A Metabolic Lens on Phytoplankton Physiology

By

Craig McLean

B.S. and B.A., University of Arkansas (2016)

Submitted to the Department of Earth, Atmospheric, and Planetary Sciences

In partial fulfilment of the requirements for the degree of

Doctor of Philosophy in the field of Chemical Oceanography

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and the

WOODS HOLE OCEANOGRAPHIC INSTITUTION

September 2021

Author ……

Joint Program in Oceanography/Applied Ocean Sciences and Engineering
Massachusetts Institute of Technology
and Woods Hole Oceanographic Institution
August 20, 2021

Certified By ……

Elizabeth B. Kujawinski
Senior Scientist, Marine Chemistry and Geochemistry, WHOI
Thesis Supervisor

Accepted By ……

Colleen M. Hansel
Senior Scientist, Marine Chemistry and Geochemistry, WHOI
Chair, Joint Committee on Chemical Oceanography

A Metabolic Lens on Phytoplankton Physiology

By

Craig McLean

Submitted to the Department of Earth, Atmospheric, and Planetary Sciences

Massachusetts Institute of Technology

and Woods Hole Oceanographic Institution

On August 20, 2021, In partial fulfilment of the requirements for the degree of

Doctor of Philosophy in the field of Chemical Oceanography

Abstract

Phytoplankton are communities of diverse groups of prokaryotic and eukaryotic single-celled organisms responsible for nearly 50% of global primary production. The relative abundance of individual groups changes dynamically in response to environmental perturbations. Recent studies suggest that such changes are primarily driven by the distinct physiological responses employed by each group towards a particular perturbation. Although knowledge of some of these responses has come to light in recent years, many aspects of their metabolisms remain unknown. We attempt to address this gap by studying the metabolism of several phytoplankton groups using metabolomics. Firstly, we developed a method to enhance the analysis of untargeted metabolomics data. Secondly, we constructed two conceptual models describing how metabolism of the raphidophyte *Heterosigma akashiwo* responds to phosphorus and nitrogen stress. These conceptual models revealed several new stress response mechanisms not previously reported in other phytoplankton. Finally, we compared the metabolic changes of several distinct phytoplankton groups to uncover possible adaptation and acclimations that distinguish them. This analysis revealed several pathways and metabolites that represent the studied groups. The contributions of these pathways and metabolites towards physiology may support the ecological fitness of these organisms.

Thesis supervisor: Elizabeth B. Kujawinski

Title: Senior Scientist, Marine Chemistry and Geochemistry, WHOI

Acknowledgements

First, and foremost, I would like to thank my advisor, Liz Kujawinski, for giving me the opportunity to work with her and contribute to her research program. She has been a thoughtful, patient, and kind mentor, always making herself available to discuss research and non-research problems and provide honest guidance. During times of hardship, Liz was always understanding of my choices and supportive. I would also like to thank my committee members Sonya Dyhrman, Penny Chisholm and Ben Temperton, as well as my defense chair, Julie Huber. Sonya provided an incredible perspective over phytoplankton ecology that helped me understand the value of our work together far beyond my initial understanding of the field. Penny's wisdom always helped me see my work from a different angle and find ways to focus even the smallest details of a project. Ben always challenged me to put forth the best possible work due to his experiences as a computational scientist.

Working with all of the members of the Kujawinski Lab has been a splendor and a privilege during this time of scientific and personal growth. In particular, Krista Longnecker has been a rock to me in the lab. Her persistent guidance and never-ending support never once ceased to amaze me. Melissa Kido Soule was incredibly generous with her willingness to share her vast expertise on mass spectrometry to help me develop as a scientist. I would not have half my work had it not been for her willingness to share these secrets. I greatly enjoyed learning from Winn Johnson, Harriet Alexander, Gretchen Swarr, Brittany Widner, Alex Frank, Yuehan Lu, Katie Halloran, Noah Germolus, Erin McParland, Laura Weber, and Mike Mazzotta.

The community of scientists at Parsons at MIT for all the good times and hard questions. Many many many people made themselves available to me as resources of growth throughout the duration of my PhD since my very first day of class. Special thank you's are owed to Gabriel Leventhal, Annie Yu, and David Albores Angeles. You all truly challenged me to be better and provided critical yet constructive feedback towards my own scientific development. The way I think about sciences is in no small part due to you all. I am grateful for uncountably many from Parsons for their friendships. These connections would not have been possible without Phil Gschwend, Eileen Covey, and Kris Kipp, who helped me join this community.

Beyond science, my experience working with the MIT Communications Lab helped me become far more competent than I ever believed could be possible. The leadership of Sarah Gluck and Diana Chien are both very much responsible for this, as both provided me with generous amounts of mentorship and guidance. Additionally, working with a dynamic group of peers didn't hurt either. All of my fellow fellows always pushed me to be better with their regular high standards of excellence.

None of this work would have been possible without a variety of funding sources. I was supported for three years by a National Science Foundation Graduate Research Fellowship and one year with a GEM fellowship. The research was carried out with grants from the MIT Microbiome Center (Award ID #6936800, EBK), the Simons Foundation (Award ID #509034, EBK), the Gordon and Betty Moore Foundation (Award ID #3304 EBK), the National Science

Foundation (Award ID #OCE-0619608 to EBK and OCE-1057447 to EBK and MCKS) and the WHOI Ocean Ventures Fund.

Table of Contents

Chapter 1

Introduction to the roles of phytoplankton in marine ecology

1.1 General Overview

The global carbon cycle characterizes the movement of carbon within and between four reservoirs: terrestrial land, the deep earth, the atmosphere, and the ocean. Among all reservoirs, the ocean makes up the largest sink of carbon with approximately 40,000 Gt C stored within its depths [1]. In addition to storage, the ocean serves a key role in the uptake of atmospheric CO_2, as studies show that nearly one half of the total carbon fixation takes place within the euphotic zone of the ocean [2]. Carbon storage at depth and its fixation in the surface ocean are connected through the activity of the biological pump [3]. The term, "biological pump," describes how the biological driven sequestration of atmospheric carbon to the deep ocean. At its base, phytoplankton are responsible for the influx of atmospheric carbon. Phytoplankton make up one percent of the total plant biomass due to their high turnover and produce one half of the worlds oxygen.

Beyond their role in the global carbon cycle, some phytoplankton also form blooms. Blooms occur when the level of phytoplankton biomass is uncharacteristically high for a given water body. Noxious phytoplankton blooms, otherwise known as harmful algal blooms (HABs), have increased in frequency over the past 40 years due to increased climate change and eutrophication [4]. These blooms can be benign or harmful, the latter of which impacts finishing and tourist economies and public health [5, 6]. Over recent years, harmful algal blooms have caused hundreds of millions of dollars in economic damages [7]. Part of the threat imposed by these harmful algal blooms is due to the release of neurotoxins like brevotoxins and saxitoxins, which have caused human mortality [7]. Other sources of economic harm stems from fish kills, due to the depletion of oxygen in aquatic ecosystems following a rapid influx of fixed carbon by the blooming phytoplankton [8]. Phytoplankton blooms are often studied in relation to the availability of dissolved nutrients like phosphorus or nitrogen, as studies suggest that these nutrients play a key role in bloom formation [4].

In addition to their roles in bloom formation, nitrogen and phosphorus are two of the most widespread limiting nutrients to primary production across marine ecosystems [9-11]. Nitrogen can enter marine ecosystems via the activity of nitrogen-fixing bacteria, episodic upwelling from depth [12, 13] riverine input, while phosphorus arrives through riverine inputs,

atmospheric deposition, and episodic upwelling [9]. Both of these nutrients make up critical macromolecules required for life including amino acids, cofactors, energy intermediates, membrane lipids, and nucleic acids. The physiological need for these macromolecules has resulted in the development of ratios to describe a phytoplankton's access to these elements [3]. On bulk scales, the canonical Redfield Ratio (C:N:P = 106:16:1) is considered as the ideal physiological quota of these elements. However, the physiological quotas of different species and groups vary greatly [14, 15]. These differences have been attributed to different physiological demands in macromolecular constructs, like ribosomes [16]. Such differences can impact physiology. Consider the case of ribosomes. A greater abundance of ribosomes would facilitate faster translations of mRNA into peptides; however it would come at the trade-off of a higher demand of nitrogen and phosphorus relative to carbon required for ribosome and peptide biosynthesis. The diversity of observed stoichiometric ratios suggests various distinct metabolic strategies are utilized by phytoplankton.

The stoichiometric heterogeneity of phytoplankton is echoed by a great diversity of physiological processes that distinguish these organisms. For example, they can vary several orders of magnitude in size [17], employ distinct strategies to respond to their environment [18-20], and foster unique chemical strategies to interact with neighboring organisms [21, 22]. These differences are mirrored genotypically. The organisms making up phytoplankton communities originate from highly diverse lineages spanning millions of years of evolution [23] and produce distinct gene expression profiles [24-26]. Additionally, studies have noted various cases of distinct genomic rearrangements between phytoplankton. Some interesting examples include the enrichment in non-coding DNA within the cyanobacteria *Trichodesmium erythraeum* [27] or the dynamic chromosomal rearrangement of dinoflagellate *Prorocentrum micans* [28].

