Pacific Northwest
NATIONAL LABORATORY

PNNL-36694

Bayesian Framework for Predicting and Controlling Metabolic Phenotypes in Microbial System

September 2024

Andrew McNaughton
James Pino
Shant Mahserejian
August George
Connah Johnson
Pavlo Bohyutski
Vladislav Petyuk
Jeremy Zucker

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-36694

0BBayesian Framework for Predicting and Controlling Metabolic Phenotypes in Microbial System

September 2024

Andrew McNaughton
James Pino
Shant Mahserejian
August George
Connah Johnson
Pavlo Bohyutski
Vladislav Petyuk
Jeremy Zucker

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

To improve titers, rates and yields for sucrose production in an engineered strain of *Synechococcus elongatus* PCC7942, we employed Bayesian metabolic control analysis to transcriptomics and external metabolomics data generated for various phases during the circadian clock. Top overexpression candidates included sodium-dependent bicarbonate transporter (H2cO3_Nat_syn), and UTP—glucose-1-phosphate uridylyltransferase (GALUi). Top repression candidates included Glycogen/starch synthetases, ADP-glucose type (GLCS3), Glutamate racemase (GLUR), and ribonucleoside diphosphate reductase (RNDR1).

Summary

This project leverages the advanced Bayesian Metabolic Control Analysis (BMCA) methodology to optimize metabolic engineering in *Synechococcus elongatus* for enhanced sucrose production. Building upon tools developed under the DOE Agile BioFoundry and extended through the Predictive Phenomics Initiative, this application employs a reduced model of *S. elongatus*, consisting of 396 reactions and 927 metabolites, validated against experimental data from 30 different circadian conditions.

Methodology: The BMCA framework constructs a low-fidelity kinetic model of *S. elongatus* metabolism using linear-logarithmic kinetics to predict steady-state internal metabolite concentrations and metabolic fluxes (Visser and Heijnen 2003). By integrating input variables (extracellular metabolite concentrations and RNA expression) with output variables (steady-state fluxes and internal metabolite concentrations), the methodology estimates posterior distributions for kinetic parameters consistent with observed data. Experimental data, including transcriptomics and metabolomics, were used to estimate uptake and excretion rates of key metabolites, which were then mapped to corresponding reactions in the metabolic model.

Key Results: The posterior distributions of kinetic parameters were estimated using automatic differentiation variational inference with the PyMC library, optimized for convergence. The posterior predictive distribution closely matched the experimental data, validating the model's accuracy. Metabolic Control Analysis (MCA) identified six enzymes with significant control over sucrose productivity. The bicarbonate transporter showed the highest flux control coefficient (FCC), making it a primary target for overexpression, while glycogen synthase and glutamate racemase were identified as candidates for repression to increase sucrose flux.

Comparative analysis showed that BMCA outperformed single-data-modality approaches, which often resulted in false positives due to insufficient data integration. The BMCA framework demonstrated a more reliable pathway for identifying metabolic engineering targets, providing actionable predictions for optimizing sucrose production in *S. elongatus*.

Conclusions: The application of BMCA in this project highlights its robust capabilities in integrating multi-modal data to generate accurate metabolic engineering predictions. By accounting for uncertainty and variability in kinetic parameters, BMCA offers a superior method for metabolic optimization. This project provides a comprehensive framework for enhancing productivity in engineered microbial systems, contributing significantly to advances in metabolic engineering and bio-production.

Acknowledgments

This research was supported by Predictive Phenomics Initiative, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

BMCA:	Bayesian Metabolic Control Analysis
MCA:	Metabolic Control Analysis
FCC:	Flux Control Coefficients
ELBO:	Evidence Lower Bound

Contents

Figures

Tables

1.0 Introduction

This report details the application of Bayesian Metabolic Control Analysis (BMCA) to the cyanobacterium *Synechococcus elongatus*, with the specific aim of optimizing its sucrose production capabilities. Leveraging methodologies initially developed under the DOE Agile BioFoundry and further extended under the Predictive Phenomics Initiative, this study showcases advanced techniques in systems biology and metabolic engineering to enhance the bio-production potential of *S. elongatus*.

