Yao *et al. Microbial Cell Factories* (2023) 22:13
https://doi.org/10.1186/s12934-022-02004-y

Microbial Cell Factories

RESEARCH

Open Access

Novel context-specific genome-scale modelling explores the potential of triacylglycerol production by *Chlamydomonas reinhardtii*

Haoyang Yao, Sanjeev Dahal and Laurence Yang*

Abstract

Gene expression data of cell cultures is commonly measured in biological and medical studies to understand cellular decision-making in various conditions. Metabolism, affected but not solely determined by the expression, is much more difficult to measure experimentally. Finding a reliable method to predict cell metabolism for expression data will greatly benefit metabolic engineering. We have developed a novel pipeline, OVERLAY, that can explore cellular flux-omics from expression data using only a high-quality genome-scale metabolic model. This is done through two main steps: first, construct a protein-constrained metabolic model (PC-model) by integrating protein and enzyme information into the metabolic model (M-model). Secondly, overlay the expression data onto the PC-model using a novel two-step nonconvex and convex optimization formulation, resulting in a context-specific PC-model with optionally calibrated rate constants. The resulting model computes proteomes and intracellular flux states that are consistent with the measured transcriptomes. Therefore, it provides detailed cellular insights that are difficult to glean individually from the omic data or M-model alone. We apply the OVERLAY to interpret triacylglycerol (TAG) overproduction by *Chlamydomonas reinhardtii*, using time-course RNA-Seq data. We show that OVERLAY can compute *C. reinhardtii* metabolism under nitrogen deprivation and metabolic shifts after an acetate boost. OVERLAY can also suggest possible 'bottleneck' proteins that need to be overexpressed to increase the TAG accumulation rate, as well as discuss other TAG-overproduction strategies.

Keywords Genome-scale modelling, *C. reinhardtii*, Metabolic engineering, Optimization, Systems biology

Background

Microalgae have long been a promising class of organism as a synthetic biological chassis due to their high growth rate, efficient photosystem, and simplicity in cultivation. Because of its carbon fixation capability, it is also believed that algal products have less carbon dioxide emissions and are more sustainable in large-scale production [1]. Within all microalgae, *Chlamydomonas reinhardtii* is the best-studied organism in terms of genome annotation and molecular mechanisms, making it the reference organism for studying algal lipid metabolism [2]. Numerous molecular tools for *C. reinhardtii*, including its chloroplast, have also been developed by the research community, making its production of recombinant proteins easier than any other algae [3, 4]. As a result, *C. reinhardtii* has been utilized to make a wide variety of chemicals in the lab and industry. On the high value-added end, non-native proteins are expressed and

*Correspondence:
Laurence Yang
laurence.yang@queensu.ca
Department of Chemical Engineering, Queen's University, 19 Division St, Kingston K7L 2N9, Canada

produced in *C. reinhardtii* as pharmaceutics such as vaccines, antibiotics, and nutritional supplements [5]. This has been the more economically profitable and fruitful direction, and some standard workflows and toolkits have been established [6]. On the other hand, efforts are made to produce biofuels such as biodiesel, biohydrogen, and bio-alcohol from algae, which are chemicals closely related to the primary metabolism [7]. Some remarkable progress is made that increases *C. reinhardtii* lipid and starch contents by up to 2.5-fold by relatively simple modifications [1]. For example, Rengel et al. showed that overexpressing the acetyl-coenzyme-A (acetyl-CoA) synthetase gene can achieve up to 2.4-fold triacylglycerol (TAG) than the control group [8]. Some sophisticated studies have achieved an 8-fold hydrocarbon increment from the controlled *C. reinhardtii*, using gene knockout, heterologous expression, and triparental conjugation technique [9]. Specifically, Yunus et al. boosted fatty acid conversion to fatty alkane and fatty alkene by introducing enzymes such as FAP and UndA/B [9]. They also increased system fatty acid levels by deleting gene *aas*, thus preventing fatty acid consumption for cellular acyl-ACP replenishment [9]. Moreover, the metabolic response of *C. reinhardtii* to various medium conditions, such as nitrogen deprivation, acetate concentration, or light intensity, is a practical topic to increase TAG production further. Bogaert et al. showed that all biomass components including fatty acids increase in concentration per cell in response to supplementation with high acetate concentration [10].

Despite these findings, algal starch derivatives, lipid derivatives, and hydrogen are not yet economically feasible substitutes for fossil fuels on the market. Most of the experimental studies focus on modifying a few genes or adding a few chemical species into the medium without dramatically changing the cell from the wild type. It is a missed opportunity, as optimizing these new strains or potentially applying them in conjunctions can achieve a much higher yield with minor added costs. The optimization usually requires quantitative measurements from the phenotype, such as RNA-sequencing, proteomic data, and extracellular metabolomics, which are available from many existing studies. Developing an in-silico workflow would greatly assist researchers in systematically understanding the cellular metabolism from phenotype measurements, which is critical to optimize current biofuel-producing strategies and suggesting novel gene targets.

Genome-scale modelling (GEM) is an in-silico tool to systematically simulate cellular expression and metabolism, which is now widely used in biotechnology and infectious disease research. GEM is a species-specific biological reaction network, usually reconstructed by researchers from an annotated genome of the organism. Genome-scale metabolic model (M-model), the most basic yet accessible GEM that focuses exclusively on predicting metabolic fluxes, is the metabolic subnetwork with equilibrium constraint applied. M-model is a mathematical linear optimization problem (LP) (Methods, Eqs. (1)–(3)) and can usually be solved within 0.1 seconds using flux balance analysis (FBA) in COBRA Toolbox [11]. Noticeably, FBA is an algorithm that finds extreme flux values within the feasible metabolic range, and it is not explicitly designed for integrating measured expression data, which is referred to as context-specific modelling. In principle, context-specific modelling has better utilities in metabolic engineering than generic M-model, due to the former being constrained directly by omic data to resemble in vivo conditions. Existing algorithms for context-specific modelling are centred mainly around two approaches: limiting the flux of reactions with lowly expressed genes (i.e., GIMME), and defining and supporting a set of core reactions with highly expressed genes (i.e., mCADRE) [12–14]. Most of these existing methods require user-specified parameters such as expression thresholds, making them less objective and less accessible to a wider community. More importantly, the qualitative 'highly/lowly/not expressed' criteria is likely too coarse for investigating TAG production, as the target flux is inherently low. Consequently, it is harder for researchers to study insights from the modelling by these algorithms, as well as suggesting metabolic engineering strategies.

In this study, we develop a computational pipeline, OVERLAY to better address these challenges. We first formulated a protein-constrained metabolic model (PC-model) starting from the published *C. reinhardtii* M-model iCre1355 and chloroplast specific M-model iGR774 [15, 16]. On top of metabolism, PC-model has protein and enzyme concentrations as variables and can be solved using the FBA algorithm with additional benefits. We formulated protein constraint similar to Yurkovich et al., by adding protein concentrations and enzyme concentrations as variables into the M-model, which constrains respective metabolic fluxes [17]. Moreover, expression data from other studies were overlaid onto the PC-model for novel context-specific modelling, which can predict the respective metabolic state using FBA and flux variability analysis (FVA). The workflow of OVERLAY is demonstrated by Fig. 1, which consists of multiple automated algorithms. This will be especially helpful for optimizing bulk material productions from *C. reinhardtii*, and the TAG

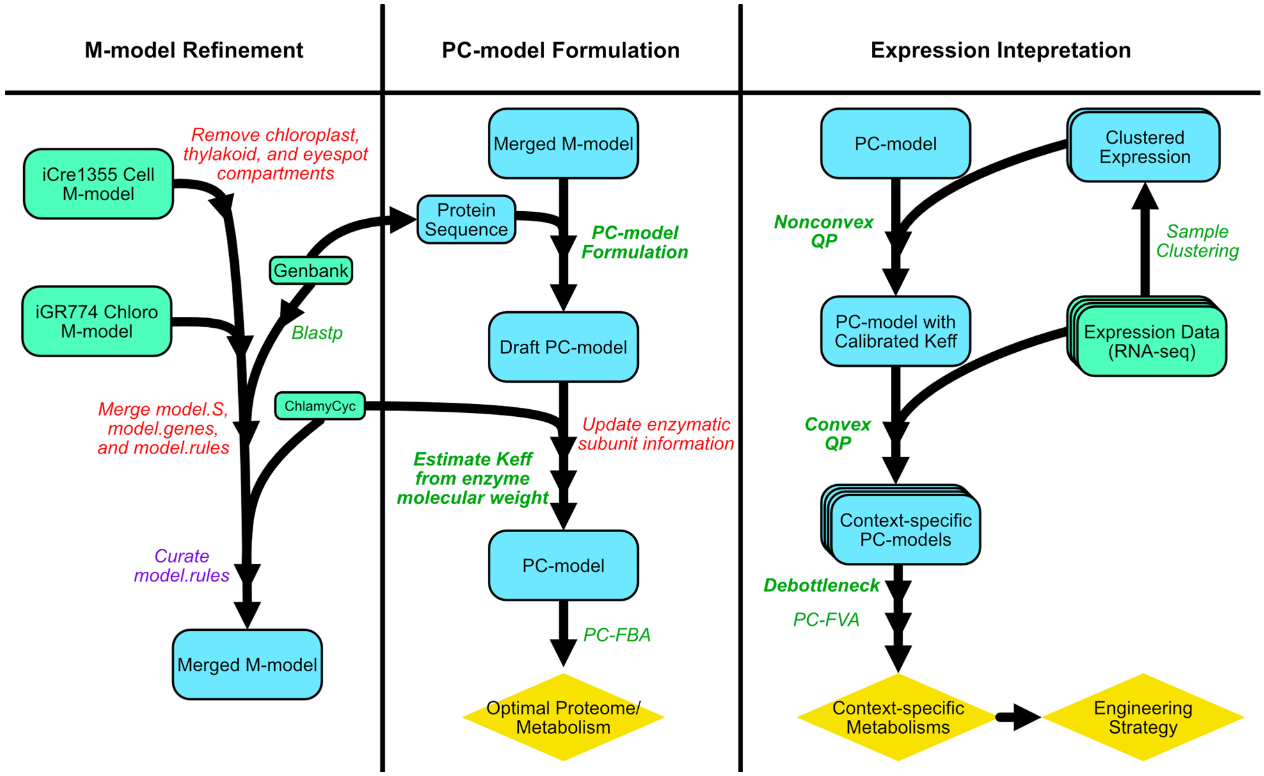


Fig. 1 Schematic of the OVERLAY computational pipeline. Boxed texts are files and data, where green rectangle, blue rectangle, and yellow diamond denote starting materials, intermediate steps, and results, respectively. Red italic texts refer to manual procedures, and purple italic texts refer to semi-manual steps with some script aids. In contrast, automatic procedures that are done by the only script are shown in green italic, and methods uniquely developed in OVERLAY are in bold green font

accumulation case study is done using RNA-seq data from other studies to show the efficacy of OVERLAY.