Phenotypic and genotypic differences have been posited to support the observed coexistence of these organisms [19]. Understanding phytoplankton coexistence is a longstanding question in ecology, and has been termed as the 'paradox of plankton' [29]. This paradox stems from the apparent violation of the principle of competitive exclusion, which states that organisms competing for similar resources should not coexist. Instead, a winner-

take-all scenario should lead to a homogenous community. As photosynthesizers, phytoplankton are competing among one another for similar resources, yet coexist in diverse communities throughout the world. Many explanations for this have emerged over time, including life history differences and species oscillations. However, in recent years many examples of research have highlighted how different groups of phytoplankton employ distinct metabolic responses to a common stimuli [30, 31]. Such studies have posited that niche partitioning, or the utilization of diverse non-overlapping resources, employed through these metabolic changes may be establishing coexistence.

The synthesis above highlights several of the contemporary research areas involving phytoplankton. Among these, several outstanding questions remain, such as what are the bottom-up and top-down controls on bloom formation, how will the efficiency of the biological pump change under future scenarios, how is coexistence established, and how will rates of photosynthesis change when faced with the increasing threats of climate change. In order to better understand these questions, we must be able to predict how phytoplankton communities will respond to environmental perturbations. Determining the factors causing phytoplankton communities to change will help policy makers and scientists forecast both the role of the ocean as a sink of carbon under future climactic scenarios and potential dangers to coastal businesses and communities from blooms.

Predicting changes in phytoplankton community structure following a perturbation requires an improved understanding of the unique physiological differences among phytoplankton. Such physiological differences are mediated by response in metabolism. Gauging the diversity of metabolisms among phytoplankton is challenging due to the morphological and ecological complexity of these organisms. For example, the term phytoplankton describes a diverse set of cohabiting single-celled organisms comprising both eukaryotes and prokaryotes [17]. Additionally, acquiring knowledge on metabolism is further inhibited by the paucity of reference genomes and the biases of existing medical databases towards model organisms that affect human health [17, 23, 32]. Recently, the advent of field-based metagenomic and metatranscriptomic techniques have greatly improved our capacity to define and classify these responses. Metatranscriptomics measurements seek to profile the

composite mRNA of a given field sample. Recent advances in reference sequencing databases specific to phytoplankton have enabled the mapping of field-collected reads to a wide variety of reference organisms [23]. Gene-based methods, however, remain limited in their capacity to serve as indicators of ongoing metabolic activity [33]. This is partly due to limited reference genomic material, as a total of 9 reference genomes exist for eukaryotic phytoplankton, in contrast to the myriad of phytoplankton without representation [32]. Additionally, gene-based methods can only report relative abundances that can be biased by the overall size of the sample.

Orthogonal approaches, such as metabolomics, or the study of all the compounds within a cell, organism, or tissue, reveal finger prints of biochemical processes [34]. Hence these studies may uncover the biological mechanisms driving responses to environmental perturbations. The use of metabolomics to understand phytoplankton has burgeoned over the past years. For example, these techniques have been used to study how stress induces changes in metabolism of diatoms and coccolithophores [35-37], to determine the allocation of luxury nitrogen within diatoms [38], and how particulate and dissolved organic matter changes by depth [39, 40] and under a diel cycle [41]. Additionally, these approaches can help reconstruct pathways lacking complete gene annotations [42]. Thus, metabolomics is a promising approach to understand metabolic dynamics within organisms lacking reference genetic material.

Several challenges have limited the widespread dissemination of metabolomic approaches to study marine phytoplankton communities. For one, unlike genetic techniques, metabolomic measurements lack taxonomic specificity, making assignments of the biological sources of signals challenging. Without this information, uncovering the mechanisms driving the *in situ* dynamics of metabolites remain difficult, with simultaneous limits on our ability to understand phytoplankton ecology. Secondly, the availability of computational resources lags behind the analytical advances of these approaches, making analysis of this data challenging and time consuming. I attempt to address these challenges with the work of this PhD. Each section that follows introduces the problem addressed within each chapter of the thesis. Together, these pieces bridge the gap in our ability to apply metabolomics techniques to study the *in situ* ecology of phytoplankton. Although the work presented herein does not include *in*

situ measurements, the results may serve as a foundation towards the understanding of such data.

1.2 Automating untargeted data analysis facilitates downstream analysis

The first contribution described by this work was the development of an algorithm designed to facilitate the analysis of metabolomics data. The application of metabolomics to study phytoplankton communities has grown over the past few years, yet still remains underdeveloped relative to other 'omics techniques. Despite this, many discoveries relevant to phytoplankton ecology have emerged from this method, including the recognition of molecular adaptive responses to stress or the clarification of mechanisms driving cell-cell interactions [43-45]. Often times, advances in mass spectrometry fostered these discoveries, specifically improvements in instrument sensitivity, accuracy, and data collection capacity [46]. Parallel advances in computational tools historically followed to fulfill the potential of analytical improvements [47].

The analysis of metabolomics data requires many steps. Prior to any statistical method, raw data from untargeted metabolomics experiments must be processed to generate a spreadsheet of viable observations. Processing extracts chemical signals from electrical noise and corrects for retention time drift across samples. The performance of processing algorithms depends on the selection of algorithmic parameters that capture the structure of the data, such as matrix effects and differences in analytical platforms [48]. No universal set of parameters exists for all datasets and parameters must always be optimized prior to analysis to avoid noise inflation within the feature table [49].

Tuning parameters manually is prohibitively time consuming due to the high number of possible numerical combinations. To address this challenge, I designed a novel parameter optimization algorithm, AutoTuner. I describe the implementation and validation of AutoTuner along with a detailed comparison between it and other existing methods during the second chapter of this thesis. Compared to previous methods, AutoTuner was over one thousand times faster and more reliable across a variety of measures in its function. This method facilitated the analysis of datasets used within later chapters.

1.3 Identifying the metabolic differences between well studied and less-studied phytoplankton reveals the importance of broadening our knowledge of phytoplankton metabolism

The second contribution of this thesis describes an analysis characterizing how the metabolic response caused by acute shortages of nitrogen (N) and phosphorus (P) varied between well studied phytoplankton like diatoms and coccolithophores and a relatively less-studied phytoplankton, the raphidophyte *H. akashiwo. H. akashiwo* populations are distributed ubiquitously within subtropical environments [8, 50], and can cause harmful algal blooms which have caused significant economic losses [51]. Both N- and P-stress are known to be important drivers of *H. akashiwo* blooms [52, 53].

The unique biological make up of phytoplankton groups contributes towards their ability to respond to scarce nutrients and form blooms. These responses are driven by changes in metabolism that lead to differences in physiology. These strategies vary from reallocation of intracellular nutrients [35, 45, 54], reduction of nutrient quotas needed for growth [55, 56], or increased production of dissolved inorganic nutrient transporters [24, 25, 45]. Their composition can have significant impacts on fitness of the organism, or the extent to which an organism is adapted to survive in a particular environment. To date, the metabolic changes associated with nutrient availability have been identified in only a few well-studied phytoplankton (*e.g.*, diatoms [57-59], coccolithophores [37, 60, 61]), leaving significant gaps in our understanding of nutrient responses in other bloom-forming groups like raphidophytes.

To overcome this gap, I evaluated whether metabolism data from well-studied phytoplankton described the physiological response to nutrient stress of *H. akashiwo*. We examined N- and P-stress metabolism using a combination of metabolomics and transcriptomics data for this organism. The findings include the detection of several novel nutrient stress response pathways and processes inherent to *H. akashiwo* and distinct from other organisms that hint at unique ecological strategies in this organism. Additionally, I show that common nutrient stress biomarkers also serve as indicators of stress within *H. akashiwo*. Together, these results provide a mechanistic understanding of *H. akashiwo* stress response

and suggest that a broader understanding of phytoplankton metabolism is necessary to understand how phytoplankton communities will adapt to a changing ocean.

1.4 Metabolic differences between phytoplankton reveal group-specific acclimatory and adaptive aspects of their metabolisms

The third contribution of this thesis evaluates the similarity in metabolomic profiles between four distinct phytoplankton. Concurrently, it evaluates their unique responses to nitrogen (N) and phosphorus (P) stress. N and P availability is predicted to decrease in the future due to climate change, causing a reduction in phytoplankton global primary production [9, 10, 62]. However, these studies often regard phytoplankton as a monolithic unit, and do not take into account the distinct metabolic capacities of these organisms. Phytoplankton communities are made up of a myriad of phylogenetically distinct organisms [23]. Recent studies show that genotypically, these organisms are highly dissimilar and respond to nutrients differently from one another [23-25]. Such observations suggest that each phytoplankton group has acquired unique adaptations to their environments and applies distinct acclimations to overcome challenges from external disturbances [63]. Adaptations characterize distinguishing physiological properties of phytoplankton acquired through the course of evolution, while acclimations describe ephemeral responses to temporal changes in the environment. Without an understanding of the various adaptive and acclimatory strategies defining distinct groups of phytoplankton, scientists cannot begin to understand the extent to which primary production will be impacted due to climatic changes.

The physiological consequences of adaptations and acclimations manifest themselves through changes in metabolism. Hence, the unique aspects of metabolism distinguishing these organisms are likely to represent defining characteristics of each. In this study, we explored the metabolism of different phytoplankton species within and between phyla and how each species responds to N- and P-stress. To accomplish this aim, we cultured four organisms from 3 phyla, representing globally important phytoplankton groups under replete, phosphorus-stress, and nitrogen-stress growth conditions and performed metabolomics on their cells. We uncovered several distinct metabolic processes that may be group-specific adaptations by comparing

individual groups. Additionally, we demonstrate acclimations by showing how each organism responds to stress. These observations provide a foundation towards understanding the metabolic drivers distinguishing the physiology of phytoplankton and how individual groups respond to stress.

