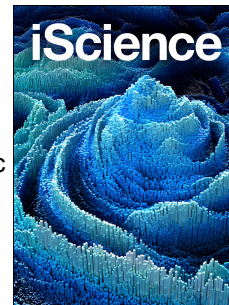


Journal Pre-proof



A hybrid flux balance analysis and machine learning pipeline elucidates the metabolic response of cyanobacteria to different growth conditions

Supreeta Vijayakumar, Pattanathu Kaja-Mohideen Sheikh Mujibur Rahman, Claudio Angione

PII: S2589-0042(20)31015-4

DOI: <https://doi.org/10.1016/j.isci.2020.101818>

Reference: ISCI 101818

To appear in: *ISCIENCE*

Received Date: 26 May 2020

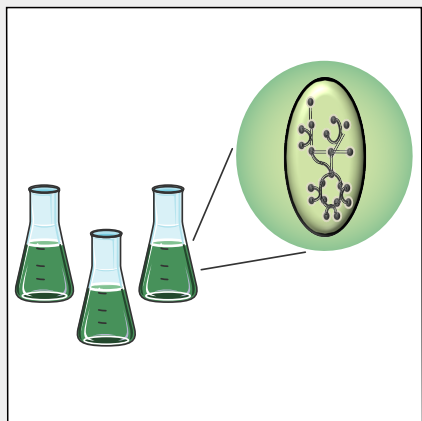
Revised Date: 23 October 2020

Accepted Date: 13 November 2020

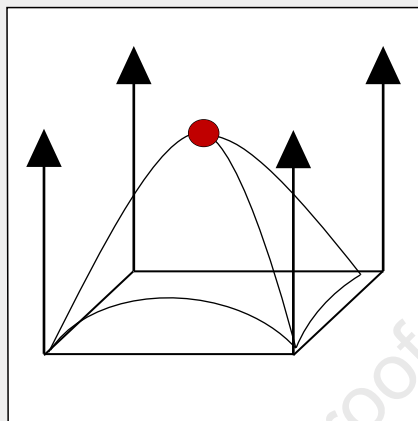
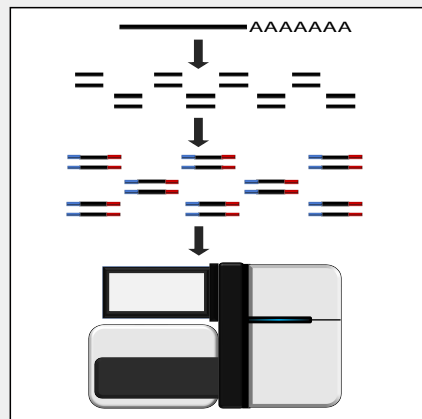
Please cite this article as: Vijayakumar, S., Kaja-Mohideen Sheikh Mujibur Rahman, P., Angione, C., A hybrid flux balance analysis and machine learning pipeline elucidates the metabolic response of cyanobacteria to different growth conditions, *ISCIENCE* (2020), doi: <https://doi.org/10.1016/j.isci.2020.101818>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

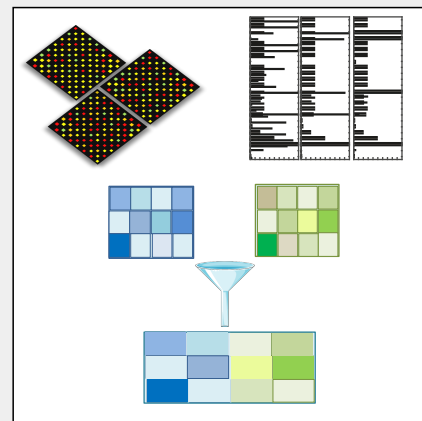
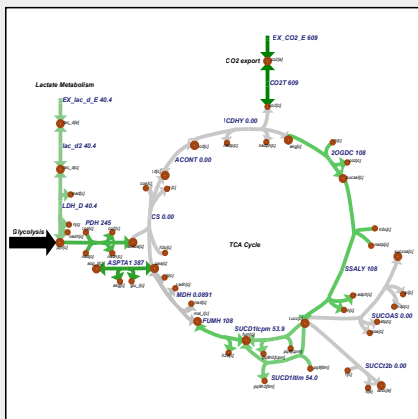
© 2020 The Author(s).



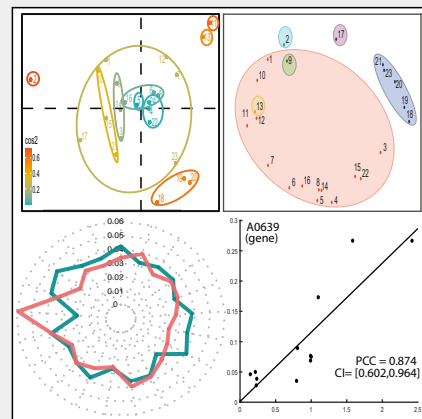
Cell culture and
RNA sequencing



Regularized bi-level
flux balance analysis



Multiomic data integration
and machine learning



A hybrid flux balance analysis and machine learning pipeline elucidates the metabolic response of cyanobacteria to different growth conditions

Supreeta Vijayakumar^a, Pattanathu Kaja-Mohideen Sheikh Mujibur Rahman^{b,c}, Claudio Angione^{*a,d,e}

^aDepartment of Computer Science and Information Systems, Teesside University, Middlesbrough, North Yorkshire, TS1 3BX, UK

^bCentre for Enzyme Innovation, Institute of Biological and Biomedical Sciences, School of Biological Sciences, University of Portsmouth, Portsmouth, Hampshire, PO1 2UP, UK

^cTara Biologics, Woking, Surrey, GU21 6BP, UK

^dCentre for Digital Innovation, Teesside University, Middlesbrough, TS1 3BX, UK

^eHealthcare Innovation Centre, Teesside University, Middlesbrough, TS1 3BX, UK

Keywords: Genome-scale metabolic models, Machine learning, Flux balance analysis, Cyanobacteria, Microalgae, Multi-omics

1. Summary

Machine learning has recently emerged as a promising tool for inferring multi-omic relationships in biological systems. At the same time, genome-scale metabolic models (GSMMs) can be integrated with such multi-omic data to refine phenotypic predictions. In this work, we use a multi-omic machine learning pipeline to analyze a GSMM of *Synechococcus* sp. PCC 7002, a cyanobacterium with large potential to produce renewable biofuels. We use regularized flux balance analysis (FBA) to observe flux response between conditions across photosynthesis and energy metabolism. We then incorporate principal component analysis (PCA), *k*-means clustering and LASSO regularization to reduce dimensionality and extract key cross-omic features. Our results suggest that combining metabolic modeling with machine learning elucidates mechanisms used by cyanobacteria to cope with fluctuations in light intensity and salinity that cannot be detected using transcriptomics alone. Furthermore, GSMMs introduce critical mechanistic details that improve the performance of omic-based machine learning methods.

2. Introduction

In the field of systems biology, several approaches have been proposed to capture the enormous complexity of biological systems by utilizing mathematical modeling and computational methods, with the goal of amalgamating the information required to build and refine predictive models. The challenges

presented by such an undertaking are numerous and persistent owing to the size, format, scale and variation of the disparate data types. Amongst these, metabolism is currently the only biological layer that can be modeled genome-wide (O'Brien et al., 2015; Haas et al., 2017). Constraint-based reconstruction and analysis (COBRA) methods are commonly used to express metabolic flux through biochemical reactions based on knowledge of reaction stoichiometry. Flux balance analysis (FBA) is particularly suitable for modeling metabolic networks at the genome scale, as the definition of kinetic parameters and metabolite concentrations is not a key requisite.

In recent years, genome-scale metabolic models (GSMMs) have been integrated with multiple data types, including omics, codon usage, enzyme costs, and limited resource availability (Abedpour and Kollmann, 2015; Opdam et al., 2017; Kashaf et al., 2017; Wortel et al., 2018; Tian and Reed, 2018; Angione, 2019). This serves to exploit the large volume of experimental data being generated from high-throughput omics technologies. In doing so, additional constraints can be applied during FBA in order to shrink the solution space (Reed, 2012), thus providing a more accurate representation of metabolic capability as a greater number of factors can be considered to explain cellular behavior. This can prove useful in refining phenotypic predictions across various environmental conditions (Vijayakumar et al., 2017; Sánchez et al., 2017; van der Ark et al., 2017; Angione, 2018), and can predict steps to engineer an organism in a way that optimizes the production of certain metabolites, which is highly applicable in many fields of industrial biotechnology including the production of biofuels, biosurfactants and pharmaceuticals (Angione et al., 2015; Dougherty et al., 2017; Huang et al., 2017; Fatma et al., 2018; Occhipinti et al., 2018).