Results

Refined PC-FBA reveals optimal chloroplast metabolism, cellular metabolism, and transportation

We consider *C. reinhardtii* PC-model to be a superior version of the basic M-model, as it can be used for simple analyses such as FBA with better accuracy and more utilities. The total proteome budget has defaulted to a constant of 150 mg per gram of dry cell weight (gDW) (see "Methods" for detailed explanations). A noticeable advantage of PC-model is that exchange reaction boundaries do not need to be set manually. This allows accurate phenotype simulation without uptake flux measurements, thus further offering a convincing comparison of flux networks between different metabolic modes. For example, by assuming sufficient lighting, the photon exchange lower bound can be opened to −1000 mmol/gDW/h for any metabolic mode, and the exchange fluxes are solved by the respective optimization. Only the mixotrophic acetate uptake lower bound is manually constrained to −2 mmol/gDW/h to mimic a limited substrate availability.

We used PC-model to simulate the optimal growth strategy in autotrophic, mixotrophic and heterotrophic conditions. We compute metabolic shifts between these conditions focusing on chloroplast metabolism and transportation and interactions with the mitochondria (Fig. 2a). FBA of PC-model (PC-FBA) enabled us to compute the optimal flux-corresponding proteome allocation, of which proteome maps are generated under each growth condition (Fig. 2b–d) [18]. Photosynthesis-associated proteins accounted for 43.2–80.2% of proteome mass in all three conditions (Fig. 2b–d). Correspondingly, photosynthesis largely drives shifts in overall flux states (Fig. 2a).

Our model qualitatively reproduced major and subtle shifts determining the optimal electron flow through photosystems I and II in different conditions. *C. reinhardtii* thylakoid can choose between circular electron flow (CEF) of photosystem I only and linear electron flow (LEF) through both photosystem II and photosystem I, while LEF is the more energy-efficient option [19]. Consistent with this knowledge, the autotroph and heterotroph utilized LEF exclusively to maximize efficiency, as seen by zero flux from Fd to Cyt-b6f (the circular step of CEF) (Fig. 2a). Meanwhile, the mixotroph utilizes CEF, diverting 6% (0.80 mmol/gDW/h out of 13.10 mmol/gDW/h) of electron flux away from NADH/NADPH production back into the CEF. This optimal flux state suggests that CEF can result in faster growth over pure LEF under certain conditions. The result may explain how CEF and LEF are used to balance ATP/reducing power ratio for carbon fixation, as reported by Chaux et al. [20].

Another observation is that, assuming optimal growth, the mixotroph has a higher photosystem activity than the other growth modes: 13.10 mmol/gDW/h electron flux from Cyt-b6f to PC for mixotroph compared to 11.89 and 7.18 mmol/gDW/h for autotroph and heterotroph, respectively (Fig. 2a). These differences are explained by PC-FBA. The autotroph needs a large portion of the proteome budget to conduct other anabolism, such as carbon fixation and gluconeogenesis. These anabolic processes are highly proteome inefficient in comparison with directly consuming organic carbon substrate, rendering autotroph with limited proteome budget for photosystem complexes. For the heterotroph, it is more optimal to spend the proteome budget on consuming acetate, which serves as both an organic carbon source and an energy source. The mixotroph has limited acetate that might be enough for organic carbon to not be forced to run the inefficient anabolism, but insufficient as an energy source, thus having the most potent photosystem to harvest energy. We note that these simulations represent growth-optimized metabolic states. FBA and PC-FBA accurately predict metabolism of adaptively evolved phenotypes [21–23]; however, without additional data-driven constraints these simulations may not resemble wild-type behavior. Indeed, chloroplast content is generally observed to be lower in mixotrophic than autotrophic conditions [24, 25]. Therefore, additional biological mechanisms that are outside the PC-FBA model scope, such as regulation and photoinhibition, may play a significant role in mixotrophic metabolism.

PC-FBA also offers insightful chloroplast metabolism and transport simulations, especially regarding carbon fixation and triose-phosphate transport. As verified by other studies, 9 moles of ATP and 6 moles of NADH are required to produce 1 mole of triose-phosphate from carbon dioxide through the Calvin cycle [26]. Due to this high energy consumption, carbon fixation appears to be a suboptimal growth strategy compared to acetate uptake and is only active when the acetate supply is insufficient.

Under heterotrophic growth, the chloroplast is a net consumer of organic carbon, which is transported as 3-phosphoglycerate. Noticeably, the chloroplast in all phototrophic modes uptakes 3-phosphoglycerate while excreting other triose-phosphate (Fig. 2a). Being the energy supplier of the cell, autotrophic and mixotrophic chloroplast mainly excretes the energy-compact glyceraldehyde-3-phosphate, which has been a phenomenon reported by other studies [26]. The mixotroph has the most active chloroplast, exporting more ATP and reducing power in the form of glyceraldehyde-3-phosphate and oxaloacetate/malate exchange. In all conditions, the chloroplast also consumes various amino acids while producing lipid precursors and six-carbon sugars, which are not shown in the figure in detail.

Our PC-model provides valid mitochondrion fluxomic and cellular exchange simulations for all growth modes. The mixotroph and heterotroph simulations used the glyoxylate shunt in the mitochondria (Fig. 2a), as reported previously by Johnson et al. [26]. This process generates excess oxaloacetate for heterotroph, which is converted to phosphoenolpyruvate and exported to the cytosol from the mitochondria. Additionally, the heterotroph excretes formate, which is only present when proteome constraints are applied (see Additional file 6: Table S1). Thus, the proteome constraints are required to correctly predict respiro-fermentation, or overflow metabolism, as observed in multiple organisms including *C. reinhardtii* [27, 28]. In particular, the optimal heterotroph allocates 43.2% of proteome mass to photosynthesis (Fig. 2d), compared to 80.2% in the mixotroph (Fig. 2c). The reduced photosynthesis protein budget in the heterotroph is allocated instead partially toward glycolysis and TCA cycle proteins (total 15.0%) (Fig. 2d).

Meanwhile, the autotroph shows several contrasting metabolic activities to the heterotroph. The autotrophic mitochondrion has very little activity, mostly powered by importing pyruvate from the cytosol. Instead, it generates phosphoenolpyruvate in the chloroplast, which is transported to the mitochondrion (Fig. 2a). The autotroph

(See figure on next page.)

Fig. 2 PC-FBA simulation results of autotrophic, mixotrophic, and heterotrophic *C. reinhardtii* growth mode. The optimal metabolic fluxes are shown in **a**, where autotrophic, mixotrophic, and heterotrophic fluxes are denoted by green, blue, and red numbers, respectively. All fluxes are shown in mmol/gDW/h. A negative flux value means the flux is flowing in the opposite direction of the arrow. The dotted lines show the electron flux. All metabolites are shown in BiGG ID. Complex/enzyme abbreviations in thylakoid: PSII, photosystem II; PQ, plastoquinone/plastoquinol; Cyt b6f, cytochrome b6f complex; PC, plastocyanin; PSI, photosystem I; Fd, ferredoxin; FNR, ferredoxin NADP+ reductase; and ATPSh, CF0F1 ATP synthase. The optimal proteome of three growth modes is shown by proteome maps of **b**, **c**, and **d**. Each small polytope is a single protein, and its area denotes the relative abundance. Likely coloured polytope is classified under the same subsystem, which is written in white text. Only modelled proteins are drawn. *Some reactions of the TCA cycle are outside of mitochondria

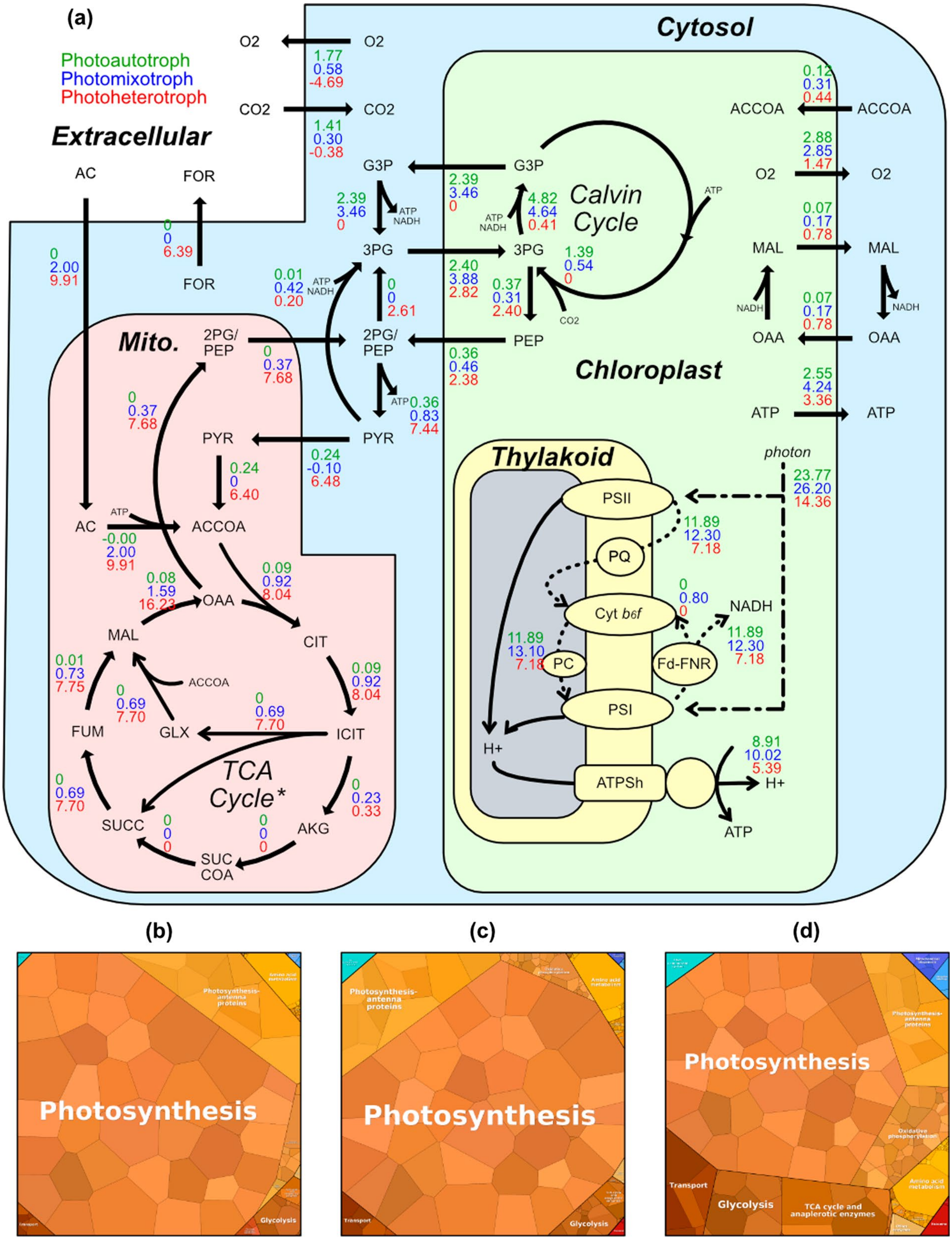


Fig. 2 (See legend on previous page.)

does share some characteristics with the mixotroph, such as consuming carbon dioxide and producing oxygen. These PC-FBA simulations suggest that the mixotrophic flux network is more well-balanced and might be the ideal candidate for metabolic engineering.