1.1 Context and Significance

Synechococcus elongatus is a photosynthetic microorganism known for its robustness and genetic tractability, making it an ideal candidate for metabolic engineering. Optimizing sucrose production in *S. elongatus* holds significant potential for a variety of applications, including biofuel production, bioplastics, and other industrial biochemicals. By applying Bayesian methodologies, BMCA provides a robust framework to integrate multi-modal experimental data, address uncertainties, and deliver high-confidence predictions for metabolic engineering.

1.2 Objectives

The primary objective of this report is to demonstrate the effectiveness of the BMCA methodology in identifying actionable metabolic engineering strategies to enhance sucrose production in *S. elongatus*. This involves constructing a low-fidelity kinetic model using linear-logarithmic kinetics to predict steady-state internal metabolite concentrations and fluxes based on enzyme expression and media conditions.

1.3 Approach

Our approach encompasses several critical steps:

1. **Kinetic Model Construction**: Developing a reduced model of *S. elongatus* metabolism, encompassing 396 reactions and 927 metabolites, adapted from a recently published model (Broddrick, et al. 2019) and engineered to produce sucrose (Wang, Betenbaugh and Young 2023).

2. **Data Integration**: Integrating experimental data, including transcriptomics and metabolomics, from 30 different circadian conditions to inform the model.

3. **Posterior Estimation**: Utilizing automatic differentiation variational inference implemented in the PyMC Python library to estimate posterior distributions of kinetic parameters (Salvatier, Wiecki and Fonnesbeck 2016).

4. **Metabolic Control Analysis**: Conducting MCA to propagate uncertainties in kinetic parameters and identify critical control points in the metabolic network that influence sucrose production (St. John, et al. 2019).

1.4 Expected Outcomes

By employing BMCA, we aim to identify key enzymes and metabolic pathways that significantly influence sucrose production in *S. elongatus*. This detailed understanding will

guide targeted genetic modifications to enhance bioproduction efficiency. The results will demonstrate the efficacy of BMCA in providing high-confidence predictions for metabolic engineering, thereby contributing valuable insights and tools for broader applications in bio-manufacturing and systems biology.

In summary, this report presents a focused application of Bayesian Metabolic Control Analysis to optimize sucrose production in *Synechococcus elongatus*. It highlights an advanced methodology to streamline and enhance metabolic engineering efforts, aligning with the DOE's vision for innovative and sustainable biotechnological advancements.

2.0 Methods

2.1 Bayesian metabolic control analysis

In this application, we employed the Bayesian metabolic control analysis (BMCA) methodology previously developed under the DOE Agile BioFoundry (St. John, et al. 2019) and extended it under the Predictive Phenomics Initiative. In BMCA, a low-fidelity kinetic model of microbial metabolism is constructed leveraging linear-logarithmic kinetics (Visser and Heijnen 2003). With known kinetic parameters, a kinetic model enables the expected steady-state internal metabolite concentrations and metabolic fluxes to be estimated as a function of enzyme expression and media conditions. With measurements of both the input variables (extracellular metabolite concentrations and enzyme expression) and the output variables (steady-state fluxes and internal metabolite concentrations), posterior distributions in the kinetic parameters that are consistent with the observed data can then be estimated.

In this application, we used a reduced model of *S. elongatus* (396 reactions, 927 metabolites) adapted from a recently published model and experimental data for 30 samples under various circadian conditions to demonstrate the ability of the method to generate actionable metabolic engineering predictions (Broddrick, et al. 2019). This model was augmented with the ability to secrete sucrose based on a previously published engineered strain. (Lin, Zhang and Pakras 2020) (Wang, Betenbaugh and Young 2023). See **Figure 5**.

