# A Protein Blueprint of the Diatom $CO_2$-Fixing Organelle

Onyou Nam[1], Caroline McKenzie[1], Adam Dowle[2], Matthew Dowson[1], James Barrett[1], Luke C. M. Mackinder[1]

[1]Centre for Novel Agricultural Products, Department of Biology, University of York, York YO10 5DD, United Kingdom
[2]Department of Biology, University of York, York YO10 5DD, United Kingdom

Author for Correspondence: luke.mackinder@york.ac.uk

## Abstract:

Diatoms are central to the global carbon cycle. At the heart of diatom carbon fixation is an overlooked organelle called the pyrenoid, where concentrated $CO_2$ is delivered to densely packed Rubisco. Diatom pyrenoids fix approximately one-fifth of global $CO_2$ but virtually nothing is known about this organelle in diatoms. Using large-scale fluorescence protein tagging and affinity purification-mass spectrometry we generate a high-confidence spatially-defined protein-protein interaction network for the diatom pyrenoid. Within our pyrenoid interactome are 10 proteins with no known function. Six form a static shell encapsulating the Rubisco matrix of the pyrenoid, with the shell critical for pyrenoid structural integrity and potentially acting as a $CO_2$ diffusion barrier. Although no conservation at a sequence level, the diatom pyrenoid shares some structural similarities to prokaryotic carboxysomes. Collectively, our results support the convergent evolution of pyrenoids across the two main plastid lineages and uncovers a major structural and functional component of global $CO_2$ fixation.

## Introduction

Approximately one-third of global carbon fixation takes place in pyrenoids[1]. Pyrenoids are biomolecular condensates of the principle $CO_2$-fixing enzyme Rubisco found in the chloroplasts of algae.[2] There are two major chloroplast lineages, the green and red plastids with algae within these lineages proposed to have convergently evolved pyrenoids.[3,4] Nearly all of our knowledge of pyrenoid function and composition comes from the model terrestrial green plastid alga *Chlamydomonas reinhardtii.*[2,5] *C. reinhardtii* is a powerful model organism[6] but global carbon fixation is primarily driven by oceanic red plastid containing algae, such as diatoms, where our knowledge is still in its infancy.[7–9] Diatoms are responsible for up to 20% of global net primary production in the modern ocean, are estimated to fix ~70 gigatons $CO_2$ $yr^{-1}$, and are fundamental for the long-term storage of carbon by driving the flux of organic material from the ocean surface to sediments.[10,11]

Pyrenoids are found at the heart of algal $CO_2$ concentrating mechanisms (CCMs).[2,7] CCMs overcome the slow diffusion of $CO_2$ in water and the catalytic limitations of Rubisco by actively pumping inorganic carbon from the external environment into the cell and releasing it as $CO_2$ in the pyrenoid where it can be fixed by tightly packaged Rubisco. In *C. reinhardtii* a disordered linker protein, EPYC1, condenses Rubisco to form the liquid-liquid phase separated matrix of the pyrenoid.[1,12,13] A shared Rubisco binding motif found in EPYC1 and

numerous other pyrenoid components enables targeting to and structural organization of the pyrenoid. With specialized proteins containing this motif proposed to link the matrix to specialized traversing thylakoids called pyrenoid tubules and link the matrix to the surrounding starch sheath.[14] Inorganic carbon in the form of $HCO_3^-$ is shuttled into the pyrenoid tubules by bestrophin-like proteins[15] where a carbonic anhydrase converts it to membrane permeable $CO_2$ that can leak out and be fixed by Rubisco in the pyrenoid matrix.[5,15,16] The surrounding starch acts as a diffusion barrier to minimize $CO_2$ leakage out of the pyrenoid.[17,18] Fluorescent protein tagging,[14,19] affinity purification followed by mass spectrometry (APMS)[19] and proximity labeling[20] have enabled a high-confidence pyrenoid proteome to be determined with multiple components now functionally characterized. This has enabled *C. reinhardtii* proto-pyrenoid engineering in plants[21] and a parts-list of components that should theoretically enable the engineering of a functional pyrenoid-based CCM to enhance plant photosynthesis.[17,22]

Diatom pyrenoids have some shared features with the *C. reinhardtii* pyrenoid including condensed Rubisco and traversing thylakoids, it is also proposed that they may function analogously with some, although limited, conservation of inorganic carbon delivery proteins.[8] However, it is still unclear the level of conservation of structural proteins.[7,8] The only structural component identified so far for the diatom pyrenoid is PYCO1 a Rubisco linker protein found in the pyrenoid of the pennate diatom *Phaeodactylum tricornutum* and suggested to be responsible for phase separating Rubisco to form the pyrenoid matrix.[9] However, PYCO1 is not widely conserved, being absent in centric diatoms, and its functional importance is yet to be determined. Outside of PYCO1 most previous diatom CCM research has focused on inorganic carbon uptake. In *P. tricornutum* several candidates belonging to the SLC4 family of transporters have been proposed for $HCO_3^-$ uptake at the plasma and chloroplast membranes.[23,24] Bestrophin-like proteins are also implicated in the shuttling of $HCO_3^-$ into the thylakoid lumen where a θ-type carbonic anhydrase has been identified that is thought to function by releasing $CO_2$ from the thylakoid membranes that traverse the pyrenoid matrix.[25,26] Less is known about inorganic carbon uptake in the centric diatom *Thalassiosira pseudonana*. SLC4 candidates have been implicated in inorganic carbon uptake across both the plasma and thylakoid membranes.[8,27] Recently two bestrophin-like proteins were localized to the *T. pseudonana* pyrenoid, likely in the pyrenoid penetrating thylakoid (PPT)[28,29] and a θ-type carbonic anhydrase 2 confirmed to be located in the PPT.[8] Although, supporting functional data is missing for these proteins. Absent from diatom pyrenoids is a starch sheath encapsulating the pyrenoid. In *C. reinhardtii* flux balance modeling of pyrenoid function[17] and analysis of starch mutants[18] indicate that a diffusion barrier is essential for efficient pyrenoid function to minimize $CO_2$ leakage. How the diatom pyrenoid minimizes $CO_2$ leakage is a substantial outstanding question.

Here we rapidly advance our knowledge of the diatom pyrenoid by developing an iterative fluorescent protein tagging followed by APMS approach in *T. pseudonana* that belongs to a global biogeochemically important genus. These data enabled us to build a high-confidence diatom pyrenoid interactome, identifying multiple new pyrenoid proteins, many with no previously known functional domains. A family of these proteins form a static shell that encapsulates the Rubisco matrix and are critical for pyrenoid structural integrity and potentially act as a $CO_2$ diffusion barrier. Our findings give novel insight into a global biogeochemically important organelle and provide additional molecular parts for engineering a CCM into crop plants to improve productivity.

## Results and discussion:

**Rubisco co-immunoprecipitation mass spectrometry to identify diatom pyrenoid components**

Although diatoms play a central role in global biogeochemical cycles, very little is known about the diatom pyrenoid. In *T. pseudonana* cells have two chloroplasts surrounded by 4 membranes, with each chloroplast containing a single centrally positioned lenticular-shaped pyrenoid (Fig. 1A). As a starting point to understand *T. pseudonana* pyrenoid composition, we performed co-immunoprecipitation coupled with mass spectrometry (coIPMS) with the main pyrenoid component, Rubisco, as a bait protein (Fig. 1B). To capture Rubisco, we used an antibody raised to a conserved 12 amino acid surface-exposed peptide on the Rubisco large subunit (rbcL) (Fig. S1). By comparing two independent coIPMS experiments, consisting of 2 and 3 technical replicates respectively, against non-antibody control experiments, we identified 36 putative pyrenoid components out of a total of 1167 number of detected proteins identified with two or more spectral counts (Table S1). For these exploratory experiments, we applied a relaxed cut-off based on the fold-change enrichment of bestrophin-like protein 2 (BST2) that we had previously localized to the pyrenoid[28] and would expect to only have a weak enrichment due to it predicted to be membrane bound (Fig. 1B). Top hits from our rbcL coIPMS experiments were then fed into an iterative tagging, localization, and APMS framework to rapidly build a spatially defined pyrenoid proteome (Fig. 1C).

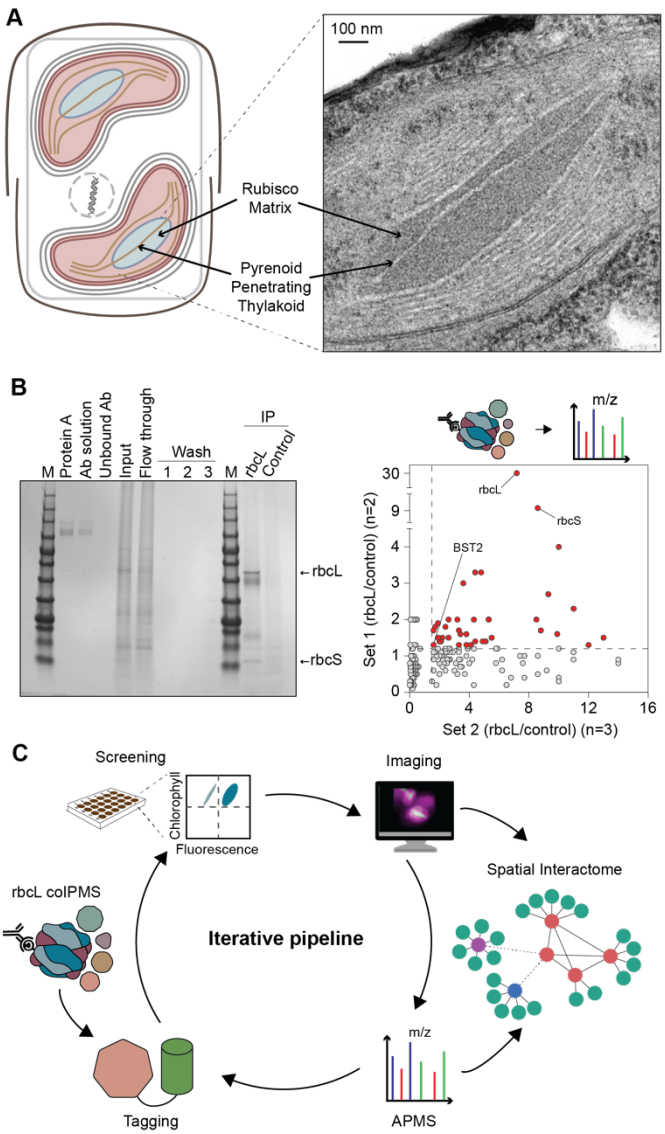



**Figure 1. Identification of candidate diatom pyrenoid components.**

A. The diatom *T. pseudonana* has two chloroplasts each containing a single pyrenoid traversed by a specialized membrane continuous with the thylakoid membranes called the pyrenoid penetrating thylakoid (PPT).

B. Rubisco co-immunoprecipitation followed by mass spectrometry (coIPMS) to identify candidate pyrenoid components. Left: example Rubisco large subunit (rbcL) coIP with different steps ran on an SDS-PAGE stained with coomassie. Right: Two separate experiments each consisting of either two or three technical coIPs we performed and the antibody bound fractions submitted for mass spectrometry. Cut-offs (dashed lines) were determined based on bestrophin-like protein 2 (BST2) that was previously shown to be pyrenoid localized.

C. An iterative pipeline to determine the pyrenoid spatial interactome. Rubisco coIP data fed into a high-throughput fluorescence protein tagging pipeline. Pyrenoid confirmed proteins were then used for affinity purification followed by mass spectrometry (APMS). Interactors were tagged to confirm pyrenoid localization and used in subsequent APMS rounds.

## Development of a high-throughput tagging pipeline in *T. pseudonana* identifies multiple novel pyrenoid components

To enable rapid cycling through our iterative pipeline, we set out to establish high-throughput fluorescent protein tagging and screening in diatoms. We initially adapted our Golden Gate

133 Modular Cloning-based episomal assembly framework[28] to be 96-well compatible and
134 combined it with multi-well diatom transformation via bacterial conjugation. We coupled this
135 with 48-well plate strain maintenance and 96-well plate flow cytometry screening for clonal
136 fluorophore-fusion expressing lines (Fig. 2A). As coIPMS data on a small-scale is inherently
137 noisy with both false positives and false negatives[30] we applied our tagging pipeline to confirm
138 if our coIPMS hits were bona fide pyrenoid proteins. Nourseothricin-positive transformants
139 were picked and screened for positive fluorescence using flow cytometry. Due to typical
140 mosaic colony presence[31] either multiple rounds of screening or screening of several
141 independent colonies was completed. We found that screening 8–12 colonies would typically
142 yield a stable cell population with >90% of cells mEGFP positive (Fig. S2). Positive lines were
143 subsequently imaged by confocal microscopy (Fig. 2B and Fig. S3). From the initially identified
144 Rubisco-interacting proteins, we saw multiple uncharacterized proteins localizing to distinct
145 sub-regions of the pyrenoid (Fig. 2C; see below for further discussion). However, several
146 candidates were localized to organelles adjacent to the pyrenoid, such as the chloroplast
147 stroma, mitochondria, and endoplasmic reticulum (Fig. S3) indicating that our coIPMS data
148 contains false positives.
149 To accurately determine sub-pyrenoid localization of components we developed a
150 dual-tagging approach using two spectrally compatible fluorophores. We first developed a
151 pyrenoid matrix marker for co-localization. In green algae, nuclear-encoded rbcS-fluorescent
152 protein fusions have been powerful for understanding sub-pyrenoid spatial organization [19] and
153 for determining the liquid-like properties of the pyrenoid.[12] As the rbcS of diatoms is chloroplast
154 encoded and no *T. pseudonana* chloroplast transformation protocol is available, we wondered
155 if we could target an episomal expressed rbcS-mEGFP to the pyrenoid. Using the N-terminal
156 signal and transit peptide sequences from the nuclear encoded chloroplast localized BST2
157 protein,[32] we successfully targeted rbcS to the pyrenoid, allowing us to clearly define the
158 pyrenoid matrix (Fig. 2C). Second, we tested assembling two target genes with different
159 fluorophores on the same episome. We decided to initially validate the BST2 localization by
160 making an episome with BST2-mEGFP and rbcS-mScarlet-I, a fluorophore we had previously
161 validated using our episome system.[28] This gave a clear pyrenoid localization with the rbcS
162 signal extending outside of the BST2 signal supporting a PPT localization of BST2 (Fig. 2D).
163

