Mapping the biochemical landscape of rubisco

Authors

Noam Prywes[1,2], Naiya R. Philips[3], Luke M. Oltrogge[2,3], Benoit de Pins[4], Aidan E. Cowan[3,5], Leah J. Taylor-Kearney[6], Hana A. Chang[3], Laina N. Hall[7], Abhishek Bhatt[3,8], Patrick M. Shih[1,6,9,10], Ron Milo[4], 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]Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot 76100, Israel
[5]Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Emeryville, CA 94608, USA
[6]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA 94720, USA.
[7]Biophysics, University of California, Berkeley, Berkeley, CA 94720, USA.
[8]School of Medicine, University of California, San Diego, La Jolla, CA 92092, USA
[9]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.
[10]Feedstocks Division, Joint BioEnergy Institute, Emeryville, CA 94608, USA.

*Corresponding author, email: dsavage@berkeley.edu

Abstract

The enzyme rubisco catalyzes the first step of carbon assimilation in photosynthesis. Despite the massive flux of CO_2 passing through this active site over billions of years, extant rubisco has relatively slow kinetics and is prone to off-target activity. In many growth regimes, this limits photosynthesis *in planta.* Many attempts have been made to improve the kinetic parameters of rubisco with limited success, potentially due to biochemical trade-offs. To understand the structural basis of constraints on rubisco, a comprehensive map of rubisco at the individual amino-acid level is needed. To that end we performed a deep mutational scan using a rubisco-dependent *E. coli* strain. By titrating CO_2 concentrations it was possible to determine estimations for both catalytic rate (k_{cat}) and substrate affinity (K_M) of >99% of rubisco point mutants. Some positions were found to act as “rheostats” where some amino-acid substitutions reduced - while others improved - affinity for CO_2. No individual point mutation was found to substantially improve the catalytic rate, but a number of highly phylogenetically conserved positions were found to tolerate mutations, indicating that a large portion of rubisco’s sequence space remains unexplored by nature and may serve as a resource for future protein engineering efforts. Together, these biochemical measurements inform our understanding of biochemical tradeoffs and will assist in future efforts to improve rubisco catalytic properties.

Introduction

Every year ~100 gigatons of carbon (~⅕ of the mass of the biosphere[1]) are fixed by photoautotrophic organisms, including plants, through the most abundant enzyme on earth, ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco). The mechanism of rubisco catalysis was the subject of intensive research over several decades. Rubisco is a TIM barrel protein that, in its most simple form, assembles into a homodimer with two identical active sites along the dimer-dimer interface. In other organisms, this homodimeric unit assembles into higher order structures, most notably into the octameric form found in cyanobacteria, algae and plants[2]. Regardless of form, the rubisco active site remains constant and is notable for a carbamylated lysine residue and magnesium ion which are both essential for catalysis.

Rubisco catalysis, which is slow compared to many other central carbon metabolic enzymes[3], can limit carboxylation flux under certain conditions[4]. To make matters worse, rubisco also catalyzes an oxygenation side-reaction that requires organisms to operate a costly carbon rescue pathway known as

photorespiration. Photosynthetic engineering has been emerging as a possible tool in agricultural improvement[5–7] The role of rubisco engineering in this effort remains unclear[4,8], with some examples of modest improvements in plant growth both in the laboratory[9,10] and in the field[11]. While some species have faster or more specific rubisco variants, extremely fast rubisco isoforms have not been found or engineered.

Rubisco engineering has struggled to improve plant growth for two reasons - 1) heterologous expression of faster rubisco variants in plants typically results in low protein expression and therefore reduced carboxylation flux[12] and 2) rubisco variants across organisms have a narrow range of biochemical parameters when compared to other enzymes[13]. It has been proposed that rubisco kinetics are constrained by inescapable tradeoffs (e.g. between carboxylation rate and CO_2 affinity) resulting from the chemical mechanism at the active site[14,15], though the precise structural explanation for these tradeoffs has not been determined.

Despite these concerns, many attempts have been made to improve rubisco carboxylation rates and specificities against oxygenation by rational engineering or laboratory evolution, resulting in a handful of improvements[9]. In recognition of the need to assay more diverse rubisco variants across the tree of life, recent efforts have focused measurements on specific taxa such as diatoms [16]. Relatedly, a recent biochemical survey of bacterial rubiscos discovered a variant from a betaproteobacterium with the fastest catalytic rate observed to date[17]. A significant technical hurdle in engineering rubisco is the low throughput of obtaining accurate measurements for all four primary biochemical parameters: catalytic rates (k_{cat}s) for carboxylation and oxygenation as well as CO_2 and O_2 affinities.

Alternatively, a number of *in vivo* systems have been developed for high-throughput rubisco engineering and biochemical analysis in *E. coli*. Since *E. coli* is not naturally dependent on rubisco for growth, strains have been engineered such that rubisco carboxylation or oxygenation activity alleviates a metabolic stress caused by rubisco's sugar substrate, ribulose-1,5-bisphosphate (RuBP)[18]. More recently, rubisco-dependent *E. coli* strains have been developed that specifically require rubisco carboxylation activity for growth. In previous work both endogenous isoforms of ribose phosphate isomerase were knocked out of the *E. coli* genome resulting in the *Δrpi* strain (Fig. 1A). This strain grows at a normal rate in rich media but cannot grow when glycerol is provided as the only carbon source because ribulose-5-phosphate accumulates with no outlet[19]. The combined actions of phosphoribulokinase and rubisco convert this otherwise dead-end metabolite into 3-phosphoglycerate, which can feed back into central carbon metabolism.

