A Protein Blueprint of the Diatom CO_2-Fixing Organelle

Onyou Nam[1,2], Sabina Musial[1,2], Manon Demulder[3], Caroline McKenzie[1,2], Adam Dowle[1], Matthew Dowson[1,2], James Barrett[1,2], James N. Blaza[4], Benjamin D. Engel[3], Luke C. M. Mackinder[1,2]

[1]Department of Biology, University of York, York YO10 5DD, UK.
[2]Centre for Novel Agricultural Products, Department of Biology, University of York, York YO10 5DD, UK.
[3]Biozentrum, University of Basel, Spitalstrasse 41, 4056 Basel, Switzerland.
[4]York Structural Biology Laboratory, University of York, York, YO10 5DD, UK.

Authors for Correspondence: luke.mackinder@york.ac.uk and ben.engel@unibas.ch

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 interaction network are 10 proteins with no known function. We show that six of these form a static shell encapsulating the Rubisco matrix of the pyrenoid, with the shell critical for pyrenoid structural integrity, shape, and function. Although not conserved at a sequence level, the diatom pyrenoid shares some architectural similarities to prokaryotic carboxysomes. Collectively, our results support the convergent evolution of pyrenoids across the two main plastid lineages and uncover 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 principal 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 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.[7,8] However, it is still unclear the level of conservation of structural proteins.[9] 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.[22,23] 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.[18] Although, supporting functional data is missing for these proteins in *T. pseudonana*. 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[28] 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*, which belongs to a global biogeochemically important genus. These data enabled us to build a high-confidence diatom pyrenoid interaction network, 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 is critical for pyrenoid shape, structural integrity, and CCM function. Our findings provide new insight into a globally 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, each surrounded by four 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), using the main pyrenoid component, Rubisco, as a bait protein (Fig. 1B). To immunoprecipitate 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 two and three technical replicates, respectively, against non-antibody control experiments, we identified 36 putative pyrenoid components out of a total of 1167 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 being predicted to be a membrane protein (Fig. 1B). Top hits from our rbcL coIPMS experiments were then fed into an iterative fluorescent protein tagging, localization, and APMS framework that we used to rapidly build a spatially defined pyrenoid proteome (Fig. 1C).

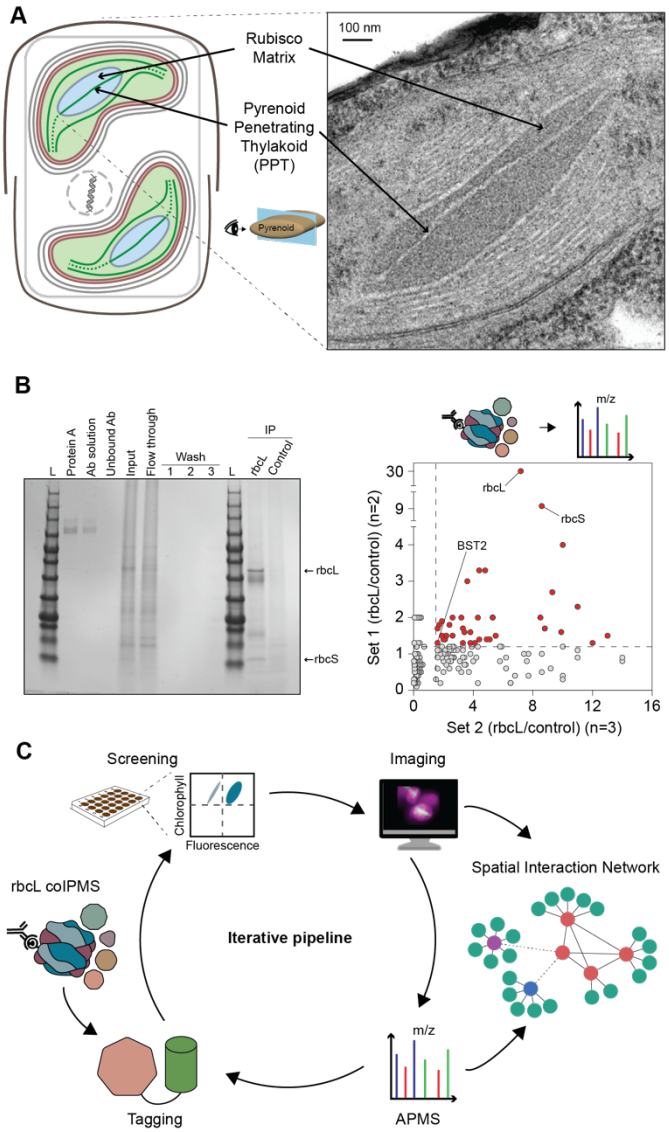


Figure 1. Identification of candidate diatom pyrenoid components.
A. The diatom *T. pseudonana* has two chloroplasts each containing a single pyrenoid traversed by a specialized membrane called the pyrenoid penetrating thylakoid (PPT). The connection between the PPT and the wider thylakoid network is unresolved.
B. Rubisco co-immunoprecipitation followed by mass spectrometry (coIPMS) to identify candidate pyrenoid components. Left: example Rubisco large subunit (rbcL) coIP with different steps resolved by SDS-PAGE. Right: Two experiments consisting of either two or three technical coIPs were 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. Ab, antibody; L, protein ladder.
C. An iterative pipeline to determine the pyrenoid spatial interaction network. RbcL 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

141 Modular Cloning-based episomal assembly framework[30] to be 96-well compatible and
142 combined it with multi-well diatom transformation via bacterial conjugation. We coupled this
143 with 48-well plate strain maintenance and 96-well plate flow cytometry screening for clonal
144 fluorophore-fusion expressing lines (Fig. 2A). As coIPMS data on a small-scale is inherently
145 noisy with both false positives and false negatives[31] we applied our tagging pipeline to 22
146 Rubisco coIPMS hits to validate if they were bona fide pyrenoid proteins. Nourseothricin-
147 positive transformants were picked and screened for positive fluorescence using flow
148 cytometry. Due to typical mosaic colony presence[12], either multiple rounds of screening or
149 screening of several independent colonies was required to identify stable mEGFP expressing
150 lines. We found that screening 8–12 colonies would typically yield a stable cell population with
151 >90% of cells mEGFP positive (Fig. S2). Positive lines were subsequently imaged by confocal
152 microscopy (Fig. 2B, Fig. S3 and S4). From the initially identified Rubisco-interacting proteins,
153 we saw 13 proteins localizing to distinct sub-regions of the pyrenoid (Fig. 2B and Fig. S4; see
154 below for further discussion). However, 9 candidates were either localized to chloroplast sub-
155 regions adjacent to the pyrenoid (4), or to distinct non-chloroplast regions (5) (Fig. S3)
156 indicating that our coIPMS data contains false positives.
157 To accurately determine sub-pyrenoid localization of components, we developed a
158 dual-tagging approach using two spectrally compatible fluorophores. We first developed a
159 pyrenoid matrix marker for co-localization. In green algae, nuclear-encoded rbcS-fluorescent
160 protein fusions have been powerful for understanding sub-pyrenoid spatial organization[31] and
161 for determining the liquid-like properties of the pyrenoid.[32] As the rbcS of diatoms is chloroplast
162 encoded, and no *T. pseudonana* chloroplast transformation protocol is available, we wondered
163 if we could target an episomal expressed rbcS-mEGFP to the pyrenoid. Using the N-terminal
164 signal and transit peptide sequences from the nuclear encoded chloroplast localized BST2
165 protein,[28] we successfully targeted rbcS to the pyrenoid, allowing us to clearly define the
166 pyrenoid matrix (Fig. 2C). Second, we tested assembling two target genes with different
167 fluorophores on the same episome. We decided to initially validate the BST2 localization by
168 making an episome with BST2-mEGFP and rbcS-mScarlet-I, a fluorophore we had previously
169 validated using our system.[33] Using this approach, we observed a clear pyrenoid localization
170 of BST2, with the rbcS signal extending outside that of BST2, supporting a PPT localization of
171 BST2 (Fig. 2D).
172

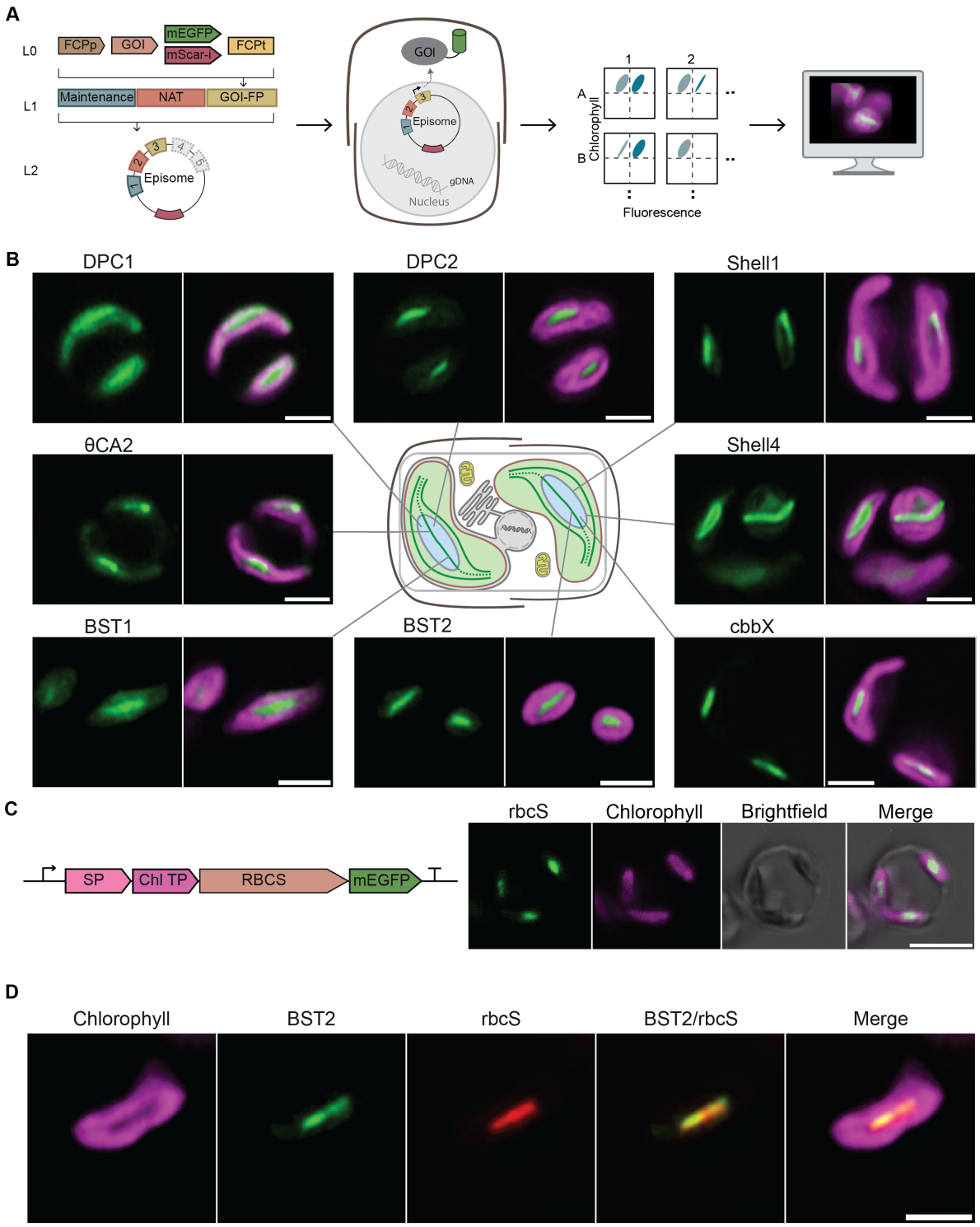


Figure 2. High-throughput fluorescence protein tagging identifies multiple new diatom pyrenoid components.

A. Candidate genes were rapidly cloned using Golden Gate modular cloning, transformed into *T. pseudonana* via bacterial conjugation and screened by flow cytometry for fluorescence prior to imaging. Gene of interest (GOI) expression was driven by the fucoxanthin chlorophyll a/c-binding protein promoter (FCPp) and terminated using the FCP terminator (FCPt).

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 interaction network

To rapidly expand the pyrenoid proteome we developed and optimized an APMS pipeline in *T. pseudonana* and used our GFP-tagged pyrenoid proteins as baits in this pipeline. Lines expressing GFP-tagged proteins were typically grown at ambient CO_2, where the CCM is fully active.[19,34] 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 defined by comparing both CompPASS[35] and SAINT analysis[28] 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.

Our initial 13 pyrenoid localized proteins 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;[25,26] θCA2 most likely involved in CO_2 release from HCO_3^- within the PPTs;[36] cbbX a nuclear encoded red-type Rubisco activase[37] that until now has not been localized in algae with red plastids;[1,38] 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. 2B; see below). A subset of these pyrenoid-localized components were utilized for APMS using our iterative pipeline (Fig. 1C, Fig. 3B and Table S4). Subsequently, two additional components, DPC3 and DPC4, which had strong interactions with Shell4 and rbcS but again with no sequence predictable function, were localized and fed into our APMS pipeline. Whereas DPC4 was found throughout the pyrenoid with a matrix-type localization, DPC3 appeared to encapsulate the pyrenoid similar to the Shell proteins (Fig. 3C). Combining the data and using our stringent interaction scoring approach enabled us to expand and build a high-confidence pyrenoid interaction network 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. 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,[39–41] potentially indicating that pyrenoid assembly in *T. pseudonana* could be based on different biophysical principles to both green algal and pennate diatom pyrenoids. The localization of DPC2 solely to the pyrenoid

(Fig. 2B), its Rubisco interaction (Fig. 3B), and its relatively high abundance from whole cell mass spectrometry (Fig. 3D and Table S5) make DPC2 a strong candidate for future studies.

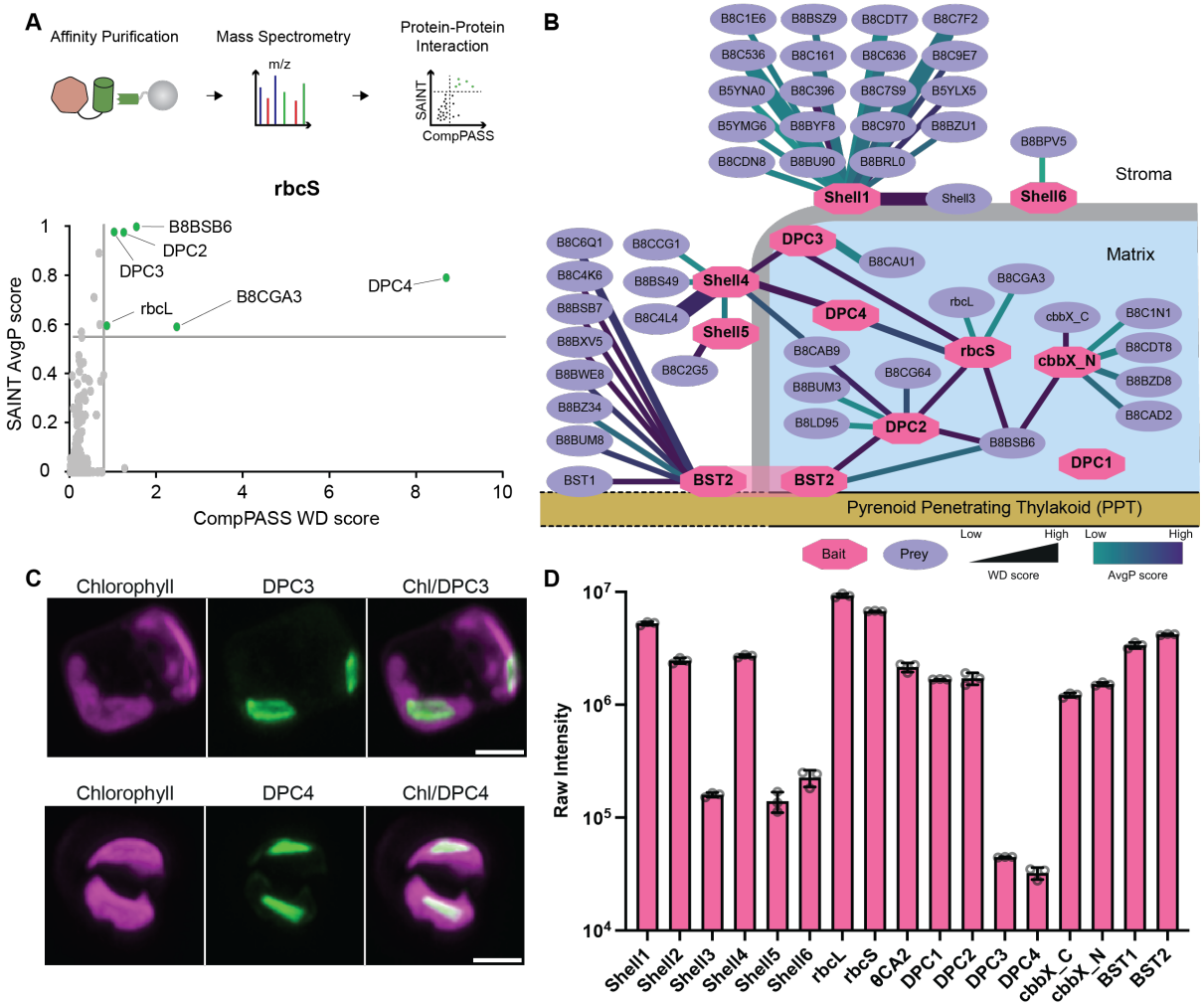


Figure 3. A spatial interaction network of the diatom pyrenoid.

A. Top: 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. Bottom: analysis example for RbcS APMS. 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 interaction network of the pyrenoid. Where available, 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 in the APMS study.

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

D. Whole cell mass spectrometry raw intensity values of pyrenoid confirmed proteins. Data is from wildtype cells ran in triplicate. Error bars: S.D. of the mean.