1.5 Coalescence of work

Each of the presented chapters builds upon the previous work to allow us to reach the goal of obtaining a greater understanding of how metabolism impacts *in situ* phytoplankton communities. The order in which they were presented represents the experimental sequence that lead to each discovery. The discoveries here reduce both the technical uncertainties associated with untargeted metabolomics measurements by introducing new methods for the data analysis and by ensuring the fidelity and robustness of biological signals. In addition, they provide a scaffold to interpret how metabolism serves phytoplankton communities within a local and global context. Together, they reveal an unmet need for further experiments gauging how phytoplankton respond to nutrient stress. The combination of these works establishes a foundation that may then be expanded towards the application of these methods in the field and serve as a bridge between cellular biochemistry and ecology.

Chapter 2

AutoTuner: High fidelity, robust, and rapid parameter selection for metabolomics data processing

Published as: McLean, C., Kujawinski, E.B., 2020, AutoTuner: High fidelity, robust, and rapid parameter selection for metabolomics data processing. Analytical Chemistry 92, 8, 5724-5732.

2.1 Introduction

Metabolomics is the study of all the compounds present within a cell, organism, or tissue. Such investigations provide a holistic snapshot of the activity within a biological matrix, and have led to a myriad of discoveries ranging from the elucidation of novel biochemical pathways, to the recognition of molecular adaptive responses to stress, to the clarification of mechanisms driving cell-cell interactions [43-45]. Advances in mass spectrometry fostered these discoveries, specifically improvements in instrument sensitivity, accuracy, and data collection capacity [43, 46, 64]. Parallel advances in computational tools have historically followed to fulfill the potential of analytical improvements [47].

Prior to data analysis, raw data from untargeted metabolomics experiments must be processed to generate a features table. Features are defined as peaks with unique mass to charge (*m/z*) and retention time values, with relative abundances determined by their height or area. Processing is critical to extract chemical signals from electrical noise and to correct for retention time drift across samples [65]. A variety of untargeted data processing methods exist [66-69], including two commonly used tools: MZmine2 [70] and XCMS [71]. Although these tools reliably extract true features from complex data, their performance depends on the selection of algorithmic parameters that capture the structure of the data, such as matrix effects and differences in analytical platforms [48, 72, 73]. No universal set of parameters exists for all datasets, hence parameter optimization must occur prior to analysis to avoid noise inflation within the feature table [49, 74, 75].

Tuning parameters manually is prohibitively time consuming due to the high number of possible numerical combinations. To overcome this challenge, several methods exist to identify optimal dataset-specific parameters [76-78]. These methods each use distinct optimization functions based on maximizing or minimizing a numerical value. Each approach iteratively runs XCMS peak-picking and retention time correction algorithms until they identify a set of parameters that optimizes a desired criterion. For example, isotopologue parameter optimization (IPO), the most commonly-used parameter selection tool, scores groups of parameters by the number of features detected after XCMS that contain a naturally-occurring ^{13}C isotopologue. Many separate XCMS runs are required to find ideal parameters, sometimes

taking weeks to complete [76-78]. Currently, these parameter selection algorithms depend on high performance computing resources. As users continue to adopt ultra-high pressure liquid chromatography systems and rapid scanning mass spectrometers, the size and abundance of data from these platforms will preclude the use of unscalable parameter selection algorithms to users without access to high performance computing resources [79, 80].

We designed a novel parameter optimization algorithm, AutoTuner, to ameliorate these challenges. The method performs statistical inference on raw data in a single step in order to make parameter estimates as opposed to iteratively checking estimates. Further, it complements recent tools focused on generating higher-confidence feature annotations [81-84]. AutoTuner is capable of selecting parameters for seven continuously valued parameters required for centWave peak selection algorithm used by both MZmine2 and XCMS, and it determines a key parameter for grouping in XCMS. AutoTuner is freely distributed through BioConductor as an R package.

2.2 Theory and Design of AutoTuner

2.2.1 Algorithm Overview.

AutoTuner makes estimates for the following mass spectrometry peak-picking and grouping algorithms parameters: ***Group difference***, ***ppm***, ***S/N Threshold, Scan count, Noise***, ***Prefilter intensity, and Minimum/Maximum Peak-width***. We chose to optimize these parameters because they have the greatest influence on the number and quality of post-processing features and have the greatest number of possible values [77, 85]. We chose to optimize centWave peak-picking parameters over other peak-picking methods, as centWave is the recommended method for processing high-resolution untargeted data, which is increasingly becoming the standard for untargeted metabolomics [46]. See Table 1 for a description of parameters and their matching arguments in XCMS. AutoTuner makes estimates in three steps (Figure 1):

1. **TIC Peak Detection:** A user identifies peaks within each sample's total ion chromatogram (TIC), which is the plot of integrated ion intensities within the mass spectrometer over time.

2. **Parameter Estimation Within EICs of Each TIC Peak:** AutoTuner isolates predicted extracted ion chromatograms (EICs) within each identified TIC peak. An EIC is a plot of one or more selected *m/z* values in a series of mass spectra. AutoTuner applies statistical inference on all EIC peaks to estimate parameters in an unsupervised manner.
3. **Dataset-Wide Parameter Estimates:** AutoTuner integrates all peak-specific estimates into a dataset-wide set.

2.2.2 Total Ion Chromatogram (TIC) Peak Detection

To identify TIC peaks, AutoTuner first applies a sliding window analysis, which detects peaks by testing if an upcoming scan's intensity is greater than an intensity threshold determined by the average and standard deviation of a fixed number of prior scans. To ensure the correct peak bounds are retained, AutoTuner generates a linear model from the first three and last three points bounding each TIC peak. If the model fails to calculate an R^2 value greater than or equal to 0.8 or to reach a local R^2 maximum, AutoTuner expands the ending bound by one scan and reruns the model until the model meets either criterion. The time difference of a TIC peak's final bounds represents its width.

AutoTuner groups all TIC peaks originating from distinct samples whose maxima co-occur within each other's retention-time bounds. It then determines the time differences between intensity maxima of all pairs of grouped peaks. AutoTuner returns the largest time difference as the estimate for the ***Group difference*** parameter that is used in the grouping step following peak-picking. Because highly complex datasets may contain distinct sample-specific peaks occurring at similar retention times, AutoTuner may overestimate this parameter. Prioritizing the inclusion of experimental replicates within AutoTuner would limit this issue. The overestimation of ***Group difference*** does not affect downstream parameter estimation, as future estimates do not involve comparisons across samples and instead focus on properties of individual EICs. At this point, AutoTuner has only collected data to estimate the ***Group difference*** parameter.

2.2.3 Parameter Estimation Within EICs of Each TIC Peak

AutoTuner estimates remaining parameters (***ppm***, ***S/N Threshold***, ***Scan count***, ***Noise, Prefilter intensity, and Minimum/Maximum Peak-width***) from raw data contained within each individual TIC peak. A central assumption in AutoTuner is that TIC peaks represent chromatographic regions enriched in chemical ions relative to electrical noise [65].

*Error (**ppm**).* First, AutoTuner sorts all *m/z* values detected in mass spectra contained within the bounds of a TIC peak. AutoTuner bins *m/z* values if the difference in mass of two adjacent *m/z* values is below a user-provided threshold. AutoTuner stores unbinned *m/z* values as noise peaks. Because peaks of true features are made up of *m/z* values within adjacent scans (scan continuity criterion), AutoTuner sorts each bin by scan number to check that this criterion holds for the binned *m/z* values. In the case where two or more *m/z* values are retained from a single scan, only the *m/z* value with the lowest difference in mass to the previous scan's mass is retained. If multiple *m/z* values occur within the first scan of the bin, the difference in mass is calculated for the next adjacent scan's *m/z* value instead. AutoTuner stores *m/z* values within bins that fail the scan continuity criterion as noise peaks, similar to the noise removal step earlier.

AutoTuner estimates the parts per million (***ppm***) error parameter from the remaining bins by distinguishing between bins formed by random associations of noise peaks and those of hypothesized true features. To do this, AutoTuner first calculates the ***ppm*** of all *m/z* values within bins. AutoTuner then builds an empirical distribution of ***ppm*** values using a Gaussian kernel density estimator (KDE) defined by:

$$KDE(x_i) = \frac{1}{n}\Sigma_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

where

$$K(x) = \frac{1}{\sqrt{2 * \pi}} e^{\frac{-x^2}{2}} \tag{2}$$

and x is the set of all observations, x_i is the i[th] observation, n is the number of observations, and h is a measure of smoothness for the empirical distribution. The function between ***ppm*** and absolute error is not surjective, meaning two identical absolute mass error values can have distinct ***ppm*** values. Thus, we hypothesize that the ***ppm*** value of noise peaks should be scattered widely, while ***ppm*** values of real features should be within a narrow range [72]. Hence, we expect that by using a user-provided mass difference threshold larger than an instrument-defined threshold, the KDE will have a long-tail and a high narrow peak representing the instrument-dependent ***ppm*** of real features and a shorter smaller down-stream peak representing the ***ppm*** from erroneously binned *m/z* values (Figure S1).