*Lead Contact and Corresponding Author
 Email address: c.angione@tees.ac.uk (Claudio Angione*)

2.1. Modelling and Metabolic Engineering in Cyanobacteria

Cyanobacteria is a phylum of oxygenic, phototrophic microalgae that need to adapt to constant fluctuations in temperature, salinity, light intensity (or irradiance), and nutrient availability, amongst other factors (Montgomery, 2017; Blanco-Ameijeiras et al., 2018; Gunde-Cimerman et al., 2018). Metabolic engineering is helping to develop cyanobacteria into photoautotrophic biofactories that can act as production hosts (chassis) for alcohols, carbohydrates, organic acids, fatty acid derivatives, isoprenoids, and many other chemicals (Noreña-Caro and Benton, 2018). However, since such approaches are generally designed with heterotrophic organisms in mind, the metabolic features unique to photoautotrophs must be considered, e.g. pathways relating to photosynthesis and CO₂ fixation (Carroll et al., 2018).

Synechococcus sp. PCC 7002 is a fast-growing cyanobacterium that flourishes in both freshwater and marine environments, owing to its ability to tolerate high light intensity and a wide range of salinities. Harnessing the properties of cyanobacteria has become an important goal in recent years owing to their potential to serve as biocatalysts for the production of renewable biofuels (Hendry et al., 2016). Metabolic modeling of two cyanobacteria, *Arthrospira* and *Synechocystis*, has successfully characterized the use of photosynthetic electron transport components in different light conditions (Toyoshima et al., 2020).

In an industrial setting, *Synechococcus* sp. PCC 7002 has been recommended as the ideal chassis for the mass cultivation of microalgae for biotechnological applications owing to its ease of genetic manipulation as well as its tolerance for high salinity, light intensity and temperature (Pade and Hagemann, 2014; Clark et al., 2018). These are highly desirable traits in micro-algae as they enable cultures to maintain a rapid growth rate in open raceway ponds as well as in photobioreactors, which operate at high temperatures (Ruffing et al., 2016). Within the *Synechococcus* genus, a comparative analysis of slow- and fast-growing strains in terms of their active reactions under phototrophic conditions has been proposed to better inform their development into production hosts (i.e. strain optimization), primarily through maximizing their growth rates (Hendry et al., 2019). In a recent study, Song et al. (2015) completed an integrative analysis of metabolic and gene co-expression networks in *Synechococcus* sp. PCC 7002 by integrating expression data from either continuous cultures or existing studies into a GSMM and deriving fluxes using E-Fmin flux minimization (Song et al., 2014) and MOMA (Segre et al., 2002). Further studies have examined temporal variations in response to varying light intensity and associated conditional dependencies (Rügen et al., 2015; Reimers et al., 2016). These

need to be accounted for as constraints in genome-scale metabolic models designed to simulate the phototrophic growth in cyanobacteria over diurnal cycles and tackle issues associated with resource allocation (Vijayakumar and Angione, 2017).

Genome-scale isotopic non-stationary metabolic flux analysis (INST-MFA) has been utilized to estimate internal metabolic fluxes more accurately in *Synechococcus elongatus* UTEX 2973, towards the aim of establishing factors affecting phototrophic metabolism under optimal growth conditions (Hendry et al., 2019). Similarly, MOMA and INST-13C MFA were used to establish carbon partitioning at intracellular branching points in the central metabolism of a glycogen-deficient *Synechococcus* sp. PCC 7002 mutant (Hendry et al., 2017). Such models benefit significantly from constraints designed using experimentally-measured uptake or growth rates for the identification of alternative reactions responsible for the synthesis of metabolites and differences in pathway recruitment and utilization (e.g. for carbon conversion to biomass).

The current state of strain-specific metabolic modeling in cyanobacteria, and the potential of fluxomic data and metabolic engineering have been recently discussed elsewhere (Angermayr et al., 2015; Oliver et al., 2016; Hendry et al., 2020; Luan et al., 2020; Babel and Young, 2020; Mukherjee et al., 2020; Hitchcock et al., 2020). Within the *Synechococcus* genus, a comparative analysis of slow- and fast-growing strains in terms of their active reactions under phototrophic conditions has been proposed to better inform their development into production hosts (i.e. strain optimization), primarily through maximizing their growth rates (Hendry et al., 2019). A number of novel, non-model strains of *Synechococcus* that have been developed include *Synechococcus* UTEX 2973 (Yu et al., 2015), PCC 11801 (Jaiswal et al., 2018), PCC 11802 (Damini et al., 2020), PCC 11901 (Włodarczyk et al., 2019) and BDU 130192 (Ahmad et al., 2020).

2.2. Multi-omic Data Integration in Microalgae

In recent years, synthetic biology has facilitated the modeling of biological processes for genetic engineering. Using synthetic biology tools, algal strains have been designed according to highly-specific environmental conditions and yield requirements. Synthetic biologists have been successful in assembling genetic material and manipulating the lipid content of microalgae, as well as maximizing biomass accumulation and biofuel yield (Jagadevan et al., 2018). These results are promising for the biofuels industry from the microalgal perspective (Randhawa et al., 2017). In the context of microalgae, the alteration of lipid biosynthesis pathways through the induction of a stress response to a change in environment (such as temperature, nutrient limitation, salinity) is a common prac-

tice to enhance the production of target compounds, including those that are used to produce workable bio-fuels (Rawat et al., 2013).

Omics approaches have made a significant contribution to the understanding of the molecular processes of microalgae. Furthermore, the discoveries that omics studies have made - e.g. the identification of genes involved in specific processes - may be vital to the engineering of enhanced microalgae. Through the understanding of transcription levels and gene activation data gathered from transcriptomics, the effectiveness of genetic alterations can be measured as previously achieved for other organisms, allowing for optimization of the target product. For example, if the new gene insert is operating at its optimum, the transcriptomic data should show an increase in the mRNA of the target gene when compared to the wild type (Randhawa et al., 2017). Based on genomic and transcriptomic data, Wang et al. (2019) recently identified a series of neutral sites on the chromosome of *Synechococcus* sp. PCC 7002 for the introduction of novel heterologous genes or pathways without disruption.

Omics techniques can also provide valuable insights into alterations of lipid synthesis pathways that occur as a result of stress conditions in microalgae. Metabolomics studies assess the low molecular metabolite end products and are highly indicative of response to stresses. Previously, global transcriptomic, proteomic and metabolomic analyses have aided in identifying adaptations for cyanobacterial salt tolerance in *Synechocystis* sp. PCC 6803 (Pandhal et al., 2009; Wang et al., 2016). An omic-combination approach would allow for optimization of algal engineering, as the data gathered from transcriptomics should show an increase in transcription in the gene of interest that coincides with a reduction in metabolism caused by stress (such as nutrient limitation) highlighted by metabolomics - if the expression of the gene of interest is linked to a metabolic process. The application of omic studies can not only ascertain the effectiveness of any genetic modification, but can also be used to optimize the scale-up process. With the use of spatial and temporal omics studies of systems such as raceways used for algal growth, a deeper understanding of how algae will perform in various areas of the raceway can be gained, allowing for process optimization (Randhawa et al., 2017).

2.3. Aims and Objectives

In this work, we present a pipeline combining metabolic modeling with statistical and machine learning tools (Figure 1) for analyzing a genome-scale metabolic model of the cyanobacterium *Synechococcus* sp. PCC 7002. We characterize *Synechococcus* adaptation mechanisms using an updated GSMM of *iSyp702* containing 728 genes (Hendry et al., 2016), implementing multi-objective flux balance analysis

with quadratic regularization. We then apply machine learning techniques to identify functionally important genes and reactions. These include PCA, k -means clustering and LASSO regression, which serve not only to identify biologically-significant gene transcripts and fluxes, but also to relate these features more effectively to growth-promoting or growth-limiting conditions provided by the initial expression data.

Our goal is to show whether, in a predictive setting, features derived from the metabolic model can add meaningful information to the features derived from the transcriptomic data (Zampieri et al., 2019). Therefore, for each method, we will consider the predictions yielded using three sets of features: (i) gene expression only, (ii) fluxes only, and (iii) gene expression and fluxes combined.

Through LASSO regression, we find that using flux rates to predict growth rates is more effective than using gene transcript values alone. This suggests that GSMMs provide critical details in terms of stoichiometry, and the involvement of genes in reactions determines the rate of cellular phototrophic growth as well as other modes of energy utilization (e.g. heterotrophy, mixotrophy) in various environmental conditions.

3. Results

As highlighted above, our goal is to reconnect metabolism to growth and other cellular objectives using a data-driven multiview approach that yields biologically reasonable predictions. The results of PCA and k -means clustering for flux data are included in Figure 2, whereas results of these analyses for gene transcript data in isolation and gene transcript data combined with fluxes are detailed in Figure 3. Additionally, the results of the pathway-wide analysis of principal components are provided in Figure 4. The highest positive/negative Pearson correlation coefficient (PCC) values for transcript- and flux-only datasets are given in Figure 5, whereas mean absolute PCCs for metabolic subsystems or pathways are shown in Figure 6(c) with the number of reactions in each subsystem specified in Figure 6(d). A list of all nonzero LASSO coefficients and the top ten positive/negative correlation coefficients are given in the Supplemental Information, with the full calculation of these coefficients provided in Supplementary Data 2. An interpretation of the results for each technique used is provided below.

