A map of the rubisco biochemical landscape

Authors

Noam Prywes[1,2], Naiya R. Philips[3], Luke M. Oltrogge[2,3], Sebastian Lindner[4], Yi-Chin Candace Tsai[5], Benoit de Pins[6], Aidan E. Cowan[3,7], Leah J. Taylor-Kearney[8], Hana A. Chang[8], Laina N. Hall[9], Daniel Bellieny-Rabelo[1,10], Hunter M. Nisonoff[11], Rachel F. Weissman[3], Avi I. Flamholz[12], David Ding[1,2], Abhishek Y. Bhatt[3,13], Patrick M. Shih[1,8,14,15], Oliver Mueller-Cajar[5], Ron Milo[6], David F. Savage[1,2,3,*]

[1]Innovative Genomics Institute, University of California; Berkeley, California 94720, USA;
[2]Howard Hughes Medical Institute, University of California; Berkeley, California 94720, USA;
[3]Department of Molecular and Cell Biology, University of California; Berkeley, California 94720, USA;
[4]University of Heidelberg; 69047 Heidelberg, Germany
[5]School of Biological Sciences, Nanyang Technological University; Singapore 637551, Singapore
[6]Department of Plant and Environmental Sciences, Weizmann Institute of Science; Rehovot 76100, Israel
[7]Joint BioEnergy Institute, Lawrence Berkeley National Laboratory; Emeryville, CA 94608, USA
[8]Department of Plant and Microbial Biology, University of California, Berkeley; Berkeley, CA 94720, USA.
[9]Biophysics, University of California, Berkeley; Berkeley, CA 94720, USA.
[10]California Institute for Quantitative Biosciences (QB3), University of California; Berkeley, CA 94720, USA
[11]Center for Computational Biology, University of California, Berkeley; Berkeley, CA, USA
[12]Division of Biology and Biological Engineering, California Institute of Technology; Pasadena, CA 91125
[13]School of Medicine, University of California, San Diego; La Jolla, CA 92092, USA
[14]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory; Berkeley, CA 94720, USA.
[15]Feedstocks Division, Joint BioEnergy Institute; Emeryville, CA 94608, USA.

*Corresponding author, Email: savage@berkeley.edu

Abstract

Rubisco is the primary CO_2 fixing enzyme of the biosphere yet has slow kinetics. The roles of evolution and chemical mechanism in constraining the sequence landscape of rubisco remain debated. In order to map sequence to function, we developed a massively parallel assay for rubisco using an engineered *E. coli* where enzyme function is coupled to growth. By assaying >99% of single amino acid mutants across CO_2 concentrations, we inferred enzyme velocity and CO_2 affinity for thousands of substitutions. We identified many highly conserved positions that tolerate mutation and rare mutations that improve CO_2 affinity. These data suggest that non-trivial kinetic improvements are readily accessible and provide a comprehensive sequence-to-function mapping for enzyme engineering efforts.

Introduction

Plants, algae and photosynthetic bacteria together fix ~100 gigatons of carbon annually using ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco), the most abundant enzyme on earth (*1*). Rubisco catalysis, which is slow compared to many other central carbon metabolic enzymes (*2*), is thought to limit photosynthesis under common conditions (*3*). Rubisco is also prone to a side-reaction with oxygen leading to the hypothesis that this apparent inefficiency is in fact a careful balance of multiple biochemical tradeoffs between rate, affinity and promiscuity (*4*–*7*).

Efforts to engineer improvements to rubisco have been hampered by the low throughput of obtaining accurate measurements for its parameters including catalytic rate for carboxylation ($k_{cat,C}$, hereafter k_{cat}), CO_2 affinity (K_C) and specificity for CO_2 vs. O_2 ($S_{C/O}$). A concentrated effort across several decades has produced several hundred biochemical measurements of natural and mutant rubiscos (*4*–*7*). Collection of these measurements has been biased towards plant rubiscos and the diversity of natural rubiscos remains undersampled. Library screens and rational mutations have been used in the past to increase rubisco activity.

These efforts often resulted in improved expression (*8*) but rarely led to fundamental biochemical improvements (*9*).

Protein engineering has benefited in recent years from the introduction of machine learning approaches. One goal of such efforts is to train models with labeled protein sequence-function data from high throughput functional screens (*10*–*15*). Enzyme engineering with machine learning presents an additional challenge: ideally, functional data would be decomposed into individual catalytic parameters measured in high throughput either *in vitro (16)* or *in vivo (14)*.

Here, we have developed a selection assay in *E. coli* to estimate the carboxylation fitness of >99% (8760/8835) of the single amino acid mutants of the model Form II rubisco from *Rhodospirillum rubrum* (Fig. S1). Ribose phosphate isomerase was knocked out to generate *Δrpi*, a strain which only grows on glycerol when it expresses functional rubisco (Fig. S2). We then generated a barcoded library of single-amino acid mutations of the *R. rubrum* rubisco, which we assayed in high-throughput using *Δrpi*. By varying the CO_2 concentrations of the growth environment, we were able to estimate the CO_2 affinities ($\widetilde{K_C}$) of 65% (5687) of the rubisco variants, a subset of which we went on to validate *in vitro*. This screen revealed a very small minority of mutations which improved affinity for CO_2 ≈3-fold. These affinities have never before been observed among bacterial rubiscos, and are more typical of the Form I rubiscos found in plants and algae.

Results and Discussion

High-throughput characterizations of rubisco variants

The rubisco-dependent *E. coli* strain, *Δrpi*, cannot grow when glycerol is provided as the only carbon source because ribulose-5-phosphate accumulates with no outlet (*17*). The combined actions of phosphoribulokinase (which produces the five-carbon rubisco substrate; PRK) and rubisco rescue growth by converting this otherwise dead-end metabolite into 3-phosphoglycerate, which can feed back into central carbon metabolism (Fig. 1A, S2A, for similar selection systems see (*18*, *19*)).

We first confirmed that the growth rate of *Δrpi* was quantitatively related to known *in vitro* enzyme behavior (Fig. 1B, S2). Expression of rubisco driven by an inducible promoter demonstrated that growth rates increased with the rubisco concentration, indicating that increased enzyme concentration led to higher fitness (Fig S2B,C, S3). Similarly, we observed faster growth in the presence of higher CO_2 concentrations (Fig S2B,D). We next assessed whether growth-based selection correlated with biochemical behavior. Previous work on *R. rubrum* rubisco identified 77 mutants spanning <1% to 100% of wild-type catalytic rate (Supplemental Data File 1). Growth of a subset of these mutants was tested and found to correlate with reported catalytic rates (Fig. S4). Together, these results are consistent with glycerol growth of *Δrpi* being limited by rubisco carboxylation flux, which is determined by enzyme kinetics – k_{cat}, K_C – as well as enzyme and [CO_2] concentrations.

We next constructed a library of all single amino acid substitutions to the model Form II rubisco from *R. rubrum* (Fig. S3A). This library was cloned into a selection plasmid containing *PRK*, barcoded, and bottlenecked to ~500,000 colonies. Long read sequencing was used to map barcodes to mutants (Fig S5B, S6) and determined that the final library contained ≈180,000 barcodes, representing 8760 mutants or >99% of the designed library (Fig S6C-F).