AutoTuner subsamples the empirical distribution of all ***ppm*** values to speed downstream calculations when calculated ***ppm*** values are abundant (> 500). To do this, AutoTuner checks the similarity between the original distribution comprising all ***ppm*** values and seven distributions comprising one-half of all ***ppm*** values randomly sampled from the total. Seven was chosen arbitrarily. The distance between the original distribution and each subsampled distribution is calculated using the Kullback-Leibler divergence (KLD), a function that calculates the information theoretic gain required to describe both distributions. A KLD value of 0.5 represents an increase of one-half bit of information required to store the two distributions. If a KLD value of 0.5 or greater is not calculated across any comparison, AutoTuner replaces the original distribution with one consisting of half as many ***ppm*** values subsampled randomly from the original, and repeats the subsampling.

AutoTuner then calculates an outlier score for each ***ppm*** value to distinguish between error values derived from real features and those derived from random associations of noise using the following outlier score function [86]:

$$Score(x_i) = \frac{KDE(x_i)}{\frac{1}{|C|}\Sigma_{x_j \in C} KDE(x_j)} \quad (3)$$

where C represents the largest cluster of error values and x_i is the i[th] observation similar to (1). To identify this cluster, AutoTuner uses k-means clustering, a data partitioning technique used

to separate a set of observations into k-many groups. Either the gap statistic or a user-provided variance-explained threshold is used to determine the appropriate number of clusters [87]. Using C ensures that the density of each calculated ***ppm*** value is normalized by the density of the true error values (Figure S1).

The ***ppm*** estimate is calculated by the following:

$$ppm_{estimate} = \max(x) + 3 * \mathrm{sd}(x) \tag{4}$$

where x is any calculated ***ppm*** value with outlier scores above 1, and $sd(x)$ is the standard deviation of all x values. An outlier score value above 1 indicates that the density of that particular *x* is at least as great as the expected value of the density of all elements within C. The statistical properties of probability distributions inspired this heuristic, as the sum of a probability distribution's mean and three times its standard deviation provides an upper bound containing 99.7 percent of the total distribution area [86].

Signal-to-noise threshold. We calculate the maximum intensity of each bin as well as the mean and standard deviation of the intensity of all noise features occurring within two peak widths from the original bin to estimate the signal to noise (S/N) threshold, similar to Myers et al.[72] First, AutoTuner subtracts the maximum intensity of each bin ($x_{bin,i}$) from the mean intensity of adjacent noise (μ_{noise}). AutoTuner retains the bin if this difference is greater than three times the standard deviation (σ_{noise}) of adjacent noise intensity values:

$$x_{bin,i} - \mu_{noise} > 3\sigma_{noise} \tag{5}$$

AutoTuner calculates the ***S/N Threshold*** from the smallest observed value of bin and noise intensity difference divided by the standard deviation of noise intensity across all bins passing the above threshold:

$$S/N_{threshold} = \min(k) \text{ where } \frac{\mu_{bin\ i} - \mu_{noise}}{\sigma_{noise}} = k_i \tag{6}$$

Remaining Parameters. AutoTuner sets the ***Scan count*** estimate as the minimum number of scans across all bins. AutoTuner estimates ***Noise*** and ***Prefilter intensity*** parameters by first determining the minimum integrated bin and single scan intensities. Then, it returns 90 percent

of the magnitude of these values as the estimate to ensure that no AutoTuner-detected bin is removed during actual peak-picking. The ***Minimum Peak-width*** represents the lowest number of scans within any bin multiplied by the duty cycle of the instrument. To estimate the ***Maximum Peak-width***, AutoTuner expands bins at the boundaries of the TIC peak. The expansion continues until an adjacent scan does not contain a *m/z* value whose error against the mean *m/z* of the bin is below the estimated ***ppm*** threshold. A correlation checks to ensure that adjacent *m/z* values are not coming from noise after a bin has been expanded by 3 scans. For this, AutoTuner requires an absolute Spearman correlation coefficient of 0.9 between scans and intensity values for expansion to continue. AutoTuner returns the ***Maximum Peak-width*** across bins.

2.2.4 Dataset-Wide Parameter Estimates

AutoTuner uses the average of all ***ppm*** and ***S/N Threshold*** values weighed by the number of bins within each TIC peak to return dataset-wide estimates for these parameters. For dataset-wide values of ***Scan count***, ***Noise***, ***Prefilter intensity***, and ***Minimum Peak-width***, AutoTuner returns the minimum values from all bins detected. The maximum calculated ***Group difference*** parameter represents the dataset wide parameter estimate. The average of each sample's maximal peak-width represents the ***Maximum Peak-width*** estimate.

2.3 Experimental Demonstration

2.3.1 Materials

We chose a suite of 85 metabolites that represent compounds expected in metabolomic experiments, including cofactors, amino acids, and secondary metabolites. Of these, 41 ionized exclusively in negative mode, 28 ionized exclusively in positive mode, and 16 ionized in both modes. See Table S1 for a complete list of standards.

We prepared stock solutions of each metabolite standard in water or a 1:1 mix of methanol and water at 1000 mg mL^{-1}, unless constrained by solubility. Some standards required the addition of ammonium hydroxide or formic acid for dissolution. We stored stock solutions in the dark at -20°C. We created a standard metabolite mix (10 mg ml^{-1}) from the stock solutions and diluted it with Milli-Q water to obtain four solutions with standard concentrations

of 500 ng mL^{-1}. We obtained standards at the highest grade available through Sigma Aldrich (MO, USA) with the exception of dimethylsulfoniopropionate (DMSP), purchased from 21 Research Plus Inc. (NJ, USA).

2.3.2 Mass Spectrometry

We analyzed four replicates of the standard mixes with ultra-high-performance liquid chromatography (UPLC; Accela Open Autosampler and Accela 1250 Pump (Thermo Scientific)), coupled via heated electrospray ionization (H-ESI) to an ultrahigh resolution tribrid mass spectrometer (Orbitrap Fusion Lumos (Thermo Scientific)). We performed chromatographic separation with a Waters Acquity HSS T3 column (2.1 × 100 mm, 1.8 µm) equipped with a Vanguard pre-column, both maintained at 40°C. We used mobile phases of (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile at a flow rate of 0.5 mL min^{-1} to elute the column. The gradient started at 1% B for 1 min, ramped to 15% B from 1-3 min, ramped to 50% from 3-6 min, ramped to 95% B from 6-9 min, held until 10 min, ramped to 1% from 10-10.2 min, and finally held at 1% B (total gradient time 12 min). We made separate positive and negative ion mode autosampler injections of 5 µL. We set electrospray voltage to 3600 V (positive) and 2600 V (negative), and source gases to 55 (sheath) and 20 (auxiliary). We set the heated capillary temperature to 375°C and the vaporizer temperature to 400°C. We acquired full scan MS data in the Orbitrap analyzer (mass resolution 120,000 FWHM at m/z 200), with an automatic gain control (AGC) target of 4e5, a maximum injection time of 50 sec, and a scan range of 100-1000 m/z. We set the AGC target value for fragmentation spectra at 5e4, and the intensity threshold at 2e4. We collected all data in profile mode.

2.3.3 Validation Data

We used two published datasets to validate AutoTuner's performance on experimental data: (1) a bacterial culture experiment[88], MetaboLights [89] identifier MTBLS157, and (2) a rat fecal microbiome, by direct contact with authors (Table 2) [90].

2.3.4 Data Processing and Quality Control

We converted all raw data files from their proprietary formats to mzML files using msconvert [91]. All computing of mzML files took place within an Ubuntu Xenial 16.04 Google Cloud instance with 8 CPUs and 10Gb of memory. During time comparisons, we used 8 and 1 CPU(s) to obtain IPO and AutoTuner data processing parameters, respectively. We used an *m/z* mass error threshold of 0.005 Daltons for AutoTuner, because this absolute error was sufficiently large enough to return a broad range of error values (in ppm) greater than those of the mass analyzers used to generate the validation data (See Table 2).

We used XCMS and centWave to generate feature tables for each dataset [71, 92], and CAMERA for isotopologue and adduct detection [93]. Table S2 contains parameters used for processing. For the standards, we searched for most abundant parent ion within EICs (See Table S1). We confirmed the presence of a metabolite standard within feature tables if a feature had an intensity above 1e4, was within an exact mass error of 5 ppm of the parent ion, and had a retention time error of 5 seconds from the EIC peak. We identified $^{12}C_{n-1}{}^{13}C_1$ and $^{12}C_{n-2}{}^{13}C_2$ isotopologue peaks as features with exact mass error of 5 ppm of the parent ion isotopologue masses (1.0033 for $^{13}C_1$ and 2.0066 for $^{13}C_2$). Additionally, we required that peaks matching *m/z* values of isotopologues also had retention time error less than 5 seconds from the $^{12}C_n{}^{13}C_0$ peak. Prior to any identification, we confirmed that standards contained isotopologue peaks by visually inspecting raw data. For the culture experiment, the data was subjected to quality control as described previously [94]. Briefly, we removed features detected in process blanks, features detected within only one replicate, and features representing isotopologues and adducts. Additionally, we removed features with coefficient of variation values above 0.4 across six pooled samples. We defined overlapping features in AutoTuner- and IPO-parameterized feature tables to be those with ppm error below 5 and retention time differences less than 20 seconds. The culture experiment allowed a higher retention time correction because it relied on data collected with HPLC compared to the standards which were analyzed with a UPLC system.