Investigating lipid accumulation in nitrogen-deprived *C. reinhardtii* by OVERLAY

Based on insights gained from PC-FBA simulations, we further investigated mixotrophic conditions for bulk metabolite overproduction. The accumulation of TAG, a useful and value-added industrial compound, has been studied extensively in *C. reinhardtii* by inducing nitrogen deprivation. In particular, Goodenough et al. investigated a *sta6* (unable to form starch) strain of *C. reinhardtii*, which showed enhanced TAG accumulation under nitrogen deprivation with acetate boosting 48 h later [29]. The study collected time-course RNA-Seq over four days of culture and discovered highly complex gene expression dynamics: 425 genes up-regulated and over 850 genes down-regulated in response to acetate [29]. Here, we use our OVERLAY to decipher how these complex gene expression dynamics drive flux changes that ultimately lead to enhanced TAG production.

OVERLAY Constructs context-specific models with calibrated rate constants

We first used convex QP only to fit each of the 16 time-course samples onto the PC-model, resulting in 16 context-specific PC-models. Noticeably, enzymatic rate constant (K_{eff}) values are difficult to determine because most values are not experimentally available. We assume initially that K_{eff} are centred around a basal value of $K_{\text{eff}}^{\text{avg}} = 65s^{-1}$, and they are proportionally scaled to the SASA [30, 31] (Methods, Nonconvex QP). Across all samples, the best-fitted proteome vectors are consistent with RNA-seq data, with R^2 ranging between 0.950 and 0.963, with a median $R^{2*} = 0.958$ (see Additional file 1: Fig. S1a for the complete plot). For example, sample 4 (time = 4 h) has $R^{2*} = 0.954$, with 56 outliers (≥ 3 times inconsistency) and 16 far outliers (≥ 10 times inconsistency) out of 1495 proteins (Fig. 3a).

Our OVERLAY optimally tuned r, or equivalently all K_{eff}, to achieve the best fit of simulated proteomes to the RNA-Seq, subject to carefully formulated constraints (see "Methods"). We clustered 16 RNA-seq samples into four groups (Additional file 2: Fig. S2), which were used to estimate a single K_{eff} vector representing all samples. This results in an improved best-fitted proteome from the original with a median $R^{2*} = 0.966$ across 16 samples (Additional file 1: Fig. S1b). Fig. 3b has 48 outliers, 15 far outliers, and a higher R^{2*} than using convex QP only. Noticeably, the fitting improvement is achieved by varying K_{eff} only slightly from $K_{\text{eff}}^{\text{ori}}$. According to Fig. 3c, the distributions of K_{eff} before and after nonconvex QP adjustment are similar, although a few enzymes are assigned much higher K_{eff} than before. Only 214 out of 1222 enzymes have a K_{eff} different from K_{eff}^{ori} due to extra constraints on OVERLAY (Methods, Eq. (19)), which are placed to reduce the number of total adjustments to K_{eff}.

Additionally, OVERLAY helps to quality control the metabolic reconstruction, especially regarding its gene-reaction association. We identified a set of proteins whose abundances could not match measurements across all samples. Because we allowed for adjusted rate constants, we hypothesized that the reason for these inconsistencies

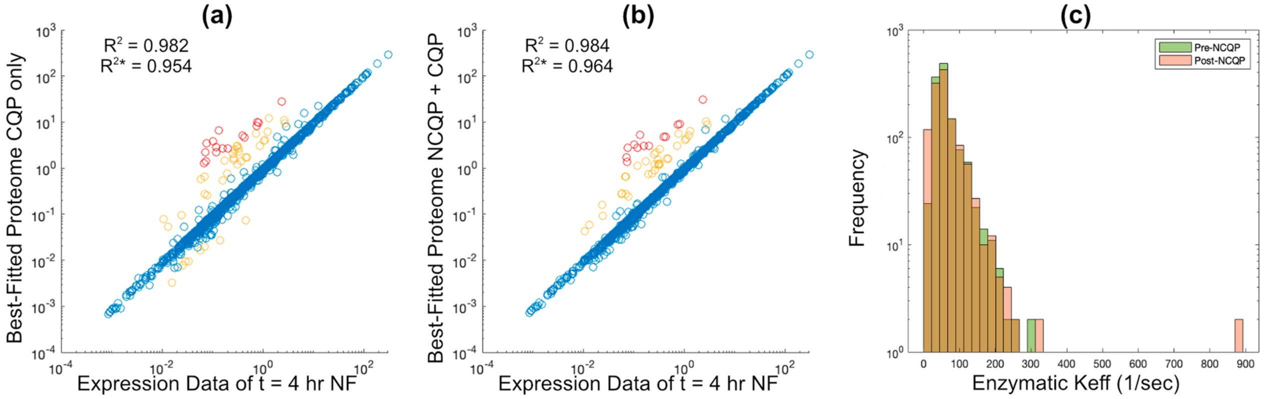


Fig. 3 Consistency of simulated proteomes to transcriptomics, and estimated rate constants by OVERLAY. Best-fitted proteome versus measured transcriptomes at t = 4 h before (**a**) and after (**b**) the enzymatic rate constant adjustment by nonconvex QP. R^{2*} is computed using log-transformed data for simulated proteomes and measured transcriptomes, whereas R^2 is computed without log-transforming. Outliers are denoted in yellow (≥ 3 times) and red (≥ 10 times). R^2 values are computed using all points, including outliers. **c** Demonstrates K_{eff} values before calibration using OVERLAY (pre-NCQP) and after calibration (post-NCQP). Here, $K_{\text{eff}} = r \cdot K_{\text{eff}}^{\text{avg}}$

is due to mis-annotations in the original model reconstruction. We manually inspected all 15 proteins that are far outliers in at least 8 out of 16 samples and compared them with their annotated functions in ChlamyCyc and ALGAEPATH (see Additional file 6: Table S2a for the full list and details) [32, 33]. Indeed, we found that 5/15 proteins had incorrect gene–protein-reaction associations in the reconstruction (see Additional file 6: Table S2a). Of the remaining ten inconsistent proteins, we found potential isozymes for four proteins. We found a total of eight potential isozymes (Additional file 6: Table S2b), which are promising candidates for future studies.

Context-specific PC-models providing new metabolic insights

The main merit of the context-specific PC-model is converting expression data to metabolic fluxes, which are insightful both independently and comparatively across samples. For example, the maximum TAG production rate is slightly reduced by ammonium-free medium and slightly promoted by the acetate boost (Fig. 4a), yet it does not translate to the 'actual' accumulation rate, as any point on the bar is possible for *C. reinhardtii* to operate on.

Using the final PC-model, including calibrated k_{eff}, we performed a systems-level analysis of dynamic shifts in the proteome and fluxome for mixotrophic TAG production. Given the best-fit proteome for every RNA-Seq sample, we computed the corresponding fluxome using protein-constrained flux variability analysis (PC-FVA) with TAG production rate constrained to $\geq$ 0%, 50%, 90%, and 99% of the maximum (Fig. 4a). For each reaction, we computed the Spearman rank correlation (ρ) between the max flux from PC-FVA and the total abundance of all transcripts associated with the reaction. From this procedure, among all 1876 enzymatic reactions, we classify 130 reactions as expression-dependent ($\rho \geq 0.8$), 218 reactions as expression-correlated ($0.5 \leq \rho < 0.8$), and 1528 reactions as expression-independent ($\rho < 0.5$), while spontaneous reactions are always expression independent (Additional file 3: Fig. S3).