The experimental data consists of transcriptomics and external metabolomics measurements and is depicted graphically in **Figure 1**. The spent media and time at collection were used to estimate strain-specific uptake and excretion rates for key measured extracellular metabolites. Metabolomics and transcriptomics measurements indicate relative changes in abundance of key intermediate species and were mapped back to their appropriate identifier in the metabolic model.

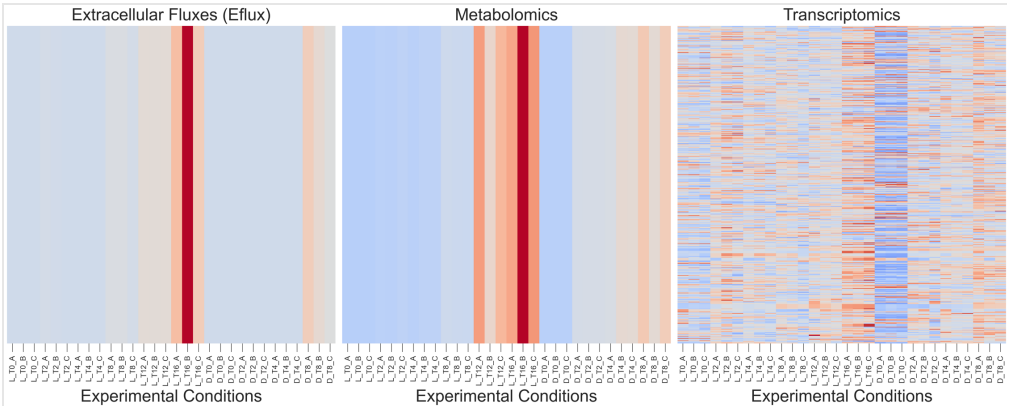

Figure 1: Depiction of the data collected in the multi-modal *S. elongatus* experimental campaign, including extracellular fluxes, metabolomics and proteomics. Rows represent measured quantities, while columns indicate different experimental conditions. Red values indicate expression levels higher than the mean for each measured quantity, while blue indicates a decrease relative to the mean. Extracellular fluxes were calculated using our E-Flux algorithm.

3.0 Results and Discussion

3.1 Results

Due to the size of the kinetic model considered, posterior distributions in kinetic parameters as a function of the observed data was estimated using automatic differentiation variational inference as implemented in the PyMC Python library, using the PyTensor backend (Salvatier, Wiecki and Fonnesbeck 2016). The model was optimized until convergence of the evidence lower bound score using the Adagrad optimizer (**Figure 2**).

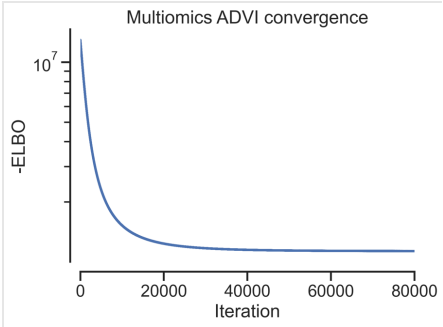

Figure 2: Optimization of the ELBO score.

The posterior predictive distribution (PPD) of the model shows the ability of the model to reproduce the variability found in the experimental dataset. The PPD of the fitted model closely reproduces the measured steady-state transcriptomics data within the unclipped shaded region (**Figure 3**). Outside this region, predicted transcriptomics measurements fluxes were based on clipped.