3.1. Regularized Flux Balance Analysis

Compared to transcriptomics, metabolic flux data modeled at the genome-scale provide a more comprehensive, condition-specific view of the phenotype. Therefore, we mapped each RNA-Seq profile measured in 24 growth conditions to a *Synechococcus*

GSMM, and we employed a regularized FBA to obtain condition-specific flux distributions (see Transparent Methods in Supplemental Information). To calculate the flux rates more accurately for each condition, several lower and upper bounds were adjusted prior to performing FBA, according to specific growth media and other requirements described for each growth condition (Ludwig and Bryant, 2011, 2012a,b). The full details of these growth conditions (including composition of growth media, optical density at the time of cell harvestation, mode of energy utilization, availability of oxygen/carbon dioxide, light intensity, salinity and temperature) are listed in Table S1. The full specification of constraints for each growth condition is given in Supplementary Data 3.

From the transcriptomic studies listed in Table S1, there were a number of genes that were not transcribed in the control condition but were transcribed specifically under perturbed conditions. Many of these genes have yet to be assigned a particular functional category or encode hypothetical proteins, but many more have been linked to specific pathways and compounds and some have been associated with the adaptation of *Synechococcus* sp. PCC 7002 to atypical environmental or growth conditions.

As shown in Figure 2, apart from the standard control, the highest fluxes through the ATP maintenance reaction (when ATP maintenance is set as the secondary objective) were among conditions that limit growth, such as phosphate limitation, 30°C and oxidative stress. In dark anoxic, low salinity, heat shock, phosphate limitation and mixotrophic conditions, there was no flux for the biomass reaction. However, for all the objective pairs, the flux through the biomass during high light intensity and OD 0.4 (optical density) was higher than the control condition (0.19 mmol gDW⁻¹ h⁻¹ and 0.093 mmol gDW⁻¹ h⁻¹ respectively, compared to 0.053 mmol gDW⁻¹ h⁻¹). The biomass is likely to be higher at OD 0.4 than OD 0.7 due to adjustment of the photon constraint, which allowed for more transmission of light at lower OD. Apart from the dark anoxic and low O₂ conditions, all fluxes through photosystem II were negligible, but the fluxes through photosystem I were still maintained (0.058 mmol gDW⁻¹ h⁻¹ in the high light intensity condition as opposed to 0.016 mmol gDW⁻¹ h⁻¹ in the control).

When photosystem I was set as the secondary objective, a low amount of flux through ATP maintenance reaction was retained in phosphate-limited, heat shock, and low salinity conditions (approximately 0.0002 mmol gDW⁻¹ h⁻¹). When photosystem II was set as the secondary objective, the highest fluxes through the photosystem II reaction were given by the phosphate limitation, mixotrophic, and low salinity conditions (0.016 mmol gDW⁻¹ h⁻¹).

Lack of light is likely to be the greatest contribut-

ing factor to decrease in growth as low oxygen concentration does not seem to stunt growth, as the proportional decrease in biomass was lower relative to the standard control conditions. On the other hand, there appears to be little to no flux for the biomass or the photosystem I reaction in the dark conditions. This is supported by Vu et al. (2013), who reported that lower yields under dark conditions may be due to the limited generation of energy (ATP) and reductant (NADPH) from glycogen in the absence of photoautotrophic growth. When optical density was varied through the batch growth conditions (OD 0.4, 1.0, 3.0, and 5.0), the transmission of light through the cultures decreased as the dry cell weight (DCW) increased. Equal reduction in transcript levels for the photosynthetic apparatus was previously observed in all macronutrient-limited conditions studied (Ludwig and Bryant, 2012a). We infer that heat shock, mixotrophic growth and phosphate limitation have the largest effect on reducing growth rate, as there was a complete impairment of biomass production predicted across all of our objectives for these conditions. This is in line with reported findings (Ludwig and Bryant, 2012a), where perturbations caused by phosphate limitation had a greater impact on the global transcription pattern than observed for high irradiance or dark treatments.

Synechococcus sp. PCC 7002 is known to possess one of the greatest tolerances for high light intensity among cyanobacteria (with an upper limit of approximately 2000 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$) (Xiong et al., 2015). This was evident from our predictions for all three pairs of objectives, where flux through the biomass pathway during high light intensity was slightly higher than the control condition (0.192 mmol gDW⁻¹ h⁻¹ compared to 0.053 mmol gDW⁻¹ h⁻¹). Although the fluxes through photosystem II were disrupted, the fluxes through photosystem I were still maintained (0.058 mmol gDW⁻¹ h⁻¹ in the high light intensity condition as opposed to 0.016 mmol gDW⁻¹ h⁻¹ in the control). Heat shock resulted in no fluxes through any of the four reactions within all three objective pairs. It was previously reported that transcript levels for genes encoding photosystem I decreased slightly in cells grown at high salinity and remained constant at low salinity (Ludwig and Bryant, 2012b). On the other hand, it was found that transcript levels for genes encoding photosystem II did not change in response to fluctuations in salinity (Ludwig and Bryant, 2012b).

3.2. Multi-omic Principal Component Analysis (PCA)

It can be argued that analyzing single-omic data alone has limited relevance in the context of metabolic processes, since it does not capture the full complexity

of the phenotype in relation to environmental variability. The hybrid approach proposed in this work connects transcriptomic and fluxomic data using a data-driven multi-view approach that supports machine learning algorithms to yield more accurate predictions (Culley et al., 2020). Considering the vast dimensionality of multi-omic models, the identification of biologically-meaningful information can prove to be challenging. As a non-parametric statistical technique, principal component analysis (PCA) was incorporated into our workflow for identifying patterns and genes/reactions responsible for the most variance in the datasets (Brunk et al., 2016).

The principal component analysis indicates the proportion of variance exhibited by fluxes in the first two dimensions for each objective pair. For all three pairs of objectives in Figure 2(a-c), over 68.31% of the variance can be explained by the first dimension when considering flux data alone. As shown in plot (a), the high light intensity condition contributed the highest score for the first dimension, accounting for the vast majority of the variance. On the other hand, plots (b) and (c) showed that high light intensity, phosphate limitation, mixotrophic, low salinity were the highest scoring conditions in the first dimension. For the second dimension, the highest score was given by high salinity, iron limitation, urea, 30°C and oxidative stress in plot (a), and high light intensity, phosphate limitation, mixotrophic and low salinity conditions in plots (b) and (c).

When considering only the gene transcript data (Figure 3(a)), the combined proportion of variance that could be accounted for by the first two dimensions was vastly reduced (only 35.18% compared to 75.70-86.27% for flux data). The conditions with the largest scores for the first dimension were sulfate and iron limitation, followed by oxidative stress, 30°C, and high salinity. Once again, sulfate and iron limitation were the highest in the second dimension, along with phosphate limitation, nitrogen limitation, dark anoxic/oxic and the last phase of the batch growth (OD 5.0).

When using a combined dataset of both gene transcripts and fluxes (Figure 3(c), (e) and (g)), the total proportion of variance that could be explained in two dimensions for all three objective pairs was lower than using transcript data alone (31.43-32.07%). The highest scores were given by iron and sulfate limitation in the first dimension and by dark oxic, dark anoxic, OD 5.0, iron limitation, and sulfate limitation in the second dimension.

A full list of gene transcript and calculated fluxes are included in Supplementary Data 1. For a list of the top ten contributions of genes and reactions to the principal components, we refer the reader to the Supplemental Information (Tables S3, S4, and S5).

3.2.1. Pathway-Level Analysis of Principal Components

To further examine the most metabolically significant pathways or cellular processes, we also performed a pathway-level PCA whilst categorizing genes and reaction by their main function. Owing to the varying number of reactions within each pathway, both the pathway sum and average contribution to the variance from the first two principal components were calculated. In Figure 4(a-b), the sum of all contributions to variance within each pathway or COG (Cluster of Orthologous Groups) category is summarized. For the gene transcripts (a), the COGs with the highest sum of variance within the first two principal components were poorly characterized (with general or unknown function). It can be observed that for each pair of flux objectives in (b), the pathways that contribute the most to the first and second components were similar: cofactor and vitamin metabolism, nucleotide metabolism, energy metabolism, lipid metabolism, amino acid metabolism, carbohydrate metabolism, and transport metabolism. These pathways can be directly linked to cellular growth since many of their products are biomass precursors or compounds that can be catabolized to produce energy, i.e. carbohydrates, proteins and fats.