2.3.5 Statistical Analyses

We applied several distinct statistical methods to summarize the various pieces of data used to validate AutoTuner. We used R programming language to perform all analyses (CRAN R-Project). We used a Kolmogorov-Smirnov Test (KS-test) to compare empirical cumulative distribution functions. We used the hypergeometric test to compare MS^2 enrichment of IPO- and AutoTuner-specific features against features observed in the intersect of the two datasets. In order to estimate the robustness of AutoTuner parameter estimation, we performed a Monte Carlo experiment by running AutoTuner on distinct subsets of the data. We first randomly selected 7 subsets of 11 samples to compare the variability across parameters. We used the coefficient of variation from estimates within each group as a measure of variability. We also performed linear regressions on these values to find trends between estimate variability and sample numbers used for estimates. We randomly selected 3 to 9 samples from each of these subsets 55 times. In total, there were 385 estimates for each group of 3-9 samples, resulting in a total of 2695 separate runs of AutoTuner per dataset. We performed a sensitivity analysis to determine the downstream data processing effect different values on mzDiff, the only continuous valued centWave parameter not optimized by AutoTuner, had on the returned feature table. To accomplish this, we counted the number of unique features between pairs of feature tables generated with mzDiff parameters varying by a value of 0.001.

2.4 Results

2.4.1 AutoTuner Accuracy and Comparison to IPO

At this time, the only open source method for automated selection of peak-picking parameters for XCMS is isotopologue parameter optimization (IPO) [78]. IPO uses a gradient descent algorithm that requires users to iteratively run centWave with different combinations of parameters until the set that maximizes a scoring function is identified. We used 5 distinct metrics to compare the accuracy, speed, and downstream data structure of IPO- and AutoTuner-derived parameters. The metrics include the accuracy, number of features, the peak areas and shapes of EIC peaks only detected using parameters from one of the two methods, and MS^2 count.

We searched for 85 known chemical standards (a total of 101 possible ions) within feature tables generated with IPO- and AutoTuner-derived parameters to test the influence of each parameter selection method on data processing accuracy (Figure 2, Table 2, Table S1). We detected 82 and 100 standards within the feature table generated with IPO- and AutoTuner-derived parameters, respectively. Figure S2 provides an example of compounds that were only detected with AutoTuner and were absent when the IPO-derived parameters were used. These results were robust to the choice of intensity thresholds (Figure S3).

Additionally, we enumerated all features matching $^{12}C_{n-1}{}^{13}C_1$ and $^{12}C_{n-2}{}^{13}C_2$ isotopologues of standards to determine the influence of parameter values on detection of lower intensity features (Figure 2). We only considered these isotopologues if the $^{12}C_n{}^{13}C_0$ ion was present within the feature tables derived with method-specific parameters. We detected 80 out of 81 and 38 out of 64 possible $^{12}C_{n-1}{}^{13}C_1$ and $^{12}C_{n-2}{}^{13}C_2$ peaks, respectively, within the AutoTuner-derived feature table. We detected 46 out of 68 and 8 out of 59 possible $^{12}C_{n-1}{}^{13}C_1$ and $^{12}C_{n-2}{}^{13}C_2$ peaks, respectively, within the IPO-derived feature table.

We first compared the number of features from culture data generated with parameters from each algorithm to understand the influence of parameter selection on downstream data properties (Figure 3A, Figure S5A). Each feature table contained a distinct number of total features following processing and quality control (Table S3). In positive ion mode, AutoTuner-derived parameters detected fewer unique features (203) compared to 2606 unique features detected with IPO-derived parameters (Figure 3A), while sharing 1022 features between them. A similar situation was observed in negative ion mode where AutoTuner detected 540 unique features compared to 3420 unique features found with IPO-derived parameters, while sharing 904 features (Figure S5A).

We then compared the structural differences between features exclusively detected using IPO- and AutoTuner-derived parameters. We created an empirical cumulative distribution function (CDF) to compare the distribution of peak areas (Figure S5, Figure S5B) and maximum observed continuous wavelet transform (CWT) coefficients (Figure 3B, Figure S5C) of all EIC peaks belonging to features outside of the intersect. The maximum observed CWT coefficient increases with peak steepness, may provide a measure of a peak's chromatographic resolution

(Figure 3: see inset). The CDF of each metric was significantly different in positive (KS-Test, Area: $p < 10^{-6}$; CWT: $p < 10^{-4}$, n = 203) and negative (KS-Test, area: $p < 10^{-14}$, CWT: $p < 10^{-8}$, n = 540) ionization mode data. Applying the same test on unbalanced comparisons (e.g., negative ion mode: 3420 IPO- vs. 540 AutoTuner-unique features) was more highly significant than using equivalent numbers of features obtained through subsampling.

Next, we compared the abundance of features with MS/MS spectra within each unique feature table because features with MS/MS spectra can be compared to spectral datasets and authenticated standards, thus enabling potential identification. In total, we observed more features with MS/MS spectra within the feature table generated with IPO-derived parameters compared to that generated with AutoTuner-derived parameters (positive: 448 vs. 115; negative: 686 vs. 121, both for IPO vs. AutoTuner, respectively). However, this is due primarily to the greater number of features in the IPO-derived table. Indeed, relative to total features, IPO-derived features were less likely to have associated MS/MS spectra than features within the intersect of both datasets (Hypergeometric Test, Negative ion mode: $p < 10^{-10}$, Positive ion mode: $p < 10^{-10}$). A similar comparison revealed that MS/MS enrichment was not significantly different between AutoTuner-derived features and those within the intersect (Table S4).

Finally, using all three of the test datasets, we compared the time required to run AutoTuner and IPO (Table 3). After accounting for number of CPUs, AutoTuner ran hundreds to thousands of times faster.

2.4.2 Testing Robustness of AutoTuner Estimation

Figure 4 shows coefficient of variation and estimates values for each 11-sample subset for the parameter ***ppm*** obtained from Monte Carlo analysis on culture and community datasets. Figures S6-S13 show the complete set of results from the Monte Carlo analysis for all parameters. For all parameter estimates, the variability of estimation decreased linearly with the number of samples used under both ionization modes (Figure 4A and Figures S6-S13). The rendered parameter estimates were consistent with expectations based on the mass analyzer used to generate the data (Table 2). With the exception of the ***Maximum Peak-width*** parameter estimate in the community dataset and the ***Noise*** estimate in the negative culture

data, all parameters had a coefficient of variation (CV) less than or equal to 0.1 when using 9 samples to obtain estimates (Figures S6, S8, S10, and S12).

2.5 Discussion

AutoTuner is a robust, rapid, and high-fidelity estimator of untargeted mass spectrometry data processing parameters. Its unique design improves upon previous methods by providing a scalable framework to handle large datasets, reducing runtime, and generating high-accuracy parameters that retain known features. AutoTuner's ease of use make it an ideal candidate to include within existing data processing pipelines [95-98].

AutoTuner's high accuracy indicates that its parameter selection is based on true data features. One possible explanation for the lower accuracy of IPO is that the peak-width of the missing standards was below the ***Minimum Peak-width*** parameter selected by IPO (Table S2 and Figure S2).

AutoTuner parameter estimates were robust across all datasets and ionization modes. Some parameters like ***ppm***, ***Noise***, ***S/N Threshold***, ***Prefilter intensity***, and ***Scan count*** reflect systematic properties inherent to the platform chromatography, mass analyzer, and/or sample matrix [98]. Other parameters like ***Maximum peak-width*** are more specific to each sample; hence increasing the total number of samples used to estimate parameters always strengthened their robustness. The low CV values for parameter estimates suggests that using a subset of samples to generate estimates returns a set representative for all samples. Based on our results, we recommend the use of 9 and 12 samples to generate estimates in culture and community datasets, respectively. For most of the parameters estimated here, 9 samples proved sufficient to obtain estimates with CV values less than 0.1. The 12-parameter recommendation originated from extrapolating the linear fits of these data to obtain 0.1 CV values for remaining parameters that failed this criterion (Figure S6, S8, S10, and S12). We were unable to check the robustness of the ***Group difference*** parameter estimate, as this parameter is estimated through a non-automatable cross sample comparison during the TIC peak detection step of the algorithm.

Although other algorithms return more parameter estimates than AutoTuner, the parameters calculated by AutoTuner represent continuous valued ones with the greatest

possible number of choices. Performing a parameter sweeping optimization like previous approaches to estimate the remaining parameters after fixing the AutoTuner derived ones reduces the total combinations of available centWave parameters from a space of at least 2^4*5^8 possible choices of parameters to one of 40. This is because the centWave algorithm, used by both XCMS and MZmine2 data processing tools, requires tuning of 11 distinct parameters. Of these, 8 are continuously valued, meaning that they can be any real number. The remainder are either boolean values or can be one of a few discrete choices (Table S5). The reduction of the total number of combinations is achieved by optimizing 7 of the 8 continuous valued parameters. In regards to the last continuous parameter not optimized by AutoTuner, mzDiff, we performed a sensitivity analysis to show that distinct values had minimal effect on the returned feature table (Table S6). Future contributions towards AutoTuner's design can help the estimation of additional parameters not covered within its current design.

AutoTuner's low runtime indicates that the algorithm is scalable (Table 3). As more and more data is generated due to increases in analytical throughput and dataset size, AutoTuner will remain a tractable option to generate estimates of metabolomics data processing parameters [46, 99]. Because AutoTuner estimates parameters much faster than IPO, and IPO was shown to perform at a faster rate than software preceding it, we surmise that AutoTuner is the fastest parameter selection algorithm available [78].

Evaluating quality between culture dataset feature tables generated with IPO- and AutoTuner-derived parameters is impossible without performing a complete validation of all possible features. Such analyses are time consuming, labor intensive, and beyond the scope of this manuscript. However, the measured properties of these datasets may provide some expectation for practitioners of metabolomics of how the data generated using each method may differ.

