parameters were used with k_c = 3x10^{-5} cm/s and cytosolic pH 8.3. To examine the effect of suboptimal CCM characteristics, we tested the model in a low pH (cytosolic pH 7.3) and high carboxysome permeability ($k_c = 10^{-2}$ cm/s) regime separately. Both of these parameter regimes produce "efficient carboxylation" in the sense that Rubisco is saturated and carbonic anhydrase is unsaturated. Finally, cyanobacteria are known to induce alternative transporters when CO_2 is limiting [209] and may also incur additional costs in the maintenance of Na^+ gradients. To account for the likelihood of increased transport costs in limiting CO_2 conditions, we calculated also the total energetic cost in a model where transport costs 2 H^+ per fixation (1/2 our baseline value of 4 H^+) and 8 H+ per HCO_3^- transported (2x baseline). The range drawn in Figure B.10 is determined by the costs calculated for these four regimes: (i) optimal parameters, (ii) low cytosolic pH, (iii) high carboxysome permeability (iv) half-cost HCO_3^- transport and (iv) double-cost transport.

The above analysis suggests that the main energetic benefit of the CCM arises in low or limiting CO_2 concentrations. Near atmospheric CO_2 levels the cost of the specific and C_3 Rubisco approaches the minimum cost of the CCM (which, as discussed above, depends heavily on the cost ascribed to HCO_3^- transport). On a purely energetic basis, it may even be preferable to use a highly specific carboxylating enzyme in atmosphere (Figure B.10). It is worth noting that the specific Rubisco does have a lower carboxylation rate, however. So even though the cost per fixation is about equal, the absolute carboxylation rate will be slower for the selective Rubisco. Given the strong interest in engineering C_3 plants to use the cyanobacterial CCM [187, 232] this analysis suggests that a more detailed understanding of the CCM is required to predict how it might benefit a higher plant. For example, due to uncertainty about the exact cytosolic pH and carboxysome permeability, the total cost of fixing carbon through CCM is uncertain and may be 2-3 fold greater than the optimal cost calculated in Figure 3.3. As the CCM requires integral membrane proteins for transport, biogenesis of a 200+ MDa carboxysome organelle [56, 62] and unknown regulatory mechanisms controlling C_i fluxes, the energetic cost of CCM biogenesis and maintenance may not be negligible. Conversely, if the carboxysome shell is highly selectively permeable [65, 154, 237] then the CCM may be more energetically efficient than calculated here.

B.8 Analysis of Model Results and Assumptions

Parameter Sensitivity Analysis

We can completely analyze the behavior of the CCM system in terms of multiple parameters by considering the analytic solutions described above and detailed below. To give further intuition for which parameters are most important, we performed a sensitivity analysis of the cost of the CCM at cytosolic pH 7.3 (the pH measured in the light) and pH 8.3 (the value measured in the dark) in Figure B.11. For all parameters tested, the total cost of the CCM is always higher at pH 7.3 than pH 8.3. However, the sensitivity of the model to changes in certain parameters (e.g. the membrane permeability to H_2CO_3) depends on the pH.

Several parameters that might intuitively seem to strongly affect CCM performance – namely the cell membrane permeability to CO_2 (k_m^C) and selective carboxysome permeability to HCO_3^- ($k_c^H > k_c^C$) – do not affect CCM efficiency at all (Figure B.11). Indeed, order of magnitude changes in k_m^C changes the cost of fixation by less than 1%. Assuming 1000-fold selectivity for HCO_3^- over CO_2 at the carboxysome shell has a similarly small effect (as discussed above). In Chapter 3 we required 30 mM cytosolic HCO_3^-. Although Rubisco is not saturated when the cytosolic HCO_3^- pool is < 10 mM (as discussed above), the per-fixation cost of the CCM is not drastically affected by smaller cytosolic HCO_3^- pools. Indeed, lowering the cytosolic HCO_3^- concentration to 5, 15 and 25 mM pools affects the cost by $<$ 20% (in a pH-dependent manner, Figure B.11). As a result of the non-saturation of Rubisco, however, the total fixation rate will be very low when cytosolic $[HCO_3^-] < 10$ mM.

Changing the membrane permeability to H_2CO_3 (k_m^H), in contrast, has a dramatic effect on the cost of fixation (even if we re-optimize the carboxysome permeability to C_i at each tested k_m^H). Simultaneously increasing the permeability of the carboxysome to both CO_2 and HCO_3^- also has a very large effect. Finally we test the sensitivity of our analysis to a 2-fold decrease and increase in the cost of each HCO_3^-. As transport is a major factor determining the cost of fixation through the CCM, changing the HCO_3^- transport cost has a nearly proportional effect on the total cost of fixation.

The Effect of Selectivity at the Carboxysome Shell

For most of our modeling efforts here, we assumed that the carboxysome is equally permeable to CO_2 and HCO_3^-. Recent structures of the shell proteins of both α and β-carboxysome shell proteins provide a potential mechanistic basis for differential permeability of CO_2 and HCO_3^-: the pores of shell proteins typically carry positive charge, which might increase the rate of HCO_3^- transit relative to CO_2. Indeed, recent experimental evidence suggests that other protein compartments (e.g. the propanediol utilization compartment of *S. enterica*) are selectively permeable [65, 114]. It is impossible, however, to estimate the scale of such selectivity on the basis of these structures and experiments: it is unclear whether the carboxysome would prefer HCO_3^- over CO_2 by $\approx$10x, 100x, 1000x or not at all. For this reason, we initially treated equal permeability (i.e. no selectivity at the carboxysome shell) as a

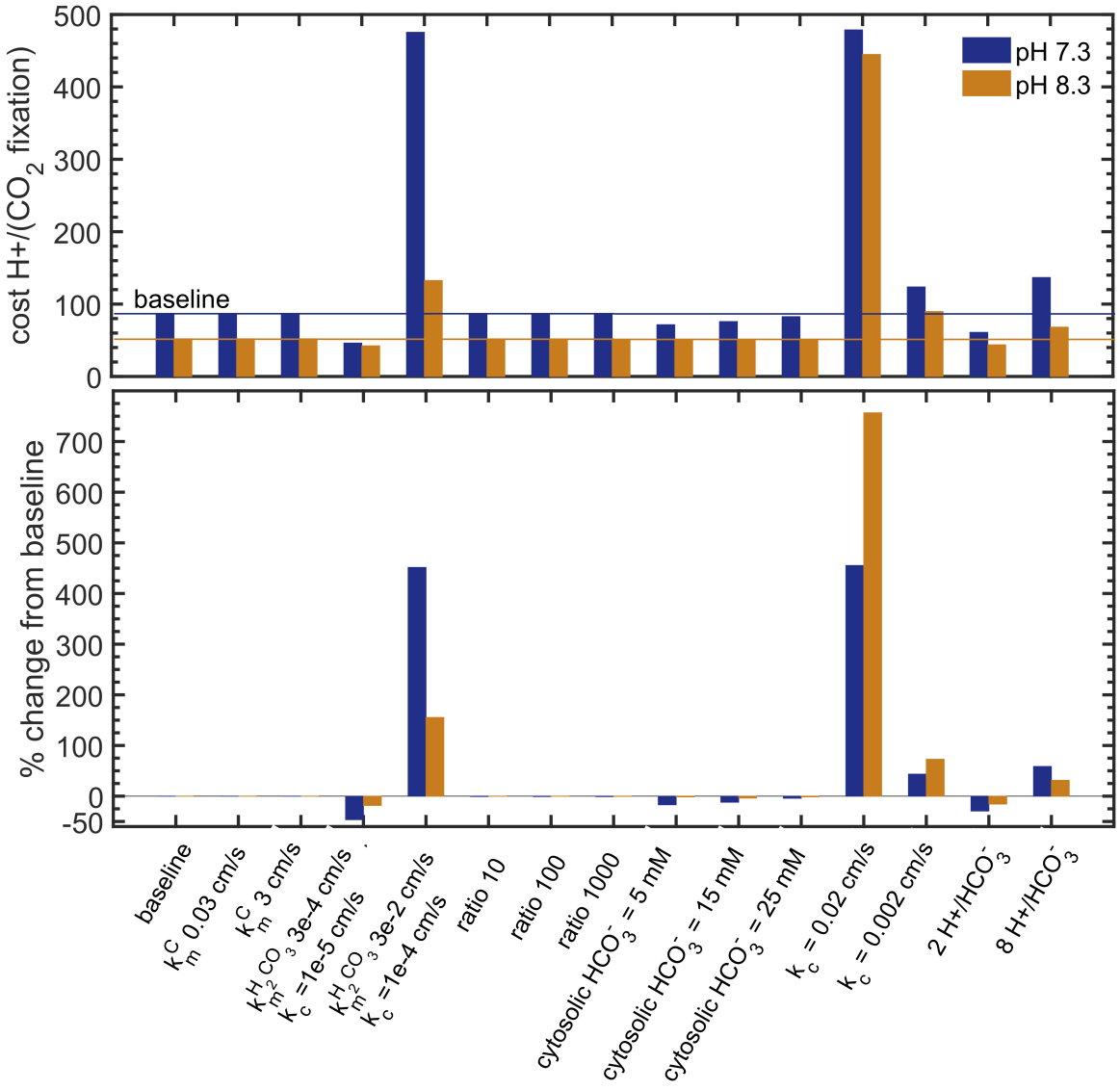


Figure B.11: Sensitivity analysis of the pH-aware CCM model. Cost and relative change from baseline for varying different parameters. For some parameters the carboxysome permeability is also changed to remain at optimal carboxysome permeability, as otherwise the main effect would be to shift away from optimum permeability.

parsimonious assumption to begin modeling from. Based on this assumption we derived an optimal carboxysome permeability coefficient of $\approx 3 \times 10^{-5} \frac{\text{cm}}{s}$ for a cytosolic pH $= 8$.

Intuitively, it seems that the "optimal" permeability of the carboxysome to HCO_3^- is ∞ cm/s while the "optimal" permeability to CO_2 is 0 cm/s. These permeabilities would maximize HCO_3^- uptake and minimize loss of CO_2, ensuring that every carbon entering the carboxysome is fixed. We note at the outset that this intuition is compelling but has

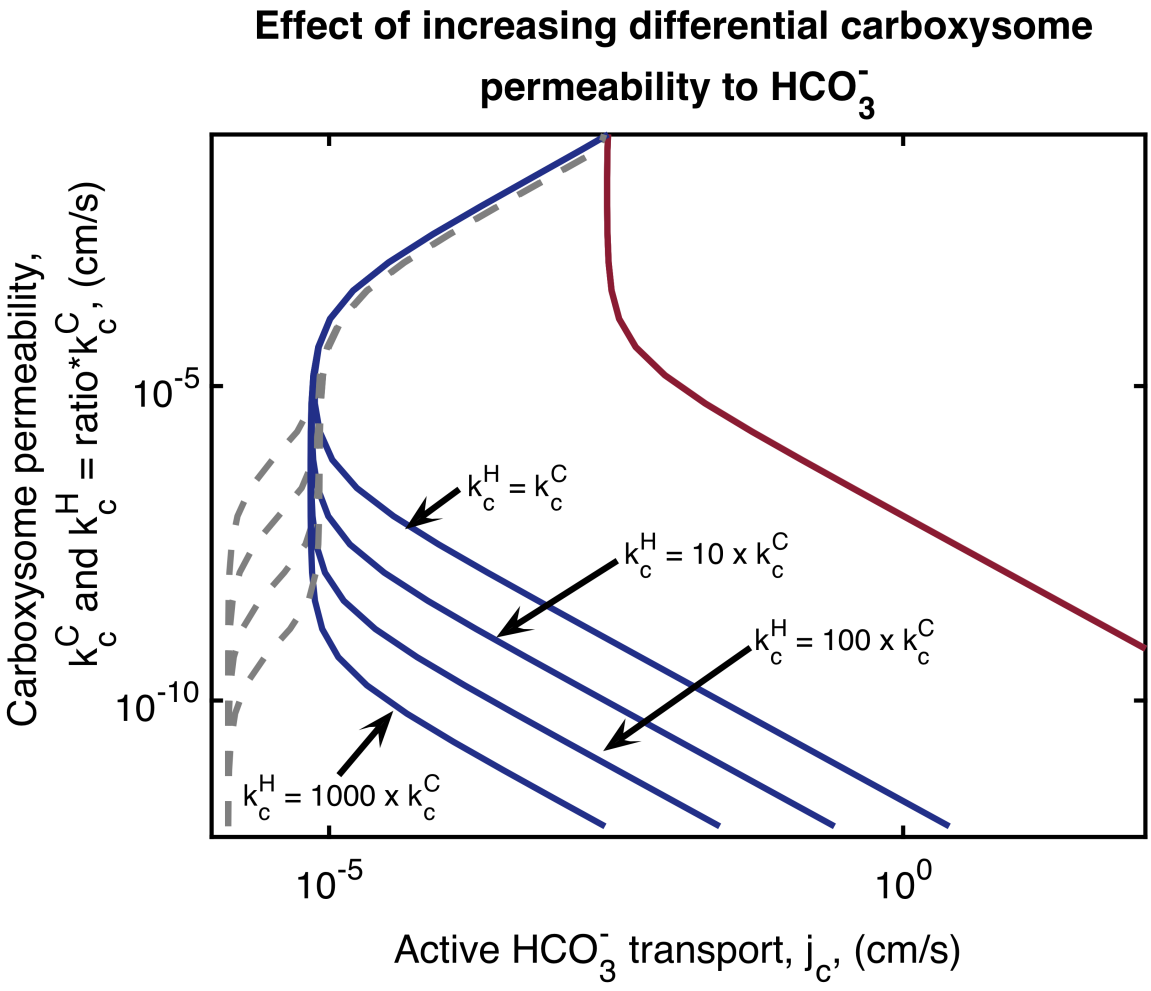


Figure B.12: Selective uptake of HCO_3^- into the carboxysome does not alter the optima in the CCM performance space. Increased carboxysome permeability to HCO_3^- relative to CO_2 mostly affects the transport required rates for carboxysome permeabilities to CO_2 that are beneath the optimum. The optimal single carboxysome permeability (i.e. the permeability that minimizes the amount of HCO_3^- uptake required to saturate Rubisco) changes by less than 5% by permiting selective HCO_3^- uptake.

a clear flaw: HCO_3^- and CO_2 are similar in size and both small enough to pass through the pores of the carboxyome shell. Moreover, CO_2 is the smaller molecule and so it is certain that both molecules have finite, non-zero permeabilities to the carboxysome shell. Nonetheless, the core intuition remains persuasive: very high HCO_3^- permeability and very low CO_2 permeability should maximize the efficiency of the CCM. We used our model to rigorously test this intuition and also to quantitatively evaluate our null assumption of equal permeability at the carboxysome shell.

As is clear from Figures 3.4 and B.12, selectivity at the carboxysome shell does not

substantially improve the performance of the CCM. To explain this non-intuitive result, we begin by noting that the best possible case for the CCM is that the carboxysomal carbonic anhydrase brings the cytosolic HCO_3^- pool into equilibrium inside the carboxysome. Since C_i is supplied to the carboxysome in the form of HCO_3^-, non-equilibrium implies excess bicarbonate and equilibrium yields the highest possible carboxysomal CO_2 concentration (in the absence of any energetic activation of the carboxysomal carbonic anhydrase, e.g. ATP coupling, of which there is no evidence). For this reason, several papers have postulated that the carboxysome might hold a more acidic pH than the cytosol - i.e. in order to increase the equilibrium CO_2 concentration (see below).

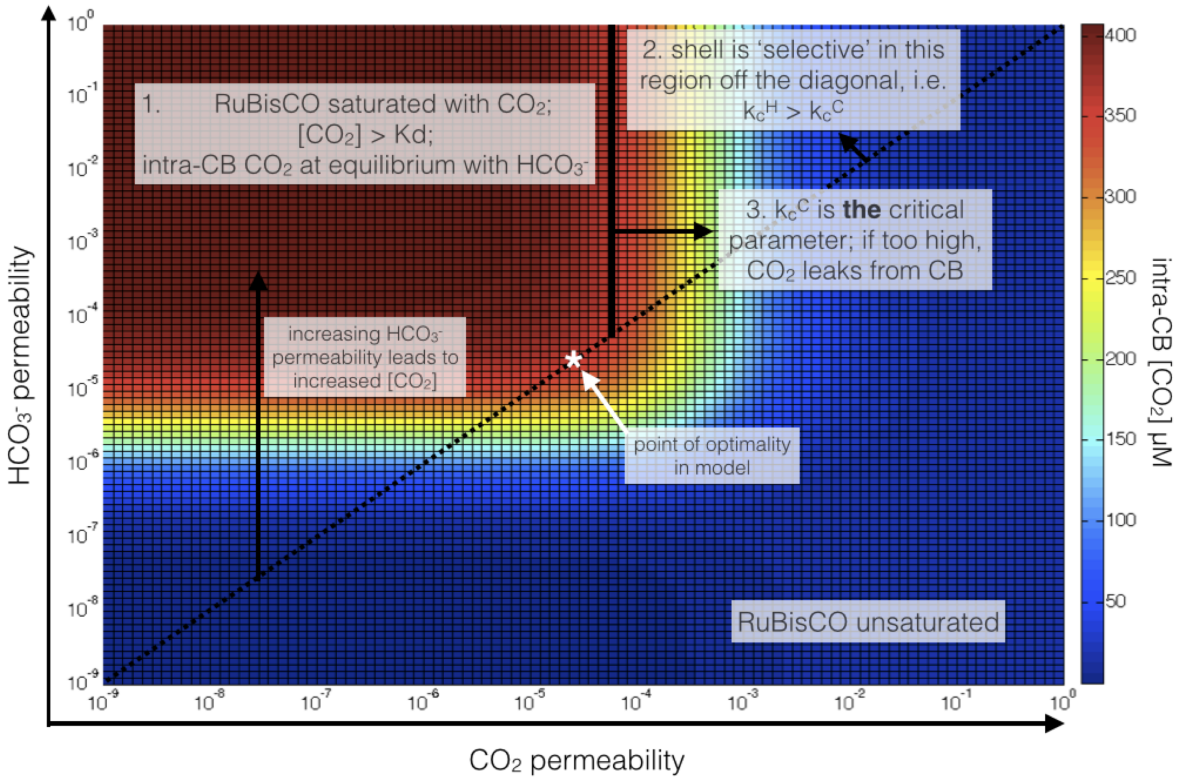


Figure B.13: Selective uptake of HCO_3^- and selective retention of CO_2 in the carboxysome does not increase the maximal carboxysomal CO_2 concentration. This plot shows the effect of varying both carboxysome permeability to CO_2 (X-axis) and carboxysome permeability to HCO_3^- (Y-axis). Color indicates the resulting carboxysomal CO_2 concentration, which plateaus in the upper-left region of the plot at $\approx$ 400 μM, the concentration resulting from equilibrating 30 mM HCO_3^- with CO_2 at pH 8. The x = y diagonal is the line of equal carboxysome permeability to CO_2 and HCO_3^-. This line passes through the red region of maximal carboxysomal CO_2 concentration, indicating that a single permeability is sufficient to achieve maximum Rubisco carboxylation. Note also that the carboxysome permeability to CO_2 (k_c^C) is the critical parameter in this plot: increasing k_c^C beyond $\approx 10^{-4}$ cm/s causes a precipitous drop in the carboxysomal CO_2 concentration that cannot be rescued by selective uptake of HCO_3^- into the carboxysome.

Our analysis shows that it is possible to achieve a CO_2 concentration very near this maximal value without any selectivity (Figures 3.4 & B.12). If we then imagine increasing k_c^H (the

carboxysome permeability to HCO_3^-) by tenfold while leaving k_c^C (the carboxysome permeability to CO_2) at the optimum single permeability value, it is clear that this cannot increase the carboxysomal CO_2 concentration very much – the optimal single permeability already achieved a carboxysomal CO_2 concentration very near the equilibrium value. Furthermore, the optimum CO_2 permeability (the k_c^C value allowing for the minimum C_i uptake flux) does not change appreciably by allowing 10, 100 or 1000-fold selectivity (Figure B.12). If k_c^C were to increase any further, it would allow CO_2 to leak from the carboxysome into the cytosol, where the CO_2 concentration is roughly 10 μM (Figure B.13). In summary, selectivity at the carboxysome shell has only a small effect on the carboxysomal CO_2 concentration. Rather, the carboxysome permeability to CO_2 is the critical parameter affecting the efficiency of the CCM, with $k_c^C > 10^{-4} \frac{\text{cm}}{s}$ greatly increasing the total cost of CO_2 fixation (Figures B.12 & B.13).

Effect of Carboxysome Geometry on Small Molecule Entry

As explained above, our model demands a carboxysome permeability to CO_2 on the order 10^{-5} cm/s in order for the CCM to function efficiently. In this section we ask whether this value is consistent with the observed geometry and structure of the carboxysome. The carboxysome shell is composed of hexameric and pentameric proteins, the vast majority of which contain a central pore approximately 0.4-0.5 nm in diameter [154]. Presumably, the shell blocks passage of small molecules into and out of the carboxysome lumen everywhere except at the pore. We approximate the scale of this effect as follows.

The surface area of the carboxysome shell can be calculated assuming a regular icosahedral geometry (i.e. 20 equilateral triangular sides). Cryo-electron tomography of the α-carboxysome has shown that the diameter of *H. neapolitanus* carboxysomes ranges from 88-108 nm and the diameter of *Synechococcus* WH8102 carboxysomes ranges from 114-137 nm [135, 263]. For this calculation, we will assume a carboxysome diameter of 100 nm for simplicity. The edge length of a regular icosahedron equals diameter/1.9, giving 53 nm for a 100 nm carboxysome diameter, yielding total surface area of $\frac{20\sqrt{3}}{4} \times (53nm)^2 = 2.4 \times 10^4$ nm^2. Notably, there is some variation in carboxysome diameter across species and some heterogeneity in diameter even among carboxysomes isolated from the same species. Electron microscopy studies suggest that the α-carboxysome diameter might be as small as 80 nm or as large as 120 nm [135, 263]. The β-carboxysome from *S. elongatus* 7942 is substantially larger, with a 175 nm diameter (with a wider size dispersions as well). Given these diameter estimates, the carboxysome surface area could range from 1.5×10^4 nm^2 on the low end to 7.3×10^4 nm^2. For simplicity, we will proceed under the assumption that the carboxysome diameter is 100 nm, reflecting measurements from α-carboxysomes.

Pentameric proteins are thought to cap the 12 vertices of the icosahedral carboyxsome while hexamers form the faces of the shell [330]. Based on the size of individual hexamers in crystal structures, it is estimated that there are roughly 40 hexamers per face [135, 330]. Therefore, there are $40 * 20 + 12 = 812$ pores on the shell. If each has a diameter of 0.4 nm then there are $812 \times \pi \times 0.2^2 = 102$ nm^2 of pore area of the shell, comprising about

0.4% of the carboxysome surface area. Notably, carboxysome shell proteins with larger pores have been found. For example, Csos1D has a pore ≈1 nm in diameter. If all the pores on the carboxysome had 1 nm diameter, upwards of 3% of surface area would be passable. However, Csos1D shows up very faintly on denaturing protein gels of *H. neapolitanus* and *Prochlorococcus* MED4 carboxysomes, suggesting that it is only a minor constituent of the shell [156, 250]. Therefore, we proceed assuming that about 0.5% of the shell surface is accessible for small molecule diffusion.

Ignoring any specific interaction between the carboxysome and the small molecules traversing its shell (e.g. RuBP, HCO_3-, CO_2, O_2), we would expect a 200-fold reduction in diffusional flux simply due to occlusion of surface area. This calculation enables an upper bound estimate of the permeability of the carboxysome shell as follows. First we calculate a velocity on the assumption that the shell constitutes no barrier at all across the length of the pore (≈2nm).

$$v = \frac{D}{l} = \frac{10^{-5}\frac{cm^2}{s}}{2 \times 10^{-7}\text{cm}} = 50\frac{\text{cm}}{s}$$

This velocity should be reduced according to the fraction of 2 nm slices through the carboxysome shell that are passable, which we calculated about as 1/200. This yields an estimated carboxysome permeability coefficient of 0.25 cm/s for a non-interacting molecule small enough to pass a pore of radius 0.4 nm.

We note that this number is roughly four orders of magnitude higher than the optimal single carboxysome permeability calculated through our model for a cytosolic pH of 8 ($3 \times 10^{-5}\frac{\text{cm}}{s}$) and about 20-fold higher than the maximum carboxysome permeability that allows for efficient carboxylation at pH 8 ($10^{-2}\frac{\text{cm}}{s}$). This 20-to-1000-fold difference between permeabilities calculated based on geometry and optimality principles presents us with a challenge: how can we reconcile these very different estimates?

Many molecular interactions in the bacterial cytosol could affect carboxysome permeability. For example, it is plausible that the negatively charged HCO_3^- interacts favorably with positive charge in and around the carboxysome pore and, as such, enters the carboxysome more quickly than CO_2. It is also possible that some unknown secondary molecule binds to carboxysome pores and blocks a fraction of the time, thereby decreasing the overall permeability of the shell and putting it into the regime we calculated from the model. As the permeability characteristics of the carboxysome shell appear to be crucial to the functioning of the CCM, this upper bound estimation of the carboxysome permeability highlights the need for a better physical and biochemical understanding of the factors that alter the permeability of protein shells. A more detailed discussion of how secondary molecules could affect carboxysome permeability is given in Appendix C.

Assumption of pH Equilibrium Across the Carboxysome Shell

Our pH-aware model requires pH values for all compartments in the model, namely the extracellular space, the cytosol and the carboxysome. In Chapter 3, we varied the intracel-

lular pH at fixed extracellular pH and above, we varied the extracellular pH while fixing the intracellular pH. In all cases, we assumed throughout that the carboxysomal pH was equal to the cytosolic pH. This assumption has some empirical support: using a pH-sensitive GFP fused to the carboxysomal Rubisco, it was shown that a pH gradient across the carboxysome shell collapses within 10 milliseconds or less [190]. In this section we evaluate whether a pH gradient across the carboxysome shell is plausible.