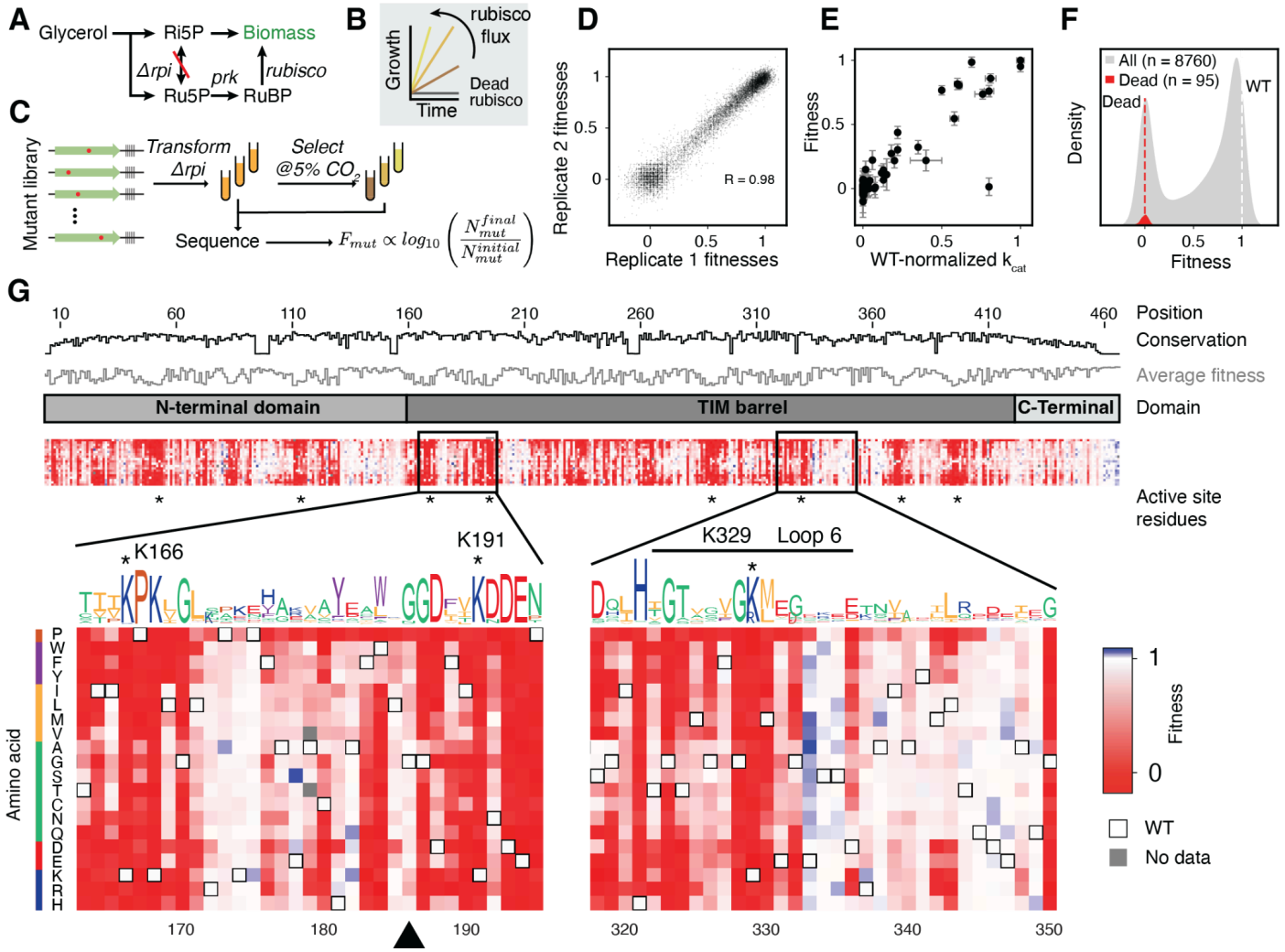

Figure 1: A deep mutational scan individually characterizes all single amino-acid mutations in rubisco. A) A summary of the metabolism of *Δrpi* the rubisco-dependent strain. **B)** *Δrpi* grows with a rate that is proportional to the flux through rubisco. **C)** Schematic of the library selection. A library of rubisco single amino-acid mutants was transformed into *Δrpi* then selected in minimal media with supplemented glycerol at elevated CO_2. Samples were sequenced before and after selections and barcode counts were used to determine the relative fitness of each mutant. **D)** Correspondence between 2 example biological replicates, each point represents the median fitness among all barcodes for a given mutant. **E)** Fitness of 77 mutants with measurements in previous studies compared to the catalytic rates measured in those studies (k_{cat}). The outlier I190T, see supplemental text for discussion. **F)** Histogram of all variant fitnesses (grey) were normalized between values of 0 and 1 with 0 representing the average of fitnesses of mutations at a panel of known active-site positions (red distribution, average is plotted as a red dashed line) and 1 representing the average of WT barcodes (white dashed line). **G)** A heatmap of variant fitnesses. Conservation by position and the sequence logo were determined from a multiple sequence alignment of all rubiscos. Black triangle indicates G186, an example of a position with high conservation that is mutationally tolerant.

This library was transformed into *Δrpi* to assess mutant fitness (Fig. 1C). Mutant fitness is defined by the relative growth rate of *Δrpi* expressing that mutant. Three independent library transformations were grown in selective conditions and grown for ≈7 divisions in 5% CO_2 (equivalent to ≈1200 μM CO_2 in solution; wild-type K_C = 150 uM). Short read sequencing quantified barcode abundance before and after selection (see supplemental methods). Mutant fitness was calculated by normalizing pre- and post-selection $\log_{10}$ read-count ratios to a panel of known catalytically dead mutants and all wild-type barcodes (see Methods). Nine replicate experiments were performed with an average pairwise Pearson coefficient of 0.98 (Fig. 1D, S7).

We compared mutant fitness measurements against 77 catalytic rate values taken from the literature (Fig. 1E, Supplemental Data File 1) as well as 35 *in vitro* measurements from purified mutants (Fig. S8B) and

observed a linear relationship. Overall, we observed a bimodal distribution of mutant effects (Fig. 1F) with mutant fitnesses clustering near wild-type (neutral mutations) and catalytically dead variants (*12*, *20*).

We measured fitness values for >99% (8760 out of 8835) of amino-acid substitutions (Fig. 1G, S6F, S9). Fewer than 0.14% mutations appeared more fit than WT, and when they did it was by a small amount (Fig. 1F) and 72.76% were found to be deleterious. Mutations at known active-site positions had very low fitness (e.g. K191, K166, K329, residues with asterisks Fig. 1G bottom), and mutations to proline were more deleterious on average than any other amino acid (Fig. S12). Phylogenetic conservation and average fitness at each position tended to anti-correlate (Fig. 1G top tracks, 2D, S13) consistent with previous studies (*21*, *22*) however, several positions appeared to be both highly conserved and mutationally tolerant (Fig. 1G black triangle).

Mutational sensitivity varies across the rubisco structure

Our fitness assays revealed that some regions of the rubisco structure are much more sensitive to mutation than others (Fig. 2A.B). For example, residues on the solvent exposed faces of the structure are more tolerant to mutation, as expected, while active site and buried residues typically do not tolerate mutations well. A notable region of interest is Loop 6 of the TIM barrel, which is known to fold over the active site during substrate binding and to participate in catalysis (Fig. S1C, Fig. 2C inset). Despite this key role in catalysis, some residues in this loop are highly tolerant to mutation (e.g. E331 and E333), though the active-site residue K329 is highly sensitive (Fig. 2C).