When considering the unique features identified by each algorithm in the culture dataset, AutoTuner found fewer features in each case (Figure 3, Figure S4). This may be due to the selection of different ***ppm error*** parameters between AutoTuner and IPO; the ***ppm error*** thresholds selected by IPO were higher under each ionization mode than those selected by AutoTuner (Table S2). AutoTuner's lower ***ppm error*** estimates do not appear to be too

stringent, as they are between 4 and 6 times greater than the instrument-recommended error threshold of 0.5 ppm and they are consistent with recommendations by the 'centWave' developers [92]. The size of the processed data using AutoTuner-derived parameters was in line with previous work validating the metabolome of *Escherichia coli* after performing stringent isotope labeling experiments and quality control filtering [81]. AutoTuner feature selection does not appear to be biased towards higher intensity features, because the standard dataset processed with AutoTuner-derived parameters contained a high percentage of possible ^{13}C isotopologues. The paucity of size-validated metabolome datasets precludes further evaluation of the feature number comparison. Within the AutoTuner-derived feature tables, those features unique to AutoTuner-parameters were enriched in MS/MS relative to the unique IPO-derived features. We stress that features with MS/MS spectra cannot be assumed to be more or less important features within a metabolomics data set; however, the presence of these spectra enhances down-stream identification efforts and may be desirable to some investigators.

AutoTuner has several avenues for possible improvement. First, AutoTuner could be parallelized to reduce computation time by a factor of the total number of CPUs used. Second, additional algorithms may be implemented to optimize parameters not covered here. One drawback from the speed gained in the computation through its "divide-and-conquer" approach comes at a loss of comparing EIC peaks across samples to estimate retention time correction algorithms. This challenge leaves room for the implementation of additional algorithms. Third, the replacement of the sliding window analysis with a more sophisticated and sensitive peak detection approach may eliminate the need for user input during the first portion of AutoTuner. However, this automation comes at the cost of manual inspection of the raw data. We support manual inspection of the raw data, because it provides a quality control check for the data generation steps leading up to the analysis. AutoTuner provides several built-in plotting functions to facilitate this evaluation step. Despite these minor caveats, AutoTuner is a viable and time-saving option to determine proper data processing parameters for untargeted metabolomics data.

2.6 Figures

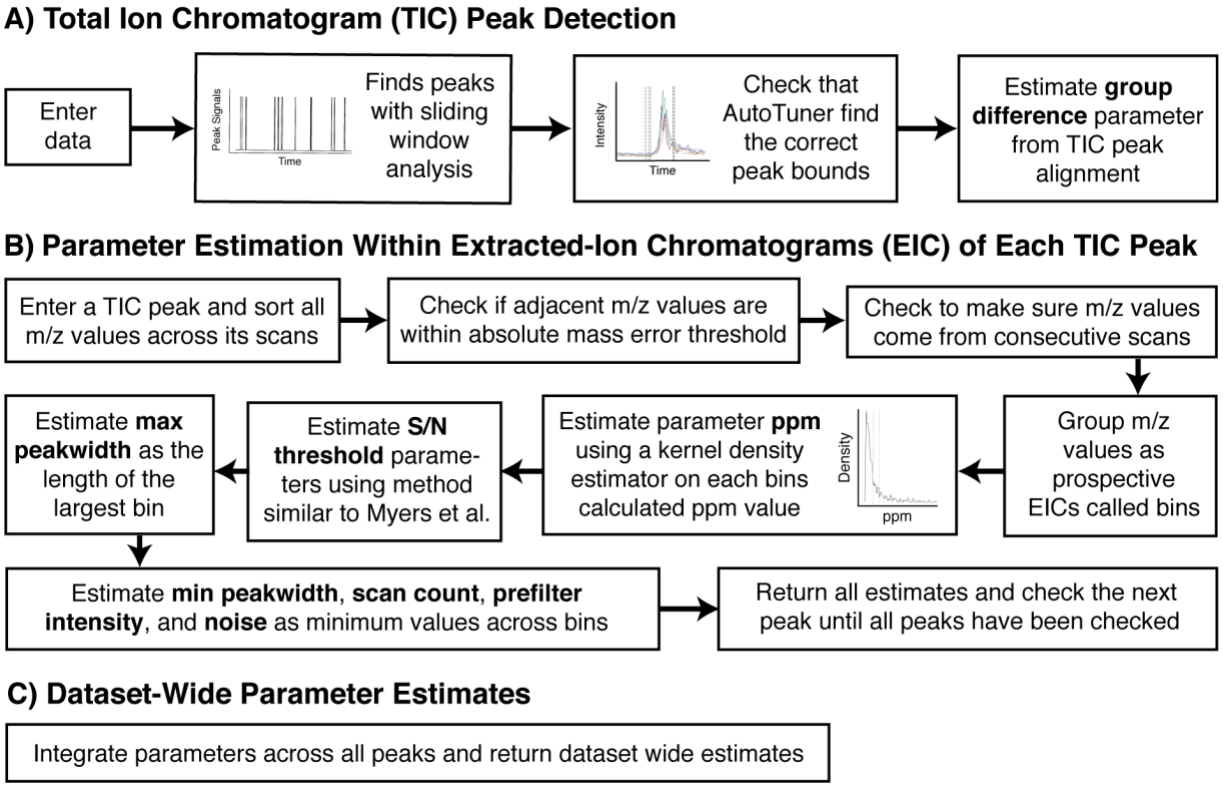


Figure 1. Schematic of the three stages of the AutoTuner algorithm. (A) Total Ion Chromatogram (TIC) Peak Detection requires user input and is focused upon identifying peaks within each sample's TIC. The user directly adjusts a signal processing sliding window analysis to identify peaks within the TIC. (B) Parameter Estimation Within Extracted-Ion Chromatograms (EICs) of Each TIC Peak iteratively looks at each peak to make parameter estimates from EICs. (C) Dataset-Wide Parameter Estimates aggregates results from the second stage to provide an ideal set of parameters for the entire dataset. Parameters estimated are in bold. The R package vignette at BioConductor provides an example on how to use the algorithm.

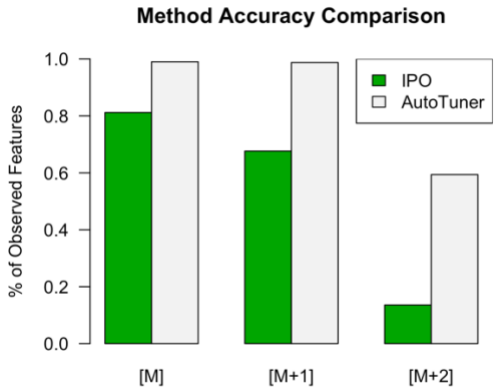


Figure 2. AutoTuner and IPO accuracy comparison. Percentages were determined from number of detected standard peaks relative to total possible set. M denotes $^{12}C_n{}^{13}C_0$ isotopologue, [M+1] denotes $^{12}C_{n-1}{}^{13}C_1$ isotopologue and [M+2] denotes $^{12}C_{n-2}{}^{13}C_2$ isotopologue. We normalized percentages by the total number of possible detectable peaks based on the detection of ^{12}C standards ($^{12}C_n{}^{13}C_0$).

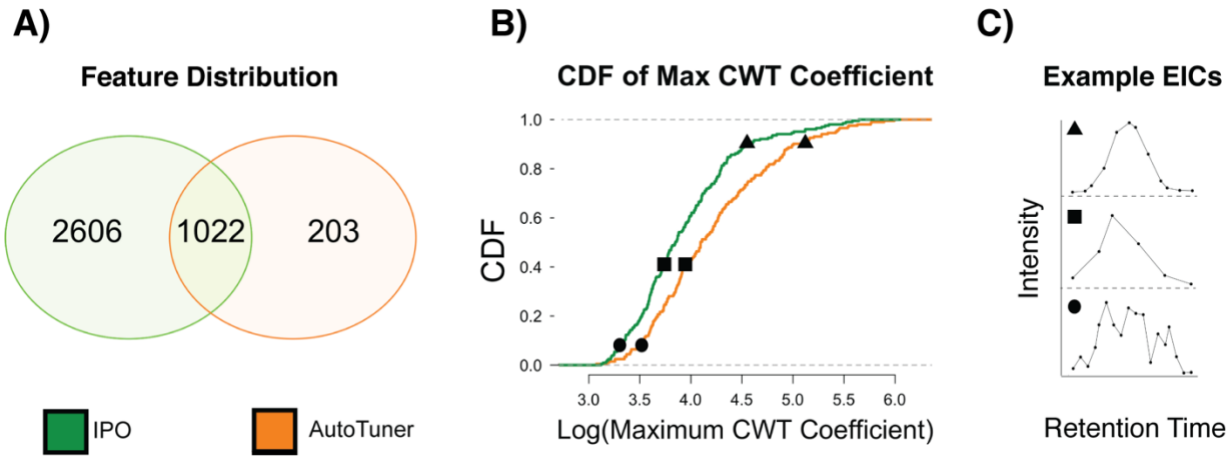


Figure 3. Comparing the differences between positive ion mode data generated by AutoTuner and IPO on culture dataset. A) portrays the overlap in the number of m/z-rt features generated by both methods. Features with an error of 5 ppm and retention time error of 20 seconds are placed in the intersect. B compares the differences in structural properties for the maximum continuous wavelet transform coefficient (CWT) between peaks detected only within AutoTuner (orange) and IPO (green). Both curves are empirical cumulative distribution functions (CDF) of the calculated metric. C shows three randomly selected EIC peaks that fall on distinct regions of the maximum CWT empirical cumulative distribution function to demonstrate how this metric influence peak shapes. The EIC shape reflects the maximal CWT rather than parameterization method. The curves were significantly different (KS-test, $p < 10^{-4}$, n = 203). Results for positive mode data area CDF and negative data were similar to this data and are found in figure S4 and S5, respectively.