The radar plots in Figure 4(c-f) depict the average contributions to the variance within each pathway for the first and second principal components. The average contribution was higher for reactions (fluxes) than genes since the number of genes in each COG category was greater than the number of reactions in each subsystem of the GSMM. The pathways with the highest average contributions for gene transcripts were nucleotide/amino acid metabolism in the first component and chromatin structure and dynamics in the second component. For all three objective pairs, amino acid, aminoacyl-tRNA and peptidoglycan biosynthesis were relevant in the first component (with an average contribution between 0.26-0.3). For the second component, the pathways with the largest contribution varied for each objective. Coenzyme and thiamine metabolism both had an average contribution greater than 0.45 for the second component in relation to ATP and photosystem I fluxes, with the addition of hydrogen metabolism for ATP, and pyridine metabolism for photosystem I. On the other hand, purine and nucleotide metabolism had the highest contributions in the second component for photosystem II. However, many of these pathways contained only one or two reactions, which caused these results to be skewed; for example, the purine metabolism pathway had an average contribution of 1.36, but only contains one nucleotide phosphodiesterase reaction (PDE2).

Finally, to characterize the PCA in the context of

single reactions, we analyzed the principal component coordinates for all growth conditions against different reaction fluxes selected from the top ten contributors to the variance in each of the three objective pairs (see Figure 4(g-l)). These plots confirmed that a large part of the variance could be explained in the first principal component, as the first component coordinates showed a near-perfect correlation with flux (> 0.99). The second principal component displayed a less consistent but still strongly positive correlation between coordinates and flux values. The set of reactions with the highest contributions to variance in the first and second components were completely different, but three main functional categories could be identified among these reactions. IODP and GARFT can be linked to nucleotide metabolism, ASPTA1 and ILEABC to amino acid metabolism and PDH and NADH.PQ9t1m to energy metabolism, respectively.

3.3. Clustering

k -means is a clustering algorithm that computes clusters whilst iteratively minimizing the sum of squared Euclidean distances between each observation and its respective cluster mean (McLachlan et al., 2008). To assess whether the generated multi-omic datasets could identify clusters of growth conditions according to the respective omic responses, we applied k -means to the set of 24 growth conditions, considering gene expression, flux rates, and the combined expression/flux dataset.

For the flux data (Figure 2(d-f)), the partitioning of k -means clusters varied depending on the pair of FBA objectives for which fluxes were calculated. Following silhouette analysis, the number of clusters set for plots (d), (e), (f) in Figure 2 and plots (b), (d), (f) and (h) in Figure 3 was $k = 6$. The full list of members of each cluster is reported in the Supplemental Information.

When combining both transcript and flux data (Figure 3(d), (f) and (h)), the clusters formed were less distinct. This suggests that fluxes could help to contribute more biological insights into metabolic reactions (through the metabolic network) that are not available in the transcriptomic data. Nevertheless, through the k -means analysis with transcripts-only and the combined multi-omic dataset of transcripts and fluxes, we conclude that clustering techniques benefit from analyzing the flux and transcript datasets in isolation rather than combining them, as this avoids an increase in data dimensionality that cannot be easily reduced.

In most of the k -means plots in Figure 3, changing the objective pair used for FBA did not result in a significant difference in the clusters formed. However, there was sometimes a demarcation between clusters of conditions that limited growth (e.g. low light, nutrient limitation) and those that promoted

growth (e.g. high light availability, nutrient supplementation). In some instances, certain conditions (e.g. heat shock, high light intensity, iron limitation) were isolated within a single cluster. Furthermore, reducing the number of dimensions in the data following PCA could serve to reduce noise and make the definition of clusters even clearer.

3.4. LASSO Regression

Regression-based algorithms can ascertain a mapping function, given a number of continuous output variables y , and a number of real-valued or discrete input variables x . For each input variable, a coefficient is estimated with linear regression, determining the importance of the input variable towards predicting the output variable. In our study, the least absolute shrinkage and selection operator (LASSO) was used to select a subset of the input variables by minimizing the number of nonzero coefficients. It employs L1 regularization, which penalizes the sum of absolute values of all the coefficients; this sets the coefficients of unnecessary or recursive features equal to zero, resulting in a sparser matrix (Tibshirani, 1996). The regularization enables the identification of important predictors, elimination of redundant predictions, and generation of shrinkage estimates with lower predictive errors.

The LASSO regression identified genes and reactions in the model that are strongly related to *in-vivo* growth rates through the retention of non-zero predictor coefficients (see Table S2). The full calculation of LASSO coefficients is provided in Supplementary Data 2.

A complete list of non-zero predictors retained by LASSO for each dataset is given in the Supplemental Information. The functional classifications were provided by CyanOmics in the case of genes (CY Category and CY Sub Category) or the *subsystems* field within the model GSMM for reactions. In Tables S6 and S8 (where both transcript and flux datasets or only the gene transcripts are considered), the same non-zero predictors (genes) were retained, irrespective of the objective-pair used for FBA. The genes yielding positive coefficients were associated with photosynthesis and respiration or post-translational modification of proteins. When applying the LASSO algorithm to the flux-only dataset (see Table S7), the non-zero coefficients retained were primarily related to the metabolism of nucleotides, cofactors and vitamins and pathways relating to energy generation, such as carbohydrate and amino-acid metabolic pathways. Such co-factors are often comprised of metal ions, for which there are numerous transport and exchange reactions, e.g. cobalt and manganese.

3.5. Correlation Analysis

The Pearson correlation coefficients were calculated to ascertain the strength of the association between transcripts and/or flux rates and growth across different conditions (see Table S2). The absolute Pearson correlation coefficients were sorted in descending order, and the top ten positive/negative correlation coefficients for each dataset are listed in the Supplemental Information. Figure 5 shows the highest positive and negative Pearson correlation coefficients for transcript- and flux-only datasets. The gene A0639 encodes a phycocyanin-associated phycobilisome rod-core linker polypeptide, which is an important component of the photosynthetic apparatus. This confirms that photosynthesis and energy metabolism are directly correlated with cellular growth.

For the Biomass-ATP maintenance flux objectives, the selection of reactions encoding succinate dehydrogenase in the cytoplasmic and thylakoid membranes (SUCD1Icpm and SUCD1Itlm) demonstrated the importance of the TCA cycle in the generation of energy for biomass accumulation. To illustrate this, a comparison of flux values between the nitrogen-limited and urea-supplemented growth conditions is provided in Figure 6(a) and (b). It can be seen in Table S2 that the limitation or supplementation of a nitrogen source had a direct effect on the growth rate. Cyanobacteria have long been known to possess a unique TCA cycle where an alternative reaction homologous to 2OGDH is used to convert alpha-ketoglutarate into succinyl semialdehyde (2OGDC), which is subsequently converted into succinate (via succinate-semialdehyde dehydrogenase, i.e. SSALY) (Zhang and Bryant, 2011; Steinhauser et al., 2012). 2OGDC and SSALY were found to carry negligible flux under phototrophic conditions in (Hendry et al., 2016), which was supported by the flux values derived for the standard control for our simulations ($3.599 \text{ mmol gDW}^{-1} \text{ h}^{-1}$ for both reactions), as well as the growth-limiting conditions such as nitrogen limitation (Figure 6(a)). This suggests that succinate dehydrogenation plays an important role in growth; it is known to still take place in dark, anoxic conditions (McNeely et al., 2010) where there is an increased flux towards succinate during the dark period, driving ATP production through respiratory electron transport (Sarkar et al., 2019).

Figure 6(c) shows the mean absolute PCC values among reactions within each subsystem, whereas Figure 6(d) shows the number of reactions within a given range of PCC values for each subsystem/pathway listed in the model, to account for the differing number of reactions in each pathway in the model. The pathway with the largest mean absolute correlation across all the flux objectives was the folate metabolism. In *Synechocystis* sp. PCC 6803, folate is synthesized from chorismate and is known to be important for cellular processes such as DNA replication, repair, and

methylation, in addition to being a vital precursor for the biosynthesis of certain amino acids, co-factors, nucleotides, and tRNAs (Mills et al., 2020). The highest mean PCC in the biomass-ATP maintenance flux pair corresponded with the proline biosynthesis pathway/reaction. Interestingly, it has been found that proline accumulation is highly induced in stress conditions in cyanobacteria, especially high salinity (increased NaCl concentration) since it plays a role in osmoprotection, antioxidative defense and signaling (Hayat et al., 2012; Pingkhanont et al., 2019). Across all flux objectives, the majority of reactions in Figure 6(d) have a correlation value between $[-0.1, 0.1[$ or $[0.3, 0.5[$ - implying most reactions had little to no significant correlation or a moderately positive correlation. The strongest positive correlation values $[0.5, 0.7[$ were found between the Biomass - ATP maintenance flux pair and the growth rates. It can be seen that these reactions were classified under various pathways, i.e. carbohydrate, amino acid, energy, transport- and exchange-related metabolism.