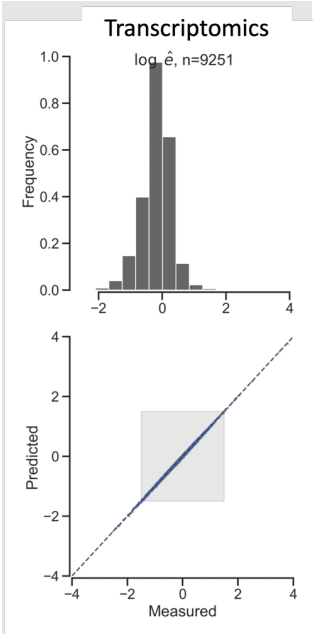

Figure 3: Posterior predictive distribution of the fitted model. The transcriptomics closely match the experimentally measured values. Transcriptomics data are in log-transformed, dimensionless units relative to the reference condition (L_T16_B). Measured fluxes were calculated under a fixed growth rate with strain-specific optimal sucrose secretion rates, using

E-flux2 to constrain strain-specific intracellular fluxes with transcriptomics data mapped to enzyme activity.

With a kinetic model and estimated probability distributions in kinetic parameters, we can next conduct the Metabolic Control Analysis (MCA) portion of the BMCA framework (St. John, et al. 2019). Here, we propagate the uncertainty in the estimated kinetic parameters to the metabolic design strategies suggested by MCA. In **Figure 4**, we show the 94% highest posterior density regions of flux control coefficients (FCCs) sucrose export calculated from the posterior distribution. FCCs capture the systems-level regulation of changing enzyme concentration on steady-state metabolic flux. The results show that six enzymes in the sucrose synthesis pathway are predicted to have the highest overall control on sucrose productivity. The largest flux control coefficient was the Bicarbonate transport (Na/HCO3 cotransport periplasm). Other overexpression candidates ordered by decreasing control over sucrose flux include Sulfate transport via diffusion (SO4tex), UTP-glucose-1-phosphate uridylyl transferase (GALUi), and Ribonucleoside-diphosphate reductase (RNDR1). Since SO4tex does not rely on a gene, it is not a target for overexpression.

We also look at the most negative FCC's, which represent targets for knockout or repression. In order of decreasing control over sucrose flux is Glycogen synthase (GLCS3) and Glutamate racemase (GLUR).

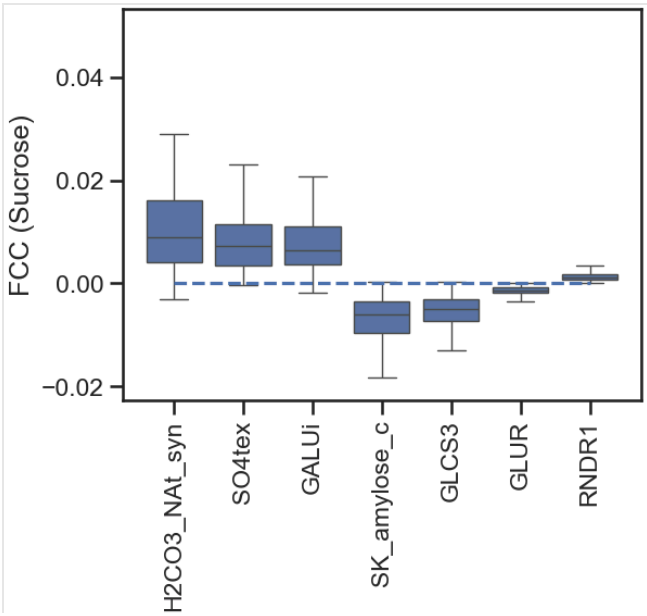

Figure 4: Posterior distributions in 3HP flux control coefficients. A positive FCC indicates that an increase in the corresponding enzyme concentration will increase sucrose production, while a negative FCC indicates that a decrease in enzyme concentration will increase sucrose. Error bars indicate the 94% highest posterior density region. While the *SK_amylose_c* reaction has a negative FCC, it is a product of modeling and not a target for under expression.

The performance of the BMCA methodology was then compared to approaches using only a single data modality. A simple approach for determining enzyme targets that does not require additional experimental data is by searching for transcripts whose expression is highly correlated with higher sucrose flux. A list of the five most positively correlated and most negatively correlated genes is shown in **Figure 6**. Many of these genes are distantly or unrelated to sucrose production, underscoring that correlation may not imply causation.