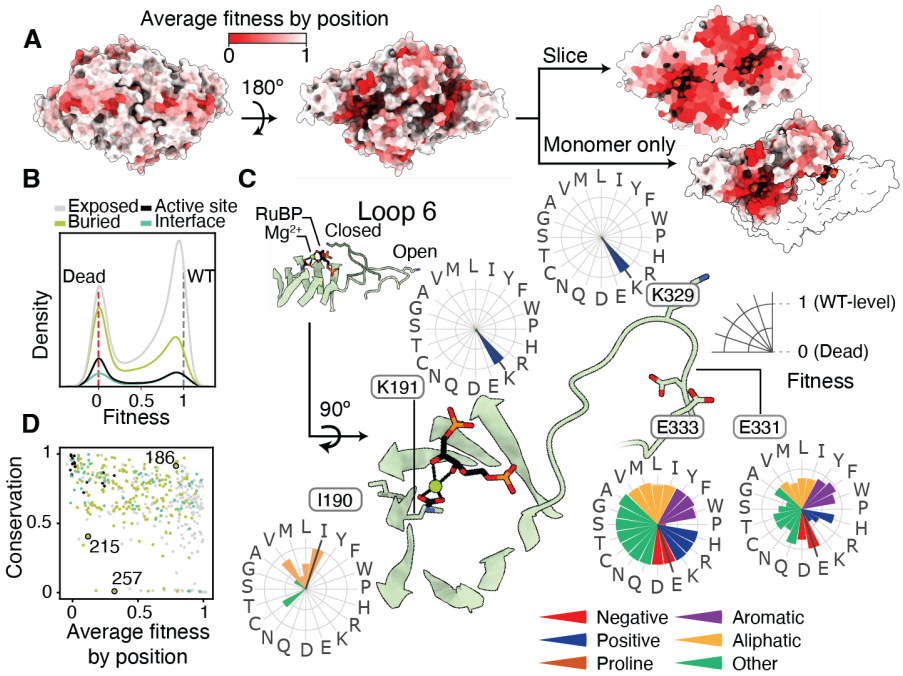

Figure 2: Fitness values provide structural, functional and evolutionary insights in rubisco. A) Structure of *R. rubrum* rubisco homodimer (Protein Data Bank (PDB) ID: 9RUB) colored by the average fitness value of a substitution at every site. **B)** Histograms of variant effects for amino-acids in different parts of the homodimer complex. **C)** Comparison of average fitness at each position against phylogenetic conservation among all rubiscos. Positions colored by the same scheme as part B. Positions 215 and 257 form a tertiary interaction, position 186 is highly conserved with no known function. **D)** Close-up view of the active-site and the mobile loop 6 region. Radar plots show the fitness effects of all mutations at a given position.

We expected that conserved positions would not tolerate mutations well. Consistent with this common hypothesis, the average fitness value at each position was negatively correlated with sequence conservation (Fig. 2C and S13). There were, however, many outliers with a number of positions being highly conserved yet

showing high mutational tolerance (e.g. G186, Fig. 2D top right corner). Selection in alternative conditions may reveal what selective forces have maintained high conservation at those positions(*23*). Positions with low conservation and low mutational tolerance may indicate a recently evolved, but critical, function (*21*, *22*); for example, M215 and H257 (Fig. 2D) are in contact in the *R. rubrum* structure but are absent in Form I sequences (Fig. S13).

Enzyme activity and affinity can be inferred by substrate titration

Enzyme fitness is determined by the underlying biochemical parameters including catalytic rates and affinities. In order to measure these parameters individually we performed a substrate titration on the whole library of mutations in tandem (Fig. 3A). Mutant fitness values varied overall with increasing [CO_2] (Fig. 3B, S14, S15) and some mutants' fitnesses were strongly affected (Fig. 3C). We fit the data to a Michaelis-Menten model of catalysis to estimate effective maximum rates ($\widetilde{V}_{max}$) and CO_2 half-saturation constants ($\widetilde{K}_C$) (*14*). This fitting (Fig. 3D, see Methods) generated $\widetilde{V}_{max}$ and $\widetilde{K}_C$ estimates for every mutant (Fig. 3G S10 and S11). We judged the reliability of the estimates by the coefficient of variation (standard deviation over the mean; σ/μ) of 1100 fits of the data for each mutation (see Methods); we focus here on the 65% of the mutants (5687) that had a coefficient of variation under 1 (*21*). The remaining 35% are primarily mutants with low fitness values (Fig. S16) which may fail to fold altogether, though at higher expression levels or in combination with other mutations it may yet be possible to produce reliable estimates of their effects on rate and affinity.

We validated our $\widetilde{K}_C$ estimates by purifying a set of 7 mutants chosen to span a range of predicted $\widetilde{K}_C$ values and measuring their CO_2 affinities *in vitro* (Fig. 3E). Unexpectedly, for several mutants, the *in vitro*-measured K_C values were substantially lower (i.e. tighter affinity) than expected from our prior estimates based on fitness data. For example, the $\widetilde{K}_C$ of V266T was ≈130μM but K_C was determined to be ≈80μM CO_2 (Fig. 3G highlighted box, Fig. 3F).

Our estimates of $\widetilde{V}_{max}$ correlated with fitness (r = 0.93, Fig. S16) indicating that it is the primary driver of rubisco flux. However, $V_{max} = k_{cat}$ x [rubisco] so variation in V_{max} can have two potential causes: rubisco expression level and k_{cat}. $\widetilde{V}_{max}$ estimates report the product of those two factors.

We further found that $\widetilde{V}_{max}$ and $\widetilde{K}_C$ estimates anticorrelate for variants with near-WT kinetics where the estimates are most reliable (Fig. 3H). This correlation implies that, in the absence of selective pressure, the majority of single amino acid mutations harm CO_2 affinity and V_{max} in tandem. As there is no binding site for CO_2 in the enzyme (*24*), this trend may be related to subtle changes in the electronics of the active site or the geometry of the bound sugar substrate before bond-formation with CO_2. It is also possible that these effects are caused by changes to enzyme stability.

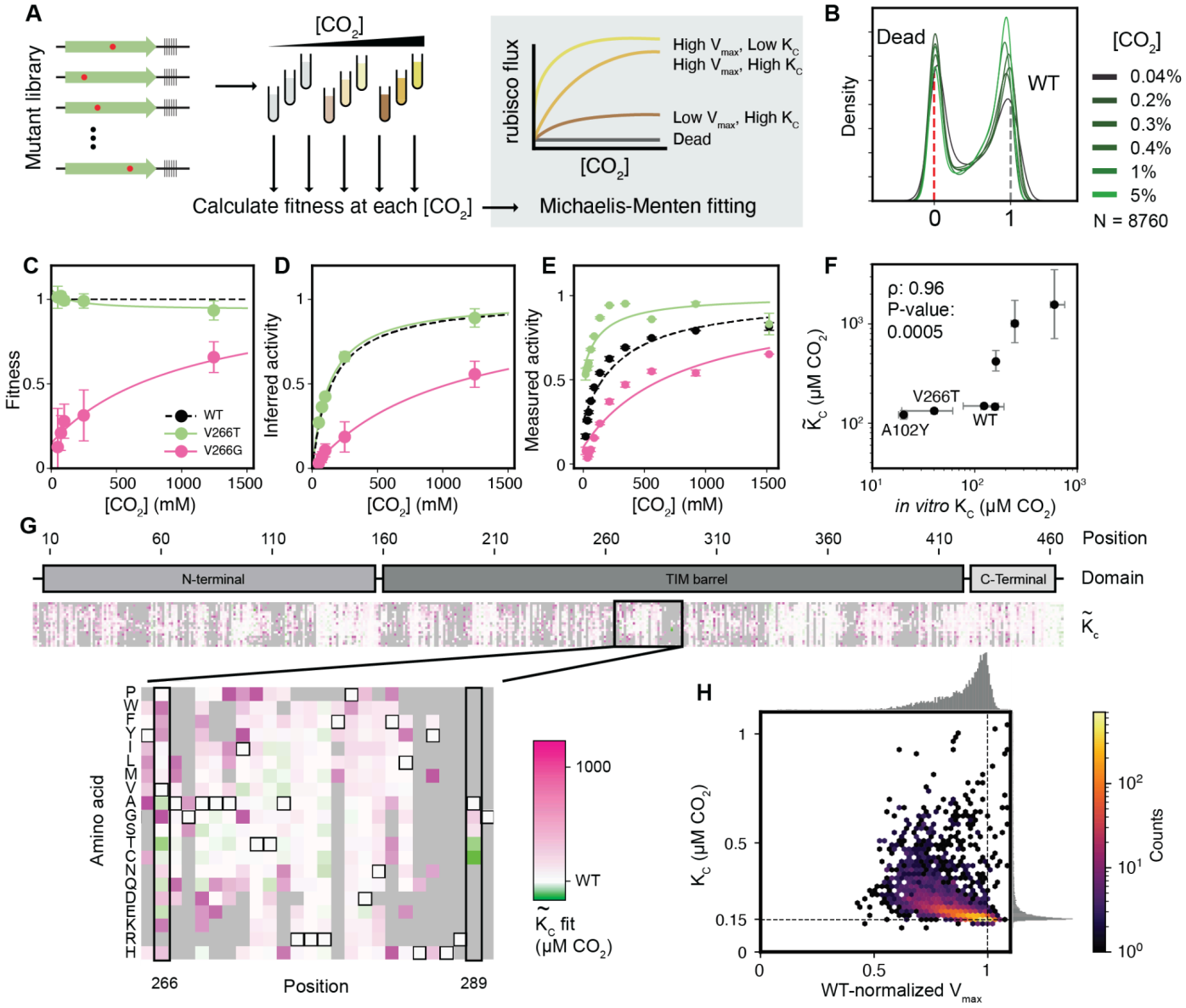

Figure 3: $\widetilde{K_C}$ and $\widetilde{V_{max}}$ can be inferred from fitness across a CO_2 titration. A) Schematic of rubisco selection in [CO_2] titration and some examples of inferred Michaelis-Menten curves of mutants with varying K_C and V_{Max}. **B)** Histograms of variant fitnesses at different [CO_2]. **C)** Measured fitnesses at different [CO_2] for two mutants. **D)** The same data as in **C** plotted under the assumptions of the Michaelis-Menten equation. **E)** Individually measured rubisco kinetics for the same two mutants from **C** and **D**. **F)** Comparison between *in vitro*-measured rubisco K_C and those inferred from fitness values ($\widetilde{K_C}$). **G)** Heatmap of $\widetilde{K_C}$ values for all mutants where the coefficient of variation is <1 (N = 5687 mutants, 65% of total). Two positions with high-affinity mutations are highlighted in the inset below. Variants where the $\widetilde{K_C}$ fits had a coefficient of variation above 1 are in gray. **H)** Two dimensional histogram of $\widetilde{K_C}$ and $\widetilde{V_{max}}$ from **G** with hexagonal bins. Dashed lines represent the WT values.

Three mutations (A289C, A102Y, V266T), caused strong improvements in CO_2 affinity *in vivo* (Fig. 3G, 4A). Other mutations at these same positions also reduced affinity (e.g. V266G, A102F, A289G, Fig. 3C-G). These three positions are not part of the active site and sit near the C_2 axis of the rubisco homodimer interface (Fig. 4B). In this region of the structure, residues are in closest proximity to "themselves" - i.e. to their counterpart residue in the other monomer of the homodimer. The role these amino acids play in CO_2 entry into the active site, active site conformation, or electrostatics remains unclear.

In vitro measurements confirmed that V266T and A102Y possess improved CO_2 affinities (we were unable to purify A289C). This correspondence between $\widetilde{K_C}$ measured *in vivo* and K_C measured *in vitro* stands in contrast to mutations with $\widetilde{V_{max}}$ where followup biochemistry (Fig. S8B, Supplementary Data 1) did not reveal faster k_{cat} values. Variants with improved $\widetilde{V_{max}}$ were likely improved through higher protein expression. Our $\widetilde{K_C}$ predictions were isolated from expression effects, because mutants were judged individually by their relative performance across a CO_2 titration, and were thus more accurate. V266T and A102Y both exhibit roughly-proportional reductions in catalytic rate (Fig 4C, Table S2). The k_{cat} and K_C measurements place these mutants outside of the range heretofore measured among bacterial Form II variants and at the edge of the distribution of plants and algae.

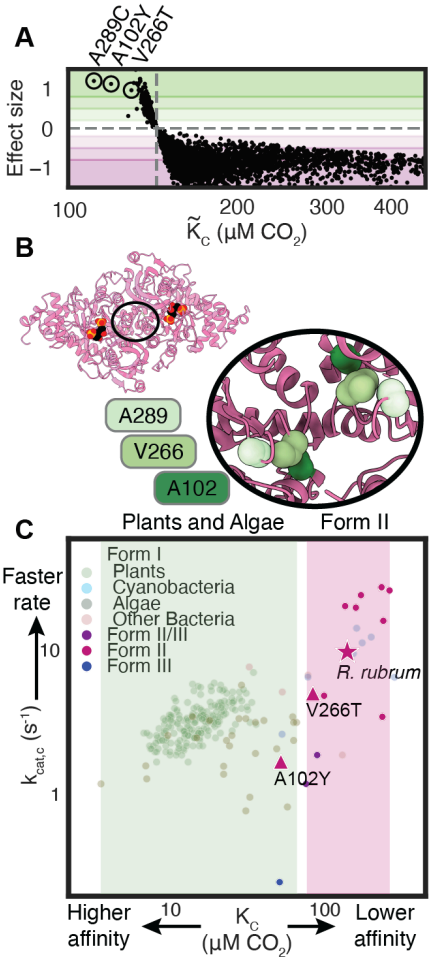

Figure 4: Single amino-acid mutations can traverse the functional landscape. A) $\widetilde{K_C}$ vs. effect size for each mutant. Effect size is the difference between the mutant $\widetilde{K_C}$ and WT K_C divided by the coefficient of variation of $\widetilde{K_C}$. **B)** PDB structure 9RUB with inset of a zoom on the C_2 symmetry axis. Each position appears twice due to proximity to the C_2 axis. **C)** k_{cat} vs. K_C of the indicated mutants vs. all measured rubiscos from(*6*, *25*). Shaded regions indicate known ranges of $\widetilde{K_C}$ values for plants and algae in green and Form II bacterial rubiscos in pink. WT *R. rubrum* is represented by a star while mutants A102Y and V266T are triangles.

Conclusion

Among the narrow range of sequences measured here it was possible to identify mutants with substantially improved CO_2 affinity, suggesting the enzyme parameter landscape is rugged with apparent gain-of-function readily accessible. Form I plant rubiscos typically share <50% identity with Form II bacterial rubiscos (>200 mutations, Fig S17) and are thought to have evolved under a different set of selective

constraints. Furthermore, Form I and II rubiscos have different oligomeric states and Form II rubiscos lack the small subunit characteristic of Form I, so it is surprising that it is possible to traverse the functional space between them with just one amino acid change. In *R. rubrum*, the present-day sequence evolved under constraints including endogenous regulation, environmental selective pressure and possible tradeoffs between enzymatic parameters.

Various tradeoffs have been proposed in the catalytic mechanism of rubisco (*4*, *6*), including one between catalytic rate and CO_2 affinity (*5*). The reductions in k_{cat} observed in the mutants with the highest CO_2 affinity is consistent with such a tradeoff (Fig. 4C). A selection of a library of higher order mutants which spans a wider range of rubisco functional possibilities could confirm or reject a tradeoff. The tradeoffs in bacterial rubiscos may also constrain the evolution of plant rubiscos. However, previous work comparing the sequence-to-function map of related proteins found substantial context-dependence on the effects of mutations (*12*). Due to advancements in expressing plant rubiscos in *E. coli (26)*, it may be possible to use this assay to understand the biochemical constraints of the organisms which are responsible for nearly all of terrestrial photosynthesis (*27*).

The overall space of rubiscos remains largely unexplored, raising the question of whether natural evolution has already produced rubiscos optimized for every environment. A higher throughput exploration of sequence space may reveal regions which are constrained by different tradeoffs and produce substantial engineering improvements.

Materials and Methods

Strains, plasmids and primers

Strains:

Cloning was performed in a combination of *E. coli* TOP10 cells, DH5α and NEB Turbo cells. Protein expression was carried out using BL21(DE3). *Δrpi* was previously produced from the BW25113 strain by knocking out *rpiA* from the Keio strain lacking *rpiB* as well as the *edd* gene. The latter deletion makes the strain rubisco-dependent when grown on gluconate, a feature we did not make use of in this study.

Plasmids:

Sequences and further details about plasmids used in this study can be found in supplemental data file S3.

pUC19_rbcL

The rubisco mutant library was assembled in a standard pUC19 vector. This plasmid was used as a PCR template for each of the 11 sublibrary ligation destination sites.

NP-11-64-1

Selections were conducted using a plasmid designed for this study with a p15 origin, chloramphenicol resistance, LacI controlling rubisco expression, TetR controlling PRK expression and a barcode.

NP-11-63

Protein overexpression in BL21(DE3) cells was conducted using pET28 with a SUMO domain upstream of the expressed gene (*25*). pSF1389 is the plasmid that expresses the necessary SUMOlase, bdSENP1, from *Brachypodium distachyon*.

Primers:

All primers were purchased from IDT and the oligo pool was purchased from TWIST. For sequences see supplemental data file S3.

Library design and construction

The *R. rubrum* rubisco sequence was codon-optimized for *E. coli* and systematically mutated via the scheme outlined in Fig. S5. The rubisco gene was split into 11 pieces. For each of those pieces (≈200 bp each) all point mutants were designed and synthesized as oligonucleotide pools. 11 oligo sub-library pools, containing all single mutants within their respective ≈200 bp region, were purchased from Twist Bioscience and

each sub-library was amplified individually using Kapa Hifi polymerase with a cycle number of 15. Each rubisco gene fragment was inserted into a corresponding linearized pUC19 destination vector, containing the remainder of the rubisco sequence flanking the insert, via golden gate assembly. This assembly generated 11 sub-libraries of the full-length *R. rubrum* rubisco gene with each sub-library containing a ≈200 bp region including all single mutants. Each of these 11 rubisco libraries were separately transformed into *E. coli* TOP10 cells and in each case >10,000 transformants were scraped from agar plates to ensure oversampling of the ≈1,000 variants in each sublibrary. Plasmids were purified from each sub-library and mixed together at equal molar ratios to generate the full protein sequence library.

In order to produce the final library for assay, a selection plasmid containing an induction system for rubisco and PRK (Tac- and Tet-inducible, respectively) was amplified with primers that included a random 30 nucleotide barcode. The linearized plasmid amplicon and the library were cut with BsaI and BsmBI, respectively, ligated together and transformed into TOP10 cells. Plasmid was purified by scraping ≈500,000 colonies and transformed in triplicate into *Δrpi* cells. These transformations were grown in 2XYT media into log phase (OD = 0.6) and frozen as 25% glycerol stocks.

Long-read sequencing analysis

The plasmid library was cut with SacII and sent for Sequel II PacBio sequencing. Reads were aligned and grouped by their barcodes. All reads of a given barcode were aligned and a consensus sequence was obtained using SAMtools(*28*). Consensus sequences were retained if they were WT or had one mutation that matched the designed library. Any mutation in the backbone invalidated a barcode. A lookup table was generated to link each barcode to its associated mutation. The *in silico* procedures described in this study are publicly available at https://github.com/SavageLab/rubiscodms.

Library characterization and screening

Selections were performed by diluting 200 uL of glycerol stock with OD of ≈0.25 into 5 mL of M9 minimal media with added chloramphenicol (25 µg/mL), glycerol (0.4%), 20 µM IPTG and 20 nM anhydrotetracycline. These cultures were grown at 37 °C in different CO_2 concentrations until they reached an OD at 5 mL of 1.2 +/- 0.2. This corresponds to a 100-fold expansion of the cells, i.e. between 6 and 7 doublings.

Cultures before and after selection were spun down and we lysed the cells and performed a standard plasmid extraction protocol using QIAprep Spin Miniprep Kit (QIAGEN, Hilden, Germany). Illumina amplicons were generated by PCR of the barcode region. These amplicons were sequences using a NextSeq™ P3 kit

Calculation of variant enrichment

Variant enrichments were computed from the log ratio of barcode read counts. The enrichment calculations include two processing parameters: a minimum count threshold (c_{min}) and a pseudocount constant (α_p). The count threshold is the minimum number of barcode reads that must be observed either pre- or post-selection for the barcode to be included in the enrichment calculation. The pseudocount constant is used to add a small positive value to each barcode count to circumvent division by zero errors. We use a pseudocount value that is weighted by the total number of reads in each condition. For the j^{th} variant and the individual barcodes, *i*, passing the threshold condition the variant enrichment is calculated as,

Eq. 1
$$e_j = median\left(log_{10}\left(\frac{N_{f,i}+N_{f,tot}\alpha_p}{N_{0,i}+N_{0,tot}\alpha_p}\right) - log_{10}\left(\frac{N_{f,tot}}{N_{0,tot}}\right)\right)$$

To identify optimal values for these parameters, we computed the variant enrichments across a 2D parameter sweep of c_{min} and α_p to find the combination that resulted in the maximum mean Pearson correlation coefficient across all replicates at each condition. These were c_{min} = 5 and α_p = 3.65e-7 (average of 0.3

pseudocounts) leading to a correlation coefficient of 0.978. Variant enrichment, e_j, was then calculated for every mutant using Eq. 1.

The variant enrichments were then normalized such that wild-type has an enrichment value of 1 in all conditions and catalytically dead mutants have a median enrichment of 0. For the "dead" variant enrichment we computed the median enrichment for all mutations at the catalytic positions K191, K166, K329, D193, E194, and H287. The normalized enrichments at each condition were computed as,

Eq. 2
$$e_{j,\,norm} = \frac{e_j - \tilde{e}_{dead}}{e_{wt} - \tilde{e}_{dead}}$$

where e_j is the enrichment of the j^{th} variant as given in Eq. 1, e_{wt} is the wild-type enrichment, and $\tilde{e}_{dead}$ is the median enrichment across all mutants of the catalytic residues listed above.

Michaelis Menten fits to enrichment data

The DMS library enrichments across different CO_2 concentrations were used to estimate Michaelis-Menten kinetic parameters for every variant. Guided by the linear relationship between growth rate and k_{cat} observed in Fig. 1D we assume that the cell growth rate is proportional to the rubisco enzyme velocity to derive the CO_2 titration fits (see SI, Derivation of Michaelis-Menten Fit).

Eq. 3
$$e_{mut,\,norm}\left([CO_2]\right) = \frac{V_{max,\,mut}\left(K_{C,wt} + [CO_2]\right)}{V_{max,wt}\left(K_{C,mut} + [CO_2]\right)}$$

$\tilde{V}_{max,mut} / \tilde{V}_{max,WT}$ is the ratio of mutant maximum velocity relative to wild-type, $\tilde{K}_{C,wt}$ is the wild-type K_C for which we used the value 149 μM, and $\tilde{K}_{C,\,mut}$ is the mutant K_C. The titration curves in triplicate for each variant were fit to Eq. 3 using non-linear least squares curve fitting while requiring both V_{max} and $K_{C,\,mut}$ to be positive.

We noted that the $\tilde{K}_C$ fits to certain variants – particularly ones with low $\tilde{V}_{max}$ – were sensitive to the choice of processing parameters c_{min} and α_p. Given the semi-arbitrary nature of these parameters, this is clearly an undesirable dependence and engenders low confidence in the inferred $\tilde{K}_C$ values. To account for this uncertainty we conducted a parameter sweep (with 11 different c_{min} values linearly spaced between 0 and 50, and 10 α_p values log spaced between 1e-9 and 1e-6), and computed the variant enrichments for all combinations of these parameters. Then we performed 10 bootstrap subsamplings of the replicates for all parameter sets and performed the ratiometric Michaelis-Menten fit.From this set of 1100 $\tilde{K}_C$ fit values for each variant we computed a quartile-based coefficient of variation that was used as a figure of merit for the $\tilde{K}_C$.

Multiple sequence alignment

An MSA of the broader rubisco family beyond Form II rubiscos was created using the profile HMM homology search tool jackhmmer (*29*). Starting with the *R. rubrum* rubisco sequence, jackhmmer applied five search iterations with a bit score threshold of 0.5 bits/residue against the UniRef100 database of non-redundant protein sequences (*30*). To compute phylogenetic conservation at each position, for each possible amino acid we computed the fraction of the total sequences that had that amino acid at the corresponding position of the MSA. The phylogenetic conservation is the maximum fraction, where the maximum is taken over all possible amino acids. Thus, if a position has an alanine in 90% of the sequences of the MSA, the phylogenetic conservation will be 0.9.

Protein purification

E. coli BL21(DE3) cells were transformed with pET28 (encoding the desired rubisco with a 14x His and SUMO affinity tag) and pGro plasmids. Colonies were grown at 37°C in 100mL of 2XYT media under

Kanamycin selection (50 µg/ml) to an OD of 0.3-1. 1 mM arabinose was added to each culture and then incubated at 16°C for 30 minutes. Protein expression was induced with isopropyl-b-D-thiogalactopyranoside (IPTG, Millipore) at 100 uM and cells were grown overnight at 16°C. Cultures were spun down (15 min; 4,000 g; 4°C) and purified as reported (*25*). Briefly, cultures were spun down and lysed using BPER-II™. Lysates were centrifuged to remove insoluble fraction. His-tag purification using Ni-NTA resin (Thermo Fisher, Massachusetts, United States) was performed and rubisco was eluted by SUMO tag cleavage with bdSUMO protease (as produced in Davidi et al. 2020). Purified proteins were concentrated and stored at 4 °C until kinetic measurement (within 24 hr). Samples were run on an SDS-PAGE gel to ensure purity.

Rubisco spectrophotometric assay

Both k_{cat} and K_C measurements use the same coupled-enzyme mixture wherein the phosphorylation and subsequent reduction of 1,3-bisphosphoglycerate, the product of RuBP carboxylation, was coupled to NADH oxidation which can be followed through 340 nm absorbance. Following (*31*) and (*25*) the reaction mixture (Table S1) contains buffer at pH 8, $MgCl_2$, DTT, 2 mM ATP, 10 mM creatine phosphate, 0.5mM NADH, 1mM EDTA and 20U/mL each of PGK, GAPDH and creatine phosphokinase. Reaction volumes are 150 µL and samples are shaken once before absorbance measurements begin. Absorbance measurements are collected on a SPARK plate reader with O_2 and CO_2 control (TECAN). The extinction coefficient of NADH in the plate reader was determined through a standard curve of NADH solutions of known concentration (determined by a genesys20 spectrophotometer with a standard 1 cm pathlength, Thermo Fisher). Absorbance over time gives a rate of NADH oxidation and therefore a carboxylation rate. Because rubisco produces 2 molecules of 3-phosphoglycerate for every carboxylation reaction we assume a 2:1 ratio of NADH oxidation rate to carboxylation rate.

Table S2. Assay mix composition.

Component	Assay concentration	Source
EPPS buffer pH 8.0	100 mM	Alfa Aesar (Cat # J61296)
MgCl2	20 mM	Sigma Aldrich (Cat # M2670-500G)
Dithiothreitol	0.5 mM	Bio Basic Canada inc. (Cat # DB0058)
ATP	2 mM	Sigma Aldrich (Cat # A3377-5G)
Phosphocreatine	10 mM	Sigma Aldrich (Cat # 27920-5G)
NADH	1.7 mM	Merck (Cat # 481913-1GM)
Carbonic anhydrase	0.1 mg/mL	Sigma Aldrich (Cat # C3934-100MG)
Creatine phosphokinase	20 U/mL	Sigma Aldrich (Cat # C3755-35KU)
Glyceraldehyde 3-phosphate dehydrogenase	20 U/mL	Sigma Aldrich (Cat # G2267-10KU)
3-Phosphoglyceric phosphokinase	20 U/mL	Sigma Aldrich (Cat # P7634-5KU)

Spectrophotometric measurements of k_{cat}

The carboxylation rate (k_{cat}) of each rubisco was measured using methods established previously (*25*). Briefly, rubisco was activated by incubation for 15 minutes at room temperature with CO_2 (4%) and O_2 (0.4%) and added (final concentration of 80 nM) to aliquots of appropriately-diluted assay mix (see Table S2)

containing different CABP concentrations pre-equilibrated in a plate reader (Infinite® 200 PRO; TECAN) at 30°C, under the same gas concentrations. After 15 min, RuBP (final concentration of 1 mM) was added to the reaction mix and the absorbance at 340 nm was measured to quantify the carboxylation rates. A linear regression model was used to plot reaction rates as a function of CABP concentration. The k_{cat} was calculated by dividing y-intercept (reaction rates) by x-intercept (concentration of active sites). Protein was purified in triplicate for k_{cat} determination.

Spectrophotometric measurements of K_C

Purified rubisco mutants were activated (40 mM bicarbonate and 20 mM MgCl2) and added to a 96-well plate along with assay mix (Table S2, in this case the same concentration of Hepes pH 8 buffer was used but EPPS can be substituted). Bicarbonate was added for a range of concentrations (1.5, 2.5, 4.2, 7, 11.6, 19.4, 32.4, 54, 90 and 150mM). Plates and RuBP were pre-equilibrated at 0.3% O_2 and 0% CO_2 at room temperature. RuBP was added to a final concentration of 1.25 mM with water serving as a control for each replicate. NADH oxidation was measured by A340 as in the kcat assay. Absorbance curves were analyzed using a custom script to perform a hyper-parameter search to choose a square in which to take the slope as carboxylation rate that best represented the majority of the monotonic decrease in A_{340}. K_C was derived by fitting the Michaelis-Menten curve using a non-linear least squares method. Error bars were determined depending on replicates: (1) Multi day replicates: Michaelis-Menten fits were made for each replicate, std error and median was calculated based on these fits (2) Triplicates: Absorbance data was fit 100 times using different hyperparameters. Michaelis-Menten fits of each set of rates were calculated and the median K_C value was plotted. Error values were determined from the K_C values of the hyperparameters one standard deviation above and below the median. Standard deviations and medians were calculated based on technical replicates. Subsequently, three different fits were made: one based on the median, one based on the lowest reaction rate and one based on the highest reaction rate for each point.

Radiometric measurements of K_C and k_{cat}

$^{14}CO_2$ fixation assays were conducted as in (Davidi et al. (*25*)) with minor modifications. Assay buffer (100 mM EPPES-NaOH pH 8, 20 mM $MgCl_2$, 1 mM EDTA) was sparged with N_2 gas. Rubisco, purified as described above, was diluted to ~10 µM (quantified using UV absorbance) in assay buffer. It was then diluted with one volume of assay buffer containing 40 mM $NaH^{14}CO_3$ to activate. 0.5 mL reactions were conducted at 25°C in 7.7 ml septum-capped glass scintillation vials (Perkin-Elmer) with 100 µg/mL carbonic anhydrase, 1 mM RuBP and $NaH^{14}CO_3$ concentrations ranging from 0.4 to 17 mM (which corresponds to 15 to 215 µM CO_2). The assay was initiated by the addition of a 20 µL aliquot of activated rubisco and stopped after 2 minutes by the addition of 200 µL 50% (v/v) formic acid.

The specific activity of $^{14}CO_2$ was measured by performing a 1 hour assay at the highest $^{14}CO_2$ concentration containing 10 nmoles of RuBP. Reactions were dried on a heat block, resuspended in 1 mL water and mixed with 3 mL Ultima Gold XR scintillant for quantification with a Hidex scintillation counter.

The rubisco active-site concentration used in each assay was quantified in duplicate by a [^{14}C]-2-CABP binding assay. 10 µL of the ~10 µM rubisco solution was activated in assay buffer containing 40 mM cold $NaHCO_3$ (final volume 100 µl) for at least ten minutes.1.5 µL of 1.8 mM ^{14}C-carboxypentitol bisphosphate was added and incubated for at least one hour at 25°C. [^{14}C]-2-CABP bound rubisco was separated from free [^{14}C]-2-CPBP by size exclusion chromatography (Sephadex G-50 Fine, gE Healthcare)) and quantified by scintillation counting.

The data was fit to the Michaelis-Menten equation using the concatenated data of 3-4 experiments performed on different days.

Quantification of soluble enzyme concentration via Immunoblot

Δrpi strain with WT rubisco was grown under selective conditions (overnight at 37 °C in M9 media with 0.4% glycerol and 20 nM aTc) with varying IPTG concentrations at 5% CO2 for 24h. Afterwards, turbid cultures were spun down (10 min; 4,000 g; 4°C) culminating in roughly 20 mg pellet per sample. Pellets were lysed with 200 µL of BPER II and supernatant was transferred into a fresh tube and mixed with SDS loading dye. BioRad RTA Transfer Kit for Transblot Turbo Low Fluorescence PVDF was used in combination with the Trans-Blot® Turbo™ Transfer System. Nitrocellulose Membrane was carefully cut between 50 and 70 kDa post-blocking using a razor blade. Primary Anti-RbcL II Rubisco large subunit Form II Antibody from Agrisera (1:10000) and DnaK Antibody from Abcam (1:5000) were incubated separately. Secondary HRP-conjugated antibodies Donkey anti-mouse for DnaK (Santa Cruz Biotechnology) and Goat pAB to RB IgG HRP (Abcam were both used at 1:10000. Subsequently BioRad Clarity Max Western ECL Substrates were applied and the final results were imaged using a GelDoc.

Acknowledgements

We thank Niv Antonovsky and Arren Bar-Even for taking part in formulating the basis for this work as well as Naama Tepper and Shira Amram for originally conceiving of and producing the *Δrpi* strain respectively. We thank Philip Romero, Nat Thompson, Leon Fedotov, Orren Saltzman, Eden Prywes, Stacia Wyman, Bin Yu and Jack Desmarais for essential help in the process of data analysis. For their assistance in the process of generating and validating the DMS library we thank Andrew Glazer, Kenneth Matreyek, Jesse Bloom and Kim Reynolds. Additionally we thank Julia Tartaglia for the use of her sequencing primers and Netra Krishnappa for assistance in running NGS samples. We would like to thank Elaine Meng for assistance using ChimeraX. Finally we thank Flora Wang for technical assistance over the weekends.

Funding:

National Institutes of Health grant K99GM141455-01 (NP)

DFS is an Investigator of the Howard Hughes Medical Institute

U. S. Department of Energy, Physical Biosciences Program, Award Number DE-SC0016240 (DFS)

Author contributions:

Conceptualization: NP, AIF, DFS

Methodology: NP, NRP, LMO, SL, DD, OMC, RM, DFS

Investigation: NP, NRP, SL, YCT, BdP, AEC, LJTK, HAC, LNH, DBR, HMN, RFW, AYB

Visualization: NP, LMO, SL, DFS

Funding acquisition: NP, DFS

Project administration: NP, DFS

Supervision: NP, PMS, OMC, RM, DFS

Writing – original draft: NP, LNH, AIF, DFS

Competing interests: DFS is a co-founder and scientific advisory board member of Scribe Therapeutics.

Data and materials availability: All data are available in the main text or the supplementary materials.

References

1. Y. M. Bar-On, R. Phillips, R. Milo, The biomass distribution on Earth. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6506–6511 (2018).

2. A. Bar-Even, R. Milo, E. Noor, D. S. Tawfik, The Moderately Efficient Enzyme: Futile Encounters and Enzyme Floppiness. *Biochemistry* **54**, 4969–4977 (2015).

3. A. Wu, J. Brider, F. A. Busch, M. Chen, K. Chenu, V. C. Clarke, B. Collins, M. Ermakova, J. R. Evans, G. D. Farquhar, B. Forster, R. T. Furbank, M. Groszmann, M. A. Hernandez-Prieto, B. M. Long, G. Mclean, A. Potgieter, G. D. Price, R. E. Sharwood, M. Stower, E. van Oosterom, S. von Caemmerer, S. M. Whitney, G. L. Hammer, A cross-scale analysis to understand and quantify the effects of photosynthetic enhancement on crop growth and yield across environments. *Plant Cell Environ.* **46**, 23–44 (2023).

4. G. G. B. Tcherkez, G. D. Farquhar, T. J. Andrews, Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 7246–7251 (2006).

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

6. A. I. Flamholz, N. Prywes, U. Moran, D. Davidi, Y. M. Bar-On, L. M. Oltrogge, R. Alves, D. Savage, R. Milo, Revisiting Trade-offs between Rubisco Kinetic Parameters. *Biochemistry* **58**, 3365–3376 (2019).

7. C. Iñiguez, S. Capó-Bauçà, Ü. Niinemets, H. Stoll, P. Aguiló-Nicolau, J. Galmés, Evolutionary trends in RuBisCO kinetics and their co-evolution with CO2 concentrating mechanisms. *Plant J.* **101**, 897–918 (2020).

8. N. Prywes, N. R. Phillips, O. T. Tuck, L. E. Valentin-Alvarado, D. F. Savage, Rubisco Function, Evolution, and Engineering, *Annu. Rev. Biochem.* (2023)pp. 385–410.

9. R. H. Wilson, H. Alonso, S. M. Whitney, Evolving Methanococcoides burtonii archaeal Rubisco for improved photosynthesis and plant growth. *Sci. Rep.* **6**, 22284 (2016).

10. A. J. Faure, J. Domingo, J. M. Schmiedel, C. Hidalgo-Carcedo, G. Diss, B. Lehner, Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).

11. D. Ding, A. Y. Shaw, S. Sinai, N. Rollins, N. Prywes, D. F. Savage, M. T. Laub, D. S. Marks, Protein design using structure-based residue preferences. *Nat. Commun.* **15**, 1639 (2024).

12. L. Gonzalez Somermeyer, A. Fleiss, A. S. Mishin, N. G. Bozhanova, A. A. Igolkina, J. Meiler, M.-E. Alaball Pujol, E. V. Putintseva, K. S. Sarkisyan, F. A. Kondrashov, Heterogeneity of the GFP fitness landscape and data-driven protein design. *Elife* **11** (2022).

13. S. Thompson, Y. Zhang, C. Ingle, K. A. Reynolds, T. Kortemme, Altered expression of a quality control protease in E. coli reshapes the in vivo mutational landscape of a model enzyme. *Elife* **9** (2020).

14. M. A. Stiffler, D. R. Hekstra, R. Ranganathan, Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* **160**, 882–892 (2015).

15. W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, R. Ranganathan, An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).

16. C. J. Markin, D. A. Mokhtari, F. Sunden, M. J. Appel, E. Akiva, S. A. Longwell, C. Sabatti, D. Herschlag, P. M. Fordyce, Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science* **373** (2021).

17. A. I. Flamholz, E. Dugan, C. Blikstad, S. Gleizer, R. Ben-Nissan, S. Amram, N. Antonovsky, S. Ravishankar, E. Noor, A. Bar-Even, R. Milo, D. F. Savage, Functional reconstitution of a bacterial CO2 concentrating mechanism in Escherichia coli. *Elife* **9** (2020).

18. M. R. Parikh, D. N. Greene, K. K. Woods, I. Matsumura, Directed evolution of RuBisCO hypermorphs through genetic selection in engineered E.coli. *Protein Eng. Des. Sel.* **19**, 113–119 (2006).

19. O. Mueller-Cajar, M. Morell, S. M. Whitney, Directed evolution of rubisco in Escherichia coli reveals a specificity-determining hydrogen bond in the form II enzyme. *Biochemistry* **46**, 14067–14074 (2007).

20. K. S. Sarkisyan, D. A. Bolotin, M. V. Meer, D. R. Usmanova, A. S. Mishin, G. V. Sharonov, D. N. Ivankov, N. G. Bozhanova, M. S. Baranov, O. Soylemez, N. S. Bogatyreva, P. K. Vlasov, E. S. Egorov, M. D. Logacheva, A. S. Kondrashov, D. M. Chudakov, E. V. Putintseva, I. Z. Mamedov, D. S. Tawfik, K. A. Lukyanov, F. A. Kondrashov, Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).

21. E. M. Jones, N. B. Lubock, A. J. Venkatakrishnan, J. Wang, A. M. Tseng, J. M. Paggi, N. R. Latorraca, D. Cancilla, M. Satyadi, J. E. Davis, M. M. Babu, R. O. Dror, S. Kosuri, Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. *Elife* **9** (2020).

22. S. Subramanian, K. Gorday, K. Marcus, M. R. Orellana, P. Ren, X. R. Luo, M. E. O'Donnell, J. Kuriyan, Allosteric communication in DNA polymerase clamp loaders relies on a critical hydrogen-bonded junction. *Elife* **10** (2021).

23. D. Mavor, K. A. Barlow, D. Asarnow, Y. Birman, D. Britain, W. Chen, E. M. Green, L. R. Kenner, B. Mensa, L. S. Morinishi, C. A. Nelson, E. M. Poss, P. Suresh, R. Tian, T. Arhar, B. E. Ary, D. P. Bauer, I. D. Bergman, R. M. Brunetti, C. M. Chio, S. A. Dai, M. S. Dickinson, S. K. Elledge, C. V. M. Helsell, N. L. Hendel, E. Kang, N. Kern, M. S. Khoroshkin, L. L. Kirkemo, G. R. Lewis, K. Lou, W. M. Marin, A. M. Maxwell, P. F. McTigue, D. Myers-Turnbull, T. L. Nagy, A. M. Natale, K. Oltion, S. Pourmal, G. K. Reder, N. J. Rettko, P. J. Rohweder, D. M. C. Schwarz, S. K. Tan, P. V. Thomas, R. W. Tibble, J. P. Town, M. K. Tsai, F. S. Ugur, D. R. Wassarman, A. M. Wolff, T. S. Wu, D. Bogdanoff, J. Li, K. S. Thorn, S. O'Conchúir, D. L. Swaney, E. D. Chow, H. D. Madhani, S. Redding, D. N. Bolon, T. Kortemme, J. L. DeRisi, M. Kampmann, J. S. Fraser, Extending chemical perturbations of the ubiquitin fitness landscape in a classroom setting reveals new constraints on sequence tolerance. *Biol. Open* **7** (2018).

24. S. Gutteridge, M. A. J. Parry, C. N. G. Schmidt, J. Feeney, An investigation of ribulosebisphosphate carboxylase activity by high resolution1H NMR. *FEBS Lett.* **170**, 355–359 (1984).

25. D. Davidi, M. Shamshoum, Z. Guo, Y. M. Bar-On, N. Prywes, A. Oz, J. Jablonska, A. Flamholz, D. G. Wernick, N. Antonovsky, B. Pins, L. Shachar, D. Hochhauser, Y. Peleg, S. Albeck, I. Sharon, O. Mueller-Cajar, R. Milo, Highly active rubiscos discovered by systematic interrogation of natural sequence diversity. *EMBO J.*, doi: 10.15252/embj.2019104081 (2020).

26. H. Aigner, R. H. Wilson, A. Bracher, L. Calisse, J. Y. Bhat, F. U. Hartl, M. Hayer-Hartl, Plant RuBisCo assembly in E. coli with five chloroplast chaperones including BSD2. *Science* **358**, 1272–1278 (2017).

27. Y. M. Bar-On, R. Milo, The global mass and average rate of rubisco. *Proc. Natl. Acad. Sci. U. S. A.*, doi: 10.1073/pnas.1816654116 (2019).

28. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021).

29. L. S. Johnson, S. R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).

30. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, UniProt Consortium, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**,

926–932 (2015).

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

32. S. Gutteridge, G. Lorimer, J. Pierce, Details of the reactions catalysed by mutant forms of rubisco. *Plant Physiol. Biochem.* **26**, 675–682 (1988).