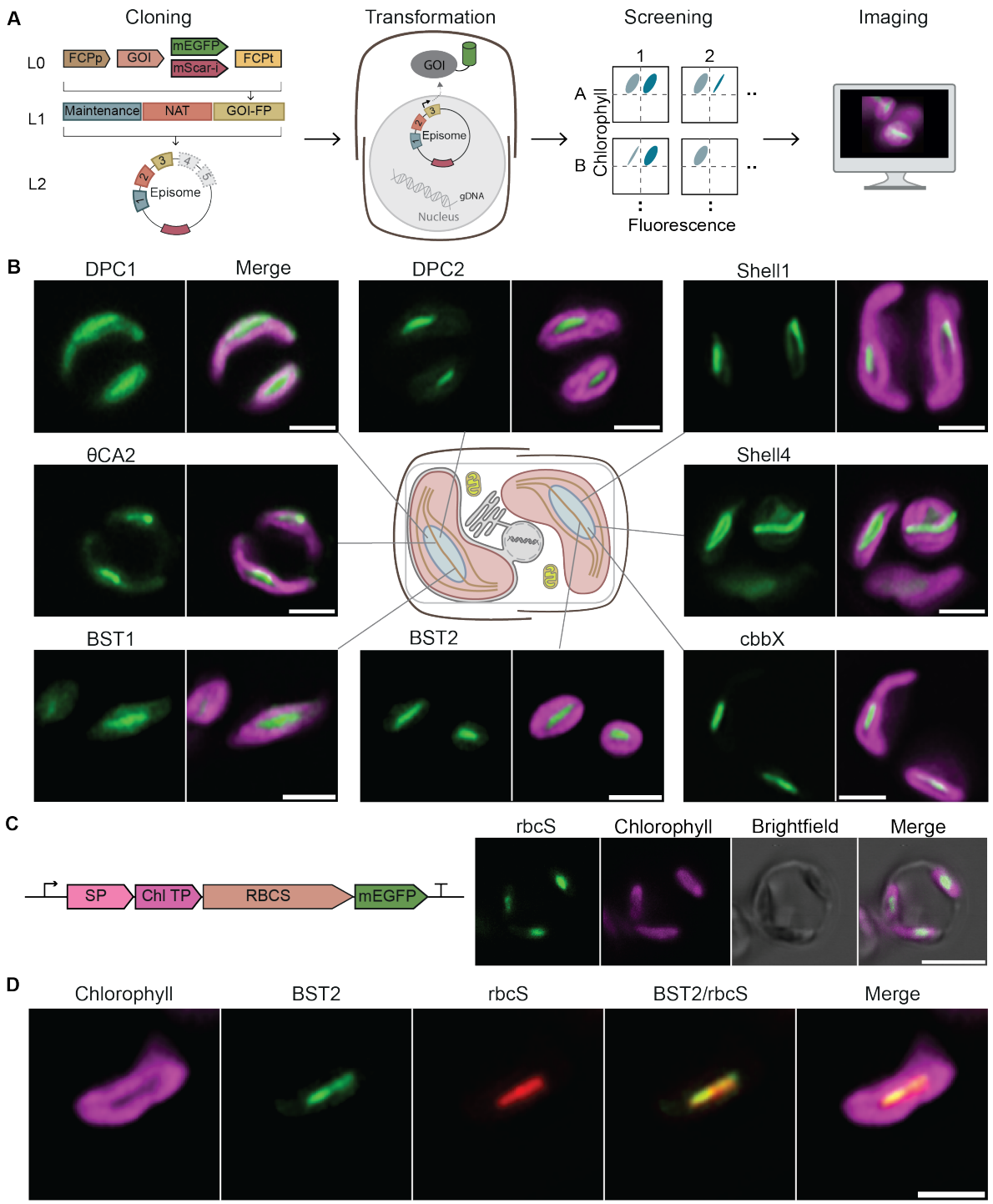



**Figure 2. High-throughput fluorescence protein tagging identifies multiple new diatom pyrenoid components.**

A. Candidate genes are rapidly cloned using Golden Gate modular cloning, transformed into *T. pseudonana* via bacterial conjugation and screened by flow cytometry for fluorescence prior to imaging.

B. Multiple proteins localized to the canonical pyrenoid region within the chloroplast. Newly identified pyrenoid components with no sequence-predicted function are termed Diatom Pyrenoid Components (DPCs) or Shells. Green: mEGFP fusion protein; Magenta: chlorophyll autofluorescence. Scales bars: 2 µm.

C. Development of a pyrenoid matrix marker. By using the BST2 signal and transit peptide sequences we could express chloroplast encoded rbcS from nuclear episomes and target it to the chloroplast. Scale bar: 5µm.

D. Co-localization of BST2-mEGFP (green) and rbcS-mScarlet-I (red) expressed from the same episome. Scale bar: 2 µm.

**Establishing large-scale APMS to build a pyrenoid interactome**

To rapidly expand the pyrenoid proteome we developed and optimized an APMS pipeline in *T. pseudonana* and used our GFP-tagged pyrenoid proteins as bait proteins in this pipeline. Lines expressing GFP-tagged proteins were typically grown at ambient-$CO_2$, where the CCM is fully active.[33] In triplicate, GFP-trap nanobodies were used to enrich for target proteins from cell lysate and their interactors determined via LC-MS/MS (Fig. 3A, Table S2). Protein-protein interactions were stringently determined by comparing both CompPASS[19,34] and SAINT analysis[35] scores (Fig. 3A, Table S3) that use different weighting criteria to identify true interactors from non-specific background using label-free proteomic quantitation data. We set interaction confidence thresholds based on the known interaction of rbcS with rbcL, which resulted in proteins in the top 2.2% for CompPASS and top 1% for SAINT being designated as high confidence interactors.

From our initial Rubisco coIPMS we GFP tagged 22 proteins and 13 localized to the pyrenoid. These included the Rubisco small subunit rbcS; BST1 and BST2 that are bestrophin-like proteins proposed to be involved in $HCO_3^-$ uptake into the PPTs;[28] θCA2 most likely involved in $CO_2$ release from $HCO_3^-$ within the PPTs;[25,26] cbbX a nuclear encoded red-type Rubisco activase[36] that until now has not been localized in algae with red plastids;[37] and 8 uncharacterized proteins with no clear function. These newly identified GFP-tagged pyrenoid components were called Diatom Pyrenoid Components 1 and 2 (DPC1 and DPC2) and Shell 1-6. Whilst DPC1 and DPC2 were predominantly in the pyrenoid matrix the initial tagging of Shell1 and Shell4 showed that they may encapsulate the pyrenoid (Fig. 1C; see below). A subset of these pyrenoid localized components were utilized for APMS using our iterative pipeline (Fig. 1C and 3). Subsequently, two additional components, DPC3 and DPC4, that had strong interactions with Shell4 and rbcS but again with no sequence predictable function were localized to the pyrenoid and also fed into our pipeline. Combining the data and using our stringent scoring approach enabled us to expand and build a high-confidence pyrenoid interactome for *T. pseudonana* built from 11 baits and containing 46 additional protein nodes linked by 57 interaction edges (Fig. 3B). In the network interaction confidence can be further interpreted by CompPASS and SAINT score magnitude (line thickness and color respectively in Fig. 3B) as well as the number of connecting edges with baits.

RbcS and DPC2 appear to be key hub proteins, each linking four nodes that further link to confirmed pyrenoid components. Although shell proteins were identified in our initial rbcL coIPMS data, and detected in our rbcS APMS they did not fall above the stringent thresholds set to be defined as high-confidence interactors with either rbcS or rbcL. This indicates that shell proteins may interact with the Rubisco matrix via an intermediary protein. The shell-like pattern displayed by DPC3 (Fig. 3C) suggests that this protein could be acting as a Rubisco matrix-shell adaptor or potentially an additional shell component. How Rubisco is packaged into the pyrenoid is unknown in *T. pseudonana*. With the absence of an EPYC1 or PYCO1 homolog to phase separate Rubisco, it was hypothesized that an alternative repeat protein could be fulfilling this role. Unexpectedly none of the pyrenoid proteins identified in our study contain a repeated sequence with the expected frequency of ~60 amino acids,[1] potentially indicating that pyrenoid assembly in *T. pseudonana* could be based on different biophysical principles to both green algal and pennate diatom pyrenoids.

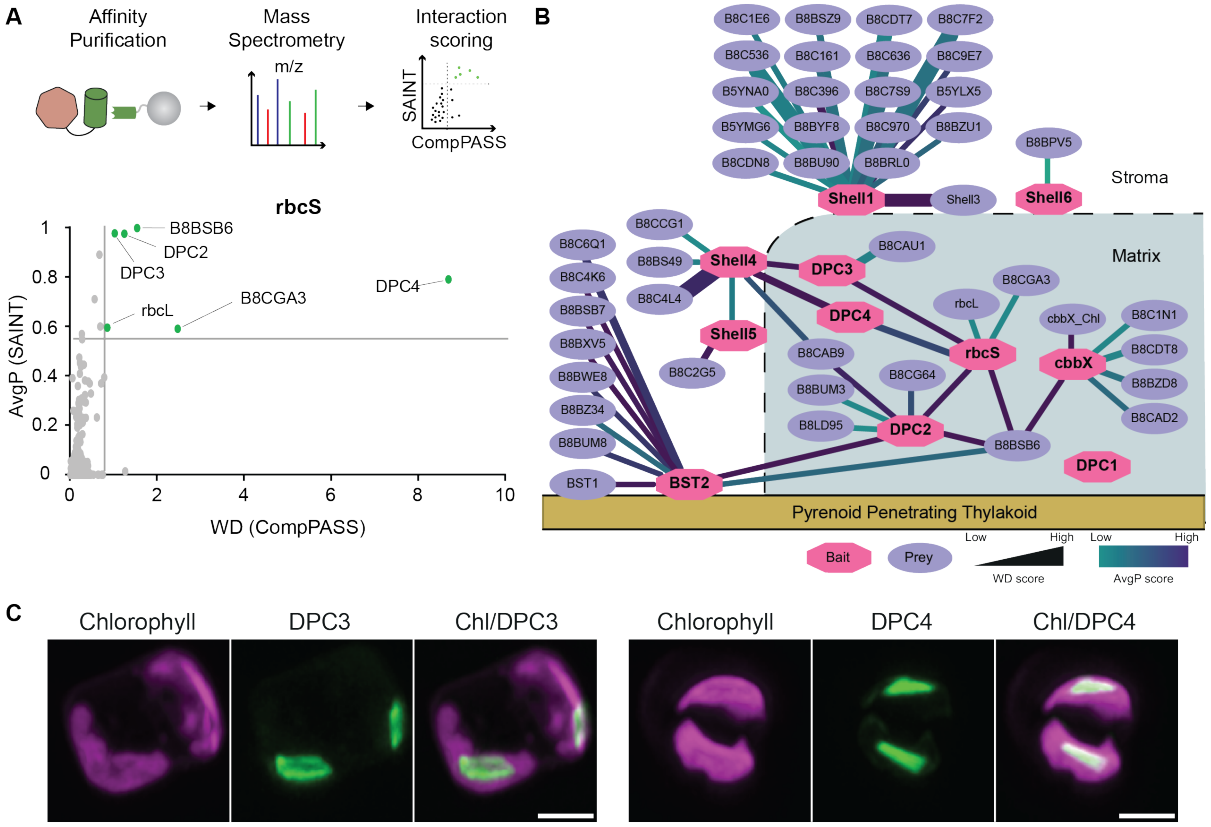



**Figure 3. A spatial interactome of the diatom pyrenoid.**

A. mEGFP tagged proteins were affinity purified using GFP nanobodies, interactors identified using mass spectrometry and high-confidence protein-protein interactions determined using SAINT and CompPASS analysis. RbcS is not present in the plot due to bait spectral counts being set to zero prior to SAINT and CompPASS analysis.

B. Spatially determined protein-protein interactome of the pyrenoid. Positioning is based on localization determined by confocal microscopy. Due to 96% similarity of the mature proteins of Shell1 and Shell2 all peptides were assigned to Shell1.

C. Confocal localization of DPC3 and DPC4 confirming pyrenoid localization (max intensity z-stack projections). Green: mEGFP fusion protein; Magenta: chlorophyll. Scale bars: 2µm.

## Six proteins encapsulate the pyrenoid matrix

In our initial tagging we were intrigued to see that two proteins with no annotated functional domains appeared to encapsulate the pyrenoid matrix (Fig. 2C: Shell1, Shell4). BLAST analysis identified four additional homologs in the *T. pseudonana* genome that all contain a predicted beta-sheet domain (Fig. 4A and Fig. S4). We explored the distribution of these proteins across different evolutionary lineages. BLAST analysis against the NCBI database identified homologs in the stramenopiles (including other diatoms), pelagophytes and haptophytes; all of which are photosynthetic algae that contain secondary endosymbiotic red plastids (Fig. 4B). However, Shell proteins are absent in the rhodophytes, the primary endosymbiotic red plastid lineage, suggesting that Shell proteins were not present in the engulfed red alga but may have been present in the genome of the heterotrophic host prior to endosymbiosis or evolved after engulfment. Further supporting a role in pyrenoid function of shell proteins, TEM data from available literature indicates that algae found within the Shell protein containing clades all contain pyrenoids.[2]

To elucidate the precise sub-regions of the six Shell proteins we co-expressed rbcS-mScarlet-I with GFP-tagged Shell proteins (Fig. 4C and Fig. S5). Analysis of fluorescence

intensity of perpendicular transects of the pyrenoid (Fig. 4D and Fig. S5) and max intensity z-stack projections (Fig. 4E and Fig. S5) indicate that all six proteins encapsulate the Rubisco matrix of the pyrenoid. Intriguingly, they can be grouped into two localization patterns. Generalizing the pyrenoid shape to be an elliptic cylinder with curved ends (Fig. S6), Shell1,2,3 and 6 encapsulate the pyrenoid matrix around the high-curvature edges, whereas Shell4 and 5 radially encapsulate the pyrenoid matrix (Fig. 4C,D,E and Fig. S5 and S6). Further supporting different pyrenoid surface localizations AlphaFold2 model comparisons of Shell1 with each of the other Shell proteins indicates that Shell2,3 and 6 are structurally closer to Shell1 than Shell4 and 5 (Fig. S7). With small differences in rotational angles potentially enabling different curvature.