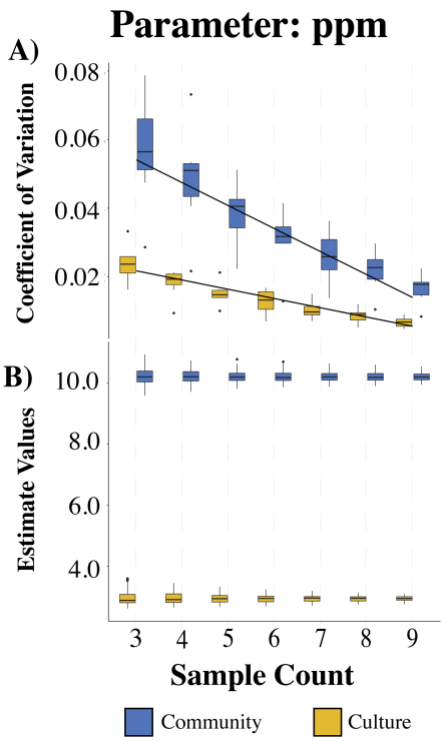


Figure 4. Results from Monte Carlo experiment for parameter ***ppm*** in positive ion mode data. A) depicts the distribution of coefficient of variation from parameters within 11 sample group, while B) shows the distribution of all estimates for ppm. Blue bars describe data collected by qTOF instrument (community data) while yellow bars describe FT-ICR-MS data (culture data). See figures S5-12 for results on other parameter estimates in each ionization mode and dataset.

2.7 Tables

Table 1. Parameters estimated through AutoTuner algorithm. We chose to optimize these parameters due to their influence on the number and quality of features returned following XCMS data processing [76, 84]. Table S4 gives more information on these parameters.

Parameter	Description	XCMS parameter name	Functionality	Application
Group difference	Expected retention time deviation of an mz/rt feature between samples	bw	Grouping	XCMS
ppm	Parts per million (ppm) error threshold used to bin consecutive mass intensities across adjacent scans into a single peak	ppm	centWave (peak-picking)	XCMS & MZmine2
S/N Threshold	The minimum ratio between peak and average noise intensity required to retain a feature	snthresh	centWave (peak-picking)	XCMS & MZmine2
Scan count	Minimum number of scans required to retain a peak	Prefilter scan	centWave (peak-picking)	XCMS & MZmine2
Noise	Numerical threshold used to filter out noise from true masses	noise	centWave (peak-picking)	XCMS & MZmine2
Prefilter intensity	Minimum integrated intensity to required retain a peak	Prefilter intensity	centWave (peak-picking)	XCMS & MZmine2
Peak-width	The width of a chromatographically resolved peak	min/max peakwidth	centWave (peak-picking)	XCMS & MZmine2

Table 2. Information on the datasets used to test AutoTuner's performance. The mass spectrometers and liquid chromatography systems herein are some of the most commonly used analytical platforms for untargeted metabolomics [46].

Dataset	Reference	Access	Mass Spectrometer	Liquid Chromatography	Sample Number	Ionization Mode
Standards	(current project)	(current project)	Orbitrap Fusion Lumos (Thermo)	Ultra-high Performance Liquid Chromatography (Accela 2015 Pump - Thermo)	4	Pos/Neg
Culture	[88]	MetaboLights MTBLS157	Hybrid Linear Ion Trap 7T Fourier Transform Ion Cyclotron Resonance (Thermo)	High Performance Liquid Chromatography (Surveyor MS Pump Plus - Thermo)	45	Pos/Neg
Community	[90]	Contributing Author	Time-Of-Flight Tandem Mass Spectrometer (Xevo-G2 waters)	Ultra-high Performance Liquid Chromatography (Acquity - Waters)	90	Pos/Neg

Table 3. Run-times for AutoTuner and IPO required to run 6 common samples collected in positive (+) and negative (-) ionization modes. All system time measurements were done on an 8 CPUs and 10Gb of memory Ubuntu Xenial 16.04 Google Cloud instance. IPO ran on 8 CPUs, while AutoTuner ran on 1 CPU. The ratio accounts for the total computing power used to run both algorithms.

Algorithm	**Culture (-)**	**Culture (+)**	**Standards (-)**	**Standards (+)**	**Community (-)**	**Community (+)**
AutoTuner	2 min	9 min	2 min	3 min	25 min	26 min
IPO	7 hr 23 min	28 hr 40 min	31 hr 56 min	28 hr 5 min	38 hr 4 min	21 hr 27 min
Ratio (Auto/IPO)	1479	1518	6970	4238	715	396
Samples Used	6	6	4	4	6	6

2.8 Supplementary Material

2.8.1 Supplementary Figures

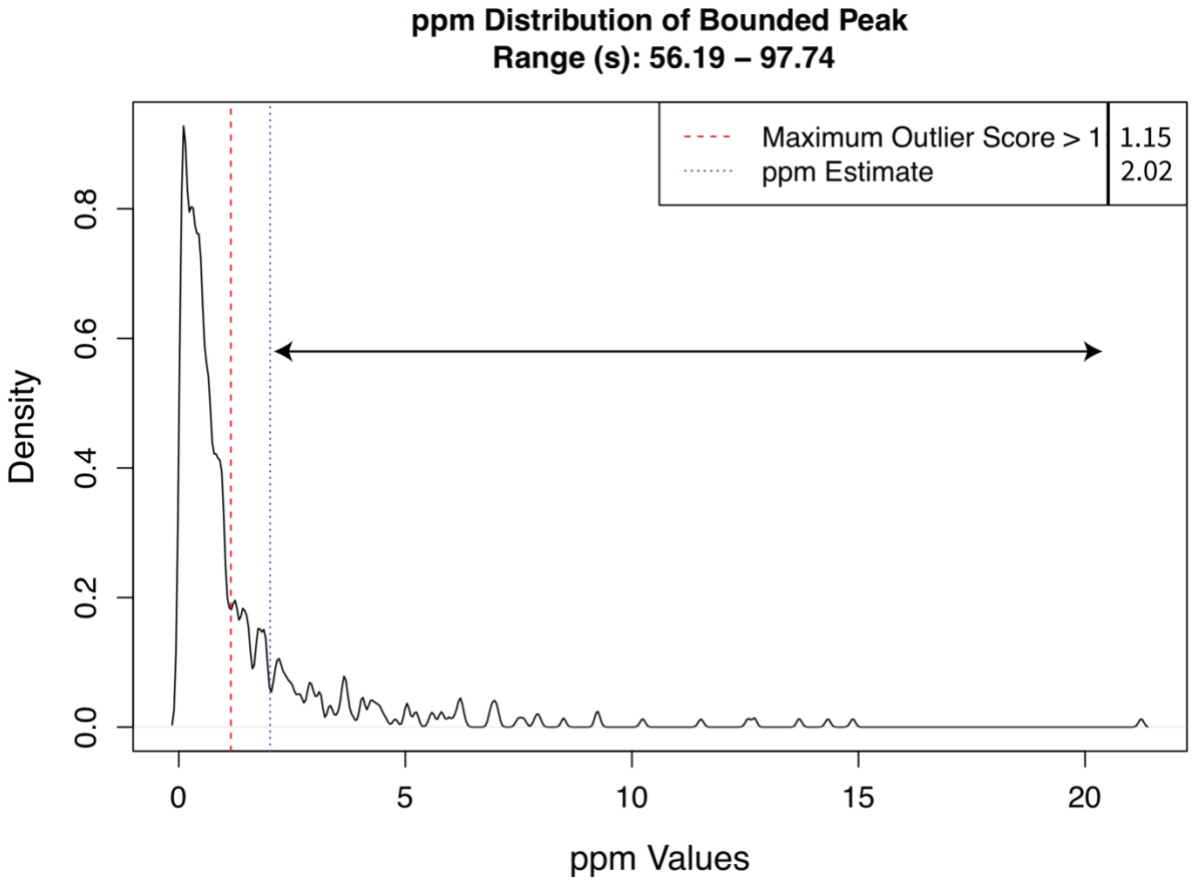


Figure S1. Example of AutoTuner-generated ppm error distribution. Such plots are returned by the algorithm to check quality of estimates. Red line represents the maximum ppm error value with an outlier score greater than 1 (see equation 3). In this example, a ppm error value of 1.15 meets this criterion (see legend). Blue line represents the ***ppm*** error parameter estimate described in equation 4, or 2.02 in this example (see legend). The "Range" subtitle represents the original chromatographic bounds of the TIC peak used to obtain estimates. The peaks under the arrow are assumed to originate from ppm values calculated from random associations of noise rather than from true features.