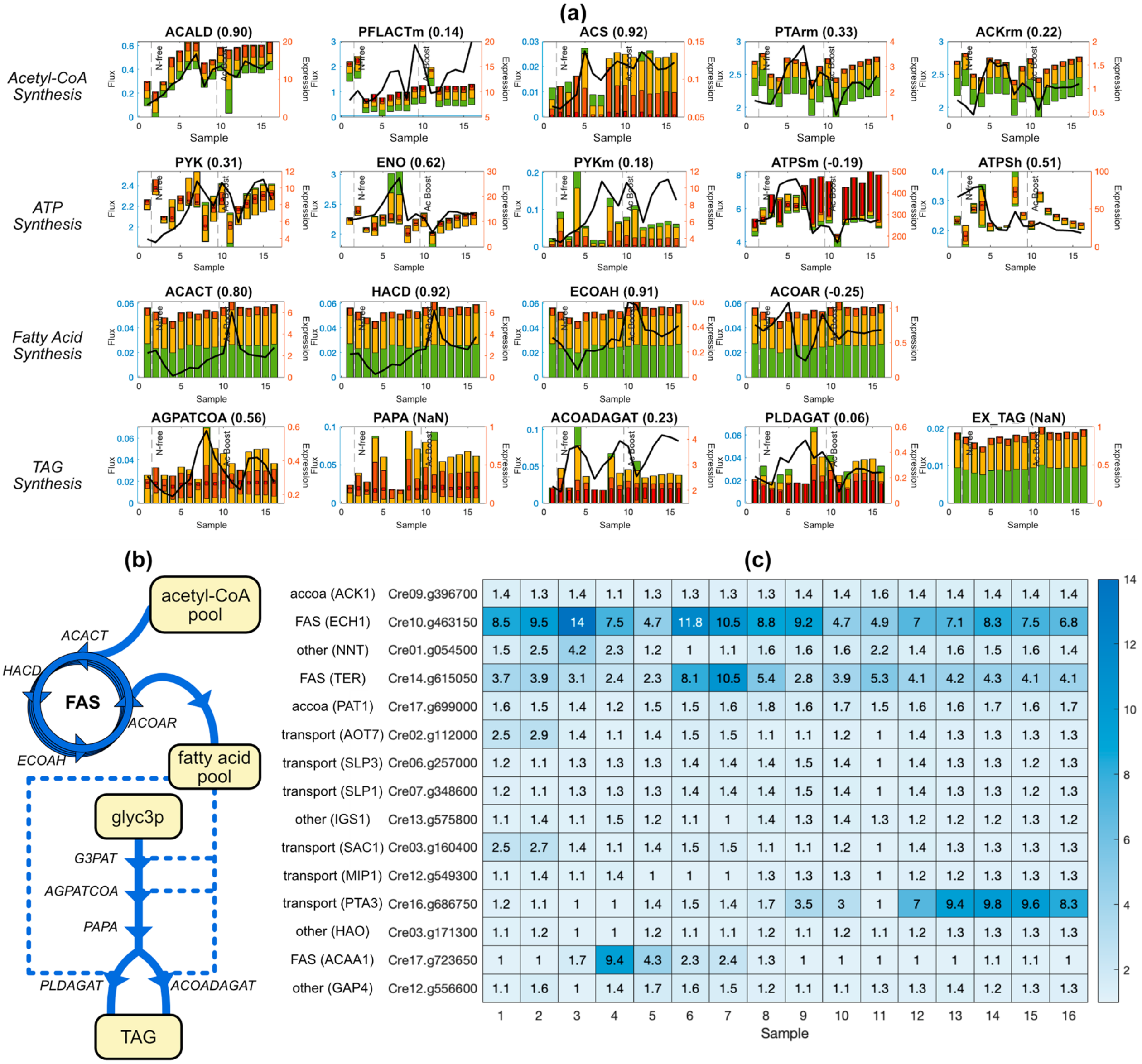
From these PC-FVA results, especially with high optimum percentages (orange and red bars in Fig. 4a), we find that the key reactions for TAG production can be categorized into acetyl-CoA synthesis, ATP synthesis, *De novo* synthesis of free fatty acids, and TAG synthesis (Fig. 4a).

Acetyl-CoA synthesis The majority of acetyl-CoA was supplied from acetaldehyde dehydrogenase (ACALD), formate C-acetyltransferase (PFLACTm), and phosphotransacetylase (PTArm). Of these reactions, ACALD is the only expression-dependent (Spearman rank $\rho = 0.9$) reaction (Fig. 4a). ACALD is associated solely with the gene Cre17.g746997 and shows high expression correlation even for PC-FVA computations with TAG production $\geq$ 99% of the maximum (Fig. 4 red bars). TAG production is strongly dependent on ACALD flux, which in turn is strongly expression-dependent. Thus, Cre17.g746997 is an overexpression candidate to increase acetyl-CoA supply.

On the other hand, PFLACTm and PTArm fluxes are uncorrelated with gene expression (Spearman rank $\rho < 0.5$). PFLACTm, producing acetyl-CoA by converting pyruvate to formate, is likely dictated by the upstream pyruvate mass balance. PTArm catalyzes the highest flux of acetyl-CoA production and does not appear to be dictated by either its own (Cre09.g396650 or Cre17.g699000) expression or acetate kinase (ACKrm) (Cre09.g396700 or Cre17.g709850) expression. Acetyl-CoA synthetase (ACS) produces acetyl-CoA using acetate like PTArm, and it also coincides with the expression. However, its flux is nearly 100-fold lower than PTArm. Furthermore, maximum TAG dilution is achieved when ACS flux is low, and PTArm flux is high (Fig. 4a). This result suggests that ACS flux can be controlled through gene expression and that to maximize TAG production, its expression should be repressed.

ATP synthesis ATP production is dominated by cytosolic pyruvate kinase (PYK) and mitochondrion ATP synthase (ATPSm), while the chloroplast ATP synthase (ATPSh) activity is relatively low across all samples. This mode of ATP synthesis observed for TAG production contrasts sharply with PC-FBA simulations of optimal growth that showed large proteome allocation toward photosystems I and II (Fig. 2a). The acetate boost starts from sample 10—prior to the boost, ATPSh flux shows moderate variability and low flux relative to mitochondrial ATPS. After the acetate boost, ATPSh flux variability is even narrower, and TAG production is maximized with lowered ATPSh flux. The good correlation of ATPSh flux with gene expression (Spearman rank $\rho = 0.51$) indicates that *C. reinhardtii* is programmed to reduce photosystem-based ATP synthesis under mixotrophic growth with nitrogen limitation. Indeed, the dynamic regulation of ATP production between mitochondria and chloroplast has been studied, but no transcription factor is found [34].

PYK produces ATP by converting phosphoenolpyruvate to pyruvate. PYK flux and protein allocation differ notably between the optimal mixotrophic and heterotrophic flux networks (2PG/PEP to PYR in the cytosol in Fig. 2A). During TAG production, PYK flux is not strongly correlated with gene expression, indicating that other constraints, such as mass balance, determine its flux. Specifically, enolase (ENO) produces phosphoenolpyruvate, which is the primary substrate of PYK.



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
accoa (ACK1)	Cre09.g396700	1.4	1.3	1.4	1.1	1.3	1.3	1.3	1.3	1.4	1.4	1.6	1.4	1.4	1.4	1.4	1.4
FAS (ECH1)	Cre10.g463150	8.5	9.5	14	7.5	4.7	11.8	10.5	8.8	9.2	4.7	4.9	7	7.1	8.3	7.5	6.8
other (NNT)	Cre01.g054500	1.5	2.5	4.2	2.3	1.2	1	1.1	1.6	1.6	1.6	2.2	1.4	1.6	1.5	1.6	1.4
FAS (TER)	Cre14.g615050	3.7	3.9	3.1	2.4	2.3	8.1	10.5	5.4	2.8	3.9	5.3	4.1	4.2	4.3	4.1	4.1
accoa (PAT1)	Cre17.g699000	1.6	1.5	1.4	1.2	1.5	1.5	1.6	1.8	1.6	1.7	1.5	1.6	1.6	1.7	1.6	1.7
transport (AOT7)	Cre02.g112000	2.5	2.9	1.4	1.1	1.4	1.5	1.5	1.1	1.1	1.2	1	1.4	1.3	1.3	1.3	1.3
transport (SLP3)	Cre06.g257000	1.2	1.1	1.3	1.3	1.3	1.4	1.4	1.4	1.5	1.4	1	1.4	1.3	1.3	1.2	1.3
transport (SLP1)	Cre07.g348600	1.2	1.1	1.3	1.3	1.3	1.4	1.4	1.4	1.5	1.4	1	1.4	1.3	1.3	1.2	1.3
other (IGS1)	Cre13.g575800	1.1	1.4	1.1	1.5	1.2	1.1	1	1.4	1.3	1.4	1.3	1.2	1.2	1.2	1.3	1.2
transport (SAC1)	Cre03.g160400	2.5	2.7	1.4	1.1	1.4	1.5	1.5	1.1	1.1	1.2	1	1.4	1.3	1.3	1.3	1.3
transport (MIP1)	Cre12.g549300	1.1	1.4	1.1	1.4	1	1	1	1.3	1.3	1.3	1	1.2	1.2	1.3	1.3	1.3
transport (PTA3)	Cre16.g686750	1.2	1.1	1	1	1.4	1.5	1.4	1.7	3.5	3	1	7	9.4	9.8	9.6	8.3
other (HAO)	Cre03.g171300	1.1	1.2	1	1	1.2	1.1	1.1	1.2	1.1	1.2	1.1	1.2	1.3	1.3	1.3	1.3
FAS (ACAA1)	Cre17.g723650	1	1	1.7	9.4	4.3	2.3	2.4	1.3	1	1	1	1	1	1.1	1.1	1
other (GAP4)	Cre12.g556600	1.1	1.6	1	1.4	1.7	1.6	1.5	1.2	1.1	1.1	1.3	1.3	1.4	1.2	1.3	1.3

Fig. 4 Various results of expression data interpretation through context-specific PC-model. **a** Is a selected collection of PC-FVA results across 16 samples regarding acetyl-CoA synthesis, ATP synthesis, *De novo* fatty acid synthesis (FAS), and TAG synthesis pathway reactions. The bar plot shows the variability of each metabolic reaction and is coupled to the left y-axis in mmol/gDW/h. Green, yellow, orange, and red bars reflect the flux variability at 0%, 50%, 90%, and 99% of maximum TAG synthesis rate (i.e., EX_TAG flux), respectively. The black line plot shows the expression level of the reaction and is coupled to the right y-axis. In the case of isozymes presence, the black line plots the numerical sum of all isozyme levels. The calculated Spearman's rank correlation (ρ) between the FVA result and expression is listed in the bracket. Vertical dashed lines divide time-series samples into pre-wash, N-free, and post-boost phases. **b** Demonstrates a simplified reaction network around FAS and the TAG synthesis pathway. Different fatty acid, their derivatives, and different TAG are not differentiated. **c** Shows the suggested overexpression level (folds) of the top 15 bottleneck proteins to maximize TAG production while minimizing deviation from the measured RNA-Seq (through the protein abundance constraints). Gene symbols and functional categories of each protein are shown to the left of each gene ID (row labels)

ENO is correlated with expression of Cre12.g513200 (Spearman rank $\rho = 0.62$). In turn, PYK flux is highly correlated with ENO (Spearman rank $\rho = 0.994$); therefore, both PYK and ENO flux can be controlled by regulating the Cre12.g513200 gene.

De novo fatty acid and TAG synthesis An intuitive way to maximize TAG production would be to overexpress its direct biosynthesis genes. This strategy has been applied for multiple algae species, including *C. reinhardtii* [35]. TAG is synthesized from glycerol 3-phosphate by a sequence of five reactions: glycerol-3-phosphate (G3PAT), 1-hexadecanoyl-sn-glycerol 3-phosphate O-acyltransferase (AGPATCOA), phosphatidate phosphatase (PAPA), acyl-CoA diacylglycerol acyltransferase (ACOADAGAT), and phospholipid diacylglycerol acyltransferase (PLDAGAT). Three acyl groups are attached in the process, using acyltransferase reactions (G3PAT, AGPATCOA, ACOADAGAT, PLDAGAT). TAG accumulation has been increased by overexpressing AGPATCOA in *Cyanidioschyzon merolae* [36], PLDAGAT in *C. reinhardtii* [35] and *Phaeodactylum tricornutum* [37]. These strategies work by pulling carbon flux to TAG.