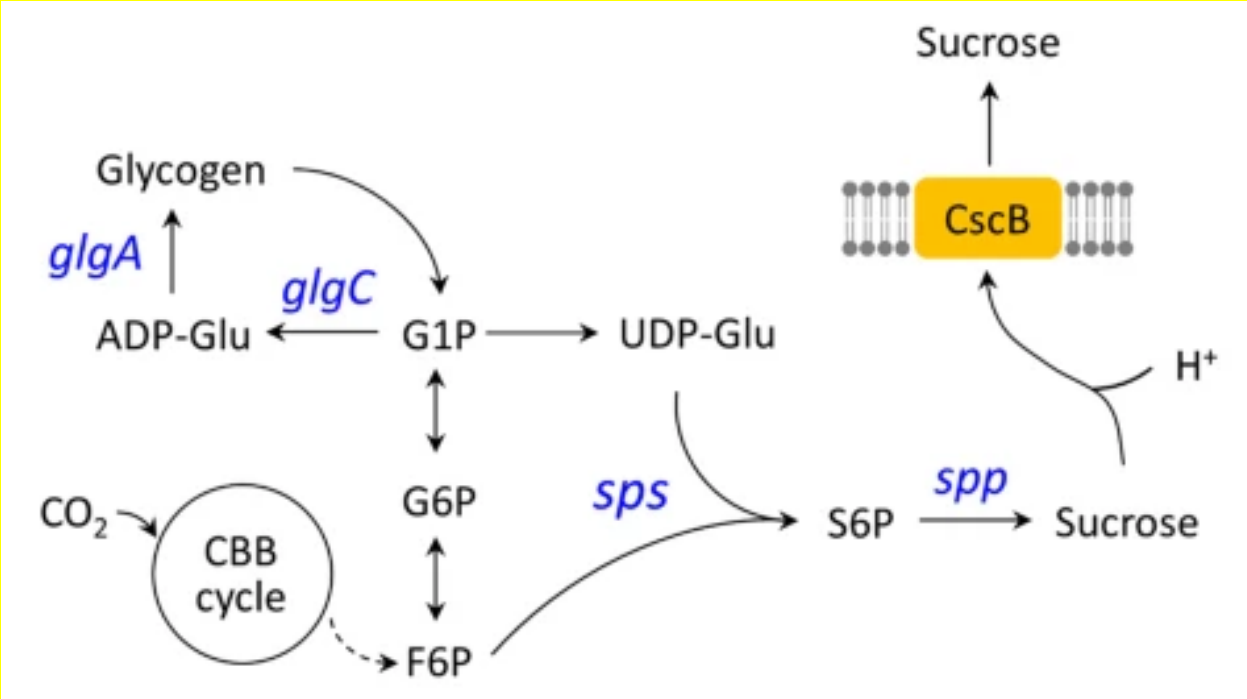

Figure 5: Engineered pathways for sucrose production. This diagram highlights the sucrose production pathway engineered into *S. elongatus.* Particularly the *CscB* gene which allows for sucrose to be exported from the cell. We can visualize how some suggested modifications from the FCC plot may affect this pathway, such as the under expression of GLCS3 reaction representing the *glgA* gene.