4. Discussion

In this work, we showed how using a hybrid multi-view approach with multi-omic data and machine learning to yield metabolically significant fluxes enabled the identification of trends in data that were not apparent using solely transcriptomic data. We used condition-specific flux balance analysis to obtain flux distributions with norm-2 regularized bi-level optimization.

The flux distributions obtained for four key reactions showed clear differences in pathway activity across the various conditions and also between the three pairs of objectives used. When comparing the results across the types of datasets used, it is clear that complex metabolic and phenotypic outcomes as a result of adaptation to a changing environment are difficult to predict from gene expression alone. Condition-specific metabolic models within a machine learning framework allowed for the detection of coordinated responses shared between different data types, as well as the variation in responses across different growth conditions. Although a large number of studies express the maximization of biomass as the only objective when performing FBA, it is imperative to recognize that in reality most organisms have multiple objectives to satisfy. It has been well established that the activity of biosynthetic and energy-generating pathways increases with the growth rate (Bernstein et al., 2014), which led us to implement multi-level regularized optimization in our pipeline, considering more than one objective function.

Specifically, when calculating the flux distribution across conditions, biomass was chosen as the primary objective, while the secondary objective was set to

ATP maintenance, photosystem I or photosystem II, to reflect the main cellular goals of cyanobacteria. Biomass was chosen as a primary objective to represent the maximization of growth rate and cellular yields (Feist and Palsson, 2010; Yuan et al., 2016; Lakshmanan et al., 2019), which is a critical consideration for the production of biofuels by cyanobacteria as this informs the substrate uptake rates and maintenance requirements that indicate fundamental cellular growth requirements. The chosen secondary objectives are key pathways involved in energy metabolism during photosynthesis. Simulating the cost of ATP maintenance can help to examine the energy required for sustaining metabolic activity even in the absence of growth. The incorporation of the photoexcitation reactions occurring within photosystems I and II served to characterize how flux under various conditions reflects the light harvesting and energy transfer via photon absorption through these complexes. Thus, solving the quadratic optimization problem for multiple pairs of objectives helped to resolve trade-offs by considering the conditions and constraints affecting each of these objectives (Sajitz-Hermstein and Nikoloski, 2016; Occhipinti et al., 2020).

Our results suggest that it is worth using model-generated flux data that incorporates transcriptomics to conduct machine learning analyses. The flux data was initially informed by transcriptomic data as the condition-specific gene expression profiles were generated by combining them with a baseline GSMM for *Synechococcus* sp. PCC 7002; in this way, gene transcripts already constituted an important component of the flux balance analysis. Furthermore, reducing the number of dimensions in the data following PCA can serve to reduce noise and make the definition of clusters even clearer. In addition to this, a reduced set of predictors were identified as being related to growth as a result of the LASSO regularization. Specifically, the identification of reactions by LASSO as key features which are of potential use for the prediction of growth rates supports the inclusion of metabolic fluxes as features for future applications of regression techniques with smaller, more concise sets of flux data.

The reactions identified as being strongly correlated with the growth rate in the flux datasets (SUCD1Itlm, SUCD1Icpm, ME2) suggest that fluxes can help to gain more biological insights into machine learning analyses. Since a different, unrelated set of genes displayed a strong correlation with the growth rate, it is evident that analyzing both transcriptomic and fluxomic data provides a more complete picture of cyanobacterial metabolism than single-omic analyses. In particular, the role of metal transport pathways in cyanobacteria was significant since they are highly relevant in the context of photosynthesis. The detection of latent, biologically significant patterns and adaptive mechanisms to fluctuations in light intensity and

salinity elucidates the maintenance of metabolic efficiency at the cellular level, as well as the attainment of multiple cellular objectives.

Algal engineering supplemented with data from multi-omic studies can contribute to informing the scale-up of these processes. Such multi-omic data are sensitive enough to detect the effect of stress on metabolism. Metabolic engineers could apply this pipeline to test more strategies *in-silico* when developing the optimal production host, or to analyze multi-omic outputs (both independently and in combination with other omic data). In this regard, the use of transcriptomic data to characterize fluxomic predictions elucidates many of the unique mechanisms employed by *Synechococcus* sp. PCC 7002 when adapting to changes in light intensity, salinity, and other conditions. In the case of cyanobacteria, we also emphasize the importance of assessing model inputs in accordance with specific growth conditions prior to conducting flux balance analysis. These contribute to the organism's underlying objective of maintaining metabolic efficiency for phototrophic growth and light-dependent photosynthesis. As a result of predicting and classifying metabolic profiles in various growth conditions, our approach sheds light on the cross-omic mechanisms of its adaptation process, which enables survival across a wide range of environments and stress conditions.

4.1. Limitations of the Study

The availability of exact measurements for the various growth conditions could yield more precise flux predictions. For example, in our case, the exact photon absorbance of the *Synechococcus* sp. PCC 7002 cultures was not available. Hence, the photon uptake constraints were approximated using dry cell weight and photon consumption based on the availability of light. Likewise, the setting of nutrient uptake rates was approximated based on data provided by *in-vivo* experiments rather than measured directly.

Furthermore, in this study we adopted linear transformations and linear methods, where possible. This was with the goal of maximizing the biological interpretability of the predictions, using quadratic terms for regularization only. However, different dimensionality reduction or clustering methods could be implemented, e.g. to elucidate any further non-linear relationship among the omic elements.

Finally, there is further potential for other types of omic data to be integrated into the model (e.g. from proteomic or metabolomic datasets). It is expected that integrating further omic datasets, i.e. further data views in our multi-view machine learning setting, could produce even more detailed insights into metabolic adaptations, or better support existing findings derived from transcriptomic and fluxomic data.

5. Methods

All methods used in this study can be found in the accompanying Transparent Methods section in the Supplemental Information file.

6. Resource Availability

6.1. Lead Contact

Dr Claudio Angione, email: c.angione@tees.ac.uk

6.2. Materials Availability

The study did not generate new unique reagents or other materials.

6.3. Data and Code Availability

The complete source code of our pipeline is freely available on GitHub at <https://github.com/Angione-Lab/Synechococcus7002-metabolic-modelling>.

The RNA-Seq data integrated into the model were from transcriptomic datasets uploaded by Ludwig and Bryant (2011, 2012a,b). These were acquired directly from the Cyanomics dataset (Yang et al., 2015), but have since been made available on the NCBI Sequence Read Archive (SRA: “SRP007372”; “SRP013965”; “SRP066851”).

7. Acknowledgements

We would like to thank Dr John Hendry from Penn State University for providing advice on fixing boundary constraints for the *Synechococcus* sp. PCC 7002 GSMM. We would also like to thank Dr Guido Zampieri from the University of Padova for discussions on the methodology. CA and PR received funding from BBSRC, grants CBMNet-PoC-D0156. PR would like to thank Research England for funding support through the Expanding Excellence in England (E3) scheme. CA would like to acknowledge the support from UKRI Research England’s THYME project.

8. Author Contributions

Conceptualization, S.V. and C.A.; Data curation, S.V.; Formal analysis, S.V.; Funding Acquisition, P.R. and C.A.; Investigation, S.V.; Methodology, S.V. and C.A.; Project administration, C.A.; Resources, P.R. and C.A.; Software, S.V. and C.A.; Supervision, C.A.; Validation, S.V.; Visualization, S.V., C.A.; Writing — original draft, S.V., P.R. and C.A.; Writing — reviewing and editing, S.V. and C.A.

9. Declaration of Interests

The authors declare no competing interests.

10. List of Figures

Figure 1: **The multi-omic machine learning pipeline for prediction and classification of *Synechococcus* metabolic features.** (i) RNA sequencing data obtained from *Synechococcus* sp. PCC 7002 cells grown under 23 growth conditions (Ludwig and Bryant, 2011, 2012a,b). (ii) Data downloaded from CyanOmics (Yang et al., 2015). (iii) Starting from a model recently published by Hendry et al. (2016), the condition-specific GSMMs of *Synechococcus* sp. PCC 7002 are generated by integrating omics data, and three pairs of objectives are optimized for each condition-specific model. (iv) Bi-level regularized FBA is conducted using quadratic programming to compute regularized flux distributions. (v) Transcriptomic, fluxomic and multi-omic (combination) datasets are preprocessed for the machine learning analysis. (vi) PCA, k -means clustering, LASSO regression and correlation analysis are applied to identify latent cross-omic patterns in the metabolic adaptation mechanisms. These techniques are applied and compared across three sets of omic features: gene transcripts, condition-specific flux rates, and a combination of both omics.

Figure 2: **Flux distributions with fluxomic PCA and k -means.** (Top): Flux distributions in the 24 growth conditions considered in this study. Flux distributions for four key reactions: ATP maintenance, photosystem I, photosystem II and biomass when running FBA using three different pairs of objectives (indicated at the bottom of each plot). Conditions 1-24 correspond to those detailed in Table S1. To better visualize the differences in flux between conditions, flux values were normalized by dividing by the maximal flux (i.e. the flux value for the control condition) for that reaction across all conditions. The full list of flux rates is reported in the Supplemental Information. Regularized FBA correctly predicts reduced growth in sub-optimal conditions and the highest biomass flux is given by the high light condition. (Bottom): Fluxomic principal component analysis (PCA) and k -means. PCA (panels a-c) and k -means clustering (panels d-f) were conducted using the entire flux distribution (742 reactions). Panels (a) and (d) are associated with Biomass - ATP maintenance fluxes, (b) and (e) with Biomass - Photosystem I fluxes and (c) and (f) with Biomass - Photosystem II fluxes. For PCA plots, growth conditions are colored according to their \cos^2 value, which indicates the contribution of the first two components to the squared distance of each condition to the origin (Abdi and Williams, 2010). The higher the \cos^2 value, the greater the proportion of contribution to the total distance, meaning that the importance of the principal components is greater for that condition. For k -means, data are clustered by condition (where the colors of ellipses represent different clusters) and the number of clusters ($k = 6$) was selected following silhouette analysis. Due to co-location of conditions in the two-dimensional plot, not all overlapping points are visible, but the cluster associated with each condition is labeled. The full list of growth conditions and their respective k -means clusters are reported in the Supplemental Information.

Figure 3: **Transcriptomic and multi-omic principal component analysis (PCA) and k -means.** PCA and k -means clustering conducted for 3187 gene transcripts (a-b) and both transcripts and fluxes, with biomass - ATP maintenance (c-d), biomass - photosystem I (e-f) and biomass - photosystem II (g-h) as objective pairs. In the PCA plots (a,c,e and g), growth conditions are colored according to their \cos^2 value, indicating the contribution of the first two components to the squared distance of each condition to the origin (Abdi and Williams, 2010). For k -means, data are clustered by condition (where the color of the ellipses represents different clusters) and the number of clusters was determined following silhouette analysis ($k = 6$). When compared to using the transcriptomic dataset alone, the combined proportion of variance for PCA in the first two dimensions was slightly higher when gene transcript data was used in isolation than when it was combined with fluxes. For k -means clustering, the change in objective pair used for FBA did not result in a significant difference in the clusters formed. However, there is a clear demarcation between clusters of conditions that limit growth (e.g. low light, sulfate limitation) and those which promote growth (e.g. high light, nitrate supplementation).

Figure 4: **Pathway-based PCA to identify pathway and reaction contribution to variance across conditions.** (a-b) Component sums by pathway. Principal component contributions summed across reactions within each COG category/pathway, in order to decompose the metabolic function of the main contributors to variance. The total percentage of contribution to variance of the first two principal components is also given for each dataset: (a) Gene transcripts, (b) Fluxes calculated with objective functions Biomass - ATP maintenance, Biomass - Photosystem I, and Biomass - Photosystem II. (c-f) Component contribution by pathway. Average principal component contributions within each pathway, calculated across all gene transcripts (c) and fluxes for each objective function pair: (d) Biomass - ATP maintenance, (e) Biomass - Photosystem I, (f) Biomass - Photosystem II. (g-l) Interpreting PCA coordinates with fluxes. The Pearson correlation coefficient (PCC) was calculated between the principal component coordinates and the flux values across the conditions in three pairs of objectives. The coordinates for the first principal component (x axis) and flux (y axis) across the 24 conditions were plotted for the following reactions: (g) inorganic diphosphatase (IODP) for Biomass - ATP maintenance, (h) aspartate transaminase (ASPTA1) for Biomass - Photosystem I, and (i) pyruvate dehydrogenase (PDH) for Biomass - Photosystem II. For the second principal component, the reactions were: (j) L-isoleucine transport via ABC system (ILEABC) for Biomass - ATP maintenance, (k) NADH dehydrogenase type II in the thylakoid membrane (NADH_PQ9t1m) for Biomass - Photosystem I, and (l) phosphoribosylglycinamide formyltransferase (GARFT) for Biomass - Photosystem II. The PCC with their respective 95% confidence intervals (CI) are displayed within each plot.

Figure 5: **Top PCCs with their respective 95% confidence intervals (CI) between gene transcript/reaction flux data (x axis) and growth rates (y axis).** Left to right: Black - top correlated genes, Green - top correlated reactions when maximizing Biomass-ATP maintenance flux, Red - top correlated reactions when maximizing Biomass-Photosystem I flux, Blue - top correlated reactions when maximizing Biomass-Photosystem II flux. Tables S9, S10, S11, S12, and S13 list the top ten genes/reactions in the dataset that are positively or negatively correlated with growth and their respective PCC values. Additional figures of top ten genes/reactions and their respective PCC values are provided in the Supplemental Information (Figures S1 - S4).

Figure 6: **Flux map comparison and PCC values by subsystem.** Comparison between flux values in the TCA cycle for (a) nitrogen limitation and (b) urea supplementation. The SUCD1It1m and SUCD1Icpm reactions encoding succinate dehydrogenase were identified as having a strong positive correlation with the growth rate for the Biomass - ATP maintenance objective pair. (c) Mean absolute Pearson correlation coefficients (PCC) calculated between 12 experimental growth rates and their corresponding condition-specific GSMM reaction fluxes within each metabolic subsystem/pathway in the *Synechococcus* sp. PCC 7002 GSMM. The highest mean absolute correlations were identified for folate metabolism, proline and amino acid biosynthesis. (d) Reactions within each model subsystem sorted into classes of PCC values obtained between growth rates and flux rates in each objective pair (Biomass - ATP maintenance, Biomass - Photosystem I, Biomass - Photosystem II). The Biomass - ATP maintenance pair yielded the highest positive PCC values [0.5, 0.7] for reactions within the carbohydrate, amino acid, energy, transport, and exchange metabolic pathways.

11. Supplemental Information

Supplemental Information. Document containing Transparent Methods, Supplemental Results, Figures S1-S4 and Tables S1-S13.

Supplementary Data S1 - Full Transcript and Flux Data, Related to Figures 2 and 3. Microsoft Excel workbook containing RPKM transcript reads for each growth condition, fold changes calculated for transcripts, complete flux rates for three objective pairs and fold changes calculated for flux rates.

Supplementary Data S2 - Growth Rates, LASSO and Pearson Correlation, Related to Figure 5. Microsoft Excel workbook containing: (i) A list of 24 growth conditions with growth rates and doubling times calculated for 11 conditions relative to cultures grown under standard conditions. (ii) Fitted least-squares regression beta coefficients and mean squared errors for LASSO conducted with gene transcripts and/or flux rates and growth rates. (iii) Pearson correlation coefficients calculated between gene transcripts and/or flux rates and growth rates listed with p -values, lower and upper bounds according to the 95% confidence interval and coefficients of determination (R^2). (iv) Mean absolute Pearson correlation coefficients within each group of reactions in each subsystem of the GSMM.

Supplementary Data S3 - Photon Unit Conversions and Flux Constraints, Related to Figure 2. Microsoft Excel workbook containing: (i) Calculation of photon uptake rate for FBA in different growth rates. (ii) Full list of modified flux constraints (model.lb and model.ub) used to perform FBA.

References

- Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley interdisciplinary reviews: computational statistics 2, 433–459.
- Abedpour, N., Kollmann, M., 2015. Resource constrained flux balance analysis predicts selective pressure on the global structure of metabolic networks. BMC systems biology 9, 88.
- Ahmad, A., Pathania, R., Srivastava, S., 2020. Biochemical characteristics and a genome-scale metabolic model of an indian euryhaline cyanobacterium with high polyglucan content. Metabolites 10, 177.
- Angermayr, S.A., Rovira, A.G., Hellingwerf, K.J., 2015. Metabolic engineering of cyanobacteria for the synthesis of commodity products. Trends in biotechnology 33, 352–361.
- Angione, C., 2018. Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism. Bioinformatics 34, 494–501.
- Angione, C., 2019. Human systems biology and metabolic modelling: A review—from disease metabolism to precision medicine. BioMed Research International 2019.
- Angione, C., Costanza, J., Carapezza, G., Lió, P., Nicosia, G., 2015. Multi-target analysis and design of mitochondrial metabolism. PLoS one 10, e0133825.
- van der Ark, K.C., van Heck, R.G., Dos Santos, V.A.M., Belzer, C., de Vos, W.M., 2017. More than just a gut feeling: constraint-based genome-scale metabolic models for predicting functions of human intestinal microbes. Microbiome 5, 78.
- Babele, P.K., Young, J.D., 2020. Applications of stable isotope-based metabolomics and fluxomics toward synthetic biology of cyanobacteria. Wiley Interdisciplinary Reviews: Systems Biology and Medicine 12, e1472.
- Bernstein, H.C., Konopka, A., Melnicki, M.R., Hill, E.A., Kucek, L.A., Zhang, S., Shen, G., Bryant, D.A., Beliaev, A.S., 2014. Effect of mono-and dichromatic light quality on growth rates and photosynthetic performance of *Synechococcus* sp. pcc 7002. Frontiers in microbiology 5, 488.
- Blanco-Ameijeiras, S., Moisset, S.A., Trimborn, S., Campbell, D.A., Heiden, J.P., Hassler, C.S., 2018. Elemental stoichiometry and photophysiology regulation of *Synechococcus* sp. pcc 7002 under increasing severity of chronic iron limitation. Plant and Cell Physiology 59, 1803–1816.
- Brunk, E., George, K.W., Alonso-Gutierrez, J., Thompson, M., Baidoo, E., Wang, G., Petzold, C.J., McCloskey, D., Monk, J., Yang, L., et al., 2016. Characterizing strain variation in engineered *e. coli* using a multi-omics-based workflow. Cell systems 2, 335–346.
- Carroll, A.L., Case, A.E., Zhang, A., Atsumi, S., 2018. Metabolic engineering tools in model cyanobacteria. Metabolic engineering .
- Clark, R.L., McGinley, L.L., Purdy, H.M., Korosh, T.C., Reed, J.L., Root, T.W., Pfleger, B.F., 2018. Light-optimized growth of cyanobacterial cultures: Growth phases and productivity of biomass and secreted molecules in light-limited batch growth. Metabolic engineering 47, 230–242.
- Culley, C., Vijayakumar, S., Zampieri, G., Angione, C., 2020. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. Proceedings of the National Academy of Sciences 117, 18869–18879.
- Damini, J., Annesha, S., Shinjinee, S., Swati, M., Pakrasi, H.B., Wangikar, P.P., 2020. A novel cyanobacterium *synechococcus elongatus* pcc 11802 has distinct genomic and metabolomic characteristics compared to its neighbor pcc 11801. Scientific reports 10.
- Dougherty, B.V., Moutinho Jr, T.J., Papin, J., 2017. Accelerating the drug development pipeline with genome-scale metabolic network reconstructions. volume 6. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.
- Fatma, Z., Hartman, H., Poolman, M.G., Fell, D.A., Srivastava, S., Shakeel, T., Yazdani, S.S., 2018. Model-assisted metabolic engineering of *Escherichia coli* for long chain alkane and alcohol production. Metabolic engineering 46, 1–12.
- Feist, A.M., Palsson, B.O., 2010. The biomass objective function. Current opinion in microbiology 13, 344–349.
- Gunde-Cimerman, N., Plemenitaš, A., Oren, A., 2018. Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations. FEMS microbiology reviews 42, 353–375.
- Haas, R., Zelezniak, A., Iacovacci, J., Kamrad, S., Townsend, S., Ralser, M., 2017. Designing and interpreting ‘multi-omic’ experiments that may change our understanding of biology. Current Opinion in Systems Biology 6, 37–45.
- Hayat, S., Hayat, Q., Alyemeni, M.N., Wani, A.S., Pichtel, J., Ahmad, A., 2012. Role of proline under changing environments: a review. Plant signaling & behavior 7, 1456–1466.
- Hendry, J.I., Bandyopadhyay, A., Srinivasan, S., Pakrasi, H.B., Maranas, C.D., 2020. Metabolic model guided strain design of cyanobacteria. Current opinion in biotechnology 64, 17–23.
- Hendry, J.I., Gopalakrishnan, S., Ungerer, J., Pakrasi, H.B., Tang, Y.J., Maranas, C.D., 2019. Genome-scale fluxome of *Synechococcus elongatus* utex 2973 using transient ^{13}C -labeling data. Plant physiology 179, 761–769.
- Hendry, J.I., Prasanna, C., Ma, F., Möllers, K.B., Jaiswal, D., Digmurti, M., Allen, D.K., Frigaard, N.U., Dasgupta, S.,

- Wangikar, P.P., 2017. Rerouting of carbon flux in a glyco-gen mutant of cyanobacteria assessed via isotopically non-stationary ^{13}C metabolic flux analysis. *Biotechnology and bioengineering* 114, 2298–2308.
- Hendry, J.I., Prasannan, C.B., Joshi, A., Dasgupta, S., Wangikar, P.P., 2016. Metabolic model of *Synechococcus* sp. pcc 7002: Prediction of flux distribution and network modification for enhanced biofuel production. *Bioresource Technology* 213, 190–197.
- Hitchcock, A., Hunter, C.N., Canniffe, D.P., 2020. Progress and challenges in engineering cyanobacteria as chassis for light-driven biotechnology. *Microbial biotechnology* 13, 363–367.
- Huang, Z., Lee, D.Y., Yoon, S., 2017. Quantitative intracellular flux modeling and applications in biotherapeutic development and production using cho cell cultures. *Biotechnology and bioengineering* 114, 2717–2728.
- Jagadevan, S., Banerjee, A., Banerjee, C., Guria, C., Tiwari, R., Baweja, M., Shukla, P., 2018. Recent developments in synthetic biology and metabolic engineering in microalgae towards biofuel production. *Biotechnology for biofuels* 11, 185.
- Jaiswal, D., Sengupta, A., Sohoni, S., Sengupta, S., Phadnavis, A.G., Pakrasi, H.B., Wangikar, P.P., 2018. Genome features and biochemical characteristics of a robust, fast growing and naturally transformable cyanobacterium *synechococcus elongatus* pcc 11801 isolated from india. *Scientific reports* 8, 1–13.
- Kashaf, S.S., Angione, C., Lió, P., 2017. Making life difficult for *clostridium difficile*: augmenting the pathogen’s metabolic model with transcriptomic and codon usage data for better therapeutic target characterization. *BMC systems biology* 11, 25.
- Lakshmanan, M., Long, S., Ang, K.S., Lewis, N.E., Lee, D.Y., 2019. On the impact of biomass composition in constraint-based flux analysis. *bioRxiv* , 652040.
- Luan, G., Zhang, S., Lu, X., 2020. Engineering cyanobacteria chassis cells toward more efficient photosynthesis. *Current opinion in biotechnology* 62, 1–6.
- Ludwig, M., Bryant, D.A., 2011. Transcription profiling of the model cyanobacterium *Synechococcus* sp. strain pcc 7002 by next-gen (solid™) sequencing of cDNA. *Frontiers in microbiology* 2, 41.
- Ludwig, M., Bryant, D.A., 2012a. Acclimation of the global transcriptome of the cyanobacterium *Synechococcus* sp. strain pcc 7002 to nutrient limitations and different nitrogen sources. *Frontiers in microbiology* 3, 145.
- Ludwig, M., Bryant, D.A., 2012b. *Synechococcus* sp. strain pcc 7002 transcriptome: acclimation to temperature, salinity, oxidative stress, and mixotrophic growth conditions. *Frontiers in microbiology* 3, 354.
- McLachlan, G.J., Bean, R.W., Ng, S.K., 2008. Clustering, in: *Bioinformatics*. Springer, pp. 423–439.
- McNeely, K., Xu, Y., Bennette, N., Bryant, D.A., Dismukes, G.C., 2010. Redirecting reductant flux into hydrogen production via metabolic engineering of fermentative carbon metabolism in a cyanobacterium. *Applied and environmental microbiology* 76, 5032–5038.
- Mills, L.A., McCormick, A.J., Lea-Smith, D.J., 2020. Current knowledge and recent advances in understanding metabolism of the model cyanobacterium *Synechocystis* sp. pcc 6803. *Bioscience reports* 40, BSR20193325.
- Montgomery, B.L., 2017. Seeing new light: recent insights into the occurrence and regulation of chromatic acclimation in cyanobacteria. *Current opinion in plant biology* 37, 18–23.
- Mukherjee, B., Madhu, S., Wangikar, P.P., 2020. The role of systems biology in developing non-model cyanobacteria as hosts for chemical production. *Current opinion in biotechnology* 64, 62–69.
- Noreña-Caro, D., Benton, M.G., 2018. Cyanobacteria as photoautotrophic biofactories of high-value chemicals. *Journal of CO₂ Utilization* 28, 335–366.
- O’Brien, E.J., Monk, J.M., Palsson, B.O., 2015. Using genome-scale models to predict biological capabilities. *Cell* 161, 971–987.
- Occhipinti, A., Eyassu, F., Rahman, T.J., Rahman, P.K., Angione, C., 2018. In silico engineering of *Pseudomonas* metabolism reveals new biomarkers for increased biosurfactant production. *PeerJ* 6, e6046.
- Occhipinti, A., Hamadi, Y., Kugler, H., Wintersteiger, C., Yordanov, B., Angione, C., 2020. Discovering essential multiple gene effects through large scale optimization: an application to human cancer metabolism. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* .
- Oliver, N.J., Rabinovitch-Deere, C.A., Carroll, A.L., Nozzi, N.E., Case, A.E., Atsumi, S., 2016. Cyanobacterial metabolic engineering for biofuel and chemical production. *Current opinion in chemical biology* 35, 43–50.
- Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D.C., Lewis, N.E., 2017. A systematic evaluation of methods for tailoring genome-scale metabolic models. *Cell systems* 4, 318–329.
- Pade, N., Hagemann, M., 2014. Salt acclimation of cyanobacteria and their application in biotechnology. *Life* 5, 25–49.
- Pandhal, J., Noirel, J., Wright, P.C., Biggs, C.A., 2009. A systems biology approach to investigate the response of *synechocystis* sp. pcc6803 to a high salt environment. *Saline Systems* 5, 1–16.
- Pingkanont, P., Tarasuntisuk, S., Hibino, T., Kageyama, H., Waditee-Sirisattha, R., 2019. Expression of a stress-responsive gene cluster for mycosporine-2-glycine confers oxidative stress tolerance in *Synechococcus elongatus* pcc 7942. *FEMS microbiology letters* 366, fnz115.
- Randhawa, K.S., Relph, L.E., Armstrong, M.C., Rahman, P.K., 2017. Biofuel production: tapping into microalgae despite challenges. *Biofuels* 8, 261–271.
- Rawat, I., Kumar, R.R., Mutanda, T., Bux, F., 2013. Biodiesel from microalgae: a critical evaluation from laboratory to large scale production. *Applied energy* 103, 444–467.
- Reed, J.L., 2012. Shrinking the metabolic solution space using experimental datasets. *PLoS Computational Biology* 8, e1002662.
- Reimers, A.M., Knoop, H., Bockmayr, A., Steuer, R., 2016. Evaluating the stoichiometric and energetic constraints of cyanobacterial diurnal growth. *arXiv preprint arXiv:1610.06859* .
- Ruffing, A.M., Jensen, T.J., Strickland, L.M., 2016. Genetic tools for advancement of *Synechococcus* sp. pcc 7002 as a cyanobacterial chassis. *Microbial Cell Factories* 15, 190.
- Rügen, M., Bockmayr, A., Steuer, R., 2015. Elucidating temporal resource allocation and diurnal dynamics in phototrophic metabolism using conditional fba. *Scientific Reports* 5.
- Sajitz-Hermstein, M., Nikoloski, Z., 2016. Multi-objective shadow prices point at principles of metabolic regulation. *Biosystems* 146, 91–101.
- Sánchez, B.J., Zhang, C., Nilsson, A., Lahtvee, P.J., Kerkhoven, E.J., Nielsen, J., 2017. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular systems biology* 13, 935.
- Sarkar, D., Mueller, T.J., Liu, D., Pakrasi, H.B., Maranas, C.D., 2019. A diurnal flux balance model of *Synechocystis* sp. pcc 6803 metabolism. *PLOS Computational Biology* 15, e1006692.
- Segre, D., Vitkup, D., Church, G.M., 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences* 99, 15112–15117.
- Song, H.S., McClure, R.S., Bernstein, H.C., Overall, C.C., Hill, E.A., Beliaev, A.S., 2015. Integrated in silico analyses of regulatory and metabolic networks of *Synechococcus* sp. pcc 7002 reveal relationships between gene centrality and essentiality. *Life* 5, 1127–1140.
- Song, H.S., Reifman, J., Wallqvist, A., 2014. Prediction of metabolic flux distribution from gene expression data based

- on the flux minimization principle. PLoS One 9, e112524.
- Steinhauser, D., Fernie, A.R., Araújo, W.L., 2012. Unusual cyanobacterial tca cycles: not broken just different. Trends in plant science 17, 503–509.
- Tian, M., Reed, J.L., 2018. Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. Bioinformatics 34, 3882–3888.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) , 267–288.
- Toyoshima, M., Toya, Y., Shimizu, H., 2020. Flux balance analysis of cyanobacteria reveals selective use of photosynthetic electron transport components under different spectral light conditions. Photosynthesis Research 143, 31–43.
- Vijayakumar, S., Angione, C., 2017. Multi-omic data integration elucidates *Synechococcus* adaptation mechanisms to fluctuations in light intensity and salinity, in: International Conference on Bioinformatics and Biomedical Engineering, Springer. pp. 220–229.
- Vijayakumar, S., Conway, M., Lió, P., Angione, C., 2017. Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. Briefings in bioinformatics 19, 1218–1235.
- Vu, T.T., Hill, E.A., Kucek, L.A., Konopka, A.E., Beliaev, A.S., Reed, J.L., 2013. Computational evaluation of *Synechococcus* sp. pcc 7002 metabolism for chemical production. Biotechnology journal 8, 619–630.
- Wang, M., Luan, G., Lu, X., 2019. Systematic identification of a neutral site on chromosome of *Synechococcus* sp. pcc 7002, a promising photosynthetic chassis strain. Journal of Biotechnology 295, 37–40.
- Wang, Y., Chen, L., Zhang, W., 2016. Proteomic and metabolomic analyses reveal metabolic responses to 3-hydroxypropionic acid synthesized internally in cyanobacterium *synechocystis* sp. pcc 6803. Biotechnology for biofuels 9, 209.
- Włodarczyk, A., Selão, T.T., Norling, B., Nixon, P.J., 2019. Unprecedented biomass and fatty acid production by the newly discovered cyanobacterium *synechococcus* sp. pcc 11901. BioRxiv , 684944.
- Wortel, M.T., Noor, E., Ferris, M., Bruggeman, F.J., Liebermeister, W., 2018. Metabolic enzyme cost explains variable trade-offs between microbial growth rate and yield. PLoS computational biology 14, e1006010.
- Xiong, Q., Feng, J., Li, S.t., Zhang, G.y., Qiao, Z.x., Chen, Z., Wu, Y., Lin, Y., Li, T., Ge, F., et al., 2015. Integrated transcriptomic and proteomic analysis of the global response of *Synechococcus* to high light stress. Molecular & Cellular Proteomics 14, 1038–1053.
- Yang, Y., Feng, J., Li, T., Ge, F., Zhao, J., 2015. Cyanomics: an integrated database of omics for the model cyanobacterium *Synechococcus* sp. pcc 7002. Database 2015, bau127.
- Yu, J., Liberton, M., Cliften, P.F., Head, R.D., Jacobs, J.M., Smith, R.D., Koppenaar, D.W., Brand, J.J., Pakrasi, H.B., 2015. *Synechococcus elongatus* utex 2973, a fast growing cyanobacterial chassis for biosynthesis using light and co 2. Scientific reports 5, 8132.
- Yuan, H., Cheung, C., Hilbers, P.A., van Riel, N.A., 2016. Flux balance analysis of plant metabolism: the effect of biomass composition and model structure on model predictions. Frontiers in plant science 7, 537.
- Zampieri, G., Vijayakumar, S., Yaneske, E., Angione, C., 2019. Machine and deep learning meet genome-scale metabolic modeling. PLoS computational biology 15.
- Zhang, S., Bryant, D.A., 2011. The tricarboxylic acid cycle in cyanobacteria. Science 334, 1551–1553.



1. Cells cultured under various growth

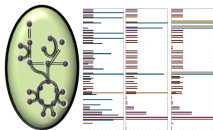
Journal Pre-proof

- RNA extracted and sequenced
- Data deposited in CyanOmics



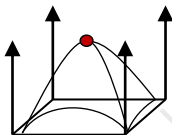
2. RNA-Seq collected from CyanOmics

- RPKM (Reads per kilobase per million)
- Normalization



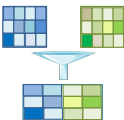
3. Data integrated into the genome-scale metabolic model (GSMM) of *Synechococcus* spp. PCC 7002

- Condition-specific transcript data loaded into METRADE
- Three pairs of objective vectors (f, g) select reactions where flux is optimized



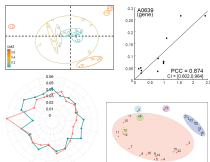
4. Constraint-based modeling and regularizes flux balance analysis to compute condition-specific fluxes

- Flux simulated using quadratic programming and L2 regularization
- Metabolic goals: biomass, photosystems I & II, and ATP maintenance reactions



5. Multi-omic datasets generated as input for subsequent ML analysis

- Transcriptomic and fluxomic dataset (individually)
- Transcriptomic-fluxomic datasets combined (at various stages)



6. Machine learning approaches identify key multi-omic features contributing to variance and predicting growth rates

- PCA reduces data dimensionality
- k-means clusters growth conditions
- LASSO regression eliminates redundant features
- Correlation associates omic predictors with growth

- . A pipeline for metabolic modeling in *Synechococcus* sp. PCC 7002 is presented
- . Metabolic fluxes display clear differences in pathway activity across conditions
- . Omic-informed GSMMs provide critical mechanistic details within machine learning
- . Combining GSMM and machine learning improves methods based on transcriptomics alone

Journal Pre-proof