In the well-characterized green algal pyrenoid, chloroplast synthesized starch forms a sheath that encapsulates the Rubisco matrix and acts as a $CO_2$ leakage barrier to enhance CCM efficiency.[17,18] In diatoms the main carbohydrate storage molecule is chrysolaminaran that is stored in cytosolic vacuoles[38] with no clear carbohydrate barrier surrounding the diatom pyrenoid. We hypothesize that the shell proteins, instead of starch, could be forming a diffusion barrier in diatoms to minimize $CO_2$ leakage. This has potential structural analogies to the cyanobacterial carboxysomes where a protein shell encapsulates Rubisco and is proposed to act as a $CO_2$ barrier to minimize leakage.[39,40] In addition, carboxysome shell proteins are essential for carboxysome biogenesis and shape.[41,42] We set out to further understand the importance of the shell proteins for pyrenoid function and more broadly understand diatom pyrenoid functionality.

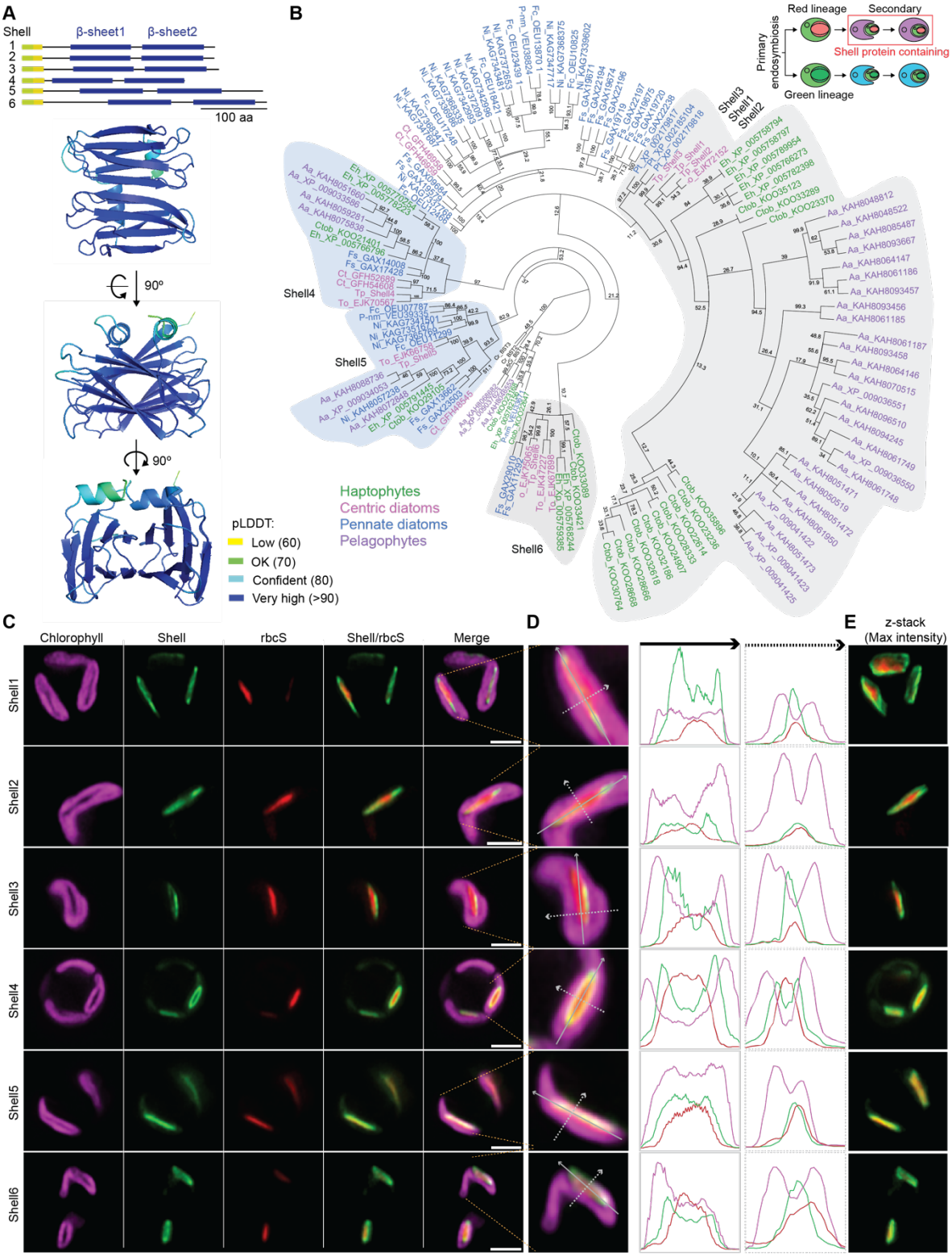



**Figure 4. Six proteins form a shell that encapsulates the diatom pyrenoid matrix**

A. *T. pseudonana* has six Shell protein homologs that all contain two beta-sheet regions. Top: domain alignment of Shell homologs, Green is signal peptide, yellow is transit peptide. Bottom: AlphaFold2 structure prediction of Shell1. pLDDT: predicted local distance difference test.

B. Phylogenetic analysis indicates that shell proteins are found in red plastid secondary endosymbionts including stramenopiles, pelagophytes and haptophytes. Bootstrap values are shown. Cartoon inset represents the two main plastid lineages and where Shell proteins are present.

C. Co-localization of Shell proteins with Rubisco. Scale bars: 2μm.

D. Fluorescence intensity cross-sections of Shell and rbcS fluorescence across the pyrenoid. Level of Zoom changes between images.
E. Z-stack max intensity projections of C.

**The pyrenoid matrix and shell proteins have minimal mobility in the pyrenoid**

To understand the dynamics of pyrenoid components, we leveraged our generated tagged lines. If the shell proteins act as a $CO_2$ barrier we would expect them to form a static assembly around the Rubisco matrix. We investigated the mobility of Shell1 and Shell4 by fluorescence recovery after photobleaching (FRAP). Supporting the role of the Shell proteins forming a structural barrier around the pyrenoid matrix both Shell1 and Shell4 showed minimal mixing after photobleaching (Fig. 5A, B).

In *C. reinhardtii* the pyrenoid matrix has liquid-like properties with Rubisco, EPYC1 and Rubisco activase (RCA1) all showing rapid recovery in FRAP experiments on the timescale of ~30 s after photobleaching.[12] Surprisingly, FRAP experiments on both *T. pseudonana* rbcS and cbbX showed minimal mixing (Fig. 5C, D). This aligns with recent *in vitro* and *in vivo* data where *in vitro* phase separated Rubisco from the pennate diatom *P. tricornutum* has slow mixing within droplets with full recovery not seen after 30 minutes and PYCO1 is immobile *in vivo.*[9] This opens a considerable question on how cbbX can sufficiently access inhibited Rubisco to reactivate it. Collectively, this indicates that the diatom pyrenoid has different mesoscale properties to the Chlamydomonas pyrenoid and that once Rubisco is assembled into the pyrenoid there is minimal mixing. This aligns with the cyanobacterial carboxysome where upon Rubisco packing it forms ordered arrays.[43,44]

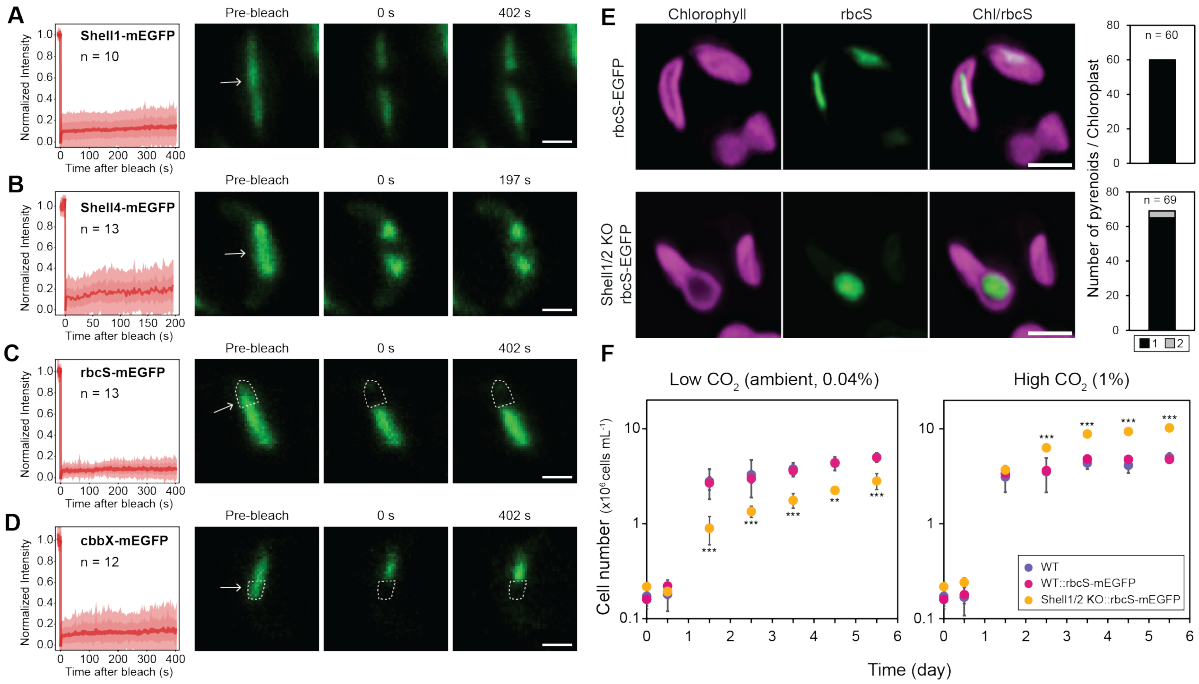



**Figure 5: Shell proteins are static on the pyrenoid surface and essential for pyrenoid morphology**

A-D: Fluorescence recovery after photobleaching (FRAP) experiments for different diatom pyrenoid components. Arrows indicate photobleached regions. Scale bars: 1 µm. Plots show S.E.M. and S.D. of the mean.

313 E: Tagged rbcS in WT (top) and a Shell1/2 knock-out line. Scale bars: 2 μm. Bar charts indicate
314 pyrenoid number per chloroplast in WT::rbcS-mEGFP and Shell1/2 knock-out::rbcS-mEGFP
315 lines.
316 F: Growth rate of WT, WT::rbcS-mEGFP and Shell1/2 knock-out::rbcS-mEGFP lines.
317 Statistical analyses were performed using Student's t-test (**$p<0.01$, *** $p<0.001$).
318
319 **Shell1 and Shell2 are essential for pyrenoid structure**
320 In addition to the Shell proteins potentially providing a barrier to minimize $CO_2$ leakage from
321 the pyrenoid matrix we hypothesized they may also be required for pyrenoid structural integrity.
322 To test this, we used our MoClo Golden Gate system to simultaneously GFP tag rbcS and
323 delete Shell1/2. Due to Shell1/2 being abundant shell components and having 93% DNA
324 sequence similarity sgRNAs were designed that targeted both genes (Fig. S8). Edited lines
325 were grown under high $CO_2$ conditions, selected for by GFP fluorescence, and gene editing
326 confirmed by Sanger sequencing (Fig. S8). Microscopy images of a Shell1/2 knock-out line
327 that showed biallelic editing of both Shell1 and Shell2 failed to form a lenticular pyrenoid and
328 instead typically possessed a single spherical pyrenoid per chloroplast (Fig. 5E). This
329 suggests that Shell1/2 are required for the lenticular shape of the pyrenoid and in their absence
330 the pyrenoid assembles into a sphere. Although speculative, spherical condensation of
331 Rubisco suggests that liquid-liquid phase separation may have a role in pyrenoid matrix
332 assembly. The *shell1/2* mutant had significantly reduced growth at atmospheric $CO_2$ that
333 could be fully rescued by supplying elevated $CO_2$ (Fig. 5F). Interestingly, the mutant grew
334 faster than wild-type at high $CO_2$ suggesting that the Shell may be minimizing $CO_2$ diffusion to
335 Rubisco in the pyrenoid when there is an excess of $CO_2$ in the surrounding environment.
336 These data support that the diatom Shell is critical for pyrenoid structural integrity and function
337 and may act as a $CO_2$ diffusion barrier.
338
339 **A model of the diatom pyrenoid-based $CO_2$ concentrating mechanism**
340 Combining our data with previous experimental data we propose a structural and functional
341 model for the pyrenoid (Fig. 6). The SLC4 family proteins contribute to sodium dependent
342 $HCO_3^-$ transport at the plasma membrane.[7,23,24] Ci transport across the four chloroplast
343 membranes is still unknown, SLC4 transporters are also proposed to play a role here[27] along
344 with the carbonic anhydrase LCIP63[45] and the V-type ATPase.[46] Ci delivery to the pyrenoid is
345 potentially analogous to *C. reinhardtii* via channeling of $HCO_3^-$ into the thylakoid lumen via
346 BST1 and BST2.[28] $HCO_3^-$ is then dehydrated to $CO_2$ via θCA2[25,26] where it can diffuse to
347 Rubisco packaged within the pyrenoid. A doublet of thylakoids enter the pyrenoid where they
348 often branch within the pyrenoid matrix, potentially to increase surface area for $CO_2$ delivery.
349 $CO_2$ leakage out of the pyrenoid is then minimized by the proteinaceous shell that
350 encapsulates the pyrenoid. With the shell also critical for maintaining the lenticular shape of
351 the pyrenoid.
352