T. pseudonana has six Shell homologs, and Shell proteins are found across algal lineages containing red plastid secondary endosymbionts

In our initial tagging, we were intrigued to see that two proteins with no annotated functional domains appeared to encapsulate the pyrenoid matrix (Fig. 2B: Shell1, Shell4). 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, which is stored in cytosolic vacuoles[40–43] with no clear carbohydrate barrier surrounding the diatom pyrenoid. Pyrenoid modeling of inorganic carbon fluxes has shown that a CO_2 diffusion barrier is critical for an efficient CCM,[17] thus opening up the question of how diatoms and other algal lineages that lack starch encapsulated pyrenoids operate efficient CCMs.[43–45] Although protein encapsulation of pyrenoids has not previously been identified, cyanobacterial Rubisco-containing carboxysomes have a well-characterized protein shell that is proposed to act as a CO_2 barrier to minimize leakage,[44–46] and is essential for carboxysome biogenesis and shape.[46] Instead of starch, we hypothesized a protein shell could have an analogous function in diatoms. We set out to further understand the importance of the Shell proteins for pyrenoid function.

BLAST analysis of Shell1 and Shell4 identified four additional homologs in the *T. pseudonana* genome that all contain two predicted beta-sheet domains (Fig. 4A and Fig. S5) and are predicted to be structurally similar to each other but distinct from DPC3 (Fig. S6). We explored the distribution of these proteins across different evolutionary lineages. BLAST analysis against the NCBI database identified homologs in stramenopiles (including other diatoms), pelagophytes and haptophytes; all of which are photosynthetic algae that contain red plastids originated from secondary endosymbiotic events (Fig. 4B and Table S6). However, Shell proteins are absent in the rhodophytes, the primary endosymbiotic red plastid lineage that was engulfed by a heterotrophic host to form the red plastid secondary endosymbiotic lineages. This suggests that Shell proteins were either not present in the engulfed red alga but may have been present in the genome of the heterotrophic host prior to endosymbiosis, were present in the engulfed red alga but have since been lost in rhodophytes, or evolved after engulfment. Given that Shells are not found in extant non-photosynthetic organisms and are not present in pyrenoid-containing rhodophytes, it seems plausible that they evolved following the secondary endosymbiotic event. 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 possess pyrenoids.[47]

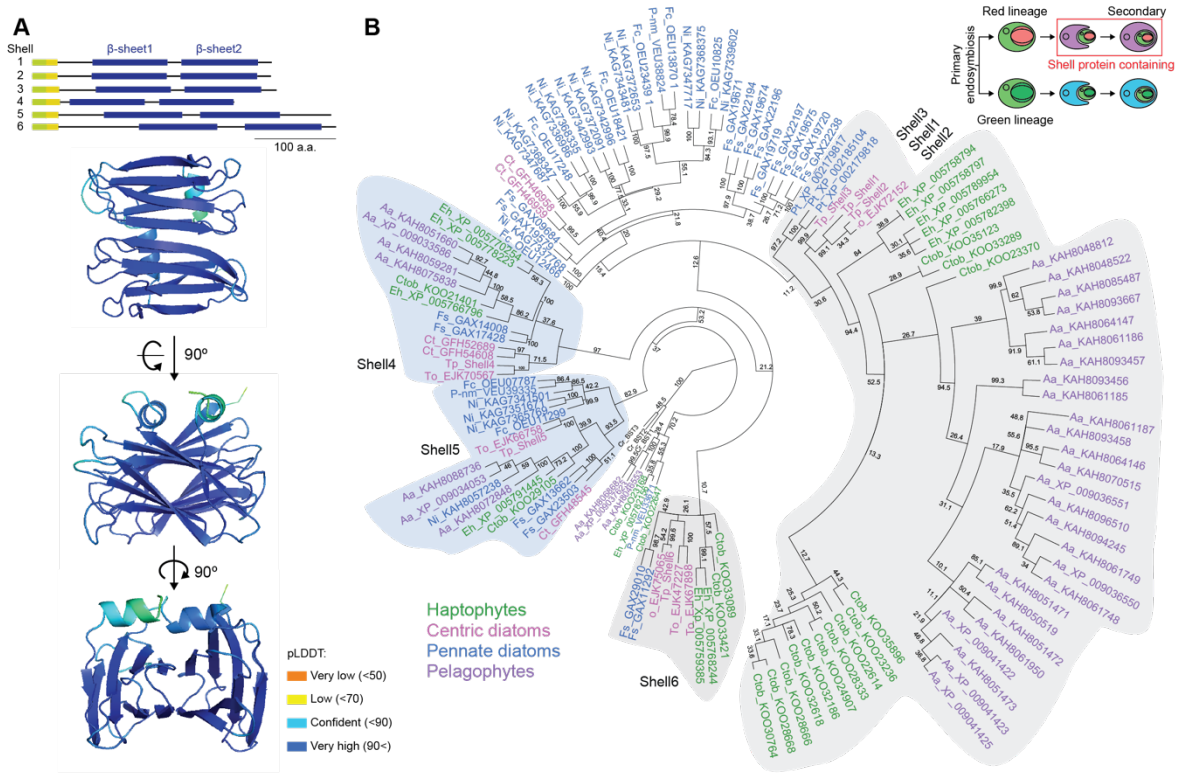


Figure 4. *T. pseudonana* has six Shell homologs and Shell proteins are found across red plastid secondary endosymbionts

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. The N-terminal and C-terminal disordered regions are not shown due to low confidence structural prediction.

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. Aa, *Aureococcus anophagefferens*; Ct, *Chaetoceros tenuissimus*; Ctob, *Chrysochromulina tobinii*; Eh, *Emiliania huxleyi*; Fs, *Fistulifera solaris*; Fc, *Fragilariopsis cylindrus*; Ni, *Nitzschia inconspicua*; Pt, *Phaeodactylum tricornutum*; P-nm, *Pseudo-nitzschia multistriata*; To, *Thalassiosira oceanica*; Tp, *Thlassiosira pseudonana*. *Chlamydomonas reinhardtii* bestrophins (Cr_BST1-3) have been used as an outgroup.

Shell proteins encapsulate the pyrenoid matrix and localize to distinct sub-regions

To elucidate the precise sub-regions of the six Shell proteins, we co-expressed rbcS-mScarlet-I with mEGFP-tagged Shell proteins (Fig. 5A and Fig. S7). Analysis of fluorescence intensity of pyrenoid transects (Fig. 5B and Fig. S7) and max intensity z-stack projections (Fig. 5C and Fig. S7) indicate that all six proteins encapsulate the Rubisco matrix of the pyrenoid. To prioritize Shell proteins for further investigation, we looked at their relative abundance in our whole cell mass spectrometry data performed on wildtype cells (Fig. 3D and Table S5). The most abundant homologs were Shell1 and Shell2, which share 92% identity, and Shell4, which shares 34% identity with Shell1 (Fig. S5B). As the diatom pyrenoid shape can be generalized as an elliptic cylinder with curved ends (Fig. 1), the apparent localization of the Shell proteins is highly dependent on the orientation of the chloroplast and pyrenoid during imaging (Fig. 5A-C and Fig. S7), making the assignment to sub-pyrenoid surface regions challenging. To explore if Shell1 and Shell4 occupy the same regions of the shell we co-expressed Shell1 and Shell4 with different fluorescent tags. Imaging shows that they localize

to distinct regions of the pyrenoid periphery, suggesting that they may have different roles in the formation or function of the shell (Fig. 5D-F).

Although Shell proteins were identified in our rbcL coIPMS and rbcS APMS data, they did not fall above the stringent thresholds we defined for high-confidence interactors with either rbcS or rbcL. Their pyrenoid-peripheral localization and low confidence Rubisco interaction suggest Shell proteins could interact indirectly with the Rubisco matrix. The shell-like pattern displayed by DPC3 (Fig. 3C) raises the possibility that this protein could be acting as an intermediary Rubisco matrix-shell adaptor or potentially an additional shell component. The co-expression of DPC3 with Shell4 shows that they co-localize (Fig. 5G-I). The relatively low protein abundance of DPC3 in comparison to Shell1 and Shell4 (Fig. 3D) indicates that DPC3 is most likely a minor protein constituent of the pyrenoid surface.

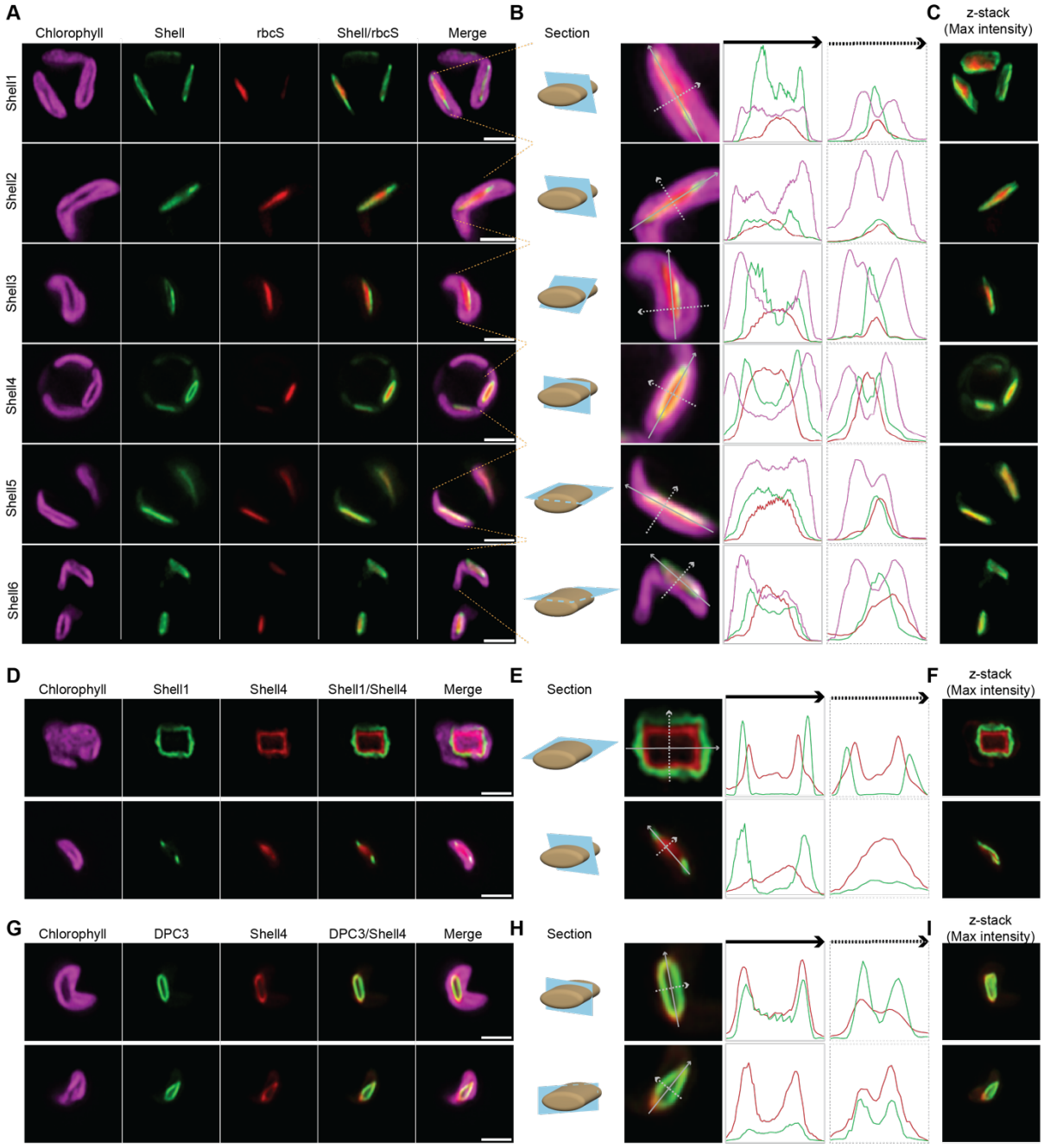


Figure 5. Shell proteins encapsulate the pyrenoid matrix, and Shell1 and Shell4 occupy distinct regions of the pyrenoid surface

A. Co-localization of Shell proteins with Rubisco.
B. Fluorescence intensity cross-sections of chlorophyll, Shell and rbcS across the pyrenoid.
C. Z-stack max intensity projections of C.
D. Co-localization of Shell1 and Shell4 in two cells with different pyrenoid orientations.
E. Fluorescence intensity cross-sections of Shell1 and Shell4 across the pyrenoid.
F. Z-stack max intensity projections of E.
G. Co-localization of DPC3 and Shell4 in two cells with different pyrenoid orientations.
H. Fluorescence intensity cross-sections of DPC3 and Shell4 across the pyrenoid.
I. Z-stack max intensity projections of H.
Scale bars: 2μm. For B, E and H the level of Zoom changes between images.

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

To understand the dynamics of pyrenoid components, we leveraged our tagged lines. If the shell proteins form a lattice around the Rubisco matrix, similar to carboxysome shell proteins in assembled carboxsomes,[47,48] then we would expect them to form a static layer. We investigated the mobility of Shell1 and Shell4 by fluorescence recovery after photobleaching (FRAP). Both Shell1 and Shell4 showed minimal rearrangement after photobleaching (Fig. 6A, B). Although this does not provide direct evidence that they form a CO_2 diffusion barrier, the static nature of the pyrenoid shell is consistent with our current understanding of both cyanobacterial carboxysome shell and green algal starch sheath CO_2 diffusion barriers.

In *C. reinhardtii*, the pyrenoid matrix has liquid-like properties. Rubisco, EPYC1 and Rubisco activase (RCA1) all show rapid rearrangement in FRAP experiments on the timescale of ~30 s after photobleaching.[48,49] In *T. pseudonana* Rubisco is densely packed in the pyrenoid but does not formed ordered arrays.[50] Surprisingly, FRAP experiments on both *T. pseudonana* rbcS and cbbX showed minimal rearrangement over minute timescales (Fig. S8). This observation aligns with recent data from the pennate diatom *P. tricornutum*, where *in vitro* phase-separated Rubisco demonstrates minimal rearrangement over similar timescales and the linker protein PYCO1 appears immobile *in vivo.*[51] This lack of dynamic mixing opens a considerable question about how cbbX can sufficiently access inhibited Rubisco to reactivate it. Collectively, this indicates that the diatom pyrenoid has different mesoscale properties to the *C. reinhardtii* pyrenoid, and that once Rubisco is assembled into the pyrenoid there is minimal rearrangement. This observation aligns with those of the cyanobacterial carboxysome, where Rubisco forms ordered arrays upon packing.[43,51]

Shell1 can self-assemble *in vitro* to form tubes and sheets

To form an encapsulating protein layer around the Rubisco matrix, we would expect that Shell proteins can undergo higher order self-assembly. To test this, we purified recombinantly-expressed Shell1 (Fig. S9A) from *E. coli* and performed cryogenic electron microscopy (cryo-EM). The formation of tubes and sheets (Fig. 6C and S9B) was observed in the micrographs. This observation is similar to the self-assembly of bacterial microcompartment hexamers, the main building block of carboxysome shells, which have been shown to self-assemble into both sheets[44,52] and tubes *in vitro*.[47] The ability to self-assemble into stable higher order structures likely explains the limited mobility of Shell1 observed *in vivo* and supports the hypothesis that Shell proteins could provide a continuous, encapsulating layer at the pyrenoid surface to function as a CO_2 barrier, control metabolite flux, and/or provide structural integrity to the pyrenoid.

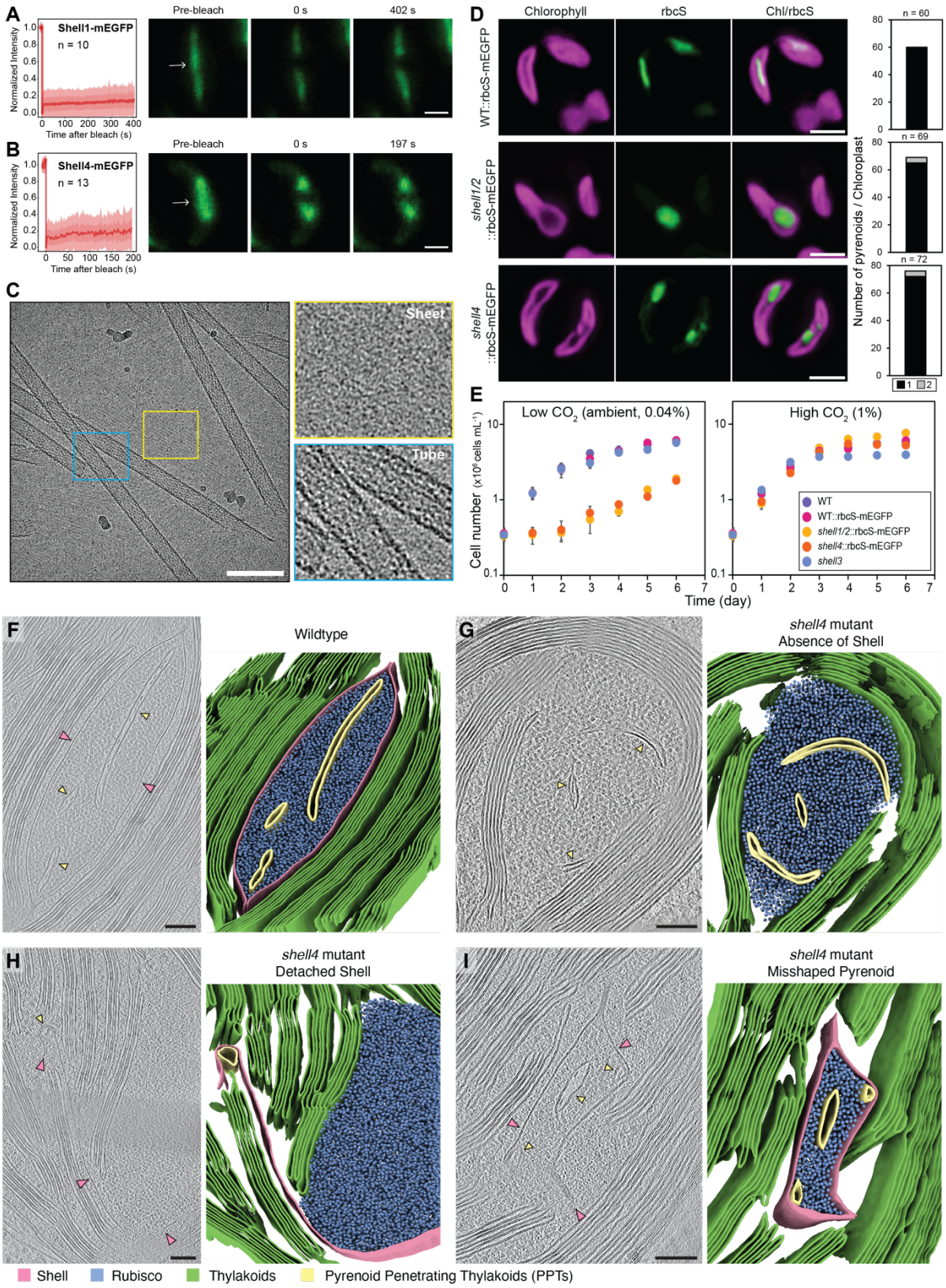


Figure 6: Shell proteins are static on the pyrenoid surface, can self-assemble into sheets and tubes, and are essential for CCM function and pyrenoid architecture

A-B. Fluorescence recovery after photobleaching (FRAP) experiments for Shell1 (A) and Shell4 (B). Arrows indicate photobleached regions. Scale bars: 1 µm. Plots show S.E.M. and S.D. of the mean.