Ideally, a system where growth depends on rubisco carboxylation could be used to systematically map the relationship between rubisco protein sequence and function. Previous studies have achieved such mappings for enzymes like GFP[20], DHFR[21] and β-Lactamases[22] through deep mutational scans (DMSs). A deep mutational scan (DMS) is the systematic assessment of many mutants of a protein simultaneously using a phenotypic screen followed by deep sequencing[23]. The ultimate goal of such studies is to generate a mapping from protein sequence space to functional space. Sequence space is the collection of all possible protein sequences, each of which maps to functional space which can be defined independently for any function. For instance, wild-type *R. rubrum* rubisco, the starting point for this study, has been previously characterized *in vitro* along several dimensions resulting in an estimate of catalytic rate (k_{cat}) of ~$7.5s^{-1}$, a substrate affinity of ~150μM CO_2, a specificity of carboxylation vs. oxygenation of ~10 in addition to over measurements. In a literature survey we found 77 single-mutants that had been characterized along one or more of those dimensions and 9 double mutants. Over 100 other Form II bacterial rubiscos haven been similarly characterized[17] but the distance in sequence space between *R. rubrum* and any of these other wild sequences can be more than 100 mutations (out of ~450 total amino-acids) and, as with all protein space, remains largely unexplored. It is impossible to measure even a small fraction of this space, but it is possible to learn, for any given protein family, rules about which mutations are acceptable, which will ablate function and even higher order interactions between different amino-acids in the structure. Typically, the first step in this process is the determination of protein function for single-mutants of a selected model protein. Here, using the

Δrpi strain we performed a deep mutational scan and quantitatively assessed the fitness of *E. coli* strains expressing all point mutants of the *R. rubrum* rubisco.

Results

Δrpi grows in proportion to rubisco catalytic rate

In order to provide estimates for rubisco k_{cat} *in vivo* it was necessary to demonstrate that the growth of *Δrpi* in glycerol media was quantitatively dependent on rubisco flux. By expressing rubisco under a Tac promoter we observed faster growth rates as IPTG levels were increased (Fig 1B). This indicated that increasing rubisco expression increased carboxylation flux in a way that translated into improved fitness. The growth rate reached a maximum after which it diminished slightly, which may suggest a load effect caused by overexpression of rubisco. When the same strain was grown at different levels of CO_2 (Fig 1C) growth rates increased with increasing [CO_2] (Fig 1C).

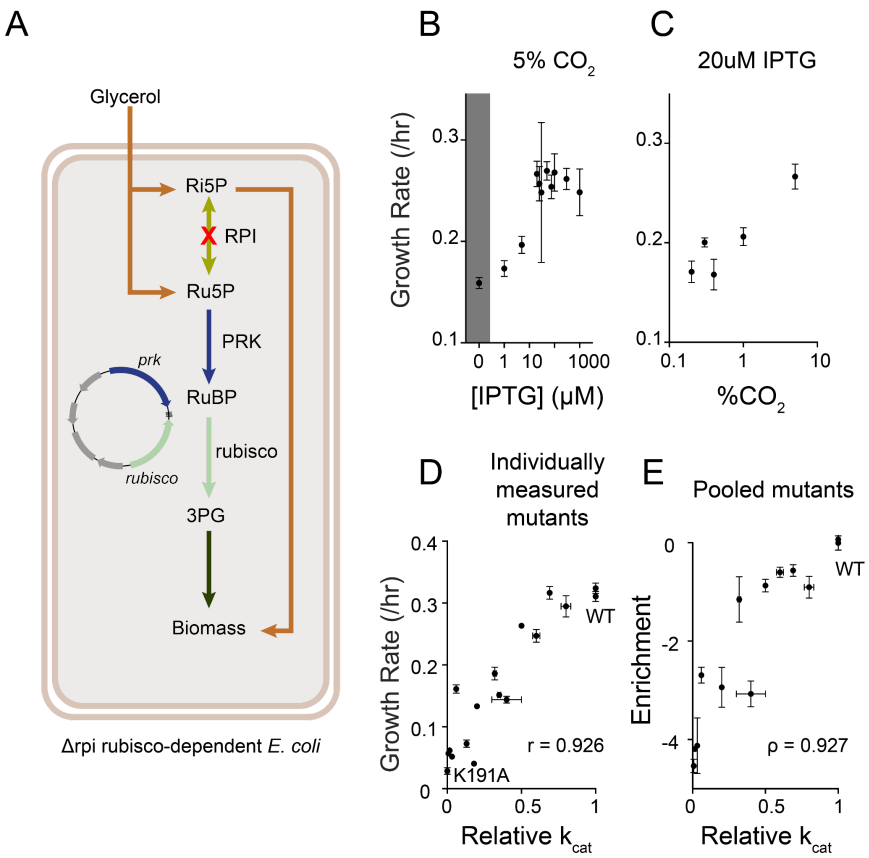

Figure 1: *Δrpi* is a rubisco-dependent *E. coli* strain with a growth rate that correlates to rubisco flux. A) Schematic of the *Δrpi* strain of rubisco-dependent *E. coli*. PRK and rubisco compensate for the deletion of RPI and rescue growth. **B,C)** Growth rates across a titration of rubisco induction by IPTG (**B**) and CO_2 (**C**) concentration. **D)** Individually measured growth rates of a set of mutants with previously-determined k_{cat}s. **E)** Mutant enrichments for the same mutants as in **D** measured in one nanopore sequencing experiment. Error bars in **D** and **E** determined as standard deviations of three or more replicates. Errors in literature values are shown from studies where they were reported.