Our simulations are consistent with these observations in that maximum TAG accumulation requires elevated expression of the acyltransferase proteins. Namely, the minimum required flux of AGPATCOA was above ~ 0.01 to achieve max TAG flux (Fig. 4a—*TAG synthesis*). The two diacylglycerol acyltransferases (ACOADAGAT and PLDAGAT) are also required to maximize TAG flux, but because either reaction can be used, each flux has a minimum requirement of zero.

Our simulations indicate that another, possibly more critical, strategy for TAG production is to provide ample acyl-CoA by overexpressing free fatty acid synthesis cycle reactions: Acetyl-CoA C-acyltransferase (ACACT), 3-hydroxyacyl-CoA dehydrogenase (HACD), and enoyl-CoA hydratase (ECOAH), and trans-2-enoyl-CoA reductase (ACOAR). These reactions showed time-course flux patterns that were nearly identical to the TAG dilution reaction (EX_TAG) (Fig. 4a). Three of these reactions (ACACT, HACD, and ECOAH) are highly correlated with gene expression (Spearman rank $\rho = 0.8$, 0.92, and 0.91). Therefore, overexpression of these genes would directly increase flux. Indeed, a sharp overexpression of these genes coincides well with the acetate boost (Fig. 4a, *De novo synthesis*). Finally, the ACOAR reaction is uncorrelated with gene expression (Spearman rank $\rho = -0.25$); however, due to mass balance constraints, its flux is entirely determined by the flux of the preceding three reactions in the cycle.

Engineering strategies for enhancing TAG productivity

Next, we developed a tool to find optimal, system-level debottlenecking strategies for metabolic engineering. This tool is formulated as a linear program (Methods): TAG production is maximized subject to all context-specific PC model constraints while also allowing for a user-defined total protein overexpression "budget" (E). The optimal solution to this problem provides the set of highest priority targets for protein overexpression. Using this tool, we identified protein overexpression strategies that were consistent with the transcriptome changes observed in the acetate-boosted experiment. For example, all three free fatty acid synthesis genes (ECH1, TER, ACAA1) that were highly correlated with reaction flux (ECOAH, HACD, ACACT) were identified as overexpression targets (Fig. 4b). ECH1 requires an average of 8.2-fold overexpression across 16 samples, the most out of all modelled genes.

Additionally, two of the acetyl-CoA synthesis genes (*ackA and PAT1*), corresponding to ACKrm and PTArm reactions, were identified as overexpression targets. Interestingly, the algorithm did not identify the gene for ACALD as an overexpression target, despite it being a key step in Acetyl-CoA synthesis. This result is consistent with the high expression levels of ACALD-associated transcripts (Fig. 4a, ACALD); therefore, no further debottlenecking is required. In fact, this result suggests that the expression levels for PATrm and ACKrm-associated genes may need to be increased further to achieve TAG production higher than that observed.

Finally, we identified additional overexpression candidates that may be candidates for future engineering. These proteins include six transporters, including those for sulphate (*SLP1, SLP3*), phosphate (*PTA3*), and amino acids (*AOT7*) (Fig. 4b). Other targets include proteins associated with amino acid biosynthesis (*IGS1*), glycolysis (*GAP4*), redox balance (*NNT*), and glyoxylate metabolism (*HAO*) (Fig. 4b).

Apart from debottlenecking TAG synthesis through overexpression, we can also block competing reactions by gene deletions to achieve the maximum potential. For example, multiple studies have shown that starchless mutants of *C. reinhardtii* exhibit significantly more TAG accumulation [38, 39]. Our flux simulations are consistent with these results, showing reduced TAG production capabilities when starch synthesis reaction is active (Additional file 4: Fig. S4, STARCH300S). Shin et al. discovered that knocking out phospholipase A2 (Cre02.g095000) would increase *C. reinhardtii* lipid productivity by up to 64% [40]. Our simulation also resembles that an active phospholipase A2 will always reduce TAG productivity by deviating diacylglycerol away from TAG synthesis (Additional file 4: Fig. S4, PLPSA2). Yunus et al. and Kato et al. showed that deleting the fatty acyl-ACP synthase would better preserve the system free fatty acid level, and therefore boost hydrocarbon accumulation in cyanobacteria [9, 41]. In *C. reinhardtii* simulation, we observed a different but similar phenomenon, wherein the reaction fatty acid CoA ligase hinders TAG production (Additional file 4: Fig. S4, FACOAL). Experimentally knocking out the three genes above forces their respective fluxes to zero, which are the flux values OVERLAY predicts to enable maximum TAG production.

Another possible TAG productivity-enhancing approach is to supplement the medium with additional carbon sources. The measured transcriptomes indicated the expression of transporters for alternative carbon substrate uptake, including L-glutamine, D-ribose, and D-lactate. By adding these carbon sources to the *in silico* growth media, we confirmed that the observed gene expression levels support the uptake of these carbon sources, albeit at slow rates (see Additional file 6: Table S3). However, our simulations indicate that supplementing these carbon sources would not boost TAG production under the current transcriptome because the added carbon can not alleviate the main bottleneck, fatty acid biosynthesis.

Discussion

C. reinhardtii as a synthetic biology chassis

Microalgae *C. reinhardtii* has been studied for decades as a cell factory, producing both bulk metabolites and, more recently, expressing heterologous genes for non-native value-added products. Although many established methods are available for bulk metabolite production, in some cases, they are still far from optimal productivity based on our simulation results. Indeed, the recent adaptive laboratory evolution of *C. reinhardtii* has increased both growth rate (by up to 300%) and product yields (DHA production by 90%) [42]. Using genome-scale modelling, especially with the context-specific PC-model pipeline, *C. reinhardtii* may become an economically-efficient cell factory for bulk chemicals after multiple iterations of optimization. On the other hand, the optimization for non-native products is more complicated, yet it can be highly impactful for the (bio)chemicals industry due to the potential for sustainable production of high-value products, especially when non-biological synthesis routes are unavailable.

Besides *C. reinhardtii* and algae in general, *Escherichia coli* has also been extensively studies as a synthetic cell chassis. The main advantages of *C. reinhardtii* over *E. coli* are its lower carbon emission and simplicity in cultivation, which are both more significant in bulk chemical productions setup. Therefore, biofuel and potential food production has been the main interest in *C. reinhardtii* cell factories. On the other hand, *E. coli* is the best studies microorganism with even more established knowledge and synthetic biology tools available to researchers. The production of certain value-added chemicals by *E. coli*, including various vitamins and nutraceuticals, has reached commercialization stage [43, 44]. We believe that OVERLAY, as a general computational tool, is also capable of helping to boost the productivity of *E. coli* cell factories.

Evaluation and application of OVERLAY

Our PC-model formulation and OVERLAY pipeline provide several advantages over existing methods. First, our PC-model computes optimal fluxomes in response to changes in the allocation of the proteome. This proteome, in turn, is highly consistent with measured transcriptomes through a sequence of convex and non-convex quadratic optimization problems. Second, to perform context-specific simulations, we do not require choosing an arbitrary gene expression threshold for turning on/off reactions based on transcript abundance—this has been a challenge in existing methods [12]. Third, our pipeline enables the identification of optimal overexpression targets. This method requires only one parameter to be adjusted: *E* (total protein overexpression budget), which can be determined using a simple procedure. Finally, our method enables using transcriptomics to quality-control genome-scale reconstructions and their annotations, which are found through persistent discrepancies between optimal protein and measured transcript abundances.

Given all the advantages offered, the application of OVERLAY can be extended from metabolic engineering to a broader scientific inquiries. Importantly, OVERLAYovercomes the shortcomings (optimality assumption and incomplete scope) of FBA and PC-FBA by imposing constraints and calibrating rate constants using transcriptomics measurements that are consistent with observed phenotypes. For example, mixotrophic regulations and metabolisms in algae are a research focus for scientists. A recent study by Vidotti et al. measured time-course transcriptomic and proteome data for autotrophic, mixotrophic, and heterotrophic *Chlorella vulgaris* [45]. These measurements can be easily incorporated onto the newest *C. vulgaris* metabolic reconstruction iCZ843 using OVERLAY for deeper metabolic insights [46]. We anticipate OVERLAY to consistently help to address scientific problems related to steady state and transition state metabolism. Additionally, because OVERLAY can find errors in knowledge-based M-models (i.e., Additional file 6: Table S2), it may help researchers to update their understanding, such as discovering new metabolic pathways or enzymatic functions. These utilities are not offered by the original M-model or even PC-model solely, as shown previously in optimal autotrophic and mixotrophic photosystem activity predictions.

In general, PC-model and OVERLAY is a simple yet effective tool for understanding and manipulating cellular metabolism through gene expression, making it potentially valuable for various applications. The prediction results can be used for practical decision-making in various research fields such as biotechnology, infectious disease, and cancer. We believe OVERLAY will benefit the system biology and metabolic engineering community.

Prospect: incorporating other omics data

With more multi-omics data measurements in recent studies, we think it is worth mentioning the potential of OVERLAY to uniquely incorporate more omics data to achieve higher modelling performance.

Kinetome

Kinetome refers to the collection of cellular enzymatic rate constants [47]. Due to the scarcity of measured enzyme kinetic, OVERLAY is designed to be capable of conducting high-quality *C. reinhardtii* modelling without kinetome data. We believe OVERLAY can incorporate available kinetome data by explicitly setting and fixing individual rate constants in the PC-model before non-convex QP. Available kinetome data would likely increase the modelling performance.

Proteome

Proteome profiling data, if available, is a substitute of transcriptomic data for OVERLAY. In this study, OVERLAY uses transcriptome data to approximate cellular proteome, which is potentially less accurate than using proteome data directly. However, RNAseq remains more accessible than proteome profiling, making it more practically useful in modelling for biotechnological applications. If proteome profiling is available but not for all proteins, it can still be useful in partially constraining metabolic flux or verifying its consistency with transcriptomic data.

Metabolome

Metabolomic data refers to the presence or concentration of intracellular metabolites. Recently, the GEM community has developed various protocols, such as MetaboTools and matTFA toolbox, to incorporate metabolomics data into M-model [48, 49]. These protocols can be adopted in conjunction with OVERLAY to improve model quality. These fall outside the scope of this study; however, they are promising directions for future work.