C. Shell1 can form tubes and sheets as seen by cryo-EM. Scale bar: 100 nm.

D: Tagged rbcS in wildtype (WT::rbcS-mEGFP, top), a *shell1/2* mutant (*shell1/2*::rbcS-mEGFP, middle) and a *shell4* mutant line (*shell4*::rbcS-mEGFP, bottom). Scale bars: 2 µm. Bar charts indicate pyrenoid number per chloroplast in corresponding line.
E: Growth rate of WT, WT::rbcS-mEGFP, *shell1/2*::rbcS-mEGFP, *shell4*::rbcS-mEGFP and a *shell3* mutant.
F-I: Cryo-electron tomography of wildtype *T. pseudonana* (F) and the *shell4*::rbcS-mEGFP line (G-I). In each panel, the left image shows a 2D slice through the tomogram, and the right image shows the corresponding 3D segmentation of the Shell structure (pink), Rubisco matrix (blue), thylakoids (green), and PPTs (yellow). Scale bars: 100 nm.

Shell1/2 and Shell4 are essential for CCM function and pyrenoid architecture

The *in vitro* self-assembly of Shell1 into sheets and tubes and the immobile nature of Shell1 and Shell4 *in vivo* suggest they may be required for pyrenoid structural integrity. To test this, we used our MoClo Golden Gate system to simultaneously tag rbcS with mEGFP and disrupt either *Shell1/2* (*Shell1* and *Shell2)* or *Shell4* expression by CRISPR. Due to *Shell1* and *Shell2* having 93% DNA sequence similarity (Fig. S5C), single guide RNAs were designed to simultaneously target both genes (Fig. S10A). Edited lines were grown under high CO_2 conditions and selected by mEGFP fluorescence. Biallelic gene editing was confirmed by Sanger sequencing (Fig. S10A,B). Imaging of rbcS-mEGFP in the *shell1/2* mutant revealed that cells lacking both *Shell1* and *Shell2* failed to form a lenticular pyrenoid and instead typically possessed a single spherical pyrenoid per chloroplast, although the presence of multiple pyrenoids was also observed (Fig. 6D). This suggests that Shell1/2 are required for the lenticular shape of the pyrenoid, and in their absence, the pyrenoid assembles into a sphere. Similarly, the *shell4* mutant failed to form a lenticular pyrenoid, forming more oval shaped pyrenoids that were subtly different in shape from the more spherical pyrenoids in the *shell1/2* mutant (Fig. 6D and S11). This hints that Shell4 may have a distinct structural role to Shell1/2, in line with their different localizations at the pyrenoid periphery. The more spherical appearance of Rubisco in the absence of the shell is consistent with surface tension effects observed in liquid-liquid phase separated condensates, suggesting dynamic Rubisco condensation may also have a role in pyrenoid matrix assembly *in vivo* in *T. pseudonana,* as shown *in vitro* for *P. tricornutum*[53] and as seen for carboxysome assembly.[47]

Consistent with their disrupted pyrenoid morphology, the *shell1/2* and *shell4* mutants had significantly reduced growth at atmospheric CO_2, which was fully rescued by supplying elevated CO_2 (Fig. 6E), demonstrating that both Shell1/2 and Shell4 are required for a functional CCM. Interestingly, a *shell3* mutant did not show abnormal pyrenoid morphology by TEM (Fig. S10C and Fig. S12) and had no growth defect at ambient CO_2 (Fig. 6E). Relative to other Shell homologs, Shell3 has a low abundance (Fig. 3D), and thus could potentially play a minor role or be compensated by Shell1/2, which have 73% amino acid sequence similarity.

In a parallel study, it was shown that *shell1/2* knock-out mutants completely lack a protein shell around the pyrenoid.[47] To see if our *shell4* mutant has a similar architectural defect, we performed cryo-electron tomography (cryo-ET).[47] Wildtype *T. pseudonana* cells contain lenticular shaped pyrenoids, where the Rubisco matrix is encapsulated in a protein shell and bisected end-to-end by one or two PPTs that contain densities in their lumen (Fig. 6F).[47] In contrast, the *shell4* mutant has misshaped pyrenoids with a diverse range of morphologies (Figs. 6G-I; S13). These include: 1) spherical pyrenoids that lack a visible shell and have abnormal PPTs that do not bisect the matrix (similar to the *shell1/2* mutant[27,53]); 2) pyrenoids with a shell that is detached from the Rubisco matrix; 3) pyrenoids that are encapsulated in a shell but have an atypical distorted shape and abnormal PPTs. Whole-cell proteomics on the *shell4* mutant showed the almost complete absence of Shell4 but largely

429 unaltered levels of Shell1/2, suggesting that the Shell we observe in the *shell4* mutant is
430 composed of Shell1/2 proteins (Fig. S14 and Table S7). Shells in the *shell4* mutants show
431 apparent decreased affinity for the Rubisco matrix (detached shells) as well as apparent
432 increased interaction with PPT membranes (attachment to PPTs located both within the matrix
433 and displaced in the stroma). Interestingly, these mutant shells were also often observed to
434 self-interact, forming double or triple layered sheets (Fig. S13). We therefore hypothesize that
435 Shell4 promotes interaction with Rubisco, whereas Shell1/2 have high affinity for PPT
436 membranes and Shell1/2 proteins in adjacent sheets. Taken together, the complete absence
437 of a shell in the *shell1/2* mutant seen by cryo-ET[28,54] and the presence of a shell but severely
438 misshaped pyrenoids in the *shell4* mutant suggest that Shell1/2 and Shell4 have distinct roles:
439 Shell1/2 form the main structural component of the shell, and Shell4 is critical for pyrenoid
440 organization, ensuring that the Rubisco matrix, PPT, and shell are correctly assembled
441 together. This is further supported by both the whole-cell proteomics, where Shell1/2 are more
442 abundant than Shell4, as well as the pyrenoid interaction network, where Shell4 directly and
443 indirectly is connected to several pyrenoid matrix components but Shell1 has no such
444 connections.
445 The diverse pyrenoid structural defects in the *shell4* mutant makes it difficult to assign
446 the high CO_2 requiring phenotype to a specific pyrenoid structural defect. The disrupted PPTs
447 proposed to be involved in inorganic carbon delivery to Rubisco could reduce inorganic carbon
448 fluxes to the Rubisco matrix. If the shell has a role in minimizing CO_2 diffusion, its disruption
449 could make the pyrenoid leaky, thereby reducing the effective CO_2 concentration around
450 Rubisco. The shell could also be critical for efficient delivery and partitioning of Calvin cycle
451 intermediates between the pyrenoid and the surrounding stroma. Finally, pyrenoid shape
452 could be important to maximize membrane contact area and to minimize diffusion distances.
453 Collectively, these data support that the diatom Shell proteins 1, 2 and 4 are critical for
454 pyrenoid architecture and CCM function, and may act as a CO_2 diffusion barrier, although
455 further experimental proof is required for the latter.
456
457 **Perspective**
458 The development of a high-throughput tagging and APMS pipeline in a model diatom
459 has enabled us to generate a spatial interaction network of the diatom pyrenoid, providing
460 novel molecular insight into how diatoms help drive the global carbon cycle. We have identified
461 and confirmed via localization 13 new pyrenoid components, of which a large number have no
462 conserved functional domains. Six of these new components constitute a protein shell that
463 encapsulates the pyrenoid and is found across diverse species with red plastids derived from
464 secondary endosymbiotic events. Knock-out of the most abundant shell components, Shell1/2
465 and Shell4, resulted in large pyrenoid structural changes and poor growth at atmospheric
466 levels of CO_2. Cryo-ET on the *shell4* mutant indicates that Shell4 is critical for correct
467 organization of the PPT, shell and matrix, whilst our data along with data in a parallel study[25,26]
468 indicates that Shell1/2 is most likely the major structural component of the shell.
469 Four additional pyrenoid components, DPC1-4, have no clear function that can be
470 predicted from their sequence. DPC3 showed a shell-like localization pattern, co-localizing
471 with Shell4 and interacting with both Shell4 and rbcS, suggesting a potential role in mediating
472 shell-Rubisco matrix interactions. The molecular mechanism of Rubisco condensation to form
473 the pyrenoid matrix is currently unknown in *T. pseudonana*, with no homologs or functional
474 analogs of EPYC1 or PYCO1 (Rubisco likers in *C. reinhardtii* and *P. tricornutum*) identified in
475 our study. This suggests that pyrenoid matrix formation may be different in *T. pseudonana*.
476 The matrix localization and interaction partners of DPC2 and DPC4 suggest that they may

477 have a central role in pyrenoid matrix assembly/function. DPC2, DPC3 and DPC4 are prime
478 targets for future characterization.
479 Close to 50% of global carbon fixation is performed by biomolecular condensates of
480 Rubisco. This includes prokaryotic cyanobacterial carboxysomes and eukaryotic algal
481 pyrenoids. Nearly all of our data of pyrenoid structure and function comes from the green
482 plastid lineage alga, *C. reinhardtii*. However, pyrenoids both between plastid lineages and
483 within plastid lineages are proposed to have convergently evolved.[40] Insights from our data
484 suggest that diatom pyrenoids have similarities to both green plastid pyrenoids and prokaryotic
485 carboxysomes. Similarities between the *T. pseudonana* and *C. reinhardtii* pyrenoid include
486 dense Rubisco packaging around specialized thylakoid membranes (PPTs) for CO_2 delivery,
487 CO_2 delivery to the PPTs via bestrophin family channels,[54] and CO_2 release within the acidic
488 lumen driven by constrained localization of a carbonic anhydrase (θCA2; refs[55] and this study)
489 within the PPTs. However, the encapsulation by a protein shell layer composed of homologs,
490 some with different sub-shell localizations, is analogous to carboxysome shell proteins.[47]
491 Additionally, the static nature of Rubisco, cbbX, and shell proteins contrasts with the dynamic
492 properties of the *C. reinhardtii* pyrenoid and aligns more with carboxysomes. Another major
493 outstanding question is the connection of the PPT with the broader thylakoid network; in both
494 our TEM and cryo-ET, we are yet to visualize an unambiguous connection as seen in nearly
495 all pyrenoid containing algae[7,23,24] including *C. reinhardtii*[27] and *P. tricornutum.*[56] If there is no
496 connection, this would align the *T. pseudonana* pyrenoid even closer to carboxysomes, and
497 require a new functional model for the *T. pseudonana* CCM. It is tempting to postulate that the
498 extension and diversification of the Shell protein family in *T. pseudonana* vs. *P. tricornutum* (6
499 homologs vs. 2 homologs; Fig. 4B) may be critical for the complete encapsulation of the
500 Rubisco matrix and PPT, with Shell4 a strong candidate that could enable this.
501 In Figure 7, we propose two models for the *T. pseudonana* CCM, a more classical
502 pyrenoid-based CCM where a PPT-thylakoid connection is present (Fig. 7A) and a completely
503 Shell-encapsulated pyrenoid that would align closer to a carboxysome system (Fig. 7B). In
504 both models, SLC4 family proteins contribute to sodium dependent HCO_3^- transport at the
505 plasma membrane.[57] The mechanism of Ci transport across the four chloroplast membranes
506 is still unknown, although SLC4 transporters are also proposed to play a role here[28] along with
507 the carbonic anhydrase LCIP63[25,26] and the V-type ATPase.[58] In the classical model, Ci
508 delivery to the pyrenoid is potentially analogous to *C. reinhardtii*, relying on channeling of
509 HCO_3^- into the thylakoid lumen via BST1 and BST2.[59] HCO_3^- could then diffuse into the PPT,
510 where it is dehydrated to CO_2 via θCA2[60,61] and diffuses to Rubisco packaged within the
511 pyrenoid. In a more carboxysome-like system, HCO_3^- accumulated in the stroma would have
512 to diffuse across the shell and subsequently be dehydrated to CO_2. CO_2 release could
513 potentially still occur via channeling into the PPT by bestrophins, and its release accelerated
514 via θCA2. In the classical pyrenoid model, like in *C. reinhardtii*, the low luminal pH of the PPTs
515 established by the light reactions of photosynthesis would provide protons for HCO_3^-
516 dehydration and the energetic driving force of the CCM.[62] In a more carboxysome-like system,
517 protons for HCO_3^- dehydration could come from the Rubisco carboxylase reaction that has
518 been modelled to produce two protons for every carboxylation reaction.[28] In both models, CO_2
519 leakage out of the pyrenoid would be minimized by the proteinaceous shell that encapsulates
520 the pyrenoid. The shell is also critical for maintaining the correct architecture of the pyrenoid,
521 including the lenticular shape of the pyrenoid that would minimize diffusion distances of CO_2
522 from the PPTs to Rubisco.

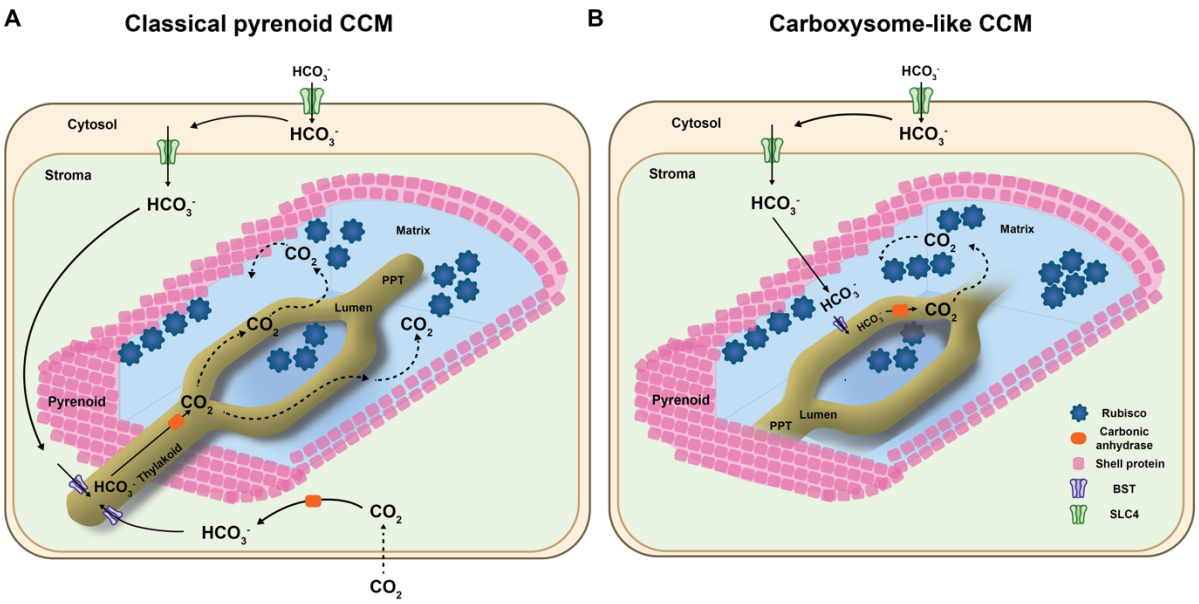


Figure 7: Structural and functional models of the *T. pseudonana* pyrenoid-based CO_2-concentrating mechanism.

A. Proposed model based on a classical pyrenoid CCM where there is a connection between the PPT and the broader thylakoid network.
B. Proposed model based on a carboxysome-like CCM, where the shell completely encapsulates the pyrenoid and the PPT is separated from the broader thylakoid network.

Models integrate data from this study with the available literature. For simplicity, the multiple membranes encapsulating the chloroplast, and components with unknown or unclear function have been omitted. See text for further discussion.

Other than Rubisco, there appears to be no sequence or structural similarities between carboxysomes and *T. pseudonana* pyrenoid proteins. This further supports the convergent evolution of pyrenoids and that a broad range of biophysical, structural and functional properties, some previously associated with carboxysomes, can be expected as more pyrenoids are characterized across diverse alga.

A core structural component of pyrenoids is a CO_2 leakage barrier, with the starch sheath in *C. reinhardtii* shown both experimentally and theoretically to be required for efficient CCM function.[17,18] As engineering of a pyrenoid into plants progresses, a major future challenge will be CO_2 diffusion barrier engineering.[17,22] This is thought to require multiple starch synthesis-related steps correctly localized to the pyrenoid periphery. The diatom Shell proteins could potentially provide an alternative biotechnology solution to this challenge.

Limitations of the Study

Well documented technical limitations of protein tagging could result in protein mislocalization and inaccurate protein-protein interactions. C-terminal tagging could result in masking interactions that could alter protein targeting and modulate native interactions with other proteins. The use of a constitutive promoter and expression of proteins in trans to native copies could also modify protein localization and interaction partners. Reported protein-protein interactions were not validated by a parallel method and may not mean a direct interaction but could be mediated through an additional component. Although DPC3 and Shell4 were shown to interact and colocalize, colocalization does not confirm a direct interaction.