We next assessed whether growth-base selection was correlated with biochemical behavior. Previous work on *R. rubrum* rubisco has identified 77 mutants that span from <1% to 100% of wild-type activity. Growth of these mutants was individually tested in *Δrpi* under selective conditions and found to correlate with reported catalytic rates (Fig. 1D,E). Together these results are consistent with a model wherein the growth of the strain is limited by rubisco V_{max}, which is determined by enzyme k_{cat}, [E] and [CO_2].In order to compare rubisco flux across a larger library of mutants we individually prepared a library of 39 plasmids containing 13 mutant rubisco genes each appended with 3 DNA barcodes. This pilot library was mixed and transformed into *Δrpi*. Nanopore sequencing was used to quantify the relative abundance of each barcode before and after overnight growth in M9 minimal media supplemented with 0.4% glycerol. Barcode enrichments were calculated relative to wild-type and correlated to the k_{cat} values from the literature (Fig. 1E). As with individually-measured growth rates, sequenced barcode enrichments correlated to known k_{cat}s for our panel of mutants. This provided the impetus to produce a scaled-up library including all possible single amino-acid mutations.

Individual fitness characterizations of rubisco single amino-acid mutations

We generated a library of all point mutants of the *R. rubrum* rubisco by cloning an set of oligo pools into plasmids containing rubisco with appropriate deletions to be filled by the oligos; each oligo was designed to contain exactly one mutation. The mutant library was then transferred to a selection plasmid containing *PRK* and barcodes were added. The library was plated onto agar and ~500,000 colonies were scraped for plasmid purification. The library was sequenced using long read sequencing in order to map barcodes to mutants and to verify that there were no mutations in the backbone or in PRK (Fig 2A). After filtering out barcodes with undesired mutations we were left with ~180,000 barcodes representing >99% of the designed library with at least 3 barcodes each and an average of 20 barcodes per mutation.

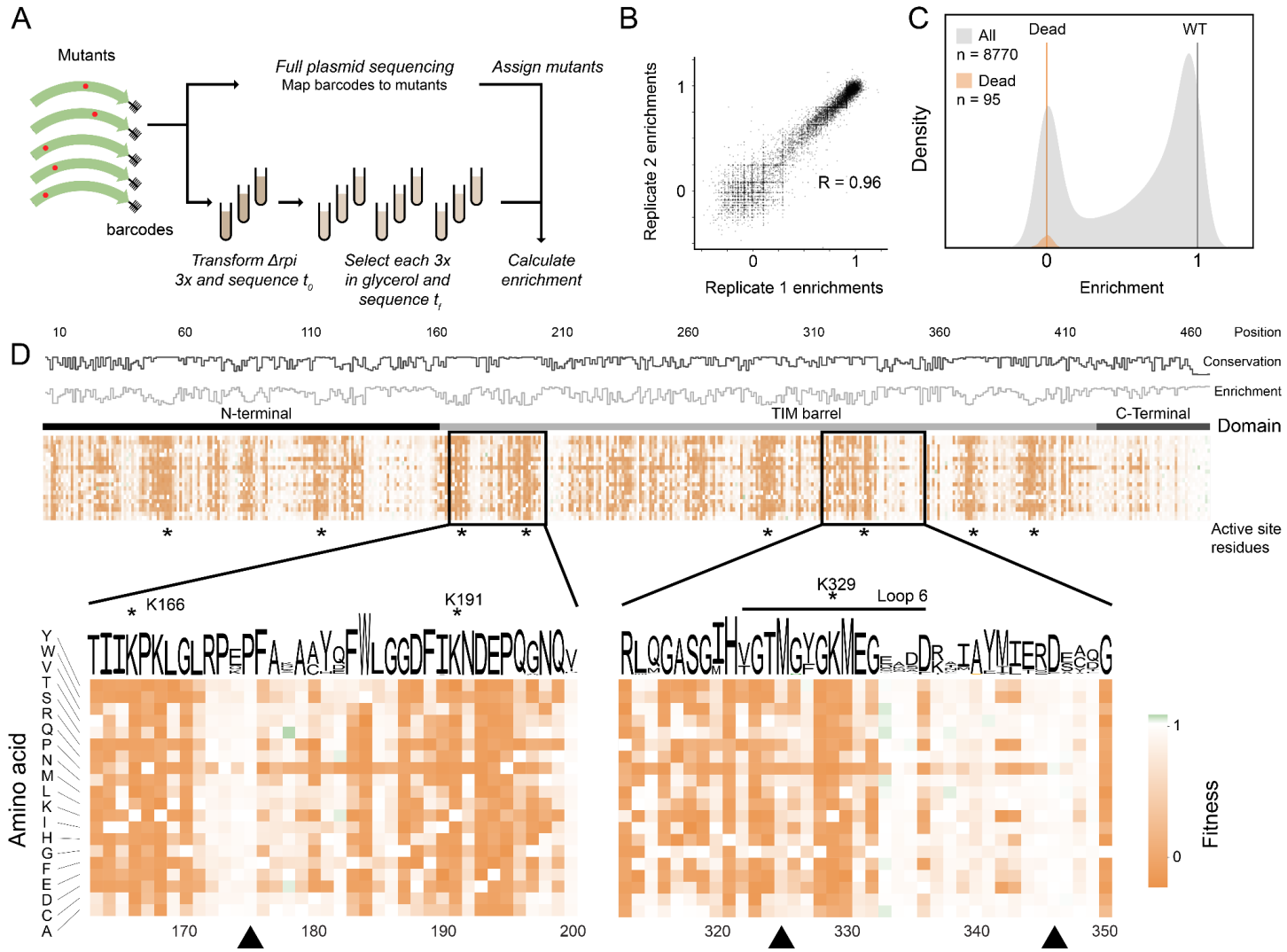

Figure 2: A deep mutational scan individually characterizes all single amino-acid mutations in rubisco A) Schematic of DMS experiment. **B)** Correspondence between 2 example biological replicates, each point represents the median enrichment among all barcodes for a given variant. **C)** Histogram of variant enrichments. Enrichments were normalized between values of 0 and 1 with 0 representing the average of enrichments of mutations at a panel of known active-site positions (orange) and 1 representing the average of WT barcodes. **D)** A heatmap of variant enrichments. Highly conserved but mutationally permissive positions indicated with black triangles at bottom.