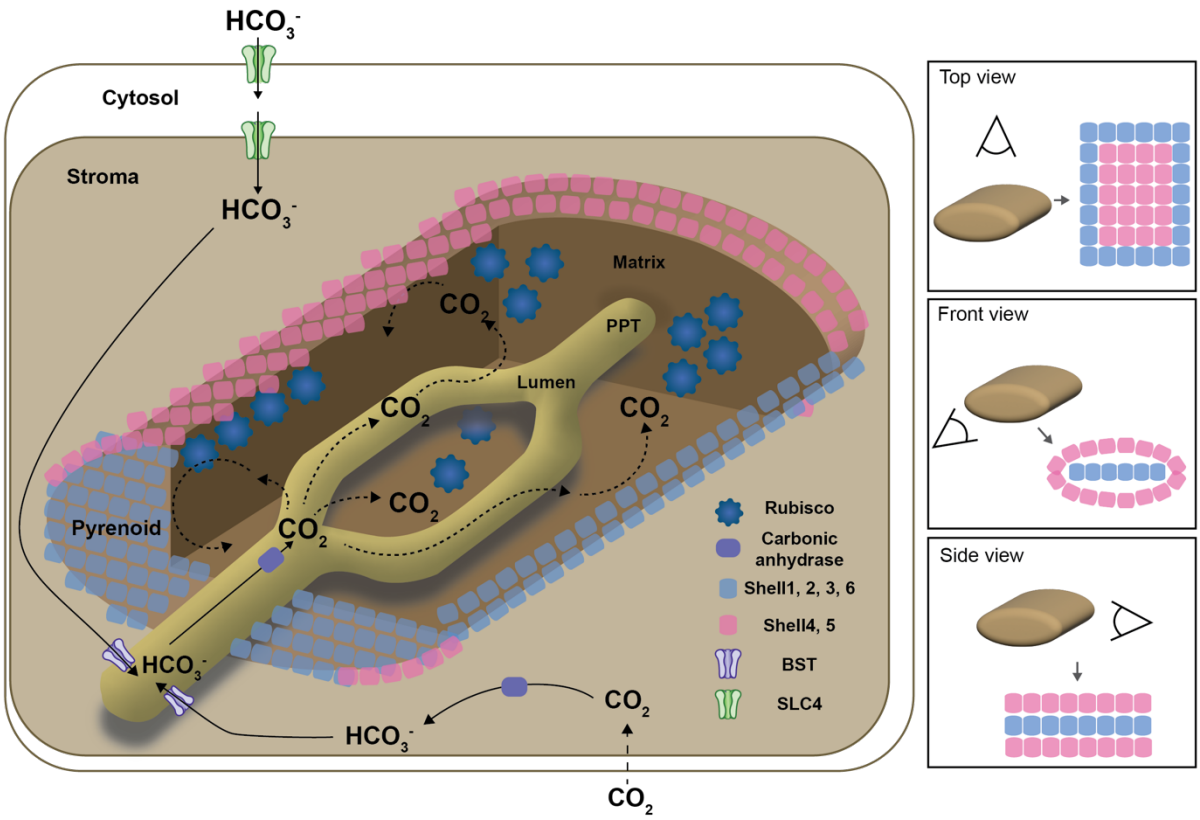



**Figure 6: Structural and functional model of the *T. pseudonana* pyrenoid-based $CO_2$-concentrating mechanism.** Proposed model based on localization data and availbale literature.

## Perspective

The development of a high-throughput tagging and APMS pipeline in a model diatom has enabled us to generate a spatial interactome of the diatom pyrenoid, providing novel molecular insight into how diatoms drive the global carbon cycle. We have identified and confirmed via localization 13 new pyrenoid components, of which a large number have no conserved functional domains. Six of these new components constitute a protein Shell that encapsulates the pyrenoid and is found across diverse red plastid secondary endosymbionts. Simultaneous knock-out of Shell1 and Shell2 resulted in large pyrenoid structural changes and poor growth at atmospheric levels of $CO_2$. Moving forward it will be important to understand the importance of the additional four Shell proteins, especially Shell4 and Shell5 that localize to different surface regions of the pyrenoid than Shell1 and Shell2. Four additional pyrenoid components, DPC1-4, have no clear sequence predictable function. As yet the condensation of Rubisco to form the pyrenoid matrix is unknown, with no EPYC1 or PYCO1 homolog or functional analog identified in our study. This suggests that pyrenoid matrix formation maybe different in centric diatoms. The matrix localization and interaction partners of DPC2 and DPC4 are suggestive that they may have a central role in pyrenoid matrix assembly/function. DPC3 showed a more Shell-like localization pattern and may mediate Shell-Rubisco matrix interactions. DPC2, DPC3 and DPC4 are prime targets for future characterization.

Close to 50% of global carbon fixation is performed by biomolecular condensates of Rubisco. This includes prokaryotic cyanobacterial carboxysomes and eukaryotic algal pyrenoids. Nearly all of our data of pyrenoid structure and function comes from the green plastid lineage alga, *C. reinhardtii*, with pyrenoids both between plastid lineages and within

380 plastid lineages proposed to have convergently evolved.[2,4] Insights from our data suggest that
381 diatom pyrenoids have similarities to both green plastid pyrenoids and prokaryotic
382 carboxysomes. Similarities between the *T. pseudonana* and *C. reinhardtii* pyrenoids include
383 dense Rubisco packaging around specialized thylakoid membranes (PPTs) for $CO_2$ delivery.
384 With $CO_2$ delivery to the PPTs via bestrophin family channels and $CO_2$ release within the acidic
385 lumen driven by constrained localization of a carbonic anhydrase within the PPTs. However,
386 the dynamics of the pyrenoid and structural aspects of the pyrenoid have more similarities to
387 carboxysomes. The encapsulation by a protein shell layer composed of homologs, some with
388 different sub-shell localization patterns is analogous to carboxysome shell proteins.[39]
389 Additionally, the static nature of Rubisco, cbbX and shell proteins contrasts to dynamic
390 properties of the *C. reinhardtii* pyrenoid and aligns more with carboxysomes. Outside of
391 Rubisco, there are no homology or structural similarities between carboxysome and *T.*
392 *pseudonana* pyrenoid components. This supports the convergent evolution of pyrenoids and
393 that a broad range of biophysical and structural properties, some previously associated with
394 carboxysomes, can be expected as more pyrenoids are characterized across diverse alga.
395 A core structural component of pyrenoids is a $CO_2$ leakage barrier, with the absence
396 of a starch sheath in *C. reinhardtii* shown both experimentally and theoretically to be required
397 for efficient CCM function.[17,18] As engineering of a pyrenoid in plants progresses a major future
398 challenge will be $CO_2$ diffusion barrier engineering.[17,22] This is thought to require multiple
399 starch synthesis related steps correctly localized to the pyrenoid periphery. The diatom Shell
400 proteins could provide an alternative biotechnology solution to this challenge.
401

402 ## Materials and methods

403

404 ### Strains and Culturing

405

406 The background *Thalassiosira pseudonana* strain for all experiments was wild-type (WT)
407 CCAP1085/12 (Scottish Culture Collection of Algae and Protozoa, equivalent to CCMP1335).
408 WT cells were axenically maintained in artificial sea water (ASW) (32 g $L^{-1}$, Instant Ocean
409 SS15-10) supplemented with half-strength (F/2) Guillard F solution[47,48] at 20°C under
410 continuous illumination of ~50 µmol photons $m^{-2}$ $s^{-1}$. All strains were grown at ambient $CO_2$
411 except *shell1/2* knock-out and Shell1-mEGFP lines, which were maintained under 1% $CO_2$.

412 ### Episome Assemblies using Golden Gate Cloning

413 Level 0, 1, and 2 (L0, L1, and L2) plasmids were assembled by Golden Gate (GG) cloning[49]
414 using the custom parts from the diatom MoClo framework.[28] The target genes without stop
415 codon were synthesized from Twist Bioscience (Table S4).

416 ### *T. pseudonana* Transformation via Bacterial Conjugation

417 Episomes were delivered to *T. pseudonana* via bacterial conjugation according to[50] with minor
418 modifications. Episome plasmids were transformed into *E. coli* (TransforMax EPI300)
419 harboring the pTA_Mob[51] mobility plasmid (gift from R. Lale) via electroporation (Bio-Rad).
420 Transformed cells were spread onto LB agar plates containing both gentamycin (10 µg $mL^{-1}$)
421 and kanamycin (25 µg $mL^{-1}$) for selection overnight at 37°C. Colonies were inoculated for
422 subsequent conjugation. Cultures (150 mL) grown at 37°C to OD600 of 0.3-0.4 were harvested

423 by centrifugation (3,000 ×g, 5 min) and resuspended in 800 µL of SOC media. Liquid grown
424 *T. pseudonana* WT culture was harvested by centrifugation (3,000 ×g, 5 min) and
425 resuspended at a concentration of $2x10^8$ cells $mL^{-1}$ in ½ASW-F/2. Equal volume (200 µL) of
426 *E. coli* and *T. pseudonana* WT cells were gently mixed by pipetting. Next the mixture of cells
427 was plated on ½ASW-F/2, 5% LB, 1% agar plates and incubated in the dark for 90 min at
428 30°C. The plates were transferred to 20°C with continuous illumination (~50 µmol photons·$m^{-2}$·$s^{-1}$)
429 and grown overnight. Next day, 500 µL of ½ASW-F/2 medium was added to the plate for
430 scraping and resuspending the cells. Up to 200 µL of resuspended cells were spread onto 1%
431 (w/v) ½ASW-F/2 agar plates with 100 µg $mL^{-1}$ nourseothricin for selection. Colonies appeared
432 after 6-14 days.

433 **Fluorescence Screening by Flow Cytometry**

434 mEGFP and mScarlet-I expression was analyzed by flow cytometry using either CytoFLEX
435 LX355 or 375 (Beckman Coulter) analyzers. Forward scattered (FSC) and side scattered
436 photons by the 488 nm laser were used to distinguish diatoms from cell culture debris. FSC-
437 height versus FSC-area signal was used to separate single events from sample aggregates.
438 Chlorophyll autofluorescence excited by 561 nm laser and emitted photons detected with
439 675/25 filter was used to ensure all the diatom cells were fully intact. mEGFP fluorescence
440 excited by the 488 nm laser was detected by an avalanche photodiode detector with 525/40
441 bandpass filter. All the data analysis was done using CytExpert software (Beckman Coulter).

442 **Microscopy**

443 All super-resolution imaging was performed using Zeiss LSM880 confocal microscope in
444 Airyscan mode with a 63x objective, 1.4 numerical aperture (NA) Plan-Apo oil-immersion lens
445 (Carl Zeiss). The rbcS-mEGFP line was imaged in confocal mode. The 20 µL of cell
446 suspension were pipetted on 8 well µ-Slide chambered coverslips (ibidi) overlaid with 180 µL
447 of 1.5% F/2-low-melting point agarose (Invitrogen) for imaging. Excitation lasers and emission
448 filters were as follows: mEGFP excitation 488 nm, emission 481-541 nm; mScarlet-I excitation
449 561 nm, emission 561-633 nm; and chlorophyll excitation 633 nm, emission 642-712 nm. All
450 the microscopic images were processed using Fiji.[52]

451 **Co-immunoprecipitation and affinity purification**

452 For rbcL coIP, 50 mL of WT *T. pseudonana* cells grown in log phase ($2\text{-}3x10^6$ cells $mL^{-1}$) were
453 harvested by centrifugation (3,000 ×g, 10 min). The pellets were resuspended in a CoIP buffer
454 (20 mM Tris-HCl pH 8.0, 50 mM NaCl, 0.1 mM EDTA, 12.5% glycerol) containing 5 mM DTT
455 and protease inhibitor cocktail tablets (PIs, cOmplete EDTA-free, Roche). Cells were lysed by
456 sonication for 3 min (ON 3 sec, OFF 12 sec). The lysates were centrifuged for 20 min (full
457 speed, 4°C) to separate the supernatant (soluble lysate) from the pellet. 200 µL of Protein A
458 (Dynabeads Protein A, Invitrogen) beads were washed twice in coIP buffer containing PIs. 32
459 µg of anti-rbcL antibody in 500 µL coIP buffer containing PIs was added to the washed beads
460 and incubated at 4°C for 2 hours. After incubation beads were washed twice in coIP buffer
461 containing PIs. For blocking, 500 µL of BSA (2 mg $mL^{-1}$) was added and incubated at 4°C for

462 1 hour. After incubation beads were washed twice in coIP buffer containing PIs. Subsequently,
463 the soluble lysates were added to protein A beads primed with antibody and incubated at 4°C
464 for 3 hours. After incubation, beads were washed three times with coIP buffer containing PIs
465 and 0.1% digitonin (SigmaAldrich). For elution, 200 µL of 1x SDS loading dye was added and
466 boiled at 95°C for 5 min. The supernatant was collected without any beads and ran on an
467 SDS-PAGE gel for ~1.5 cm. Gels were sliced for further in-gel digestion for LC-MS/MS (see
468 below).

469 For mEGFP tagged lines AP, 50 mL of GFP tagged *T. pseudonana* lines grown in exponential
470 phase (2-3x$10^6$ cells $mL^{-1}$) were harvested by centrifugation (3,000 ×g, 10 min). The pellets
471 were resuspended in an immunoprecipitation (IP) buffer (200 mM D-sorbitol, 50 mM HEPES,
472 50 mM KOAc, 2 mM $Mg(OAc)_2$, 1 mM $CaCl_2$) containing protease inhibitor cocktail tablets
473 (cOmplete EDTA-free, Roche), 2% digitonin (SigmaAldrich), 1mM PMSF, 0.5 mM NaF and
474 0.15 mM $Na_3VO_4$. Cells were lysed by sonication for 30 sec (On 3 sec, Off 15 sec) twice. The
475 lysates were centrifuged for 20 min (full speed, 4°C) and the supernatant was incubated with
476 mEGFP-Trap Agarose beads (ChromoTek) for 1 hour according to the manufacturer's
477 instructions. Subsequently, beads were washed twice with an IP buffer containing 0.1%
478 digitonin and a final wash without digitonin. All steps were performed at 4°C.