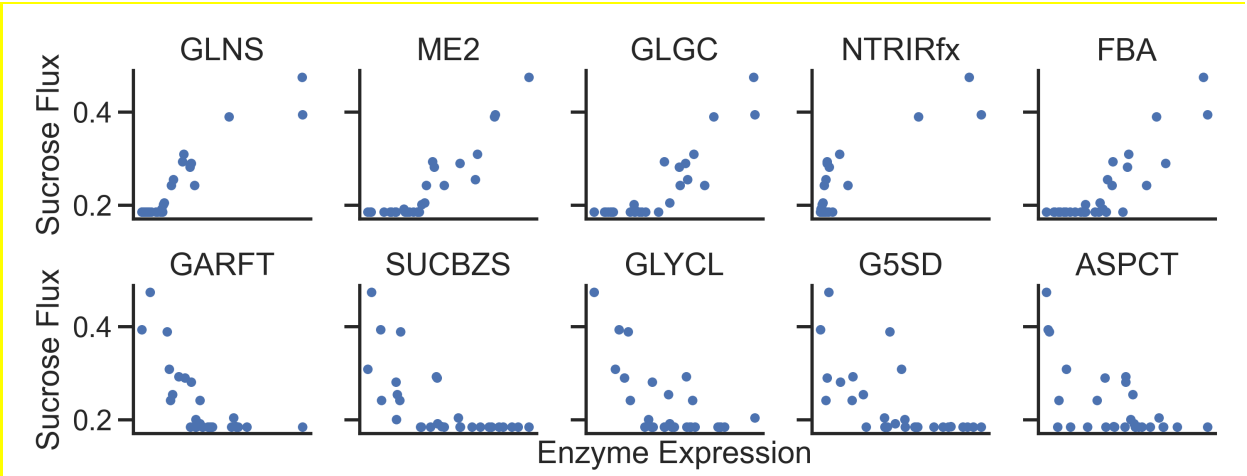

Figure 6: An analysis using only a single data modality to generate metabolic engineering predictions. Transcripts with the highest correlation with sucrose flux are shown in the top row, while transcripts with the most negative correlation are shown in the bottom row. While comparison of hit rates for these methods against BMCA would require experimental confirmation, these sets likely include more false positive targets.

Even though the observed variations in transcription expressions can potentially affect the flux of metabolic reactions within the cell (see **Table 1** for details), none of them were predicted to have a high flux control coefficient. The changes in transcription expression are related to the variations in the effect that light conditions have on expression in different phases of the circadian clock. The transcriptomics measurements were taken when the cell was experiencing circadian alterations which impact sucrose flux due to changes in storage and utilization of the energy source. Additionally, a connection between nitrogen fixation and the circadian cycle has been considered in cyanobacteria such as *Synechococcus elongatus* (Bandyopadhyay, et al. 2024). Therefore, general changes in photosynthetic pathways, which may affect the carbon and nitrogen fixation pathways, would be expected.

Table 1: The expressed mRNA transcripts can be associated to a set of metabolic reactions whom each play a role in cell. The over expressed transcripts can be directly linked to sucrose flux while under expression of the other proteins can be seen as resource management.

Protein	Descriptive Name	Reaction
GLNS	Glutamine synthetase	ATP + L-Glutamate + Ammonium ⇌ ADP + L-Glutamine + H+ + Phosphate
ME2	Malic enzyme (NADP)	L-Malate + NADP ⇌ CO2 + NADPH + Pyruvate
GLGC	Glucose-1-phosphate adenylyltransferase	ATP + D-Glucose 1-phosphate + H+ ⇌ ADPglucose + Diphosphate
NTRIRfx	Nitrite reductase (ferredoxin)	8.0 H+ + Nitrite + 6.0 Reduced ferredoxin ⇌ 2.0 H2O + Ammonium + 6.0 Oxidized ferredoxin
FBA	Fructose-bisphosphate aldolase	D-Fructose 1,6-bisphosphate ⇌ Dihydroxyacetone phosphate + Glyceraldehyde 3-phosphate
GARFT	Phosphoribosylglycinamide formyltransferase	10-Formyltetrahydrofolate + N1-(5-Phospho-D-ribosyl)glycinamide ⇌ N2-Formyl-N1-(5-phospho-D-ribosyl)glycinamide + H+ + 5,6,7,8-Tetrahydrofolate
SUCBZS	O-succinylbenzoate-CoA synthase	2-Succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate ⇌ H2O + O-Succinylbenzoate
GLYCL	Glycine Cleavage System	Glycine + NAD + 5,6,7,8-Tetrahydrofolate ⇌ CO2 + 5,10-Methylenetetrahydrofolate + NADH + Ammonium
G5SD	Glutamate-5-semialdehyde dehydrogenase	L-Glutamate 5-phosphate + H+ + NADPH ⇌ L-Glutamate 5-semialdehyde + NADP + Phosphate
ASPCT	Aspartate carbamoyltransferase	L-Aspartate + Carbamoyl phosphate ⇌ N-Carbamoyl-L-aspartate + H+ + Phosphate