The library was transformed into *Δrpi* to assess mutant fitness. Three independent library transformations were grown in selective conditions (minimal media supplemented with 0.4% glycerol) and grown for 6 to 7 divisions (24 hours) in an atmosphere of 5% CO_2 (equivalent to ~1500 μM CO_2 in solution; wild-type K_M = 150 uM). Short read sequencing was used to quantify barcode abundance before and after

selection and barcode enrichments were calculated by normalizing a panel of mutants known to completely eliminate catalytic function to 0 and all wild-type barcodes to 1. Nine replicate experiments were performed and the pairwise correlations between them had an average Pearson coefficient of 0.97 (Fig. 2B). As expected from DMS studies of other proteins, a bimodal distribution was observed (Fig. 2C) representing mutations clustered around wild-type (neutral mutations) and catalytically dead[20,24].

We measured fitness values for >99% of amino-acid substitutions and generated a heatmap of variant effects (Fig. 2D). Mutations to proline were more deleterious on average than any other amino acid (Fig. 2D). Mutations at known active-site positions showed strong signatures of depletion (e.g. K191, K166, K329, residues with asterisks). Very few mutations appeared more fit than WT, and when they did it was by a small amount. Phylogenetic conservation and average fitness at each position tended to anti-correlate as seen in previous studies[25,26].

Mutational scanning reveals mutationally insensitive regions of the structure

When colored by maximum enrichment at each position, the 3D structure of rubisco (PDB accession 9RUB) displays a strong sensitivity to mutation at each of the two active sites (Fig. 3A.B). Buried residues are also disproportionately sensitive to mutation (Fig. 3C). 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 (Fig. 3D). Loop 6 of the TIM barrel is known to fold over the active site during substrate binding and participate in catalysis. Surprisingly, many residues of this loop are insensitive to mutation, with the exception of the active-site residue K329 (Fig. 2E).

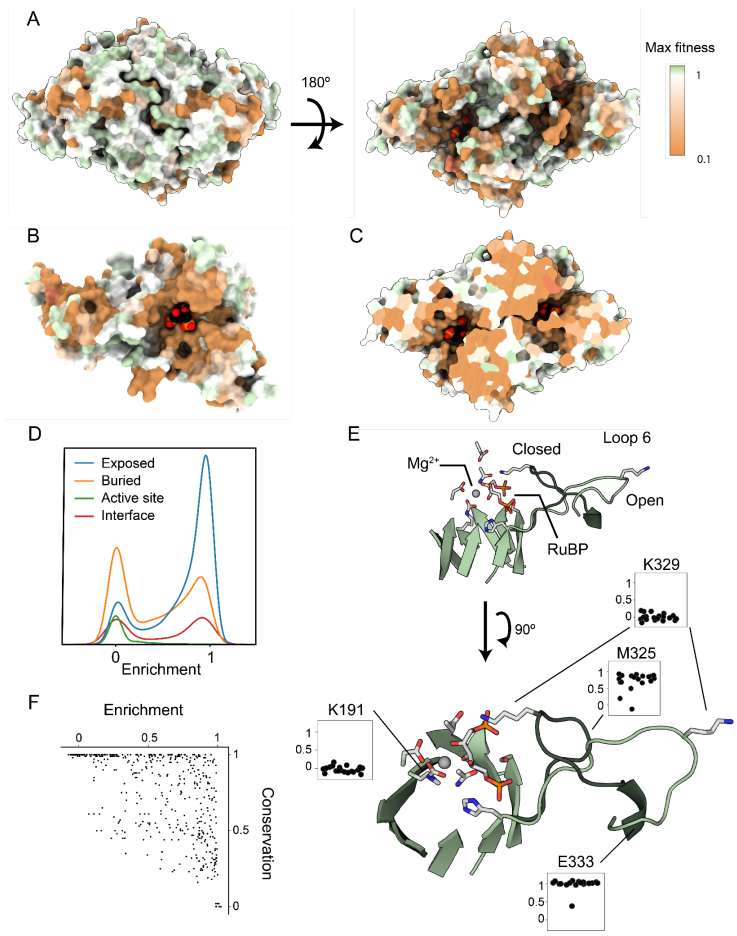

Figure 3: Fitness values provide structural, functional and evolutionary insights in rubisco A) Structure of *R. rubrum* rubisco homodimer (Protein Data Bank (PDB) ID: 9RUB) colored by the maximum enrichment value of a substitution at every site. **B)** One rubisco monomer with the substrate, RuBP, in one active site. **C)** A slice through the homodimer showing residues buried in the TIM-barrel core. **D)** Histograms of variant effects for amino-acids in different parts of the homodimer complex. **E)** Close-up view of the active-site and the mobile loop 6 region. **F)** Comparison of average enrichment at each position against phylogenetic conservation among Form II bacterial rubiscos.

We expected that highly conserved positions would not tolerate mutations well and therefore have low average enrichment values. The average enrichment value at each position generally showed anticorrelation with sequence conservation by position calculated from a sequence alignment of ~1000 bacterial rubiscos (Fig. 2D top track and 3F). There were, however, many outliers with a number of positions being highly conserved yet showing high mutational tolerance (Fig. 3F top right corner, Fig. 2D black triangles). It is not clear what selective forces have enforced such high conservation at those positions while other positions that are also not required for catalysis have experienced substantial drift; bacterial rubiscos can be found across the full diversity of bacterial phylogeny. The conserved but robust positions are not clustered in the structure in an

obvious pattern. In contrast to other proteins[25,26], we did not find any positions with low conservation and low enrichment which could have implied a specific function unique to *R. rubrum*.