479 **Mass Spectrometry**

480 For rbcL coIPMS, samples were in-gel digested with 0.2 µg Sequencing-grade, modified
481 porcine trypsin (Promega), following reduction with 1.5 mg $ml^{-1}$ dithioerythritol and alkylation
482 with 9.5 mg $mL^{-1}$ iodoacetamide. Digests were incubated overnight at 37ºC. Peptides were
483 extracted by washing three times with aqueous 50% (v:v) acetonitrile containing 0.1% (v:v)
484 trifluoroacetic acid, before drying in a vacuum concentrator and reconstituting in aqueous 0.1%
485 (v:v) trifluoroacetic acid. Peptides were loaded onto an mClass nanoflow UPLC system
486 (Waters) equipped with a nanoEaze M/Z Symmetry 100 Å $C_{18}$, 5 µm trap column (180 µm x
487 20 mm, Waters) and a PepMap, 2 µm, 100 Å, C 18 EasyNano nanocapillary column (75 mm
488 x 500 mm, Thermo). The trap wash solvent was aqueous 0.05% (v:v) trifluoroacetic acid and
489 the trapping flow rate was 15 µL $min^{-1}$. The trap was washed for 5 min before switching flow
490 to the capillary column. Separation used gradient elution of two solvents: solvent A, aqueous
491 0.1% (v:v) formic acid; solvent B, acetonitrile containing 0.1% (v:v) formic acid. The flow rate
492 for the capillary column was 330 nL $min^{-1}$ and the column temperature was 40°C. The linear
493 multi-step gradient profile was: 3-10% B over 5 mins, 10-35% B over 85 mins, 35-99% B over
494 10 mins and then proceeded to wash with 99% solvent B for 5 min. The column was returned
495 to initial conditions and re-equilibrated for 15 min before subsequent injections. The nanoLC
496 system was interfaced with an Orbitrap Fusion Tribrid mass spectrometer (Thermo) with an
497 EasyNano ionisation source (Thermo). Positive ESI-MS and MS2 spectra were acquired using
498 Xcalibur software (version 4.0, Thermo). Instrument source settings were: ion spray voltage,
499 1900-2100 V; sweep gas, 0 Arb; ion transfer tube temperature; 275°C. MS 1 spectra were
500 acquired in the Orbitrap with: 120,000 resolution, scan range: *m/z* 375-1,500; AGC target, $4e^5$;
501 max fill time, 100 ms. Data dependant acquisition was performed in top speed mode using a
502 1 s cycle, selecting the most intense precursors with charge states >1. Easy-IC was used for
503 internal calibration. Dynamic exclusion was performed for 50 s post precursor selection and a
504 minimum threshold for fragmentation was set at $5e^3$. MS2 spectra were acquired in the linear

505 ion trap with: scan rate, turbo; quadrupole isolation, 1.6 *m/z*; activation type, HCD; activation
506 energy: 32%; AGC target, $5e^3$; first mass, 110 *m/z*; max fill time, 100 ms. Acquisitions were
507 arranged by Xcalibur to inject ions for all available parallelizable time. Tandem mass spectra
508 peak lists were extracted from Thermo .raw files to .mgf format using MSConvert
509 (ProteoWizard 3.0). Mascot Daemon (version 2.6.0, Matrix Science) was used to submit
510 searches to a locally-running copy of the Mascot program (Matrix Science Ltd., version 2.7.0).
511 Peak lists were searched against the *Thalassiosira pseudonana* subsets of UniProt and NCBI
512 with common proteomic contaminants appended. Search criteria specified: Enzyme, trypsin;
513 Max missed cleavages, 2; Fixed modifications, Carbamidomethyl (C); Variable modifications,
514 Oxidation (M); Peptide tolerance, 3 ppm; MS/MS tolerance, 0.5 Da; Instrument, ESI-TRAP.
515 Peptide identifications were collated and filtered using Scaffold (5.2.0, Proteome Software
516 Inc). Peptide identifications were accepted if they could be established at greater than 51.0%
517 probability to achieve an FDR less than 1.0% by the Percolator posterior error probability
518 calculation. Protein identifications were accepted if they could be established at greater than
519 5.0% probability to achieve an FDR less than 1.0% and contained at least 2 identified peptides.

520

521 For mEGFP tagged lines APMS, samples were on-bead digested using Chromotek's
522 recommended procedure for NanoTraps: protein was digested overnight at 37°C with 25 μL
523 50 mM Tris-HCl pH 7.5, 2 M urea, 1mM DTT, 5 μg $ml^{-1}$ Sequencing Grade Modified Trypsin
524 (Promega). Peptides were eluted with 50 mM Tris-HCl pH 7.5, 2 M urea, 5 mM iodoacetamide
525 before loading onto EvoTip Pure tips for desalting and as a disposable trap column for
526 nanoUPLC using an EvoSep One system. A pre-set EvoSep 60 SPD gradient was used with
527 a 8 cm EvoSep $C_{18}$ Performance column (8 cm x 150 μm x 1.5 μm). The nanoUPLC system
528 was interfaced to a timsTOF HT mass spectrometer (Bruker) with a CaptiveSpray ionisation
529 source (Source). Positive PASEF-DDA, ESI-MS and $MS^2$ spectra were acquired using
530 Compass HyStar software (version 6.2, Thermo). Instrument source settings were: capillary
531 voltage, 1,500 V; dry gas, 3 l $min^{-1}$; dry temperature; 180°C. Spectra were acquired between
532 *m/z* 100-1,700. The following TIMS settings were applied as: 1/K0 0.6-1.60 V.s $cm^{-2}$; Ramp
533 time, 100 ms; Ramp rate 9.42 Hz. Data dependent acquisition was performed with 10 PASEF
534 ramps and a total cycle time of 1.17 s. An intensity threshold of 2,500 and a target intensity of
535 20,000 were set with active exclusion applied for 0.4 min post precursor selection. Collision
536 energy was interpolated between 20 eV at 0.5 V.s $cm^{-2}$ to 59 eV at 1.6 V.s $cm^{-2}$. Pick picking,
537 database searching, significance thresholding and peak area integration was performed using
538 FragPipe (version 19.1). Data were searched against UniProt reference proteome
539 UP000001449, appended with common contaminants and concatenated with reversed
540 sequences for false discovery calculation. Search criteria specified: Enzyme, trypsin; Max
541 missed cleavages, 2; Fixed modifications, Carbamidomethyl (C); Variable modifications,
542 Oxidation (M), Acetylation (Protein N-term); Peptide tolerance, 10 ppm; MS/MS tolerance, 10
543 ppm; Instrument, IM-MS. Peptide identifications were filtered using Percolator and
544 ProteinProphet to 1% PSM FDR, protein probabilities >99%, best peptide probability >99%
545 and a minimum of two unique peptides. Peak area quantification was extracted using IonQuant
546 with match between run applied. Feature detection tolerances were set to: MS1 mass <10
547 ppm; RT < 0.4 min; and IM (1/k0) <0.05.

548 **Interactome analysis**

549 Protein abundances quantified using MS2 spectral count measurements of fragment ions from
550 all sample triplicates were run through a CompPASS package in R Studio
551 (https://github.com/dnusinow/cRomppass/blob/master/R/cRomppass.R) and a control IP
552 inclusive variation of SAINT analysis in Ubuntu using standard parameters.[35] The WD and
553 AvgP scores respectively generated were used as measures of interaction strength between
554 bait and prey proteins. Only interactions which fell in both the top 2.2% WD score and 1%
555 AvgP score were filtered as high confidence interactors. Prior to analysis bait spectral count
556 data was set to zero to minimize data skewing due to the typically high spectral counts and
557 the inability to distinguish between mEGFP tagged bait and untagged native protein.

558 ***In vivo* fluorescence recovery after photobleaching**

559 FRAP experiments were performed using a Zeiss LSM980 confocal microscope with a 63x
560 objective 1.4 numerical aperture (NA) Plan-Apo oil-immersion lens (Carl Zeiss). Samples were
561 prepared as in confocal microscopy and overlaid with 200 µL of ibidi anti-evaporation oil. 20
562 pre-bleach images were taken prior to bleaching (60% 488 nm intensity, 1 cycle). All the
563 images were processed by Fiji.[52] The Image Stabilizer plugin (4 pyramid levels, 0.99 template
564 update coefficient) output of the brightfield images were used to stabilize the fluorescence
565 images. The mean gray values were measured for the bleached, unbleached and background
566 ROIs. Background values were subtracted from bleached and unbleached values before
567 photobleach normalization using the unbleached reference was completed. The average pre-
568 bleach intensity was used for full-scale normalization.

569 **Data Visualization**

570 Network visualization of the interactome was done using Cytoscape (version 3.10.0)
571 (https://cytoscape.org/). Phylogenetic tree was done in Geneious Prime (2023.2.1). Adobe
572 illustrator (2023) was used to generate the figures.

573 **Alphafold Structure Prediction**

574 The alphafold structure for Shell1 protein without signal peptide and chloroplast transit peptide
575 was predicted using ColabFold v1.5.2-patch: AlphaFold2 using MMseqs2
576 (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb).

577

578 **Protein model for Rubisco**

579 *Thalassiosira antarctica* Rubisco model (PDB: 5MZ2) was used to predict the peptide regions
580 utilized for raising the anti-rbcL antibody using ChimeraX.[53]

581 **Transmission Electron Microscopy**

582 50 mL of cell culture grown in log phase were harvested by centrifugation (3,000 ×g, 5 min).
583 Cells were fixed in 0.1 M cacodylate buffer (SigmaAldrich), pH 7.4, containing 2.5%
584 glutaraldehyde (SigmaAldrich), 2% formaldehyde (Polysciences) for 1 hour at room
585 temperature. After this, they were rinsed twice in phosphate-buffered saline (PBS, pH 7.4) for
586 15 min by removing the liquid at the top of the tube and leaving the cells undisturbed at the

587 bottom of the tube. The cells were secondary fixed with Osmium Tetroxide (1%, in buffer pH
588 7.2, 0.1M) for 1 hour. After rinsing twice with buffer, again removing the liquid above the cells
589 whilst leaving the cells undisturbed, the cells were dehydrated through an alcohol series (30%,
590 50, 70 90%) until it was in 100% ethanol. Each rinse for 15 min. To ensure thorough
591 dehydration the 100% ethanol step was repeated 3 times. Then the alcohol was replaced by
592 Agar low viscosity resin by placing it in increasing concentrations of resin (30% resin: 70 %
593 ethanol. 50:50, 70:30 each change was left for at least 12 hours) until it was in 100% resin.
594 Again to ensure complete resin infiltration, the 100% resin step was repeated 3 times leaving
595 it overnight between changes. The Eppendorf tubes were then placed in an embedding oven
596 and the resin polymerized at 60°C overnight. The resulting blocks were sectioned with a Leica
597 Ultracut E ultra microtome using a diatome diamond knife. The sections were then stained
598 using a saturated solution of Uranyl acetate (for 15 min) and Reynold's Lead citrate (15 min).
599 The sections were examined using a JEOL 1400 TEM.

600 **Blast search**

601 Full amino acid sequence of Shell 1-6 was used for Blastp with default settings: 'Standard
602 databases', 'Non-redundant protein sequences (nr)', 'blastp (protein-protein BLAST)' 'Max
603 target sequences=100' 'Word size =5' ' Max matches in a query range = 0' 'Matrix =
604 BLOSUM62' 'Gap Costs= Existence: 11 Extension: 1' 'Compositional adjustments =
605 Conditional compositional score matrix adjustment' and 'Expect threshold =1'. Hits were
606 sorted by highest to lowest alignment length, and a cut-off length of 100 amino acids was
607 employed.

608 **Phylogeny**

609 The phylogenetic tree was build using Geneious Prime (2023.2.1). Selection of 136 proteins
610 for MAFFT alignment defaults (Algorithm = 'Auto', Scoring matrix = 'BLOSUM62', Gap open
611 penalty = '1.53', Offset value = '0.123') Selection of MAFFT alignment. Generated
612 phylogenetic tree selecting RAxML: Protein Model = 'GAMMA BLOSUM62', Algorithm =
613 'Rapid Bootstrapping', Number of starting trees or bootstrap replicates = '1000', Parsimony
614 random seed = '1'. Generated a consensus tree selecting Consensus Tree Builder, Create
615 Consensus Tree, Support Threshold % = 0, Topology Threshold % = 0, Burn-in % = 0, Save
616 tree(s) separately. Added *C. reinhardteii* BST1-3 sequences (Cre16.g662600 (BST1)
617 Cre16.g663400 (BST2) Cre16.g663450 (BST3) to root the tree.

618 **Data availability**

619 All mass spectrometry and proteomic identification data is referenced in ProteomeXchange
620 (PXD045418) and can be accessed via MassIVE (MSV000092867).

621 **Author contributions:** O.N. and L.C.M.M. designed and supervised the study. O.N. carried
622 out the experiments. C.M. and O.N. analyzed the phylogenetic tree. A.D., M.D. and J.B.
623 provided bioinformatics and data analysis support. A.D. oversaw the mass spectrometry and
624 peptide mapping. O.N., C.M., and L.C.M.M. analyzed and interpreted the data. O.N. created
625 the figures. O.N. and L.C.M.M. wrote the manuscript with input from all authors.

**Acknowledgements:** Authors would like to thank Oliver Mueller-Cajar for kindly sharing the anti-rbcL antibody and members of the Mackinder Lab for fruitful discussions. The Technology Facility in the Department of Biology for the access/support for microscopy and flow cytometry. Thanks to Glenn Harper for optimizing the TEM. The York Centre of Excellence in Mass Spectrometry was created thanks to a major capital investment through Science City York, supported by Yorkshire Forward with funds from the Northern Way Initiative, and subsequent support from EPSRC (EP/K039660/1; EP/M028127/1). L.C.M.M. was supported by a UKRI Future Leader Fellowship (MR/T020679/1), EPSRC funding (EP/W024063/1) as part of the York Physics of Pyrenoids Project (YP3), and BBSRC funding (BB/S015337/1 and BB/X004953/1).

**References:**

1. Mackinder, L.C.M., Meyer, M.T., Mettler-Altmann, T., Chen, V.K., Mitchell, M.C., Caspari, O., Freeman Rosenzweig, E.S., Pallesen, L., Reeves, G., Itakura, A., et al. (2016). A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle. Proc. Natl. Acad. Sci. U. S. A. *113*, 5958–5963.