Conclusion

In this work, we developed a computational pipeline OVERLAY for building a context-specific, protein-constrained genome-scale model (PC-model), starting from metabolic reconstruction and transcriptomics data. We showcase the utility of PC-model for deciphering how complex gene expression dynamics drive system-level fluxome shifts in *C. reinhardtii* using published time-course transcriptomics data. Using PC-FBA, we recapitulate metabolic hallmarks of autotrophic, heterotrophic, and mixotrophic growth. Importantly, the protein constraints are required to accurately simulate respiro-fermentation (overflow) metabolism. We then use time-course RNA-Seq data to investigate the over-production of triacylglycerol (TAG) in response to acetate supplementation under nitrogen limitation [29]. Our pipeline generated context-specific models for each experimental time point (over four days of culture), with very high consistency between 1495 modelled proteins and measured transcriptomes (R^2 between 0.95 to 0.963, median $R^2 = 0.958$). We then determined which metabolic fluxes were controlled by gene expression. By comparing simulated fluxes and measured transcriptomes across the 16 time-course RNA-Seq samples, we could categorize all gene-associated reactions into 130 expression-dependent (Spearman rank $\rho \geq 0.8$), 218 expression-related (Spearman rank $0.5 \leq \rho < 0.8$, and 1528 expression-independent (Spearman rank $\rho < 0.5$) reactions.

To enable researchers to systematically identify optimal overexpression targets, we developed a novel optimization-based tool. Using the tool, we identified key gene expression bottlenecks for TAG overproduction. The tool recapitulated known bottlenecks (e.g., the acyltransferase steps in TAG biosynthesis). Furthermore, we identified several novel overexpression targets to improve TAG overproduction further, including genes encoding sulphate, phosphate, and amino acid transporters; glyoxylate metabolism, and redox balancing.

Material and methods

Merging iCre1355 and iGR774 metabolic model and curation

Instead of using *C. reinhardtii* M-model iCre1355 only, we decided to plug in a newer chloroplast M-model, which contains a more up-to-date understanding of chloroplast metabolism. We merged M-model iCre1355 (cellular model) and iGR774 (chloroplast model) by first deleting all chloroplast metabolites and reactions in the cellular model. The chloroplast model was slightly modified (see Additional file 7: Table S4), and its transported reactions were matched to the dead-end transportations in the cellular model (see Additional file 7: Table S5–S7). The newly merged M-model has 1354 genes, 2641 reactions, and 2240 metabolites. We curated the gene-reaction rules (stored in the model as `model.rules` field) using complex data from ChlamyCyc 8.0 [32]. Only enzyme complexes with multiple subunits were curated, but not any protein-monomer enzymes (Additional file 7: Table S8). The starch metabolism pathway and appropriate rules were also added to the merged model (Additional file 7: Table S4). We also opened the lower bound of reaction GAPDHi and GAPDH_nadp to allow reverse reactions (Additional file 7: Table S4). The script written to merge and modify M-models is `MergedModel.m`, which calls functions in COBRA Toolbox on MATLAB to load and manipulate M-models [11, 50]. We used the Kyoto Encyclopedia of Genes and Genomes (KEGG) and BiGG Models as general references in M-model modifications [51, 52].

Obtaining and processing expression data for case study

Raw reads of RNA-seq data (E-GEOD-56505) for the TAG case study were downloaded as FASTQ files [29]. We downloaded the NCBI genome assembly `GCF_000002595.2.gbff` and parsed the GenBank file into a FASTA reference transcript [53]. Reads were aligned using Bowtie2 with default settings and quantified using Samtools and Salmon [54–56]. The complete quantified vector is denoted as $t_{cp} \in \mathbf{R}^{17713}$, which is further parsed a modelled transcript vector $t \in \mathbf{R}^{1495}$.

Protein constraints implementation

Our PC-model formulation is shown below.

$$\max_{v,p,x,e} c^T v \tag{1}$$

$$s.t.\ Sv = 0 \tag{2}$$

$$v^{\mathrm{lb}} \leq v \leq v^{\mathrm{ub}} \tag{3}$$

$$Cx \leq p \tag{4}$$

$$e_{\mathrm{for}} + e_{\mathrm{rev}} = B \cdot diag(r) \cdot x \tag{5}$$

$$-K_{\mathrm{eff}}^{\mathrm{avg}} I e_{\mathrm{rev}} \leq v \leq K_{\mathrm{eff}}^{\mathrm{avg}} I e_{\mathrm{for}} \tag{6}$$

$$0 \leq p \leq p^{\mathrm{ub}} \tag{7}$$

$$p^T d \leq P, \tag{8}$$

where $v \in \mathbb{R}^{2641}$, $p \in \mathbb{R}^{1495}$, $x \in \mathbb{R}^{1000}$, and $e_{\mathrm{for}}, e_{\mathrm{rev}} \in \mathbb{R}^{1876}$ denote metabolic flux, proteome concentration, complex concentration, and enzyme concentration, respectively.

By adopting this formulation, we assumed the following:

1. The total amount of metabolic proteome may not exceed a weight fraction of the dry weight, which is further referred to as the 'proteome budget' and denoted by a scalar (P) in mg/gDW.
2. Each annotated gene in the M-model is transcribed and translated to a unique protein whose molecular weight can be estimated by its protein sequence.
3. Rate constant of a certain enzyme is fixed regardless of reactions. This will greatly reduce the complexity of the problem, especially the nonconvex quadratic programming problem (nonconvex QP) in the later section.
4. Enzyme concentration upper bounds but not forces the respective reaction flux. Enzymes are currently not compartmentalized.

The protein constraints were implemented in the M-model by adding four sets of variables and four sets of constraints. Variables are defined as follows:

1. Protein dilution: protein concentrations (or abundances) in nmol/gDW. Proteins are uniquely defined for each gene in the M-model.
2. Complex formation: complex concentrations in nmol/gDW. We define 'complex' as a unique protein combination that can sufficiently catalyze any single reaction. The list of complexes is obtained by parsing rules in M-model (`parseGeneRule.m`).
3. Enzyme formation: enzyme concentrations in nmol/gDW. We define 'enzyme' as a collection of indifferent complexes that can catalyze a certain reaction. A pair of forwarding and reverse enzymes are added for each enzymatic reaction, and no enzyme is added for spontaneous reactions.
4. Enzyme dilution: One dilution reaction for each forward or reverse enzyme.

Extra constraints were added to the model as follows:

1. Each complex may not exceed the abundance of available protein subunit, according to C (4). C is a matrix containing complex subunit information. Excess proteins are allowed.
2. The sum of forward and reverse enzyme equals the total complex, according to r and B (5). B is Boolean matrix mapping complexes and enzymes and further enzymatic reactions. Vector r denotes the ratio between the rate constant of each complex and the average enzymatic rate constant, or $K_{\mathrm{eff}}^{i} = r_i \cdot K_{\mathrm{eff}}^{\mathrm{avg}}$. We first estimated r as below (`estimateKeffFromMW.m`):

$$r_i^{\mathrm{ori}} = \left(\frac{X_i}{\frac{1}{N}\sum_{i=1}^{N} X_i} \right)^{3/4}, \quad X = C^{-1} d, \tag{9}$$

which is scaled according to the enzymatic surface area as other studies [30, 31].

3. Enzymatic reaction fluxes are restrained by respective forward and reverse enzyme levels through the average rate constant of $65s^{-1}$.
4. Protein concentrations are collectively constrained by the proteome budget P, according to protein molecular weight vector d in mg/nmol. We assumed P being a constant across all growth conditions, and the weight fraction of total proteome to be $600mg/g\mathrm{DW}$. Modelled proteome weight fraction within the total proteome ($\%modelled$) can be estimated using the complete transcript t_{cp} and modelled transcript t, as well as molecular mass vector for complete transcript d_{cp} and for modelled transcript d. Thus, P is approximated as below.

$$\begin{aligned} P &\approx 600 \cdot \%modelled \\ &= 600 \cdot \left(\frac{t^T d}{t_{cp} d_{cp}} \cdot \frac{\text{length}(t)}{\text{length(t)} - \text{length}(\text{NaN } t)} \right), \end{aligned} \tag{10}$$

where NaN t denotes the collection of modelled transcripts of which are not present in t_{cp}. This may happen either for genes in mitochondria and chloroplast genome, or due to the presence of genes in the M-model whose identifiers do not map to any genes in the transcriptomics data. The estimated P are shown by Additional file 5: Fig. S5, and we chose $P = 150$ for this dataset.

We collected a complete protein sequence FASTA file using NCBI genome assembly *Chlamydomonas reinhardtii* v5.5 (`GCF_000002595.2.gbff`), *Chlamydomonas reinhardtii* chloroplast reference genome (NC_005353.1), and *Chlamydomonas reinhardtii* mitochondrial reference genome (NC_001638.1) [53]. This FASTA was constructed by extracting all locus tags and respective protein sequences into a plain text file (`fastaParsing.m`). It was used to calculate the molecular mass of modelled proteins (`calcProteinMM.m`). The PC-model construction processes above are also automated in a MATLAB file as `pcModel.m`. The solving time of PC-FBA is around 0.3 seconds on our device, which is six times more than its respective FBA.

Overlaying processed RNA-Seq data onto PC-model using convex QP

We proposed a methodology to interpret the underlying cellular metabolism for a given RNA-seq data using the PC-model (`overlayMultiomicsData.m`). Assuming the proteome vector is similar to the mRNA vector, we formulated a quadratic objective function, subject to constraints (2)–(8):

$$\min_{v,p,x,e} \left(diag(w)(p-t)\right)^T (p-t), \tag{11}$$

where t denotes the modelled transcript abundance vector, and w is a weighting vector for each transcript. This finds the proteome vector closest to the measured transcriptome while maintaining underlying metabolic feasibility. We defined

$$w_j = \begin{cases} \frac{1}{t_j}, & t_j > 0 \\ 1, & t_j = 0 \\ 0, & \text{NaN } t_j, \end{cases} \tag{12}$$

which increases the weighting of lowly transcribed and un-transcribed genes. This is essential to keep the unexpressed proteins absent from the context-specific model, although other weighting functions might be feasible too. The expression measurement was unavailable for some modelled proteins (NaN t_j), in which case the weighting was assigned to zero. t was scaled to satisfy

$$t^T d = P, \tag{13}$$

which put t and p into the same magnitude. This guaranteed the possibility that objective function (11) might reach zero value from solving.