Enzyme activity and affinity can be inferred by substrate titration

We performed selections in different CO_2 concentrations in order to determine if substrate affinity could also be measured in high throughput (Fig. 4A). Enrichments were overall higher with increasing [CO_2] (Fig. 4B) indicating that some mutants appear functional only at higher [CO_2]. We observed that while many mutants had a constant relationship between their enrichment values and [CO_2], others appeared to vary (Fig. 4C). The data were fit under the assumption that a straightforward relationship existed between the catalytic activity and substrate affinity of each mutant and its variable enrichment values with increasing substrate concentration as shown previously for β-Lactamases[22]. By fitting the data to this assumption we could generate inferred Michalis-Menten titrations (Fig. 4D). From these inferences we generated V_{max} and K_M fitting estimates for every mutant (Fig. 4J).

We compared these biochemical parameters against 78 values taken from the literature and measured in this study (Fig. 4F-I). For V_{max} we observed a linear relationship (Fig. 4F) after WT-normalization - with a single outlier (I190T). For further confirmation we purified 35 mutants selected to vary across the range of relative V_{max} estimates and measured their k_{cat} values *in vitro* (Fig. 4G). These measurements confirmed the trend observed with the data from the literature; most of the points showed a correlation with our estimates and I190T was still an outlier, though we found that its k_{cat} was lower than that previously reported.

Only 12 measurements of CO_2 affinity (K_M) for various mutants have been reported in the literature but it was possible to observe a positive correlation between our calculated K_M values and the *in vitro*-measured K_Ms (Fig. 4H). This correlation was not as straightforward and linear as that for V_{max} so we purified an additional set of 7 mutants for validation, once again chosen to span a range of predicted values. By measuring rubisco activity in a titration of dissolved CO_2 concentrations (Fig. 4E) we produced a set of K_M values that correlated to the K_M estimates.

In surveys of rubisco biochemical measurements, a trade-off has been repeatedly proposed between the catalytic rate and affinity for CO_2; this trade-off would mean that improvements in catalytic rate would come at the cost of reduced CO_2 affinity, and vice versa[13–15,27]. Thus far measurements of natural rubisco sequences have been used to determine if there exists an optimal ratio or k_{cat} to K_M that different species evolve towards. In this study we only examine single-mutants and do not select fitter mutants from less fit ones, therefore the measurements in this study do not represent optimized enzymes. Still, it is possible to plot all inferred V_{max}s against all K_Ms (Fig. 4K). The relationship that emerges is a negative correlation when rates approach that of wild-type. This correlation suggests that mutants with sub-WT rates are also less capable of binding CO_2. As there is no binding site for CO_2 in the enzyme[28], this trend may be related to subtle changes in the geometry of the bound sugar substrate within the active site before bond-formation with CO_2. The positive correlation observed between V_{max} and K_M at low V_{max} values is likely artifactual due to poor unstable fitting. K_M values with high coefficients of variation are shown partially transparent (see methods).

One of the surprising observations in this analysis was the presence of a number of "rheostat" positions, where different mutations had a variety of different effects on the inferred K_M. Some positions, like V266, had substitutions that substantially improved affinity (e.g. V266T) while other substitutions reduced it (V266G) (Fig. 4C-E). V266 is not especially close to the active site and sits near the C2 axis of the rubisco homodimer at the interface. What role V266 plays in CO_2 entry into the active site, active site geometrical rearrangement or electrostatics remains unclear.