Several factors complicate the direct use of the above measurements to evaluate the possibility that the carboxysome shell might hold a pH gradient. First, the GFP chromophore responds to changes in pH on the millisecond timescale [322] and so it is likely that the above measurement represents an upper bound on the timescale of pH equilibration across the carboxysome shell. Moreover, cells harboring GFP-containing carboxysomes had a mild HCR (high-CO_2 requiring) phenotype [190], suggesting that these mutant carboxysomes may have different permeability characteristics than native carboxysomes.

The above *in vitro* measurement also does not account for the possibility that steady state fluxes producing and consuming H^+ in the carboxysome result in a pH across the shell. Let's proceed on the assumption that the carboxysome lumen is more acidic than the cytosol, as this would increase the CO_2 concentration and improve the Rubisco kinetics relative to pH 8 (Figure B.1 & B.4). In this case, the Rubisco reaction is the only flux producing H^+ in the carboxysome (producing two new carboxylic acids for each carboxylation), while the CA reaction and H^+ leakage across the shell both consume H^+.

Supposing that the CA and Rubisco fluxes are equal (i.e. every dehydrated HCO_3^- is subsequently carboxylated) then H^+ would be produced in the carboxysome at a rate equal to the CA flux or $\frac{1}{2}$ the Rubisco flux (CA dehydration consumes 1 H+ while Rubisco carboxylation produces 2 H^+). We note that this is the best case for the "relatively acidic carboxysome" hypothesis because Rubisco cannot carboxylate faster than CA produces CO_2. The maximum carboxylation rate of the carboxysomal Rubisco is roughly 10 s^{-1} per active site. As there are ≈2000 Rubisco active sites in the carboxysome, 2×10^4 H^+ are produced per second in the carboxysome at maximum and the timescale of maximal H^+ production is ≈50 s.

We can compare this to the timescale for H+ diffusion across the carboxysome as follows. The diffusion coefficient for protons in water is about threefold greater than water itself ($D_{H_2O} \approx 2300\mu m^2/s$, $D_{H^+} \approx 7000\mu m^2/s$). Protons diffuse more quickly than water because they can "hop" between adjacent water molecules [3]. Diffusion coefficients in this range yield a timescale of $\tau = \frac{x^2}{6\ D} = \frac{(0.1\ \mu m)^2}{6 \times 7000 \mu m^2/s} \approx 240\ ns$ to traverse the $0.1\mu m = 100nm$ diameter of the carboxysome. We note that this is a lower bound estimate of the timescale because the carboxysome interior is densely packed with protein (400-600 mg/ml or 2-3 times the cytosolic protein concentration in most organisms), which would slow the diffusion of both water and protons. Notably, there is precedence for protein pores – e.g. the aquaporins - that allow conduction of small substrates, including water, with highly reduced proton permeabilities [299].

The diffusional timescale calculated above is about 200-fold faster than the maximal rate

of H^+ production in the carboxysome. However, if the carboxysome shell is impassable to H^+ everywhere except the pores, the timescale of diffusional H^+ escape from the carboxysome would be reduced by about 200 fold (as calculated above), putting it in the same order-of-magnitude as the maximal rate of H^+ production. This calculation suggests that it is possible to maintain a steady state where the carboxysomal pH differs from the cytosol. Ideally, the carboxysomal pH would be more acidic than the cytosol so as to produce a greater carboxysomal CO_2 concentration (Figure B.1) and to operate near the Rubisco pH optimum (pH 7.8, Figure B.4).

The Effect of ΔpH Across the Carboxysome Shell

Proceeding on the assumption that it is possible for the carboxysome to maintain a different pH than the cytosol, we now use the model to evaluate whether a different carboxysomal pH would improve the performance of the CCM. Changing the pH in the carboxysome has two effects on the CCM: (i) it affects the equilibrium composition of C_i inside the carboxysome and (ii) the kinetics of the carboxysomal enzymes. A more acidic pH < 8 would increase the equilibrium CO_2 concentration (Figure B.1) and increase Rubisco's maximum carboxylation rate (Figure B.4).

In Figure B.14 we show that a relatively acidic carboxysome would result in a higher degree of Rubisco saturation at the same rate of cellular HCO_3^- uptake. A carboxysomal pH ≈ 7 appears to minimize the amount of cellular HCO_3^- uptake required to saturate Rubisco (Figure B.14). However, if the carboxysomal pH is too acidic it would deleteriously affect Rubisco kinetics, eliminating the otherwise beneficial effect of acidifying the carboxysome lumen. A relatively basic carboxysomal pH > 8.5 would deleteriously affect CCM performance on two fronts: the equilibrium CO_2 concentration is reduced and Rubisco's maximum carboxylation rate vanishes with increasing pH [18].

When Can CO_2 Scavenging Effect CCM Efficiency?

It is often posited that the facilitated CO_2 uptake systems diagrammed in Figure 3.1 act to scavenge CO_2 leaking from the carboxysome [107, 231]. In this section, we examine the parameter regimes in which active, vectorial $CO_2 \rightarrow HCO_3^-$ conversion activity acts as facilitated uptake (resulting in a flux of CO_2 into the cell) or scavenging (reducing the leakage of CO_2 out of the cell). The conversion activity as modeled does not (and cannot) discriminate between CO_2 that has recently diffused into the cell and CO_2 that was produced in the carboxysome and leaked out. In other words, we do not "trace" the history of each CO_2 molecule, or even model each CO_2 molecule explicitly (the model operates on concentrations). Therefore, we must consider the net CO_2 flux at the cell membrane to evaluate the role of scavenging.

When there is a net flux of CO_2 out of the cell there is no net facilitated uptake (by definition). If there is zero net flux of CO_2 at the cell membrane, then the cell is scavenging 100% of the CO_2 leaking out of the carboxysome. If there is net flux of CO_2 into the cell,

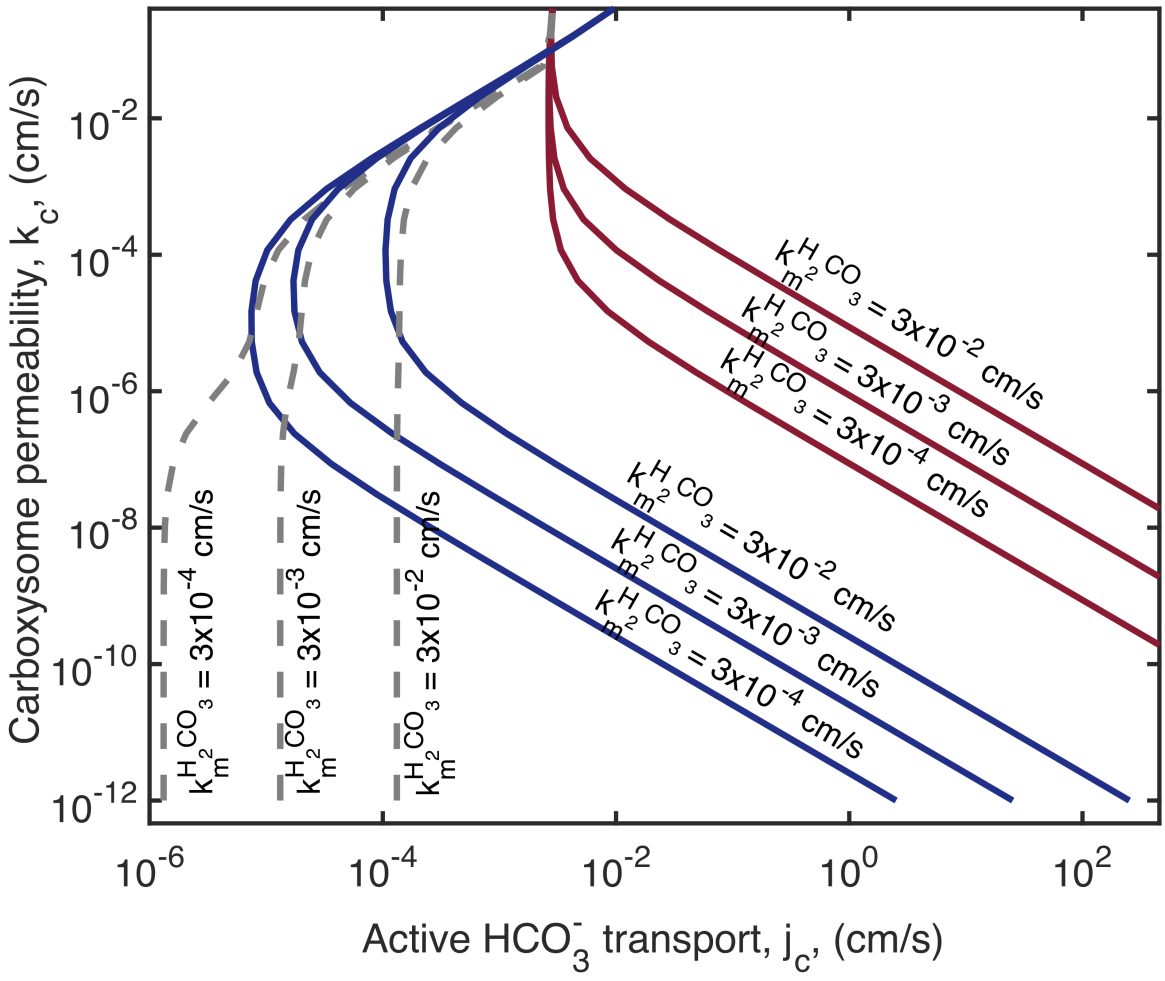


Figure B.14: A relatively acidic carboxysome would improve CCM performance. Panel (A) shows the effect of varying the carboxysomal pH and HCO_3^- transport on the degree of Rubisco saturation. The cytosolic pH was taken to be 8. When the carboxysomal pH is relatively acidic compared to cytosol (pH $\approx$ 7) less HCO_3^- transport is required to saturate Rubisco because (i) Rubisco has an increased carboxylation k_{cat} at this pH and (ii) the equilibrium between CO_2 and HCO_3^- favors CO_2 more than at pH 8. Even allowing for a pH gradient across the carboxysome, the minimum cytosolic HCO_3^- concentration (black lines) that allows for saturation of the carboxysomal Rubisco is $\approx$ 10 mM. Panel (B) shows that a CCM employing a more CO_2-specific Rubisco (carboxylation $K_{\text{M}} = 169$ μM) would still require $\approx$ 10 mM cytosolic HCO_3^- to saturate the carboxysomal enzyme. In both panels the teal region at the bottom denotes the portion of the phase space where the total CO_2 fixation flux is at most 10% of the carboxylation flux in reference conditions (pH 8 in the cytosol and carboxysome). This region is labeled to emphasize that the total CO_2 fixation flux vanishes at basic pH because the carboxylation k_{cat} is projected to vanish in that regime (see Figure B.4A).

then the conversion mechanism scavenging all CO_2 that would otherwise leak out of the cell and also reducing the cytosolic CO_2 concentration beneath the external concentration and facilitating CO_2 uptake.

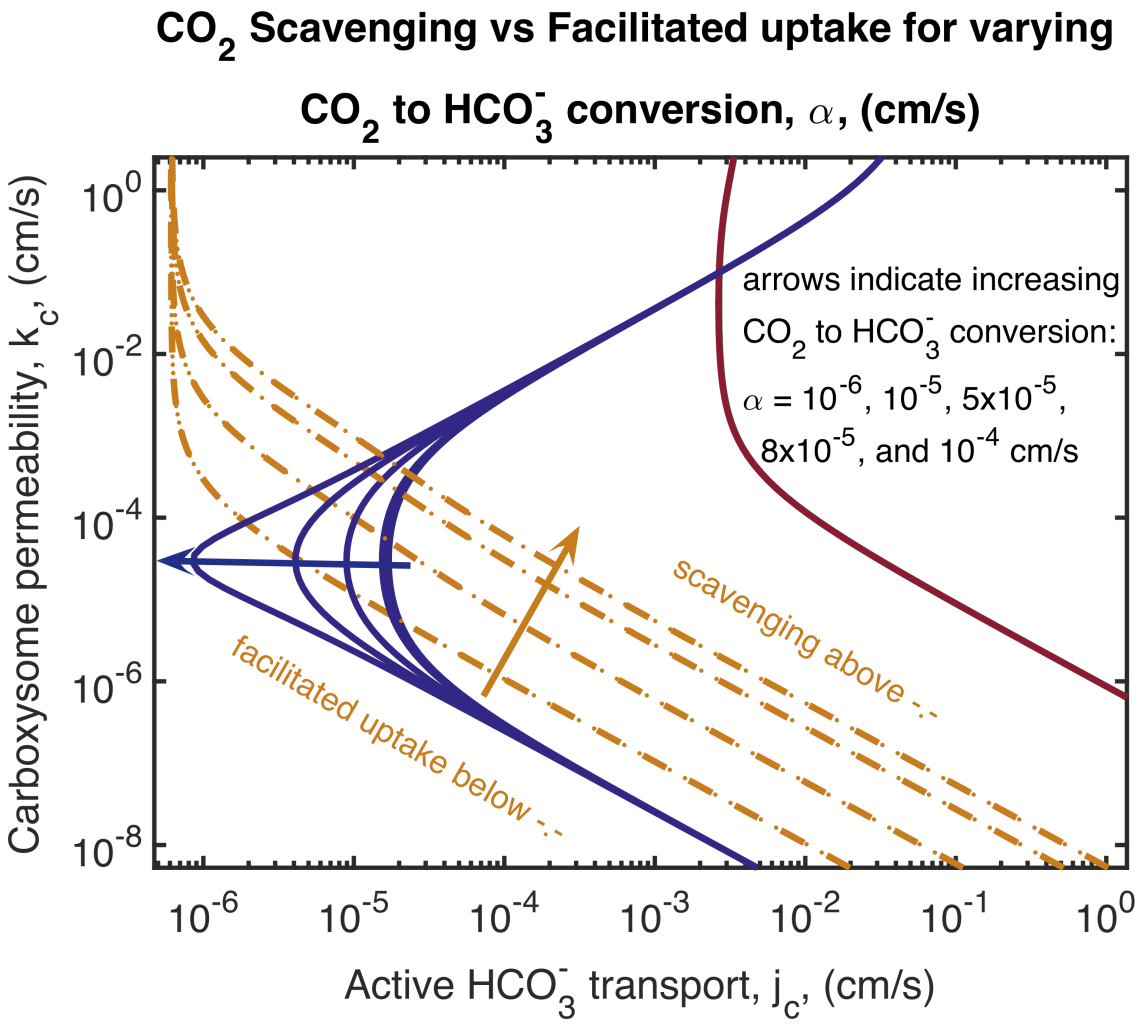


Figure B.15: Energetically activated CO_2 hydration could improve CCM efficiency by substituting for HCO_3^- transport. It is sometimes argued that the facilitated uptake of CO_2 (as diagrammed in Figure 3.1) could serve to scavenge CO_2 leaking from the carboxysome, thus improving the performance of a leaky carboxysome. The figure shows the effect of energetically activated CO_2 hydration on the CCM functionality space in Figure B.6. Rubisco is saturated to the right of the blue lines and carbonic anhydrase is saturated to the right of the red line. Orange dashed lines demarcate where the CO_2 flux across the membrane equals zero. Each line is calculated for a different active hydration velocity α, with α increasing in the direction of the orange and blue arrows. Above (to the right) of an orange dashed line there is net CO_2 leaking from cell, and the main function of the facilitated uptake system is scavenging. Below (to the left) of an orange line the net CO_2 uptake, and conversion is both scavenging and facilitating CO_2 uptake. Notice that scavenging only serves to reduce the required HCO_3^- transport velocity beneath the orange line, consistent with the demonstration in Figure B.11 that the energetic cost of fixing carbon through the CCM is not sensitive to changes in the CO_2 permeability to the cell membrane.

From Figure B.15 it is evident that increasing vectorial $CO_2 \rightarrow HCO_3^-$ conversion decreases the need for active HCO_3^- uptake at optimal carboxysome permeability. We can also determine when the conversion mechanism switches primarily scavenging to also acting as facilitated uptake. The orange dashed lines show the active HCO_3^- transport rate for each

carboxysome permeability k_c value where the net CO_2 flux at the cell membrane is zero. Above this line the system is scavenging and below it the system is scavenging and facilitating uptake. We have plotted these curves for increasing $CO_2 \rightarrow HCO_3^-$ conversion strengths α.

For $\alpha = 10^{-6}\frac{cm}{s}$, the system does not facilitate uptake of CO_2 at the optimal carboxysome permeability. At this point, the minimum active HCO_3^- transport rate required to saturate RuBiSCO is approximately the same as without the CO_2 conversion mechanism. For higher α the conversion system is scavenging at the optimal carboxysome permeability. In this case the system requires less HCO_3^- transport because CO_2 conversion is acting as facilitated uptake, and contributing to the internal inorganic carbon pool. Note that the absolute flux from scavenging and facilitated uptake will vary drastically over this space.

In Chapter 3 we showed that the loss of CO_2 from leakage out of the cell is about an order of magnitude smaller than HCO_3^- leakage (without considering vectorial CO_2 conversion). At optimal carboxysome permeability, scavenging will have little effect because there is just not very much CO_2 in the cytosol to scavenge. At higher carboxysome permeabilities, though, the scavenging mechanism could compensate and matter much more. Thus the relative importance of the CO_2 conversion mechanism for scavenging or facilitated uptake depends highly on the as yet experimentally undetermined carboxysome permeability.

Cytosolic C_i Pool Required for Efficient CO_2 Fixation

Recent experiments by Whitehead et al. used a membrane inlet mass-spectrometry approach (MIMS) and measured $\approx$ 5 mM C_i pools inside both α- and β-cyanobacteria [320]. 5 mM C_i concentrations (measured by MIMS) were found to be sufficient to saturate photosynthetic O_2 production of both organisms. These MIMS measurements are inconsistent with our assumption of 30 mM cytosolic HCO_3^-. We therefore investigated whether our model allows for Rubisco saturation at lower cytosolic HCO_3^- concentrations. Figure B.16 shows the iso-lines of cytosolic HCO_3^- and demonstrates that that these measurements are also incompatible with our model: the minimum cytosolic HCO_3^- concentration that allows for Rubisco saturation is $\approx$ 10 mM. We note that this effect is not obvious in the sensitivity analysis in Figure B.11 because that figure considers the energetic cost per-carboxylation but ignores the total carboxylation flux per cell, which is much reduced when Rubisco is not saturated.

Whitehead et al. suggest two mechanisms that might allow for Rubisco saturation at such C_i levels: employing a Rubisco with higher affinity for CO_2 ($K_M \approx 170\mu M$) and maintaining a relatively acidic carboxysomal pH to increase the CO_2 concentration [320]. Figure B.14B demonstrates that even the combination of these two mechanisms – a lower K_M and a more acidic carboxysomal pH – is not sufficient to saturate Rubisco when the cytosolic HCO_3^- concentration is only 5 mM. We conclude by noting that Whitehead et al. also measured the cytosolic C_i using a silicon-oil centrifugation technique and found much higher saturating cytosolic C_i concentrations of 15-20 mM using this technique. The authors themselves note that the MIMS technique allows the cells to rest in the dark, which might lead to systematic

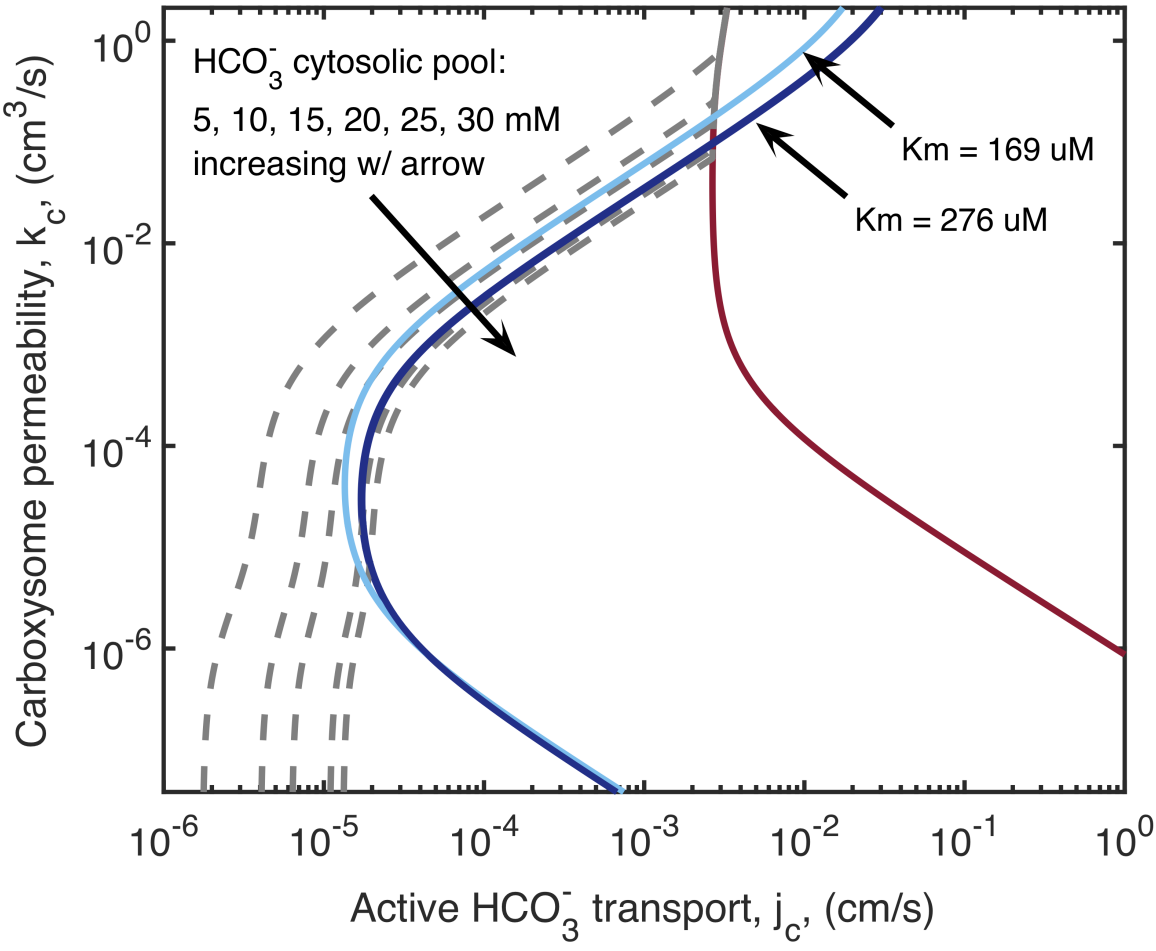


Figure B.16: The CCM model requires at 10-15 mM cytosolic HCO_3^- to saturate Rubisco. The two-dimensional "CCM functionality space" shows carboxysome permeability and HCO_3^- transport rates resulting in different cytosolic HCO_3^- concentrations. Grey dashed lines indicate where the HCO_3^- pool in the cytosol is 5, 10, 15, 20, 25, and 30 mM respectively. The blue lines indicate where Rubisco is saturated and the red line indicates where carbonic anhydrase is saturated. It is clear that $\geq$ 10 mM cytosolic HCO_3^- is to saturate Rubisco, given our threshold for Rubisco saturation.

underestimation of the C_i pool due to continued fixation in the dark [320]. Our modeling results are in agreement with this assessment – 5 mM cytosolic C_i is likely insufficient to saturate the carboxysomal Rubisco while a concentration near 15 mM is sufficient and so more parsimonious.

Cell Surface Area Required for C_i Transport

Our model cyanobacterium is spherical with a radius of 0.5 microns and a surface area of 3×10^{-8} cm^2 $= 3 \times 10^6$ nm^2. The primary HCO_3^- transporter in *S. elongatus* PCC 7942,

SbtA, is a single integral membrane protein with 10 transmembrane helices [231]. LacY has 12 transmembrane helices and an oval shape with a single-side surface area of approximately of 14 nm^2 (BioNumbers BNID 102929). If we assume a maximum rate of $1000 \frac{\text{molecules}}{transporter \times s}$ and a per-transporter surface area of 10 nm^2, similar to LacY, then the total surface area required for transport equals

$$V_{\text{import}} \frac{\text{molecules}}{\text{cell} \times \text{sec}} \times 10 \frac{\text{nm}^2}{\text{transporter}} \times 10^{-3} \frac{\text{transporters} \times \text{sec}}{\text{molecule}}$$

Where V_{import} is the rate of HCO_3^- transport into the cell. According to our original model, the lowest V_{import} allowing for efficient carboxylation was $3.3 \times 10^{-4} \frac{\text{picomoles}}{cell \times sec} = 2 \times 10^8 \frac{\text{molecules}}{cell \times sec}$, which yields a surface area requirement of 2x10^6 nm^2, representing an implausible 60% of total cell surface area.

According to the pH-aware model allows a V_{import} that is roughly 1000-fold lower at pH 8. The lowest V_{import} producing efficient carboxylation at pH 8 is on the order of $10^{-7} \frac{\text{picomoles}}{cell \times sec} = 6 \times 10^5 \frac{\text{molecules}}{cell \times sec}$ (Figure 3.2 and Figure B.5). This updated estimate implies a surface area requirement of 6×10^3 nm^2 or 0.2% of cell surface area. Not only is this a plausible fraction of surface area to dedicate to transport, but there is also substantial "headroom:" the fraction could be five or tenfold greater without forcing us to question the validity of the model.

B.9 Net Primary Photosynthetic Productivity Due to Cyanobacteria

About 45% of the global net primary photosynthetic productivity (NPP) is due to photosynthesis in the oceans, where cyanobacteria are major contributors to NPP. By contrast, nearly all of the NPP on land is due to macroscopic plants [217, 99]. Some previous reviews and papers attribute up to 25% of global NPP to the cyanobacteria [217, 238] based on the assessment that 50% of oceanic carbon fixation is due to cyanobacteria. Tracing these references it seems that the 50% number is drawn from two studies of Prochlorococcus in the equatorial Pacific [168, 169]. These values are likely artificially high because they focus on one of the regions of the world's oceans where Prochlorococcus is most abundant (as shown in Figure 3 of [104]).

Based on a review of the recent literature, we conclude that a number closer to 10% is more accurate. Some have estimated the cyanobacterial contribution as 10% of oceanic NPP [253] or 25% [104]. As carbon fixation is estimated to be ≈45% of total NPP, cyanobacteria contribute 5-12% of global NPP.