555 Where possible we validated two independently tagged lines for localization, however
556 in some cases only a single stable line was generated. For *shell4* mutants, four independent
557 lines were sequence verified for biallelic editing. Both pyrenoid shape disruption by rbcs-
558 mEGFP imaging and growth defects at ambient CO_2 were seen across all edited lines, and a
559 single representative line was chosen for further detailed studies. For *shell1/2* mutants,
560 multiple lines were generated that resulted in pyrenoid shape disruption by rbcs-mEGFP
561 imaging, however only one was confirmed for biallelic editing of both *Shell1* and *Shell2*; this
562 line was prioritized for further studies.
563 Whilst we propose that the shell may act as a CO_2 barrier analogous to the
564 carboxysome shell, we provide no direct evidence for this. We also do not understand how
565 the Rubisco substrate, ribulose 1,5-bisphosphate (RuBP), and product, 3-phosphoglycerate
566 (PGA), cross the shell. Although we show that Shell1 and Shell4 do not colocalize, we do not
567 understand how the other Shell components are orientated relatively to each other, whether
568 Shell proteins other than Shell1 can homo-oligomerize, and whether Shell proteins can hetero-
569 oligomerize. All our images are snapshots of living cells. Proteins could change localization
570 and interactions depending on cell-cycle state and during pyrenoid division. Further, the
571 dynamics of Shell1, Shell4, Rubisco, and cbbX may change at different states of the cell cycle
572 to enable pyrenoid division and pyrenoid growth. Finally, how the pyrenoid matrix is assembled
573 is still an open question.
574
575 ## Materials and methods
576
577 ### Strains and Culturing
578
579 The background *Thalassiosira pseudonana* strain for all experiments was wildtype (WT)
580 CCAP1085/12 (Scottish Culture Collection of Algae and Protozoa, equivalent to CCMP1335).
581 WT cells were axenically maintained in artificial sea water (ASW) (32 g L^{-1}, Instant Ocean
582 SS15-10) supplemented with half-strength (F/2) Guillard F solution[64] at 20°C under continuous
583 illumination of ~50 µmol photons m^{-2} s^{-1}. All strains were grown at ambient CO_2 except *shell1/2*
584 deletion, *shell4* deletion and Shell1-mEGFP lines, which were maintained under 1%
585 CO_2. Exponentially growing cultures were used for growth assays. Once the pre-cultures
586 reached 2-4 x 10^6 cell mL^{-1}, cells were harvested by centrifugation at 3000 x*g* for 10 min. Then
587 the cells were resuspended at a concentration of 3 x 10^5 cell mL^{-1} in 10 mL of fresh media. To
588 ensure optimal gas exchange, cultures were grown in 6-well plates swirling twice a day and
589 incubated in growth chambers with water-saturated 0.04% CO_2 (ambient air, LC) or 1% CO_2
590 (HC) in air. Cell density was monitored daily by counting cells using a digital haemocytometer
591 (Countess II FL, Life Technology). All experiments were performed in biological triplicates.

592 ### Episome Assemblies using Golden Gate Cloning

593 Level 0, 1, and 2 (L0, L1, and L2) plasmids were assembled by Golden Gate (GG) cloning[65]
594 using the custom parts from the diatom MoClo framework.[35] Using genome version
595 ASM14940v2, target genes without stop codons were synthesized by Twist Bioscience (Table
596 S4). The regulatory elements, fucoxanthin chlorophyll a/c-binding protein (FCP) promoter and
597 terminator, and the fluorescent protein (FP) tags (mEGFP and mScarlet-I) harboring L0
598 plasmids were used to build L1 plasmids together with the gene-of-interest (GOI) for FP
599 tagging. Subsequently, L1 plasmids were assembled into L2 plasmids (episomes) in the

600 following order: episomal maintenance elements (position 1, P1),[65] nourseothricin (NAT)
601 resistance cassette (P2), FP tagged GOI (P3), and for dual-FP tagging (P4). For the CRISPR-
602 Cas9 knock-out episomes, Cas9 occupied P3 and two sgRNAs for each target both under the
603 U6 promoter occupied P4 and P5. P6 was used for simultaneous knock-out and rbcS-mEGFP
604 tagging. Each GG assembly was performed in 20 µL containing 40 fmols of each component
605 with 10x ligase buffer (NEB), 10 units T4 DNA ligase (NEB), and 10 units restriction enzyme
606 (BsaI for L1 assembly or BpiI for L0/L2 assembly, Thermo Fisher Scientific). The reaction was
607 incubated in a thermocycler by switching between 37°C and 16°C for 5 min intervals in a total
608 of 20-30 cycles, followed by 37°C for 5 min and terminated after incubating at 65°C for 20 min.
609 3 µL of the reaction were transformed into 50 µL chemically competent DH5α *E. coli* cells.

610 ***T. pseudonana* Transformation via Bacterial Conjugation**

611 Episomes were delivered to *T. pseudonana* via bacterial conjugation according to[66] with minor
612 modifications. Episome plasmids were transformed into *E. coli* (TransforMax EPI300)
613 harboring the pTA_Mob[67] mobility plasmid (gift from R. Lale) via electroporation (Bio-Rad).
614 Transformed cells were spread onto LB agar plates containing both gentamycin (10 µg mL^{-1})
615 and kanamycin (25 µg mL^{-1}) for selection overnight at 37°C. Colonies were inoculated for
616 subsequent conjugation. Cultures (150 mL) grown at 37°C to OD600 of 0.3-0.4 were harvested
617 by centrifugation (3,000 x*g*, 5 min) and resuspended in 800 µL of SOC media. Liquid grown *T.*
618 *pseudonana* WT culture was harvested by centrifugation (3,000 x*g*, 5 min) and resuspended
619 at a concentration of $2x10^8$ cells mL^{-1} in ½ASW-F/2. Equal volume (200 µL) of *E. coli* and *T.*
620 *pseudonana* WT cells were gently mixed by pipetting. Next the mixture of cells was plated on
621 ½ASW-F/2, 5% LB, 1% agar plates and incubated in the dark for 90 min at 30°C. The plates
622 were transferred to 20°C with continuous illumination (~50 µmol photons m^{-2} s^{-1}) and grown
623 overnight. Next day, 500 µL of ½ASW-F/2 medium was added to the plate for scraping and
624 resuspending the cells. Up to 200 µL of resuspended cells were spread onto 1% (w/v) ½ASW-
625 F/2 agar plates with 100 µg mL^{-1} nourseothricin for selection. Colonies appeared after 6-14
626 days.

627 **Fluorescence Screening by Flow Cytometry**

628 mEGFP and mScarlet-I expression was analyzed by flow cytometry using either CytoFLEX
629 LX355 or 375 (Beckman Coulter) analyzers. Forward scattered (FSC) and side scattered
630 photons by the 488 nm laser were used to distinguish diatoms from cell culture debris. FSC-
631 height versus FSC-area signal was used to separate single events from sample aggregates.
632 Chlorophyll autofluorescence excited by 561 nm laser and emitted photons detected with
633 675/25 filter was used to ensure all the diatom cells were fully intact. mEGFP fluorescence
634 excited by the 488 nm laser was detected by an avalanche photodiode detector with 525/40
635 bandpass filter. All the data analysis was done using CytExpert software (Beckman Coulter).

636 **Microscopy**

637 Fluorescence imaging was performed using a Zeiss LSM880 confocal microscope with a 63x
638 objective, 1.4 numerical aperture (NA) Plan-Apo oil-immersion lens (Carl Zeiss). All imaging
639 was done in Airyscan mode except for the rbcS-mEGFP line, which was imaged in confocal
640 mode. 20 µL of cell suspension were pipetted on 8 well µ-Slide chambered coverslips (ibidi)

641 overlaid with 180 µL of 1.5% F/2-low-melting point agarose (Invitrogen) for imaging. Excitation
642 lasers and emission filters were as follows: mEGFP excitation 488 nm, emission 481-541 nm;
643 mScarlet-I excitation 561 nm, emission 561-633 nm; and chlorophyll excitation 633 nm,
644 emission 642-712 nm. All the microscopic images were processed using Fiji.[68]

645 **Co-immunoprecipitation and affinity purification**

646 For rbcL coIP, 50 mL of WT *T. pseudonana* cells grown in log phase (2-3 x 10^6 cells mL^{-1})
647 were harvested by centrifugation (3,000 x*g*, 10 min). The pellets were resuspended in CoIP
648 buffer (20 mM Tris-HCl pH 8.0, 50 mM NaCl, 0.1 mM EDTA, 12.5% glycerol) containing 5 mM
649 DTT and protease inhibitor cocktail tablets (PIs, cOmplete EDTA-free, Roche). Cells were
650 lysed by sonication for 3 min (ON 3 sec, OFF 12 sec). The lysates were centrifuged for 20 min
651 (20,817 x*g*, 4°C) to separate the supernatant (soluble lysate) from the pellet. 200 µL of Protein
652 A (Dynabeads Protein A, Invitrogen) beads were washed twice in coIP buffer containing PIs.
653 32 µg of anti-rbcL antibody in 500 µL coIP buffer containing PIs was added to the washed
654 beads and incubated at 4°C for 2 hours. After incubation beads were washed twice in coIP
655 buffer containing PIs. For blocking, 500 µL of BSA (2 mg mL^{-1}) was added and incubated at
656 4°C for 1 hour. After incubation beads were washed twice in coIP buffer containing PIs.
657 Subsequently, the soluble lysates were added to protein A beads primed with antibody and
658 incubated at 4°C for 3 hours. After incubation, beads were washed three times with coIP buffer
659 containing PIs and 0.1% digitonin (SigmaAldrich). For elution, 200 µL of 1x SDS loading dye
660 was added and boiled at 95°C for 5 min. The supernatant was collected without any beads
661 and ran on an SDS-PAGE gel for ~1.5 cm. Gels were sliced for further in-gel digestion for LC-
662 MS/MS (see below).

663 For mEGFP tagged lines AP, 50 mL of GFP tagged *T. pseudonana* lines grown in exponential
664 phase (2-3 x 10^6 cells mL^{-1}) were harvested by centrifugation (3,000 x*g*, 10 min). The pellets
665 were resuspended in an immunoprecipitation (IP) buffer (200 mM D-sorbitol, 50 mM HEPES,
666 50 mM KOAc, 2 mM $Mg(OAc)_2$, 1 mM $CaCl_2$) containing protease inhibitor cocktail tablets
667 (cOmplete EDTA-free, Roche), 2% digitonin (SigmaAldrich), 1mM PMSF, 0.5 mM NaF and
668 0.15 mM Na_3VO_4. Cells were lysed by sonication for 30 sec (On 3 sec, Off 15 sec) twice. The
669 lysates were centrifuged for 20 min (20,817 x*g*, 4°C) and the supernatant was incubated with
670 mEGFP-Trap Agarose beads (ChromoTek) for 1 hour according to the manufacturer's
671 instructions. Subsequently, beads were washed twice with an IP buffer containing 0.1%
672 digitonin and a final wash without digitonin. All steps were performed at 4°C.

673 **Mass Spectrometry**

674 For rbcL coIPMS, samples were in-gel digested with 0.2 µg Sequencing-grade, modified
675 porcine trypsin (Promega), following reduction with 1.5 mg ml^{-1} dithioerythritol and alkylation
676 with 9.5 mg mL^{-1} iodoacetamide. Digests were incubated overnight at 37ºC. Peptides were
677 extracted by washing three times with aqueous 50% (v:v) acetonitrile containing 0.1% (v:v)
678 trifluoroacetic acid, before drying in a vacuum concentrator and reconstituting in aqueous 0.1%
679 (v:v) trifluoroacetic acid. Peptides were loaded onto an mClass nanoflow UPLC system
680 (Waters) equipped with a nanoEaze M/Z Symmetry 100 Å C_{18}, 5 µm trap column (180 µm x
681 20 mm, Waters) and a PepMap, 2 µm, 100 Å, C 18 EasyNano nanocapillary column (75 mm
682 x 500 mm, Thermo). The trap wash solvent was aqueous 0.05% (v:v) trifluoroacetic acid and

683 the trapping flow rate was 15 µL min^{-1}. The trap was washed for 5 min before switching flow
684 to the capillary column. Separation used gradient elution of two solvents: solvent A, aqueous
685 0.1% (v:v) formic acid; solvent B, acetonitrile containing 0.1% (v:v) formic acid. The flow rate
686 for the capillary column was 330 nL min^{-1} and the column temperature was 40°C. The linear
687 multi-step gradient profile was: 3-10% B over 5 mins, 10-35% B over 85 mins, 35-99% B over
688 10 mins and then proceeded to wash with 99% solvent B for 5 min. The column was returned
689 to initial conditions and re-equilibrated for 15 min before subsequent injections. The nanoLC
690 system was interfaced with an Orbitrap Fusion Tribrid mass spectrometer (Thermo) with an
691 EasyNano ionisation source (Thermo). Positive ESI-MS and MS2 spectra were acquired using
692 Xcalibur software (version 4.0, Thermo). Instrument source settings were: ion spray voltage,
693 1900-2100 V; sweep gas, 0 Arb; ion transfer tube temperature; 275°C. MS 1 spectra were
694 acquired in the Orbitrap with: 120,000 resolution, scan range: *m/z* 375-1,500; AGC target, $4e^5$;
695 max fill time, 100 ms. Data dependent acquisition was performed in top speed mode using a
696 1 s cycle, selecting the most intense precursors with charge states >1. Easy-IC was used for
697 internal calibration. Dynamic exclusion was performed for 50 s post precursor selection and a
698 minimum threshold for fragmentation was set at $5e^3$. MS2 spectra were acquired in the linear
699 ion trap with: scan rate, turbo; quadrupole isolation, 1.6 *m/z*; activation type, HCD; activation
700 energy: 32%; AGC target, $5e^3$; first mass, 110 *m/z*; max fill time, 100 ms. Acquisitions were
701 arranged by Xcalibur to inject ions for all available parallelizable time. Tandem mass spectra
702 peak lists were extracted from Thermo .raw files to .mgf format using MSConvert
703 (ProteoWizard 3.0). Mascot Daemon (version 2.6.0, Matrix Science) was used to submit
704 searches to a locally-running copy of the Mascot program (Matrix Science Ltd., version 2.7.0).
705 Peak lists were searched against the *Thalassiosira pseudonana* subsets of UniProt and NCBI
706 with common proteomic contaminants appended. Search criteria specified: Enzyme, trypsin;
707 Max missed cleavages, 2; Fixed modifications, Carbamidomethyl (C); Variable modifications,
708 Oxidation (M); Peptide tolerance, 3 ppm; MS/MS tolerance, 0.5 Da; Instrument, ESI-TRAP.
709 Peptide identifications were collated and filtered using Scaffold (5.2.0, Proteome Software
710 Inc). Peptide identifications were accepted if they could be established at greater than 51.0%
711 probability to achieve an FDR less than 1.0% by the Percolator posterior error probability
712 calculation. Protein identifications were accepted if they could be established at greater than
713 5.0% probability to achieve an FDR less than 1.0% and contained at least 2 identified peptides.

714 For mEGFP tagged lines APMS, samples were on-bead digested using Chromotek's
715 recommended procedure for NanoTraps: protein was digested overnight at 37°C with 25 µL
716 50 mM Tris-HCl pH 7.5, 2 M urea, 1mM DTT, 5 µg ml^{-1} Sequencing Grade Modified Trypsin
717 (Promega). Peptides were eluted with 50 mM Tris-HCl pH 7.5, 2 M urea, 5 mM iodoacetamide
718 before loading onto EvoTip Pure tips for desalting and as a disposable trap column for
719 nanoUPLC using an EvoSep One system. A pre-set EvoSep 60 SPD gradient was used with
720 a 8 cm EvoSep C_{18} Performance column (8 cm x 150 µm x 1.5 µm). The nanoUPLC system
721 was interfaced to a timsTOF HT mass spectrometer (Bruker) with a CaptiveSpray ionisation
722 source (Source). Positive PASEF-DDA, ESI-MS and MS^2 spectra were acquired using
723 Compass HyStar software (version 6.2, Thermo). Instrument source settings were: capillary
724 voltage, 1,500 V; dry gas, 3 l min^{-1}; dry temperature; 180°C. Spectra were acquired between
725 *m/z* 100-1,700. The following TIMS settings were applied as: 1/K0 0.6-1.60 V.s cm^{-2}; Ramp
726 time, 100 ms; Ramp rate 9.42 Hz. Data dependent acquisition was performed with 10 PASEF
727 ramps and a total cycle time of 1.17 s. An intensity threshold of 2,500 and a target intensity of

728 20,000 were set with active exclusion applied for 0.4 min post precursor selection. Collision
729 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,
730 database searching, significance thresholding and peak area integration was performed using
731 FragPipe (version 19.1). Data were searched against UniProt reference proteome
732 UP000001449, appended with common contaminants and concatenated with reversed
733 sequences for false discovery calculation. Search criteria specified: Enzyme, trypsin; Max
734 missed cleavages, 2; Fixed modifications, Carbamidomethyl (C); Variable modifications,
735 Oxidation (M), Acetylation (Protein N-term); Peptide tolerance, 10 ppm; MS/MS tolerance, 10
736 ppm; Instrument, IM-MS. Peptide identifications were filtered using Percolator and
737 ProteinProphet to 1% PSM FDR, protein probabilities >99%, best peptide probability >99%
738 and a minimum of two unique peptides. Peak area quantification was extracted using IonQuant
739 with match between run applied. Feature detection tolerances were set to: MS1 mass <10
740 ppm; RT < 0.4 min; and IM (1/k0) <0.05.

741 For whole-cell MS samples were ran in triplicate. 50 mL of WT cells were grown in ambient
742 CO_2 conditions and harvested at the log phase (2-3 x 10^6 cells mL^{-1}) by centrifugation (3,000
743 x*g*, 10 min). Shell4 mutant lines were grown in 1% CO_2. Cells were lysed by sonication (3 min:
744 ON 3 s, OFF 12 s) in CoIP buffer containing 5 mM DTT and PIs followed by centrifugation for
745 20 min (20,817 x*g*, 4°C). Supernatant was run on an SDS-PAGE gel for ~1.5 cm. In-gel
746 digestion was performed with the addition of 0.2 mg sequencing-grade, modified porcine
747 trypsin (Promega), following reduction with dithioerythritol and alkylation with iodoacetamide.
748 Digests were incubated overnight at 37°C. Peptides were extracted by washing three times
749 with aqueous 50% (v:v) acetonitrile containing 0.1% (v:v) trifluoroacetic acid, before drying in
750 a vacuum concentrator and reconstituting in aqueous 0.1% (v:v) trifluoroacetic acid.