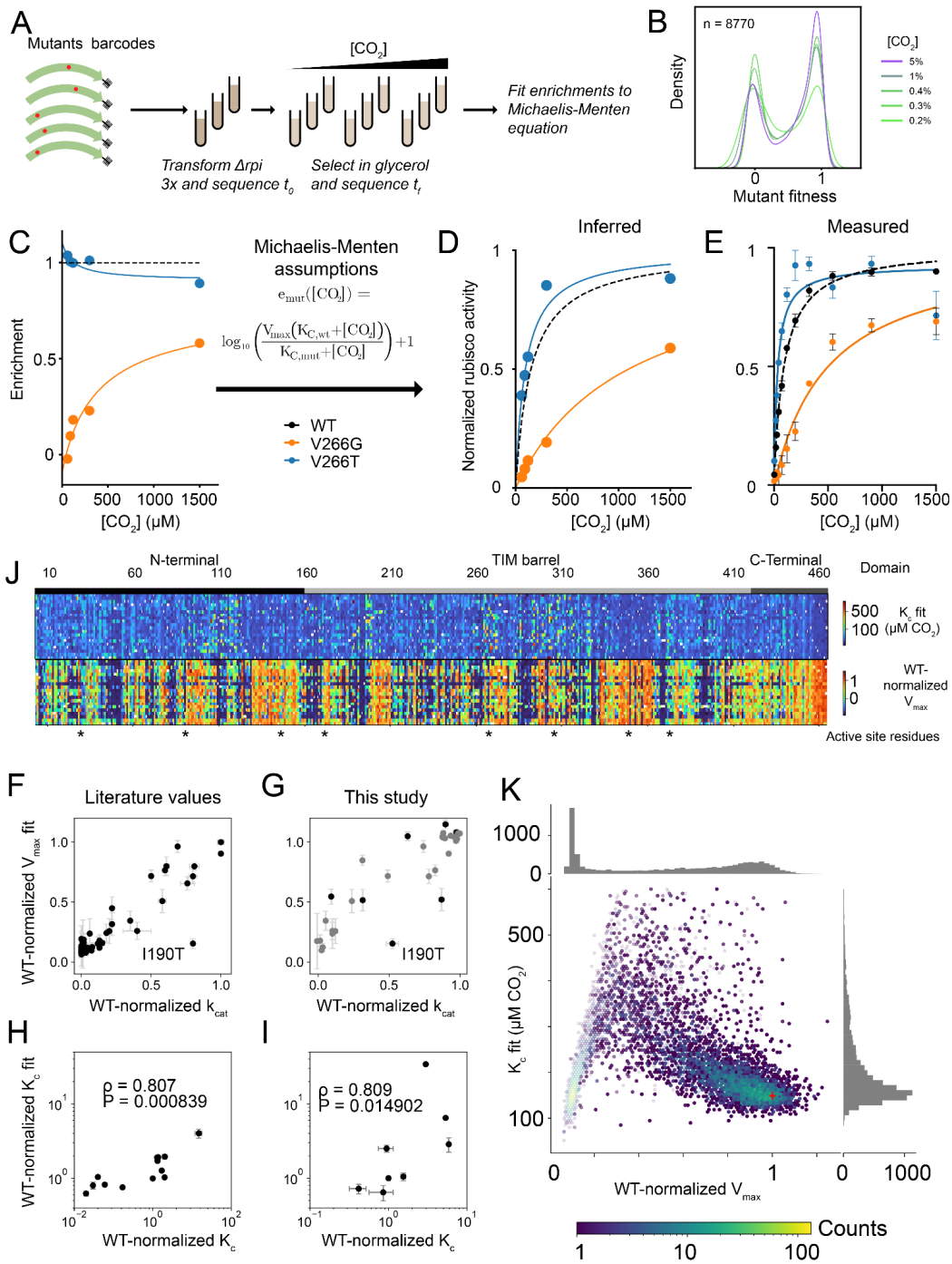

Figure 4: K_M and V_{max} can be inferred from enrichments across a CO_2 titration A) Schematic of rubisco selection in [CO_2] titration. **B)** Histograms of variant enrichments at different [CO_2]. **C)** Measured enrichments at different [CO_2] for two mutants followed by the fit equation used to extract K_M and V_{max} values. **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-I)** Comparisons between biochemically measured rubisco kinetic parameters and those same parameters as inferred from enrichment values. WT-normalized V_{max} values in **F** and **G**, K_M values in **H** and **I**. Measurements from the literature in **F** and **H**, values measured in this study in **G** and **I**. Black points in **G** were purified 3 independent times, all other data are from individual purifications. **J)** Heatmaps for K_M and V_{max} values for all mutants. **K)** Two dimensional histogram of Michaelis-Menten fits from **J** with hexagonal bins and a log-scale colormap. The transparency of each bin depicts an aggregate measure of fit confidence and is derived from the mean coefficient of variation of the K_M for all bin members. Mean cov ≤ 0.1 is fully opaque while mean cov > 0.5 is 90% transparent. WT values are shown as a red cross.

Discussion

In this study we have demonstrated that an engineered *E. coli* strain, *Δrpi*, can provide a readout of rubisco kinetic parameters and we have produced estimates for V_{max} and K_M for nearly all single amino-acid substitutions of one enzyme. One limitation of this study is that we have not disentangled V_{max} into its two components, enzyme concentration and k_{cat}. Other studies have found ways to deconvolute similar parameters in different proteins[29] and a similar approach may be possible here, perhaps by repeating the selection at different inducer concentrations, at different temperatures or else by co-expressing a folding chaperone to quantify the effects of protein misfolding.

Rubisco has 4 biochemical parameters of interest, K_M and k_{cat} for carboxylation, which we have studied here, and the same values for the oxygenation side reaction. In this study we did not attempt to control O_2

levels, but it is likely that this study could be repeated with variable O_2 in order to measure those parameters, along with the derived parameter of specificity ($S_{C/O} = k_{cat,C}K_O/k_{cat,O}K_C$). Since O_2 levels have pleiotropic effects it would be critical to control growth carefully and to provide an alternative electron acceptor at low O_2 levels.

Further libraries with multiple mutations can help to elucidate allosteric networks within the enzyme[30] and libraries that start with rubiscos from other species will help to determine the relative evolvabilities of various isoforms[20]. Libraries of form I rubiscos from plants are especially interesting due to their relevance to photosynthetic engineering in agriculture. Due to differential expression between rubiscos from different species in *E. coli* it is unlikely that comparisons between highly divergent sequences will report on biochemical differences and will likely reflect differences in intracellular enzyme concentrations. This effect could be corrected for through an independent measurement of rubisco abundance through, for example, a fluorescent protein fusion and fluorescence activated cell sorting.

The existence of positions that can have highly variable effects on K_M can be explained in a number of ways. Rubisco does not have a standard Michaelis complex like many other enzymes, CO_2 isn't bound before it reacts with RuBP. Rather, CO_2 either forms a covalent bond with RuBP upon colliding with the enzyme or it fails to do so and diffuses away. "Rheostat" residues could affect the K_M by modulating active site electrostatics, affecting the entry of CO_2 or altering the active site geometry. Further analysis of these positions is necessary.

The existence of highly conserved but mutationally permissive positions seems to imply one of two explanations. Either those positions are not permissive in the native hosts due, for example, to some sort of protein-protein interaction, or rubisco evolution is uneven in pace between positions. Since rubisco is conserved across a wide range of bacterial genomes, neither of these explanations is especially satisfying.

The relationship between V_{max} and K_M has critical implications for the overall permitted functional space of rubisco. Recent meta-analyses of rubisco kinetics have found that these two parameters do not clearly correlate in a simple trade-off[13,27]. Here we do not observe any trade-off, but that is for the simple reason that we have not applied any selection against non-optimal sequences. Further studies could select for better-performing variants and then map the V_{max} and K_M of those variants to more comprehensively trace out the limits of rubisco biochemical space.

While a small subset of mutants appeared to have higher V_{max} values than WT in this study, we were unable to validate any of those results *in vitro*. It may be the case that *R. rubrum* has a rubisco that sits at an extremely narrow fitness optimum, though additional validation of potentially improved mutations is necessary to confirm that result. Future mutant libraries will determine how rare improved variants are in sequence space.

Materials and Methods

Strains

Cloning was performed in a combination of Turbo cells, top10 cells and NEB DH5α. For protein expression we used 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 (in order to make the strain rubisco-dependent when grown on gluconate, this was not a relevant feature in this study).