These pathways can indeed be seen to alter through considering the overexpressed proteins. Glutamine synthetase (GLNS) is a step-in nitrogen metabolism that acts to remove a buildup of nitrogen in the organism which occurs during a day circadian phase (Bandyopadhyay, et al. 2024). Carbon fixation and photosynthetic processes can be connected in the pyruvate metabolism with Malic enzyme (ME2) replenishing the NADP energy source, as can be seen during photosynthetic pathways in plant systems. This connection between carbon fixation and photosynthesis is also evident with the changes in expression of Fructose-bisphosphate aldolase (FBA), demonstrating a change in how carbon is processed as part of the Calvin cycle. The change in sucrose metabolism can be seen more directly through the overexpression of Glucose-1-phosphate adenylyl transferase (GLGC) showing the processing of sucrose

derivatives. The impact of the circadian system on nitrogen fixation is shown through overexpression of Nitrite reductase (NTRIRfx) which acts to reduce the nitrogen source (Nitrite) in ammonium. This process occurs during light intensive (day) conditions, to provide a pathway to reduce ferredoxin, and the correlation with sucrose flux suggests a decreased sucrose flux during low light (night) conditions.

The under expressed proteins seen in the measurements are not directly related to sucrose synthesis. Therefore, an under expression can be seen as a cost saving or resource control procedure. are involved in the pyrimidine and purine biosynthesis pathways. Aspartate carbamoyl transferase (ASPCT) is involved in pyrimidine biosynthesis which produces high energy molecules, CTP, ATP used for cell signaling and sucrose transport. These molecules are however also generated during day conditions so are not required during the circadian phase transition. Similarly, Phosphoribosyl glycinamide formyl transferase (GARFT), which generates purine for purine biosynthesis pathway and ATP production, is also not necessary. The photosynthetic associated proteins O-succinylbenzoate-CoA synthase (SUCBZS) and Glycine Cleavage System (GLYCL) are also adversely change under the circadian shift and would act contrary to the carbon and nitrogen fixation. Finally, Glutamate-5-semialdehyde dehydrogenase (G5SD), which while involved in catabolism and redox balance, is not directly connected to sucrose metabolism that is driving the expression profiles.

3.2 Discussion

The application of Bayesian Metabolic Control Analysis (BMCA) to Synechococcus elongatus aimed at optimizing sucrose production has yielded significant insights and actionable metabolic engineering strategies. By constructing a low-fidelity kinetic model incorporating linear-logarithmic kinetics and integrating multi-modal experimental data, this study effectively demonstrated the power of BMCA in predicting and enhancing metabolic fluxes under various circadian conditions.

3.2.1 Main outcomes

1. **Model Validation**: The kinetic model, encompassing 396 reactions and 927 metabolites, was successfully validated against experimental data from 30 different circadian conditions. The posterior predictive distribution was shown to closely reproduce observed transcriptomics data, confirming the model's accuracy.

2. **Key Enzyme Identification**: Metabolic Control Analysis (MCA) identified key enzymes with significant flux control coefficients (FCCs) over sucrose production. The bicarbonate transporter was highlighted as the top candidate for overexpression to enhance sucrose flux, while glycogen synthase and glutamate racemase were identified as primary targets for repression.