751 S-Trap™ micro spin column (PROFITI) digestions were performed using the manufacture's
752 protocol. Briefly, protein in 5% SDS (w:w) lysis buffer was reduced with tris(2-
753 carboxyethyl)phosphine and alkylated with methyl methanethiosulfonate before acidification
754 with phosphoric acid and dilution into 100 mM triethylammonium bicarbonate (TEAB) in 90%
755 (v:v) methanol. Protein was washed on s-Trap with the same buffer five times before digestion
756 for 2 h at 47°C with 2 mg Promega Trypsin/Lys-C mixed protease in 50 mM TEAB.

757 Peptides were loaded onto EvoTip Pure tips for nanoUPLC using an EvoSep One system. A
758 pre-set 30SPD gradient was used with a 15 cm EvoSep C_{18} Performance column (15 cm x
759 150 mm x 1.5 mm).

760 The nanoUPLC system was interfaced to a timsTOF HT mass spectrometer (Bruker) with a
761 CaptiveSpray ionisation source. Positive PASEF-DIA, nanoESI-MS and MS^2 spectra were
762 acquired using Compass HyStar software (version 6.2, Bruker). Instrument source settings
763 were: capillary voltage, 1,500 V; dry gas, 3 l/min; dry temperature; 180°C. Spectra were
764 acquired between m/z 100-1,700. DIA windows were set to 25 Th width between m/z 400-
765 1201 and a TIMS range of 1/K0 0.6-1.60 V.s/cm2. Collision energy was interpolated between
766 20 eV at 0.65 V.s/cm2 to 59 eV at 1.6 V.s/cm2.

767 LC-MS data, in Bruker .d format, was processed using DIA-NN (1.8.2.27) software and
768 searched against an *in silico* predicted spectral library, derived from the *Thalassiosira*
769 *pseudonana* subset of UniProt (11934 protein sequences). Search criteria were set to maintain

770 a false discovery rate (FDR) of 1% with heuristic protein inference. High-precision quant-UMS
771 was used for extraction of quantitative values within DIA-NN. Peptide-centric output in .tsv
772 format, was pivoted to protein-centric summaries using KNIME 5.1.2 and data filtered to
773 require protein q-values < 0.01 and a minimum of two peptides per accepted protein. Raw
774 proteomic mass spectrometry data and search results files are referenced in
775 ProteomeXchange (PXD052522) and can be accessed via MassIVE (MSV000094846)
776 (doi:10.25345/C53J39C2Q).

777 **Protein-protein interaction network analysis**

778 Protein abundances quantified using MS2 spectral count measurements of fragment ions from
779 all sample triplicates were run through a CompPASS package in R Studio
780 (https://github.com/dnusinow/cRomppass/blob/master/R/cRomppass.R) and a control IP
781 inclusive variation of SAINT analysis in Ubuntu using standard parameters.[69] The WD and
782 AvgP scores respectively generated were used as measures of interaction strength between
783 bait and prey proteins. Only interactions which fell in both the top 2.2% WD score and 1%
784 AvgP score were filtered as high confidence interactors. Prior to analysis bait spectral count
785 data was set to zero to minimize data skewing due to the typically high spectral counts and
786 the inability to distinguish between mEGFP tagged bait and untagged native protein.

787 ***In vivo* fluorescence recovery after photobleaching**

788 FRAP experiments were performed using a Zeiss LSM980 confocal microscope with a 63x
789 objective 1.4 numerical aperture (NA) Plan-Apo oil-immersion lens (Carl Zeiss). Samples were
790 prepared as in confocal microscopy and overlaid with 200 µL of ibidi anti-evaporation oil. 20
791 pre-bleach images were taken prior to bleaching (60% 488 nm intensity, 1 cycle). All the
792 images were processed by Fiji.[70] The Image Stabilizer plugin (4 pyramid levels, 0.99 template
793 update coefficient) output of the brightfield images were used to stabilize the fluorescence
794 images. The mean gray values were measured for the bleached, unbleached and background
795 ROIs. Background values were subtracted from bleached and unbleached values before
796 photobleach normalization using the unbleached reference was completed. The average pre-
797 bleach intensity was used for full-scale normalization.

798 **Data Visualization**

799 Network visualization of the interaction network was done using Cytoscape (version 3.10.0)
800 (https://cytoscape.org/). The phylogenetic tree was done in Geneious Prime (2023.2.1). Adobe
801 Illustrator (2023), Excel (Microsoft) and Prism 10 (GraphPad) were used to generate the
802 figures.

803 **AlphaFold Structure Prediction**

804 AlphaFold structure predictions for Shell 1-6 and DPC3 without signal peptide and chloroplast
805 transit peptides were done using ColabFold v1.5.2-patch: AlphaFold2 using MMseqs2
806 (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb).

807

808 **Protein model for Rubisco**

809 *Thalassiosira antarctica* Rubisco model (PDB: 5MZ2) was used to predict the peptide regions
810 utilized for raising the anti-rbcL antibody using ChimeraX.[71]

811 **Transmission Electron Microscopy**

812 50 mL of cell culture grown in log phase were harvested by centrifugation (3,000 x*g*, 5 min).
813 Cells were fixed in 0.1 M cacodylate buffer (SigmaAldrich), pH 7.4, containing 2.5%
814 glutaraldehyde (SigmaAldrich), 2% formaldehyde (Polysciences) for 1 hour at room
815 temperature. They were then rinsed twice in phosphate-buffered saline (PBS, pH 7.4) for 15
816 min by removing the liquid at the top of the tube and leaving the cells undisturbed at the bottom
817 of the tube. The cells were secondary fixed with Osmium Tetroxide (1%, in buffer pH 7.2, 0.1M)
818 for 1 hour. After rinsing twice with buffer, again removing the liquid above the cells whilst
819 leaving the cells undisturbed, the cells were dehydrated through an ethanol series of 30, 50,
820 70, 90 and 100%, with each rinse for 15 min. To ensure thorough dehydration the 100%
821 ethanol step was repeated 3 times. The ethanol was then replaced by Agar low viscosity resin
822 by placing it in increasing concentrations of resin (30% resin: 70 % ethanol. 50:50, 70:30 each
823 change was left for at least 12 hours) until it was in 100% resin. To ensure complete resin
824 infiltration, the 100% resin step was repeated 3 times leaving it overnight between changes.
825 The Eppendorf tubes were then placed in an embedding oven and the resin polymerized at
826 60°C overnight. The resulting blocks were sectioned with a Leica Ultracut E ultra microtome
827 using a diatome diamond knife. The sections were then stained using a saturated solution of
828 Uranyl acetate (for 15 min) and Reynold's Lead citrate (15 min). The sections were imaged
829 using a JEOL 1400 TEM.

830 **Blast search**

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

838 **Phylogeny**

839 The phylogenetic tree was build using Geneious Prime (2023.2.1). 136 proteins were used for
840 MAFFT alignment using default settings (Algorithm = 'Auto', Scoring matrix = 'BLOSUM62',
841 Gap open penalty = '1.53', Offset value = '0.123'). Phylogenetic tree analysis was performed
842 using RAxML with the following settings: Protein Model = 'GAMMA BLOSUM62', Algorithm =
843 'Rapid Bootstrapping', Number of starting trees or bootstrap replicates = '1000', Parsimony
844 random seed = '1'. A consensus tree was generated by selecting Consensus Tree Builder,
845 Create Consensus Tree, Support Threshold % = 0, Topology Threshold % = 0, Burn-in % = 0,
846 Save tree(s) separately. *C. reinhardtii* BST1-3 sequences (Cre16.g662600, BST1;
847 Cre16.g663400, BST2; Cre16.g663450, BST3) were added to root the tree.

848 **Shell1 purification**

849 A modified pOPT expression plasmid[72] containing *shell1* with N-terminal His6-MBP tag and
850 C-terminal GST tag was transformed into *E. coli* expression strain C41 DE3. Cells were grown
851 in 2 L of liquid LB medium with 100 µg mL^{-1} carbenicillin until OD_{600} reached 0.6 when the
852 culture was induced with 1 mM IPTG for 6 h. Harvested cells were resuspended in a lysis
853 buffer. The lysis buffer contained 50 mM Tris-HCl, pH 8.0, 1x cOmplete EDTA-free protease
854 inhibitor with 500 mM NaCl and 2 mM PMSF. The cells were lysed by cell disruption (30 kPSI)
855 and the lysate cleared by 10 min of centrifugation at 7,000 x*g* and 30 min at 20,000 x*g* before
856 being filtered with a 0.2 µm filter. The supernatant was applied onto a 5 ml HisTrap Fast Flow
857 nickel affinity column (Cytiva) equilibrated with buffer A (50 mM Tris-HCl pH 8.0, 500 mM NaCl)
858 followed by 4CV wash with lysis buffer. The bound protein was then eluted using a linear
859 gradient of imidazole up to 500 mM concentration. The elution was then treated with TEV
860 protease and 3C protease while dialysed overnight at 4°C into 50 mM Tris-HCl, pH 8.0, 500
861 mM NaCl buffer. The sample was then applied to both a HisTrap Fast Flow nickel column and
862 a GSTrap column to remove the tags. The Shell protein remained in the flowthrough. The
863 protein was then concentrated using an Amicon Ultra concentrator 10,000 MW, applied onto
864 a Superdex 200 pg 16/600 Column equilibrated with buffer A to further ensure the purity of the
865 sample, and concentrated again with an Amicon Ultra Concentrator if needed.
866

867 **Shell1 cryo-electron microscopy grid preparation and data collection**

868 QUANTIFOIL® R 1.2/1.3 grids were glow discharged with PELCO easiGlow system using a
869 pressure of 0.26 mBar, for 60 seconds at a 20 mA current. Shell1 protein sample in 50 mM
870 Tris-HCl (pH 8.0), 50 mM NaCl at 2.3 mg mL^{-1} concentration was applied onto the grid,
871 blotted for 8 s using force 10 with a FEI Mark IV Vitrobot (ThermoFisher Scientific) in a
872 chamber at 4°C and 95% relative humidity. Grids were then rapidly plunge frozen in liquid
873 ethane.

874 Data was collected on a 200 kV Glacios cryo-electron microscope with a Falcon IV direct
875 electron detector at the University of York. Automated data collection was performed using
876 EPU in AFIS mode (ThermoFisher Scientific). Micrographs were collected at a nominal
877 magnification of 120,000x and total electron dose of 50 $e^{-}/Å^{2}$ with pixel size of 1.2 $Å^{2}$. Total
878 exposure was 8 s. The 100 µm objective aperture was inserted and the C2 aperture was 50
879 µm. Defocus values were: -2.0, -1.6, -1.4, -1.2, -1.0.

880 **Cellular cryo-electron tomography**

881 *T. pseudonana* cultures were grown as described above, with ΔtpPyShell4 being
882 supplemented with 5 mM NaCO3 in the F/2 medium. Cells were sedimented at 800 x*g* for 5
883 min prior to vitrification, 4 µL of cell suspension was applied to 200-mesh R2/1 carbon-film
884 covered copper grids (Quantifoil Micro Tools), and grids were blotted and plunge-frozen using
885 a Vitrobot Mark IV (Thermo Fisher Scientific). EM grids were clipped into Autogrid supports
886 (Thermo Fischer Scientific) and loaded into Aquilos 2 FIB-SEM instruments (Thermo Fischer
887 Scientific), where they were thinned with a Gallium ion beam as previously described.[73] The
888 resulting EM grids with thin lamellae were transferred to a transmission electron microscope
889 for tomographic imaging. Datasets were acquired on three different microscopes: a 300 kV
890 Titan Krios G4i microscope (Thermo Fisher Scientific), equipped with a Selectris X post-

column energy filter (Thermo Fisher Scientific) and a Falcon 4 direct electron detector (Thermo Fisher Scientific) ("M1"); a 300 kV Titan Krios G3i microscope (Thermo Fisher Scientific) equipped with a BioQuantum post-column energy filter (Gatan) and a K3 direct electron detector (Gatan) ("M2"); and Titan G4 equipped with a monochromator, a Selectris X energy filter, and a Falcon 4 camera ("M3"). Tilt-series were obtained using SerialEM 3.8 software[74] or Tomo5 (Thermo Fisher Scientific). In all cases, tilt-series were acquired using a dose-symmetric tilt scheme,[75] with 2° steps totaling 60 tilts per series. Each image was recorded in counting mode with ten frames per second. For M1 and M3, data was acquired in EER mode. The target defocus of individual tilt-series ranged from −2 to −5 µm. Total dose per tilt series was approximately 120 e-/Å2. Image pixel sizes for microscopes M1, M2, and M3 were 2.42 Å, 2.143/2.685 Å, and 2.38 Å, respectively.

Tomograms were subsequently reconstructed. First, TOMOMAN Matlab scripts (version 0.6.9 https://github.com/williamnwan/TOMOMAN/; https://doi.org/10.5281/zenodo.4110737)[76] were used to preprocess the tomographic tilt series data. Raw frames were then aligned using MotionCor2 (version 1.5.0)[77] before dose-weighting the tilt-series series,[66] followed by manual removal of bad tilts. The resulting tilt-series (binned 4 times) were aligned in IMOD (version 4.11)[74] using patch tracking and were reconstructed by weighted back projection. Cryo-CARE (version 0.2.1)[75] was applied on reconstructed tomogram pairs from odd and even raw frames to enhance contrast and remove noise. Snapshots of denoised tomograms were captured using the IMOD 3dmod viewer. Denoised tomograms were used as input for automatic segmentation using MemBrain.[76] The resulting segmentations were manually curated in Amira (version 2021.2; Thermo Fisher Scientific). Rubisco enzymes were template matched using PyTOM (version 0.0.6)[77] on bin4 tomograms. The resulting segmentations and particle coordinates were visualized in ChimeraX.[66]

Data availability

All mass spectrometry and proteomic identification data are referenced in ProteomeXchange (PXD052522) and can be accessed via MassIVE (MSV000094846).

Author contributions: O.N. and L.C.M.M. designed and supervised the study. O.N. carried out the experiments unless otherwise mentioned. M.D. and B.E. performed cryo-ET of the *shell4* mutant. S.M. purified the Shell1 protein and carried out cryo-EM with support from J.N.B. C.M. and O.N. performed the phylogenetic analysis of Shell proteins. A.D., M.D. and J.B. provided bioinformatics and data analysis support. A.D. oversaw the mass spectrometry and peptide mapping. O.N., C.M., and L.C.M.M. analyzed and interpreted the data. O.N. created the figures. O.N. and L.C.M.M. wrote the manuscript with input from all authors.

Acknowledgements: We thank Oliver Mueller-Cajar for kindly sharing the anti-rbcL antibody, the Technology Facility in the York Department of Biology for the access/support for microscopy and flow cytometry, and members of the Mackinder Lab for fruitful discussions. Thanks to Glenn Harper for optimizing the TEM. Cryo-ET was performed at the University of Basel using instrumentation provided by the BioEM core facility and computation provided by the sciCORE high-performance computing center. 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). B.D.E acknowledges funding from ERC consolidator grant “cryOcean” (fulfilled by the Swiss State Secretariat for Education, Research and Innovation, M822.00045). 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. Preprint, 10.1007/s11120-023-01007-z 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.

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

12. Freeman Rosenzweig, E.S., Xu, B., Kuhn Cuellar, L., Martinez-Sanchez, A., Schaffer,

M., Strauss, M., Cartwright, H.N., Ronceray, P., Plitzko, J.M., Förster, F., et al. (2017). The Eukaryotic CO2-Concentrating Organelle Is Liquid-like and Exhibits Dynamic Reorganization. Cell *171*, 148-162.e19.

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

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

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

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

17. Fei, C., Wilson, A.T., Mangan, N.M., Wingreen, N.S., and Jonikas, M.C. (2022). Modelling the pyrenoid-based CO2-concentrating mechanism provides insights into its operating principles and a roadmap for its engineering into crops. Nat Plants *8*, 583–595.

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

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

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

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

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

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

24. Nawaly, H., Matsui, H., Tsuji, Y., Iwayama, K., Ohashi, H., Nakajima, K., and Matsuda, Y. (2023). Multiple plasma membrane SLC4s contribute to external HCO3- acquisition during CO2 starvation in the marine diatom Phaeodactylum tricornutum. J. Exp. Bot. *74*, 296–307.

25. Kikutani, S., Nakajima, K., Nagasato, C., Tsuji, Y., Miyatake, A., and Matsuda, Y.

(2016). Thylakoid luminal θ-carbonic anhydrase critical for growth and photosynthesis in the marine diatom Phaeodactylum tricornutum. Proc. Natl. Acad. Sci. U. S. A. *113*, 9828–9833.

26. Shimakawa, G., Okuyama, A., Harada, H., Nakagaito, S., Toyoshima, Y., Nagata, K., and Matsuda, Y. (2023). Pyrenoid-core CO2-evolving machinery is essential for diatom photosynthesis in elevated CO2. Plant Physiol. *193*, 2298–2305.

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

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

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

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

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

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

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

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

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

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

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

38. Barrett, J., Naduthodi, M.I.S., Mao, Y., Dégut, C., Musiał, S., Salter, A., Leake, M.C., Plevin, M.J., McCormick, A.J., Blaza, J.N., et al. (2024). A promiscuous mechanism to phase separate eukaryotic carbon fixation in the green lineage. bioRxiv. 10.1101/2024.04.09.588658.

39. Kroth, P.G., Chiovitti, A., Gruber, A., Martin-Jezequel, V., Mock, T., Parker, M.S.,

1059 Stanley, M.S., Kaplan, A., Caron, L., Weber, T., et al. (2008). A model for carbohydrate
1060 metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole
1061 genome analysis. PLoS One *3*, e1426.

1062 40. Rae, B.D., Long, B.M., Badger, M.R., and Price, G.D. (2013). Functions, compositions,
1063 and evolution of the two types of carboxysomes: polyhedral microcompartments that
1064 facilitate CO_2 fixation in cyanobacteria and some proteobacteria. Microbiol. Mol. Biol.
1065 Rev. *77*, 357–379.

1066 41. Kerfeld, C.A., and Melnicki, M.R. (2016). Assembly, function and evolution of
1067 cyanobacterial carboxysomes. Curr. Opin. Plant Biol. *31*, 66–75.

1068 42. Cameron, J.C., Wilson, S.C., Bernstein, S.L., and Kerfeld, C.A. (2013). Biogenesis of a
1069 bacterial organelle: the carboxysome assembly pathway. Cell *155*, 1131–1140.