Plasmids

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

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

Library design and construction

The *R. rubrum* rubisco sequence codon-optimized for *E. coli* was systematically mutated. The rubisco gene from *R. rubrum* was codon optimized and split into 11 pieces. For each of those pieces (~200 bp each) all point mutants were designed as oligos. These 11 oligo pools containing all single mutants within their respective ~200bp region were purchased from TWIST Bioscience™ and each sub-library was amplified individually using Kapa Hifi polymerase with a cycle number of 15 to reduce PCR bias. Each rubisco gene fragment was inserted into a corresponding PCR-linearized pUC19 destination vector containing the codon optimized, WT sequence of rubisco flanking the insert using golden gate assembly. This assembly generated 11 sub libraries of the full-length *R. rubrum* rubisco gene with each library containing a ~200 bp region including all single mutants. Each of these 11 rubisco libraries were separately transformed into *E. coli* top10 cells; >10,000 transformants were scraped from agar plates to ensure adequate library coverage. Plasmids were purified from each sub-library and mixed together at equal molar ratios to generate a library of all single mutants of rubisco.

In order to produce the final library for selection, the selection plasmid containing induction systems for rubisco and PRK (Tac and Tet respectively) was amplified with primers that included a random 30 nt barcode. The linearized plasmid and the library were cut with BsaI and BsmBI respectively, ligated together and transformed into top10 cells. Plasmid was purified from a scrape of ~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 glycerol stocks.

Library characterization and screening

Selections were performed by diluting 200uL glycerol stocks with ODs of 0.25-0.3 into 5mL of M9 minimal media with added chloramphenicol, thiamine, glycerol (0.4%) and 20nM anhydrotetracycline. These cultures were grown in different IPTG and CO2 concentrations until they reached an OD at 5 mL of 1.2. This corresponds to a 100-fold expansion of the cells or between 6 and 7 doublings.

Cultures before and after selection were spun down and plasmid samples were extracted. 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 jth variant and the individual barcodes, *i*, passing the threshold condition the variant enrichment is calculated as,

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

To select optimal values for these parameters we computed the variant enrichments across a 2D array and found the combination that resulted in the maximum mean Pearson correlation coefficient across all condition replicates; these were c_{min} = 5 and α_p = 3.65e-7 (average of 0.3 pseudocounts) leading to a correlation coefficient of 0.978.

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 jth 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 fit each CO_2 titration enrichments to the log ratio of Michaelis-Menten velocities.

Eq. 3 $$e_{mut}([CO_2]) = \left(\frac{V_{max}(K_{M,wt} + [CO_2])}{K_{M,mut} + [CO_2]}\right) + 1$$

V_{max} is the ratio of mutant maximum velocity relative to wild-type, $K_{M,wt}$ is the wild-type K_M for which we used the value 149 μM, and $K_{M,\,mut}$ is the mutant K_M. The added factor of 1 is to account for the enrichment normalization since $\log_{10}(1) = 0$ for wild-type. 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 $K_{M,\,mut}$ to be positive.

We noted that the K_M fits to certain variants–particularly ones with low 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 K_M's thus obtained. 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 and Michaelis-Menten fits for all combinations of these parameters. From this set of K_M fit values for each variant we computed a coefficient of variation that was used as a figure of merit for the K_M.

Protein purification

E. coli BL21(DE3) cells were transformed with pET28 with a His and SUMO tag containing the desired rubisco and pGro plasmids. Colonies were grown at 37°C, shaken at 200 RPM, 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 16°C, 200 RPM, for 30 minutes. Protein expression was induced with IPTG (100uM) and cells were grown overnight at 16°C. Cultures were spun down (15 min; 4,000 g; 4°C) and purified as reported[17]. Briefly, cultures were spun down and lysed using BPER-II™. Lysates were centrifuged to remove insoluble fraction. His-tag purification using Ni-NTA beads was performed and pure rubisco was eluted by SUMO tag cleavage with bdSUMO protease. Purified proteins were separated from the Ni beads, 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

Measuring k_{cat}

The carboxylation rate (k_{cat}) of each rubisco was measured using methods established previously[17]. Briefly, rubisco (800 nM) was activated by incubation with CO_2 (4%) and O_2 (0.2%) and added to aliquots of assay mix containing different CABP concentrations pre-equilabrated in 4% CO_2 and 0.2% O_2. After inhibition, RuBP was added to samples and the absorbance was measured (A340). NADH oxidation rates were calculated using A_{340}. This was coupled to rubisco activity by phosphosphoglucokinase (pgk) and

glyceraldehyde 3-phosphate dehydrogenase[31] assuming a 2:1 ratio of NADH oxidation rate to carboxylation rate. Absorbance over time gives a rate of NADH oxidation and therefore a carboxylation rate. 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.

Measuring K_M

Purified rubisco mutants were activated (400 mM bicarbonate and 1 M MgCl2) and added to a 96-well plate along with assay mix (100 mM HEPES pH 8, 20 mM $MgCl_2$, 0.5 mM DTT, 2 mM ATP, 10 mM creatine phosphate, 0.5mM NADH, 1mM EDTA and 20U/mL each of pgk, gapdh and creatine phosphokinase). 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-equilabrated at 0.3%O2 and 0%CO2. 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 A_{340} as in the k_{cat} assay. Plots of absorbance over time were fitted in Prism™ using the Michaelis-Menten fit and determined K_M. The K_M was measured in triplicate for each protein.

Acknowledgements

We thank Avi Flamholz and Niv Antonovsky 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 David Ding, Philip Romero, Nat Thompson, Leon Fedotov, Orren Saltzman, Eden Prywes, Stacia Wyman 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 running our NGS samples. We would like to thank Elaine Meng for assistance using ChimeraX™ in displaying the results of our screen on the rubisco structure. Finally we thank Flora Wang for technical assistance over the weekends.