This is a convex QP problem and can be solved using commercial LP solvers such as Gurobi Optimizer or IBM ILOG CPLEX [57, 58]. The optimization took around five seconds for both solvers. Defining the best-fitting protein vector is solved to be p', and we replaced constraint (7) with constraint (14) for the base PC-model to make it sample-specific.

$$p_j \begin{cases} 0 \le p_j \le p_j^{\text{ub}}, & \text{NaN } t_j \\ (1-s)p_j' \le p_j \le p_j', & otherwise. \end{cases} \tag{14}$$

We added a slack term ($s = 0.02$) into constraint (13) for two practical reasons: leaving the proteome budget for unmeasured transcripts in the unbiased analysis and making the de-bottlenecking algorithm easier to implement. We found that eliminating the slack also indirectly restricts unmeasured proteins due to proteome budget depletion, which would impact the downstream analysis. In our practice, adjusting s within a reasonable range would not significantly affect the result.

Estimating system level enzymatic rate constants using nonconvex QP

Although it is impractical to experimentally measure all enzymatic rate constants, they can be systematically estimated using this PC-model formulation. Given the Q set of samples of the same strain under different conditions, we rewrote objective function (11) into (15). The intuition was to find the single vector r that allows the best-fitting result for all RNA-seq samples:

$$\min_{v^k,p^k,x^k,e^k,r} \sum_{k=1}^{Q} (diag(w')(p^k - t^k))^T (p^k - t^k), \tag{15}$$

subject to constraints (2)–(8) for each of the Q samples. For example, constraint (2) effectively becomes

$$\begin{pmatrix} S_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & S_Q \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_Q \end{pmatrix} = 0. \tag{16}$$

We also simplified the weighting by decorrelating w' with relative abundance, greatly speeding up the computation:

$$w'_j = \begin{cases} 0, & \text{NaN } t_j \\ 1, & otherwise. \end{cases} \tag{17}$$

Vector r was made a variable with the following constraints:

$$\frac{1}{N} \sum_{i=1}^{N} r_i = 1, \tag{18}$$

$$r_i \begin{cases} 0.1 r_i^{\text{ori}} \le r_i \le 1.9 r_i^{\text{ori}}, & \exists t_j^k \ge t_{\text{avg}}^k,\ C_{j,i} > 0 \\ r_i = r_i^{\text{ori}}, & \forall t_j^k < t_{\text{avg}}^k,\ C_{j,i} > 0. \end{cases} \tag{19}$$

The constraints above formulate a nonconvex QP that is Q times the size of the convex QP. By constraint (19), we preserved the original r_i with low subunit abundance, which prevented the solver from prematurely modifying the rate constant. This is inferring that a more comprehensive rate constant estimation can be achieved by including data from various metabolic modes while stacking up data from similar metabolic modes will benefit little; on the other hand, adding each set of data exponentially increases the computational cost. Thus, we first performed hierarchical clustering by MATLAB Statistics and Machine Learning Toolbox to categorize 16 RNA-seq samples into four groups (Additional file 2: Fig S2), which were then used group averages as 'samples' to estimate r using nonconvex QP [59]. Procedures above can be done by `overlayMultiomicsData.m` with 'keffEstimate' option set to true.

The QP was solved by Gurobi Optimizer version 9.1.2, a state-of-the-art LP solver that supports nonconvex bilinear optimizations [57]. The optimization took around 2000 seconds on a laptop with an Apple M1 chip and 16 GB of memory, although the computation time can vary widely for the same problem size with different data samples.

De-bottlenecking and unbiased network analysis

Acknowledging errors and uncertainties in the data and workflow, we applied a de-bottlenecking optimization onto data-specific PC-models to mitigate the effect of a few bottlenecking proteins without shifting the landscape (`proteinDebottleneck.m`). This was done by adding a variable term to the constraint (14), which becomes

$$p_i \begin{cases} 0 \le p_i \le p_i^{\text{ub}}, & \text{NaN } t_i \\ (1-s)p'_i \le p_i \le p'_i + \epsilon_i, & otherwise, \end{cases} \tag{20}$$

where ϵ is a variable vector with an assigned error budget E (also referred to as overexpression budget):

$$\sum \epsilon \le E, \tag{21}$$

while other constraints are the same as the data-specific PC-model, optimized to the objective function (1) using the FBA algorithm, referred to as protein-constrained flux balance analysis or PC-FBA. We conducted this LP by varying error budget values and eventually chose $E = 20$, where the curve of optimal objective values versus E reached a constant slope (see Additional file 7: Table S9). This means no single protein is blocking the objective function, and therefore it was a suitable state for the downstream analysis. The result can also be interpreted as a suggested list of proteins to overexpress. The optimization took roughly 1.0 s using Gurobi Optimizer.

To further understand the metabolic capabilities under each expression data, we used PC-FVA for an unbiased analysis. For each data-specific model, FVA of all metabolic reactions v was done to find v_{min} and v_{max} at the optimal percentages of 0%, 50%, 90%, and 99%.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12934-022-02004-y.

Additional file 1: Fig. S1. The complete record of consistency of simulated proteomes to transcriptomics before and after nonconvex QP by OVERLAY. This is an extended version of Fig. 3ab.

Additional file 2: Fig. S2. Hierarchical clustering result of 16 time-course RNA-seq sample.

Additional file 3: Fig. S3. Histogram of Spearman's ranking coefficient for all metabolic reactions. This supplements Fig. 4, where all Spearman's coefficients are calculated.

Additional file 4: Fig. S4. PC-FVA prediction results for starch synthesis reaction, phospholipase A2 reaction, and fatty acid CoA ligase reaction. This figure can be interpreted using the caption of Fig. 4a.

Additional file 5: Fig. S5. Bar plot of proteome budget estimation using dataset.

Additional file 6: Tables S1 to S3 for results section.

Additional file 7: Tables S4 to S9 for methods section.

Acknowledgements

The authors thank Professor X. Li at Chemical Engineering Department, Queen's University, for his opinions and discussions on nonconvex optimization; the lead author also thanks Z. Wang at Queen's University for her general discussion.

Author contributions

Conceptualization: HY, SD, and LY. Design of the work: HY and LY. Data acquisition: HY and SD. Interpretation: HY. Software creation: HY. Drafting and revising: HY and LY. Directing of research: LY. All authors read and approved the final manuscript.

Funding

This work was funded by the Government of Canada through Genome Canada and Ontario Genomics (OGI-207), the Government of Ontario through an Ontario Research Fund (ORF), and Queen's University.

Availability of data and materials

OVERLAY is available to download at https://github.com/QCSB/OVERLAY-Toolbox. The MATLAB code and pre-computed workspace for the work are available at https://github.com/QCSB/algal-pcFBA.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 27 October 2022 Accepted: 17 December 2022
Published online: 17 January 2023

References

1. Khan S, Fu P. Biotechnological perspectives on algae: a viable option for next generation biofuels. Curr Opin Biotechnol. 2020;62:146–52. https://doi.org/10.1016/j.copbio.2019.09.020.
2. Merchant SS, Kropat J, Liu B, Shaw J, Warakanont J. Tag, you're it! chlamydomonas as a reference organism for understanding algal triacylglycerol accumulation. Curr Opin Biotechnol. 2012;23(3):352–63.
3. Almaraz-Delgado AL, Flores-Uribe J, Pérez-España VH, Salgado-Manjarrez E, Badillo-Corona JA. Production of therapeutic proteins in the chloroplast of Chlamydomonas reinhardtii. AMB Exp. 2014;4(1):1–9.
4. Rasala BA, Mayfield SP. Photosynthetic biomanufacturing in green algae; production of recombinant proteins for industrial, nutritional, and medical uses. Photosynth Res. 2015;123(3):227–39.
5. Scaife MA, Nguyen GTDT, Rico J, Lambert D, Helliwell KE, Smith AG. Establishing Chlamydomonas reinhardtii as an industrial biotechnology host. Plant J. 2015;82(3):532–46. https://doi.org/10.1111/tpj.12781.
6. ...Crozet P, Navarro FJ, Willmund F, Mehrshahi P, Bakowski K, Lauersen KJ, Pérez-Pérez M-E, Auroy P, Gorchs Rovira A, Sauret-Gueto S, Niemeyer J, Spaniol B, Theis J, Trösch R, Westrich L-D, Vavitsas K, Baier T, Hübner W, de Carpentier F, Cassarini M, Danon A, Henri J, Marchand CH, de Mia M, Sarkissian K, Baulcombe DC, Peltier G, Crespo J-L, Kruse O, Jensen P-E, Schroda M, Smith AG, Lemaire SD. Birth of a photosynthetic chassis: a MoClo Toolkit enabling synthetic biology in the microalga Chlamydomonas reinhardtii. ACS Synth Biol. 2018;7(9):2074–86. https://doi.org/10.1021/acssynbio.8b00251.
7. Scranton MA, Ostrand JT, Fields FJ, Mayfield SP. Chlamydomonas as a model for biofuels and bio-products production. Plant J. 2015;82(3):523–31. https://doi.org/10.1111/tpj.12780.
8. Rengel R, Smith RT, Haslam RP, Sayanova O, Vila M, León R. Overexpression of acetyl-CoA synthetase (ACS) enhances the biosynthesis of neutral lipids and starch in the green microalga *Chlamydomonas reinhardtii*. Algal Res. 2018;31:183–93. https://doi.org/10.1016/j.algal.2018.02.009.
9. Yunus IS, Wichmann J, Wördenweber R, Lauersen KJ, Kruse O, Jones PR. Synthetic metabolic pathways for photobiological conversion of CO2 into hydrocarbon fuel. Metab Eng. 2018;49:201–11. https://doi.org/10.1016/j.ymben.2018.08.008.
10. Bogaert KA, Perez E, Rumin J, Giltay A, Carone M, Coosemans N, Radoux M, Eppe G, Levine RD, Remacle F, et al. Metabolic, physiological, and transcriptomics analysis of batch cultures of the green microalga chlamydomonas grown on different acetate concentrations. Cells. 2019;8(11):1367.
11. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdóttir HS, Wachowiak J, Keating SM, Vlasov V, Magnusdóttir S, Ng CY, Preciat G, Žagare A, Chan SHJ, Aurich MK, Clancy CM, Modamio J, Sauls JT, Noronha A, Bordbar A, Cousins B, El Assal DC, Valcarcel LV, Apaolaza I, Ghaderi S, Ahookhosh M, Ben Guebila M, Kostromins A, Sompairac N, Le HM, Ma D, Sun Y, Wang L, Yurkovich JT, Oliveira MAP, Vuong PT, El Assal LP, Kuperstein I, Zinovyev A, Hinton HS, Bryant WA, Aragón Artacho FJ, Planes FJ, Stalidzans E, Maass A, Vempala S, Hucka M, Saunders MA, Maranas CD, Lewis NE, Sauter T, Palsson BO, Thiele I, Fleming RMT. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nature Protocols. 2019;14(3):639–702. https://doi.org/10.1038/s41596-018-0098-2.
12. Opdam S, Richelle A, Kellman B, Li S, Zielinski DC, Lewis NE. A systematic evaluation of methods for tailoring genome-scale metabolic models. Cell Syst. 2017;4(3):318–3296. https://doi.org/10.1016/j.cels.2017.01.010.
13. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008;4(5):1000082.
14. Wang Y, Eddy JA, Price ND. Reconstruction of genome-scale metabolic models for 126 human tissues using mcadre. BMC Syst Biol. 2012;6(1):1–16.
15. Imam S, Schäuble S, Valenzuela J, López García de Lomana A, Carter W, Price ND, Baliga NS. A refined genome-scale reconstruction of Chlamydomonas metabolism provides a platform for systems-level analyses. Plant J. 2015;84(6):1239–56. https://doi.org/10.1111/tpj.13059.
16. Røkke GB, Hohmann-Marriott MF, Almaas E. An adjustable algal chloroplast plug-and-play model for genome-scale metabolic models. PLOS ONE. 2020;15(2):0229408. https://doi.org/10.1371/journal.pone.0229408.
17. Yurkovich JT, Yang L, Palsson BO. Systems-level physiology of the human red blood cell is computed from metabolic and macromolecular mechanisms. BioRxiv. 2019. https://doi.org/10.1101/797258.
18. Liebermeister W, Noor E, Flamholz A, Davidi D, Bernhardt J, Milo R. Visual account of protein investment in cellular functions. Proc Nat Acad Sci. 2014;111(23):8488–93. https://doi.org/10.1073/pnas.1314810111.
19. Arnon DI, Whatley FR, Allen MB. Assimilatory power in photosynthesis. Science. 1958;127(3305):1026–34. https://doi.org/10.1126/science.127.3305.1026.
20. Chaux F, Peltier G, Johnson X. A security network in PSI photoprotection: regulation of photosynthetic control, NPQ and O2 photoreduction by cyclic electron flow. Front Plant Sci. 2015;6.
21. Ibarra RU, Edwards JS, Palsson BO. Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 2002;420(6912):186–9.
22. Fang X, Lloyd CJ, Palsson BO. Reconstructing organisms in silico: genome-scale models and their emerging applications. Nat Rev Microbiol. 2020;18(12):731–43.
23. Yang L, Mih N, Anand A, Park JH, Tan J, Yurkovich JT, Monk JM, Lloyd CJ, Sandberg TE, Seo SW, et al. Cellular responses to reactive oxygen species are predicted from molecular mechanisms. Proc Nat Acad Sci. 2019;116(28):14368–73.
24. Smith RT, Gilmour DJ. The influence of exogenous organic carbon assimilation and photoperiod on the carbon and lipid metabolism of *Chlamydomonas reinhardtii*. Algal Res. 2018;31:122–37.
25. Roach T, Sedoud A, Krieger-Liszkay A. Acetate in mixotrophic growth medium affects photosystem ii in *Chlamydomonas reinhardtii* and protects against photoinhibition. Biochim Biophys Acta BBA Bioenerg. 2013;10:1183–90.
26. Johnson X, Alric J. Central carbon metabolism and electron transport in *Chlamydomonas reinhardtii*: metabolic constraints for carbon partitioning between oil and starch. Eukaryotic Cell. 2013;12(6):776–93. https://doi.org/10.1128/EC.00318-12.