1070 43. Long, B.M., Rae, B.D., Rolland, V., Förster, B., and Price, G.D. (2016). Cyanobacterial
1071 CO2-concentrating mechanism components: function and prospects for plant metabolic
1072 engineering. Curr. Opin. Plant Biol. *31*, 1–8.

1073 44. Turnšek, J.B., Oltrogge, L.M., and Savage, D.F. (2023). Conserved and repetitive motifs
1074 in an intrinsically disordered protein drive α-carboxysome assembly. bioRxiv,
1075 2023.07.08.548221. 10.1101/2023.07.08.548221.

1076 45. Kong, W.-W., Zhu, Y., Zhao, H.-R., Du, K., Zhou, R.-Q., Li, B., Yang, F., Hou, P.,
1077 Huang, X.-H., Chen, Y., et al. (2024). Cryo-electron tomography reveals the packaging
1078 pattern of RuBisCOs in Synechococcus β-carboxysome. Structure.
1079 10.1016/j.str.2024.05.007.

1080 46. Evans, S.L., Al-Hazeem, M.M.J., Mann, D., Smetacek, N., Beavil, A.J., Sun, Y., Chen,
1081 T., Dykes, G.F., Liu, L.-N., and Bergeron, J.R.C. (2023). Single-particle cryo-EM
1082 analysis of the shell architecture and internal organization of an intact α-carboxysome.
1083 Structure *31*, 677-688.e4.

1084 47. Shimakawa, G., Demulder, M., Flori, S., Kawamoto, A., Tsuji, Y., Nawaly, H., Tanaka,
1085 A., Tohda, R., Ota, T., Matsui, H., et al. (2023). Diatom pyrenoids are encased in a
1086 protein shell that enables efficient CO2 fixation. bioRxiv, 2023.10.25.564039.
1087 10.1101/2023.10.25.564039.

1088 48. Metskas, L.A., Ortega, D., Oltrogge, L.M., Blikstad, C., Lovejoy, D.R., Laughlin, T.G.,
1089 Savage, D.F., and Jensen, G.J. (2022). Rubisco forms a lattice inside alpha-
1090 carboxysomes. Nat. Commun. *13*, 4863.

1091 49. Ni, T., Sun, Y., Burn, W., Al-Hazeem, M.M.J., Zhu, Y., Yu, X., Liu, L.-N., and Zhang, P.
1092 (2022). Structure and assembly of cargo Rubisco in two native α-carboxysomes. Nat.
1093 Commun. *13*, 4299.

1094 50. Sutter, M., Faulkner, M., Aussignargues, C., Paasch, B.C., Barrett, S., Kerfeld, C.A.,
1095 and Liu, L.-N. (2015). Visualization of bacterial microcompartment facet assembly using
1096 high-speed atomic force microscopy. Nano Lett. *16*, 1590–1595.

1097 51. Noël, C.R., Cai, F., and Kerfeld, C.A. (2016). Purification and characterization of protein
1098 nanotubes assembled from a single bacterial microcompartment shell subunit. Adv.
1099 Mater. Interfaces *3*, 1500295.

1100 52. Wang, H., Yan, X., Aigner, H., Bracher, A., Nguyen, N.D., Hee, W.Y., Long, B.M., Price,
1101 G.D., Hartl, F.U., and Hayer-Hartl, M. (2019). Rubisco condensate formation by CcmM

1102 in β-carboxysome biogenesis. Nature. 10.1038/s41586-019-0880-5.

1103 53. Young, L.N., and Villa, E. (2023). Bringing Structure to Cell Biology with Cryo-Electron
1104 Tomography. Annu. Rev. Biophys. *52*, 573–595.

1105 54. Nigishi, M., Shimakawa, G., Yamagishi, K., Amano, R., Ito, S., Tsuji, Y., Nagasato, C.,
1106 and Matsuda, Y. (2024). Low-CO2-inducible bestrophins outside the pyrenoid sustain
1107 high photosynthetic efficacy in diatoms. Plant Physiol. 10.1093/plphys/kiae137.

1108 55. Engel, B.D., Schaffer, M., Kuhn Cuellar, L., Villa, E., Plitzko, J.M., and Baumeister, W.
1109 (2015). Native architecture of the *Chlamydomonas* chloroplast revealed by in situ cryo-
1110 electron tomography. Elife *4*, e04889.

1111 56. Jensen, E.L., Clement, R., Kosta, A., Maberly, S.C., and Gontero, B. (2019). A new
1112 widespread subclass of carbonic anhydrase in marine phytoplankton. ISME J. *13*,
1113 2094–2106.

1114 57. Yee, D.P., Samo, T.J., Abbriano, R.M., Shimasaki, B., Vernet, M., Mayali, X., Weber,
1115 P.K., Mitchell, B.G., Hildebrand, M., Decelle, J., et al. (2023). The V-type ATPase
1116 enhances photosynthesis in marine phytoplankton and further links phagocytosis to
1117 symbiogenesis. Curr. Biol. 10.1016/j.cub.2023.05.020.

1118 58. Burlacot, A., Dao, O., Auroy, P., Cuiné, S., Li-Beisson, Y., and Peltier, G. (2022).
1119 Alternative photosynthesis pathways drive the algal CO2-concentrating mechanism.
1120 Nature, 1–6.

1121 59. Long, B.M., Förster, B., Pulsford, S.B., Price, G.D., and Badger, M.R. (2021). Rubisco
1122 proton production can drive the elevation of CO2 within condensates and
1123 carboxysomes. Proc. Natl. Acad. Sci. U. S. A. *118*. 10.1073/pnas.2014406118.

1124 60. Guillard, R.R., and Ryther, J.H. (1962). Studies of marine planktonic diatoms. I.
1125 Cyclotella nana Hustedt, and Detonula confervacea (cleve) Gran. Can. J. Microbiol. *8*,
1126 229–239.

1127 61. Guillard, R.R.L. (1975). Culture of Phytoplankton for Feeding Marine Invertebrates. In
1128 Culture of Marine Invertebrate Animals: Proceedings — 1st Conference on Culture of
1129 Marine Invertebrate Animals Greenport, W. L. Smith and M. H. Chanley, eds. (Springer
1130 US), pp. 29–60.

1131 62. Weber, E., Engler, C., Gruetzner, R., Werner, S., and Marillonnet, S. (2011). A modular
1132 cloning system for standardized assembly of multigene constructs. PLoS One *6*,
1133 e16765.

1134 63. Karas, B.J., Diner, R.E., Lefebvre, S.C., McQuaid, J., Phillips, A.P.R., Noddings, C.M.,
1135 Brunson, J.K., Valas, R.E., Deerinck, T.J., Jablanovic, J., et al. (2015). Designer diatom
1136 episomes delivered by bacterial conjugation. Nat. Commun. *6*, 6925.

1137 64. Strand, T.A., Lale, R., Degnes, K.F., Lando, M., and Valla, S. (2014). A new and
1138 improved host-independent plasmid system for RK2-based conjugal transfer. PLoS One
1139 *9*, e90372.

1140 65. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T.,
1141 Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source
1142 platform for biological-image analysis. Nat. Methods *9*, 676–682.

1143 66. 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.

67. Tan, S. (2001). A modular polycistronic expression system for overexpressing protein complexes in Escherichia coli. Protein Expr. Purif. *21*, 224–234.

68. Schaffer, M., Mahamid, J., Engel, B.D., Laugks, T., Baumeister, W., and Plitzko, J.M. (2017). Optimized cryo-focused ion beam sample preparation aimed at in situ structural studies of membrane proteins. J. Struct. Biol. *197*, 73–82.

69. Mastronarde, D.N. (2005). Automated electron microscope tomography using robust prediction of specimen movements. J. Struct. Biol. *152*, 36–51.

70. Hagen, W.J.H., Wan, W., and Briggs, J.A.G. (2017). Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. J. Struct. Biol. *197*, 191–198.

71. Khavnekar, S., and Wan, W. (2024). TOMOMAN: a software package for large scale cryo-electron tomography data preprocessing, community data sharing, and collaborative computing. bioRxiv. 10.1101/2024.05.02.589639.

72. Zheng, S.Q., Palovcak, E., Armache, J.-P., Verba, K.A., Cheng, Y., and Agard, D.A. (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat. Methods *14*, 331–332.

73. Grant, T., and Grigorieff, N. (2015). Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. Elife *4*, e06980.

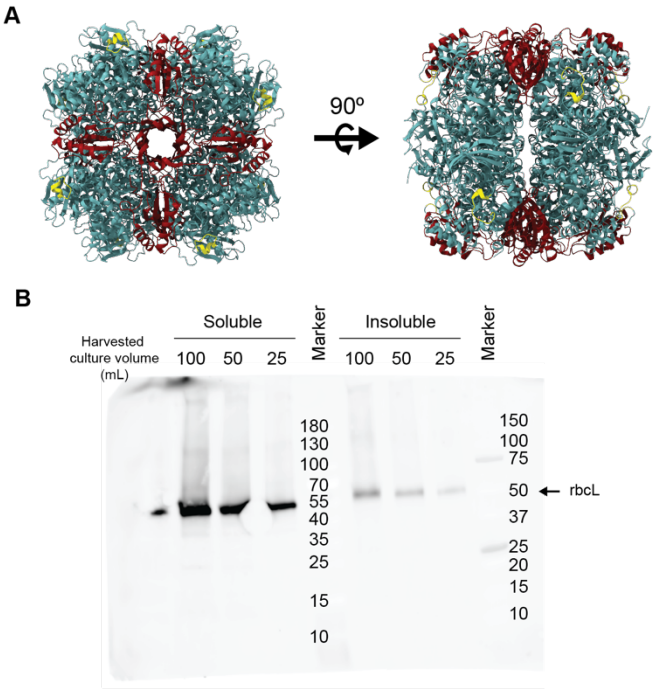
74. Mastronarde, D.N., and Held, S.R. (2017). Automated tilt series alignment and tomographic reconstruction in IMOD. J. Struct. Biol. *197*, 102–113.

75. Buchholz, T.-O., Krull, A., Shahidi, R., Pigino, G., Jékely, G., and Jug, F. (2019). Content-aware image restoration for electron microscopy. Methods Cell Biol. *152*, 277–289.

76. Lamm, L., Zufferey, S., Righetto, R.D., Wietrzynski, W., Yamauchi, K.A., Burt, A., Liu, Y., Zhang, H., Martinez-Sanchez, A., Ziegler, S., et al. (2024). MemBrain v2: an end-to-end tool for the analysis of membranes in cryo-electron tomography. bioRxiv, 2024.01.05.574336. 10.1101/2024.01.05.574336.

77. Hrabe, T., Chen, Y., Pfeffer, S., Cuellar, L.K., Mangold, A.-V., and Förster, F. (2012). PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. J. Struct. Biol. *178*, 177–188.

1177



1178

1179 **Figure S1. RbcL antibody**

1180 A. An antibody was raised against a surface exposed region (yellow) of the rbcL (green). RbcS
1181 is shown in red.

1182 B. Immunoblot using the rbcL antibody against the soluble and insoluble cell fractions of *T.*
1183 *pseudonana*.

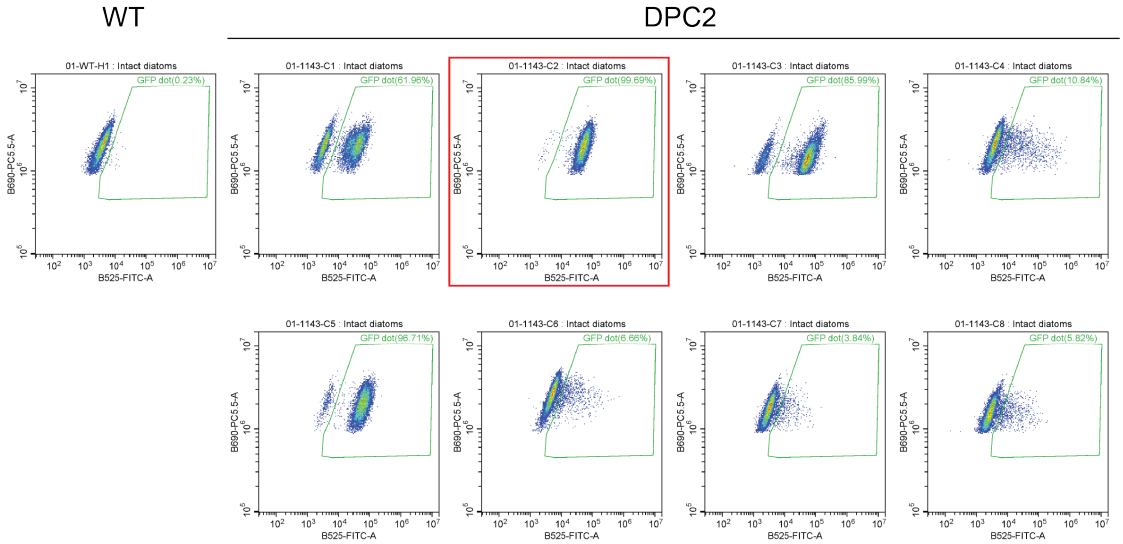


Figure S2. Flow cytometry screening to identify fluorophore fusion lines. Example screening data for Diatom Pyrenoid Component 2 (DPC2). Eight independent lines expressing DPC2-mEGFP from an episome were screened via flow cytometry. The green box outlines the gate used to quantify the percentage of cells that show mEGFP fluorescence with the percentage indicated above. The red box denotes the cell line used for subsequent imaging and APMS.

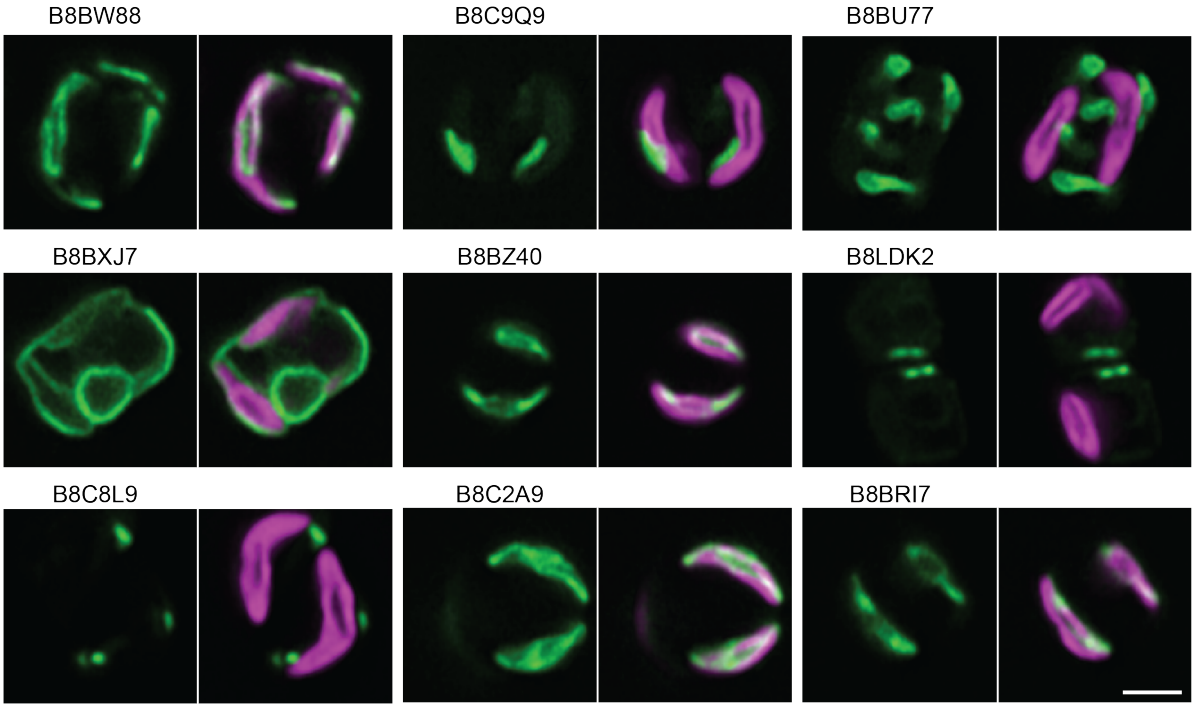


Figure S3. mEGFP fusion proteins that did not localize to the pyrenoid. Green: mEGFP fusion protein; Magenta: chlorophyll. Scale bar: 2 µm. Predicted functions/annotations: B8BW88, Aspartate/glutamate/uridylate kinase domain-containing protein; B8BXJ7, Uncharacterized protein; B8C8L9, Gfo/Idh/MocA-like oxidoreductase N-terminal domain-containing protein; B8C9Q9, Oxoglutarate/malate translocator; B8BZ40, Phosphoribulokinase; B8C2A9, S-malonyltransferase; B8BU77, Nucleoside-diphosphate kinase; B8LDK2, Glycoside hydrolase family 5 domain-containing protein; B8BRI7, 2-isopropylmalate synthase.