References

1. Bar-On YM, Phillips R, Milo R. The biomass distribution on Earth. Proc Natl Acad Sci U S A. 2018;115: 6506–6511.

2. Prywes N, Phillips NR, Tuck OT, Valentin-Alvarado LE, Savage DF. Rubisco Function, Evolution, and Engineering. Annu. Rev. Biochem. 2023. pp. 385–410. doi:10.1146/annurev-biochem-040320-101244

3. Bar-Even A, Milo R, Noor E, Tawfik DS. The Moderately Efficient Enzyme: Futile Encounters and Enzyme Floppiness. Biochemistry. 2015;54: 4969–4977.

4. Wu A, Brider J, Busch FA, Chen M, Chenu K, Clarke VC, et al. A cross-scale analysis to understand and quantify the effects of photosynthetic enhancement on crop growth and yield across environments. Plant Cell Environ. 2023;46: 23–44.

5. De Souza AP, Burgess SJ, Doran L, Hansen J, Manukyan L, Maryn N, et al. Soybean photosynthesis and crop yield are improved by accelerating recovery from photoprotection. Science. 2022;377: 851–854.

6. Kromdijk J, Głowacka K, Leonelli L, Gabilly ST, Iwai M, Niyogi KK, et al. Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. Science. 2016;354: 857–861.

7. South PF, Cavanagh AP, Liu HW, Ort DR. Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field. Science. 2019;363. doi:10.1126/science.aat9077

8. Iñiguez C, Aguiló-Nicolau P, Galmés J. Improving photosynthesis through the enhancement of Rubisco carboxylation capacity. Biochem Soc Trans. 2021;49: 2007–2019.

9. Wilson RH, Alonso H, Whitney SM. Evolving Methanococcoides burtonii archaeal Rubisco for improved photosynthesis and plant growth. Sci Rep. 2016;6: 22284.

10. Salesse-Smith CE, Sharwood RE, Busch FA, Kromdijk J, Bardal V, Stern DB. Overexpression of Rubisco subunits with RAF1 increases Rubisco content in maize. Nat Plants. 2018;4: 802–810.

11. Yoon D-K, Ishiyama K, Suganami M, Tazoe Y, Watanabe M, Imaruoka S, et al. Transgenic rice overproducing Rubisco exhibits increased yields with improved nitrogen-use efficiency in an experimental paddy field. Nature Food. 2020;1: 134–139.

12. Lin MT, Occhialini A, Andralojc PJ, Parry MAJ, Hanson MR. A faster Rubisco with potential to increase photosynthesis in crops. Nature. 2014;513: 547–550.

13. Flamholz AI, Prywes N, Moran U, Davidi D, Bar-On YM, Oltrogge LM, et al. Revisiting Trade-offs between Rubisco Kinetic Parameters. Biochemistry. 2019;58: 3365–3376.

14. Tcherkez GGB, Farquhar GD, Andrews TJ. Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. Proc Natl Acad Sci U S A. 2006;103: 7246–7251.

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

16. Young JN, Heureux AMC, Sharwood RE, Rickaby REM, Morel FMM, Whitney SM. Large variation in the Rubisco kinetics of diatoms reveals diversity among their carbon-concentrating mechanisms. J Exp Bot. 2016;67: 3445–3456.

17. Davidi D, Shamshoum M, Guo Z, Bar‐On YM, Prywes N, Oz A, et al. Highly active rubiscos discovered by systematic interrogation of natural sequence diversity. EMBO J. 2020. doi:10.15252/embj.2019104081

18. Zhou Y, Whitney S. Directed Evolution of an Improved Rubisco; In Vitro Analyses to Decipher Fact from Fiction. Int J Mol Sci. 2019;20. doi:10.3390/ijms20205019

19. Flamholz AI, Dugan E, Blikstad C, Gleizer S, Ben-Nissan R, Amram S, et al. Functional reconstitution of a bacterial CO2 concentrating mechanism in Escherichia coli. Elife. 2020;9. doi:10.7554/eLife.59882

20. Gonzalez Somermeyer L, Fleiss A, Mishin AS, Bozhanova NG, Igolkina AA, Meiler J, et al. Heterogeneity of the GFP fitness landscape and data-driven protein design. Elife. 2022;11. doi:10.7554/eLife.75842

21. Thompson S, Zhang Y, Ingle C, Reynolds KA, Kortemme T. Altered expression of a quality control protease in E. coli reshapes the in vivo mutational landscape of a model enzyme. Elife. 2020;9. doi:10.7554/eLife.53476

22. Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a function of purifying selection in TEM-1 β-lactamase. Cell. 2015;160: 882–892.

23. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11: 801–807.

24. Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, et al. Local fitness landscape of the green fluorescent protein. Nature. 2016;533: 397–401.

25. Jones EM, Lubock NB, Venkatakrishnan AJ, Wang J, Tseng AM, Paggi JM, et al. Structural and functional

characterization of G protein-coupled receptors with deep mutational scanning. Elife. 2020;9. doi:10.7554/eLife.54895

26. Subramanian S, Gorday K, Marcus K, Orellana MR, Ren P, Luo XR, et al. Allosteric communication in DNA polymerase clamp loaders relies on a critical hydrogen-bonded junction. Elife. 2021;10. doi:10.7554/eLife.66181

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

28. Gutteridge S, Parry MAJ, Schmidt CNG, Feeney J. An investigation of ribulosebisphosphate carboxylase activity by high resolution1H NMR. FEBS Lett. 1984;170: 355–359.

29. Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. Nature. 2022;604: 175–183.

30. McCormick JW, Russo MA, Thompson S, Blevins A, Reynolds KA. Structurally distributed surface sites tune allosteric regulation. Elife. 2021;10. doi:10.7554/eLife.68346

31. Kubien DS, Brown CM, Kane HJ. Quantifying the amount and activity of Rubisco in leaves. Methods Mol Biol. 2011;684: 349–362.