2. Barrett, J., Girr, P., and Mackinder, L.C.M. (2021). Pyrenoids: CO2-fixing phase separated liquid organelles. Biochim. Biophys. Acta Mol. Cell Res. *1868*, 118949.

3. Raven, J.A., Cockell, C.S., and De La Rocha, C.L. (2008). The evolution of inorganic carbon concentrating mechanisms in photosynthesis. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 2641–2650.

4. Meyer, M.T., Goudet, M.M.M., and Griffiths, H. (2020). The Algal Pyrenoid. In Photosynthesis in Algae: Biochemical and Physiological Mechanisms, A. W. D. Larkum, A. R. Grossmann, and J. A. Raven, eds. (Springer International Publishing), pp. 179–203.

5. He, S., Crans, V.L., and Jonikas, M.C. (2023). The pyrenoid: the eukaryotic CO2-concentrating organelle. Plant Cell. 10.1093/plcell/koad157.

6. Findinier, J., and Grossman, A.R. (2023). Chlamydomonas: Fast tracking from genomics. J. Phycol. 10.1111/jpy.13356.

7. Tsuji, Y., Nakajima, K., and Matsuda, Y. (2017). Molecular aspects of the biophysical CO2-concentrating mechanism and its regulation in marine diatoms. J. Exp. Bot. *68*, 3763–3772.

8. Nawaly, H., Tanaka, A., Toyoshima, Y., Tsuji, Y., and Matsuda, Y. (2023). Localization and characterization θ carbonic anhydrases in Thalassiosira pseudonana. Photosynthesis Research. 10.1007/s11120-023-01007-z.

9. Oh, Z.G., Ang, W.S.L., Poh, C.W., Lai, S.-K., Sze, S.K., Li, H.-Y., Bhushan, S., Wunder, T., and Mueller-Cajar, O. (2023). A linker protein from a red-type pyrenoid phase separates with Rubisco via oligomerizing sticker motifs. Proc. Natl. Acad. Sci. U. S. A. *120*, e2304833120.

10. Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. Science *281*, 237–240.

668 11. Benoiston, A.-S., Ibarbalz, F.M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., and
669 Bowler, C. (2017). The evolution of diatoms and their biogeochemical functions. Philos.
670 Trans. R. Soc. Lond. B Biol. Sci. *372*. 10.1098/rstb.2016.0397.

671 12. Freeman Rosenzweig, E.S., Xu, B., Kuhn Cuellar, L., Martinez-Sanchez, A., Schaffer,
672 M., Strauss, M., Cartwright, H.N., Ronceray, P., Plitzko, J.M., Förster, F., et al. (2017).
673 The Eukaryotic CO2-Concentrating Organelle Is Liquid-like and Exhibits Dynamic
674 Reorganization. Cell *171*, 148-162.e19.

675 13. Wunder, T., Cheng, S.L.H., Lai, S.-K., Li, H.-Y., and Mueller-Cajar, O. (2018). The
676 phase separation underlying the pyrenoid-based microalgal Rubisco supercharger. Nat.
677 Commun. *9*, 5076.

678 14. Meyer, M.T., Itakura, A.K., Patena, W., Wang, L., and He, S. (2020). Assembly of the
679 algal CO2-fixing organelle, the pyrenoid, is guided by a Rubisco-binding motif. Science.

680 15. Mukherjee, A., Lau, C.S., Walker, C.E., Rai, A.K., Prejean, C.I., Yates, G., Emrich-Mills,
681 T., Lemoine, S.G., Vinyard, D.J., Mackinder, L.C.M., et al. (2019). Thylakoid localized
682 bestrophin-like proteins are essential for the CO2 concentrating mechanism of
683 Chlamydomonas reinhardtii. Proc. Natl. Acad. Sci. U. S. A. *116*, 16915–16920.

684 16. Sinetova, M.A., Kupriyanova, E.V., Markelova, A.G., Allakhverdiev, S.I., and Pronina,
685 N.A. (2012). Identification and functional role of the carbonic anhydrase Cah3 in
686 thylakoid membranes of pyrenoid of Chlamydomonas reinhardtii. Biochim. Biophys.
687 Acta *1817*, 1248–1255.

688 17. Fei, C., Wilson, A.T., Mangan, N.M., and Wingreen, N.S. (2022). Modelling the
689 pyrenoid-based CO2-concentrating mechanism provides insights into its operating
690 principles and a roadmap for its engineering into crops. Nature Plants.

691 18. Toyokawa, C., Yamano, T., and Fukuzawa, H. (2020). Pyrenoid Starch Sheath Is
692 Required for LCIB Localization and the CO2-Concentrating Mechanism in Green Algae.
693 Plant Physiol. 10.1104/pp.19.01587.

694 19. Mackinder, L.C.M., Chen, C., Leib, R.D., Patena, W., Blum, S.R., Rodman, M.,
695 Ramundo, S., Adams, C.M., and Jonikas, M.C. (2017). A Spatial Interactome Reveals
696 the Protein Organization of the Algal $CO_2$-Concentrating Mechanism. Cell *171*, 133-
697 147.e14.

698 20. Lau, C.S., Dowle, A., Thomas, G.H., Girr, P., and Mackinder, L.C.M. (2023). A phase-
699 separated CO2-fixing pyrenoid proteome determined by TurboID in Chlamydomonas
700 reinhardtii. Plant Cell. 10.1093/plcell/koad131.

701 21. Atkinson, N., Mao, Y., Chan, K.X., and McCormick, A.J. (2020). Condensation of
702 Rubisco into a proto-pyrenoid in higher plant chloroplasts. Nat. Commun. *11*, 6303.

703 22. Adler, L., Díaz-Ramos, A., Mao, Y., Pukacz, K.R., Fei, C., and McCormick, A.J. (2022).
704 New horizons for building pyrenoid-based CO2-concentrating mechanisms in plants to
705 improve yields. Plant Physiol. *190*, 1609–1627.

706 23. Nakajima, K., Tanaka, A., and Matsuda, Y. (2013). SLC4 family transporters in a marine
707 diatom directly pump bicarbonate from seawater. Proc. Natl. Acad. Sci. U. S. A. *110*,
708 1767–1772.

709 24. Nawaly, H., Matsui, H., Tsuji, Y., Iwayama, K., Ohashi, H., Nakajima, K., and Matsuda,
710 Y. (2023). Multiple plasma membrane SLC4s contribute to external HCO3- acquisition

711 during CO2 starvation in the marine diatom Phaeodactylum tricornutum. J. Exp. Bot. *74*,
712 296–307.

713 25. Kikutani, S., Nakajima, K., Nagasato, C., Tsuji, Y., Miyatake, A., and Matsuda, Y.
714 (2016). Thylakoid luminal θ-carbonic anhydrase critical for growth and photosynthesis in
715 the marine diatom Phaeodactylum tricornutum. Proc. Natl. Acad. Sci. U. S. A. *113*,
716 9828–9833.

717 26. Shimakawa, G., Okuyama, A., Harada, H., Nakagaito, S., Toyoshima, Y., Nagata, K.,
718 and Matsuda, Y. (2023). Pyrenoid-core CO2-evolving machinery is essential for diatom
719 photosynthesis in elevated CO2. Plant Physiol. 10.1093/plphys/kiad475.

720 27. Matsuda, Y., Hopkinson, B.M., Nakajima, K., Dupont, C.L., and Tsuji, Y. (2017).
721 Mechanisms of carbon dioxide acquisition and CO2 sensing in marine diatoms: a
722 gateway to carbon metabolism. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*.
723 10.1098/rstb.2016.0403.

724 28. Nam, O., Grouneva, I., and Mackinder, L.C.M. (2022). Endogenous GFP tagging in the
725 diatom Thalassiosira pseudonana. bioRxiv, 2022.09.30.510313.
726 10.1101/2022.09.30.510313.

727 29. Nigishi, M., Shimakawa, G., Yamagishi, K., Amano, R., Ito, S., Tsuji, Y., Nagasato, C.,
728 and Matsuda, Y. (2023). Low-CO2 inducible bestrophins in diatom thylakoid membranes
729 sustain high photosynthetic efficacy at distant locations from the pyrenoid. bioRxiv,
730 2023.09.20.558591. 10.1101/2023.09.20.558591.

731 30. Wang, S., Wu, R., Lu, J., Jiang, Y., Huang, T., and Cai, Y.-D. (2022). Protein-protein
732 interaction networks as miners of biological discovery. Proteomics *22*, e2100190.

733 31. Huang, and Daboussi (2017). Genetic and metabolic engineering in diatoms. Philos.
734 Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160411.

735 32. Huesgen, P.F., Alami, M., Lange, P.F., Foster, L.J., Schröder, W.P., Overall, C.M., and
736 Green, B.R. (2013). Proteomic amino-termini profiling reveals targeting information for
737 protein import into complex plastids. PLoS One *8*, e74483.

738 33. Clement, R., Dimnet, L., Maberly, S.C., and Gontero, B. (2016). The nature of the CO2-
739 concentrating mechanisms in a marine diatom, Thalassiosira pseudonana. New Phytol.
740 10.1111/nph.13728.

741 34. Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human
742 deubiquitinating enzyme interaction landscape. Cell.

743 35. Choi, H., Larsen, B., Lin, Z.-Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z.S.,
744 Tyers, M., Gingras, A.-C., and Nesvizhskii, A.I. (2010). SAINT: probabilistic scoring of
745 affinity purification–mass spectrometry data. Nat. Methods *8*, 70–73.

746 36. Mueller-Cajar, O., Stotz, M., Wendler, P., Hartl, F.U., Bracher, A., and Hayer-Hartl, M.
747 (2011). Structure and function of the AAA+ protein CbbX, a red-type Rubisco activase.
748 Nature *479*, 194–199.

749 37. Meyer, M.T., Whittaker, C., and Griffiths, H. (2017). The algal pyrenoid: key unanswered
750 questions. J. Exp. Bot. *68*, 3739–3749.

751 38. Kroth, P.G., Chiovitti, A., Gruber, A., Martin-Jezequel, V., Mock, T., Parker, M.S.,
752 Stanley, M.S., Kaplan, A., Caron, L., Weber, T., et al. (2008). A model for carbohydrate

753 metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole
754 genome analysis. PLoS One *3*, e1426.

755 39. Rae, B.D., Long, B.M., Badger, M.R., and Price, G.D. (2013). Functions, compositions,
756 and evolution of the two types of carboxysomes: polyhedral microcompartments that
757 facilitate $CO_2$ fixation in cyanobacteria and some proteobacteria. Microbiol. Mol. Biol.
758 Rev. *77*, 357–379.

759 40. Kerfeld, C.A., and Melnicki, M.R. (2016). Assembly, function and evolution of
760 cyanobacterial carboxysomes. Curr. Opin. Plant Biol. *31*, 66–75.

761 41. Cameron, J.C., Wilson, S.C., Bernstein, S.L., and Kerfeld, C.A. (2013). Biogenesis of a
762 bacterial organelle: the carboxysome assembly pathway. Cell *155*, 1131–1140.

763 42. Long, B.M., Rae, B.D., Rolland, V., Förster, B., and Price, G.D. (2016). Cyanobacterial
764 CO2-concentrating mechanism components: function and prospects for plant metabolic
765 engineering. Curr. Opin. Plant Biol. *31*, 1–8.

766 43. Metskas, L.A., Ortega, D., Oltrogge, L.M., Blikstad, C., Lovejoy, D.R., Laughlin, T.G.,
767 Savage, D.F., and Jensen, G.J. (2022). Rubisco forms a lattice inside alpha-
768 carboxysomes. Nat. Commun. *13*, 4863.

769 44. Ni, T., Sun, Y., Burn, W., Al-Hazeem, M.M.J., Zhu, Y., Yu, X., Liu, L.-N., and Zhang, P.
770 (2022). Structure and assembly of cargo Rubisco in two native α-carboxysomes. Nat.
771 Commun. *13*, 4299.

772 45. Jensen, E.L., Clement, R., Kosta, A., Maberly, S.C., and Gontero, B. (2019). A new
773 widespread subclass of carbonic anhydrase in marine phytoplankton. ISME J. *13*,
774 2094–2106.

775 46. Yee, D.P., Samo, T.J., Abbriano, R.M., Shimasaki, B., Vernet, M., Mayali, X., Weber,
776 P.K., Mitchell, B.G., Hildebrand, M., Decelle, J., et al. (2023). The V-type ATPase
777 enhances photosynthesis in marine phytoplankton and further links phagocytosis to
778 symbiogenesis. Curr. Biol. 10.1016/j.cub.2023.05.020.

779 47. Guillard, R.R., and Ryther, J.H. (1962). Studies of marine planktonic diatoms. I.
780 Cyclotella nana Hustedt, and Detonula confervacea (cleve) Gran. Can. J. Microbiol. *8*,
781 229–239.

782 48. Guillard, R.R.L. (1975). Culture of Phytoplankton for Feeding Marine Invertebrates. In
783 Culture of Marine Invertebrate Animals: Proceedings — 1st Conference on Culture of
784 Marine Invertebrate Animals Greenport, W. L. Smith and M. H. Chanley, eds. (Springer
785 US), pp. 29–60.

786 49. Weber, E., Engler, C., Gruetzner, R., Werner, S., and Marillonnet, S. (2011). A modular
787 cloning system for standardized assembly of multigene constructs. PLoS One *6*,
788 e16765.

789 50. Karas, B.J., Diner, R.E., Lefebvre, S.C., McQuaid, J., Phillips, A.P.R., Noddings, C.M.,
790 Brunson, J.K., Valas, R.E., Deerinck, T.J., Jablanovic, J., et al. (2015). Designer diatom
791 episomes delivered by bacterial conjugation. Nat. Commun. *6*, 6925.

792 51. Strand, T.A., Lale, R., Degnes, K.F., Lando, M., and Valla, S. (2014). A new and
793 improved host-independent plasmid system for RK2-based conjugal transfer. PLoS One
794 *9*, e90372.

52. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

53. Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. *30*, 70–82.