3. **Efficacy of BMCA**: Comparative analysis showed that BMCA outperformed single-data-modality approaches by providing more reliable and high-confidence predictions for metabolic engineering. This underscores the robustness of BMCA in integrating experimental data and accounting for parameter uncertainties.

3.2.2 Future Work:

1. **Experimental Validation**: Future efforts will focus on experimentally validating the identified metabolic engineering targets. Genetic modifications will be implemented to overexpress or repress the key enzymes identified by BMCA, and the resulting impact on sucrose production will be quantified.

2. **Expanded Applications**: The BMCA methodology will be applied to other microbial systems and metabolic pathways to explore broader bio-production applications. This includes leveraging BMCA for optimizing the production of different biochemicals and biofuels.

3. **Model Refinement:** Applying the Consistent Reproduction of Phenotype (CROP) algorithm to essential genes in *Synechococcus* will enhance the accuracy of the model. This iterative refinement process will ensure that the metabolic models faithfully represent the real metabolic capabilities and interactions within the microbial community.

4. **Method Development:** Future innovations in BMCA will incorporate knowledge of enzyme regulation and the ability to learn from phosphoproteomics data. Additionally, we aim to compare and integrate BMCA with other machine learning algorithms within a simulated Design-Build-Test-Learn (DBTL) cycle. We also plan to extend BMCA to handle gene knockouts, further broadening its applicability in metabolic engineering.

In conclusion, this study showcases the potential of Bayesian Metabolic Control Analysis as a powerful tool for optimizing metabolic pathways in *Synechococcus elongatus*. The successful identification of key metabolic control points and the validation of the BMCA methodology pave the way for future advancements in bio-manufacturing and systems biology, aligning with the DOE's vision for innovative and sustainable technological solutions.

4.0 Bibliography

Bandyopadhyay, Anindita, Annesha Sengupta, Thanura Elvitigala, and Himadri B. Pakrasi. 2024. "Endogenous clock-mediated regulation of intracellular oxygen dynamics is essential for diazotrophic growth of unicellular cyanobacteria." *Nature Communications* 15 (3712).

Broddrick, Jared T., David G. Welkie, Denis Jallet, Susan S. Golden, Graham Peers, and Bernhard O. Palsson. 2019. "Predicting the metabolic capabilities of Synechococcus elongatus PCC 7942 adapted to different light regimes." *Metabolic Engineering* 52: 42-56.

Lin, Po-Cheng, Fuzhong Zhang, and Himadri B Pakras. 2020. "Enhanced Production of Sucrose in the Fast-Growing Cyanobacterium Synechococcus Elongatus UTEX 2973."." *Scientific Reports* 10 (1): 390. https://doi.org/10.1038/s41598-019-57319-5.

Salvatier, John, Thomas V. Wiecki, and Christopher Fonnesbeck. 2016. "Probabilistic programming in Python using PyMC3." *PeerJ Computer Science* 2: e55.

St. John, Peter C., Jonathan Strutz, Linda J. Broadbelt, Keith E. J. Tyo, and Yannick J. Bomble. 2019. "Bayesian inference of metabolic kinetics from genome-scale multiomics data." *PLOS Computational Biology* 15: e1007424.

Visser, Diana, and Joseph J. Heijnen. 2003. "Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics." *Metabolic Engineering* 5 (3): 164-176.

Wang, Bo, Cristal Zuniga, Michael T Guarnieri, Karsten Zengler, Michael Betenbaugh, and Jamey D Young. 2023. " “Metabolic Engineering of Synechococcus Elongatus 7942 for Enhanced Sucrose Biosynthesis." *Metabolic Engineering* 80 (November): 12–24. https://doi.org/10.1016/j.ymben.2023.09.002.

Appendix A – Code and Data Availability

Use Body Text for paragraphs in this section. PNNL reports use http://www.chicagomanualofstyle.org/home.html for document style. Right-click and choose open hyperlink to view the style guide.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov