# Global Landscape of Native Protein Complexes in *Synechocystis* sp. PCC 6803

Chen Xu[1], Bing Wang[1], Lin Yang[1], Lucas Zhongming Hu[2], Lanxing Yi[1], Yaxuan Wang[1], Shenglan Chen[1], Andrew Emili[2], Cuihong Wan[1]*

[1]*School of Life Sciences and Hubei Key Laboratory of Genetic Regulation and Integrative Biology, Central China Normal University, Wuhan, Hubei 430079, People's Republic of China.*

[2]*Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada*

Corresponding author:

Prof. Cuihong Wan

Address: No. 152 Luoyu Road, Wuhan, P.R. China, 430079

Phone number: 86–027–67861936

Email: ch_wan@mail.ccnu.edu.cn

**Running title:**

*Xu C et al / Synechocystis native protein complexes*

## Abstract

*Synechocystis* sp. PCC 6803 is a model organism for studying photosynthesis, energy metabolism, and environmental stress. Though known as the first fully sequenced phototrophic organism, Synechocystis sp. PCC 6803 still has almost half of its proteome without functional annotations. In this study, we obtained 291 protein complexes, including 24,092 protein-protein interactions (PPIs) among 2,062 proteins by using co–fractionation and LC/MS/MS. The additional level of PPIs information not only revealed the roles of photosynthesis in metabolism, cell motility, DNA repair, cell division, and other physiological processes, but also showed how protein functions vary from bacteria to higher plants due to the changed interaction partner. It also allows us to uncover functions of hypothetical proteins, such as Sll0445, Sll0446, Sll0447 participating in photosynthesis and cell motility, and Sll1334 regulating the expression of fatty acid. Here we presented the most extensive protein interaction data in *Synechocystis* so far, which might provide critical insights into the fundamental molecular mechanism in Cyanobacterium.

**KEYWORDS:** *Synechocystis* sp. PCC 6803; Protein complexes; Protein-protein interaction; Photosynthesis; Cyanobacterium

## Introduction

Cyanobacteria represent the phylogenetic ancestors of chloroplasts from present-day plants [1], [2]. The generation of oxygen through oxygenic photosynthesis by cyanobacteria is believed to change the chemistry of the atmosphere and life on earth by allowing aerobic organisms to develop and flourish [3]. *Synechocystis* sp. PCC 6803 (hereafter referred to as *Synechocystis*) is a unicellular photoautotrophic cyanobacterium, which is a widely used model organism for studying photosynthesis, energy metabolism, and environmental stress [4], [5]. The genome of *Synechocystis* was sequenced in 1996 [6], and its proteome has also been well analyzed in the last two decades [7], [8]. However, about two-thirds of its proteome in UniProt are "hypothetical protein", and most of which lack functional annotation.

Many intracellular biological processes rely on stable physical interactions between two or more proteins. Protein interaction is critical to understanding the fundamental molecular biology of organisms, which can be used for predicting annotation, finding new drug targets, etc. [9]. However, information on *Synechocystis*'s protein-protein interactions (PPIs) is quite a deficiency. Researchers have tried to analysis PPIs by using yeast two-hybrid assay and several kinds of prediction algorithm [10], [11], [12], [13]. However, in the STRING database, only 6510 PPIs of *Synechocystis* involved in 1876 proteins were annotated with "experiments" until January 2019, which contained PPIs that relevant information transferred from other organisms. Even for those well–known *Synechocystis* protein complexes, for example, Photosystem II (PSII) complexes that still need to identify novel assembly factors to understand its biogenesis better [14], [15]. The phototactic movement of cells was influenced by motility apparatus and light, but the link protein between photoreceptors and the motility apparatus remains uncertain [16], [17]. Thus, it is demanding to globally identify protein-protein interactions of *Synechocystis*, providing a useful resource to the biology community.

The high-throughput methods developed for systematically determining PPIs have led to global molecular interaction maps for model organisms, including *E.coli*, yeast,

worm, fly, and human [18], [19], [20], [21]. Several approaches have been developed for the identification of protein complexes at proteome scale, e.g., Yeast two-hybrid (Y2H), affinity purification followed by mass spectrometry (AP-MS), and co-fractionation coupled with mass spectrometry (CoFrac–MS). Among these methods, CoFrac-MS can rapidly detect hundreds to thousands of stable protein complexes involved in interactions of multiple proteins under native conditions [22]. CoFrac-MS assay has been broadly used to globally identify protein-protein interactions [23], [24], [25].

To reveal more protein interactions and uncover a clue of their biological functions, we applied CoFrac-MS to analyze the protein complexes of *Synechocystis*. In this work, we predicted 291 protein complexes containing 24,092 highly confident PPIs among 2062 proteins. This network facilitates our comprehensive understanding of the relationship between photosynthesis and other functions, such as carbohydrate metabolic process, signal transduction, ion transport, cell division, transcription, etc. In addition, we applied the protein interaction information to predict and confirm the new functions of proteins, such as Sll0445, Sll0446, Sll0447, and Sll1334. This work allows us to comprehensively understand the fundamental molecular organization and mechanism of *Synechocystis* and other Cyanobacteria species.

## Results and discussion

### Workflow for protein complexes identification in *Synechocystis*

The experimental workflow is similar to previous work **(Figure 1A)** [20]. Protein mixtures were separated and analyzed by size-exclusion chromatography (SEC), ion-exchange chromatography (IEC) or sucrose density gradient centrifugation (Suc-DGC). In total, 181 fractions were collected, and 2,906 proteins were identified (**Table S1**). Proteins in lysis samples were separated effectively according to their molecular weight or pI **(Figure S1)**. After fractionation, samples were analyzed using LC/MS/MS. The value of Rapp, the ratio of Mapp (proteins apparent molecular mass, **Figure S2**) to

Mmono (the predicted monomeric mass), was evaluated to estimate whether a protein involved a stable complex on SEC column [26], [27]. Rapp>=2 is used to classify proteins to be not oligomeric but within a complex, while Rapp≤0.5 suggests protein degradation during the protein extraction process (**Figure 1B**).

The SEC is advantageous to provide direct distribution of protein in different molecular weight [24]. However, SEC is not sufficient to separate the protein complexes with molecular weight beyond its valid separation range. To improve protein separation efficiency, we applied IEC as an additional separation technique. These two techniques are complementary in the way that IEC can separate protein complexes that are not distinguishable in SEC **(Figure S3)**. The reproducibility between two biological replicates was confirmed by Pearson correlation coefficients of the spectral counts for the full set of identified proteins **(Figure 1C)**.

The stable protein complexes bind tightly and can be detected by CoFrac-MS, while the unstable protein complexes might present unsatisfactory correlation profiling [19]. For example, photosystem components, NAD(P)H-quinone oxidoreductase, RubisCO complexes, and C-phycocyanin, tend to have a consistent correlation profiling **(Figure S4)**. These protein elution profiles confirm that chromatographic separation coupled with mass spectrometric protein identification and quantification could be utilized to provide evidence for the physical association of proteins in complexes. We used EPIC software toolkit for machine learning to define the final set of PPIs [28], with collected 47 gold standard complexes from UniProt and IntAct database **(Table S2)**. The output of machine learning was well organized and ready for subsequent analysis to predict protein complexes **(Figure 1D; Figure S5)**.

**Characteristic of high-confident PPIs and photosystem associated protein interactions**

In our datasets, 2,214 proteins participate in 35,028 highly confident protein pairs that were produced by machine learning **(Table S3)**. About 10% of PPIs have an overlap with the published database, including STRING and IntAct **(Figure 2A)**. Our study is the most extensive protein-protein interaction data in *Synechocystis* so far, which can

provide evidence for further functional study in the future.

Degree, the number of proteins that participated in the interactions, can reflect the critical role of this protein in the network [29]. We compared the distribution of protein degrees in *Synechocystis* with several different model organisms, such as *E.coli, S.cerevisiae, A.thaliana, and H.sapiens,* according to the Mentha database [30]. We observed that most of the proteins tend to have low degrees, and these proteins tend to be annotated with different metabolism pathways in *Synechocystis*, for example, organonitrogen, aromatic, and cellular nitrogen compound biosynthetic process **(Figure 2B; Figure S6)**. In contrast, the ribosomes and heat shock proteins present high degrees in our dataset and are conservative across different species **(Figure 2B)**. The ribosomes and heat shock proteins, such as DnaK1 and DnaJ, are usually highly conserved during species evolution and occupy an important position in metabolism [31]. The conserved proteins tend to have high degrees [32]. Additionally, the proteins that have homologous in *E.coli, A.thaliana, H.sapiens* or *S.cerevisiae* were found to have higher degrees than the proteins that unique in *Synechocystis* **(Figure 2C)**.

As a model organism to study photosynthesis, *Synechocystis* has a classical photosystem structure containing Photosystem I (PSI), PSII, Cytochrome $b_6f$ complexes, and ATP synthase complexes [5]. There are 70 proteins annotated with photosynthesis and 760 proteins were found to have direct interactions with them in our dataset **(Table S4)**. Photosynthesis associated proteins were clustered into 26 different groups based on their functions, including phosphorelay signal transduction system, potassium ion transport, DNA repair, cell division, carbohydrate metabolic process, transcription, DNA-templated, translation, chemotaxis, and cell redox homeostasis **(Figure 2D)**. The results indicate how photosynthesis influences many critical biological processes other than photosynthetic carbon fixation. For example, we observed that several CheA-like proteins interact with photosynthetic core proteins in our database. The CheA-like proteins contain phosphorelay sensor kinase activities and participate in chemotaxis in cells [16]. It suggested that cells might control motility direction by regulating the state of CheA-like proteins through phosphorelay in

photosynthesis. The protein transport by Sec complexes also had an association with photosynthesis proteins. The proteins SecD and SecF might participate in the PSII assembly process by interacting with different photosystem II core proteins with a similar function as SecY protein **(Figure 2D)**. SecY is the main transmembrane subunit of preprotein translocase that is essential for PSII assembly for co-translational pD1 insertion by interacting with YidC insertase [33], [34]. In our database, the protein ChlD was also observed to interact with SecD, which indicated that it might be functional to deliver chlorophyll to the newly synthesized pD1, carrying a similar role as protein ChlG **(Figure 2D)** [34]. Besides, we observed that proteins interacting with photosynthetic proteins are mainly localized in the cytoplasm, while this class of proteins has possible mobility or secretion according to localization identification results [35]. The results provided a clue for understanding how photosynthetic proteins to transmit signals and influence the physiological metabolism in living cells.

**The landscape of native protein complexes in *Synechocystis***

With the highly confident protein pairs assignment from machine learning, we predicted 291 protein complexes with 24,092 PPIs by using the ClusterONE algorithm [36], which is implemented in EPIC software. The set of predicted putative complexes contains well known and highly-conserved complexes such as Photosystem I protein complexes, RubisCO complexes, and NADH dehydrogenase **(Figure 3; Table S5)**. The proteins were shown in different colors according to their subcellular location, while most of the complex components have multiple annotated locations, and marked as "Overlap". The functions of predicted complexes are involved in photosynthesis, carbon absorption, nitrogen fixing, and electron transfer. Besides, we also observed some proteins with unclear molecular function annotations formed stable protein complexes, such as CRISPR3 system proteins Sll7085 – Sll7090, photosystem proteins Sll0144 – Sll0149, and ATP-dependent zinc metalloprotease FtsH complexes **(Figure 3)**.

According to computational analysis, *Synechocystis* has three types of CRISPR-Cas system, but their exact functions remain unclear [37], [38], [39]. In our dataset, we

found that some proteins of CRISPR3 can form stable complexes, and the protein interactions between different CRISPR systems were also observed **(Figure S7)**. These results suggest that different types of CRISPR systems can coordinate with each other to defend cells from virus and plasmid rather than perform function independently. The hypothetical proteins Slr0144–Slr0149 are located on the thylakoid membrane and contain putative bilin binding domains, 4-vinyl reductase (V4R) domains, and 2Fe–2S cluster binding domains. They are mainly involved in the function of photosynthetic repair [40]. The Slr0144–Slr0149 protein interactions is validated by APMS **(Figure S8)**. Among the predicted complexes, we found that the hypothetical protein, Sll0445, Sll0446, and Sll0447 formed protein complexes with pilus assembly proteins, Slr2015 and Slr2018 and photosystem complexes. The physical interactions between Sll0445 and photosynthetic proteins verified by the AP-MS experiment. The AP-MS also obtains other partners that not coordination with CoFrac-MS (**Table S6**).That might because different methods tend to catch different portions of physical interactions [41]. It is worth mentioning that the proteins Slr2015 and Slr2018 can interact with Sll0445 **(Figure 3)**. The Slr2018 is located at the plasma membrane and regulated by a cAMP receptor protein SYCRP1 that is responsible for the cell motility in *Synechocystis* [42]. The protein Slr2018 is not homologous to other known proteins, but its gene is adjacent to genes *Slr2015*, *Slr2016*, and *Slr2017*. These genes have roles in pilus morphology and motility, and their proteins were predicted to form protein complexes with photosynthetic proteins **(Figure 3)** [43]. The results reveal a close connection presents between photosynthesis and cell motility in *Synechocystis*.

**Conservative analysis of *Synechocystis* protein complexes**

Half of our predicted protein complexes are conserved with homologous in other species, like *A.thaliana* or *E.coli*, while the rest of them are unique in *Synechocystis* **(Figure S9)**. The predicted protein complexes in *Synechocystis* have a higher similarity with *A.thaliana* and *E.coli* compared to *H.sapiens* and *S.cerevisiae*. Because *Synechocystis* has a close relationship with *A.thaliana* and *E.coli* for the metabolic feature of Gram-negative bacteria and photosynthesis of green plants. The

conservativeness of protein complexes was evaluated by the number of proteins with known homologies in each complex. There are 920 proteins and 813 proteins with homologous in *E. coli* and *A. thaliana*, respectively **(Table S7)**. To estimate the evolutionary relationship of *Synechocystis* with *A.thaliana* and *E.coli*, we separated our predicted complexes into different parts according to their homologous proteins. The components of the predicted complexes were divided into four parts, homologous in *E.coli*, homologous in *A.thaliana*, homologous in both *E.coli* and *A.thaliana*, and unique in *Synechocystis* **(Figure 4A)**. Proteins having homologous in *E.coli* and *A.thaliana* retain functions as energy metabolism, organic synthesis, protein expression, and regulation, which are considered as basic activities to be conserved across most species. Proteins with these functions play important roles in maintaining the necessary life activities. Proteins with homologous in *A.thaliana* are mainly related to photosynthesis and pigment biosynthetic process, which is consistent with the fact that one of the most remarkable features shared between *Synechocystis* and *A.thaliana* was photosynthesis. Proteins with homologous in *E.coli* were primarily associated with the polysaccharide metabolic process, protein transport, and coenzyme biosynthetic process, since both *Synechocystis* and *E.coli* belong to Gram-negative bacteria. Cyanobacteria and bacteria share a high sequence homology as evidenced by 16S rDNA analysis. Thus, they have a large number of similarities in cellular structure and physiological properties [44], [45].

It is noteworthy that the function oxidation-reduction, homeostatic process and two-component signal transduction system (phosphorelay) can be elucidated in more than one of these four groups. However, the proteins that performed these functions and their interaction partners were different **(Figure 4A)**. For example, the proteins annotated as "oxidation-reduction" have homologous in *A.thaliana* and *E.coli* mainly participate in the "generation of precursor metabolites and energy" and cellular respiration processes. While the function of proteins annotated as "oxidation-reduction" with homologous in *A.thaliana* has changed and is mainly associated with photosynthesis **(Figure S10)**.

The localization of NAD(P)H-quinone oxidoreductases (NQOs), which involves electron transfer and shuttles electrons from electron donors to quinones, is different in *E.coli*, *A.thaliana*, and *Synechocystis* [46], [47], [48]. The NQOs are on the plasma membrane and participate in the respiratory process in *E.coli*, where the same complexes are found on thylakoid membrane in *A.thaliana* and *Synechocystis*, indicating that they were involved in the process of photosynthetic electron transport [49], [50]. Here we identified 14 NQQs protein (**Table S1**), and 11 of them form complexes in our final protein network (**Table S5**). However, the components and functions of NQOs are not entirely consistent in different species [51]. We compared partial NQOs protein interactions of *E.coli*, *A.thaliana*, and *Synechocystis.* We found that the protein NdhM is only found in *Synechocystis* and the protein NdhJ only exists in *A.thaliana* and *Synechocystis*. In contrast, NdhK and NdhH were conserved in three species **(Figure 4B)**. The results indicated that some components of NQOs participate in the respiratory electron transport and other components may evolve new functions. For example the proteins of NdhJ and NdhM might be involved in photosynthesis in *A.thaliana* and *Synechocystis*. The appearance of oxidation–reduction in parts of homologous with *A.thaliana* can be a result of these proteins adapt to photosynthetic damage or participate in electron transport in *Synechocystis*. Several reactive oxygen species (ROS) are produced as byproducts of aerobic metabolism in photoautotroph, which were disposed of by antioxidants [52], [53]. The presence of new proteins with oxidation-reduction function can effectively reduce the influence of oxidative stress.

Bacterioferritin is an iron–storage protein, whose ferroxidase center binds to and oxidizes $Fe^{2+}$ iron to $Fe^{3+}$ by oxygen. The complexes of bacterioferritin were studied and compared across three species. The results show that there is no ortholog in *A.thaliana* and only one ortholog in *E.coli* **(Figure 4B)**. It is suggested that other proteins might perform this function in *A.thaliana.* The photoautotrophic microorganism tends to have a massive demanding for inorganic ions, such as Fe and Mn. Because inorganic ions are usually deficient in the open ocean, a unique mechanism for the microbe to transport and store ions is required. [54], [55]. The

presence of two orthologs of bacterioferritin in *Synechocystis* indicates that *Synechocystis* may have an additional requirement for irons compared to *E.coli* **(Figure 4B)**. The protein components of bacterioferritin complexes in *Synechocystis* have annotated functions of irons transport that are absent in *E.coli* **(Figure 4B)**. Like bacterioferritin and NQOs mentioned above, the conservative analysis of predicted complexes allows us to explore the biological process within cells and how organisms adapt to environments.

**New functions of Sll0445–Sll0447 and Sll1334 predicted from protein complexes**

Figure 3 indicates that hypothetical protein Sll0445, Sll0446, Sll0447 have interactions with pilus assembly proteins and photosystem complexes. Sll0445 contains a Tubulin_2 domain, a member of clan Tubulin **(Figure S11A)**. This clan Tubulin subunits are cytoskeletal elements that are essential for intracellular transport and cell division in all eukaryotes [56]. Sll0446 contains a FtsA domain, which is an actin-like ATPase domain and co-localizes to the septal ring with FtsZ **(Figure S11A)** [57]. The Sll0447 contains a DivIC domain, which is necessary for the formation of both vegetative and sporulation septum **(Figure S11A)** [58]. We constructed interruption mutants for these three genes, respectively, through the insertion of chloramphenicol–resistance marker (CmR) into their ORFs **(Figure S11B and S11C)**. The △*sll0445*, △*sll0446*, △*sll0447* and the wild type were tested for phototactic motility. In the unidirectional white light, the wild type showed rigorous movement toward the light, while the △*sll0445*,△*sll0446*, and△*sll0447* exhibited impaired ability of phototaxis **(Figure 5A)**. The transcriptome analysis demonstrated that the transcriptional level of *sll0447* was decreased under the mutant of Sycrp1 [42], a gene involved in the phototactic movement [59], indicating that the *sll0447* was regulated by the Sycrp1 and influenced the cell motility.

The hypothetical protein, Sll0445, Sll0446 and Sll0447, not only can form protein complexes with pilus assembly proteins, but also interact with photosystem complexes. The genetic deletion of this gene shows that the removal of these protein products can affect electron transport of PSII under optimal conditions or low – temperature

conditions (20℃) **(Figure 5B and 5C)**. *Synechocystis* exhibits a typical polyphasic rise of fluorescence induction called the OJIP fluorescence transient. Similar findings for plants, green algae and cyanobacteria have been described previously [60]. Figures 5B and 5C show sharply increases in the fluorescence yields of mutants at phases J than the wild type, suggesting that the $Q_A$ reduction rate and the $Q_A^-$ oxidation rate were influenced by genetic deletion of *sll0445, sll0446* and *sll0447*. The results demonstrate that the proteins encode by *sll0445, sll0446* and *sll0447* are necessary for optimal function of photosystem and cell motility. However, it is still noteworthy that the pilus assembly proteins distributed on the plasma membrane and photosynthetic proteins localized on the thylakoid membrane. The difference of these proteins in spatial distribution causes the difficulties for proteins to localize on the thylakoid membrane to directly affect a protein complex distributed on plasma membrane unless a specific protein complex was formed. So far, we know that part of the respiratory electron transport chain in *Synechocystis* is localized on the thylakoid membrane and partially overlapped with photosynthetic electron transport chain [49]. In our dataset, the PSI subunits can form a protein complex with NADH dehydrogenase as part of the respiratory chain complexes. The main purpose of *Synechocystis* motility is to obtain plenty of light and consume much energy during chemotaxis. Presumably, the proteins of Sll0445, Sll0446, and Sll0447 may link the structure of thylakoid membrane with the plasma membrane to adjust the distribution of light on photosystem and chemotaxis.

A histidine kinase protein Sll1334 interacted with acetyl coenzyme A carboxylase (ACC) **(Figure 3)**. Sll1334 contains GAF domain, which can sense and respond to light signals **(Figure S11A)** [61]. We also observed that the transcript levels of acetyl coenzyme A carboxylase (ACC) complex were upregulated in *sll1334* interruption mutant compared to the wild type strain **(Figure 5D)**. ACC complex catalyzes the acetyl–CoA to form malonyl–CoA, which is part of lipid metabolism and conserved across different species [62]. Okada *et al*. have studied that *sll1334* may function as a suppressive regulator in the cascade and influence cell growth and gene expression involved in glycometabolism under dark conditions [63]. The mechanism is still unclear

as to how cyanobacterium adapts to dark conditions and uses glucose as the carbon source for growth. However, our results may provide new insight into this process. Carbohydrate and lipid metabolisms were considered as central metabolism in living cells and have a close relationship with each other to control basic vital activities. Okada *et al*. also have found that *Sll1330*, a histidine kinase gene, is located at the upstream of *sll1334* and influences the expression of *sll1334* [64]. The △*sll1330* and △*sll1334* did not grow well either under light activated heterotrophic growth (LAHG) or dark heterotrophic conditions compared to the wild–type but could grow under photoautotrophic and mixotrophic conditions as fast as the wild–type [63], [64]. These results indicated that carbohydrate and lipid metabolisms in *Synechocystis* were regulated by Sll1330 and Sll1334, and the light was an essential factor to influence this process. Cyanobacteria can sense various wavelengths of light to activate a range of bilin-binding photoreceptors and many vital cellular processes including growth, phototaxis, cell aggregation, and photosynthesis are tuned to adjust environmental light conditions by these photoreceptors [61]. The photosensory module of photoreceptors is found in the amino-terminal region of the protein and consists of a 'knotted' structure containing a PAS domain, GAF domain, and PHY domain, with the bilin bound by the GAF domain, but not all proteins with GAF domain are photoreceptors. Thus, it needs further confirmation whether Sll1334 is a photoreceptor and how it regulates the ACC complexes expression by sensing light.

Finally, through integrating our data with the cyanobacterial metabolism pathway, we depicted a more comprehensive working model of the phototaxis regulation, photosystem II assembly and physiological metabolism in *Synechocystis* **(Figure 6)**. Light is one of the essential elements to sustain life, especially for cyanobacteria and green plants that can fix carbon from the outer environment by photosynthesis. In order to obtain enough light, cyanobacteria have been evolved with multiple abilities to adapt to light changes. For example, light energy can be harvested by large antenna complexes–phycobilisomes and phototaxis is capable of helping *Synechocystis* locate an ideal place to collect light [16], [65]. The phototaxis can be influenced and regulated

by light, the concentration of cAMP, and the structure of pilus in *Synechocystis* [16]. The Slr2015–Slr2018 is one class of proteins involved in the process of pilus assembly and can be induced by cAMP receptor protein Sycrp1 [42]. The hypothetical proteins Sll0445, Sll0446 and Sll0447 are also induced by Sycrp1 and interact with protein pilus assembly proteins, Slr2015 and Slr2018 to regulate cell motilities **(Figure S8)** [42]. Besides, the proteins Sll0445, Sll0446, and Sll0447 can also influence photosynthesis through interactions with photosynthetic core proteins, revealing a close relationship between photosynthesis and cell motility in *Synechocystis* **(Figure 5B and 5C)**. We also provided more detailed information on the PSII assembly processes by combining our predicted PPIs with published data **(Figure 6)**. Biogenesis of PSII requires coordinated assembly of at least 20 protein subunits, as well as the incorporation of various organic and inorganic cofactors [14]. Some of the proteins were well understood during pratA–dependent PSII assembly processes. For example, proteins SecY, YidC, and CtpA can facilitate D1 maturation from pD1; proteins ChlG, HliD, and Ycf39 were involved in the process to assist the delivery of chlorophyll to newly synthesized D1 or pD1 [34], [66], [67]. In addition to these proteins, we observed the proteins SecD, SecF, and ChlD presented a good elution profiling with these proteins, which indicated that Sec complexes and ChlD might play a critical function in the maturation of PSII during pratA–dependent PSII assembly processes **(Figure 6)**. *Synechocystis* can survive under photoautotrophic or heterotrophic growth conditions. The protein Sll1334 can regulate the expression of genes involved glucolipid metabolism and GAF domain in Sll1334 may have an important role in controlling the utilization of sugars and lipids during heterotrophic growth. These proteins could expand our understanding of the regulation of cell motilities, photosystem II assembly, and glucolipid metabolism in *Synechocystis*.

## Conclusion

In this study, combing CoFrac-MS and quantitative proteomics strategies, we are able

to predict 291 protein complexes consisting of 24,092 highly confident protein-protein integrations in *Synechocystis*, which is the largest protein interaction dataset of *Synechocystis* so far. The comprehensive PPI information greatly enhanced our understanding of the molecular mechanisms of photosynthesis as well as other fundamental molecular organization in cyanobacterium.

From the predicted PPIs, most of the proteins that tend to have lower degrees involved in metabolic regulation, whereas the proteins with higher degrees are endowed with more basic and conserved functions in *Synechocystis*. By separating photosynthetic proteins interaction networks from whole PPIs, we elucidated how photosynthetic proteins connect with many other functional proteins and influence plenty of necessary biological processes. By comparing protein complexes in *Synechocystis* with other species, including *A.thaliana*, *E.coli*, *H.sapiens*, and *S.cerevisiae*, we observed protein complexes evolution and functional variations in different species. For example, the change of NQOs components in different species may be due to that photosynthetic organisms, like *Synechocystis* and *A.thaliana*, had to undertake the process of photosynthesis and recovery from photosynthetic damage. According to the predicted complexes, the hypothetical proteins Sll0445, S0446, and Sll0447 were found to build a functional connection between photosynthesis and cell motility. Photosynthetic apparatus serves as a regulator in energy metabolic process in living cells. Photosynthesis has a close relationship with chemotaxis, because one of the primary purposes of cell motility is to obtain light so that photosynthesis can use light energy to fix $CO_2$ for provided energy in the reverse of cell motility. Moreover, the expression of acetyl coenzyme A carboxylase complex was upregulated when the protein Sll1334 was depleted. Cyanobacterium was recognized as a promising microbial cell factory for producing biofuels, but current productivity in cyanobacterial systems is low. The protein Sll1334 was found to be a negative regulator of glycolipid metabolism, which provided us new insights to improve biofuel productivities by genetic modification.

The global landscape of native protein complexes in *Synechocystis* provides a

valuable new resource for the research community to find and characterize novel components of known complexes, assess whether individual proteins of interest appeared in larger complexes, and expand our knowledge of protein interaction network, which governs the life of cyanobacteria.

## Materials and methods

### Cell culture and protein extraction

The wild-type strain of *Synechocystis* was obtained from the Pasteur Culture Collection and cultured in BG11 medium at 30 °C under continuous illumination (30 μmol $m^{-2}$ $s^{-1}$). The cells were collected by centrifugation (6000g at 4℃ for 5min) when grew to exponential phase ($OD_{730}$ = 0.8 –1). Subsequently, the cells were resuspended in lysis buffer containing 20mMTris-Cl (pH 7.5), 150mM NaCl, 1%DDM (Catalog No: D4641, Merck, Darmstadt, Germany), and Complete Protease Inhibitors EDTA–free (Catalog No: 4693124001, Roche, Basel, Switzerland) and applied to sonication (5s on, 10s off) for about 5min on ice with an output of 135W. The whole cell lysate was centrifuged (12000g at 4℃ for 10min) to remove the cell debris. The protein concentration of each lysate mixture was measured using the Bradford assay.

### Size-exclusion chromatography, ion-exchange chromatography, and sucrose density gradient centrifugation

*Synechocystis* cell lysate was fractionated by Size-exclusion Chromatography (SEC), Ion exchange chromatography (IEC), and sucrose density gradient centrifugation (Suc–DGC) by using a Thermo Scientific Ultimate 3000 HPLC system. These three fractionation methods were preformed individually not sequentially. The lysates were injected (350ul per injection) onto MAbPac SEC–1 (Thermo Scientific, 5μm, 300 x 4.0mm) or Superose 6 10/300GL column (GE Life Sciences) equilibrated PBS (pH=7.2). There are 24 100ul fractions were collected by using MAbPac SEC–1, with flow rate 0.2ml $min^{-1}$. And there are 45 300ul fractions each time were collected by using Superose 6 10/300GL column, with flow rate 0.3ml $min^{-1}$. Protein standards

(thyroglobulin, BSA, Albumin egg and myoglobin) were analyzed with the same method to obtain the approximate molecular weight range across fractions. A PolyCATWAX mixed-bed ion-exchange column (200 x 4.6 mm id, 12 μm, 1500 A, Columnex, San Diego, CA, USA) was also used for lysate samples, and a 110 min salt gradient (0.12 to 1.2 M NaCl) was used to collect 43 fractions. The elution composition %A containing 10mM Tris–HCl (pH=7.6), 0.5mM DTT and 5% glycerin, while %B with additional 1.2M NaCl. Lysate samples were loaded onto a 12ml 15-70%(w/v) linear sucrose gradient, which was then centrifuged at 160,000×g for 16 h at 4°C in a Beckman MLS–50 rotor using an Optima Max ultracentrifuge (Beckman–Coulter, CA, USA), and 24 fractions were collected. In total, 181 fractions were collected.

**Affinity purification mass spectrometry**

The target protein was combined with green fluorescent protein (GFP) on its C–terminal. Cell lysis contained GFP tag strain was subjected to affinity purification by using Anti–GFP antibody (ab290, Abcam, Cambridgeshire, UK). The process of antibody purification was using GenScript Protein A MagBeads according to the manufacturer's instructions (Catalog No: L00273, GenScript, Piscataway, NJ, USA). Then, the sample was detected by mass spectrometry.

**Trypsin digestion, and peptide clean up**

Proteins from all HPLC fractions were precipitated with 10% Trichloroacetic Acid overnight at 4°C, dried at room temperature, and then dissolved in 50mM ammonium bicarbonate. Trypsin (Catalog No: V5113, Promega, Madison, WI, USA) was added at the ratio of 1:50 and incubated overnight at 37°C. Each fraction desalted using ZipTip C18 plates (Catalog No: ZTC18S960, Millipore, Darmstadt, Germany), with eluted in 80% acetonitrile, 0.1% formic acid. Peptides were dried using a LABCONCO evaporate and resuspended in 0.1% formic acid.

**LC-MS/MS and analysis of spectra**

The peptides were dissolved in the 0.1% formic acid, and analyzed by online nanoflow LC-MS/MS using an easy nanoLC-1000 system (Thermo Fisher Scientific, Waltham,

MA, USA) connected to a Q-Exactive (Thermo Fisher Scientific, Waltham, MA, USA) mass spectrometer. Peptides in 0.1% formic acid were injected onto a C18 nano–trap column using an Ultimate 3000 nano-HPLC system (Thermo Fisher Scientific) and eluted at a flow of 500nL/min with a 100min gradient from 10% solvent B (100% ACN/0.1% formic acid, vol/vol) to 90% solvent B. Peptides were ionized by nanoelectrospray ionization at 2.0 kV and analyzed by Q-Exactive mass spectrometer (Thermo Fisher Scientific) with HCD fragmentation of precursor peptides. The MS/MS spectra of the top 20 most abundant precursor ions were acquired by using a data-dependent method. The dynamic exclusion duration was set to be 40s with a repeat count of 1 and ±10 ppm exclusion window. Automatic gain control (AGC) was used to prevent overfilling of the ion trap; $5 \times 10^4$ ions were accumulated for generation of MS/MS spectra. The RAW data files were analyzed using Sequest version 2.1, supplied with the *Synechocystis* 3508 protein database from UniProt database. The mass tolerance of precursor ion was set to 10 ppm and MS/MS mass tolerance was set at 0.02 Da. The enzyme was set to trypsin, allowing up to 2 missed cleavages. The false discovery rate for peptide and protein level identifications were set at 1%, using a target–decoy based strategy.

**Data analysis**

Data analysis was performed using R Language and Python scripts, in conjunction with numpy, scikit–learn, pandas and matplotlib libraries. Elution profile for individual protein was normalized and smoothed by using scale command in R Language. The approximate apparent mass (Mapp) of all proteins identified in our dataset was calculated similar as the previous publication [26]. After that, the ratio of Mapp to the predicted monomeric mass (Mmono) was calculated. The ratio (Mapp/Mmono) which we called Rapp, accurately reflects the oligomerization state of proteins if their shape is globular. We use a cutoff value of Rapp≥2 as a useful metric to predict whether or not a protein is oligomeric. The protein with a value of Rapp≤0.5 means it may degrade during the protein extraction process and would be discarded in subsequent analysis. The protein with a value of 0.5<Rapp<2 means exit as monomeric state.

**Machine learning**

The protein elution profiles from all five experiments were used to train a random forest predictor implemented with the scikit-learn python package. We used EPIC (Elution Profile-based Inference of Complexes), a software toolkit for automated scoring of CF/MS (chromatographic fractionation of biological extracts coupled to precision mass spectrometry analysis) data for large-scale determination of high–confidence physical interaction networks and macromolecular assemblies from diverse biological specimens. This software can be obtained from https://github.com/BaderLab/EPIC. Protein pairs were scored according to five features, namely: MI, Bayes Correlation, Jaccard, Pearson Correlation Coefficient, and Apex Score. We manually collected a data set of "gold standard" protein complexes by the reference database (UniProt) for machine learning analysis, which contains 48 conserved true positive protein complexes. Protein-protein interactions (PPIs) were defined as positive if they are observed in the same protein complex. If proteins exist in the protein complex dataset but never show up in the same protein complex, then these two proteins are defined as negative PPIs. The machine learning classifier is trained on the sets of positive and negative PPIs as we defined. We create the train set by our own databases, in which only the protein pairs have at least one elution profile similarity score larger than 0.5 (among all co–fractionation experiments and all correlation metrics) will be retained. An interaction prediction score cut-off of 0.5 was used as this threshold selected protein pairs. All the proteins that used for machine learning prediction are detected with >=2 peptide spectrum matches in at least one of the experiments.

**Construction of plasmid**

The single mutants of the three *sll0445–sll0447* gene clusters and *sll1334* were generated by inserting a chloramphenicol-resistance cassette ($Cm^R$) into their ORFs. The *sll0445* and *slr0149* were cloned with C–terminal GFP-tagged forAP-MS. The DNA fragment containing the full–length targeted gene and its flanking sequences was generated by PCR with *Synechocystis* chromosome DNA as template, cloned into pMD18–T vector (Catalog No: D101A, Takara, Japan), and confirmed by sequencing.

The physical map of the *Synechocystis* genome region containing *sll0445–sll0447* or *sll1334*, indicating the insertion position of $Cm^R$ into target gene respectively, is shown in Figure S10B. The plasmid construction and primers are as follow:

Primer for Mutant Construction

M_sll0445-1 GTTCAGCGGTGATGAGTCG

M_sll0445-2 GTAAATCAAACAGGGCATG

M_sll0446-1 TGTGGCCTATACAATGTCCCAG

M_sll0446-2 AAGATATTTCTTCCAGCAAATGG

M_sll0447-1 ATCTCGTATTAAGAAAGCTTG

M_sll0447-2 TGAGCATAAACTGGACTAATG

M_sll1334-1 AGACGGTTAGAACCAACAGTCACTG

M_sll1334-2 ACAATTTGTAAGCCCTGGCGAACG

Primer for q-PCR

accA-1 AAATGTTTCGGTTAGATGTCC

accA-2 CCAAAGAATAGCCGCACA

accC-1 TTTGGTGGATGGTAACGG

accC-2 TGGCGGAAGCGGAGTTTT

rnpB-1 ACCGCTTGAGGAATTTGGTA

rnpB-2 TTAGTCGTAAGCCGGGTTCT

**Cell motility assays and measurement of photosynthetic activity**

The phototactic movement was examined as described by Wilde et al., with some modifications [68]. Strains streaked on solid BG11 medium plates, containing 0.8% (w/v) agar, 8 mM TES (pH 8.0), 5 mM glucose and 0.3% (w/v) sodium thiosulfate, were placed into opaque boxes with an open slit (3cm width) at one side. Strains were cultivated at 30°C with light intensity at 1–5 μmol photons $m^{-2}$ $s^{-1}$ and the movements were recorded at day 6. The modification of PSII photochemistry in *Synechocystis* was evaluated by JIP curve, measured by Plant Efficiency Analyzer (Hansatech, Germany).

**RNA isolation and transcriptional analysis**

About 50ml of *Synechocystis* grown in BG11 was collected by centrifugation and quickly frozen in liquid nitrogen. Total RNA was extracted with the TRIzol Reagent (Catalog No: 15596–026, Invitrogen, Waltham, MA, USA) according to the manufacturer's instructions. The Perfect Real-Time kit (Catalog No: RR047A, Takara, Japan) was used for eliminating DNA contamination and reverse transcription of cDNA. The newly synthesized cDNAs were used as a template for real-time PCR amplification. RNase P subunit B (rnpB) was used as an internal control for normalizing target gene transcriptional abundance. Primers used in these experiments are shown before.

## Data availability

All LC/MS/MS raw data related to this work will be uploaded to iProX (www.iprox.org) and are available with ID IPX0001620001. All other data was uploaded as supplementary material.

## Authors' contributions

CX performed the experiments with the association of LXY and YXW, analyzed the data with the association of BW and LZH, and wrote the manuscript. CX and LY drafted figures. SLC performed Sucrose density gradient centrifugation assays. AE participated in the discussion and guide LZH for data analysis. CHW designed and coordinated the study and contributed to draft the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

## Acknowledgments

We thank Prof. Baosheng Qiu, Prof. Haibo Jiang, and Dr. Guozheng Dai (Central China Normal University) for providing algal, cloning vectors and technical guidance. We would like to particularly thank Prof. Feng Ge (Institute of Hydrobiology, Chinese Academy of Sciences) for the valuable discussion. This work was funded by the Young Thousand Talents Program. We also thank other members of Wan lab for all their support to this project.

## References

[1] Falcon LI, Magallon S, Castillo A. Dating the cyanobacterial ancestor of the chloroplast. ISME J 2010;4:777–83.

[2] Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, et al. Evolutionary analysis of, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA 2002;99:12246–51.

[3] Schirrmeister BE, de Vos JM, Antonelli A, Bagheri HC. Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. Proc Natl Acad Sci USA 2013;110:1791–6.

[4] Xiong Q, Chen Z, Ge F. Proteomic analysis of post translational modifications in cyanobacteria. J Proteomics 2016;134:57–64.

[5] Hohmann–Marriott MF, Blankenship RE. Evolution of photosynthesis. Annu Rev Plant Biol 2011;62:515–48.

[6] Kaneko T, Stato S, Kotani H, Kaneko A, Asamizu E, Nakamura Y, et al. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp.* strain PCC 6803. II. Sequence determination of the entire genome and assignment of potential protein–coding regions. DNA research 1996;3:109–36.

[7] Gao L, Wang JL, Ge HT, Fang LF, Zhang YY, Huang XH, et al. Toward the complete proteome of *Synechocystis sp.* PCC 6803. Photosynth Res 2015;126:203–19.

[8] Castielli O, De la Cerda B, Navarro JA, Hervas M, De la Rosa MA. Proteomic analyses of the response of cyanobacteria to different stress conditions. FEBS Lett 2009;583:1753–8.

[9] Typas A, Sourjik V. Bacterial protein networks: properties and functions. Nat Rev Microbiol 2015;13:559–72.

[10] Lv Q, Ma WM, Liu H, Li J, Wang H, Wang H, et al. Genome–wide protein–protein interactions and protein function exploration in cyanobacteria. Sci Rep 2015;5:1–12.

[11] Kim WY, Kang S, Kim BC, Oh J, Cho S, Bhak J, et al. SynechoNET: integrated

protein–protein interaction database of a model cyanobacterium Synechocystis sp. PCC 6803. BMC Bioinformatics 2008;9 Suppl 1:S20.

[12] Michaut M, Kerrien S, Montecchi–Palazzi L, Chauvat F, Cassier–Chauvat C, Aude J–C, et al. InteroPORC: automated inference of highly conserved protein interaction networks. Bioinformatics 2008;24:1625–31.

[13] Sato S, Shimoda Y, Muraki A, Kohara M, Nakamura Y, Tabata S. A large–scale protein protein interaction analysis in *Synechocystis sp.* PCC 6803. DNA Research 2007;14:207–16.

[14] Heinz S, Liauw P, Nickelsen J, Nowaczyk M. Analysis of photosystem II biogenesis in cyanobacteria. Biochim Biophys Acta 2016;1857:274–87.

[15] Nickelsen J, Rengstl B. Photosystem II assembly: from cyanobacteria to plants. Annu Rev Plant Biol 2013;64:609–35.

[16] Schuergers N, Mullineaux CW, Wilde A. Cyanobacteria in motion. Curr Opin Plant Biol 2017;37:109–15.

[17] Maier B, Wong GCL. How bacteria use type IV pili machinery on surfaces. Trends Microbiol 2015;23:775–88.

[18] Babu M, Bundalovic–Torma C, Calmettes C, Phanse S, Zhang Q, Jiang Y, et al. Global landscape of cell envelope protein complexes in Escherichia coli. Nat Biotechnol 2017;36:103–12.

[19] Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell 2015;163:712–23.

[20] Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, et al. Panorama of ancient metazoan macromolecular complexes. Nature 2015;525:339–44.

[21] Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome–scale map of the human interactome network. Cell 2014;159:1212–26.

[22] Larance M, Lamond AI. Multidimensional proteomics for cell biology. Nat Rev Mol Cell Biol 2015;16:269–80.

[23] Crozier TWM, Tinti M, Larance M, Lamond AI, Ferguson MAJ. Prediction of

protein complexes in *Trypanosoma brucei* by protein correlation profiling mass spectrometry and machine learning. Mol Cell Proteomics 2017;16:2254–67.

[24] Kirkwood KJ, Ahmad Y, Larance M, Lamond AI. Characterization of native protein complexes and protein isoform variation using size–fractionation–based quantitative proteomics. Mol Cell Proteomics 2013;12:3851–73.

[25] Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, et al. A census of human soluble protein complexes. Cell 2012;150:1068–81.

[26] Aryal UK, McBride Z, Chen D, Xie J, Szymanski DB. Analysis of protein complexes in *Arabidopsis* leaves using size exclusion chromatography and label–free protein correlation profiling. J Proteomics 2017;166:8–18.

[27] Aryal UK, Xiong Y, McBride Z, Kihara D, Xie J, Hall MC, et al. A proteomic strategy for global analysis of plant protein complexes. Plant Cell 2014;26:3867–82.

[28] Hu LZ, Goebels F, Tan JH, Wolf E, Kuzmanov U, Wan C, et al. EPIC: software toolkit for elution profile–based inference of protein complexes. Nat Methods 2019;16:737–42.

[29] Ouma WZ, Pogacar K, Grotewold E. Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties. PLoS Comput Biol 2018;14:1–17.

[30] Calderone A, Castagnoli L, Cesareni G. mentha: a resource for browsing integrated protein–interaction networks. Nat Methods 2013;10:690–1.

[31] Feder ME, Hofmann GE. Heat–shock proteins, molecular chaperones, and the stress, and the stress response: evolutionary and ecological physiology. Ann Rev Physiol 1999;61:243–82.

[32] Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101–13.

[33] Linhartova M, Bucinska L, Halada P, Jecmen T, Setlik J, Komenda J, et al. Accumulation of the type IV prepilin triggers degradation of SecY and YidC and inhibits synthesis of Photosystem II proteins in the cyanobacterium *Synechocystis* PCC 6803. Mol Microbiol 2014;93:1207–23.

[34] Chidgey JW, Linhartova M, Komenda J, Jackson PJ, Dickman MJ, Canniffe DP, et al. A cyanobacterial chlorophyll synthase–HliD complex associates with the Ycf39 protein and the YidC/Alb3 insertase. Plant Cell 2014;26:1267–79.

[35] Gao LY, Ge HT, Huang XH, Liu KH, Zhang YY, Xu W, et al. Systematically ranking the tightness of membrane association for peripheral membrane proteins (PMPs). Mol Cell Proteomics 2015;14:340–53.

[36] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein–protein interaction networks. Nat Methods 2012;9:471–2.

[37] Behler J, Sharma K, Reimann V, Wilde A, Urlaub H, Hess WR. The host–encoded RNase E endonuclease as the crRNA maturation enzyme in a CRISPR–Cas subtype III–Bv system. Nat Microbiol 2018;3:367–77.

[38] Scholz I, Lange SJ, Hein S, Hess WR, Backofen R. CRISPR–Cas systems in the cyanobacterium Synechocystis sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. PLoS One 2013;8:e56470.

[39] Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, et al. RNA–guided RNA cleavage by a CRISPR RNA–Cas protein complex. Cell 2009;139:945–56.

[40] Wegener KM, Welsh EA, Thornton LE, Keren N, Jacobs JM, Hixson KK, et al. High sensitivity proteomics assisted discovery of a novel operon involved in the assembly of photosystem II, a membrane protein complex. J Biol Chem 2008;283:27829–37.

[41] Luck K, Sheynkman GM, Zhang I, Vidal M. Proteome–scale human interactomics. Trends Biochem Sci 2017;42:342–54.

[42] Yoshimura H, Yanagisawa S, Kanehisa M, Ohmori M. Screening for the target gene of cyanobacterial cAMP receptor protein SYCRP1. Mol Microbiol 2002;43:843–53.

[43] Bhaya D, Takahashi A, Shahi P, Grossman AR. Novel motility mutants of *synechocystis* strain PCC 6803 generated by in vitro transposon mutagenesis. J Bacteriol 2001;183:6140–3.

[44] Olsen GJ, Woese CR, Overbeek R. The winds of (evolutionary) change: breathing

new life into microbiology. J Bacteriol 1994;176:1–6.

[45] Woese C. Bacterial evolution. Microbiol Rev 1987;51:221–71.

[46] Gao F, Zhao J, Wang X, Qin S, Wei L, Ma W. NdhV is a subunit of NADPH dehydrogenase essential for cyclic electron transport in *Synechocystis sp.* strain PCC 6803. Plant Physiol 2016;170:752–60.

[47] Gao F, Zhao J, Chen L, Battchikova N, Ran Z, Aro EM, et al. The NDH–1L–PSI supercomplex is important for efficient cyclic electron transport in cyanobacteria. Plant Physiol 2016;172:1451–64.

[48] Matsushita K, Ohnishi T, Kaback H. NADH–ubiquinone oxidoreductases of the *Escherichia coli* aerobic respiratory chain. Biochemistry 1987;26:7732–7.

[49] Peschek GA, Obinger C, Paumann M. The respiratory chain of blue–green algae (cyanobacteria). Physiol Plant 2004;120:358–69.

[50] Ingledew W, Poole R. The respiratory chains of *Escherichia coli*. Microbiol Rev 1984;48:222–71.

[51] Melo AM, Bandeiras TM, Teixeira M. New insights into type II NAD(P)H:quinone oxidoreductases. Microbiol Mol Biol Rev 2004;68:603–16.

[52] Apel K, Hirt H. Reactive oxygen species: metabolism, oxidative stress, and signal transduction. Annu Rev Plant Biol 2004;55:373–99.

[53] Mittler R. Oxidative stress, antioxidants and stress tolerance. Trends Plant Sci 2002;7:405–10.

[54] Shcolnick S, Keren N. Metal homeostasis in cyanobacteria and chloroplasts. Balancing benefits and risks to the photosynthetic apparatus. Plant Physiol 2006;141:805–10.

[55] Morel FMM, Price NM. The biogeochemical cycles of trace metals in the oceans. science 2003;300:944–7.

[56] Nogales E, Wolf SG, Downing KH. Structure of the αβ tubulin dimer by electron crystallography. Nature 1998;391:199–203.

[57] Ent FVD, Löwe J. Crystal structure of the cell division protein FtsA from *Thermotoga maritima*. EMBO J. 2000;19:5300–7.

[58] Noirclerc–Savoye M, Le Gouellec A, Morlot C, Dideberg O, Vernet T, Zapun A. In vitro reconstitution of a trimeric complex of DivIB, DivIC and FtsL, and their transient co–localization at the division site in Streptococcus pneumoniae. Mol Microbiol 2005;55:413–24.

[59] Yoshihara S, Ikeuchi M. Phototactic motility in the unicellular cyanobacterium *Synechocystis sp.* PCC 6803. Photoch Photobio Sci 2004;3:512–8.

[60] Qiu B, Zhang A, Zhou W, Wei J, Dong H, Liu Z. Effects of potassium on the photosynthetic recovery of the terrestrial cyanobacterium, *Nostoc Flagelliforme* (Cyanophyceae) during rehydration. J Phycol 2004;40:323–32.

[61] Wiltbank LB, Kehoe DM. Diverse light responses of cyanobacteria mediated by phytochrome superfamily photoreceptors. Nat Rev Microbiol 2019;17:37–50.

[62] Huerlimann R, Heimann K. Comprehensive guide to acetyl–carboxylases in algae. Crit Rev Biotechnol 2013;33:49–65.

[63] Okada K, Horii E, Nagashima Y, Mitsui M, Matsuura H, Fujiwara S, et al. Genes for a series of proteins that are involved in glucose catabolism are upregulated by the Hik8–cascade in *Synechocystis sp.* PCC 6803. Planta 2015;241:1453–62.

[64] Tabei Y, Okada K, Makita N, Tsuzuki M. Light-induced gene expression of fructose 1,6–bisphosphate aldolase during heterotrophic growth in a cyanobacterium, *Synechocystis sp.* PCC 6803. FEBS J 2009;276:187–98.

[65] Tang K, Ding WL, Höppner A, Zhao C, Zhang L, Hontani Y, et al. The terminal phycobilisome emitter, LCM: A light–harvesting pigment with a phytochrome chromophore. Proc Natl Acad Sci USA 2015;112:15880–5.

[66] Sachelaru I, Petriman NA, Kudva R, Kuhn P, Welte T, Knapp B, et al. YidC occupies the lateral gate of the SecYEG translocon and is sequentially displaced by a nascent membrane protein. J Biol Chem 2013;288:16295–307.

[67] Anbudurai P, Mor T, Ohad I, Shestakov S, Pakrasi H. The ctpA gene encodes the C–terminal processing protease for the D1 protein of the photosystem II reaction center complex. Proc Natl Acad Sci USA 1994;91:8082–6.

[68] Wilde A, Fiedler B, Börner T. The cyanobacterial phytochrome Cph2 inhibits

phototaxis towards blue light. Mol Microbiol 2002;44:981–8.

## Figure legends

**Figure 1 Workflow for the identification of native protein complexes and data processing**

**A.** Schematic diagram of co-fractionation LC/MS/MS and machine learning. Lysates were produced containing a mixture of protein complexes, which were separated by SEC, PolyCATWAX, or Suc-DGC. Proteins in each fraction were digested with trypsin and analyzed by nano–LC-MS/MS. Putative protein-protein interactions were predicted by machine learning using EPIC toolkits. **B.** The calculation of Rapp, which can accurately reflect the oligomerization states of proteins. Mapp, apparent molecular mass; Mmono, predicted molecular mass of monomer; Rapp, the ratio of the Mapp to the Mmono. A protein with Rapp>=2 means that it has interaction with other proteins. **C.** Heat map of the Pearson correlation coefficients of the protein quantification signals in two SEC biological replicates. **D.** Receiver operating characteristic curve of machine learning.

**Figure 2 The characteristic of predicted PPIs**

**A.** Pie chart showing the overlaps of predicted co-complex PPIs with PPIs from IntAct and STRING databases. **B.** Gene Ontology analysis of proteins with different degrees, which is the number of edges that one protein links to other proteins in the network. ($P<0.05$) **C.** The degree distributions of conserved proteins and unique proteins in *Synechocystis.* The conserved protein means that it has homologous with protein in *E.coli, A.thaliana, H.sapiens,* and *S.cerevisiae.* **D.** The proteins that interact with photosynthesis were classified by their functions. Each section has three or more proteins aggregation with the same GO term. Different node colors represent the subcellular localization of the protein.

**Figure 3 Predicted protein complexes from EPIC**

The top part shows the global landscape of inferred 291 *Synechocystis* protein complexes. The protein nodes are colored according to their subcellular locations. The bottom part presents some of the known protein complex clusters, such as photosystem, ATP synthase, Ribosome, which are annotated with their name or abbreviation. Yellow

lines between proteins indicate interactions found in public databases, while gray lines demonstrate the interactions are reported only in this work.

**Figure 4 Protein complex evolution analysis**

**A.** Proteins that have homologous with either *E.coli* or *A.thaliana*, homologous with both *E.coli* and *A.thaliana*, or unique in *Synechocystis* were grouped according to their representative biological processes. **B.** Example of NAD(P)H–quinone oxidoreductase and bacterioferritin protein-protein interaction variation across species. Physical interactions in *E.coli* and *A.thaliana* were collected from the Mentha database. InParanoid orthologs of all three species are depicted with the same colors. Proteins that have no homologs in other species are shown in red.

**Figure 5 Sll0445–Sll0447 involved in photosynthesis and cell motility and Sll1334 regulated lipid metabolisms**

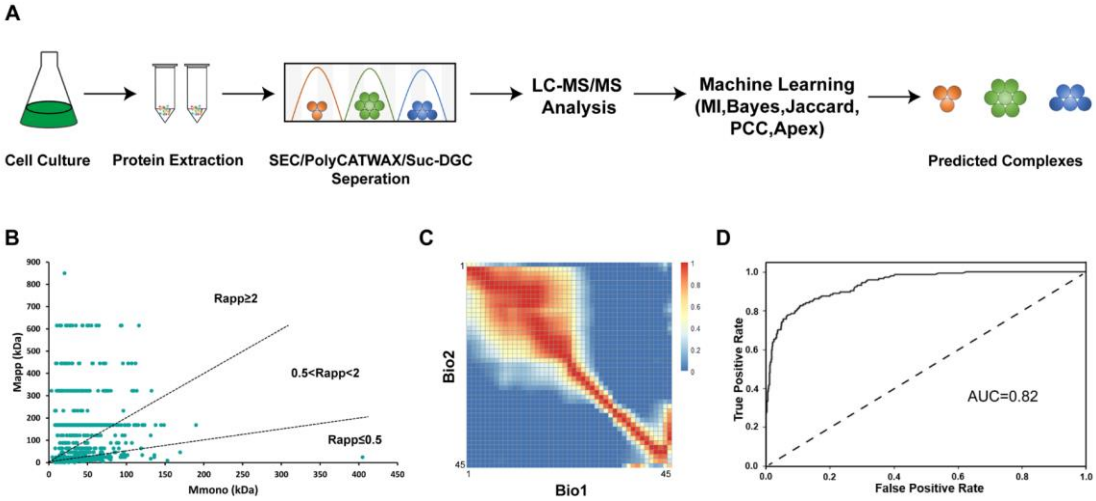
**A.** The motility state of wild type and mutants under unidirectional illumination (white light). The yellow arrow indicates the direction of the light source. **B–C.** OJIP curve under optimal conditions or low–temperature conditions. The J, I and P steps occurred at about 2 ms, 30 ms, and 400 ms, respectively. **D.** The transcript levels of acetyl coenzyme A carboxylase subunits by qRT–PCR in the wild–type strain and △*s111334*. Ct–values for each gene were normalized to rnpB.

**Figure 6 The model of cyanobacterial phototaxis regulation**

The model is based on our data (red color) and public knowledge, showing the phototaxis assembly and how it influenced by the structure of pilus and carbon metabolism. Proteins SecD, SecF, and ChlD present a good elution profiling with PSII assembly proteins, indicating that Sec complexes and ChlD might play a critical function in the maturation of PSII. The hypothetical proteins Sll0445, Sll0446 and Sll0447 are induced by cAMP receptor protein Sycrp1 and interact with protein pilus assembly proteins to regulate cell motilities.These hypothetical proteins can also interact with photosynthetic core proteins, revealing a close relationship between photosynthesis and cell motility in *Synechocystis*. The protein Sll1334 can regulate glucolipid metabolism and may have an important role in controlling the utilization of

sugars and lipids during heterotrophic growth and then effect on photosynthesis. pD1, precursor of D1; C, cytoplasm; L, lumen; OM, outer membrane; PDM, PratA-defined membrane; PM, plasma membrane; PP, periplasm; TM, thylakoid membrane.

**Figure 1**


A
Cell Culture
Protein Extraction
SEC/PolyCATWAX/Suc-DGC Seperation
LC-MS/MS Analysis
Machine Learning (MI,Bayes,Jaccard, PCC,Apex)
Predicted Complexes
B
Rapp≥2
0.5<Rapp<2
Rapp≤0.5
Mapp (kDa)
Mmono (kDa)
C
Bio2
Bio1
D
True Positive Rate
False Positive Rate
AUC=0.82

# Figure 2

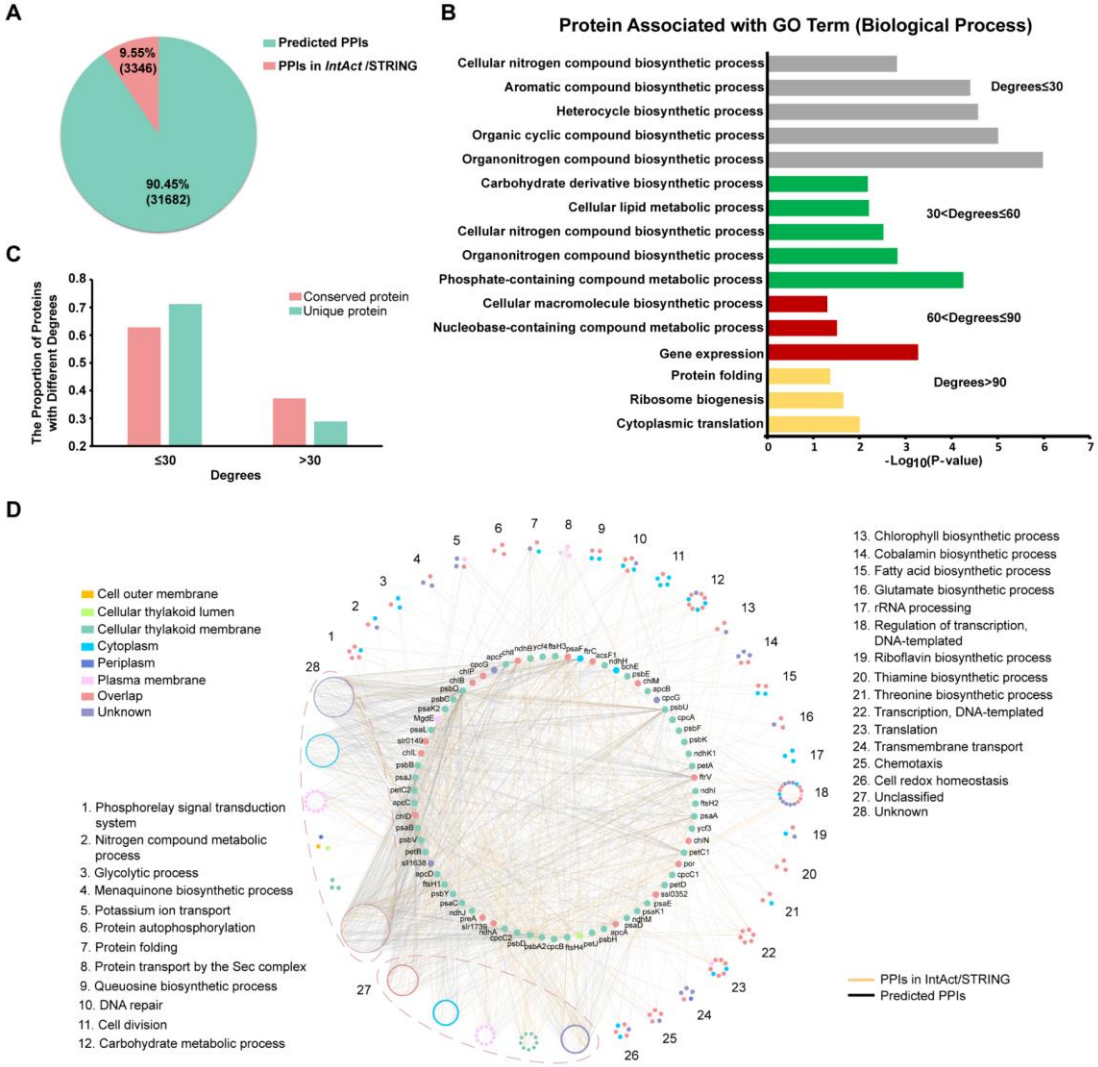
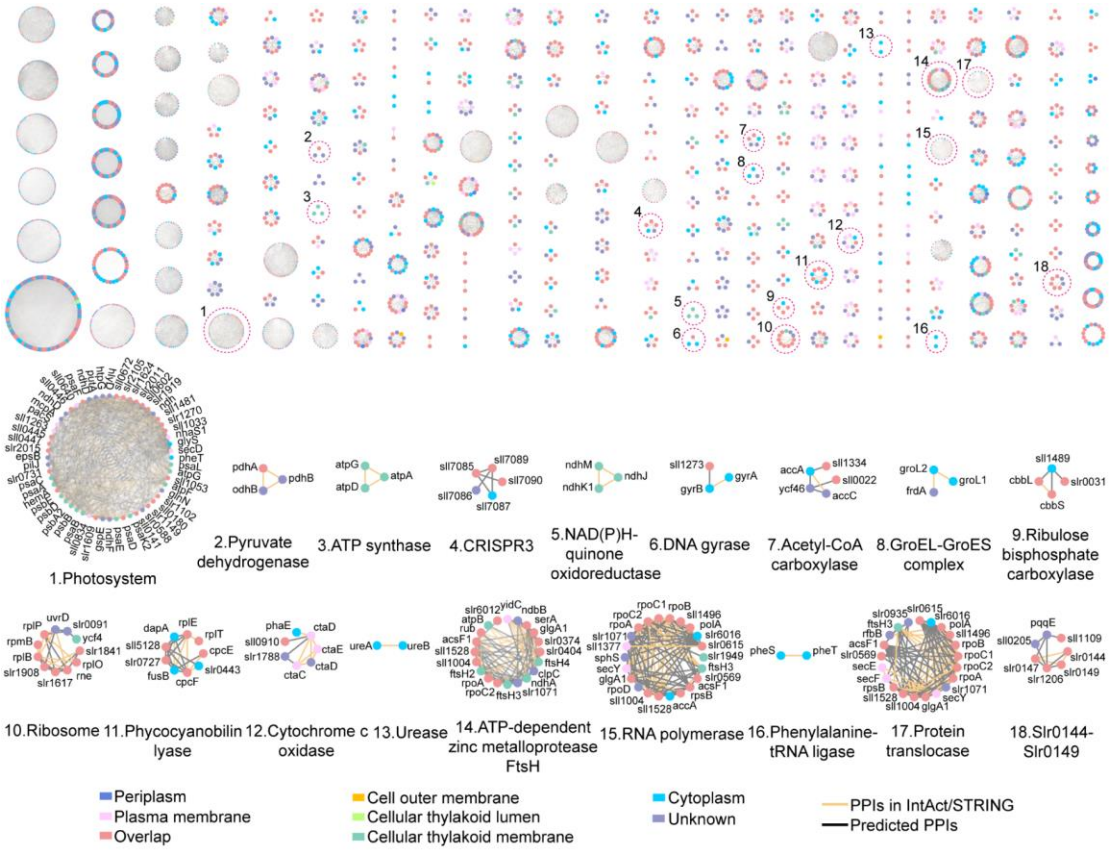

A
9.55%
(3346)
90.45%
(31682)
Predicted PPIs
PPIs in IntAct /STRING
B
Protein Associated with GO Term (Biological Process)
Cellular nitrogen compound biosynthetic process
Aromatic compound biosynthetic process
Heterocycle biosynthetic process
Organic cyclic compound biosynthetic process
Organonitrogen compound biosynthetic process
Carbohydrate derivative biosynthetic process
Cellular lipid metabolic process
Cellular nitrogen compound biosynthetic process
Organonitrogen compound biosynthetic process
Phosphate-containing compound metabolic process
Cellular macromolecule biosynthetic process
Nucleobase-containing compound metabolic process
Gene expression
Protein folding
Ribosome biogenesis
Cytoplasmic translation
Degrees≤30
30<Degrees≤60
60<Degrees≤90
Degrees>90
-Log10(P-value)
C
The Proportion of Proteins with Different Degrees
Conserved protein
Unique protein
≤30
>30
Degrees
D
Cell outer membrane
Cellular thylakoid lumen
Cellular thylakoid membrane
Cytoplasm
Periplasm
Plasma membrane
Overlap
Unknown
1. Phosphorelay signal transduction system
2. Nitrogen compound metabolic process
3. Glycolytic process
4. Menaquinone biosynthetic process
5. Potassium ion transport
6. Protein autophosphorylation
7. Protein folding
8. Protein transport by the Sec complex
9. Queuosine biosynthetic process
10. DNA repair
11. Cell division
12. Carbohydrate metabolic process
13. Chlorophyll biosynthetic process
14. Cobalamin biosynthetic process
15. Fatty acid biosynthetic process
16. Glutamate biosynthetic process
17. rRNA processing
18. Regulation of transcription, DNA-templated
19. Riboflavin biosynthetic process
20. Thiamine biosynthetic process
21. Threonine biosynthetic process
22. Transcription, DNA-templated
23. Translation
24. Transmembrane transport
25. Chemotaxis
26. Cell redox homeostasis
27. Unclassified
28. Unknown
PPIs in IntAct/STRING
Predicted PPIs

**Figure 3**


1.Photosystem
2.Pyruvate dehydrogenase
3.ATP synthase
4.CRISPR3
5.NAD(P)H-quinone oxidoreductase
6.DNA gyrase
7.Acetyl-CoA carboxylase
8.GroEL-GroES complex
9.Ribulose bisphosphate carboxylase
10.Ribosome
11.Phycocyanobilin lyase
12.Cytochrome c oxidase
13.Urease
14.ATP-dependent zinc metalloprotease FtsH
15.RNA polymerase
16.Phenylalanine-tRNA ligase
17.Protein translocase
18.Slr0144-Slr0149
Periplasm
Plasma membrane
Overlap
Cell outer membrane
Cellular thylakoid lumen
Cellular thylakoid membrane
Cytoplasm
Unknown
PPIs in IntAct/STRING
Predicted PPIs

## Figure 4

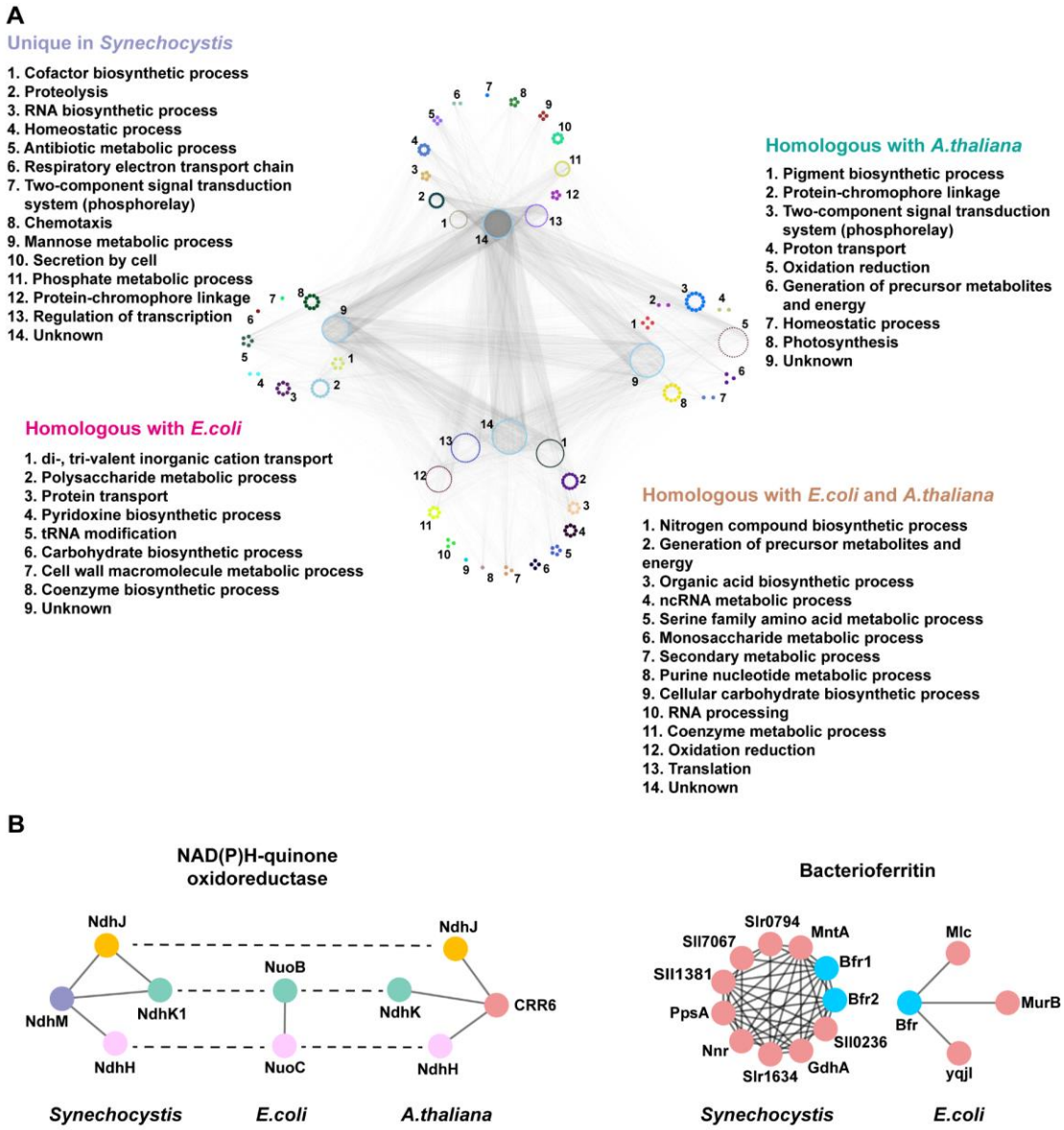

A
Unique in Synechocystis
1. Cofactor biosynthetic process
2. Proteolysis
3. RNA biosynthetic process
4. Homeostatic process
5. Antibiotic metabolic process
6. Respiratory electron transport chain
7. Two-component signal transduction system (phosphorelay)
8. Chemotaxis
9. Mannose metabolic process
10. Secretion by cell
11. Phosphate metabolic process
12. Protein-chromophore linkage
13. Regulation of transcription
14. Unknown
Homologous with A.thaliana
1. Pigment biosynthetic process
2. Protein-chromophore linkage
3. Two-component signal transduction system (phosphorelay)
4. Proton transport
5. Oxidation reduction
6. Generation of precursor metabolites and energy
7. Homeostatic process
8. Photosynthesis
9. Unknown
Homologous with E.coli
1. di-, tri-valent inorganic cation transport
2. Polysaccharide metabolic process
3. Protein transport
4. Pyridoxine biosynthetic process
5. tRNA modification
6. Carbohydrate biosynthetic process
7. Cell wall macromolecule metabolic process
8. Coenzyme biosynthetic process
9. Unknown
Homologous with E.coli and A.thaliana
1. Nitrogen compound biosynthetic process
2. Generation of precursor metabolites and energy
3. Organic acid biosynthetic process
4. ncRNA metabolic process
5. Serine family amino acid metabolic process
6. Monosaccharide metabolic process
7. Secondary metabolic process
8. Purine nucleotide metabolic process
9. Cellular carbohydrate biosynthetic process
10. RNA processing
11. Coenzyme metabolic process
12. Oxidation reduction
13. Translation
14. Unknown
B
NAD(P)H-quinone oxidoreductase
NdhJ
NdhM
NdhK1
NdhH
NuoB
NuoC
NdhJ
NdhK
CRR6
NdhH
Synechocystis
E.coli
A.thaliana
Bacterioferritin
Slr0794
MntA
Sll7067
Bfr1
Sll1381
Bfr2
PpsA
Sll0236
Nnr
GdhA
Slr1634
Mlc
MurB
Bfr
yqjI
Synechocystis
E.coli

**Figure 5**

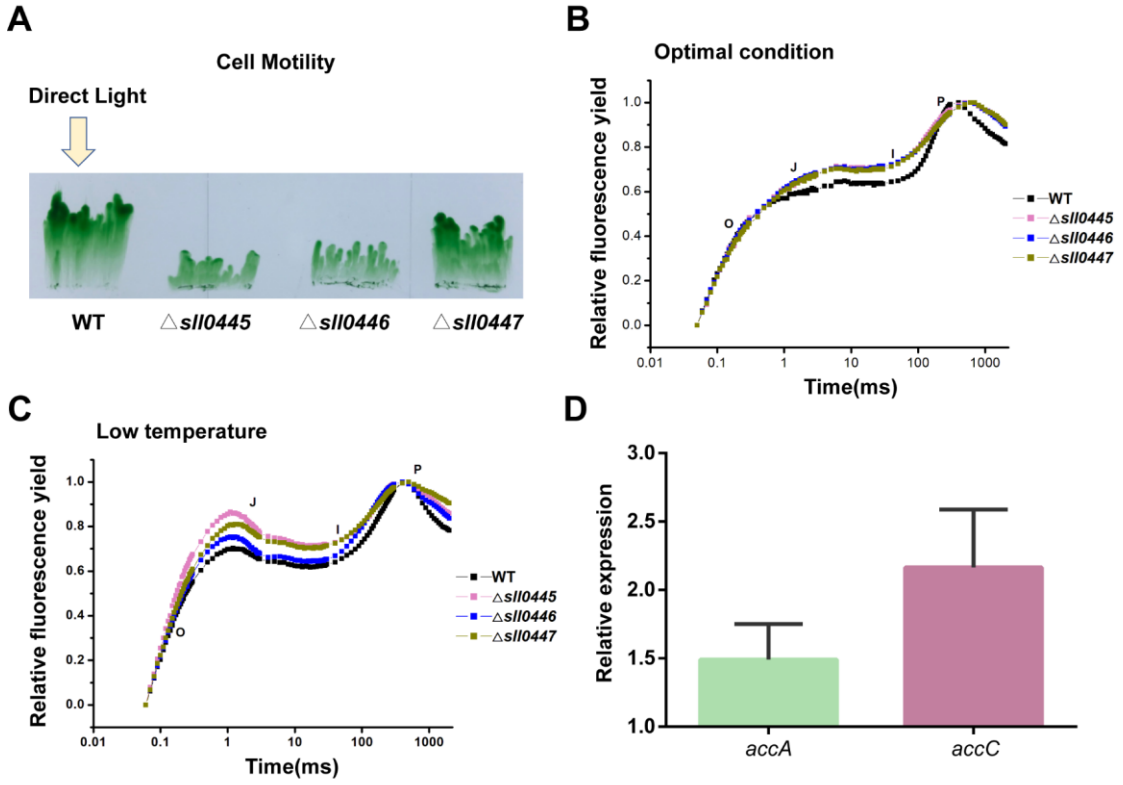
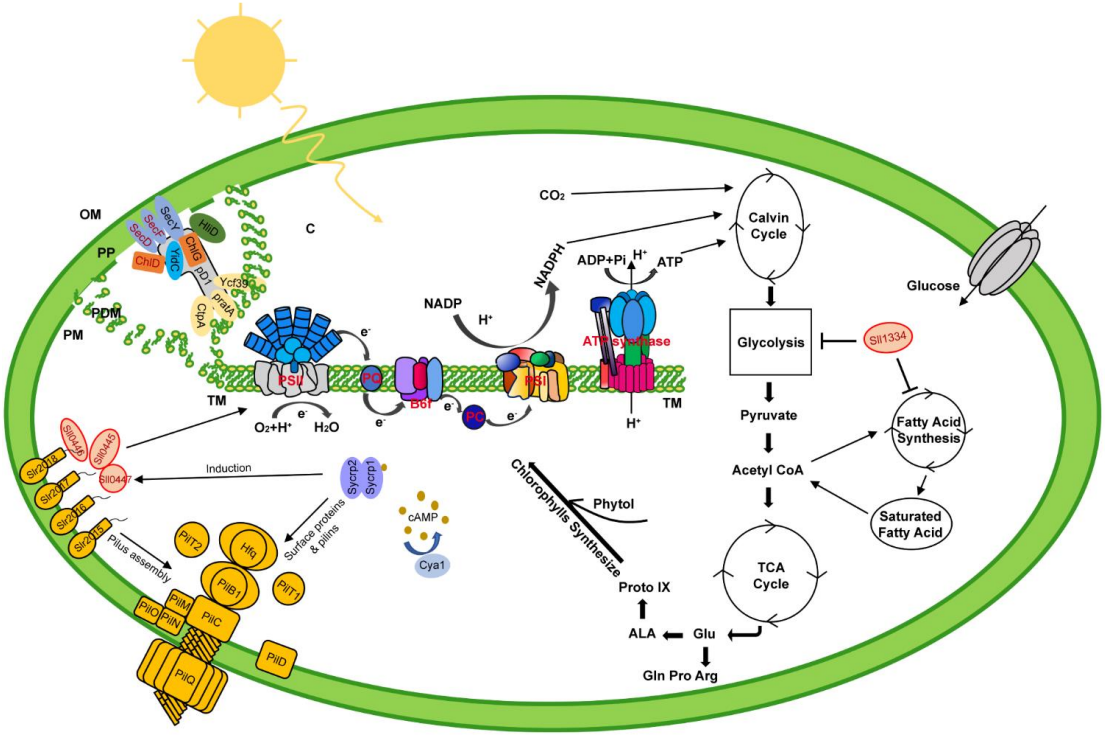

A
Cell Motility
Direct Light
WT
△sll0445
△sll0446
△sll0447
B
Optimal condition
Relative fluorescence yield
Time(ms)
O
J
I
P
WT
△sll0445
△sll0446
△sll0447
C
Low temperature
Relative fluorescence yield
Time(ms)
O
J
I
P
WT
△sll0445
△sll0446
△sll0447
D
Relative expression
accA
accC

**Figure 6**



OM
PP
PDM
PM
C
TM
SecY
SecF
SecD
ChlD
YidC
ChlG
HliD
pD1
Ycf39
pratA
CtpA
PSII
PC
B6f
PSI
PC
e-
O2+H+
H2O
NADP
H+
NADPH
ADP+Pi
H+
ATP
ATP synthase
TM
H+
CO2
Calvin Cycle
Glucose
Glycolysis
Sll1334
Pyruvate
Acetyl CoA
Fatty Acid Synthesis
Saturated Fatty Acid
TCA Cycle
Glu
Gln Pro Arg
ALA
Proto IX
Phytol
Chlorophylls Synthesize
Induction
Sll0446
Sll0445
Sll0447
Slr2018
Slr2017
Slr2016
Slr2015
Pilus assembly
Sycrp2
Sycrp1
Surface proteins & pilins
cAMP
Cya1
PilT2
Hfq
PilB1
PilT1
PilM
PilC
PilO
PilN
PilQ
PilD

## Supplementary material

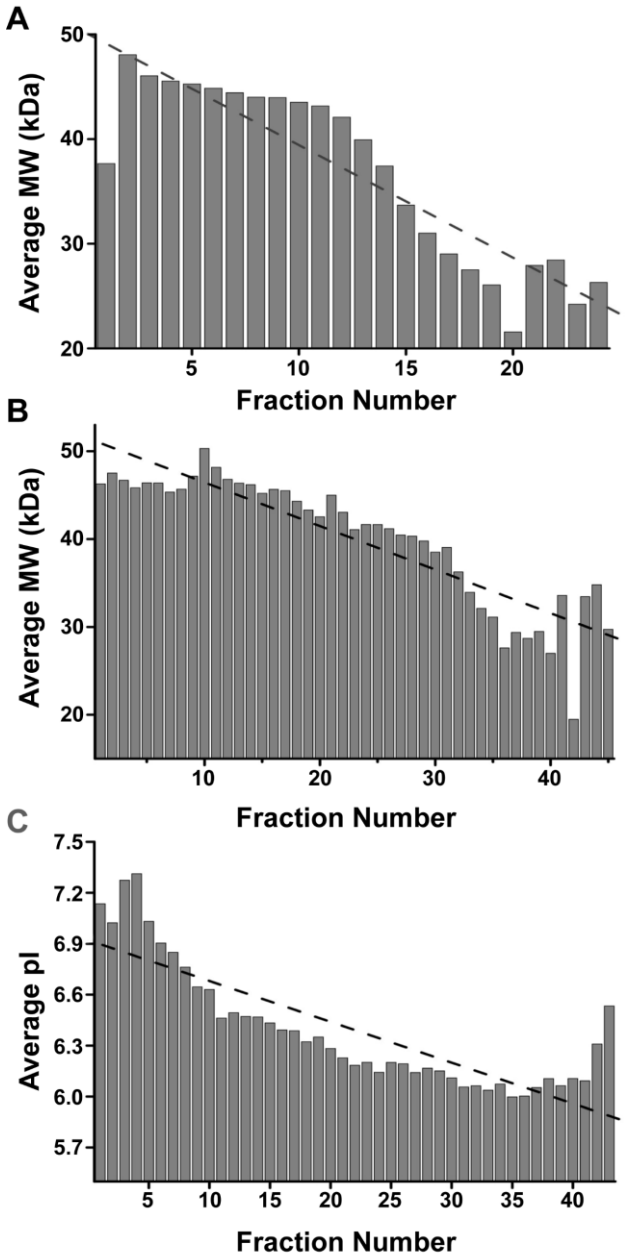



**Figure S1 The distribution of average MW or pI in all fractions on different chromatography columns**

**A.** MAbPac SEC-1 column. **B.** Superose 6 10/300GL column. **C.** PolyCATWAX mixed-bed ion-exchange column. For each fraction, the average MW or pI is average molecular weight or pI of all proteins identified in that fraction. The dash lines are trending lines.

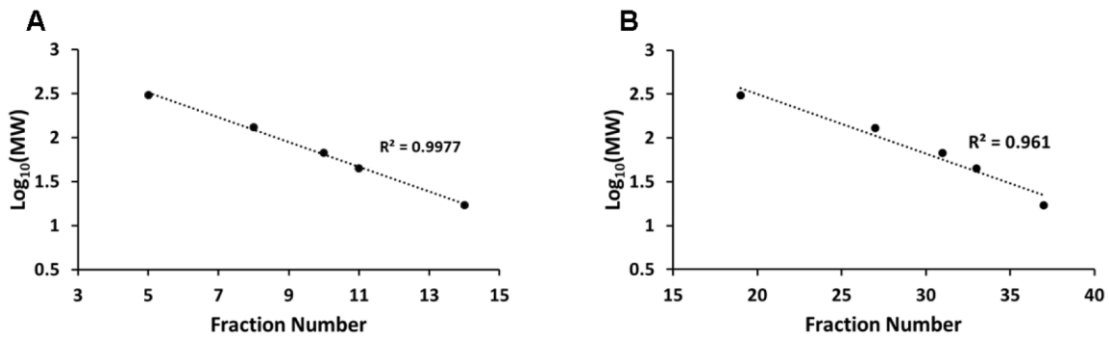



**Figure S2 Regression curve of standard proteins**

Protein standards of known molecular weight (thyroglobulin, BSA, Albumin egg and myoglobin) were separated by the SEC column, and their elution peaks were used to calculate approximate molecular weights of the fractions. **A.** MAbPac SEC–1 column. **B.** Superose 6 10/300GL column.

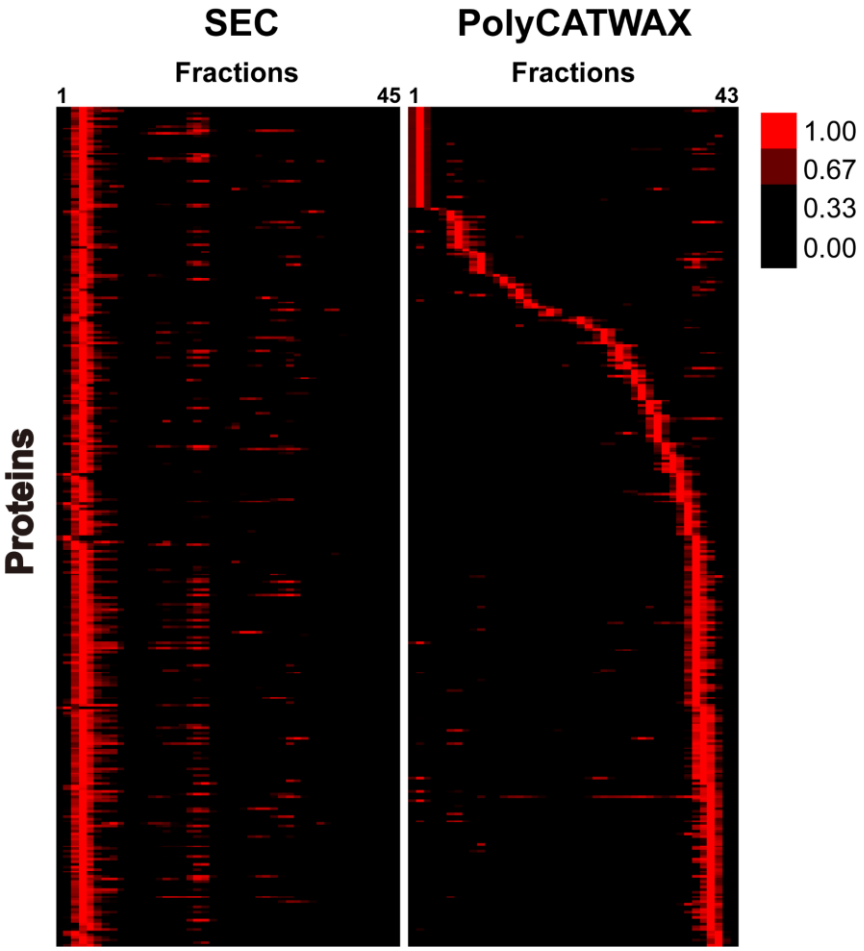



**Figure S3 The complementarity of protein elution profiling in SEC and PolyCATWAX**

The proteins that not separated effectively in SEC (left) are the protein complexes with molecular weight beyond SEC valid separation range and eluted in early fractions. However, some of these protein complexes can be separated on PolyCATWAX according to their elution profiling (right). Red color corresponds to protein abundance.

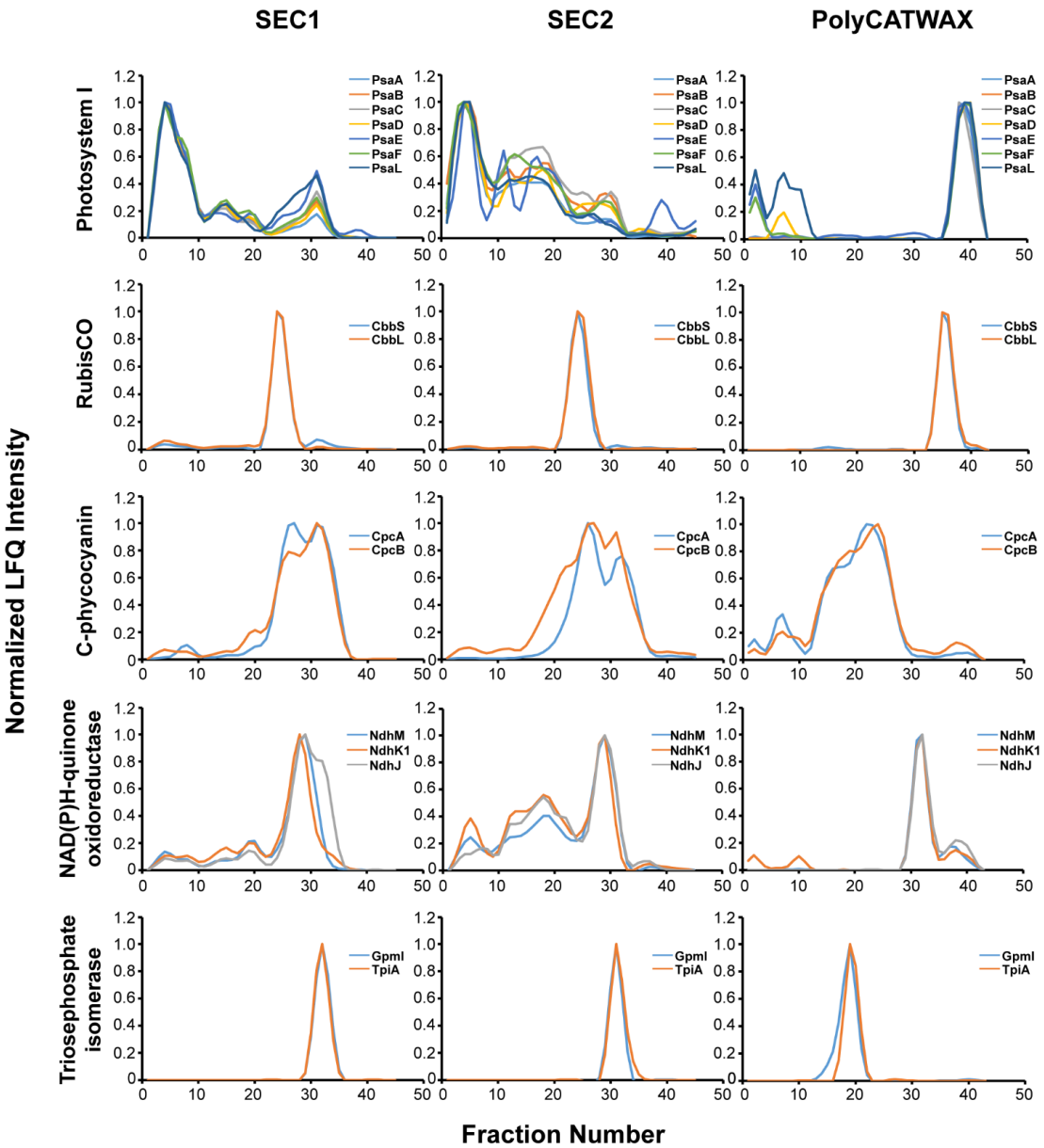



**Figure S4 Elution profiles of the components of known protein complexes**

Elution profiling of Photosystem I, RubisCO, C-phycocyanin, NAD(P)H-quinone oxidoreductase, and Triosephosphate isomerase on three different columns (SEC1: MAbPac SEC, SEC2: Superose 6 10/300GL, PolyCATWAX mixed–bed ion exchange). The elution profiling lines of proteins in one protein complex are shown in different colors. x-axis: elution fraction number, y-axis: normalized label-free quantification intensity.

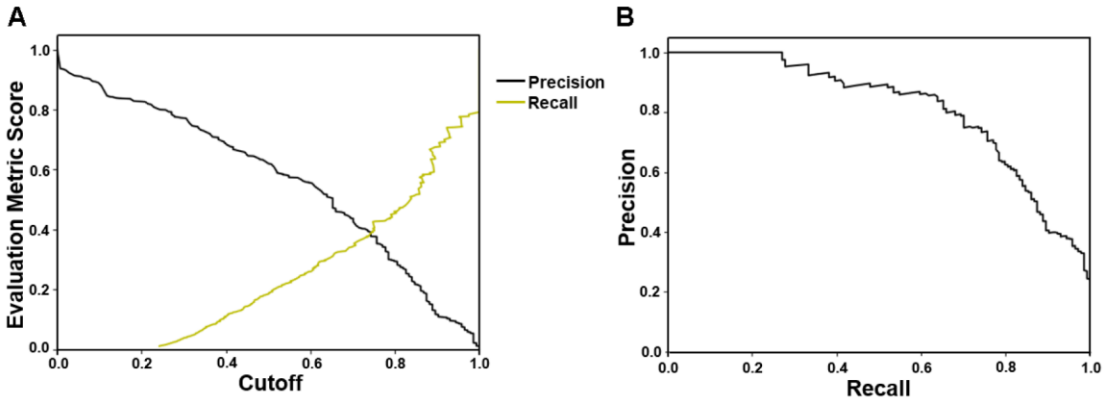



**Figure S5 Evaluation results for machine learning**

**A.** Evaluation metric score–Cutoff from experimental data. **B.** Precision–recall curve (PR) for co-complex PPI prediction from experimental data.

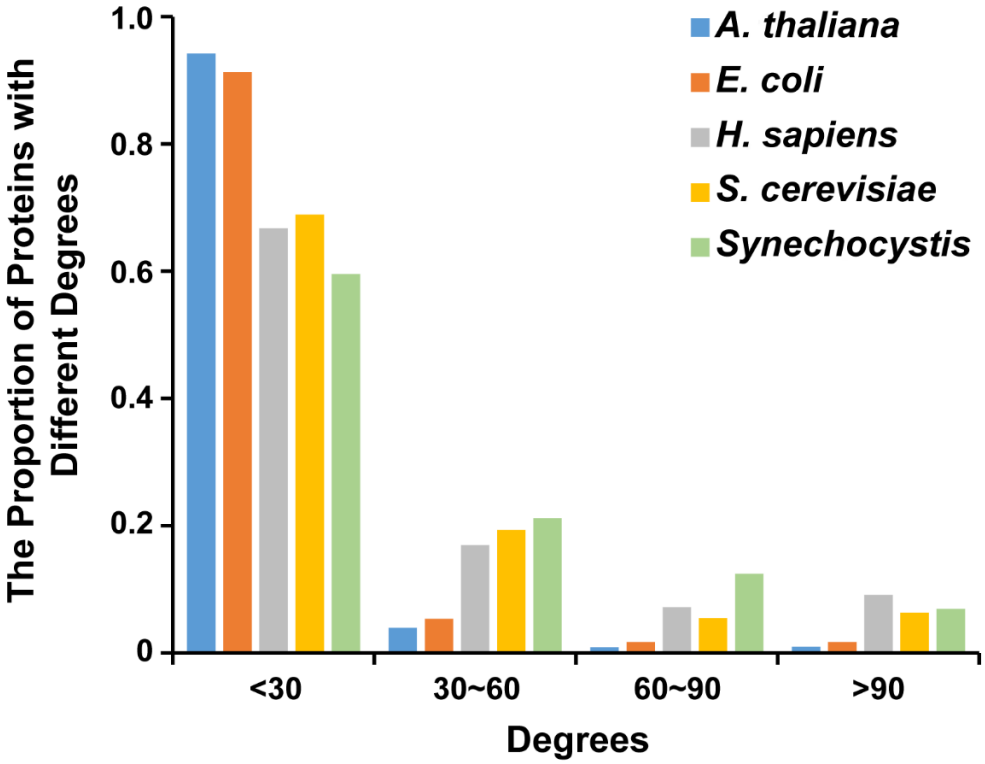



**Figure S6 The distribution of proteins with different degrees in different model organisms**

The protein-protein interaction pairs of *Synechocystis* were generated from our dataset, and PPIs of other model organisms were obtained from the Mentha database. The degree is defined as the number of edges that one protein links to other proteins in the network.

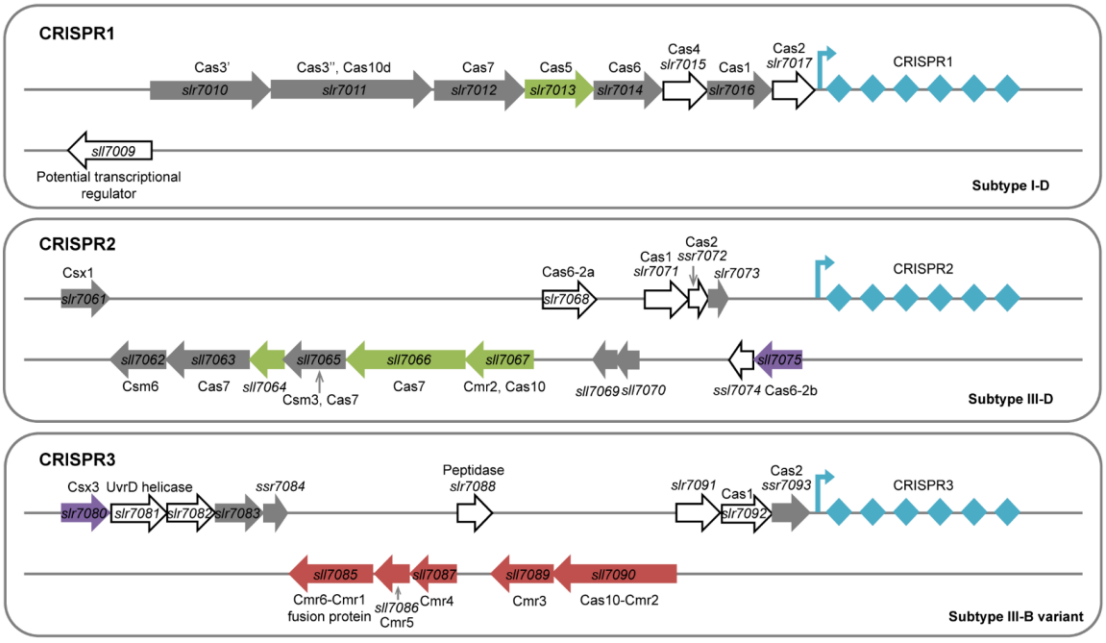
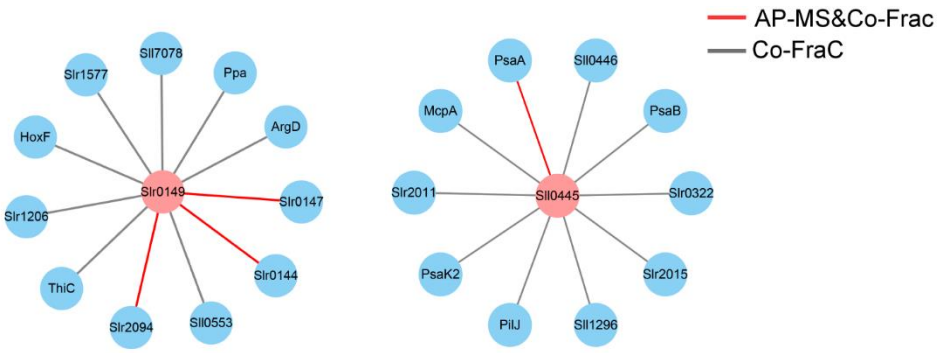



**Figure S7 Organization of the three CRISPR–cas systems in *Synechocystis***

Several cas-genes are located upstream of each of the three CRISPR arrays. Arrows in white represent proteins not identified by MS. Arrows in grey illustrate the proteins, of which no high confident physical interactions were found in our dataset. Other arrows in the same colors represent those proteins that can form complex, such as the CRISPR3 can form one complex (red color), CRISPR2 proteins can have interactions with both CRISPR 1(green) and CRISPR3 (purple).




| | Proteins | PSM Count |
|---|---|---|
| **Bait** | **Slr0149-GFP** | **12** |
| Prey | Slr0144 | 6 |
| | Slr0147 | 6 |
| | Slr2094 | 3 |
| **Bait** | **Sll0445-GFP** | **14** |
| Prey | PsaA | 5 |
| | PsaF | 8 |
| | Slr2018 | 8 |

**Figure S8 The APMS result of Slr0149**

Schematic illustration of the protein-protein interactions involved with protein Slr0149 from Co–fractionation and APMS data. And the mass spectrometry PSM (Peptide–Spectrum Match) number of the proteins co-purified with Slr0149 from GFP–tagged APMS (full data see table S6).

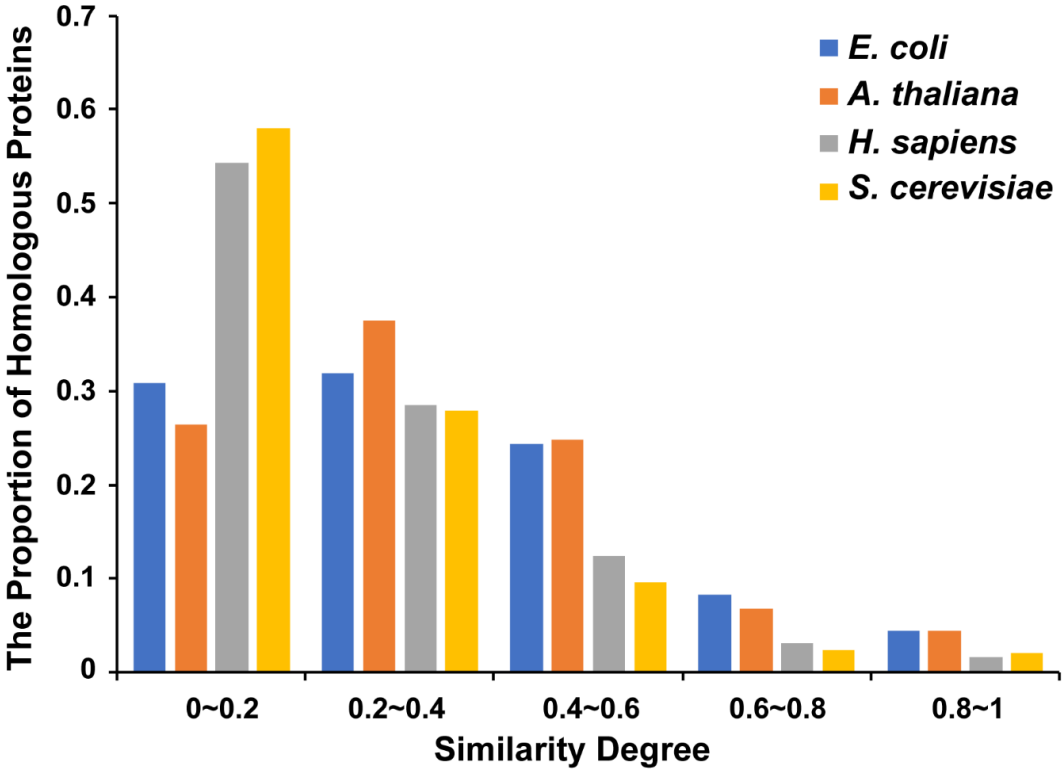



**Figure S9 Conservative analysis of predicted complexes**

The proportion of different similarity degrees of all *Synechocystis* protein complexes in this work was plotted. The similarity degree was calculated by the percentage of protein components in each complex that has homologous in other organisms.

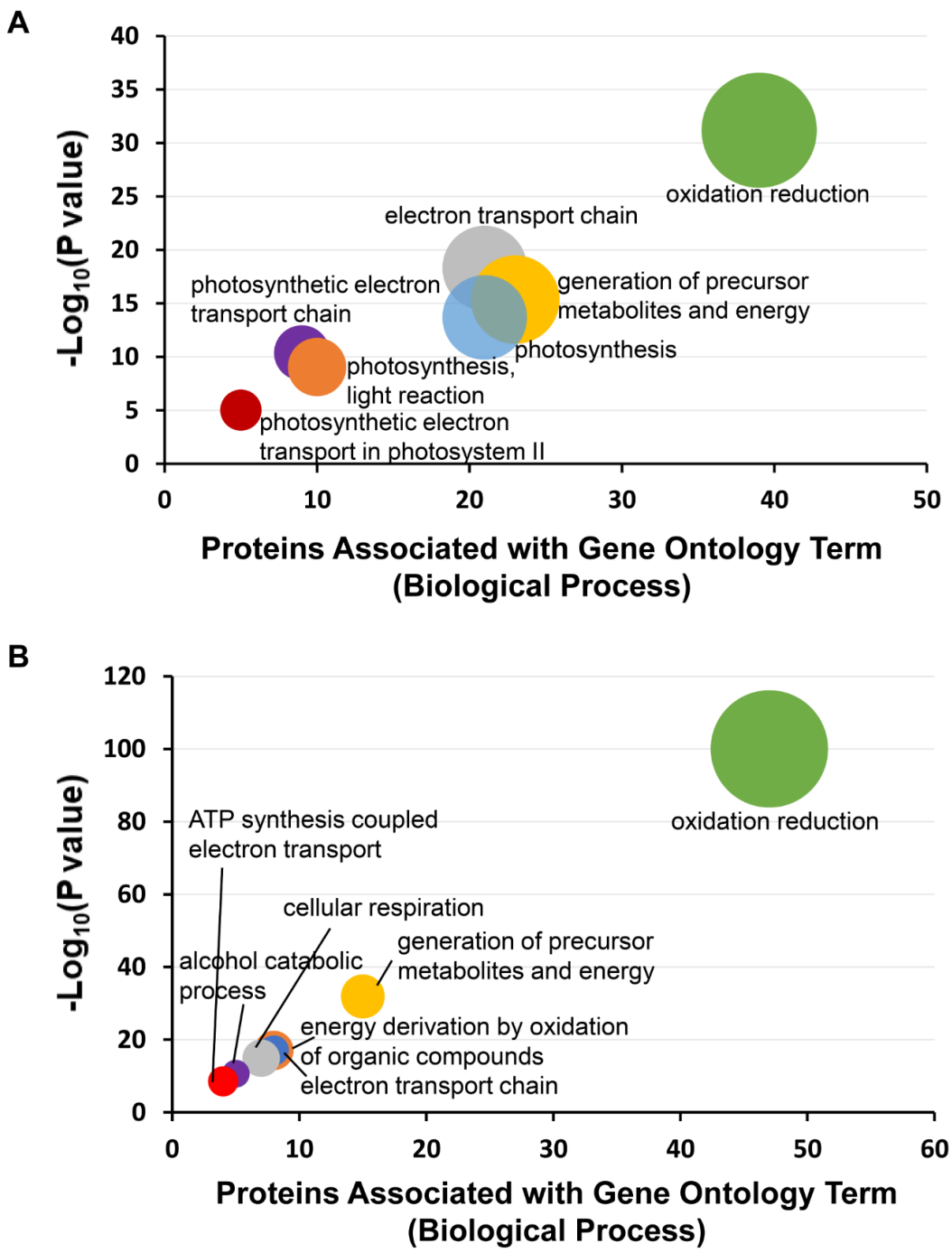



**Figure S10 Comparative analysis of proteins annotated with oxidation-reduction**

Bubble graphs demonstrate the gene ontology terms (biological process) (x-axis) plotted against the $-\log_{10}$(P-value) for oxidation–reduction related protein in homologous with the *A.thaliana* part (**A**) and homologous with *A.thaliana* and *E.coli* part (**B**), respectively.

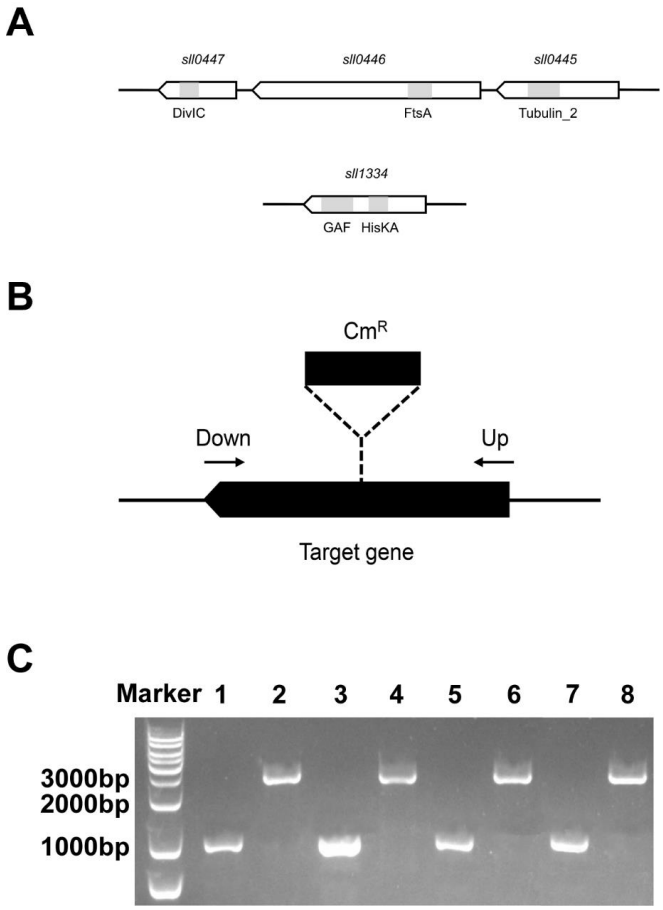



**Figure S11 Construction and detection of the *Synechocystis sll0445–sll0447* gene clusters and s*ll1334***

**A.** The predicted domains of Sll0445–Sll0447 proteins (Top) and Sll1334 (Bottom). The *sll0445* encodes a protein with a Tubulin_2 domain, the *sll0446* encodes a protein with a FtsA domain, and the *sll0447* encodes a protein with a DivIC domain. The *sll1334* encodes a protein with GAF and HisKA domains. **B.** Physical map of the *sll0445–sll0447* genes, with indications of the insertion position of inserting a chloramphenicol-resistance cassette ($Cm^R$) into their ORFs. **C.** Detection of the segregation degree of the *sll0445–sll0447* mutants by PCR amplification. Lanes 1–2 use primers sll0445up and sll0445down. Lanes 3–4 use primers sll0446up and sll0446down. Lanes 5–6 use primers sll0447up and sll0447down. Lanes 7–8 use primers sll1334up and sll1334down. Lanes 1, 3, 5, and 7: wild-type strain DNA; lane 2: sll0445::$Cm^R$ mutant DNA; lane 4: sll0446::$Cm^R$ mutant DNA; lane 6: sll0447::$Cm^R$ mutant DNA; lane 8: sll1334::$Cm^R$ mutant DNA.