B.10 Detailed Mathematical Derivations

Equations when Rubisco is saturated

The analytic solution for the CO_2 and HCO_3^- concentration in the carboxysome when Rubisco is saturated is:

$$C_{carboxysome} = \frac{N}{M} - \frac{R_c^3 V_{max} P}{3MD} \tag{B.1}$$

$$H_{carboxysome} = K_{eq}(pH) C_{carboxysome} \tag{B.2}$$

where,

$$N = (j_c + k_m^{eff}(pH_{out}))H_{out}((k_m^C + \alpha)G^C + \frac{D}{R_b^2}) + k_m^C C_{out}(k_m^{eff} G^H + \alpha G^C + \frac{D}{R_b^2}) \tag{B.3}$$

$$M = K_{eq} \times k_m^{eff} \left((\alpha + k_m^C) G^C + \frac{D}{R_b^2} \right) + k_m^C \left(k_m^{eff} G^H + \frac{D}{R_b^2} \right) + \alpha k_m^{eff} G^H \tag{B.4}$$

$$P = ((\alpha + k_m^C) G^C + \frac{D}{R_b^2})(k_m^{eff} G^H + \frac{D}{R_b^2}) \tag{B.5}$$

$$G^C = \frac{D}{R_c^2 k_c^C} + \frac{1}{R_c} - \frac{1}{R_b} \tag{B.6}$$

$$G^H = \frac{D}{R_c^2 k_c^H} + \frac{1}{R_c} - \frac{1}{R_b} \tag{B.7}$$

The derivation of this equation can be found in the supplementary material of [182]. Here we have made a few modifications: (1) kept track of the carboxysome permeability to CO_2, k_c^C, and HCO_3^-, k_c^H, independently, (2) substituted the pH dependent equilibrium constant for the carbonic anhydrase reaction, $K_{eq}(ph) = \frac{V_{ca} K_{ba}}{V_{ba} K_{ca}}$, (3) written the $CO_2 \rightarrow HCO_3^-$ reaction with α as the linear reaction rate (in [182] the linear rate was α/K_α), (4) we have replaced the membrane permeability to HCO_3^- with the effective membrane permeability to the bicarbonate pool, and designated when is dependent on the external pH, $k_m^{eff}(pH_{out})$. This term only appears once in equation 3 for N. For all other $k_m^{eff} = k_m^{eff}(pH_{in})$ values it is dependent on the pH inside the cell, so we have dropped indicating the pH dependence to simplify the formulas.

Detailed Analysis of Cell Membrane Permeability

Cell Membrane Permeabilty Compared to Diffusive Velocities

Examining equations (6-7) above, we note that for large carboxysome permeability $1/R_c$ will be the dominant term, and for smaller carboxysome permeability values the first term will

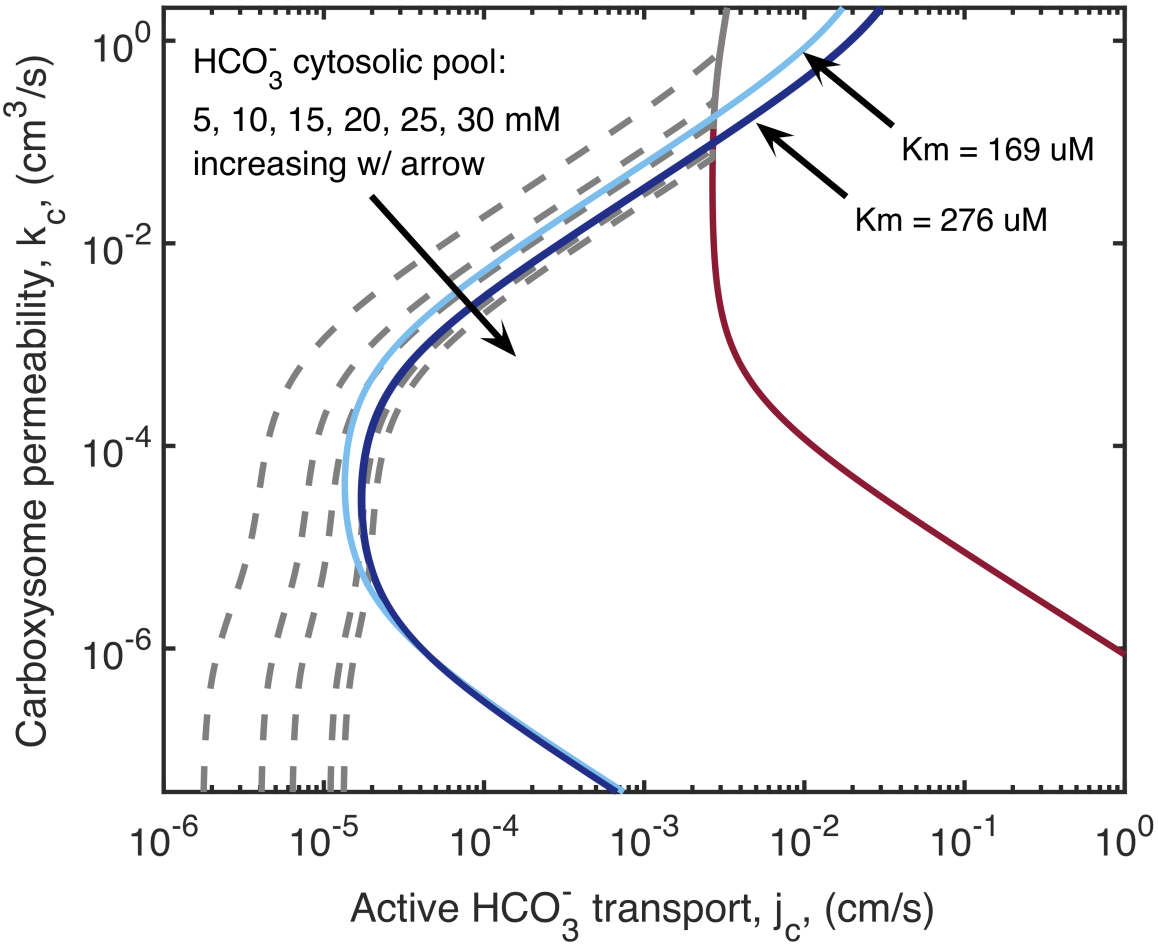


Figure B.17: Absolute membrane permeability to H_2CO_3 strongly affects CCM efficiency. This figure shows CCM functionality phase space for varying membrane permeability H_2CO_3, $k_m^{H_2CO_3}$. Blue lines indicate where Rubisco becomes saturated. Red lines indicate where carbonic anhydrase becomes saturated. Grey dashed lines indicate where the HCO_3^- pool in the cytosol is 30 mM in each case. Lines are labeled with the corresponding value of $k_m^{H_2CO_3} = 3 \times 10^{-4}$, 3×10^{-3}, 3×10^{-2} cm/s from left to right. The value of 3×10^{-3} cm/s was used throughout the rest of this work (Table B.1).

be larger and dominate. Therefore $G^C \geq 1/R_c$. Studying the equations we note that the terms $((\alpha + k_m^C)G^C + \frac{D}{R_b^2})$ appears repeatedly. We use the following argument:

$$(\alpha + k_m^C)G^C \geq (\alpha + k_m^C)/R_c >> D/R_b^2, \tag{B.8}$$

$$\text{if } (\alpha + k_m^C) >> DR_c/R_b^2 \tag{B.9}$$

For even a small 20 nm diameter ($R_c = 10^{-6}$ cm) carboxysome this will hold as $k_m^C \approx 0.3$ cm/s and $DR_c/R_b^2 = 4 \times 10^{-3}$ cm/s from the values in Table S1. So the membrane permeability to CO_2 could be an order of magnitude too high in our model and this would

still be a reasonable assumption. Therefore we will substitute

$$(\alpha + k_m^C)G^C + D/R_b^2 \approx (\alpha + k_m^C)G^C. \tag{B.10}$$

Inserting this into equations (1-5) we get

$$C_{carboxysome} = \frac{(j_c + k_m^{eff}(pH_{out}))H_{out}(k_m^C + \alpha)G^C + k_m^C C_{out}(k_m^{eff}G^H + \alpha G^C + \frac{D}{R_b^2})}{K_{eq}k_m^{eff}(\alpha + k_m^C)G^C + k_m^C\left(k_m^{eff}G^H + \frac{D}{R_b^2}\right) + \alpha k_m^{eff}G^H} - \frac{R_c^3 V_{max}(\alpha + k_m^C)G^C(k_m^{eff}G^H + \frac{D}{R_b^2})/(3D)}{K_{eq}k_m^{eff}\left(\alpha + k_m^C\right)G^C + k_m^C\left(k_m^{eff}G^H + \frac{D}{R_b^2}\right) + \alpha k_m^{eff}G^H}. \tag{B.11}$$

We can divide through by $(k_m^C + \alpha)$ to obtain:

$$C_{carboxysome} = \frac{(j_c + k_m^{eff}(pH_{out}))H_{out}G^C + \frac{k_m^C}{(k_m^C+\alpha)}C_{out}(k_m^{eff}G^H + \alpha G^C + \frac{D}{R_b^2})}{K_{eq}k_m^{eff}G^C + \frac{k_m^C}{(k_m^C+\alpha)}\left(k_m^{eff}G^H + \frac{D}{R_b^2}\right) + \frac{\alpha}{(k_m^C+\alpha)}k_m^{eff}G^H} - \frac{R_c^3 V_{max}G^C(k_m^{eff}G^H + \frac{D}{R_b^2})/(3D)}{K_{eq}k_m^{eff}G^C + \frac{k_m^C}{(k_m^C+\alpha)}\left(k_m^{eff}G^H + \frac{D}{R_b^2}\right) + \frac{\alpha}{(k_m^C+\alpha)}k_m^{eff}G^H}. \tag{B.12}$$

We now want to examine the remaining terms in the membrane permeability to CO_2, k_m^C.

Membrane Permeability to CO_2 has Little Effect.

There are two parameter groupings in equation (12) containing k_m^C:

$$\frac{k_m^C}{k_m^C + \alpha} \tag{B.13}$$

$$\frac{\alpha}{k_m^C + \alpha} \tag{B.14}$$

Therefore if $k_m^C > \alpha$ or $CO_2 \rightarrow HCO_3^-$ conversion is negligible the first term (13) reduces to 1, and the second reduces to $1/k_m^C$. We will return to the case where this conversion is not negligible later.

With these two simplifications we obtain:

$$C_{carboxysome} = \frac{(j_c + k_m^{eff}(pH_{out}))H_{out}G^C + C_{out}(k_m^{eff}G^H + \alpha G^C + \frac{D}{R_b^2})}{K_{eq}k_m^{eff}G^C + \left(k_m^{eff}G^H + \frac{D}{R_b^2}\right) + \frac{1}{k_m^C}k_m^{eff}G^H} - \frac{R_c^3 V_{max}G^C(k_m^{eff}G^H + \frac{D}{R_b^2})/(3D)}{K_{eq}k_m^{eff}G^C + \left(k_m^{eff}G^H + \frac{D}{R_b^2}\right) + \frac{1}{k_m^C}k_m^{eff}G^H}. \tag{B.15}$$

Examining equation (15), note that the only appearance of the membrane permeability to CO_2 is now in the denominator which we can rewrite as $k_m^{eff}(G^C K_{eq} + \frac{G^H}{k_m^C}) + \left(k_m^{eff} G^H + \frac{D}{R_b^2}\right)$. Using this equation, we can write a strong bound on when the membrane permeability will effect the function of the CCM.

We find k_m^C has no significant effect when $K_{eq} G^C > \frac{G^H}{k_m^C}$ or $k_m^C > \frac{G^H}{G^C K_{eq}}$. If we assume that the carboxysome permeability to CO_2 will always be smaller than or equal to the permeability to HCO_3^- ($k_c^C \geq k_c^H$) then $G^H \geq G^C$ and $\frac{G^H}{G^C} \leq 1$, so k_m^C will be negligible as long as $k_m^C > 1/K_{eq}$. For pH > 6.6, $1/K_{eq} > 0.3$ and therefore the assumed value of $k_m^C = 0.3$ will be negligible. However, if the cell operated in a lower pH regime and the membrane permeability was substantially lower to CO_2 it would begin to effect the CO_2 concentration.

Thus far we have made a series of observations about the size of terms compared to the membrane permeability to CO_2 and found that when $(\alpha + k_m^C) >> DR_c/R_b^2$, $k_m^C > \alpha$ and $k_m^C > \frac{G^H}{G^C K_{eq}} \approx 1/K_{eq}$ the CO_2 concentration in the carboxysome reduces to

$$C_{carboxysome} = \frac{(j_c + k_m^{eff}(pH_{out}))H_{out} G^C + C_{out}(k_m^{eff} G^H + \alpha G^C + \frac{D}{R_b^2})}{k_m^{eff}(G^C K_{eq} + G^H) + \frac{D}{R_b^2}} - \frac{R_c^3 V_{max} G^C (k_m^{eff} G^H + \frac{D}{R_b^2})/(3D)}{k_m^{eff}(G^C K_{eq} + G^H) + \frac{D}{R_b^2}}. \quad \text{(B.16)}$$

We can make a similar argument taking the equation for the CO_2 concentration at the cell membrane:

$$C_{cytosol}(r = R_b) = \frac{k_m^C C_{out} - (\alpha + k_m^C) C_{carboxysome}}{(\alpha + k_m^C) G^C + D/R_b^2} G^C + C_{carboxysome} \approx C_{out} \quad \text{(B.17)}$$

This means that the CO_2 leakage term will be negligible since the cytosolic CO_2 concentration will be approximately equal to the external CO_2 concentration. The HCO_3^- transport required to sustain a given internal inorgainc carbon pool will then be:

$$j_c H_{out} = \left(\frac{R_c^3}{3R_b^2} V_{max} - k_m^C (C_{out} - C_{cytosol}) - k_m^{eff} H_{out} + k_m^{eff} H_{cytosol}\right) = \left(\frac{R_c^3}{3R_b^2} V_{max} - k_m^{eff} H_{out} + k_m^{eff} H_{cytosol}\right) \quad \text{(B.18)}$$

We can calculate $H_{carboxyome} = K_{eq} C_{carboxysome}$ from equation (17), and is therefore also independent of k_m^C. In previous work we showed that

$$H_{cytosol} = \frac{(j_c + k_m^{eff}(pH_{out}))H_{out} + \frac{\alpha}{K_\alpha} C_{cytosol}(r = R_b) - k_m^{eff} H_{carboxysome}}{k_m^{eff} G^H + \frac{D}{R_b^2}} G^H \quad \text{(B.19)}$$

We have now shown that all the terms in H_{cyto} are negligibly dependent on the membrane permeability to CO_2. Therefore, the HCO_3^- transport level require to satisfy equation (18) is independent of the membrane permeability to CO_2. This observation is consistent with the low flux of CO_2 leakage in Figure 3.2.

Without Facilitated Uptake, External CO_2 has Little Effect

Unless conversion from CO_2 to HCO_3^- is large we note that the second C_{out} term in equation(15) is negligible for the regimes we study. We will revisit CO_2 uptake and recycling later. Comparing this term against the first term in the numerator, again allows us to put a quantitative description on when this regime holds. Additionally we find that when the transport of HCO_3^- is significant ($j_c > k_m^{eff}(pH_{out})$) we arrive at

$$C_{carboxysome} = \frac{j_c H_{out} G^C - R_c^3 V_{max} G^C (k_m^{eff} G^H + \frac{D}{R_b^2})/(3D)}{k_m^{eff}(G^C K_{eq} + G^H) + \frac{D}{R_b^2}} \tag{B.20}$$

$$H_{carboxysome} = K_{eq} C_{carboxysome} \tag{B.21}$$

Effect of Carboxysome Permeability

Recalling the equation for $G^C = \frac{D}{R_c^2 k_c^C} + \frac{1}{R_c} - \frac{1}{R_b}$, we can see that the carboxysome permeability to CO_2 will only matter if $\frac{D}{R_c^2 k_c^C} > \frac{1}{R_c}$. In other words the carboxysome permeability to CO_2, k_c^C, begins to effectively trap CO_2 in the carboxysome when $k_c^C < \frac{D}{R_c} \approx 2$ cm/s for our base case of a 100 nm carboxysome ($R_c = 50$ nm). Similarly $G^H \approx \frac{D}{R_c^2 k_c^H}$ when $k_c^H < \frac{D}{R_c}$. As common thinking is that $k_c^H \geq k_c^C$, $k_c^H < \frac{D}{R_c}$ may not always hold when $k_c^C < \frac{D}{R_c}$.

Different Carboxysome Peremability for HCO_3^-

An existing hypothesis in the CCM literature is that the carboxysome has differential permeability and is more permeable to HCO_3^- and less permeable to CO_2. Intuitively this would allow more HCO_3^- into the carboxysome and trap more CO_2, thereby accumulating more inorganic carbon in the form of CO_2. We use our model to test weather differential carboxysome permeability enables higher carboxysomal CO_2 concentration for the same level of HCO_3^- transport. In the following figure we show the k_c vs j_c phase space where we have plotted the carboxysome permeability to CO_2, k_c^C, on the y-axis. We plot different ratios (1, 10, 100, 1000) between k_c^C and the carboxysome permeability to HCO_3^-, $k_c^H = \text{ratio} \times k_c^C$.

Examining Figure B.13, we see that making the carboxysome more permeable to HCO_3^- does not improve the function of the CCM as drastically as on might assume. The "turn on" of CO_2 accumulation with decreasing permeability is unaffected by changes to k_c^H, and depends only on the permeability CO_2, k_c^C.The "turn off" of accumulation for lower carboxysome permeabilities is greatly effected by the permeability of the carboxysome to HCO_3^-,

k_c^H. These two effects are exactly what we previously discussed as defining the carboxysome permeability optimum.

As we start at the top of the y-axis and decrease the carboxysome permeability the following occurs: At high permeability not enough CO_2 is trapped, but HCO_3^- enters readily. As we moved to lower permeabilities CO_2 begins to be trapped, but there is a window where HCO_3^- still enters enough to supply the system. Eventually the carboxysome begins to restrict HCO_3^- entry. If the carboxysome is more permeable to HCO_3^- than to CO_2 then the window where CO_2 trapping is effective without restricting HCO_3^- entry broadens. The width of this window (on the y-axis) will also depend strongly on how much of the CO_2 is being fixed.

The "turn off" of the optimum, caused by not allowing enough HCO_3^- into the carboxysome, does slightly increase the amount of transport required to saturate Rubisco at the carboxysome optimum. The reduction in transport required, and therefore CCM cost is around 5% when going from a k_c^C to k_c^H ratio of 1 to 1000.

Effect of Membrane Permeability to H_2CO_3

The sensitivity of the cost to our assumption for the value of the membrane permeability to H_2CO_3 can be determined from the equation derived previously. If we are in a regime where CO_2 leakage is negligible, as is the regime presented in the main paper, the second line holds.

$$\begin{aligned} j_c H_{out} &= \left(\frac{R_c^3}{3R_b^2} V_{max} - k_m^C \left(C_{out} - C_{cytosol} \right) - k_m^{eff}(pH_{out}) H_{out} + k_m^{eff}(pH_{in}) H_{cytosol} \right) \\ &= \left(\frac{R_c^3}{3R_b^2} V_{max} - k_m^{eff}(pH_{out}) H_{out} + k_m^{eff}(pH_{in}) H_{cytosol} \right) \end{aligned} \quad \text{(B.22)}$$

In this equation $k_m^{eff} = k_m^{H_2CO_3} \times 10^{(pK_1 - pH)}$. Therefore, the leakage of H_{total} from the cell will depend linearly on what we assume for $k_m^{H_2CO_3}$. This linear dependence is passed on to the active HCO_3^- transport required to replenish the leaked C_i, and therefore onto the CCM cost. In Figure B.12 you can see this effect, where going from $k_m^{H_2CO_3} = 3 \times 10^{-2}$ to $k_m^{H_2CO_3} = 3 \times 10^{-3}$ (an order of magnitude change), decreases the active HCO_3^- transport needed by an order of magnitude. Decreasing to $k_m^{H_2CO_3} = 3 \times 10^{-4}$ is a little less than an order of magnitude, indicating that the linear dependence breaks down and CO_2 leakage would become important for that value. There is also an order of magnitude change in the optimal carboxysome permeability from 10^{-4} to 10^{-5} across the 2 order of magnitude change in $k_m^{H_2CO_3}$ we are checking.

Size of the Cytosolic HCO_3^- Pool

The HCO_3^- cytosolic pool we assume in our cost calculation has a large effect on the absolute values for the cost calculation. The dependence of HCO_3^- transport required to support a

given internal cytosolic pool can also be seen in the equation in the previous section.

In Figure B.16 the active HCO_3^- transport and carboxysome permeability values required to achieve a particular cytosolic pool are shown. For the Rubisco half-max values assumed in Chapter 3 $K_m = 276\ \mu M$ at internal pH 8, internal HCO_3^- cytosolic pools between 20 and 50 mM are required to saturated Rubisco. For lower K_{M} values, lower cytosolic pools would be required. Additionally, as the K_{M} values are pH dependent, the internal pH of the cell will effect when saturation takes place.

It has recently been suggested that cytosolic HCO_3^- pools of around 5-10 mM can saturate Rubisco. As was discussed in Whitehead et al., this is only possible with our given understanding of the CCM mechanism if the pH in the carboxysome is lower than the rest of the cell, or if the carbonic anhydrase does not act to bring CO_2 and HCO_3^- into equilibrium. Either of these possibilities seems physically questionable given out current understanding of the diffusion rate of protons and the mechanism of carbonic anhydrase – its speed is considered linked to its lack of directionality.

When can CO_2 Scavenging have an Effect?

Next we examine for which parameter regimes $CO_2 \rightarrow HCO_3^-$ conversion activity acts as facilitatd uptake, resulting in a flux of CO_2 into the cell, or scavenging, reducing the leakage of CO_2 out of the cell. As written, the conversion activity does not discriminate between CO_2 that has recently diffused into the cell and CO_2 which has already been in the carboxysome, but leaked back out. In other words, we do not "trace" the history of each CO_2 molecule, or even model each CO_2 molecule explicitly as we treat concentrations. Therefore, we must think about the net CO_2 flux at the cell membrane.

When there is a net flux of CO_2 out of the cell, there is by definition, no facilitated uptake. If there is zero net flux of CO_2 at the cell membrane, then the cell is scavenging 100% of the CO_2 leaking out of the carboxysome. If there is net flux of CO_2 into the cell, then not only is the conversion mechanism scavenging all CO_2 leaking out of the cell, but also reducing the cytosolic CO_2 concentration below the external concentration and facilitating uptake.

From Figure B.15 it is evident that increasing $CO_2 \rightarrow HCO_3^-$ decreases the need for active HCO_3^- uptake at optimal carboxysome permeability. We can also determine when the conversion mechanism switches from being primarily a scavenging mechanism to also acting as facilitated uptake. The orange dashed lines show the active HCO_3^- transport rate for each carboxysome permeability k_c value where the net CO_2 flux at the cell membrane is zero. Above this line the system is scavenging and below it the system is scavenging and causing facilitated uptake. We have plotted these curves for increasing $CO_2 \rightarrow HCO_3^-$ conversion strengths.

For $\alpha = 10^{-6}$ cm/s, the system is not facilitating uptake of CO_2 for the optimal permeability. At this point the minimum active HCO_3^- transport rate required to saturate RuBiSCO is approximately the same as without the CO_2 conversion mechanism. For higher α the conversion system is scavenging at the optimal carboxysome permeability. In this case

the system requires less HCO_3^- transport because CO_2 conversion is acting as facilitated uptake, and contributing to the internal inorganic carbon pool. Note that the absolute flux from scavenging and facilitated uptake will vary drastically over this space.

In Chapter 3 we showed that the loss of CO_2 from leakage out of the cell is about an order of magnitude smaller than HCO_3^- leakage without CO_2 conversion. When we are at optimal carboxysome permeability, scavenging will have little effect, because there is just not very much CO_2 in the cytosol to scavenge. At higher carboxysome permeabilities though, the scavenging mechanism could compensate and matter much more. Thus the relative importance of the CO_2 conversion mechanism for scavenging vs facilitated uptake depends highly on the yet experimentally undetermined carboxysome permeability.

To identify when facilitated uptake starts, we can find for what CO_2 concentration in the carboxysome the flux at the cell membrane is zero, or when $C_{cytosol} = C_{out}$.

$$\text{CO}_2 \text{ flux} = k_m^C(C_{cytosol}(R_b) - C_{out}). \tag{B.23}$$

Not enough is known about the mechanism of vectorial $CO_2 \rightarrow HCO_3^-$ to estimate a cost in the way we have done for HCO_3^- uptake. Therefore we cannot asses the effect of conversion on the cost.

B.11 Tables

Parameter	Definition	Value	Ref.
D	small molecule diffusion coefficient	10^{-5} cm^2/s	[107]
$k_m^{\mathbf{c}}$	cell membrane permeability to CO_2	10^{-5} cm^2/s	[116]
$k_m^{\mathbf{H_2CO_3}}$	cell membrane permeability to H_2CO_3	10^{-5} cm^2/s	[329]
$k_m^{\mathbf{HCO_3^-}}$	cell membrane permeability to HCO_3^-	10^{-5} cm^2/s	[116]
R_c	carboxysome radius	50 nm	[135, 263]
R_b	cell radius	500 nm	[182, 260]
$[\mathbf{CO_2}]$	CO_2 concentration	300 μM	Table B.4
$[\mathbf{O_2}]$	O_2 concentration (unless otherwise noted)	15 μM	Table B.4
$N_{\mathbf{Rub}}$	Rubisco active sites per carboxysome	2000	[135]
$N_{\mathbf{CA}}$	CA active sites per carboxysome	100	[122]

Table B.1: Model parameters. Other parameters are as described in [182]. Parameters listed are needed for the equations in this text and may have changed from previous work. H_2CO_3 assumed to have the same membrane permeability as formic acid (H_2CO_2) due to its identical charge and similar size.

Enzyme	k_{cat} (s^{-1})	K_{M} (μM)	Specificity ($S_{\text{C/O}}$)	Ref.
Carboxysomal Rubisco (*S. elongatus* PCC 6301)	11.6	340	43	[262]
Cytoplasmic Rubisco (*N. tabacum*)	3.4	10.7	82	[262]
High-Specificity Rubisco (*G. monolis*)	1.2	3.3	166	[262]
High-Specificity Rubisco (*G. monolis*)	4.6×10^4	9.2×10^3	NA	[122]

Table B.2: Enzyme kinetic parameters used. Parameters for the carboxysomal Rubisco and CA were used for the full CCM model throughout. pH dependence of enzyme kinetics is described above.