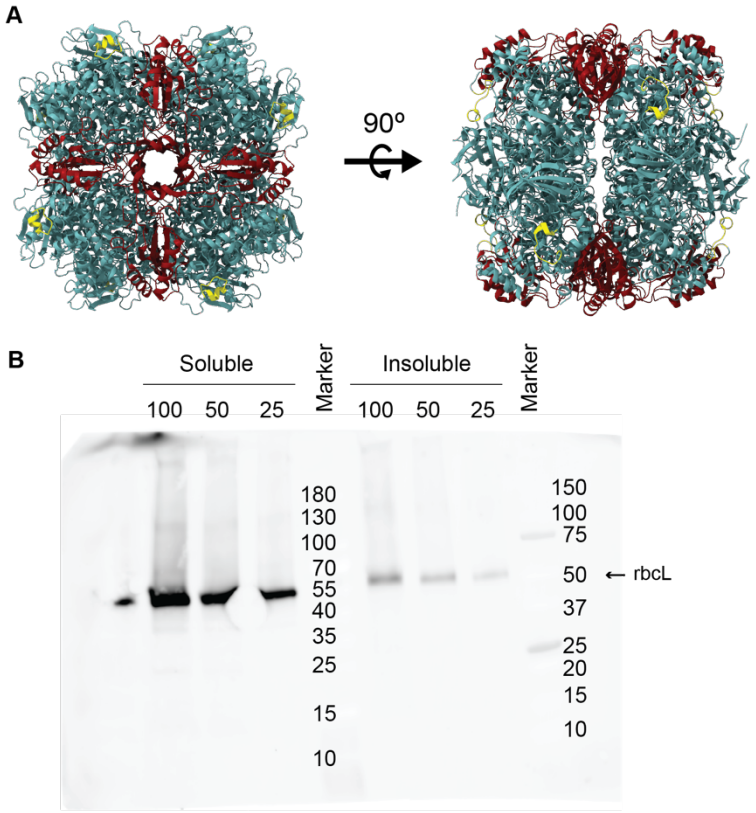



**Figure S1. RbcL antibody**

A. An antibody was raised against a surface exposed region (yellow) of the rbcL (green). RbcS is shown in red.

B. Immunoblot using the rbcL antibody against the soluble and insoluble cell fractions of *T. pseudonana*.

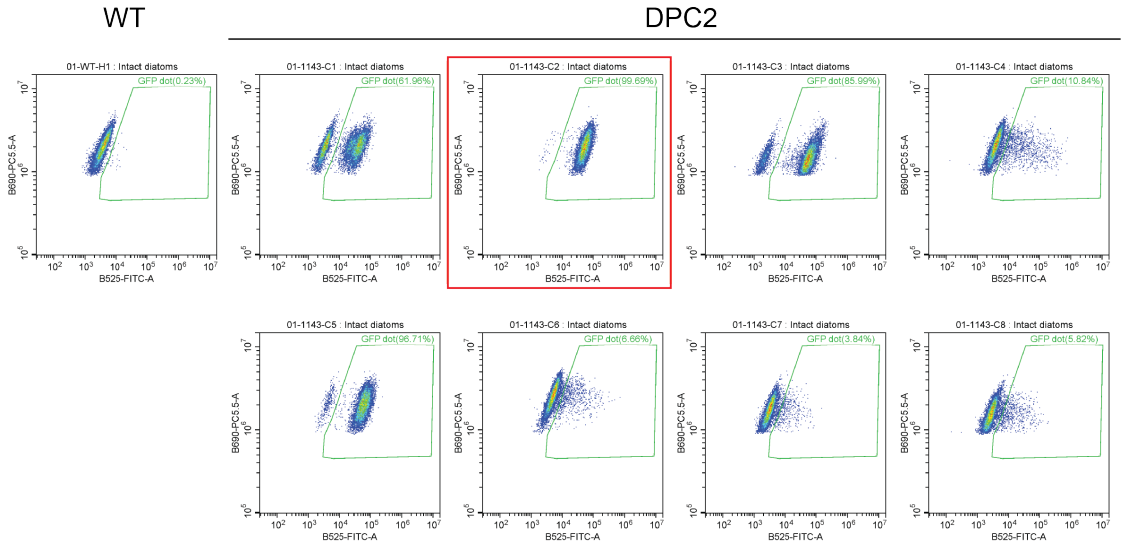
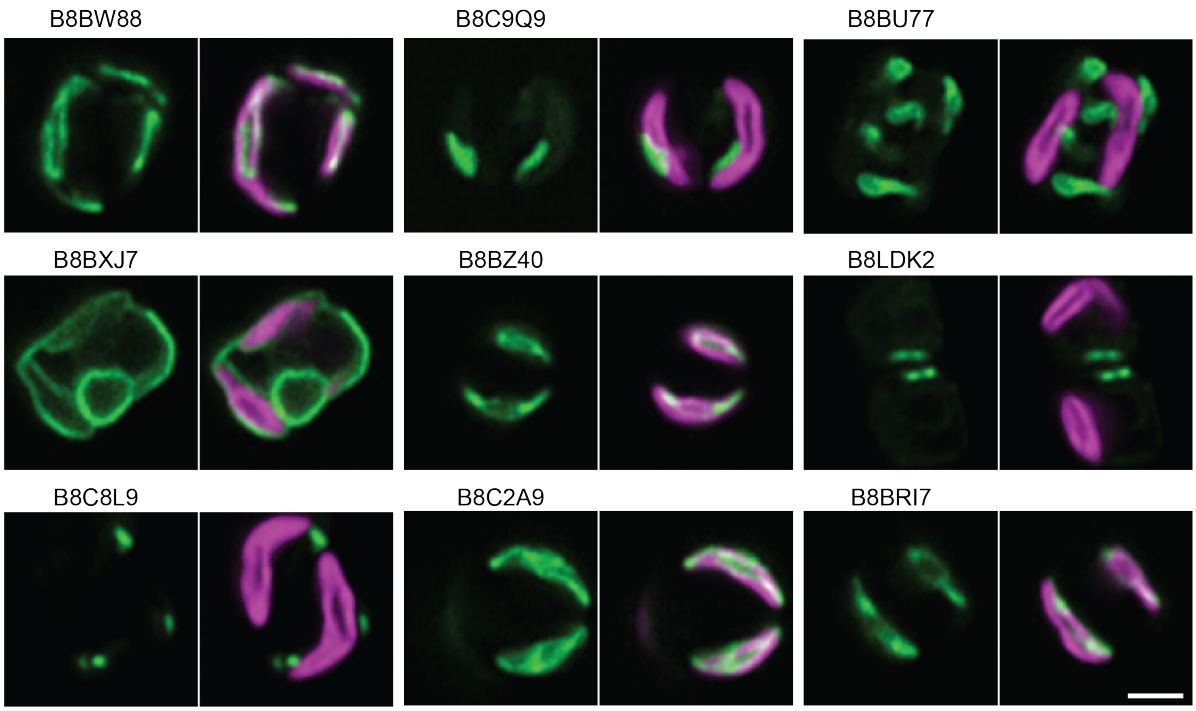



**Figure S2. Flow cytometry screening to identify fluorophore fusion lines.** Example data for Diatom Pyrenoid Component 2 (DPC2). Red box denotes cell line used for subsequent imaging and APMS.




B8BW88: Aspartate/glutamate/uridylate kinase domain-containing protein

B8BXJ7: Uncharacterized protein

B8C8L9: Gfo/Idh/MocA-like oxidoreductase N-terminal domain-containing protein

B8C9Q9: Oxoglutarate/matate translocator

B8BZ40: Phosphoribulokinase

B8C2A9: S-malonyltransferase

B8BU77: Nucleoside-diphosphate kinase

B8LDK2: Glycoside hydrolase family 5 domain-containing protein

B8BRI7: 2-isopropylmalate synthase

**Figure S3. mEGFP fusion proteins that did not localize to the pyrenoid.** Green: mEGFP fusion protein; Magenta: chlorophyll. Scales bar: 2 μm.