27. Gfeller RP, Gibbs M. Fermentative metabolism of *Chlamydomonas reinhardtii*. Plant Physiol. 1984;75(1):212–8.
28. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T. Overflow metabolism in *Escherichia coli* results from efficient proteome allocation. Nature. 2015;528(7580):99–104. https://doi.org/10.1038/nature15765.
29. Goodenough U, Blaby I, Casero D, Gallaher SD, Goodson C, Johnson S, Lee J-H, Merchant SS, Pellegrini M, Roth R, Rusch J, Singh M, Umen JG, Weiss TL, Wulan T. The path to triacylglyceride obesity in the sta6 strain of *Chlamydomonas reinhardtii*. Eukaryotic Cell. 2014;13(5):591–613. https://doi.org/10.1128/EC.00013-14.
30. O'brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol. 2013;9(1):693.
31. Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ, Liu JK, Palsson BO. Cobrame: a computational framework for genome-scale models of metabolism and gene expression. PLoS Comput Biol. 2018;14(7):1006302.
32. Hawkins C, Ginzburg D, Zhao K, Dwyer W, Xue B, Xu A, Rice S, Cole B, Paley S, Karp P, Rhee SY. Plant metabolic network 15: a resource of genome-wide metabolism databases for 126 plants and algae. J Integr Plant Biol. 2021;63(11):1888–905. https://doi.org/10.1111/jipb.13163.
33. Zheng H-Q, Chiang-Hsieh Y-F, Chien C-H, Hsu B-KJ, Liu T-L, Chen C-NN, Chang W-C. AlgaePath: comprehensive analysis of metabolic pathways using transcript abundance data from next-generation sequencing in green algae. BMC Genomics. 2014;15(1):196. https://doi.org/10.1186/1471-2164-15-196.
34. Bulté L, Gans P, Rebéillé F, Wollman F-A. ATP control on state transitions in vivo in *Chlamydomonas reinhardtii*. Biochim Biophys Acta BBA Bioenerg. 1990;1020(1):72–80. https://doi.org/10.1016/0005-2728(90)90095-L.
35. Iwai M, Ikeda K, Shimojima M, Ohta H. Enhancement of extraplastidic oil synthesis in *Chlamydomonas reinhardtii* using a type-2 diacylglycerol acyltransferase with a phosphorus starvation-inducible promoter. Plant biotechnology journal. 2014;12(6):808–19.
36. Fukuda S, Hirasawa E, Takemura T, Takahashi S, Chokshi K, Pancha I, Tanaka K, Imamura S. Accelerated triacylglycerol production without growth inhibition by overexpression of a glycerol-3-phosphate acyltransferase in the unicellular red alga cyanidioschyzon merolae. Sci Rep. 2018;8(1):1–12.
37. Zhang Y, Pan Y, Ding W, Hu H, Liu J. Lipid production is more than doubled by manipulating a diacylglycerol acyltransferase in algae. GCB Bioenergy. 2021;13(1):185–200.
38. Li Y, Han D, Hu G, Sommerfeld M, Hu Q. Inhibition of starch synthesis results in overproduction of lipids in *Chlamydomonas reinhardtii*. Biotechnol Bioeng. 2010;107(2):258–68.
39. Work VH, Radakovits R, Jinkerson RE, Meuser JE, Elliott LG, Vinyard DJ, Laurens LM, Dismukes GC, Posewitz MC. Increased lipid accumulation in the *Chlamydomonas reinhardtii* sta7-10 starchless isoamylase mutant and increased carbohydrate synthesis in complemented strains. Eukaryotic cell. 2010;9(8):1251–61.
40. Shin YS, Jeong J, Nguyen THT, Kim JYH, Jin E, Sim SJ. Targeted knockout of phospholipase a2 to increase lipid productivity in *Chlamydomonas reinhardtii* for biodiesel production. Bioresour Technol. 2019;271:368–74.
41. Kato A, Takatani N, Ikeda K, Maeda S-I, Omata T. Removal of the product from the culture medium strongly enhances free fatty acid production by genetically engineered *Synechococcus elongatus*. Biotechnol Biofuels. 2017;10(1):1–8.
42. LaPanse AJ, Krishnan A, Posewitz MC. Adaptive laboratory evolution for algal strain improvement: methodologies and applications. Algal Res. 2021;53:102122. 10.1016/j.algal.2020.102122.
43. Acevedo-Rocha CG, Gronenberg LS, Mack M, Commichau FM, Genee HJ. Microbial cell factories for the sustainable manufacturing of b vitamins. Curr Opin Biotechnol. 2019;56:18–29.
44. Yuan S-F, Alper HS. Metabolic engineering of microbial cell factories for production of nutraceuticals. Microbial cell factories. 2019;18(1):1–11.
45. Vidotti AD, Riaño-Pachón DM, Mattiello L, Giraldi LA, Winck FV, Franco TT. Analysis of autotrophic, mixotrophic and heterotrophic phenotypes in the microalgae chlorella vulgaris using time-resolved proteomics and transcriptomics approaches. Algal Res. 2020;51:102060.
46. Zuñiga C, Li C-T, Huelsman T, Levering J, Zielinski DC, McConnell BO, Long CP, Knoshaug EP, Guarnieri MT, Antoniewicz MR, et al. Genome-scale metabolic model for the green alga chlorella vulgaris utex 395 accurately predicts phenotypes under autotrophic, heterotrophic, and mixotrophic growth conditions. Plant Physiol. 2016;172(1):589–602.
47. Palsson BO, Yurkovich JT. Is the kinetome conserved? Mol Syst Biol. 2022;18(2):10782.
48. Aurich MK, Fleming RM, Thiele I. Metabotools: a comprehensive toolbox for analysis of genome-scale metabolic models. Front Physiol. 2016;7:327.
49. Salvy P, Fengos G, Ataman M, Pathier T, Soh KC, Hatzimanikatis V. pytfa and mattfa: a python package and a matlab toolbox for thermodynamics-based flux analysis. Bioinformatics. 2019;35(1):167–9.
50. MATLAB: Version 9.11.0.1769968 (R2021b). The MathWorks Inc., Natick, Massachusetts, United State 2021.
51. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021;49(D1):545–51. https://doi.org/10.1093/nar/gkaa970.
52. Norsigian CJ, Pusarla N, McConn JL, Yurkovich JT, Dräger A, Palsson BO, King Z. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. Nucleic Acids Res. 2020;48(D1):402–6. https://doi.org/10.1093/nar/gkz1054.
53. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau D, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick B, Pruitt K, Sherry S. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021;50(D1):20–6. https://doi.org/10.1093/nar/gkab1112.
54. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
55. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):008. https://doi.org/10.1093/gigascience/giab008.
56. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nat Methods. 2017;14(4):417–9. https://doi.org/10.1038/nmeth.4197.
57. Gurobi Optimization L. Gurobi optimizer reference manual 2022. https://www.gurobi.com.
58. Cplex II. V12. 9: User's Manual for CPLEX. International Business Machines Corporation 2017.
59. MATLAB: Statistics and Machine Learning Toolbox. The MathWorks Inc. 2021. https://www.mathworks.com/help/stats/.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.