Species	$\Delta_f G^\circ$(kJ/mol)	N_{H}	z
CO_2 + H_2O	-623.2	2	0
H_2CO_3	-606.3	2	0
HCO_3^-	-586.8	1	-1
CO_3^{2-}	-527.8	0	-2

Table B.3: Chemical formation energies of aqueous C_i species. N_{H} is the number of hydrogens and z is the net charge. All $\Delta_f G^\circ$values are from Alberty [7] and the value for CO_2 + H_2O denotes aqueous CO_2 and not the gaseous form.

Species	Partial Pressure	Solubility (L atm/mol)	Equilibrium Conc. (μM)
CO_2	400 ppm	29.41	13.6
O_2	21 %	769.2	273.0

Table B.4: Equilibrium concentrations of CO_2 and O_2 in the aqueous phase. Assumed atmospheric partial pressures are given for CO_2 and O_2 along with Henrys law solubility constants. The equilibrium aqueous concentration is the calculated in Molar units as concentration = (partial pressure) / (solubility).

Species	Permeability (cm/s)	Ref.
CO_2	0.3	[107, 116]
H_2CO_3	3×10^{-3}	[116, 313, 329]
HCO_3^-	$< 2 \times 10^{-7}$	[116]
CO_3^{2-}	$< 2 \times 10^{-7}$	*

Table B.5: Membrane permeability coefficients for all relevant species. H_2CO_3 is approximated using values measured for the similarly-sized formic and acetic acids (all having a net charge of 0). CO_3^{2-} assumed to be less permeable than HCO_3^-.

Species	I = 0 M	I = 0.05 M (freshwater)	I = 0.2 M (cytosol)	I = 0.75 M (seawater)
pK_1 (H_2CO_3:HCO_3^-)	3.4	3.2	3.2	3.1
pK_2 (HCO_3^-:CO_3^{2-})	10.3	10.0	9.8	8.9
pK_{eff} (CO_2:HCO_3^-)	6.4	6.2	6.1	5.9

Table B.6: Enzyme kinetic parameters used. Parameters for the carboxysomal Rubisco and CA were used for the full CCM model throughout. pH dependence of enzyme kinetics is described above.

Appendix C

Competition for Carboxysome Pores

C.1 Introduction

In deriving a mathematical model of the carboxysome-based CCM of cyanobacteria (Chapter 3), we found a conundrum. On the one hand, the carboxysome must be CO_2-tight, having a CO_2 permeability on the order of $10^{-4} - 10^{-5}$ cm/s in order for the CCM to function in an energetically-efficient manner. On the other hand, considering the geometric structure of the carboxysome predicts a CO_2 permeability on the order of $10^{-1} cm/s$. What could resolve this 3-4 order-of-magnitude discrepancy?

I propose a resolution based on the chemical makeup of the cytosol: cytosol contains high concentrations of negatively charged small molecules, e.g. those containing phophates and carboxylic acids. These molecules could bind positive charge found in the pores of carboxysome shell proteins [154], occluding the pore and impeding the passage of CO_2. The total concentration of phosphate and carboxylic acid moieties exceeds 100 mM in *E. coli*. Competition from such molecules, including those not directly involved in CO_2 fixation, would ensure that the carboxysome remains CO_2-tight and the CCM functions efficiently. The CCM itself generates cytosolic HCO_3^- concentration in excess of 10 mM by pumping C_i. Due its negative charge, HCO_3^- should also out-compete CO_2 for the carboxysome pore.

In addition to being CO_2-tight, the carboxysome must also permit HCO_3^- entry in order for the CCM to function as diagrammed in Figure 1.7. In Chapter 3 and Appendix B I showed that the carboxysome must have an HCO_3^- permeability $\geq 10^{-5}$ cm/s for the CCM to function efficiently. If the carboxysome does not admit HCO_3^- quickly enough, the carboxysomal CO_2 concentration will be too low to saturate Rubisco (Figure 3.4).

However, if the negatively charged molecules found so abundantly in cytosol block the carboxysome pore, it stands to reason that they might slow HCO_3^- entry. The "competitive pore" model offers resolution here as well: HCO_3^- is negatively charged, it can compete for pore access more easily than CO_2. it is not as easily out-competed for pore access as CO_2. Not only does the "competitive pore" model address the conundrum of the absolute carboxysome permeability, but also suggests that charge in carboxysome pores make the

CCM robust to fluctuations in the composition of cytosol. This may help explain how it can be that the CCM was laterally transferred among cyanobacteria and chemolithoautotrophic bacteria.

This document proceeds as follows. First, I will give simplified calculations that argue that a low CO_2 permeability in the range of $10^{-4} - 10^{-5}$ cm/s is required for CCM function. Second, I will show that a simple geometric calculation implies a much higher permeability number around 0.1 cm/s. Finally, I will show that a simple model of competition at carboxysome pores enables us to close this 3-5 order-of-magnitude gap.

C.2 Geometric Calclation of Carboxysome Permeability

Carboxysome Surface Area

A detailed calculation of carboxysome surface area is given in Appendix B. This calculation gives a value of 2.4×10^4 nm^2 $= 2.4 \times 10^{-10}$ cm^2 for the α-carboxysome. There is some variation in measurements of α-carboxysomes [135, 263] and β-carboxysomes are about twice as large in linear dimension (cite:sun2019), so we use a generic value of $= 5 \times 10^{-10}$ cm^2.

Effective Pore Area

With the exception of the large-pored hexamers *csos1D* and *ccmP*, which are minor constituents of the α- and β-carboxysomes respectively [156, 250, 330], carboxysome pores have typical radii of $\approx 2-3$ Å [154] and lengths of $\approx 2-4$ nm. However, CO_2 and O_2 are not that much smaller than these pores and so their size cannot be neglected. Following the logic of Hinzpeter et al. [128], we subtract off the van der Waals radius of carbon, which gives an effective radius ranging from $0.1 - 1.3$ Å.

Expected Permeability

Again following [128], we can calculate the carboxysome permeability from its geometry and diffusional properties of small molecules.

$$P = \frac{N_p \times \pi r_{\text{eff}}}{A} \frac{D}{\lambda}$$

Where $\frac{D}{\lambda}$ is an estimate of the velocity of molecular transit through a carboxysome pore. The diffusion coefficient ($D \approx 3 \times 10^{-5}$cm^2/s) is divided by the length of the pore ($\lambda \approx 3 \times 10^{-7}$ cm) and gives a value of ≈ 50 cm/s. Using representative values for all other parameters, we can calculate the expected permeability

$$P \approx \frac{800 \times \pi(10^{-8}\text{cm})^2}{5 \times 10^{-10}\text{cm}^2} \frac{3 \times 10^{-5}\text{cm}^2/\text{s}}{3 \times 10^{-7}\text{cm}} \approx 0.05\text{cm/s}$$

A more thorough version of this same calculation, based on uniform sampling from the range of plausible parameter values, is presented in Figure C.1.

C.3 CO_2 Permeability Required for CCM Function

A simplified calculation helps us gain intuition about the CO_2 permeability values required for the CCM to function. To simplify the calculation we will make a number of assumptions. First, let's assume that the total CO_2 fixation rate equals the rate of CO_2 leakage from the carboxysome at steady state. This assumption is based on the premise that the CCM is relatively energetically efficient and is borne out by measurements [132]. Remembering also that the purpose of the CCM is to saturate Rubisco with CO_2, we assume that Rubisco is saturated and all 2×10^3 active sites operate at their carboxylation $k_{\text{cat,C}}$, which is ≈ 10 s^{-1} (Chapter 2).

In order to saturate Rubisco, the carboxysomal CO_2 concentration must be well in excess of the Michaelis constant towards CO_2 ($K_C \approx 100 - 300$ μM for carboxysomal Rubiscos). For simplicity, we will assume that the carboxysomal CO_2 concentration is 500 μM, though in some cases is less than double the measured K_C. Given that, it is fair to assume that the carboxysomal CO_2 concentration greatly exceeds the cytosolic concentration. That is:

$$\Delta\text{CO}_2 = ([\text{CO}_2]_{\text{in}} - [\text{CO}_2]_{\text{out}}) \approx [\text{CO}_2]_{\text{in}}$$

Assuming steady state leakage equals fixation for a single carboxysome

$$\begin{aligned} k_{\text{cat,C}} \times N_{\text{rub}} &= N_A \times P_{\text{CO}_2} \times \Delta\text{CO}_2 \times A_C \\ &\approx N_A \times P_{\text{CO}_2} \times [\text{CO}_2]_{\text{in}} \times A_C \end{aligned}$$

Here A_C is the carboxysomal surface area in cm units, P_{CO_2} is the CO_2 permeability in cm/s, N_{rub} is the number of Rubisco active sites (≈ 2000) and N_A is Avogadro's number. We can then solve for P_{CO_2} as a function of the other parameters.

$$\begin{aligned} P_{\text{CO}_2} &= \frac{k_{\text{cat,C}} \times N_{\text{rub}}}{N_A \times [\text{CO}_2]_{\text{in}} \times A_C} \\ &\approx \frac{10s^{-1} \times 2 \times 10^3}{6 \times 10^{23}\frac{\text{molecules}}{\text{mol}} \times 500 \times 10^{-6}\frac{\text{mol}}{\text{dm}^3} \times 10^{-3}\frac{\text{dm}^3}{\text{cm}^3} \times 5 \times 10^{-10}\text{cm}^2} \\ &\approx 10^{-4}\text{cm/s} \end{aligned}$$

Figure C.1 shows that these permeability values are several orders of magnitude below those derived from geometric considerations above, irresepective of exactly what assumptions we make about the carboxysome size, pore diameter, etc. The model presented in Chapter 3 and Appendix B amounts to a more sophisticated version of this same calculation and produces a similar range for viable CO_2 permeabilities.

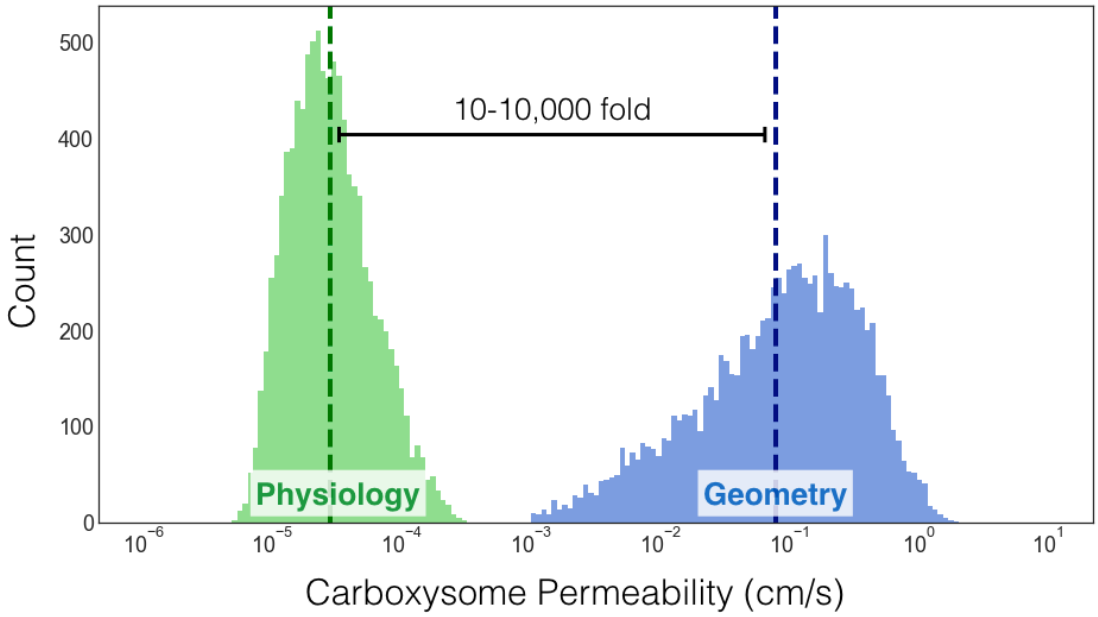


Figure C.1: Geometric and physiological calculations of carboxysome permeability to CO_2 differ by orders of magnitude. 10^4-fold rounds of uniform parameter within empirical ranges (e.g. 100-200 nm carboxysome diameter) was used to generate a distribution of values for the geometric and physiological calculations of carboxysome CO_2-permeability described in the text. Physiologically-derived values have geometric mean and median both $\approx 3 \times 10^{-5}$ cm/s. Values derived from geometric calculation have geometric mean 0.08 and median 0.1 cm/s.

C.4 Competition for Carboxysome Entry

In order to consider permeability to CO_2 and HCO_3^- independently, we decompose the permeability coefficient into three multiplicative factors.

$$P = \hat{P} \times O$$

Here $\hat{P}$ is the permeability coefficient one would expect if all pores are unobstructed and O is the fraction of pores that are "open." To simplify, we'll assume that $\hat{P}$ does not differ for CO_2 and HCO_3^-, as they have roughly the same diffusion coefficient in cytosol and both presumably fit through the pore. We proceed to calculation $O(C, H, X)$ - the fraction of unobstructed pores as a function of CO_2 (C), HCO_3^- (H), and non-specific competitor (X)

concentrations. First, however, it is important to get a sense of what these concentrations are.

Negatively Charged Metabolites are Abundant in Cytosol

Small molecules carrying negative charge, e.g. on phosphates and carboxylic acids, are ubiquitous in cellular metabolism and found at high concentration in cells of all organisms. Glutamate, for example, carries a net charge $Z \approx -1$ at pH 7 and is the most abundant metabolite in *E. coli*, assuming a concentration approaching 100 mM [36]. In human and yeast cells, glutamate is similarly abundant, assuming a concentration ≥ 40 mM (Park et al., Nature Chem. Bio. 2016). In addition, microbial metabolite concentrations correlate with the number of charged atoms ($R^2 > 0.4$, [25]) suggesting a biochemical principle: charged molecules are likely more abundant than their uncharged counterparts across the tree of life.

Using data from Bennett et al., Nature Chem. Bio 2009, I show that that the total concentration of negatively-charged small molecules exceeds 20 mM in *E. coli* regardless of how we define "small" or "negatively charged" (Figure C.2). The total concentration may exceed 100 mM if glutamate is included. In phototrophically-grown cyanobacteria, several small, negatively charged molecules have been found to be very abundant (Diamond et al., PNAS 2015; Whitehead et al., Plant Phys. 2014) [320]. These include 1-10 mM concentrations of free phosphate and several CBB cycle intermediates, including the products and substrates of the RuBisCO reaction (e.g. 3-phosphoglycerate, ribulose 1,5-bisphophate).

Competitors Block the Carboxysome Pore

As summarized in Figure C.3, we assume that the pore has no affinity for CO_2, but that it can bind HCO_3^- and a pool of negatively charged competitors X. We now calculate the fraction O from these binding energies, noting that O will differ for CO_2 and HCO_3^- because HCO_3^- can bind the pore (due to its negative charge). We will begin with the case of CO_2 and calculate $O = Pr(\text{empty})$, the probability the pore is not occupied with a competitor of CO_2.

Derivation for CO_2 Entry

Both HCO_3^- and X can plug the pore and prevent CO_2 entry as diagrammed in Figure C.3. Consider the binding of HCO_3^-

$$K_D^{HCO_3^-} = \frac{[\text{Free Pore}][HCO_3^{-free}]}{[\text{Pore} \cdot HCO_3^-]}$$

$$[\text{Pore} \cdot HCO_3^-] = \frac{[\text{Free Pore}][HCO_3^{-free}]}{K_D^{HCO_3^-}}$$

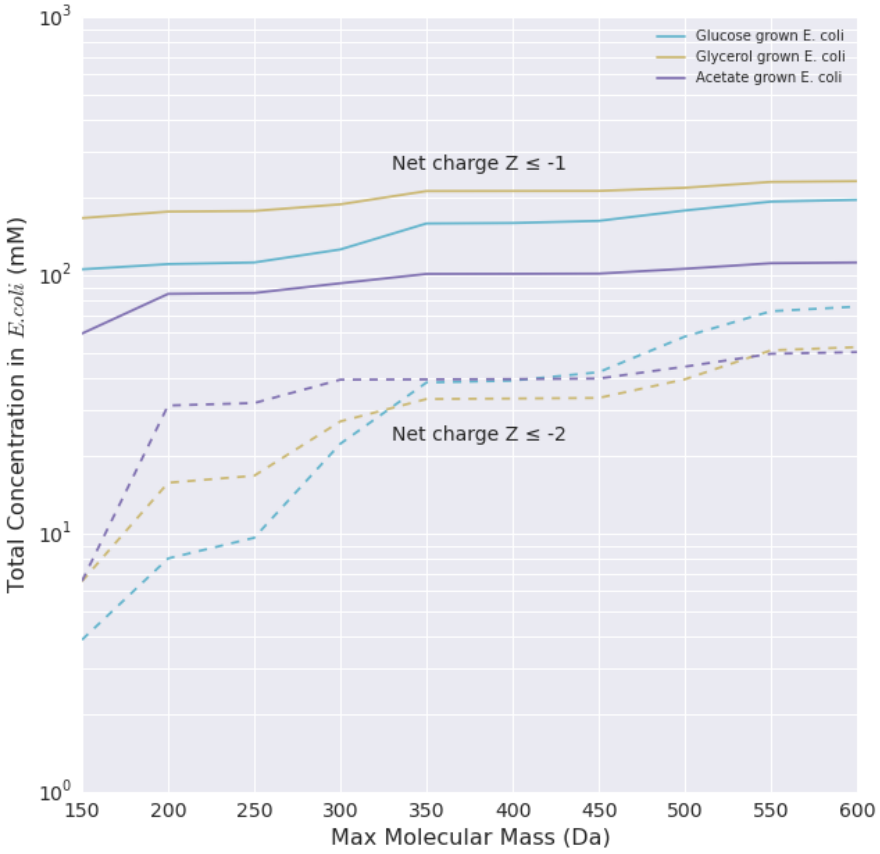
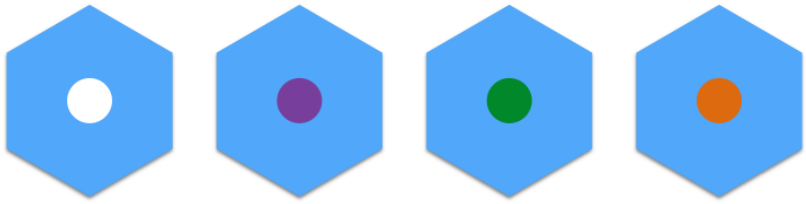


Figure C.2: Negatively charged small molecules are abundant in *E. coli* regardless of growth condition. The total concentration of small molecules with net charge $Z \leq -1$ is 100-200 mM regardless of growth condition. Small molecules with $Z \leq -2$ have a total concentration ranging from 10-50 mM depending on the maximum molecular mass considered. Data from [36]. Measurements of Eukaryotic metabolite concentrations are consistent and available in [218].

We are interested in calculating equilibrium probability the pore is empty, i.e. the degree to which competition reduces number of unoccupied pores and, therefore, the probability of CO_2 entry. This probability equals the ratio of empty pores to pores total $Pr(empty) = \frac{\text{[Free Pore]}}{\text{[Pore Tot.]}}$.



	empty	CO_2-bound	HCO_3^- bound	X bound
association energy	-	ref. E = 0	E_H	$E_X = E_H - \varepsilon$
K_D	-	-	$\exp(E_H/RT)$	$\exp(E_X/RT)$
typical concentration	-	10 uM	10-50 mM	10-200 mM

Figure C.3: Thermodynamic model of competition at the carboxysome pore. The pore binds HCO_3^- with binding energy E_H and binds non-specific competitors X more tightly, with binding energy $E_X = E_H - \epsilon$. CO_2 assumes a concentration near Henry's law equilibrium with atmosphere (≈ 15 μM), while HCO_3^- has 10-50 mM concentration due to the action of the CCM. As shown in Figure C.2, the X concentration is at least 10 mM and might exceed 100 mM.

$$\begin{aligned}
[\text{Pore Tot.}] &= [\text{Free Pore}] + [\text{Pore} \cdot HCO_3^-] + [\text{Pore} \cdot X] \\
&= [\text{Free Pore}] + \frac{[\text{Free Pore}][HCO_3^{-free}]}{K_D^{HCO_3^-}} + \frac{[\text{Free Pore}][X^{free}]}{K_D^X} \\
&= [\text{Free Pore}] \left(1 + \frac{[HCO_3^{-free}]}{K_D^{HCO_3^-}} + \frac{[X^{free}]}{K_D^X}\right)
\end{aligned}$$

Finally,

$$\begin{aligned}
Pr(empty) &= \frac{[\text{Free Pore}]}{[\text{Pore Tot.}]} \\
&= \frac{1}{\left(1 + \frac{[HCO_3^{-free}]}{K_D^{HCO_3^-}} + \frac{[X^{free}]}{K_D^X}\right)}
\end{aligned}$$

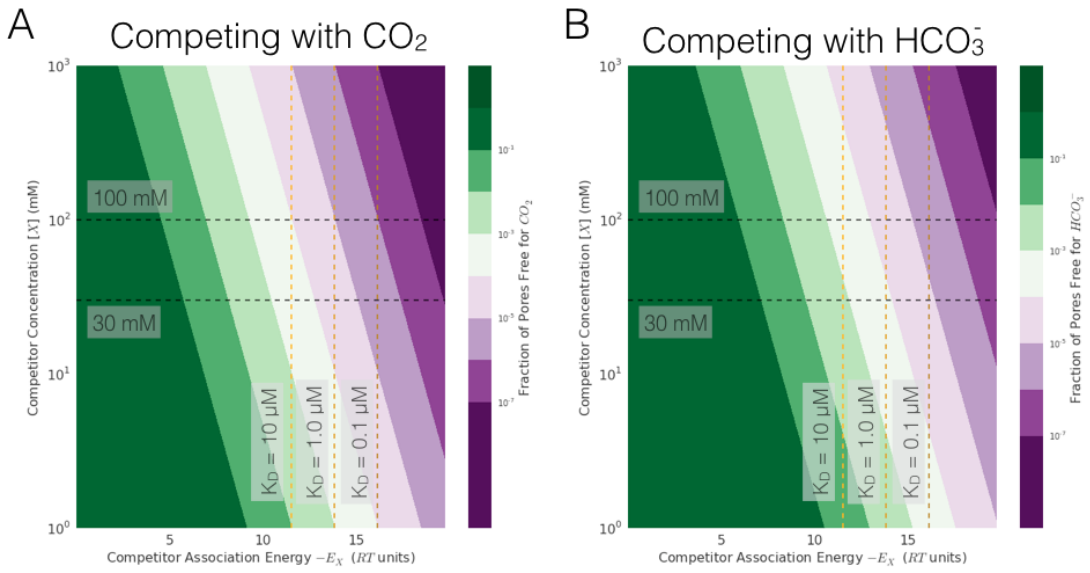


Figure C.4: Considering competition for the pore enables fast HCO_3^- entry and CO_2 retention in the carboxysome.

Derivation for HCO_3^- Entry

As bicarbonate carries negative charge and can (presumably) interact with the pore, the definition of an empty pore is somewhat different for HCO_3^- than for CO_2. Since an HCO_3^--bound pore can transit HCO_3^-, we now we define an empty pore as one not bound by X.

$$\begin{aligned} Pr(\text{empty}) &= 1 - Pr(\text{X bound}) = 1 - \frac{[\text{Free Pore}][X]}{K_D^X[\text{Pore Tot.}]} \\ &= 1 - \frac{[X]}{K_D^X\left(1 + \frac{[HCO_3^{-free}]}{K_D^{HCO_3^-}} + \frac{[X^{free}]}{K_D^X}\right)} \\ &= 1 - \frac{f}{\frac{K_D^X}{[\text{HCO}_3^-]} + f + \exp\left(-\epsilon/RT\right)} \end{aligned}$$

where $f = \frac{[X]}{[\text{HCO}_3^-]}$ is the fold excess of competitor over HCO_3^-. Notice that this expression is independent of the absolute binding energy of HCO_3^-, E_H, and instead depends only on the absolute binding energy of the competitor X and the difference in binding energy between X and HCO_3^-, ϵ. So while we must set ϵ and E_H to calculate $Pr(empty)$, we need not fix E_H explicitly although its value is implicitly set by the other two.

In Figure C.4B we assumed a $E_H \approx 4.5RT$ and calculated $Pr(free)$ for varying ϵ and $[X]$. Notice that this plot is essentially a rightward-shifted version of Figure C.4A. As such,

competitor concentrations that keep the carboxysome CO_2-tight still allow for sufficiently high HCO_3^- permeabilities on the order of $10^{-3} - 10^{-5} cm/s$. This speaks to the robustness of the CCM, suggesting that it can operate efficiently over the entire range of plausible competitor concentrations.