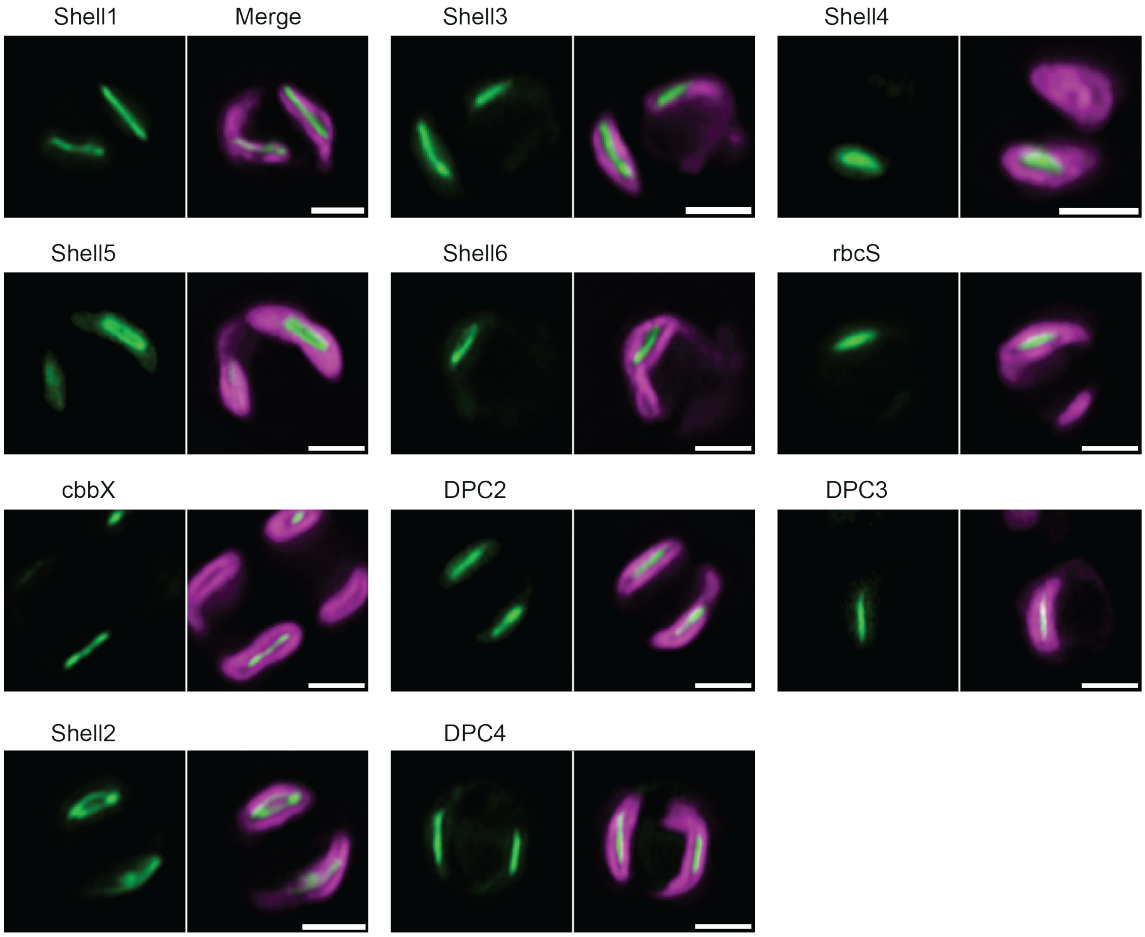


Figure S4. Images of additional mEGFP tagged lines. Images of additional pyrenoid localized proteins, and additional independently generated lines to those used in Fig. 2 and Fig. 3. Green: mEGFP fusion protein; Magenta: chlorophyll. Scale bars: 2 μm.

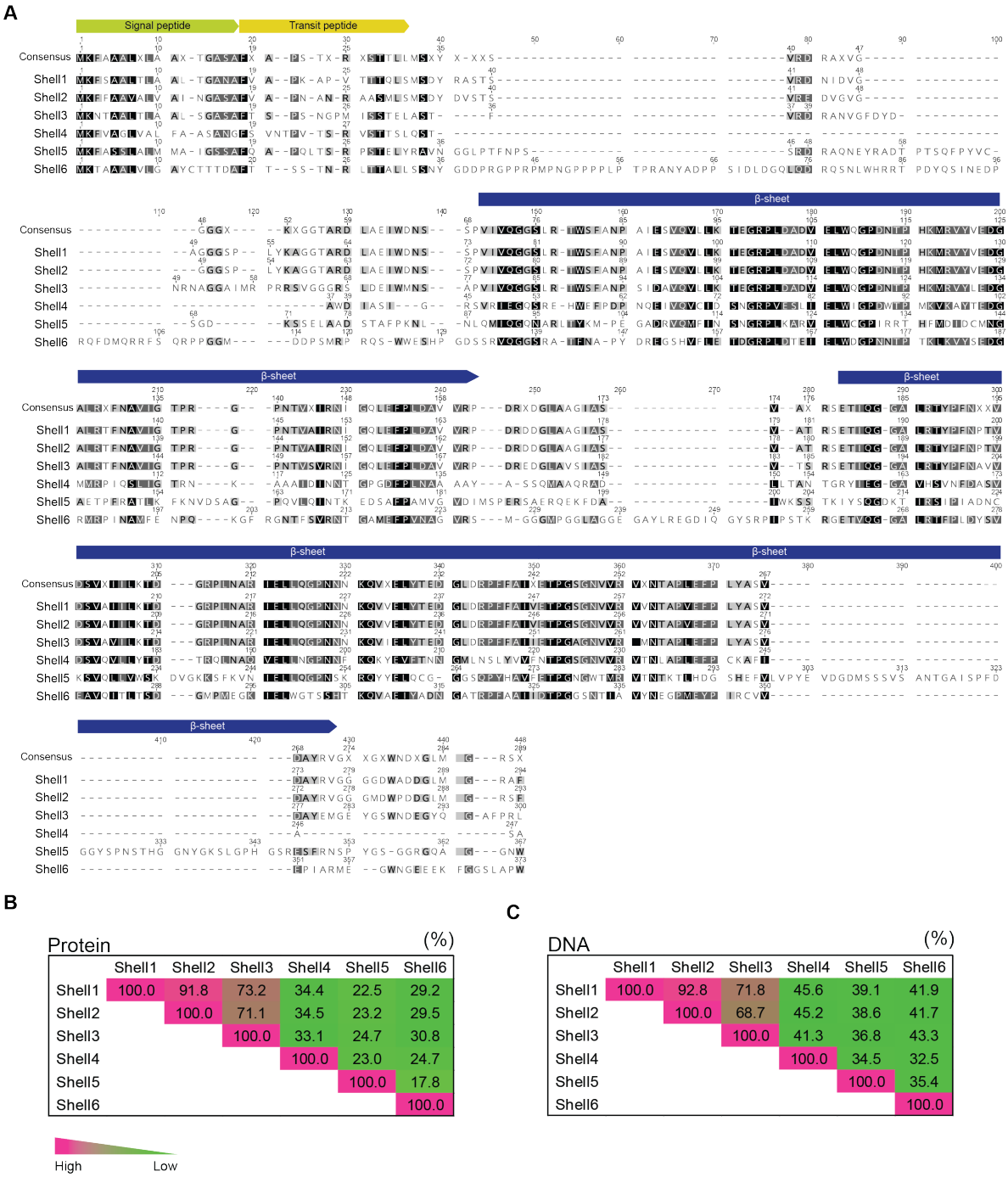


Figure S5. Shell protein amino acid alignment and sequence similarities

A. Alignment of full-length Shell amino acid sequences.

B. Protein sequence similarities between Shell homologs, including predicted signal and transit peptides.

C. Gene DNA sequence similarities between Shell homologs.

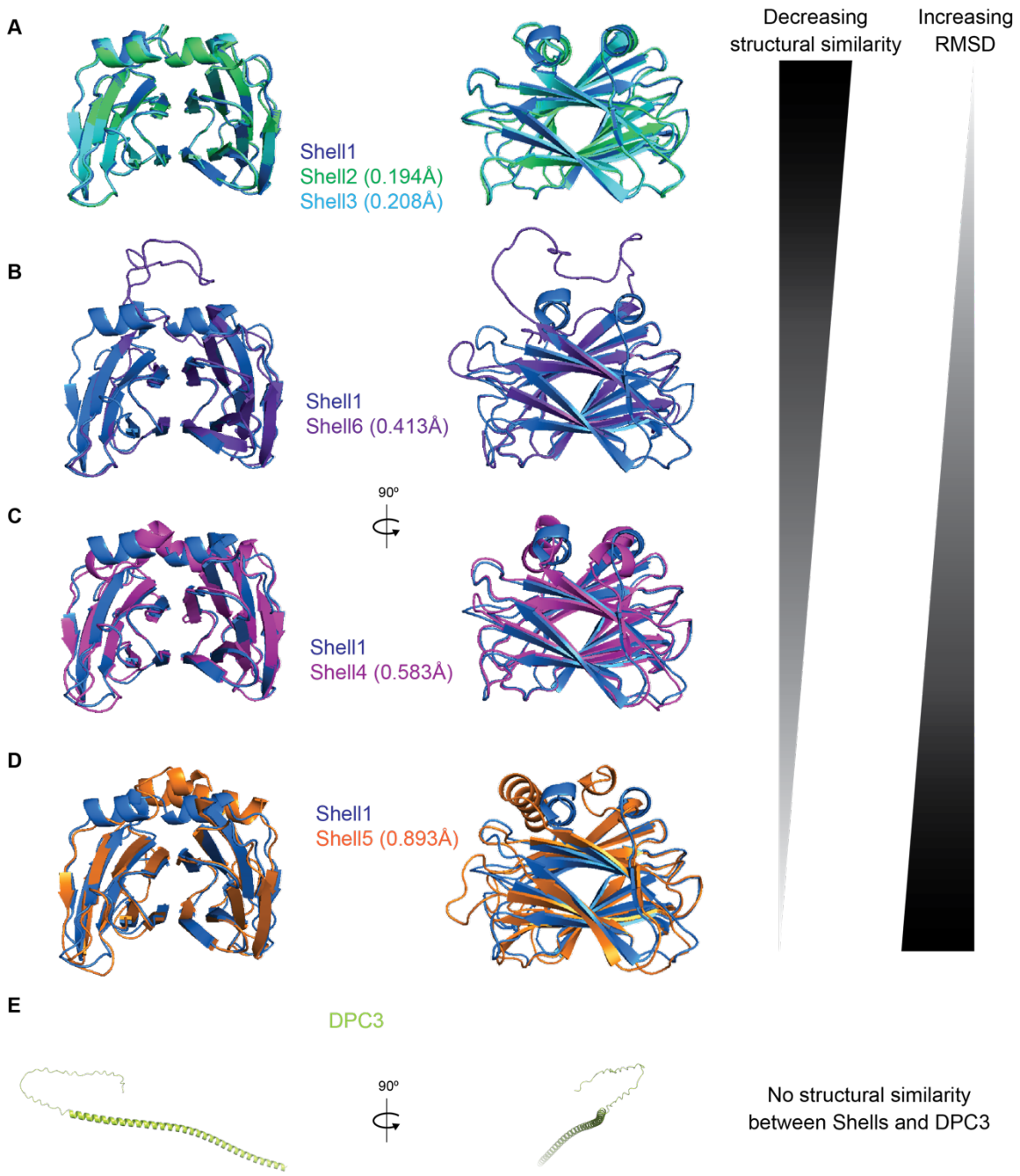
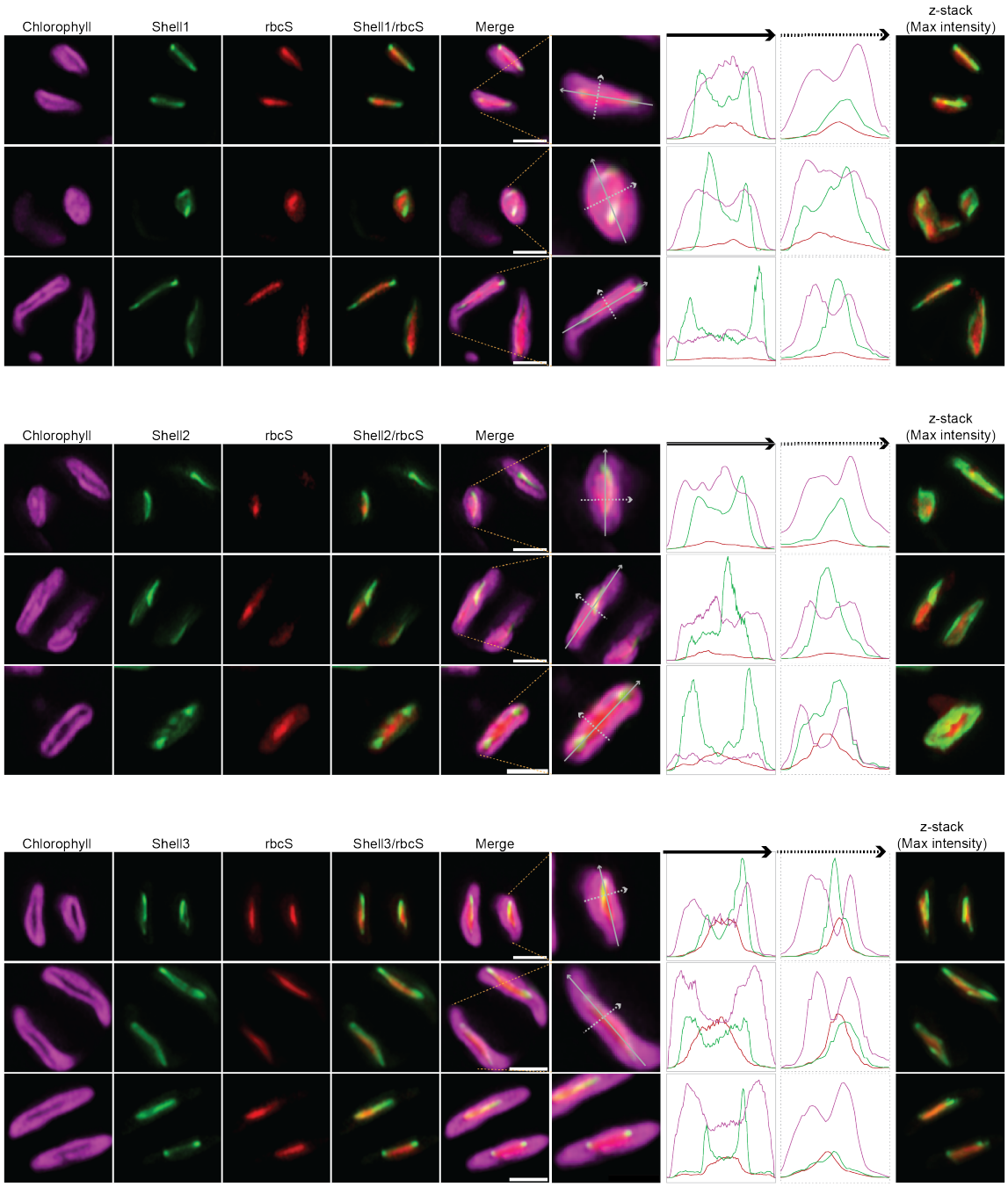


Figure S6. AlphaFold2 model comparisons of Shell proteins and DPC3. Root mean square deviation (RMSD) values relative to Shell1 are shown.

1219



1220
1221 **Figure S7 (continued on next page).**
1222
1223

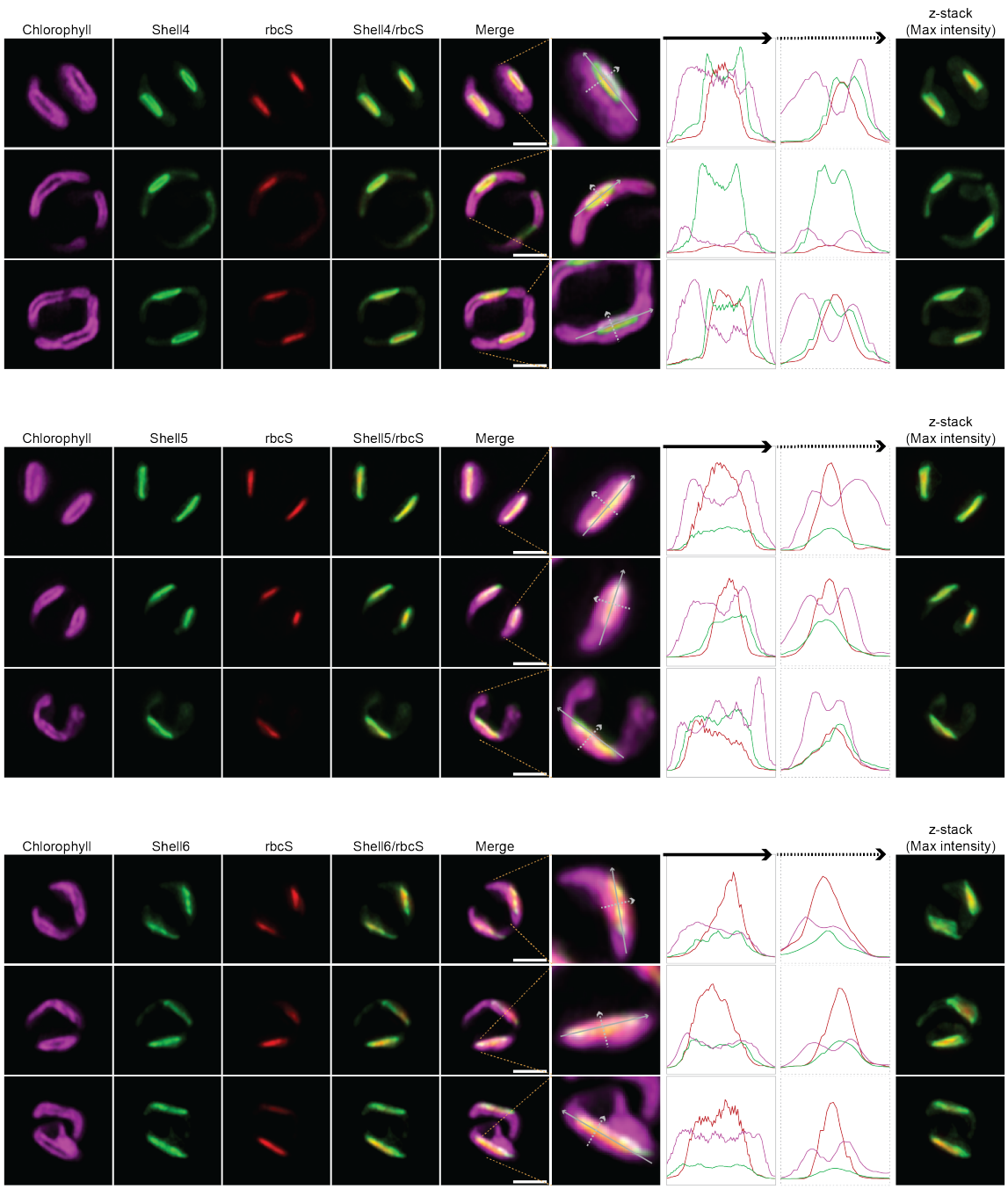


Figure S7 (continued from previous page). Co-localization of Shell proteins with Rubisco. Image columns 1-4: additional images of co-localization of Shell proteins with Rubisco. Scale bars: 2µm. Image columns 5-8: Fluorescence intensity cross-sections of chlorophyll, Shell and rbcS fluorescence across the pyrenoid. Level of Zoom changes between images. Image column 9: Z-stack max intensity projections.

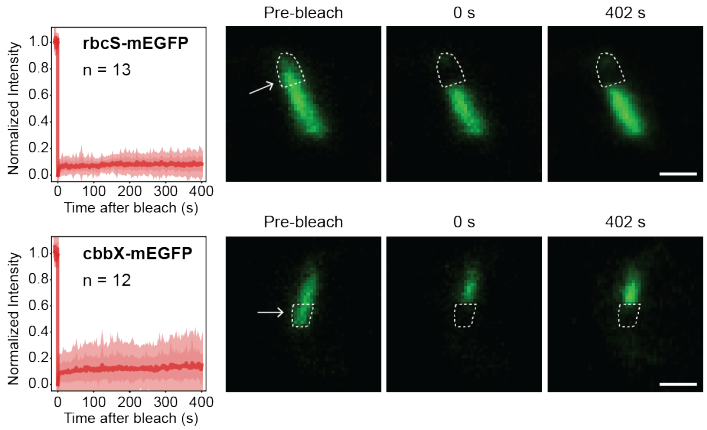


Figure S8. RbcS and cbbX are not mobile in the *T. pseudonana* pyrenoid. Fluorescence recovery after photobleaching (FRAP) experiments for rbcS and cbbX. Arrows indicate photobleached regions. Scale bars: 1 µm. Plots show S.E.M. and S.D. of the mean.

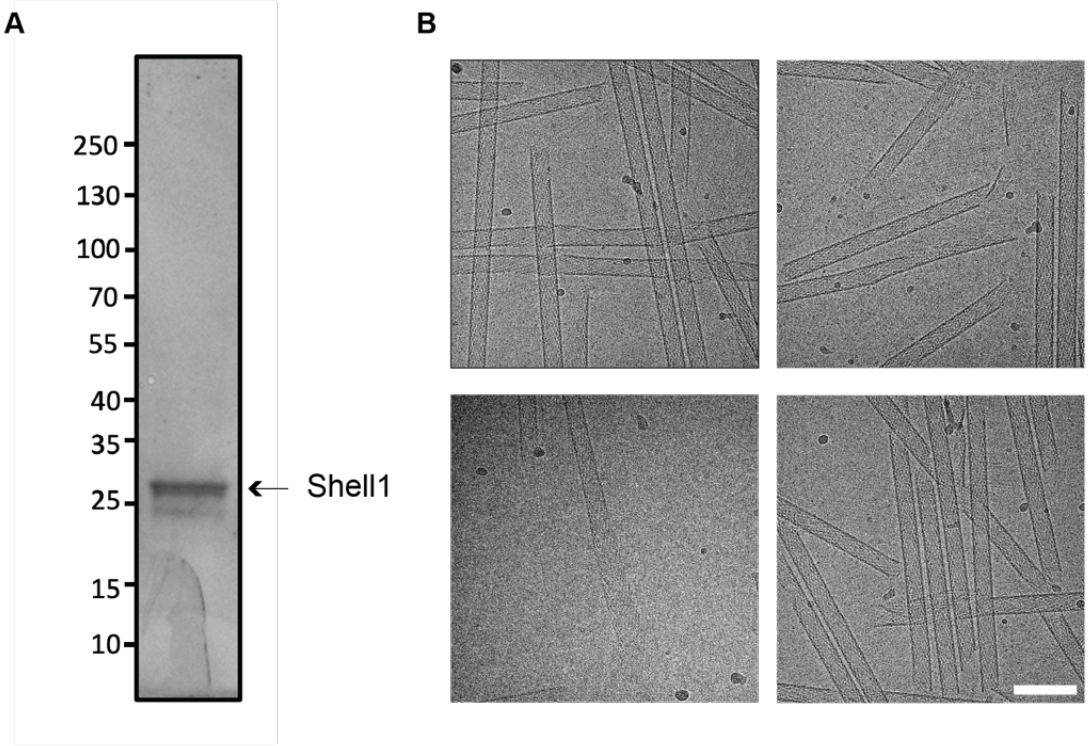


Figure S9. Shell1 forms tubes and sheets *in vitro.*
A. Coomassie stained SDS-PAGE gel of recombinant Shell1 used for cryo-EM.
B. Representative cryo-EM micrographs of Shell1 self-assembly into tubes and sheets.
Scale bar: 100 nm.

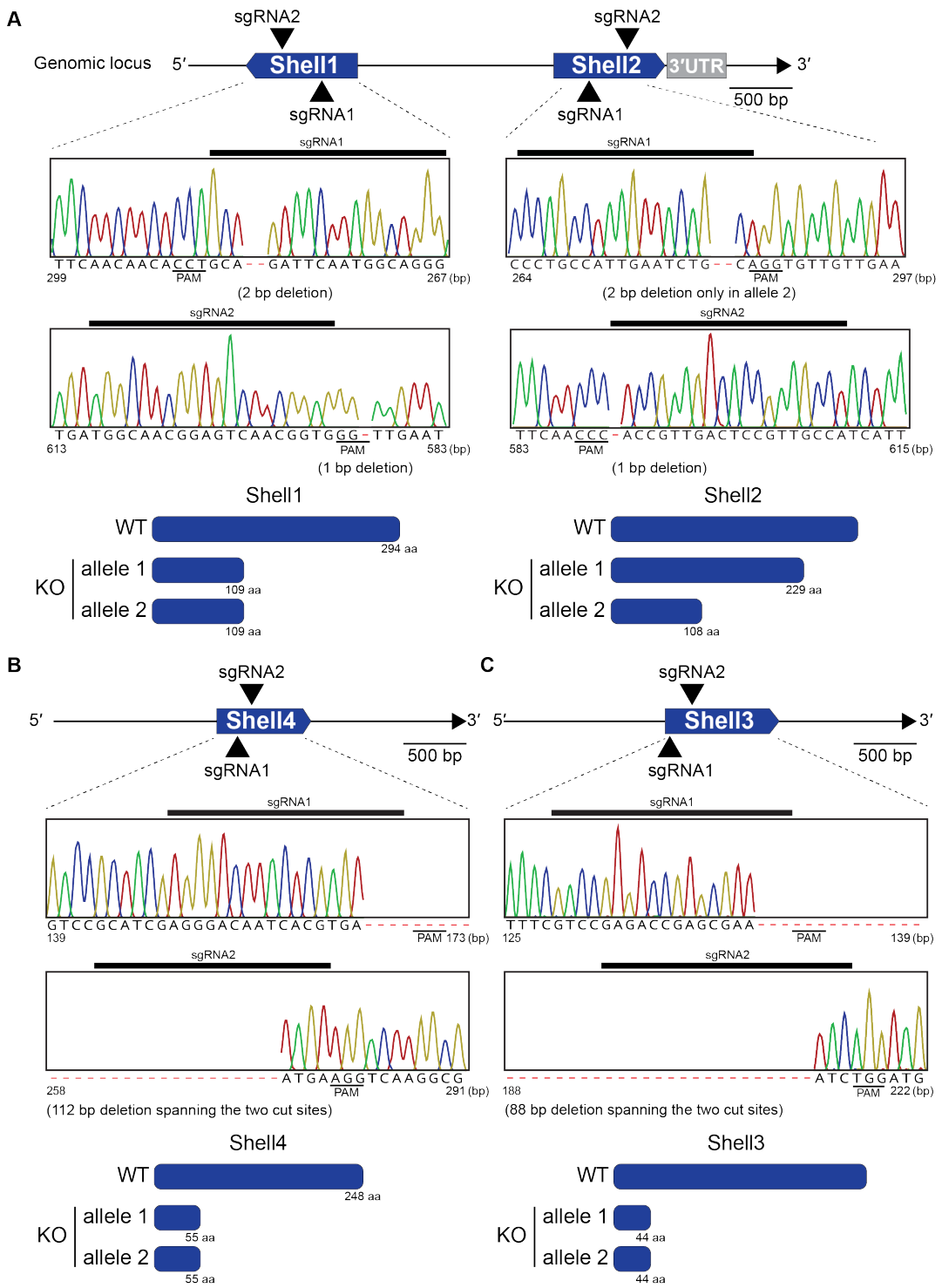


Figure S10. Shell1/2, Shell3, and Shell4 knock out genotyping.
A. Due to 93% DNA sequence similarity between *Shell1* and *Shell2*, both genes were knocked out simultaneously by designing two sgRNAs that targeted conserved sequences in *Shell1* and *Shell2*.
B. Biallelic knock out genotyping of Shell4.
C. Biallelic knock out genotyping of Shell3.

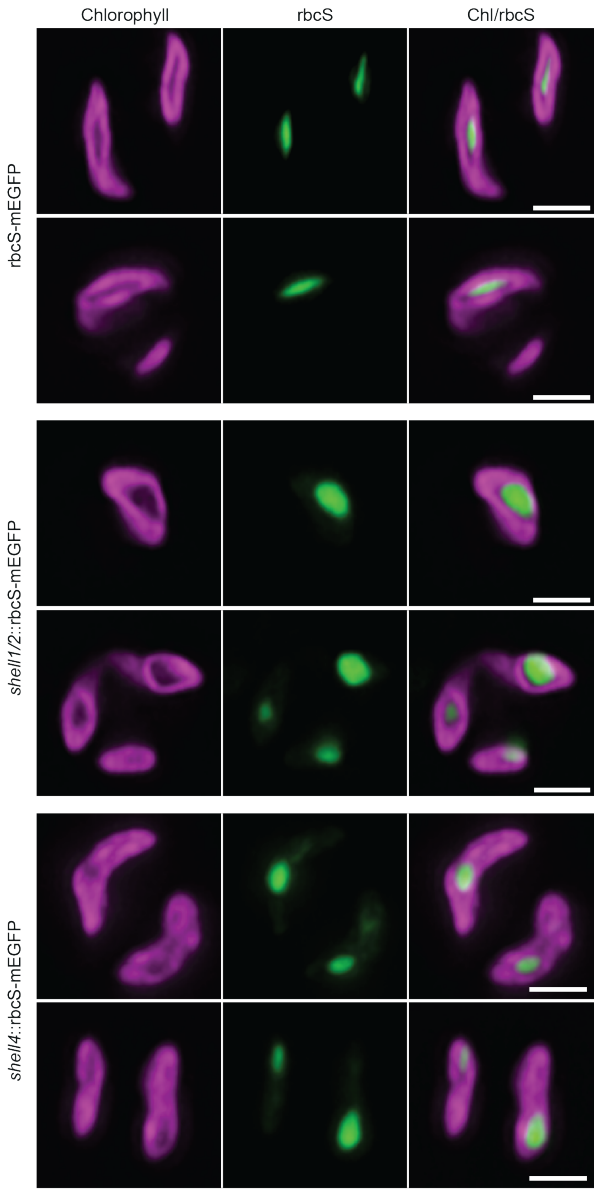


Figure S11. *Shell1/2* and *shell4* mutants have malformed pyrenoids. Additional images of WT::rbcS-mEGFP, *shell1/2*::rbcS-mEGFP and *shell4*::rbcS-mEGFP. Green: mEGFP fusion protein; Magenta: chlorophyll. Scale bars: 2 µm.

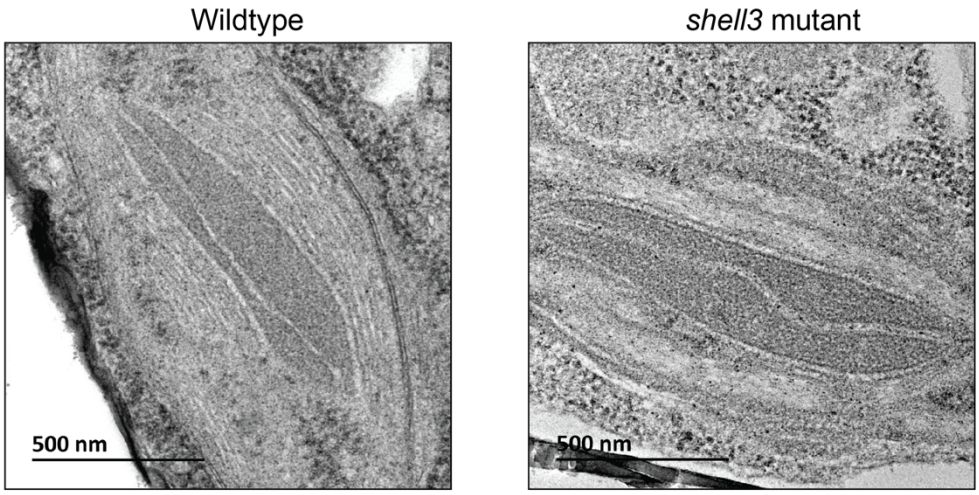


Figure S12. The *shell3* mutant has WT pyrenoid morphology. TEM images of WT and *shell3* mutant pyrenoids. The WT image is the same as used in Fig. 1.

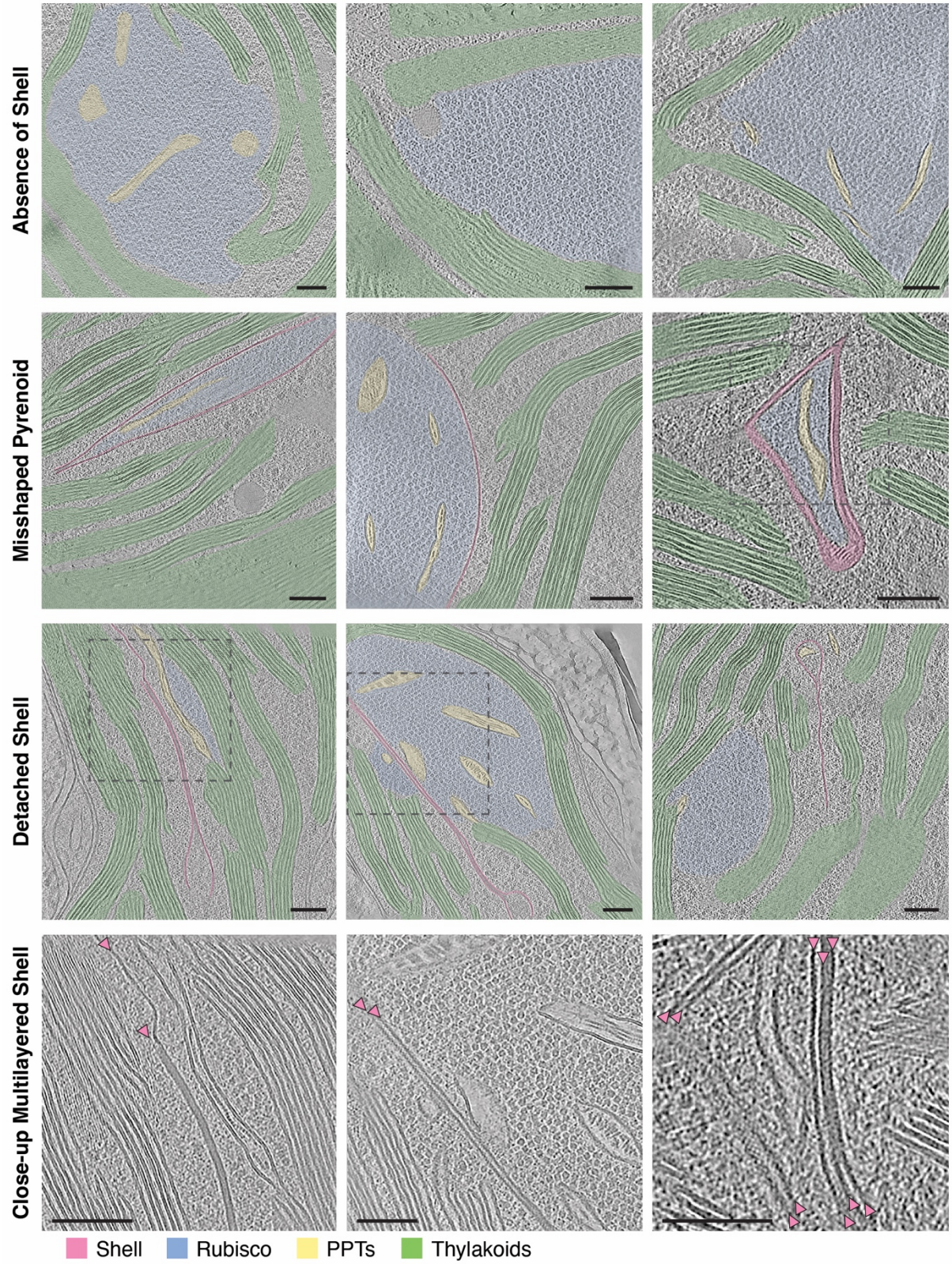


Figure S13. Additional tomograms of the *shell4* mutant pyrenoid morphologies. Additional tomogram slices of the different types of misassembled pyrenoids in the *shell4* mutant. The Shell (pink), Rubisco matrix (blue), thylakoids (green) and PPTs (yellow) are highlighted. Bottom row shows enlarge views of the boxed regions above. Scale bar: 100 nm

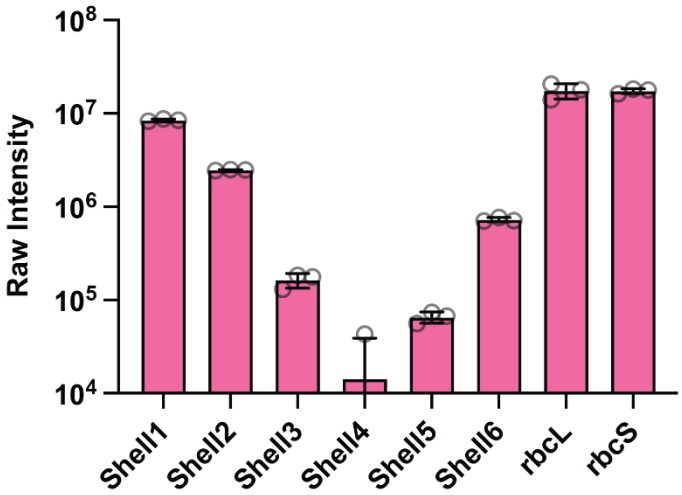


Figure S14. Whole-cell proteomics of the *shell4* mutant. Relative abundance of Shell proteins and Rubisco in the *shell4* mutant. Shell4 was detected at low levels in one of the three replicas. Error bars: S.D. of the mean.