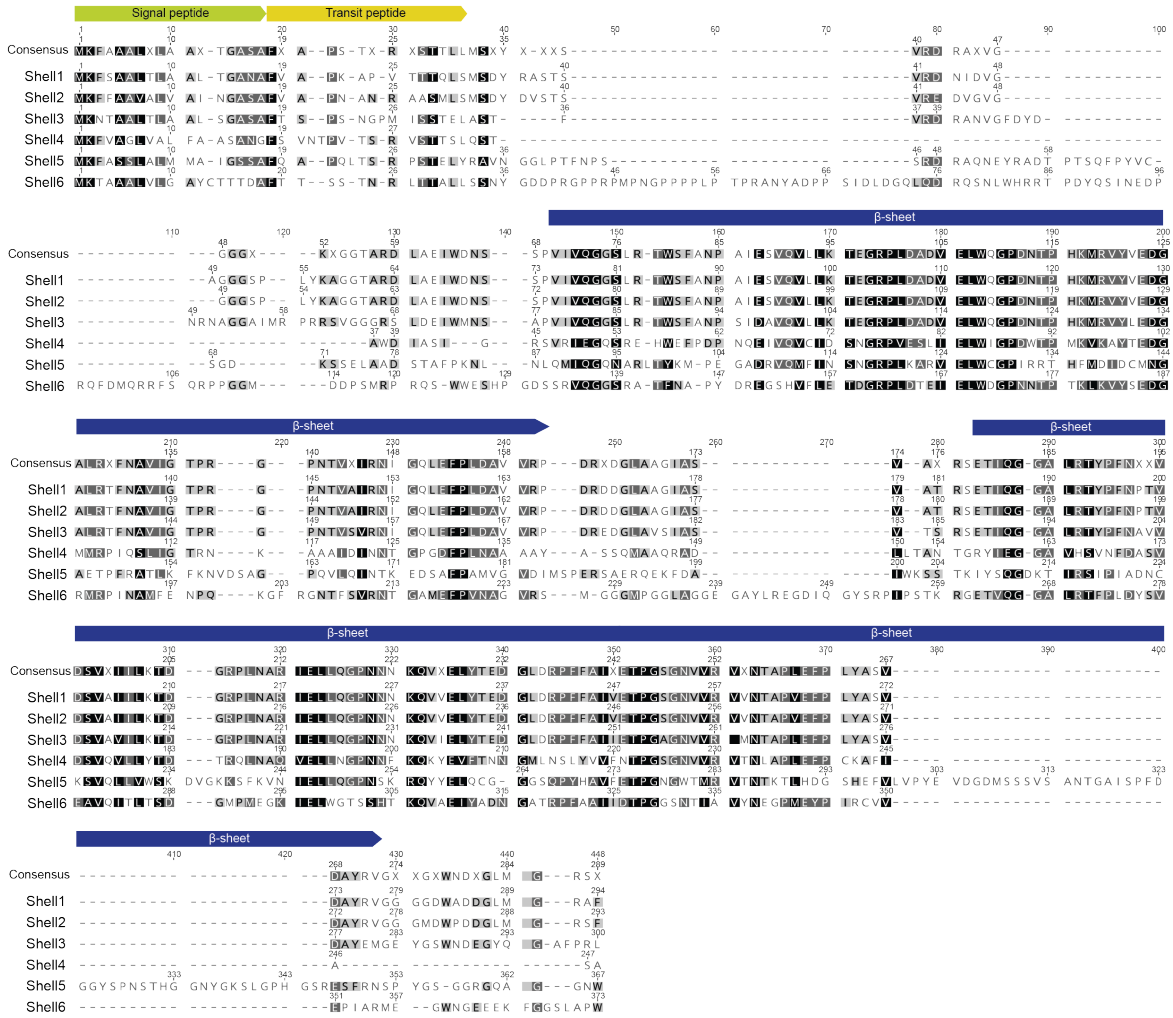



**Figure S4. Shell protein amino acid alignment.**

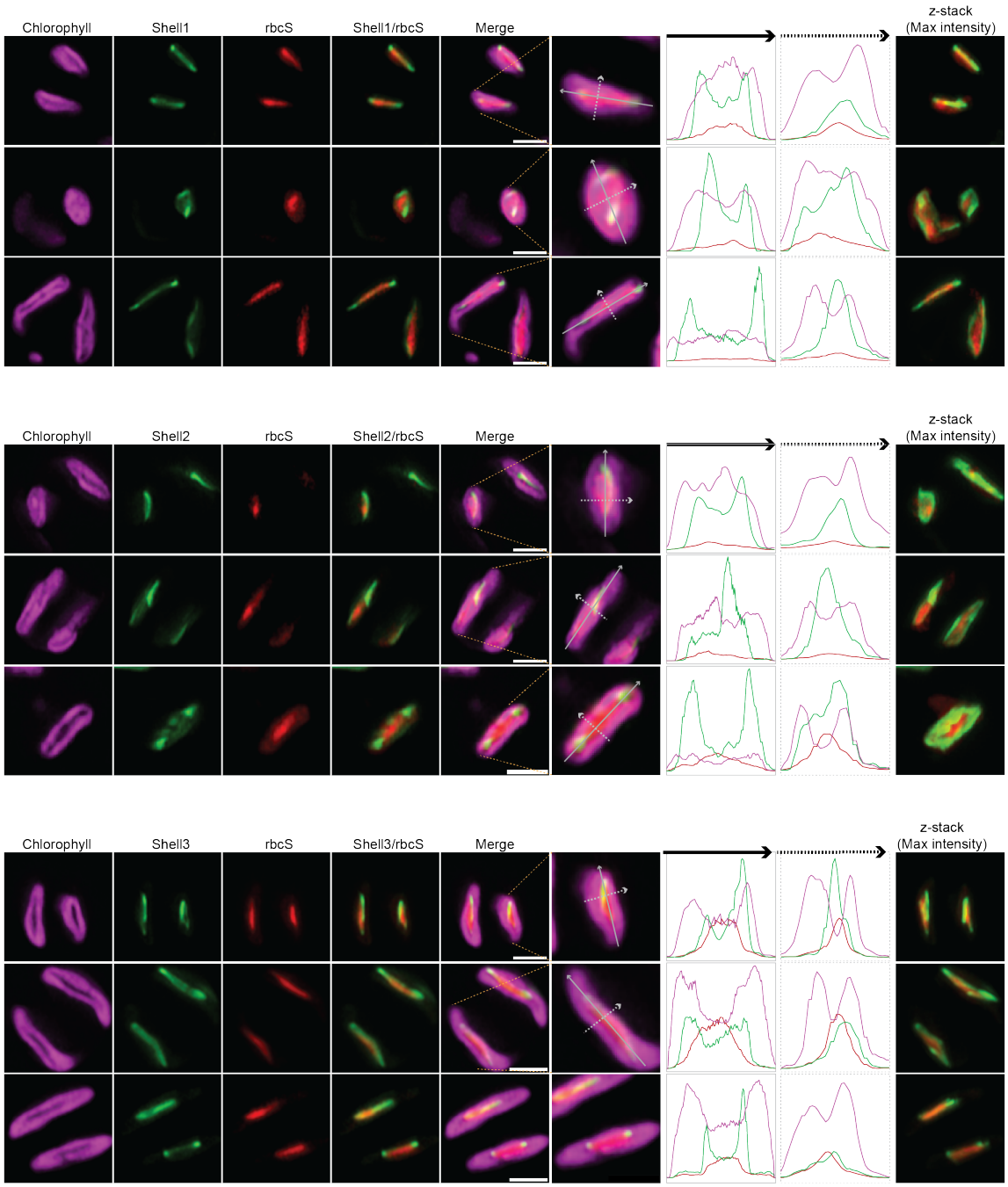



**Figure S5 (continued on next page).**

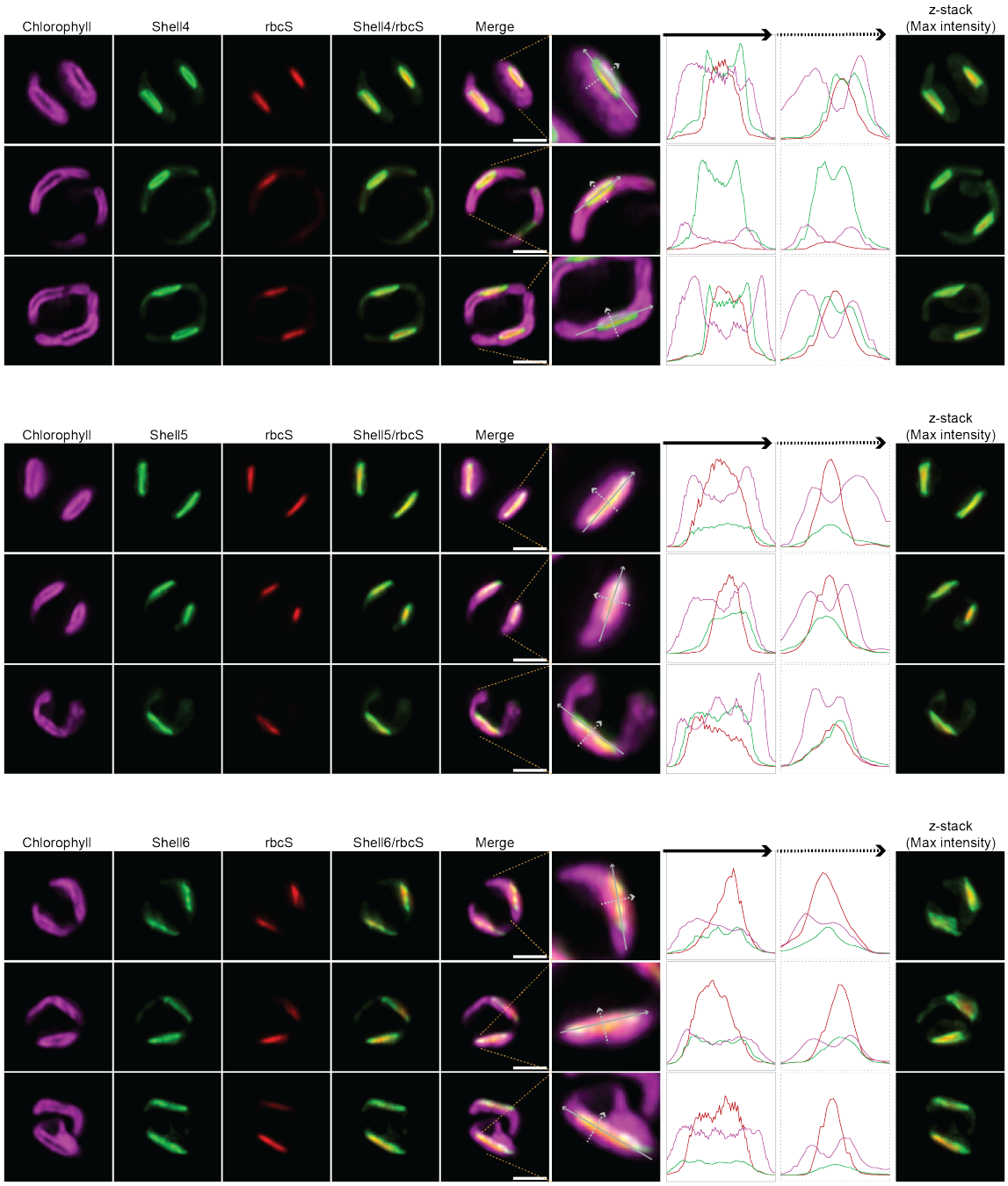



**Figure S5 (continued from previous page). Co-localization of Shell proteins with Rubisco.** Image columns 1-4: additional images of co-localization of Shell proteins with Rubisco. Scale bars: 2µm. Image columns 5-8: Fluorescence intensity cross-sections of Shell and rbcS fluorescence across the pyrenoid. Level of Zoom changes between images. Image column 9: Z-stack max intensity projections.

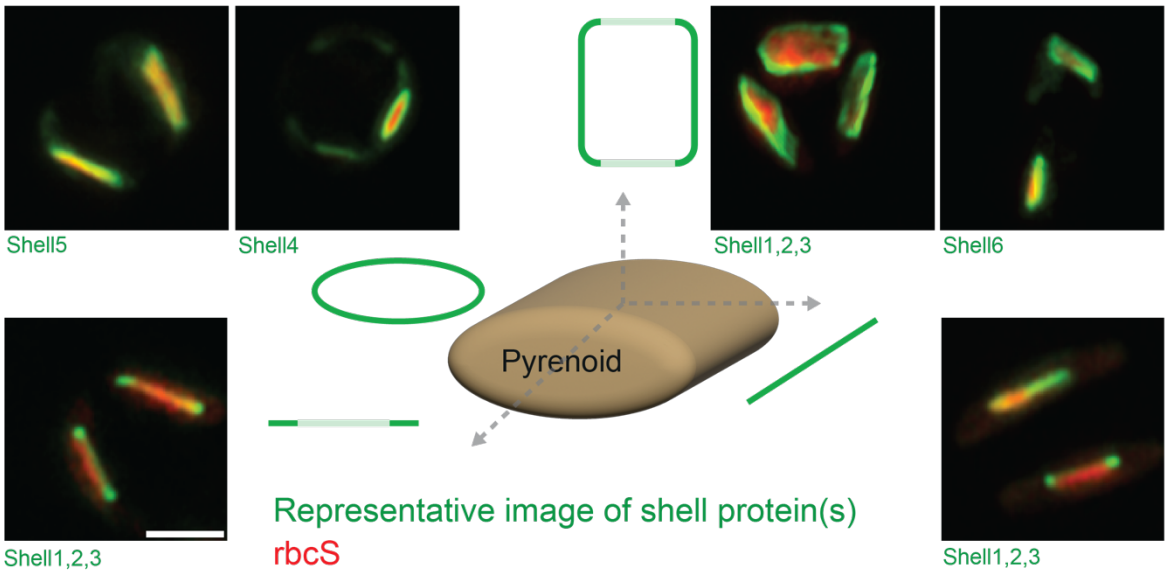



**Figure S6. Shell proteins localize to different sub-regions of the pyrenoid.** Representative images from Figure 4 illustrating how different shell proteins are organized around the pyrenoid.

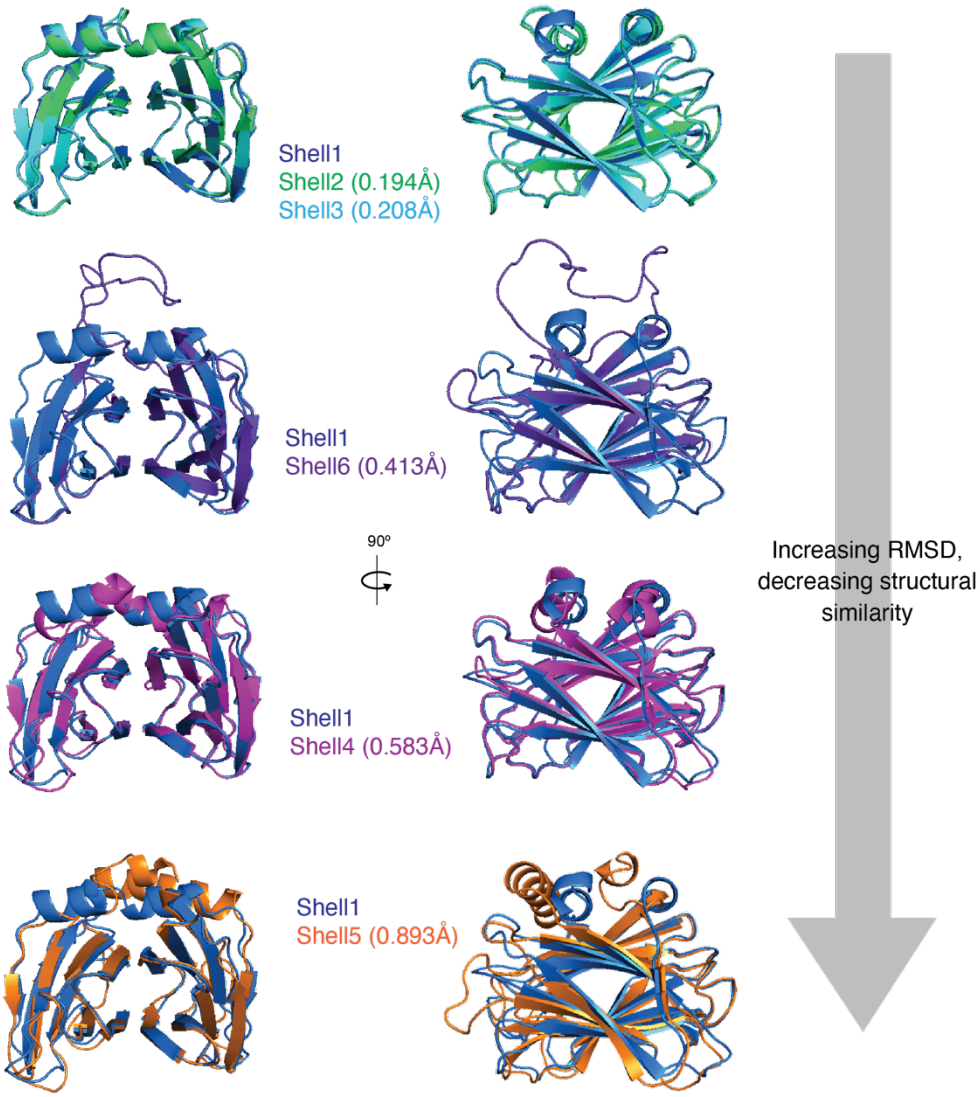



**Figure S7. AlphaFold2 model comparisons of Shell proteins.** Root mean square deviation values relative to Shell1 are shown.

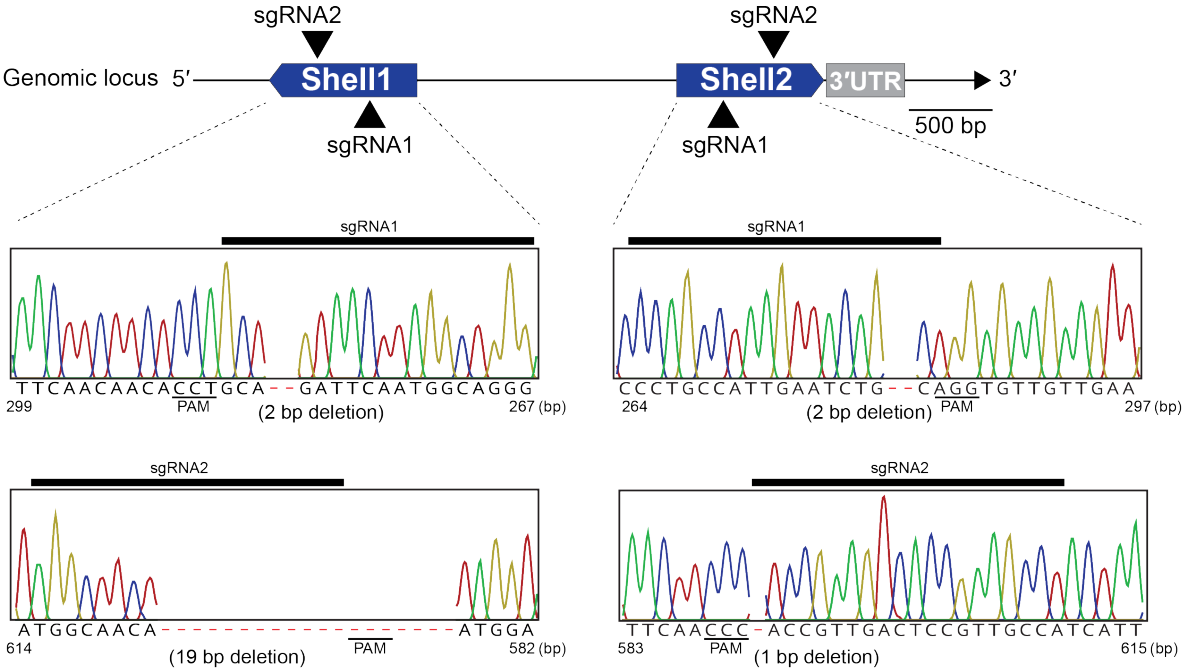



**Figure S8. Shell1/2 knock out genotyping.** Due to 93% DNA sequence similarity between Shell1 and Shell2 both genes we knocked out simultaneously by designing two sgRNAs that targeted conserved sequences in Shell1 and Shell2.