Appendix D

Comprehensive Characterization of Bacterial CCM Genes by Transposon Mutagenesis

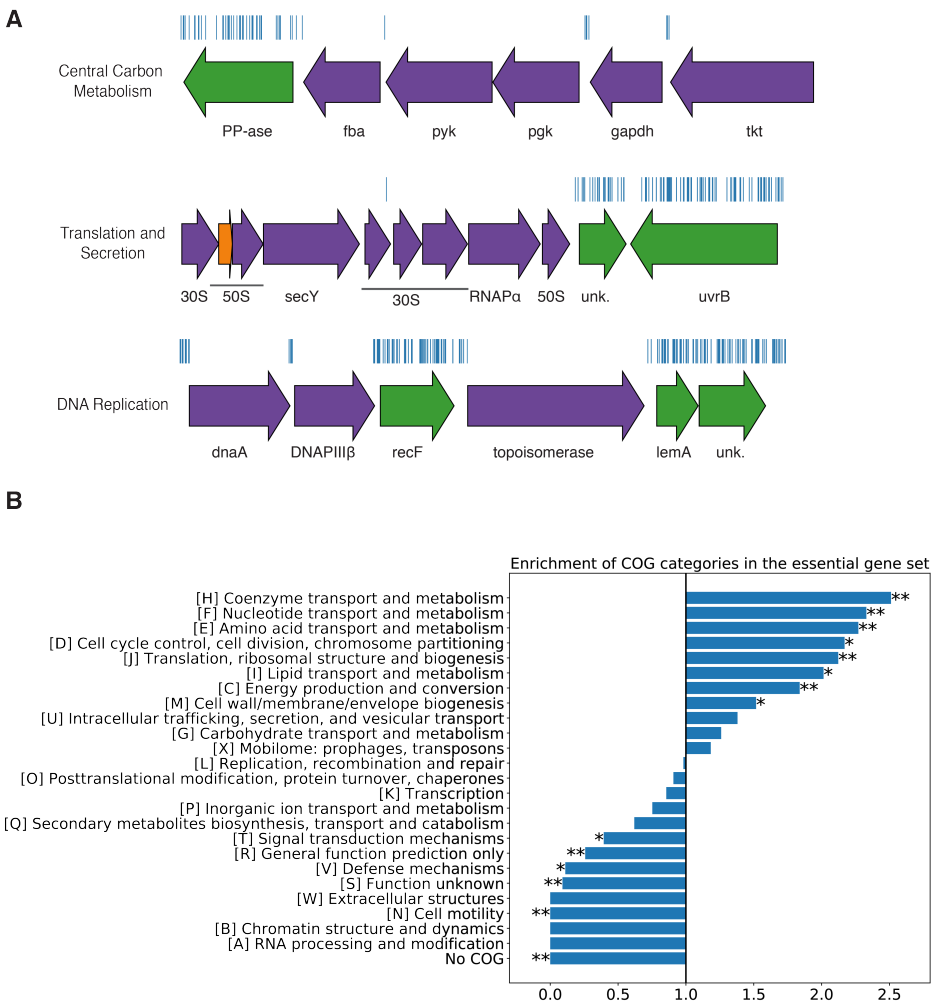


Figure D.1: The essential gene set is enriched for COGs associated with essential cellular processes. (A) Representative essential genes and nonessential genes in the *H. neapolitanus* genome. The blue track indicates the presence of an insertion. Genes in purple were called essential and genes in green are nonessential. Genes labeled "unk." are hypothetical proteins. The first genomic locus contains 5 essential genes involved in glycolysis or the CBB cycle including pyruvate kinase (pyk) and transketolase (tkt). The 8 essential genes in the second locus encodi 30S and 50S subunits of the ribosome, the secY secretory channel, and an RNA polymerase subunit. Essential genes in the third example locus include topoisomerase and DNA polymerase III β. (B) COG enrichments were calculated by dividing the fraction of genes in the essential gene set associated with this COG category by the fraction of genes in the genome associated with this category. "*" denotes that this COG is enriched (or depleted) with Bonferroni corrected $p < 0.05$ by a hypergeometric test, and "**" denotes $p < 5 \times 10^{-4}$ 5X10^{-4}. Exact p values are as follows for each category, No COG:$< 10^{-10}$, C:10^{-5}, D:10^{-2}, E:$< 10^{-10}$, F:4×10^{-6}, H:$< 10^{-10}$, I:6×10^{-4}, J:$< 10^{-10}$, M:6×10^{-3}, N:6×10^{-6}, R:6×10^{-5}, S:8×10^{-8}, T:3×10^{-2}, V:10×10^{-2}. In panel (A), the following abbreviations are used: exopolyphosphatase (PP-ase), fructose-bisphosphate aldolase class II (fba), pyruvate kinase (pyk), phosphoglycerate kinase (pgk), type I glyceraldehyde-3-phosphate dehydrogenase (gapdh), transketolase (tkt), 30S ribosomal protein (30S), 50S ribosomal protein (50S), preprotein translocase subunit SecY (SecY), DNA-directed RNA polymerase subunit alpha (RNAP), hypothetical protein (unk.), excinuclease ABC subunit UvrB (UvrB), chromosomal replication initiator protein dnaA (dnaA), DNA polymerase III subunit beta (DNAPIIIβ), DNA replication and repair protein recF (recF), DNA topoisomerase (ATP-hydrolyzing) subunit B (topoisomerase), lemA family protein (LemA).

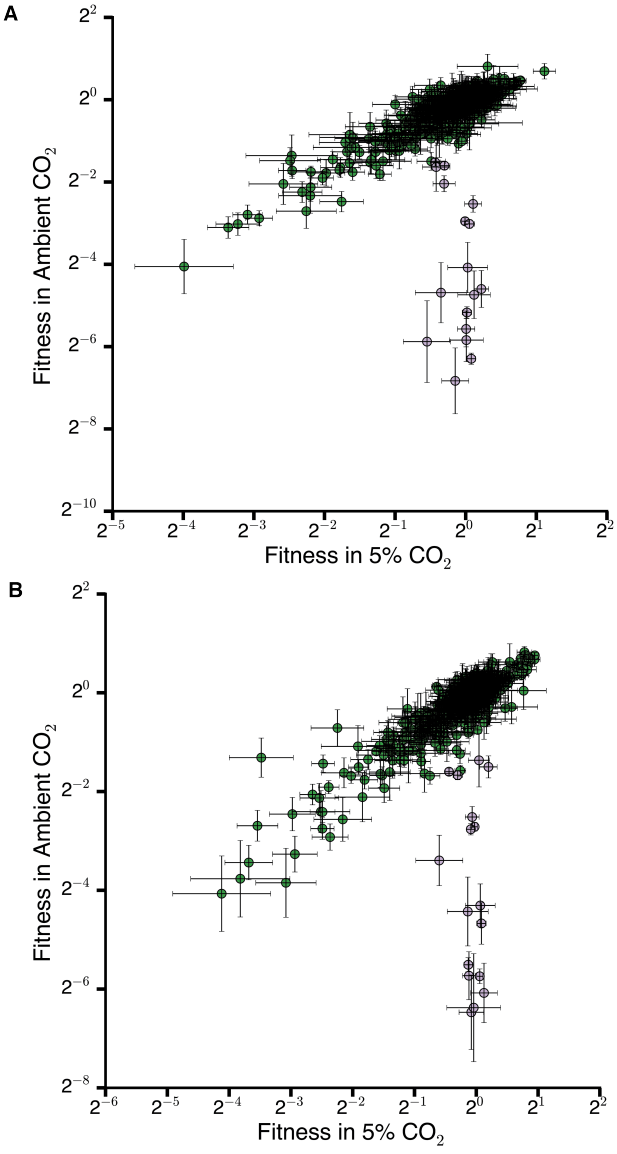


Figure D.2: Gene fitness measurements for replicates. Fitness effects of gene knockouts in 5% CO_2 as compared to ambient CO_2. The effects of single transposon insertions into a gene are averaged to produce the gene-level fitness value plotted. Error bars represent standard error of the mean. We define HCR mutants as those displaying a twofold fitness defect in ambient CO_2 relative to 5% CO_2. HCR genes are colored light purple. Panel (A) contains data from the first replicate experiment and panel (B) the second.

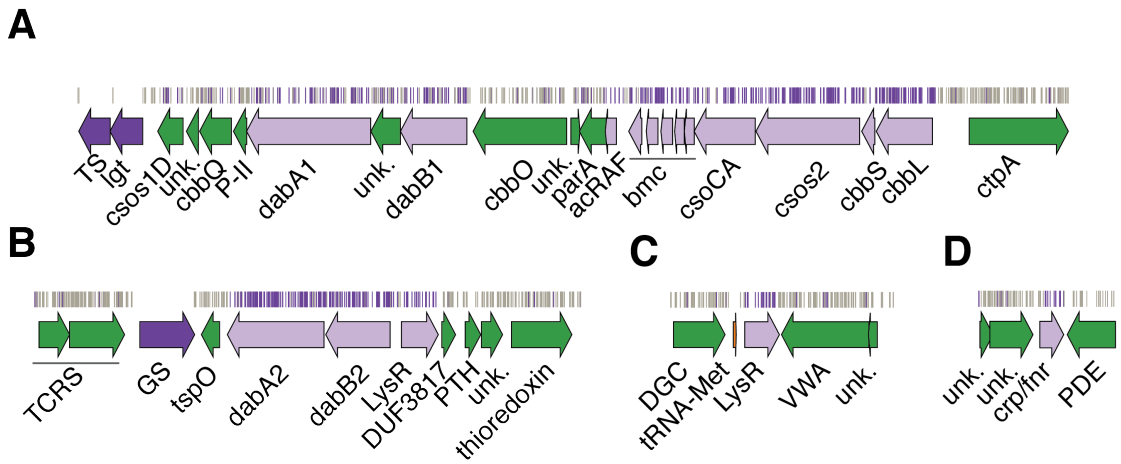


Figure D.3: Genomic context of *H. neapolitanus* HCR genes identified in our genome-wide screen. Panels A-D show regions of the *H. neapolitanus* genome containing genes annotated as HCR. Essential genes are in dark purple, HCR genes are in light purple, and other genes are in green. The top tracks show the presence of an insertion in that location. Insertions are colored colored grey unless they display a twofold or greater fitness defect in ambient CO_2, in which case they are colored purple. (A) The gene cluster containing the carboxysome operon (HNEAP_RS04660-HNEAP_RS04620) and a second CCM-associated operon. This second operon contains acRAF (HNEAP_RS04615), a FormIC associated cbbOQ-type Rubisco activase (HNEAP_RS04575 and HNEAP_RS04600), parA (HNEAP_RS04610), P-II (HNEAP_RS04580) and dabAB1 (dabA1: HNEAP_RS04585 and dabB1: HNEAP_RS04620). (B) The DAB2 operon and surrounding genomic context (lysR: HNEAP_RS01040, dabA2: HNEAP_RS01030, and dabB2: HNEAP_RS01035). (C) The genomic context of a lysR-type transcriptional regulator (HNEAP_RS05490) that shows an HCR phenotype. (D) Genomic context of a crp/fnr-type transcriptional regulator that displays an HCR phenotype (HNEAP_RS07320). Accession numbers and gi numbers for selected genes can be found in Table S1. Abbreviations: thymidylate synthase (TS), prolipoprotein diacylglyceryl transferase (lgt), Rubisco activase Rubisco activase subunits (cbbOQ), nitrogen regulatory protein P-II (P-II), ParA family protein (parA), csos1CAB and csos4AB (bmc), copper-translocating P-type ATPase (ctpA), DNA-binding response regulator and two-component sensor histidine kinase (TCRS), glutamate--ammonia ligase (GS), tryptophan-rich sensory protein (tspO), DUF3817 domain-containing protein (DUF3817), aminoacyl-tRNA hydrolase (PTH), thioredoxin domain-containing protein (thioredoxin), sensor domain-containing diguanylate cyclase (DGC), methionine tRNA (tRNA-Met), VWA domain-containing protein (VWA), diguanylate phosphodiesterase (PDE).

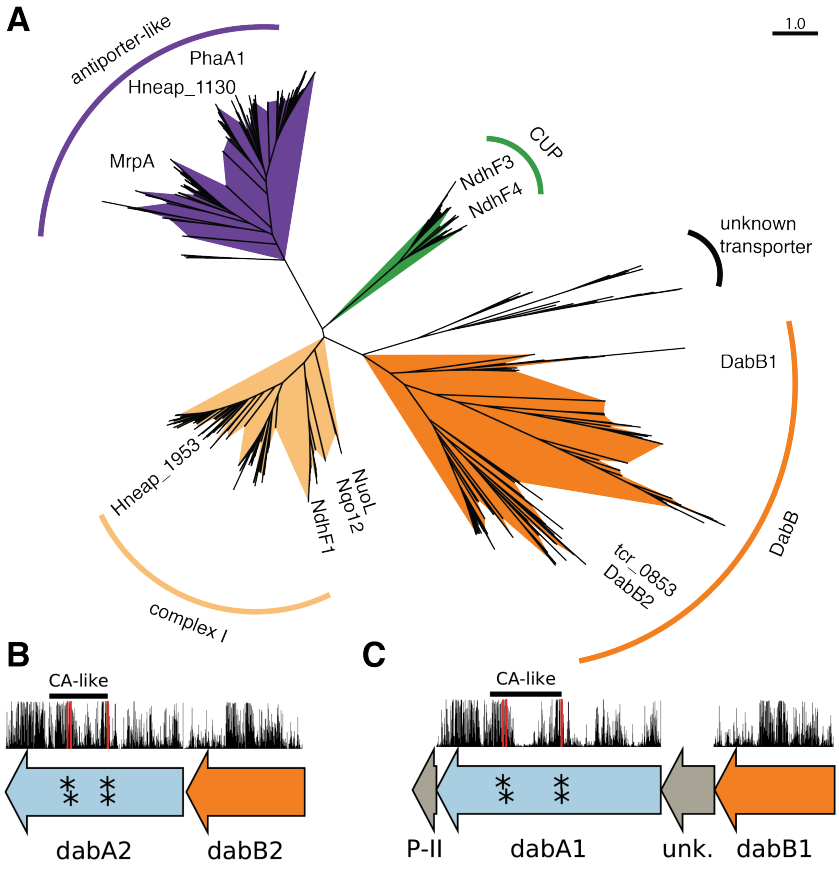


Figure D.4: PFAM0361 is a large and diverse protein family containing conserved DabB sequences. (A) PF0361 contains multiple group with different activities. Subfamilies include Mrp-family cation antiporters, proton translocating subunits of complex I, membrane subunits of CUP (CO_2 uptake protein) complexes, and DabB proteins. Due to the diversity of the family, we cannot draw conclusions about DAB mechanism from homology. Panel (A) contains an approximate maximum likelihood tree of PF0361 genes. The purple clade contains the *B. subtilis* and *S. aureus* MrpA cation antiporter subunits and the *S. meliloti* antiporter PhaA1. The light orange clade contains the known cation translocating subunits of complex I: nuoL from *E. coli*, Nqo12 from *T. thermophilus*, and NdhF1 from both *S. elongatus* PCC 7942 and *T. elongatus* BP-1. The green clade contains CUP-associated membrane subunits ndhF3 from both *S. elongatus* PCC 7942 and *T. elongatus* BP-1 and ndhF4 from from the same two species. The dark orange clade includes DabB1-2 and tcr_0853 from *T. crunogena*. We note that the clade containing DabB1-2 is distinct from that containing known complex I subunits or to mrp-family antiporters. This tree is consistent with our model, where DabB is not bound to a redox-coupled complex but rather couples redox-independent cation transport to CA activity (Figure 4.7). Scale bar indicates one substitution per site. (B-C) As dicussed in Chapter 4, DAB1 is a segment of an 11-gene operon directly downstream of the carboxysome operon that contains CCM-associated genes. Both DAB1 (B) and DAB2 (C) "operons" contain two distinct genes we term DabB and DabA. DabA is annotated as Domain of Unknown Function 2309 (DUF2309, PFAM:PF10070) and appears to be a soluble protein. Approximately one third of dabA is distantly homologous to a type II β-CA. CA-like regions are marked with a line, and the four residues expected to be involved in binding the catalytic zinc ion are marked by asterisks. The height of the asterisks has been varied to make them distinguishable despite proximity in sequence space. DabB is homologous to a cation transporter in the same family as the H+ pumping subunits of respiratory complex I (PFAM:PF00361). The DAB1 operon also contains a protein of unknown function between DabA1 and DabB1. This protein has distant homology to DabA1 but is truncated to half length. Vertical bars above the genes indicate percent conservation of that particular amino acid position in a multiple sequence alignment (Methods). Putative active site residues are in red. All active site residues are highly conserved with percent identities of greater than 99%.

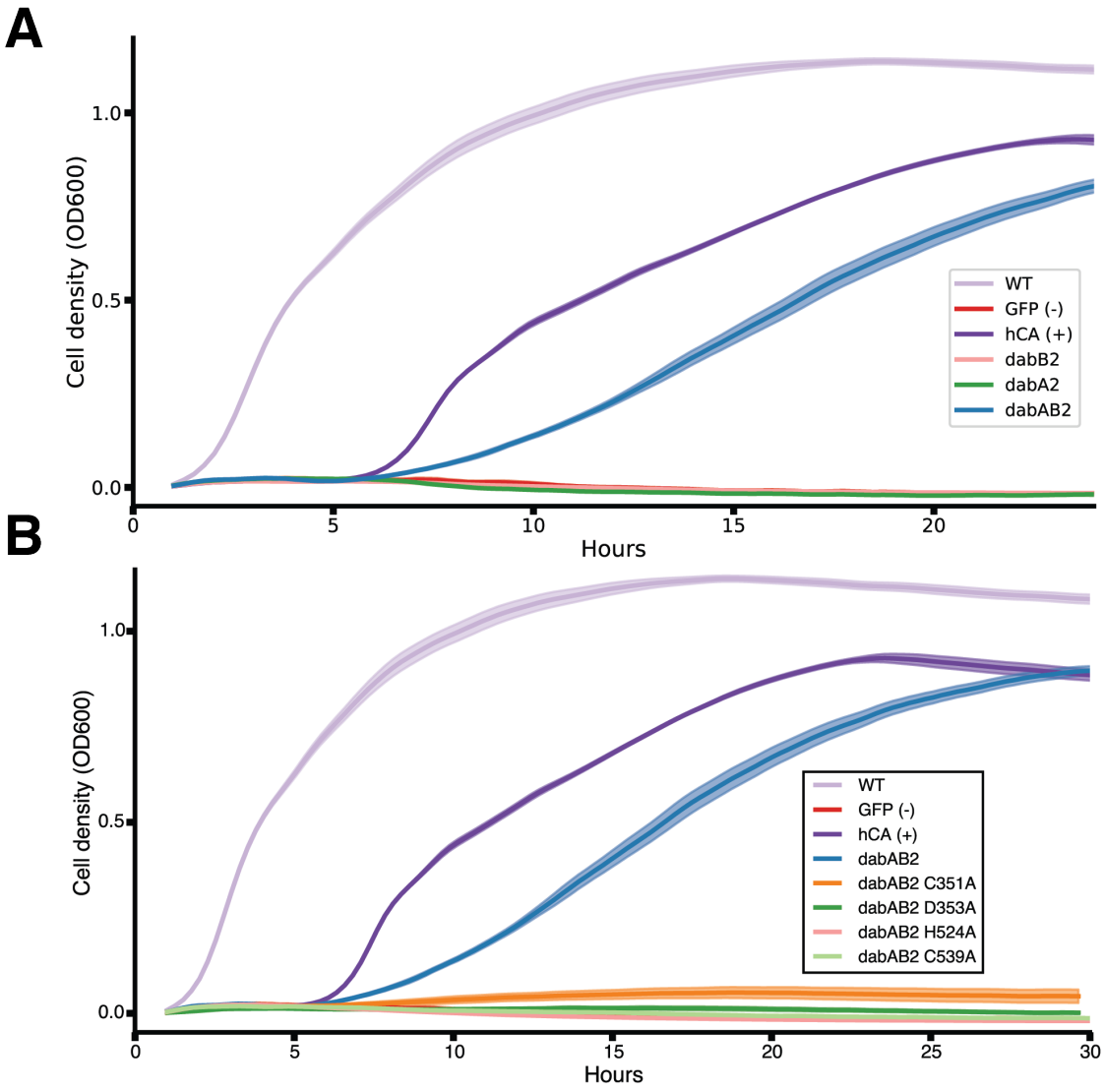


Figure D.5: Expression of DabAB2 rescues growth of CAfree *E. coli* in ambient CO_2. (A) Growth curves used to generate the growth yield values in Figure 4.4. Mean OD600 is plotted ± standard error for four replicate cultures. Wild-type *E. coli* (BW25113) and CAfree strains expressing either dabAB2 or human carbonic anhydrase II (hCA) grow in ambient CO_2 while CAfree expressing GFP, dabB2 alone, or dabA2 alone fail to grow. (B) Growth curves used to generate the growth yield values in Figure 4.5. Mean OD600 is plotted ± standard error of four replicate cultures. Wild type cells and CAfree expressing either DabAB2 or human carbonic anhydrase II (hCA) grow robustly. CAfree cells expressing putative active site mutants of DabAB2 (C351, D353, H524, or C539) grow as poorly as the negative control – CAfree expressing superfolder GFP in the same plasmid backbone.

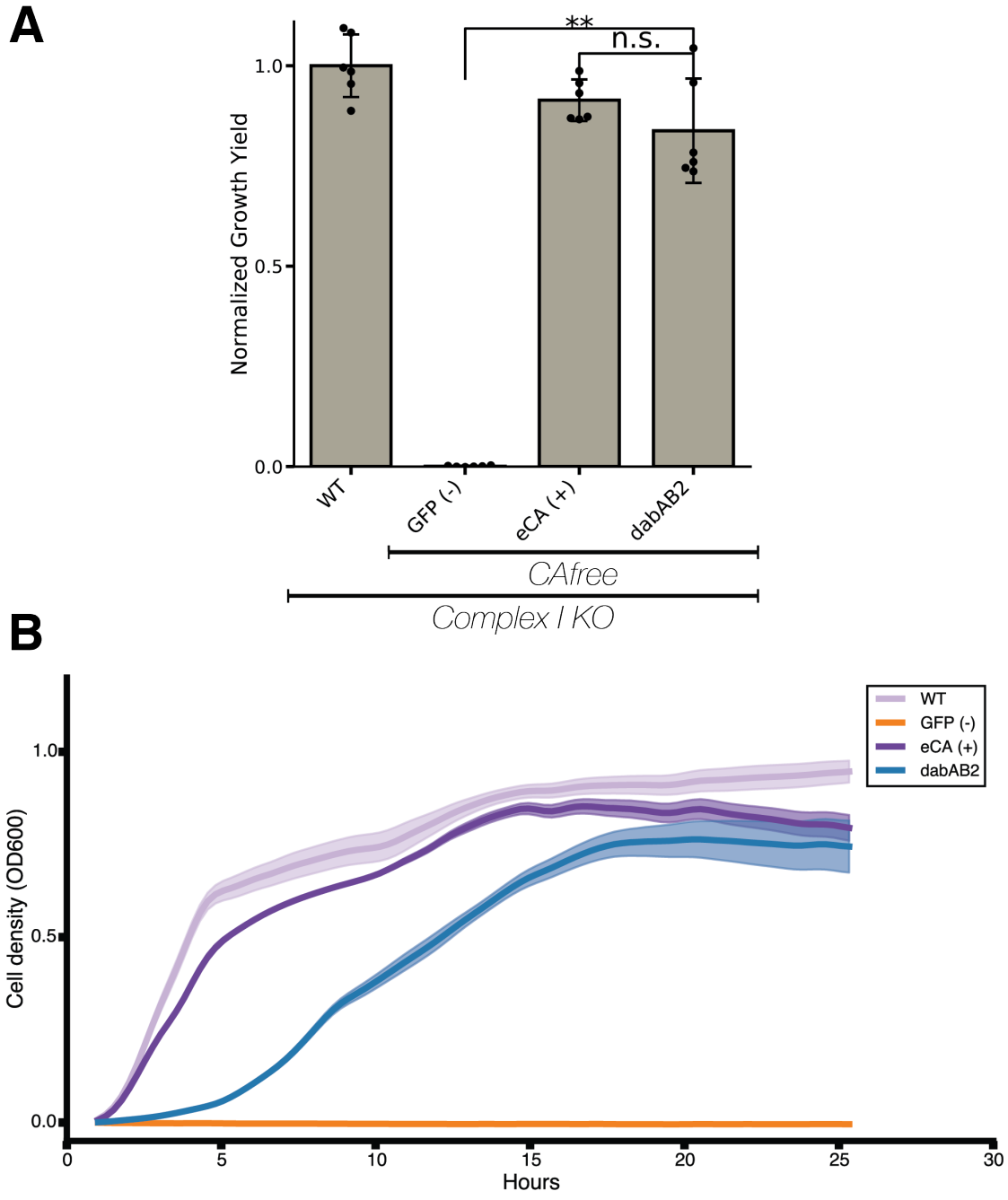


Figure D.6: DabAB2 functions indepedently of respiratory complex I. (A) DabAB2 expression rescues growth of CAfree cells in ambient CO_2 even in the absence of complex I ($\Delta nuoA\text{-}N$). A vector control expressing GFP does not rescue at all. Error bars represent standard deviation of six replicate cultures. "n.s." denotes means do not differ significantly, "*" denotes that means differ with Bonferroni corrected $p < 0.05$ by a two-tailed t-test, and "**" denotes $p < 5 \times 10^{-4}$. (B) Growth curves were used to generate the growth yield values in panel (A). Mean OD600 is plotted $\pm$ standard error of six replicate cultures. All strains are complex I knockout strains. DAB2 rescues growth of CAfree cells in the absence of complex I.

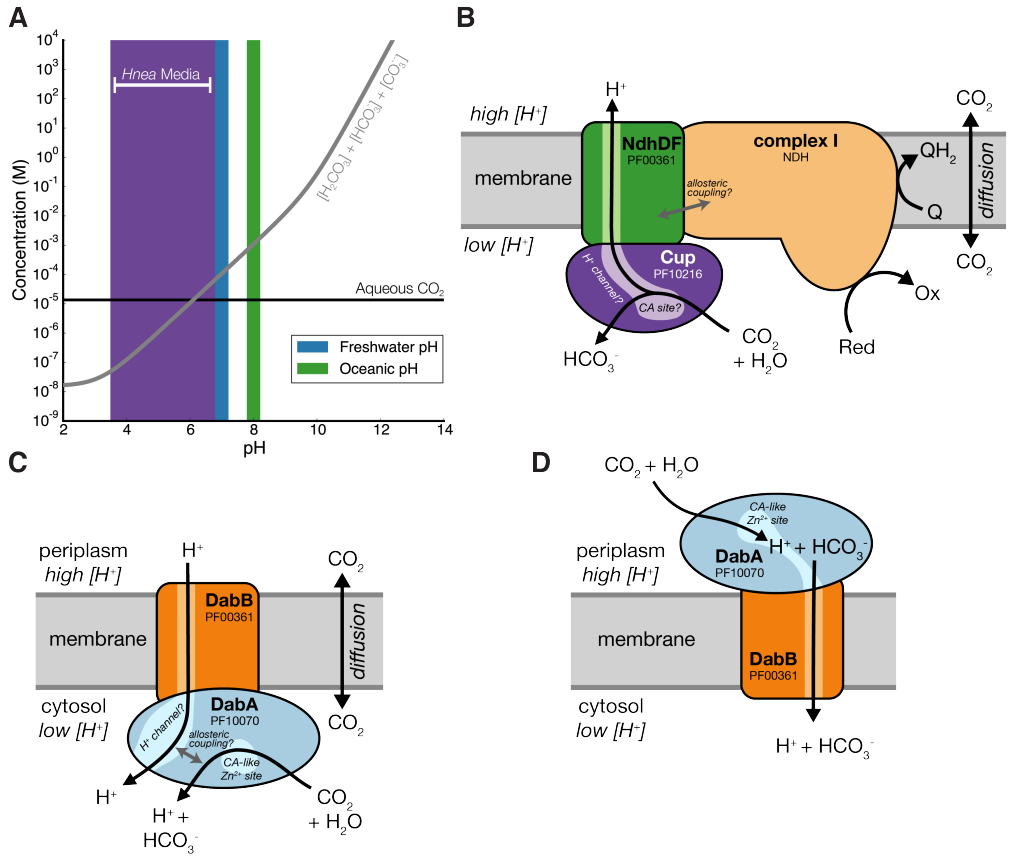


Figure D.7: Plausible biochemical models of vectorial CA activity. (A) Equilibrium concentrations of dissolved inorganic carbon as a function of pH. We assume the growth medium is in Henry's law equilibrium with present-day atmosphere (400 ppm CO_2) at 25 °C giving a soluble CO_2 concentration of $\approx$ 15 μM. Equilibrium concentrations of hydrated C_i species (H_2CO_3, HCO_3^-, CO_3^{2-}) are determined by the pH. In a oceanic pH $\approx$ 8, HCO_3^- dominates the C_i pool. HCO_3^- is also the dominant constituent of the C_i pool in freshwater, but less so (by a factor of $\approx$10 since freshwater and oceanic environments differ by $\approx$ 1 pH unit). In acid conditions (pH $<$ 6.1) CO_2 is the dominant constituent of the C_i pool. The pH of *H. neapolitanus* culture media ranges from 6.8 (when freshly made) to $\approx$3.5 at stationary phase as *H. neapolitanus* make H_2SO_4 as a product of their sulfur oxidizing metabolism. As such we expect that *H. neapolitanus* experiences environments in which it is more advantageous to pump CO_2 than HCO_3^-. (B) CupA/B proteins are CA-like subunits of a class of cyanobacterial C_i uptake systems. Cup-type systems are believed to couple electron transfer to vectorial CA activity and, potentially, outward-directed proton pumping. This model is based on the observation that Cup systems displace the two distal H^+-pumping subunits of the cyanobacterial complex I. (C) As our data are consistent with DabAB2 functioning as a standalone complex - DabAB2 do not appear to bind or require the *E. coli* complex I - we propose a model where energy for unidirectional hydration of CO_2 is drawn dissolution of a cation gradient. (D) In an alternative model, DabA is localized to the periplasm and DabB functions as a H^+:HCO_3^- symporter. Here vectorial CA activity is established by removal of the product, HCO_3^-. The process is driven by the proton-motive force. This model is not preferred because no secretion signals were identified in the DabA sequence. Moreover, the *A. ferrooxidans* genome contains an apparent DabA:DabB fusion protein where the predicted architecture of the fusion would place DabA in the cytoplasm.

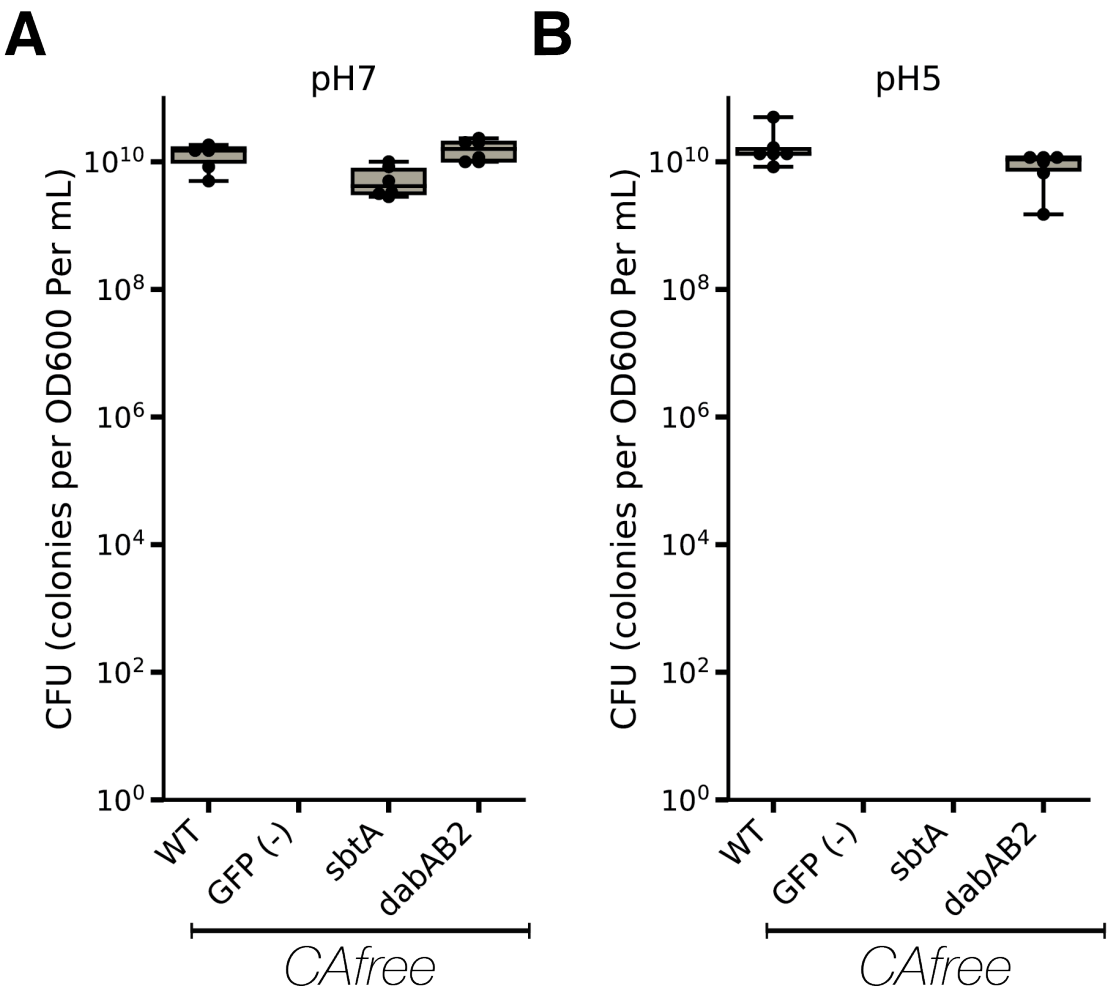


Figure D.8: pH independence of DabAB2 rescue of CAfree suggests that CO_2 is the true substrate. Colony forming units per OD600 per ml were measured by plating tenfold dilutions on LB plates with aTc induction in ambient air on media at pH 7 (A) and 5 (B). DabAB2 expression rescues growth at pH 7 and pH 5, while the cyanobacterial transporter SbtA (a known bicarbonate transporter) only rescues growth at pH 7. Whiskers represent the range of the data, box represents the interquartile range, and the middle line represents the median. Data is from 6 replicate platings of all conditions.

Bibliography

[1] Basel H Abuaita and Jeffrey H Withey. "Bicarbonate Induces Vibrio cholerae virulence gene expression by enhancing ToxT activity". en. In: *Infect. Immun.* 77.9 (Sept. 2009), pp. 4111–4120.

[2] Katrin Adamczyk et al. "Real-time observation of carbonic acid formation in aqueous solution". In: *Science* 326.5960 (2009), pp. 1690–1694.

[3] Noam Agmon. "The Grotthuss mechanism". In: *Chem. Phys. Lett.* 244.5 (Oct. 1995), pp. 456–462.

[4] Jaime Aguilera et al. "Carbonic anhydrase (Nce103p): an essential biosynthetic enzyme for growth of Saccharomyces cerevisiae at atmospheric carbon dioxide pressure". In: *Biochem. J.* 391 (2005), pp. 311–316.

[5] H Aigner et al. "Plant RuBisCo assembly in E. coli with five chloroplast chaperones including BSD2". en. In: *Science* 358.6368 (Dec. 2017), pp. 1272–1278.

[6] Eyal Akiva et al. "Evolutionary and molecular foundations of multiple contemporary functions of the nitroreductase superfamily". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 114.45 (Nov. 2017), E9549–E9558.

[7] Robert A Alberty. *Thermodynamics of Biochemical Reactions*. 1st ed. Hoboken, New Jersey: John Wiley & Sons, 2003.

[8] John F Allen and William Martin. "Evolutionary biology: out of thin air". en. In: *Nature* 445.7128 (Feb. 2007), pp. 610–612.

[9] L E Anderson. "Chloroplast and cytoplasmic enzymes. II. Pea leaf triose phosphate isomerases". en. In: *Biochim. Biophys. Acta* 235.1 (Apr. 1971), pp. 237–244.

[10] Inger Andersson. "Catalysis and regulation in Rubisco". en. In: *J. Exp. Bot.* 59.7 (Apr. 2008), pp. 1555–1568.

[11] Inger Andersson and Anders Backlund. "Structure and function of Rubisco". en. In: *Plant Physiol. Biochem.* 46.3 (Mar. 2008), pp. 275–291.

[12] Inger Andersson et al. "Crystal structure of the active site of ribulose-bisphosphate carboxylase". In: *Nature* 337.6204 (Jan. 1989), pp. 229–234.

[13] P John Andralojc et al. "Increasing metabolic potential: C-fixation". en. In: *Essays Biochem.* 62.1 (Apr. 2018), pp. 109–118.

[14] T J Andrews, G H Lorimer, and N E Tolbert. "Ribulose diphosphate oxygenase. I. Synthesis of phosphoglycolate by fraction-1 protein of leaves". en. In: *Biochemistry* 12.1 (Jan. 1973), pp. 11–18.

[15] Niv Antonovsky et al. "Sugar Synthesis from CO2 in Escherichia coli". In: *Cell* 166.1 (2016), pp. 115–125.

[16] Seth D Axen, Onur Erbilgin, and Cheryl A Kerfeld. "A taxonomy of bacterial microcompartment loci constructed by a novel scoring method". en. In: *PLoS Comput. Biol.* 10.10 (Oct. 2014), e1003898.

[17] Tomoya Baba et al. "Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection". In: *Mol. Syst. Biol.* 2 (2006), p. 2006.0008.

[18] M R Badger. "Kinetic properties of ribulose 1,5-bisphosphate carboxylase/oxygenase from Anabaena variabilis". In: *Arch. Biochem. Biophys.* 201.1 (1980), pp. 247–254.

[19] Mr Badger, David Hanson, and Gd Price. "Evolution and diversity of CO2 concentrating mechanisms in cyanobacteria". In: *Funct. Plant Biol.* 29 (2002), pp. 161–173.

[20] Murray R Badger and G James Collatz. "Studies on the kinetic mechanism of ribulose-1, 5-bisphosphate carboxylase and oxygenase reactions, with particular reference to the effect of temperature on kinetic parameters". In: *Carnegie Institute of Washington Yearbook* 76 (1977), pp. 355–361.

[21] Graeme Bainbridge et al. "Engineering Rubisco to change its catalytic properties". In: *J. Exp. Bot.* 46.special_issue (Sept. 1995), pp. 1269–1276.

[22] T S Baker et al. "The structure of form I crystals of D-ribulose-1,5-diphosphate carboxylase". en. In: *J. Mol. Biol.* 91.4 (Feb. 1975), pp. 391–399.

[23] J Timothy Ball, Ian E Woodrow, and Joseph A Berry. "A Model Predicting Stomatal Conductance and its Contribution to the Control of Photosynthesis under Different Environmental Conditions". In: *Progress in Photosynthesis Research: Volume 4 Proceedings of the VIIth International Congress on Photosynthesis Providence, Rhode Island, USA, August 10–15, 1986*. Ed. by J Biggins. Dordrecht: Springer Netherlands, 1987, pp. 221–224.

[24] Arren Bar-Even. "Daring metabolic designs for enhanced plant carbon fixation". In: *Plant Sci.* (Dec. 2017).

[25] Arren Bar-Even et al. "Hydrophobicity and charge shape cellular metabolite concentrations". en. In: *PLoS Comput. Biol.* 7.10 (Oct. 2011). Ed. by Jennifer L Reed, e1002166.

[26] Arren Bar-Even et al. "The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters". en. In: *Biochemistry* 50.21 (May 2011), pp. 4402–4410.

[27] Yinon M Bar-On and Ron Milo. "The global mass and average rate of rubisco". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 116.10 (Feb. 2019), pp. 4738–4743.

[28] Yinon M Bar-On, Rob Phillips, and Ron Milo. "The biomass distribution on Earth". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 115.25 (June 2018), pp. 6506–6511.

[29] Markus Basan et al. "Overflow metabolism in bacteria results from efficient proteome allocation for energy biogenesis". In: *Nature* 1 (2015).

[30] James A Bassham. "Mapping the carbon reduction cycle: a personal retrospective". en. In: *Photosynth. Res.* 76.1-3 (2003), pp. 35–52.

[31] Camille Bathellier et al. "Rubisco is not really so bad". en. In: *Plant Cell Environ.* 41.4 (Apr. 2018), pp. 705–716.

[32] Natalia Battchikova, Marion Eisenhut, and Eva Mari Aro. "Cyanobacterial NDH-1 complexes: Novel insights and remaining puzzles". In: *Biochimica et Biophysica Acta - Bioenergetics* 1807.8 (2011), pp. 935–944.

[33] Hermann Bauwe, Martin Hagemann, and Alisdair R Fernie. "Photorespiration: players, partners and origin". In: *Trends Plant Sci.* 15.6 (2010), pp. 330–336.

[34] "LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine". In: *Current Protocols in Bioinformatics*. Ed. by Andreas D Baxevanis et al. Vol. 9. Hoboken, NJ, USA: John Wiley & Sons, Inc., Aug. 2002, p. 277.

[35] S Belkin, R J Mehlhorn, and L Packer. "Proton gradients in intact cyanobacteria". en. In: *Plant Physiol.* 84 (1987), pp. 25–30.

[36] Bryson D Bennett et al. "Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli". en. In: *Nat. Chem. Biol.* 5.8 (June 2009), pp. 593–599.

[37] Andrew A Benson. "Following the path of carbon in photosynthesis: a personal story". en. In: *Photosynth. Res.* 73.1-3 (2002), pp. 29–49.

[38] J A Berry. "Studies of Mechanisms Affecting the Fractionation of Carbon Isotopes in Photosynthesis". In: *Stable Isotopes in Ecological Research*. Springer New York, 1989, pp. 82–94.

[39] Ralph Bertram et al. "Phenotypes of combined tet repressor mutants for effector and operator recognition and allostery". en. In: *J. Mol. Microbiol. Biotechnol.* 8.2 (2004), pp. 104–110.

[40] Javaid Y Bhat et al. "Mechanism of Enzyme Repair by the AAA+ Chaperone Rubisco Activase". en. In: *Mol. Cell* 67.5 (Sept. 2017), 744–756.e6.

[41] Maria Billini, Kostas Stamatakis, and Vicky Sophianopoulou. "Two members of a network of putative Na+/H+ antiporters are involved in salt and pH tolerance of the freshwater cyanobacterium Synechococcus elongatus". In: *J. Bacteriol.* 190.19 (2008), pp. 6318–6329.

[42] Clanton C Black and C Barry Osmond. "Crassulacean acid metabolism photosynthesis: 'working the night shift'". en. In: *Photosynth. Res.* 76.1-3 (2003), pp. 329–341.

[43] Gregory Bokinsky et al. "HipA-triggered growth arrest and β-lactam tolerance in Escherichia coli are mediated by RelA-dependent ppGpp synthesis". en. In: *J. Bacteriol.* 195.14 (July 2013), pp. 3173–3182.

[44] Walter Bonacci et al. "Modularity of a carbon-fixing protein organelle". In: *Proc. Natl. Acad. Sci. U. S. A.* 109.2 (Jan. 2012), pp. 478–483.

[45] G Bowes, W L Ogren, and R H Hageman. "Phosphoglycolate production catalyzed by ribulose diphosphate carboxylase". en. In: *Biochem. Biophys. Res. Commun.* 45.3 (Nov. 1971), pp. 716–722.

[46] George Bowes, William L Ogren, and Richard H Hageman. "pH Dependence of the Km(CO2) of Ribulose 1,5-Diphosphate Carboxylase". In: *Plant Physiology* 56 (1975), pp. 630–633.

[47] Ryan A Boyd et al. "Temperature response of Rubisco kinetics in Arabidopsis thaliana: thermal breakpoints and implications for reaction mechanisms". en. In: *J. Exp. Bot.* 70.1 (Jan. 2019), pp. 231–242.

[48] Andreas Bracher et al. "Biogenesis and metabolic maintenance of Rubisco". en. In: *Annu. Rev. Plant Biol.* 68 (Apr. 2017), pp. 29–60.

[49] F J Braun and P Hegemann. "Direct measurement of cytosolic calcium and pH in living Chlamydomonas reinhardtii cells". en. In: *Eur. J. Cell Biol.* 78.3 (Mar. 1999), pp. 199–208.

[50] Florian A Busch, Rowan F Sage, and Graham D Farquhar. "Plants increase CO2 uptake by assimilating nitrogen via the photorespiratory pathway". en. In: *Nat Plants* 4.1 (Jan. 2018), pp. 46–54.

[51] S V Caemmerer and J R Evans. "Determination of the Average Partial Pressure of CO2 in Chloroplasts From Leaves of Several C3 Plants". en. In: *Funct. Plant Biol.* 18.3 (1991), pp. 287–305.

[52] Susanne von Caemmerer, W Paul Quick, and Robert T Furbank. "The Development of C4 Rice: Current Progress and Future Challenges". en. In: *Science* 336.6089 (June 2012), pp. 1671–1672.

[53] Fei Cai et al. "The structure of CcmP, a tandem bacterial microcompartment domain protein from the β-carboxysome, forms a subcompartment within a microcompartment". en. In: *J. Biol. Chem.* 288.22 (May 2013), pp. 16055–16063.

[54] Zhen Cai et al. "Development of an activity-directed selection system enabled significant improvement of the carboxylation efficiency of Rubisco". In: *Protein Cell* (May 2014), pp. 12–18.

[55] Melvin Calvin. "The path of carbon in photosynthesis". en. In: *Science* 135.3507 (Mar. 1962), pp. 879–889.

[56] Jeffrey C Cameron et al. "Biogenesis of a bacterial organelle: the carboxysome assembly pathway". In: *Cell* 155.5 (Nov. 2013), pp. 1131–1140.

[57] A Elizabete Carmo-Silva et al. "Rubisco activities, properties, and regulation in three different C4 grasses under drought". en. In: *J. Exp. Bot.* 61.9 (May 2010), pp. 2355–2366.

[58] Thawatchai Chaijarasphong et al. "Programmed Ribosomal Frameshifting Mediates Expression of the α-Carboxysome". en. In: *J. Mol. Biol.* 428.1 (Jan. 2016), pp. 153–164.

[59] P H Chan et al. "Crystalline fraction I protein: preparation in large yield". en. In: *Science* 176.4039 (June 1972), pp. 1145–1146.

[60] M S Chapman et al. "Sliding-layer conformational change limited by the quaternary structure of plant RuBisCO". en. In: *Nature* 329.6137 (1987), pp. 354–356.

[61] M S Chapman et al. "Tertiary structure of plant RuBisCO: domains and their contacts". en. In: *Science* 241.4861 (July 1988), pp. 71–74.

[62] Anna H Chen et al. "The bacterial carbon-fixing organelle is formed by shell envelopment of preassembled cargo". In: *PLoS One* 8.9 (Jan. 2013), e76127.

[63] Jason W Chin. "Expanding and reprogramming the genetic code of cells and animals". en. In: *Annu. Rev. Biochem.* 83 (Feb. 2014), pp. 379–408.

[64] Raymond Chollet and William L Ogren. "Regulation of photorespiration in C3 and C4 species". In: *Bot. Rev.* 41.2 (1975), pp. 137–179.

[65] Chiranjit Chowdhury et al. "Selective molecular transport through the protein shell of a bacterial microcompartment organelle". In: *Proceedings of the National Academy of Sciences* (2015), p. 201423672.

[66] W Wallace Cleland et al. "Mechanism of Rubisco: The Carbamate as General Base". In: *Chem. Rev.* 98.2 (1998), pp. 549–562.

[67] W Cockburn. "Relationships between Stomatal Behavior and Internal Carbon Dioxide Concentration in Crassulacean Acid Metabolism Plants". en. In: *Plant Physiol.* 63.6 (June 1979), pp. 1029–1032.

[68] G J Collatz, M Ribas-Carbo, and J A Berry. "Coupled Photosynthesis-Stomatal Conductance Model for Leaves of C4 Plants". en. In: *Funct. Plant Biol.* 19.5 (1992), pp. 519–538.

[69] J D Cronk et al. "Crystal structure of E. coli beta-carbonic anhydrase, an enzyme with an unusual pH-dependent activity". In: *Protein Sci.* 10.5 (May 2001), pp. 911–922.

[70] Jeff D Cronk et al. "Identification of a novel noncatalytic bicarbonate binding site in eubacterial beta-carbonic anhydrase". en. In: *Biochemistry* 45.14 (Apr. 2006), pp. 4351–4361.

[71] Peter L Cummins, Babu Kannappan, and Jill E Gready. "Directions for Optimization of Photosynthetic Carbon Fixation: RuBisCO's Efficiency May Not Be So Constrained After All". en. In: *Front. Plant Sci.* 9 (Mar. 2018), p. 183.

[72] K A Datsenko and B L Wanner. "One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 97.12 (June 2000), pp. 6640–6645.

[73] Dan Davidi et al. "A Bird's-Eye View of Enzyme Evolution: Chemical, Physicochemical, and Physiological Considerations". en. In: *Chem. Rev.* (Aug. 2018).

[74] Dan Davidi et al. "Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113.12 (Mar. 2016), pp. 3401–3406.

[75] Daniel E Deatherage and Jeffrey E Barrick. "Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq". en. In: *Methods Mol. Biol.* 1151 (2014), pp. 165–188.

[76] Paramvir S Dehal et al. "MicrobesOnline: an integrated portal for comparative and functional genomics". In: *Nucleic Acids Res.* 38.Database issue (Jan. 2010), pp. D396–400.

[77] Robert J DiMario et al. "The many types of carbonic anhydrases in photosynthetic organisms". en. In: *Plant Sci.* 268 (Mar. 2018), pp. 11–17.

[78] Kimberly P Dobrinski, Dana L Longo, and Kathleen M Scott. "The carbon-concentrating mechanism of the hydrothermal vent chemolithoautotroph Thiomicrospira crunogena". en. In: *J. Bacteriol.* 187.16 (Aug. 2005), pp. 5761–5766.

[79] Jie Dong et al. "Methyl ketone production by Pseudomonas putida is enhanced by plant-derived amino acids". en. In: *Biotechnol. Bioeng.* (Apr. 2019).

[80] Robert W Dorner, Albert Kahn, and Sam G Wildman. "The proteins of green leaves. VII. Synthesis and decay of the cytoplasmic proteins during the life of the tobacco leaf". en. In: *J. Biol. Chem.* 229.2 (Dec. 1957), pp. 945–952.

[81] Zhicheng Dou et al. "CO2 fixation kinetics of Halothiobacillus neapolitanus mutant carboxysomes lacking carbonic anhydrase suggest the shell acts as a diffusional barrier for CO2". In: *J. Biol. Chem.* 283.16 (Apr. 2008), pp. 10377–10384.

[82] Jiahui Du et al. "Characterisation of Cyanobacterial Bicarbonate Transporters in E. coli Shows that SbtA Homologs Are Functional in This Heterologous Expression System". In: *PLoS One* 9.12 (Jan. 2014), e115905.

[83] Paulo Durão et al. "Opposing effects of folding and assembly chaperones on evolvability of Rubisco". en. In: *Nat. Chem. Biol.* 11.2 (Feb. 2015), pp. 148–155.

[84] Marion Eisenhut et al. "The photorespiratory glycolate metabolism is essential for cyanobacteria and might have been conveyed endosymbiontically to plants". In: *Proc. Natl. Acad. Sci. U. S. A.* 105.44 (Dec. 2008), pp. 17199–17204.

[85] I V Elanskaya et al. "Functional analysis of the Na+/H+ antiporter encoding genes of the cyanobacterium Synechocystis PCC 6803". In: *Biochemistry* 67.4 (2002), pp. 432–440.

[86] R John Ellis. "Discovery of molecular chaperones". en. In: *Cell Stress Chaperones* 1.3 (Sept. 1996), pp. 155–160.

[87] Benjamin D Engel et al. "Native architecture of the Chlamydomonas chloroplast revealed by in situ cryo-electron tomography". In: *Elife* 4 (Jan. 2015), pp. 1–29.

[88] Tobias J Erb and Jan Zarzycki. "A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme". en. In: *Curr. Opin. Biotechnol.* 49 (Feb. 2018), pp. 100–107.

[89] Tobias J Erb et al. "Carboxylation mechanism and stereochemistry of crotonyl-CoA carboxylase/reductase, a carboxylating enoyl-thioester reductase". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 106.22 (June 2009), pp. 8871–8876.

[90] G Eshel et al. "Land, irrigation water, greenhouse gas, and reactive nitrogen burdens of meat, eggs, and dairy production in the United States". In: *Proceedings of the National Academy of Sciences* (July 2014).

[91] Csaba Éva, Mária Oszvald, and László Tamás. "Current and possible approaches for improving photosynthetic efficiency". In: *Plant Sci.* (Nov. 2018).

[92] J R Evans et al. "Carbon Isotope Discrimination measured Concurrently with Gas Exchange to Investigate CO2 Diffusion in Leaves of Higher Plants". en. In: *Funct. Plant Biol.* 13.2 (1986), pp. 281–292.

[93] G Falkner and F Horner. "pH Changes in the Cytoplasm of the Blue-Green Alga Anacystis nidulans Caused by Light-dependent Proton Flux into the Thylakoid Space". en. In: *Plant Physiol.* 58.6 (Dec. 1976), pp. 717–718.

[94] Paul G Falkowski and John A Raven. *Aquatic photosynthesis*. Princeton University Press, 2013.

[95] G D Farquhar, S von Caemmerer, and J A Berry. "A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species". en. In: *Planta* 149.1 (June 1980), pp. 78–90.

[96] G D Farquhar, J R Ehleringer, and K T Hubick. "Carbon Isotope Discrimination and Photosynthesis". In: *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 40.1 (June 1989), pp. 503–537.

[97] G D Farquhar and T D Sharkey. "Stomatal Conductance and Photosynthesis". In: *Annu. Rev. Plant Physiol.* 33.1 (June 1982), pp. 317–345.

[98] G D Farquhar, von Caemmerer S, and J A Berry. "Models of photosynthesis". en. In: *Plant Physiol.* 125.1 (Jan. 2001), pp. 42–45.

[99] C B Field et al. "Primary production of the biosphere: integrating terrestrial and oceanic components". en. In: *Science* 281.5374 (July 1998), pp. 237–240.

[100] Christopher B Field. "Plant Physiology of the "Missing" Carbon Sink". en. In: *Plant Physiol.* 125.1 (Jan. 2001), pp. 25–28.

[101] Woodward W Fischer, James Hemp, and Jena E Johnson. "Evolution of Oxygenic Photosynthesis". In: *Annu. Rev. Earth Planet. Sci.* 44.1 (2016), pp. 647–683.

[102] Avi Flamholz et al. "eQuilibrator–the biochemical thermodynamics calculator". en. In: *Nucleic Acids Res.* 40.Database issue (Jan. 2012), pp. D770–5.

[103] Avi Flamholz et al. "Glycolytic strategy as a tradeoff between energy yield and protein cost". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 110.24 (June 2013), pp. 10039–10044.

[104] Pedro Flombaum et al. "Present and future global distributions of the marine Cyanobacteria Prochlrococcus and Synechococcus". In: *Proc. Natl. Acad. Sci. U. S. A.* 110.24 (2013), pp. 9824–9829.

[105] Douglas M Fowler and Stanley Fields. "Deep mutational scanning: a new style of protein science". en. In: *Nat. Methods* 11.8 (Aug. 2014), pp. 801–807.

[106] Elizabeth S Freeman Rosenzweig et al. "The Eukaryotic CO2-Concentrating Organelle Is Liquid-like and Exhibits Dynamic Reorganization". In: *Cell* 171.1 (Sept. 2017), 148–162.e19.

[107] L Fridlyand, A Kaplan, and L Reinhold. "Quantitative evaluation of the role of a putative CO2-scavenging entity in the cyanobacterial CO2-concentrating mechanism". In: *Biosystems.* 37.3 (Jan. 1996), pp. 229–238.

[108] Jeroni Galmes et al. "Temperature responses of the Rubisco maximum carboxylase activity across domains of life: phylogenetic signals, trade-offs, and importance for carbon gain". In: *Photosynth. Res.* 123.2 (2015), pp. 183–201.

[109] Jeroni Galmés et al. "Environmentally driven evolution of Rubisco and improved photosynthesis and growth within the C3 genus Limonium (Plumbaginaceae)". en. In: *New Phytol.* 203.3 (Aug. 2014), pp. 989–999.

[110] Jeroni Galmés et al. "Expanding knowledge of the Rubisco kinetics variability in plant species: environmental and evolutionary trends". en. In: *Plant Cell Environ.* 37.9 (Sept. 2014), pp. 1989–2001.

[111] E Gantt and S F Conti. "Ultrastructure of blue-green algae". en. In: *J. Bacteriol.* 97.3 (Mar. 1969), pp. 1486–1493.

[112] Christopher W Gee and Krishna K Niyogi. "The carbonic anhydrase CAH1 is an essential component of the carbon-concentrating mechanism in Nannochloropsis oceanica". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 114.17 (Apr. 2017), pp. 4537–4542.

[113] Mario Giordano, John Beardall, and John A Raven. "CO2 concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution". en. In: *Annu. Rev. Plant Biol.* 56 (2005), pp. 99–131.

[114] Jeff E Glasgow et al. "Influence of Electrostatics on Small Molecule Flux through a Protein Nanoreactor". en. In: *ACS Synth. Biol.* 4.9 (Sept. 2015), pp. 1011–1019.

[115] Benjamin Goldstein et al. "Potential to curb the environmental burdens of American beef consumption using a novel plant-based beef substitute". en. In: *PLoS One* 12.12 (Dec. 2017), e0189029.

[116] J Gutknecht, M A Bisson, and F C Tosteson. "Diffusion of carbon dioxide through lipid bilayer membranes: effects of carbonic anhydrase, bicarbonate, and unstirred layers". en. In: *J. Gen. Physiol.* 69.6 (June 1977), pp. 779–794.

[117] Martin Hagemann et al. "Evolution of photorespiration from cyanobacteria to land plants, considering protein phylogenies and acquisition of carbon concentrating mechanisms". In: *J. Exp. Bot.* 67.10 (2016), erw063–.

[118] Marshall D Hatch. "C4 photosynthesis: discovery and resolution". In: *Discoveries in Photosynthesis*. Ed. by Govindjee et al. Dordrecht: Springer Netherlands, 2005, pp. 875–880.

[119] Manajit Hayer-Hartl, Andreas Bracher, and F Ulrich Hartl. "The GroEL–GroES Chaperonin Machine: A Nano-Cage for Protein Folding". In: *Trends Biochem. Sci.* 41.1 (Jan. 2016), pp. 62–76.

[120] John M Hayes. "Fractionation of Carbon and Hydrogen Isotopes in Biosynthetic Processes". In: *Rev. Mineral. Geochem.* 43.1 (Jan. 2001), pp. 225–277.

[121] John M Hayes, Harald Strauss, and Alan J Kaufman. "The abundance of 13C in marine organic matter and isotopic fractionation in the global biogeochemical cycle of carbon during the past 800 Ma". In: *Chem. Geol.* 161.1 (Sept. 1999), pp. 103–125.

[122] Sabine Heinhorst et al. "Characterization of the carboxysomal carbonic anhydrase CsoSCA from Halothiobacillus neapolitanus". In: *J. Bacteriol.* 188.23 (Dec. 2006), pp. 8087–8094.

[123] W H Heldt et al. "Alkalization of the chloroplast stroma caused by light-dependent proton flux into the thylakoid space". en. In: *Biochim. Biophys. Acta* 314.2 (Aug. 1973), pp. 224–241.

[124] Carmen Hermida-Carrera, Maxim V Kapralov, and Jeroni Galmés. "Rubisco catalytic properties and temperature response in crops". en. In: *Plant Physiol.* 171.4 (Jan. 2016), p. 01846.2016.

[125] Elad Herz et al. "The genetic basis for the adaptation of E. coli to sugar synthesis from CO 2". In: *Nat. Commun.* 8.1 (2017), p. 1705.

[126] Sean A Higgins, Sorel V Y Ouonkap, and David F Savage. "Rapid and Programmable Protein Mutagenesis Using Plasmid Recombineering". en. In: *ACS Synth. Biol.* 6.10 (Oct. 2017), pp. 1825–1833.

[127] Sean A Higgins and David F Savage. "Protein Science by DNA Sequencing: How Advances in Molecular Biology Are Accelerating Biochemistry". en. In: *Biochemistry* (Oct. 2017).

[128] Florian Hinzpeter, Ulrich Gerland, and Filipe Tostevin. "Optimal Compartmentalization Strategies for Metabolic Microcompartments". en. In: *Biophys. J.* 112.4 (Feb. 2017), pp. 767–779.

[129] Y A Holthuijzen et al. "Energetic aspects of CO2 uptake in Thiobacillus neapolitanus". In: *Arch. Microbiol.* 147.3 (Apr. 1987), pp. 285–290.

[130] A Holzenburg et al. "Structure of D-ribulose-l,5-bisphosphate carboxylase/oxygenase from Alcaligenes eutrophyus H16". en. In: *Nature* 325.6106 (1987), pp. 730–732.

[131] J J Hopfield. "Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 71.10 (Oct. 1974), pp. 4135–4139.

[132] B M Hopkinson et al. "The Minimal CO2-Concentrating Mechanism of Prochlorococcus spp. MED4 Is Effective and Efficient". In: *Plant Physiol.* 166.4 (2014), pp. 2205–2217.

[133] Brian M Hopkinson et al. "Efficiency of the CO2-concentrating mechanism of diatoms". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.10 (Mar. 2011), pp. 3830–3837.

[134] I E Huertas et al. "Light-dependent bicarbonate uptake and CO2 efflux in the marine microalga Nannochloropsis gaditana". en. In: *Planta* 211.1 (June 2000), pp. 43–49.

[135] Cristina V Iancu et al. "The structure of isolated Synechococcus strain WH8102 carboxysomes as revealed by electron cryotomography". In: *J. Mol. Biol.* 372.3 (Sept. 2007), pp. 764–773.

[136] Alexander L Jaffe et al. "Lateral gene transfer shapes the distribution of RuBisCO among Candidate Phyla Radiation bacteria and DPANN archaea". en. In: *Mol. Biol. Evol.* (Dec. 2018).

[137] Shengyang Jin et al. "Structural insights into the LCIB protein family reveals a new group of β-carbonic anhydrases". en. In: *Proc. Natl. Acad. Sci. U. S. A.* (Dec. 2016).

[138] Julien Jorda et al. "Using comparative genomics to uncover new kinds of protein-based metabolic organelles in bacteria". In: *Protein Sci.* 22.2 (2013), pp. 179–195.

[139] D B Jordan and W L Ogren. "The CO2/O2 specificity of ribulose 1,5-bisphosphate carboxylase/oxygenase: Dependence on ribulosebisphosphate concentration, pH and temperature". en. In: *Planta* 161.4 (June 1984), pp. 308–313.

[140] Douglas B Jordan and William L Ogren. "Species variation in the specificity of ribulose biphosphate carboxylase/oxygenase". In: *Nature* 291.5815 (June 1981), pp. 513–515.

[141] H J Kane et al. "An Improved Method for Measuring the CO2/O2 Specificity of Ribulosebisphosphate Carboxylase-Oxygenase". en. In: *Funct. Plant Biol.* 21.4 (1994), pp. 449–461.

[142] Aaron Kaplan and Leonora Reinhold. "CO2 Concentrating Mechanisms in Photosynthetic Microorganisms". en. In: *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50 (June 1999), pp. 539–570.

[143] Maxim V Kapralov and Dmitry A Filatov. "Widespread positive selection in the photosynthetic Rubisco enzyme". en. In: *BMC Evol. Biol.* 7 (May 2007), p. 73.

[144] N Kawashima and S G Wildman. "Studies on fraction I protein. II. Comparison of physical, chemical, immunological and enzymatic properties between spinach and tobacco fraction I proteins". en. In: *Biochim. Biophys. Acta* 229.3 (Mar. 1971), pp. 749–760.

[145] N Kawashima and S G Wildman. "Studies on fraction-I protein. I. Effect of crystallization of fraction-I protein from tobacco leaves on ribulose diphosphate carboxylase activity". en. In: *Biochim. Biophys. Acta* 229.1 (Jan. 1971), pp. 240–249.

[146] Nobumaro Kawashima, Shiu-Yuen Kwok, and S G Wildman. "Studies on fraction-I protein III. Comparison of the primary structure of the large and small subunits obtained from five species of nicotiana". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 236.3 (June 1971), pp. 578–586.

[147] Rashad Kebeish et al. "Chloroplastic photorespiratory bypass increases photosynthesis and biomass production in Arabidopsis thaliana". In: *Nat. Biotechnol.* 25.5 (2007), pp. 593–599.

[148] Lawrence A Kelley et al. "The Phyre2 web portal for protein modeling, prediction and analysis". en. In: *Nat. Protoc.* 10.6 (June 2015), pp. 845–858.

[149] G J Kelly and E Latzko. "Inhibition of spinach-leaf phosphofructokinase by 2-phosphoglycollate". en. In: *FEBS Lett.* 68.1 (Sept. 1976), pp. 55–58.

[150] Cheryl A Kerfeld and Matthew R Melnicki. "Assembly, function and evolution of cyanobacterial carboxysomes". en. In: *Curr. Opin. Plant Biol.* 31 (June 2016), pp. 66–75.

[151] Michael J Kerner et al. "Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli". In: *Cell* 122.2 (July 2005), pp. 209–220.

[152] Raja G Khalifah. "The Carbon Dioxide Hydration Activity of Carbonic Anhydrase". In: *J. Biol. Chem.* 246.8 (1971), pp. 2561–2573.

[153] Sae Kikutani et al. "Thylakoid luminal ϑ-carbonic anhydrase critical for growth and photosynthesis in the marine diatom Phaeodactylum tricornutum". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113.35 (Aug. 2016), pp. 9828–9833.

[154] James N Kinney, Seth D Axen, and Cheryl a Kerfeld. "Comparative analysis of carboxysome shell proteins". In: *Photosynth. Res.* 109.1-3 (Sept. 2011), pp. 21–32.

[155] Helmut Kirchhoff et al. "Protein diffusion and macromolecular crowding in thylakoid membranes". en. In: *Plant Physiol.* 146.4 (Apr. 2008), pp. 1571–1578.

[156] Michael G Klein et al. "Identification and Structural Analysis of a Novel Carboxysome Shell Protein with Implications for Metabolite Transport". In: *J. Mol. Biol.* 392.2 (2009), pp. 319–333.

[157] Vijay M Krishnamurthy et al. "Carbonic anhydrase as a model for biophysical and physical-organic studies of proteins and protein- ligand binding". In: *Chem. Rev.* 108.3 (2008), pp. 946–1051.

[158] Johannes Kromdijk et al. "Improving photosynthesis and crop productivity by accelerating recovery from photoprotection". en. In: *Science* 354.6314 (Nov. 2016), pp. 857–861.

[159] Terry A Krulwich, David B Hicks, and Masahiro Ito. "Cation/proton antiporter complements of bacteria: why so large and diverse?" en. In: *Mol. Microbiol.* 74.2 (Oct. 2009), pp. 257–260.

[160] Armin Kubis and Arren Bar-Even. "Synthetic biology approaches for improving photosynthesis". en. In: *J. Exp. Bot.* (Jan. 2019).

[161] Chiara Lee et al. "A two-domain elevator mechanism for sodium/proton antiport". In: *Nature* 501.7468 (2013), pp. 573–577.

[162] Xiaobo Li et al. "A genome-wide algal mutant library and functional screen identifies genes required for eukaryotic photosynthesis". en. In: *Nat. Genet.* (Mar. 2019).

[163] Wolfram Liebermeister et al. "Visual account of protein investment in cellular functions". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 111.23 (June 2014), pp. 8488–8493.

[164] Hsiangchun Lin et al. "Targeted Knockdown of GDCH in Rice Leads to a Photorespiratory-Deficient Phenotype Useful as a Building Block for C4 Rice". en. In: *Plant Cell Physiol.* 57.5 (May 2016), pp. 919–932.

[165] Myat T Lin et al. "A faster Rubisco with potential to increase photosynthesis in crops". en. In: *Nature* 513.7519 (Sept. 2014), pp. 547–550.

[166] Chong Liu et al. "Water splitting-biosynthetic system with CO2 reduction efficiencies exceeding photosynthesis". en. In: *Science* 352.6290 (June 2016), pp. 1210–1213.

[167] Di Liu, Ramaswamy Chettiyan Seetharaman Ramya, and Oliver Mueller-Cajar. "Surveying the expanding prokaryotic Rubisco multiverse". en. In: *FEMS Microbiol. Lett.* 364.16 (Sept. 2017).

[168] Hongbin Liu, Hector A Nolla, and Lisa Campbell. "Prochlorococcus growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean". In: *Aquat. Microb. Ecol.* 12.1 (1997), pp. 39–47.

[169] Hongbin Liu et al. "Prochlorococcus growth rates in the central equatorial Pacific : An application of the fmax approach". In: *J. Geophys. Res.* 104.1 (1999), pp. 3391–3399.

[170] David B Lobell, Wolfram Schlenker, and Justin Costa-Roberts. "Climate trends and global crop production since 1980". en. In: *Science* 333.6042 (July 2011), pp. 616–620.

[171] Benedict M Long et al. "Carboxysome encapsulation of the CO2-fixing enzyme Rubisco in tobacco chloroplasts". en. In: *Nat. Commun.* 9.1 (Sept. 2018), p. 3570.

[172] Stephen P Long, Amy Marshall-Colon, and Xin-Guang Zhu. "Meeting the global food demand of the future by engineering crop photosynthesis and yield potential". en. In: *Cell* 161.1 (Mar. 2015), pp. 56–66.

[173] George H Lorimer, Murray R Badger, and T John Andrews. "The activation of ribulose-1,5-bisphosphate carboxylase by carbon dioxide and magnesium ions. Equilibria, kinetics, a suggested mechanism, and physiological implications". In: *Biochemistry* 15.3 (Feb. 1976), pp. 529–536.

[174] George H Lorimer, Yuh Ru Chen, and Fred C Hartman. "A role for the. epsilon-amino group of lysine-334 of ribulose-1, 5-bisphosphate carboxylase in the addition of carbon dioxide to the 2, 3-enediol(ate) of ribulose 1, 5-bisphosphate". In: *Biochemistry* 32.35 (1993), pp. 9018–9024.

[175] Marjorie R Lundgren, Colin P Osborne, and Pascal-Antoine Christin. "Deconstructing Kranz anatomy to understand C4 evolution". en. In: *J. Exp. Bot.* 65.13 (July 2014), pp. 3357–3369.

[176] R Lutz and H Bujard. "Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements". In: *Nucleic Acids Res.* 25.6 (Mar. 1997), pp. 1203–1210.

[177] Fangfang Ma et al. "Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation". In: *Proceedings of the National Academy of Sciences* 111.47 (2014), pp. 16967–16972.

[178] Luke C M Mackinder et al. "A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113.21 (May 2016), pp. 5958–5963.

[179] Luke C M Mackinder et al. "A Spatial Interactome Reveals the Protein Organization of the Algal CO2-Concentrating Mechanism". In: *Cell* 171.1 (Sept. 2017), 133–147.e14.

[180] Shin-Ichi Maeda, Murray R Badger, and G Dean Price. "Novel gene products associated with NdhD3/D4-containing NDH-1 complexes are involved in photosynthetic CO2 hydration in the cyanobacterium, Synechococcus sp. PCC7942". In: *Mol. Microbiol.* 43.2 (Mar. 2002), pp. 425–435.

[181] E Magid and B O Turbeck. "The rates of the spontaneous hydration of CO2 and the reciprocal reaction in neutral aqueous solutions between 0 degrees and 38 degrees". In: *Biochim. Biophys. Acta* 165.3 (1968), pp. 515–524.

[182] Niall Mangan and Michael Brenner. "Systems analysis of the CO2 concentrating mechanism in cyanobacteria". In: *Elife* 2014.3 (Apr. 2014), e02043.

[183] Mary Mangiapia et al. "Proteomic and mutant analysis of the CO2 concentrating mechanism of hydrothermal vent chemolithoautotroph Thiomicrospira crunogena". en. In: *J. Bacteriol.* (Jan. 2017).

[184] Yehouda Marcus et al. "High CO2 Requiring Mutant of Anacystis nidulans R2". In: *Plant Physiol.* 82.2 (Oct. 1986), pp. 610–612.

[185] William Martin et al. "Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus". In: *Proc. Natl. Acad. Sci. U. S. A.* 99.19 (Sept. 2002), pp. 12246–12251.

[186] Keith A Martinez 2nd et al. "Cytoplasmic pH response to acid stress in individual cells of Escherichia coli and Bacillus subtilis observed by fluorescence ratio imaging microscopy". en. In: *Appl. Environ. Microbiol.* 78.10 (May 2012), pp. 3706–3714.

[187] Justin M McGrath and Stephen P Long. "Can the cyanobacterial carbon-concentrating mechanism increase photosynthesis in crop species? A theoretical analysis". In: *Plant Physiol.* 164.4 (Apr. 2014), pp. 2247–2261.

[188] Dennis B McNevin et al. "Differences in carbon isotope discrimination of three variants of D-ribulose-1,5-bisphosphate carboxylase/oxygenase reflect differences in their catalytic mechanisms". en. In: *J. Biol. Chem.* 282.49 (Dec. 2007), pp. 36068–36076.

[189] Dennis B McNevin et al. "Measurement of (carbon) kinetic isotope effect by Rayleigh fractionation using membrane inlet mass spectrometry for CO2-consuming reactions". en. In: *Funct. Plant Biol.* 33.12 (Dec. 2006), pp. 1115–1128.

[190] Balaraj B Menon et al. "The carboxysome shell is permeable to protons". In: *J. Bacteriol.* 192.22 (Nov. 2010), pp. 5881–5886.

[191] Christophe Merlin and Millicent Masters. "Why is carbonic anhydrase essential to Escherichia coli?" In: *J. Bacteriol.* 185.21 (2003).

[192] Moritz T Meyer, Alistair J McCormick, and Howard Griffiths. "Will an algal CO2-concentrating mechanism work in higher plants?" en. In: *Curr. Opin. Plant Biol.* 31.Ccm (June 2016), pp. 181–188.

[193] Frank J Millero. *Thermodynamics of the carbon dioxide system in the oceans.* 1995.

[194] Ron Milo and Rob Phillips. *Cell Biology by the Numbers.* en. Garland Science, Dec. 2015.

[195] Yoshiko Miyagawa, Masahiro Tamoi, and Shigeru Shigeoka. "Overexpression of a cyanobacterial fructose-1,6-/sedoheptulose-1,7-bisphosphatase in tobacco enhances photosynthesis and growth". en. In: *Nat. Biotechnol.* 19.October (Oct. 2001), pp. 965–969.

[196] Douwe Molenaar et al. "Shifts in growth strategies reflect tradeoffs in cellular economics". In: *Mol. Syst. Biol.* 5.323 (Jan. 2009), p. 323.

[197] M K Morell et al. "Mutations of an active site threonyl residue promote beta elimination and other side reactions of the enediol intermediate of the ribulosebisphosphate carboxylase reaction". In: *J. Biol. Chem.* 269.11 (Mar. 1994), pp. 8091–8098.

[198] Oliver Mueller-Cajar. "The Diverse AAA+ Machines that Repair Inhibited Rubisco Active Sites". en. In: *Front Mol Biosci* 4 (May 2017), p. 31.

[199] Oliver Mueller-Cajar, Matthew Morell, and Spencer M Whitney. "Directed evolution of rubisco in Escherichia coli reveals a specificity-determining hydrogen bond in the form II enzyme". en. In: *Biochemistry* 46.49 (Dec. 2007), pp. 14067–14074.

[200] Oliver Mueller-Cajar and Spencer M Whitney. "Evolving improved Synechococcus Rubisco functional expression in Escherichia coli". en. In: *Biochem. J* 414.2 (Sept. 2008), pp. 205–214.

[201] R G Murray and S W Watson. "Structure of Nitrosocystic Oceanus and Comparison with Nitrosomonas and Nitrobacter". en. In: *J. Bacteriol.* 89 (June 1965), pp. 1594–1609.

[202] Dana C Nadler et al. "Rapid construction of metabolite biosensors using domain-insertion profiling". en. In: *Nat. Commun.* 7 (July 2016), p. 12266.

[203] Aparna Nagarajan and Himadri B Pakrasi. "Membrane-Bound Protein Complexes for Photosynthesis and Respiration in Cyanobacteria". In: *Elsevier Oceanogr. Ser.* (2001), pp. 1–8.

[204] H Nakagawa et al. "Preliminary crystallographic study of a ribulose-1,5-bisphosphate carboxylase-oxygenase from Chromatium vinosum". en. In: *J. Mol. Biol.* 191.3 (Oct. 1986), pp. 577–578.

[205] National Research Council (U.S.). Committee on Photobiology. *Photosynthetic Mechanisms of Green Plants*. en. National Academies, 1963.

[206] Zachary E R Newby et al. "A general protocol for the crystallization of membrane proteins for X-ray structural investigation". en. In: *Nat. Protoc.* 4.5 (2009), pp. 619–637.

[207] J Ninio. "Kinetic amplification of enzyme discrimination". en. In: *Biochimie* 57.5 (1975), pp. 587–595.

[208] M Nishimura and T Akazawa. "Further proof for the catalytic role of the larger subunit in the spinach leaf ribulose-1,5-diphosphate carboxylase". en. In: *Biochem. Biophys. Res. Commun.* 54.3 (Oct. 1973), pp. 842–848.

[209] Takashi Nishimura et al. "Mechanism of low CO2-induced activation of the cmp bicarbonate transporter operon by a LysR family protein in the cyanobacterium Synechococcus elongatus strain PCC 7942". In: *Mol. Microbiol.* 68.1 (Apr. 2008), pp. 98–109.

[210] Greta Nölke et al. "The integration of algal carbon concentration mechanism components into tobacco chloroplasts increases photosynthetic efficiency and biomass". en. In: *Biotechnol. J.* (June 2018), e1800170.

[211] Elad Noor et al. "An integrated open framework for thermodynamics of reactions that combines accuracy and coverage". en. In: *Bioinformatics* 28.15 (Aug. 2012), pp. 2037–2044.

[212] Alessandro Occhialini et al. "Transgenic tobacco plants with improved cyanobacterial Rubisco expression but no extra assembly factors grow at near wild-type rates if provided with elevated CO2". en. In: *Plant J.* 85.1 (Jan. 2016), pp. 148–160.

[213] William L Ogren. "Photorespiration: Pathways, Regulation, and Modification". In: *Annu. Rev. Plant Physiol.* (1984).

[214] Douglas J Orr et al. "Surveying Rubisco Diversity and Temperature Response to Improve Crop Photosynthetic Efficiency". en. In: *Plant Physiol.* 172.2 (Oct. 2016), pp. 707–717.

[215] Donald R Ort et al. "Redesigning photosynthesis to sustainably meet global food and bioenergy demand". en. In: *Proceedings of the National Academy of Sciences* 112.28 (July 2015), pp. 8529–8536.

[216] Peter Orth et al. "Structural basis of gene regulation by the tetracycline inducible Tet repressor–operator system". In: *Nat. Struct. Biol.* 7.3 (Mar. 2000), pp. 215–219.

[217] Jörg Jorg Overmann and Ferran Garcia-Pichel. "The phototrophic way of life". In: *The Prokaryotes*. Ed. by Eugene Rosenberg et al. Berlin, Heidelberg: Springer, 2013, pp. 203–257.

[218] Junyoung O Park et al. "Metabolite concentrations, fluxes and free energies imply efficient enzyme usage". In: *Nat. Chem. Biol.* May (2016).

[219] M A J Parry, A J Keys, and S Gutteridge. "Variation in the Specificity Factor of C3 Higher Plant Rubiscos Determined by the Total Consumption of Ribulose-P2". In: *J. Exp. Bot.* 40.3 (Mar. 1989), pp. 317–320.

[220] Martin A J Parry et al. "Rubisco regulation: a role for inhibitors". en. In: *J. Exp. Bot.* 59.7 (Apr. 2008), pp. 1569–1580.

[221] M T Pellicer et al. "glc locus of Escherichia coli: characterization of genes encoding the subunits of glycolate oxidase and the glc regulator protein". en. In: *J. Bacteriol.* 178.7 (Apr. 1996), pp. 2051–2059.

[222] Maria Teresa Pellicer et al. "Role of 2-Phosphoglycolate Phosphatase of Escherichia coli in Metabolism of the 2-Phosphoglycolate Formed in DNA Repair". In: *J. Bacteriol.* 185.19 (Oct. 2003), pp. 5815–5821.

[223] Kerry L Peña et al. "Structural basis of the oxidative activation of the carboxysomal gamma-carbonic anhydrase, CcmM". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 107.6 (Jan. 2010), pp. 2455–2460.

[224] G A Peschek et al. "Transmembrane Proton Electrochemical Gradients in Dark Aerobic and Anaerobic Cells of the Cyanobacterium (Blue-Green Alga) Anacystis nidulans: Evidence for Respiratory Energy Transduction in the Plasma Membrane". en. In: *Plant Physiol.* 79.1 (Sept. 1985), pp. 278–284.

[225] Thea R Pick et al. "PLGG1, a plastidic glycolate glycerate transporter, is required for photorespiration and defines a unique class of metabolite transporters". In: *Proc. Natl. Acad. Sci. U. S. A.* 110.8 (2013), pp. 3185–3190.

[226] John Pierce, George H Lorimer, and Gade S Reddy. "Kinetic mechanism of ribulosebisphosphate carboxylase: evidence for an ordered, sequential reaction". In: *Biochemistry* 25.7 (Apr. 1986), pp. 1636–1644.

[227] J Poore and T Nemecek. "Reducing food's environmental impacts through producers and consumers". en. In: *Science* 360.6392 (June 2018), pp. 987–992.

[228] G D Price and M R Badger. "Expression of Human Carbonic Anhydrase in the Cyanobacterium Synechococcus PCC7942 Creates a High CO2-Requiring Phenotype Evidence for a Central Role for Carboxysomes in the CO2 Concentrating Mechanism". In: *Plant Physiol.* 91.2 (1989), pp. 505–513.

[229] G D Price and M R Badger. "Isolation and characterization of high CO2-requiring-mutants of the cyanobacterium Synechococcus PCC7942: two phenotypes that accumulate inorganic carbon but are apparently unable to generate CO2 within the carboxysome". In: *Plant Physiol.* 91.2 (1989), pp. 514–525.

[230] G Dean Price. "Inorganic carbon transporters of the cyanobacterial CO2 concentrating mechanism". In: *Photosynth. Res.* 109 (2011), pp. 47–57.

[231] G Dean Price, Megan C Shelden, and Susan M Howitt. "Membrane topology of the cyanobacterial bicarbonate transporter, SbtA, and identification of potential regulatory loops". In: *Mol. Membr. Biol.* 28.5 (Aug. 2011), pp. 265–275.

[232] G Dean Price et al. "Advances in understanding the cyanobacterial CO2-concentrating-mechanism (CCM): functional components, Ci transporters, diversity, genetic regulation and prospects for engineering into plants". In: *J. Exp. Bot.* 59.7 (Jan. 2008), pp. 1441–1461.

[233] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. "FastTree: computing large minimum evolution trees with profiles instead of a distance matrix". en. In: *Mol. Biol. Evol.* 26.7 (July 2009), pp. 1641–1650.

[234] Joseph Priestley. "XIX. Observations on different kinds of air". In: *Philosophical transactions of the royal society of London* 62 (1772), pp. 147–264.

[235] Anneke Prins et al. "Rubisco catalytic properties of wild and domesticated relatives provide scope for improving wheat photosynthesis". en. In: *J. Exp. Bot.* 67.6 (Mar. 2016), pp. 1827–1838.

[236] Tinus Pulles. "Engineering photosynthesis: a necessary tool to protect the world's climate?" In: *Carbon Management* 8.2 (Mar. 2017), pp. 167–173.

[237] Benjamin D Rae et al. "Cyanobacterial carboxysomes: microcompartments that facilitate CO2 fixation". en. In: *J. Mol. Microbiol. Biotechnol.* 23.4-5 (Aug. 2013), pp. 300–307.

[238] Benjamin D Rae et al. "Functions, compositions, and evolution of the two types of carboxysomes: polyhedral microcompartments that facilitate CO2 fixation in cyanobacteria and some proteobacteria". In: *Microbiol. Mol. Biol. Rev.* 77.3 (2013), pp. 357–379.

[239] Benjamin D Rae et al. "Progress and challenges of engineering a biophysical carbon dioxide-concentrating mechanism into higher plants". en. In: *J. Exp. Bot.* 68.14 (June 2017), pp. 3717–3737.

[240] John A Raven. "Contributions of anoxygenic and oxygenic phototrophy and chemolithotrophy to carbon and oxygen fluxes in aquatic environments". In: *Aquat. Microb. Ecol.* 56.September (Sept. 2009), pp. 177–192.

[241] John a Raven and John F Allen. "Genomics and chloroplast evolution: what did cyanobacteria do for plants?" In: *Genome Biol.* 4.3 (Jan. 2003), p. 209.

[242] John a Raven, John Beardall, and Mario Giordano. "Energy costs of carbon dioxide concentrating mechanisms in aquatic organisms". In: *Photosynth. Res.* 121.2-3 (Sept. 2014), pp. 111–124.

[243] John A Raven, John Beardall, and Patricia Sánchez-Baracaldo. "The possible evolution and future of CO2-concentrating mechanisms". en. In: *J. Exp. Bot.* 68.14 (June 2017), pp. 3701–3716.

[244] Deepak K Ray et al. "Recent patterns of crop yield growth and stagnation". en. In: *Nat. Commun.* 3 (2012), p. 1293.

[245] B A Read and F R Tabita. "A hybrid ribulosebisphosphate carboxylase/oxygenase enzyme exhibiting a substantial increase in substrate specificity factor". en. In: *Biochemistry* 31.24 (June 1992), pp. 5553–5560.

[246] B A Read and F R Tabita. "High substrate specificity factor ribulose bisphosphate carboxylase/oxygenase from eukaryotic marine algae and properties of recombinant cyanobacterial RubiSCO containing "algal" residue modifications". en. In: *Arch. Biochem. Biophys.* 312.1 (July 1994), pp. 210–218.

[247] Betsy A Read and F Robert Tabita. "Amino acid substitutions in the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase that influence catalytic activity of the holoenzyme". In: *Biochemistry* 31.2 (Jan. 1992), pp. 519–525.

[248] Leonora Reinhold, Ronnie Kosloff, and Aaron Kaplan. "A model for inorganic carbon fluxes and photosynthesis in cyanobacterial carboxysomes". In: *Can. J. Bot.* 69.5 (May 1991), pp. 984–988.

[249] Leonora Reinhold, Menekhem Zviman, and Aaron Kaplan. "Inorganic carbon fluxes and photosynthesis in cyanobacteria—a quantitative model". In: *Progress in photosynthesis research*. Springer, 1987, pp. 289–296.

[250] Evan W Roberts et al. "Isolation and characterization of the Prochlorococcus carboxysome reveal the presence of the novel shell protein CsoS1D". In: *J. Bacteriol.* 194.4 (2012), pp. 787–795.

[251] Rutherford Ness Robertson. *Lively Membranes*. CUP Archive, 1983.

[252] C A Roeske and Marion H O'Leary. "Carbon isotope effects on enzyme-catalyzed carboxylation of ribulose bisphosphate". In: *Biochemistry* 23.25 (Dec. 1984), pp. 6275–6284.

[253] Cecile S Rousseaux and Watson W Gregg. "Interannual variation in phytoplankton primary production at a global scale". In: *Remote Sensing* 6.1 (2013), pp. 1–19.

[254] Ambrish Roy, Alper Kucukural, and Yang Zhang. "I-TASSER: a unified platform for automated protein structure and function prediction". en. In: *Nat. Protoc.* 5.4 (Apr. 2010), pp. 725–738.

[255] Benjamin E Rubin et al. "The essential gene set of a photosynthetic organism". In: *Proceedings of the National Academy of Sciences* (2015), p. 201519220.

[256] Rowan F Sage, Tammy L Sage, and Ferit Kocacinar. "Photorespiration and the evolution of C4 photosynthesis". In: *Annu. Rev. Plant Biol.* 63 (2012), pp. 19–47.

[257] Rowan F Sage et al. "Some like it hot: the physiological ecology of C4 plant evolution". en. In: *Oecologia* 187.4 (Aug. 2018), pp. 941–966.

[258] R Sander. "Compilation of Henry's law constants (version 4.0) for water as solvent". In: *Atmos. Chem. Phys.* 15.8 (2015), pp. 4399–4981.

[259] Sriram Satagopan et al. "Structure-function studies with the unique hexameric form II ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) from Rhodopseudomonas palustris". en. In: *J. Biol. Chem.* 289.31 (Aug. 2014), pp. 21433–21450.

[260] David F Savage et al. "Spatially ordered dynamics of the bacterial carbon fixation machinery". In: *Science* 327.5970 (Mar. 2010), pp. 1258–1261.

[261] Yonatan Savir and Tsvi Tlusty. "Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition". en. In: *PLoS One* 2.5 (May 2007), e468.

[262] Yonatan Savir et al. "Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 107.8 (Feb. 2010), pp. 3475–3480.

[263] Michael F Schmid et al. "Structure of Halothiobacillus neapolitanus Carboxysomes by Cryo-electron Tomography". In: *J. Mol. Biol.* 364 (2006), pp. 526–535.

[264] G Schneider et al. "Three-dimensional structure of ribulose-1,5-bisphosphate carboxylase/oxygenase from Rhodospirillum rubrum at 2.9 A resolution". en. In: *EMBO J.* 5.13 (Dec. 1986), pp. 3409–3415.

[265] I Schomburg et al. "The BRENDA enzyme information system-From a database to an expert system". en. In: *J. Biotechnol.* 261 (Nov. 2017), pp. 194–206.

[266] Thomas Schwander et al. "A synthetic pathway for the fixation of carbon dioxide in vitro". en. In: *Science* 354.6314 (Nov. 2016), pp. 900–904.

[267] Kathleen M Scott et al. "Diversity in CO2-Concentrating Mechanisms among Chemolithoautotrophs from the Genera Hydrogenovibrio, Thiomicrorhabdus, and Thiomicrospira, Ubiquitous in Sulfidic Habitats Worldwide". en. In: *Appl. Environ. Microbiol.* 85.3 (Feb. 2019).

[268] Kathleen M Scott et al. "Genomes of ubiquitous marine and hypersaline Hydrogenovibrio, Thiomicrorhabdus and Thiomicrospira spp. encode a diversity of mechanisms to sustain chemolithoautotrophy in heterogeneous environments". en. In: *Environ. Microbiol.* 20.8 (Aug. 2018), pp. 2686–2708.

[269] Matthew Scott et al. "Interdependence of cell growth and gene expression: origins and consequences". In: *Science* 330.6007 (Nov. 2010), pp. 1099–1102.

[270] J C Servaites. "pH Dependence of Photosynthesis and Photorespiration in Soybean Leaf Cells". en. In: *Plant Physiol.* 60.5 (Nov. 1977), pp. 693–696.

[271] Thomas D Sharkey. "Estimating the Rate of Photorespiration in Leaves". In: *Physiol. Plant.* 73 (1988), pp. 146–152.

[272] Robert E Sharwood et al. "Temperature responses of Rubisco from Paniceae grasses provide opportunities for improving C3 photosynthesis". en. In: *Nat Plants* 2 (Nov. 2016), p. 16186.

[273] Alon Shepon et al. "The opportunity cost of animal based diets exceeds all food losses". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 115.15 (Apr. 2018), pp. 3804–3809.

[274] D M Sherman, T a Troyan, and L a Sherman. "Localization of Membrane Proteins in the Cyanobacterium Synechococcus sp. PCC7942". In: *Plant Physiol.* 106.1 (1994), pp. 251–262.

[275] Mari Shibata et al. "Distinct constitutive and low-CO2-induced CO2 uptake systems in cyanobacteria: genes involved and their phylogenetic relationship with homologous genes in other organisms". In: *Proc. Natl. Acad. Sci. U. S. A.* 98.20 (2001), pp. 11789–11794.

[276] Mari Shibata et al. "Two CO2 uptake systems in cyanobacteria: four systems for inorganics carbon acquisition in Synechocystis sp. strain PCC 6803". en. In: *Funct. Plant Biol.* 29.3 (2002), pp. 123–129.

[277] Patrick M Shih et al. "Biochemical characterization of predicted Precambrian RuBisCO". en. In: *Nat. Commun.* 7 (Jan. 2016), p. 10382.

[278] Patrick M Shih et al. "Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 110.3 (Jan. 2013), pp. 1053–1058.

[279] Patrick M Shih et al. "Introduction of a synthetic CO2-fixing photorespiratory bypass into a cyanobacterium". In: *J. Biol. Chem.* 289.14 (Apr. 2014), pp. 9493–9500.

[280] J M Shively et al. "Functional organelles in prokaryotes: polyhedral inclusions (carboxysomes) of Thiobacillus neapolitanus". In: *Science* 182.4112 (Nov. 1973), pp. 584–586.

[281] O Shoval et al. "Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space". en. In: *Science* 336.6085 (June 2012), pp. 1157–1160.

[282] Fabian Sievers and Desmond G Higgins. "Clustal Omega for making accurate alignments of many protein sequences". en. In: *Protein Sci.* 27.1 (Jan. 2018), pp. 135–145.

[283] Jonathan J Silberg, Jeffrey B Endelman, and Frances H Arnold. "SCHEMA-guided protein recombination". en. In: *Methods Enzymol.* 388 (2004), pp. 35–42.

[284] Andrew J Simkin et al. "Simultaneous stimulation of sedoheptulose 1,7-bisphosphatase, fructose 1,6-bisphophate aldolase and the photorespiratory glycine decarboxylase-H protein increases CO2 assimilation, vegetative biomass and seed yield in Arabidopsis". en. In: *Plant Biotechnol. J.* 15.7 (July 2017), pp. 805–816.

[285] J C Sirard, M Mock, and A Fouet. "The three Bacillus anthracis toxin genes are coordinately regulated by bicarbonate and temperature". en. In: *J. Bacteriol.* 176.16 (Aug. 1994), pp. 5188–5192.

[286] Lukasz Slabinski et al. "XtalPred: a web server for prediction of protein crystallizability". en. In: *Bioinformatics* 23.24 (Dec. 2007), pp. 3403–3405.

[287] K S Smith et al. "Carbonic anhydrase is an ancient enzyme widespread in prokaryotes". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 96.26 (Dec. 1999), pp. 15184–15189.

[288] Chris R Somerville. "An early Arabidopsis demonstration resolving a few issues concerning photorespiration". In: *Plant Physiol.* 127.1 (2001), p. 3.

[289] Paul F South et al. "Bile Acid Sodium Symporter BASS6 Can Transport Glycolate and Is Involved in Photorespiratory Metabolism in Arabidopsis thaliana". en. In: *Plant Cell* 29.4 (Apr. 2017), pp. 808–823.

[290] Paul F South et al. "Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field". en. In: *Science* 363.6422 (Jan. 2019), eaat9077.

[291] Robert J Spreitzer and Michael E Salvucci. "Rubisco: structure, regulatory interactions, and possibilities for a better enzyme". In: *Annu. Rev. Plant Biol.* 53 (Jan. 2002), pp. 449–475.

[292] Robert J Spreitzer et al. "Alanine-Scanning Mutagenesis of the Small-Subunit βA-βB Loop of Chloroplast Ribulose-1, 5-Bisphosphate Carboxylase/Oxygenase: Substitution at Arg-71 Affects Thermal Stability and CO2/O2 Specificity". In: *Biochemistry* 40.19 (2001), pp. 5615–5621.

[293] G A Sprenger et al. "Anaerobic growth of Escherichia coli on glycerol by importing genes of the dha regulon from Klebsiella pneumoniae". en. In: *J. Gen. Microbiol.* 135.5 (May 1989), pp. 1255–1262.

[294] Daniel A Stolper, Niels Peter Revsbech, and Donald E Canfield. "Aerobic growth at nanomolar oxygen concentrations". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 107.44 (Nov. 2010), pp. 18755–18760.

[295] D Sültemeyer et al. "Characterisation of carbon dioxide and bicarbonate transport during steady-state photosynthesis in the marine cyanobacterium Synechococcus strain PCC7002". In: *Planta* 61 (1995), pp. 597–607.

[296] Claudiu T Supuran. "Structure and function of carbonic anhydrases". en. In: *Biochem. J* 473.14 (July 2016), pp. 2023–2032.

[297] Claudiu T Supuran and Clemente Capasso. "An Overview of the Bacterial Carbonic Anhydrases". en. In: *Metabolites* 7.4 (Nov. 2017).

[298] F Robert Tabita et al. "Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms". en. In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363.1504 (Aug. 2008), pp. 2629–2640.

[299] Emad Tajkhorshid et al. "Control of the selectivity of the aquaporin water channel family by global orientational tuning". en. In: *Science* 296.5567 (Apr. 2002), pp. 525–530.

[300] Guillaume Tcherkez. "The mechanism of Rubisco-catalysed oxygenation". In: *Plant Cell Environ.* 39.5 (2016), pp. 983–997.

[301] Guillaume G B Tcherkez, Graham D Farquhar, and T John Andrews. "Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 103.19 (May 2006), pp. 7246–7251.

[302] Guillaume G Tcherkez et al. "Commentary: Directions for Optimization of Photosynthetic Carbon Fixation: RuBisCO's Efficiency May Not Be So Constrained After All". en. In: *Front. Plant Sci.* 9 (June 2018), p. 929.

[303] Stefan Timm et al. "Glycine decarboxylase controls photosynthesis and plant growth". en. In: *FEBS Lett.* 586.20 (Oct. 2012), pp. 3692–3697.

[304] N E Tolbert. "The C2 Oxidative Photosynthetic Carbon Cycle". en. In: *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 48 (June 1997), pp. 1–25.

[305] Koichi Uemura et al. "A Rapid and Sensitive Method for Determination of Relative Specificity of RuBisCO from Various Species by Anion-Exchange Chromatography". In: *Plant Cell Physiol.* 37.3 (Apr. 1996), pp. 325–331.

[306] K Uemura et al. "Ribulose-1,5-bisphosphate carboxylase/oxygenase from thermophilic red algae with a strong specificity for CO2 fixation". en. In: *Biochem. Biophys. Res. Commun.* 233.2 (Apr. 1997), pp. 568–571.

[307] G Unden and J Bongaerts. "Alternative respiratory pathways of Escherichia coli: energetics and transcriptional regulation in response to electron acceptors". en. In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1320.3 (July 1997), pp. 217–234.

[308] Mari Valkonen et al. "Noninvasive high-throughput single-cell analysis of the intracellular pH of Saccharomyces cerevisiae by ratiometric flow cytometry". en. In: *Appl. Environ. Microbiol.* 79.23 (Dec. 2013), pp. 7179–7187.

[309] Hendrika S Van Walraven et al. "The H+/ATP coupling ratio of the ATP synthase from thiol-modulated chloroplasts and two cyanobacterial strains is four". In: *FEBS Lett.* 379.3 (1996), pp. 309–313.

[310] Arsenio Villarejo et al. "The induction of the CO2 concentrating mechanism in a starch-less mutant of Chlamydomonas reinhardtii". In: *Physiol. Plant.* 98 (1996), pp. 798–802.

[311] Juan Carlos Villarreal and Susanne S Renner. "Hornwort pyrenoids, carbon-concentrating structures, evolved and were lost at least five times during the last 100 million years". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 109.46 (Nov. 2012), pp. 18873–18878.

[312] Rungaroon Waditee et al. "Overexpression of a Na+/H+ antiporter confers salt tolerance on a freshwater cyanobacterium, making it capable of growth in sea water". In: *Proc. Natl. Acad. Sci. U. S. A.* 99.6 (2002), pp. 4109–4114.

[313] A Walter and J Gutknecht. "Permeability of small nonelectrolytes through lipid bilayer membranes". In: *J. Membr. Biol.* 90.3 (Jan. 1986), pp. 207–217.

[314] H Wang et al. "Rubisco condensate formation by CcmM in β-carboxysome biogenesis". In: *Nature* (Jan. 2019).

[315] A Weissbach, B L Horecker, and J Hurwitz. "The enzymatic formation of phosphoglyceric acid from ribulose diphosphate and carbon dioxide". en. In: *J. Biol. Chem.* 218.2 (Feb. 1956), pp. 795–810.

[316] K Werdan, H W Heldt, and M Milovancev. "The role of pH in the regulation of carbon fixation in the chloroplast stroma. Studies on CO2 fixation in the light and dark". en. In: *Biochim. Biophys. Acta* 396.2 (Aug. 1975), pp. 276–292.

[317] Mary Jane West-Eberhard, J Andrew C Smith, and Klaus Winter. "Photosynthesis, reorganized". en. In: *Science* 332.6027 (Apr. 2011), pp. 311–312.

[318] Kelly M Wetmore et al. "Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons". en. In: *MBio* 6.3 (May 2015), e00306–15.

[319] Nicole M Wheatley et al. "Structure and identification of a pterin dehydratase-like protein as a ribulose-bisphosphate carboxylase/oxygenase (RuBisCO) assembly factor in the α-carboxysome". en. In: *J. Biol. Chem.* 289.11 (Mar. 2014), pp. 7973–7981.

[320] Lynne Whitehead et al. "Comparing the in Vivo Function of α-Carboxysomes and β-Carboxysomes in Two Model Cyanobacteria". In: *Plant Physiol.* 165.1 (May 2014), pp. 398–411.

[321] S M Whitney et al. "Form I Rubiscos from non-green algae are expressed abundantly but not assembled in tobacco chloroplasts". en. In: *Plant J.* 26.5 (June 2001), pp. 535–547.

[322] Jerker Widengren, Bob Terry, and Rudolf Rigler. "Protonation kinetics of GFP and FITC investigated by FCS — aspects of the use of fluorescent indicators for measuring pH". In: *Chem. Phys.* 249.2-3 (1999), pp. 259–271.

[323] Sam G Wildman. "Along the trail from Fraction I protein to Rubisco (ribulose bisphosphate carboxylase-oxygenase)". en. In: *Photosynth. Res.* 73.1-3 (2002), pp. 243–250.

[324] Sam G Wildman and J Bonner. "The proteins of green leaves; isolation, enzymatic properties and auxin content of spinach cytoplasmic proteins". en. In: *Arch. Biochem.* 14.3 (Aug. 1947), pp. 381–413.

[325] C Willmer and M Fricker. *Stomata*. en. Springer Science & Business Media, 1996.

[326] Robert H Wilson et al. "An improved Escherichia coli screen for Rubisco identifies a protein-protein interface that can enhance CO2-fixation kinetics". en. In: *J. Biol. Chem.* (Oct. 2017).

[327] Fiona J Woodger, Murray R Badger, and G Dean Price. "Sensing of inorganic carbon limitation in Synechococcus PCC7942 is correlated with the size of the internal inorganic carbon pool and involves oxygen". en. In: *Plant Physiol.* 139.4 (Dec. 2005), pp. 1959–1969.

[328] Alex Wu et al. "Quantifying impacts of enhancing photosynthesis on crop yield". en. In: *Nat Plants* 5.4 (Apr. 2019), pp. 380–388.

[329] T-X Xiang and B D Anderson. "The relationship between permeant size and permeability in lipid bilayer membranes". In: *J. Membr. Biol.* 140.2 (June 1994), pp. 111–122.

[330] Todd O Yeates et al. "Protein-based organelles in bacteria: carboxysomes and related microcompartments". In: *Nat. Rev. Microbiol.* 6.9 (Sept. 2008), pp. 681–691.

[331] Oren Yishai et al. "The formate bio-economy". en. In: *Curr. Opin. Chem. Biol.* 35 (Dec. 2016), pp. 1–9.

[332] Jamey D Young et al. “Mapping photoautotrophic metabolism with isotopically nonstationary (13)C flux analysis”. In: *Metab. Eng.* 13.6 (Nov. 2011), pp. 656–665.

[333] Jodi N Young et al. “Large variation in the Rubisco kinetics of diatoms reveals diversity among their carbon-concentrating mechanisms”. In: *J. Exp. Bot.* 67.11 (May 2016), pp. 3445–3456.

[334] Rémi Zallot, Nils O Oberg, and John A Gerlt. “‘Democratized’ genomic enzymology web tools for functional assignment”. en. In: *Curr. Opin. Chem. Biol.* 47 (Dec. 2018), pp. 77–85.

[335] Tomáš Zavřel et al. “Quantitative insights into the cyanobacterial cell economy”. en. In: *Elife* 8 (Feb. 2019).

[336] I Zelitch. “Photorespiration: Studies with Whole Tissues”. In: *Photosynthesis II: Photosynthetic Carbon Metabolism and Related Processes*. Ed. by Martin Gibbs and Erwin Latzko. Berlin, Heidelberg: Springer Berlin Heidelberg, 1979, pp. 353–367.

[337] Xin-Guang Zhu, Stephen P Long, and Donald R Ort. “Improving photosynthetic efficiency for greater yield”. en. In: *Annu. Rev. Plant Biol.* 61 (2010), pp. 235–261.