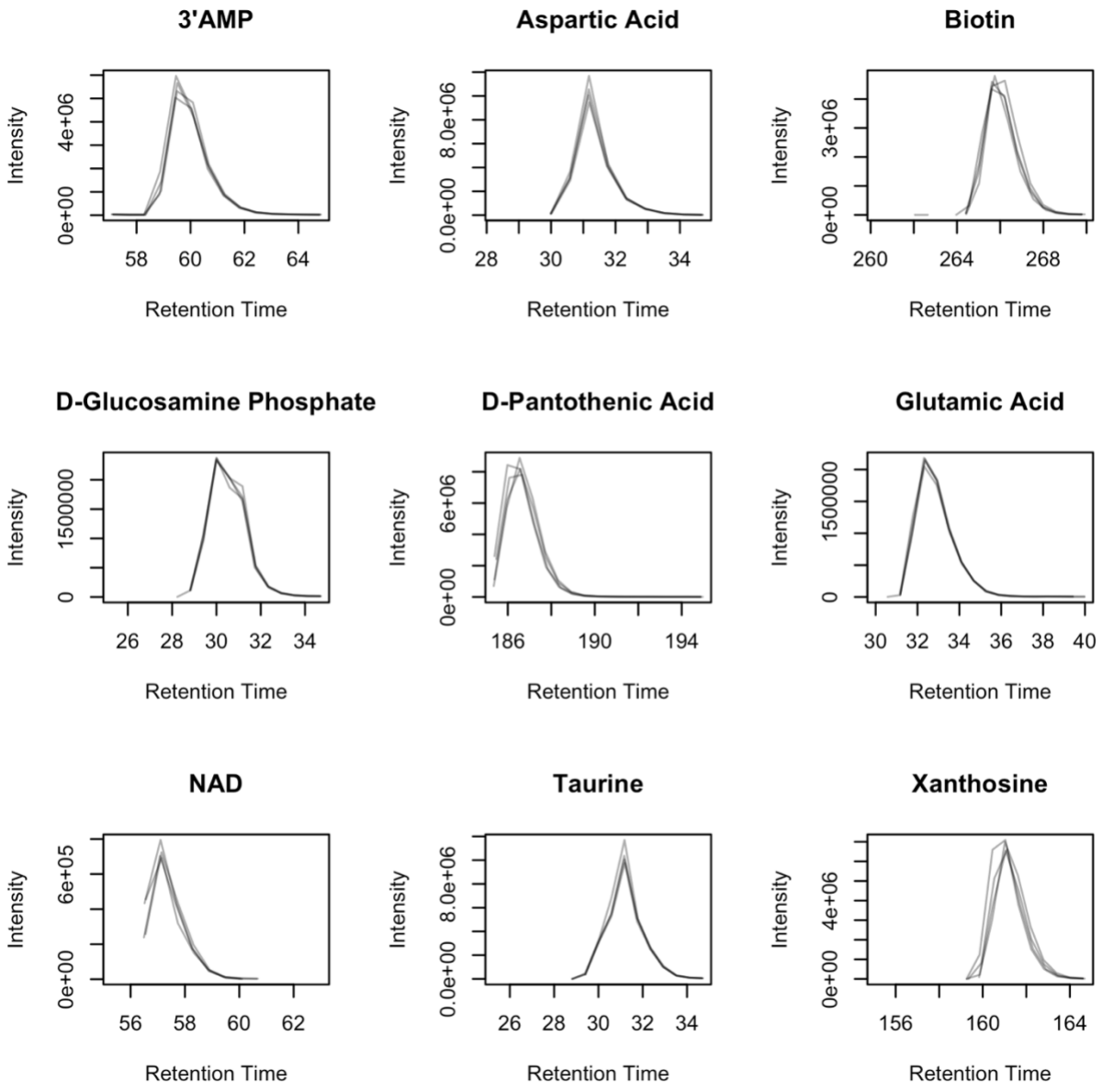


Figure S2. Example EIC peaks of standards not detected within feature table generated with IPO-derived parameters. The lines represent individual standard samples. 3'AMP = 3'-adenosine monophosphate; NAD = ß-nicotinamide adenine dinucleotide

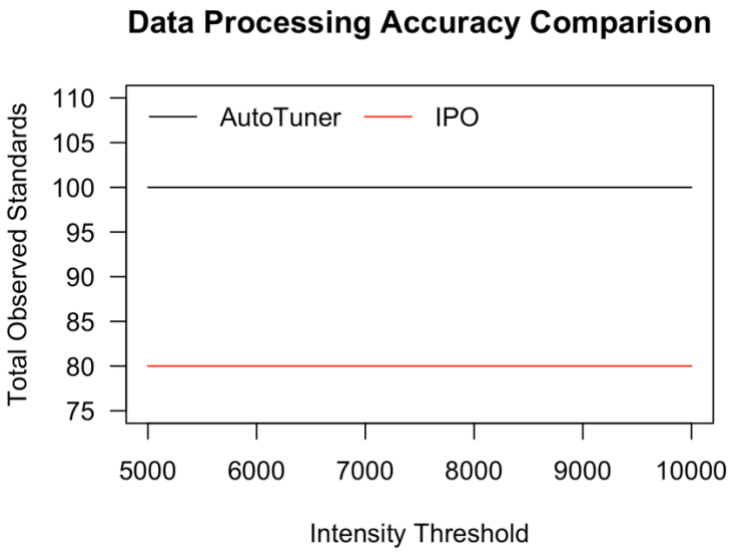


Figure S3. Impact of feature intensity threshold on standard detection. Intensity threshold varied by 5000 from 5000 until 10000. Lines indicate the number of detected standards from feature tables generated with IPO- and AutoTuner-derived parameters. The minimum intensity value observed across standards was measured at 74804.84.

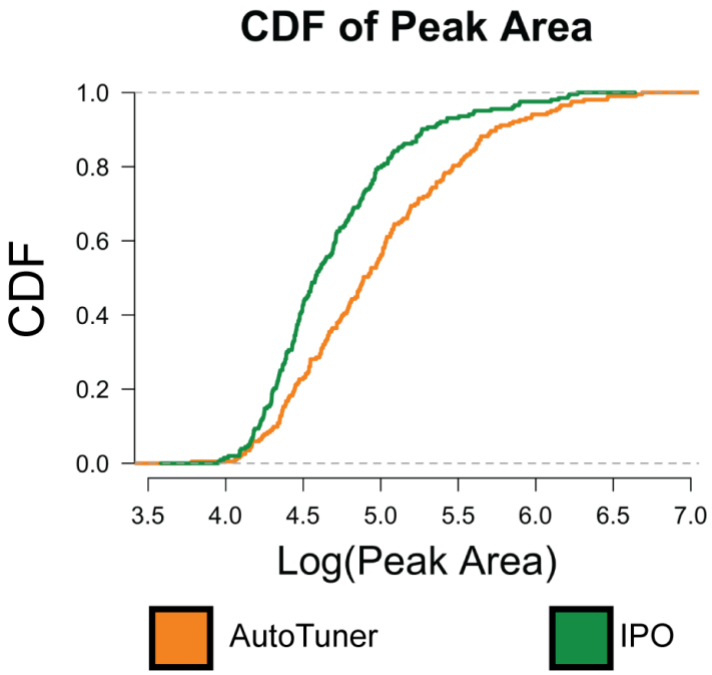


Figure S4. Positive ion mode data empirical cumulative distribution functions (CDF) comparison of peak area from EICs of features uniquely identified within feature tables generated with AutoTuner- and IPO-derived parameters. The curves were significantly different from one another (KS-test, Area: $p < 10^{-6}$; n = 203).

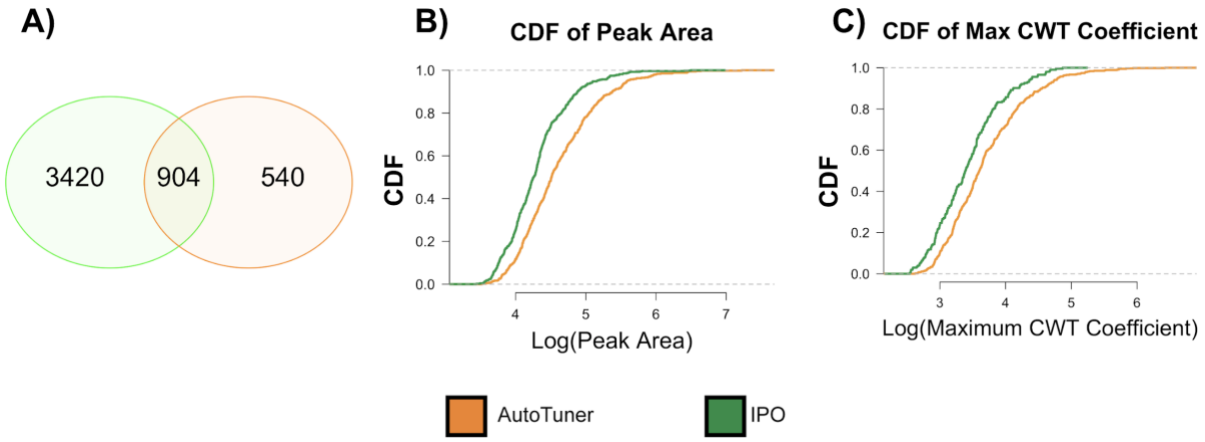


Figure S5. Negative ion mode data comparison of feature tables based on AutoTuner- and IPO-derived parameters on the culture dataset. A) portrays the overlap in the number of m/z-rt features generated by both methods. Features with an error of 5 ppm and retention time error of 20 seconds are placed in the intersect. B and C compare the differences in structural properties for the (B) peak area and (C) maximum continuous wavelet transform coefficient (CWT) between peaks detected only within AutoTuner or IPO. Both curves are empirical cumulative distribution functions (CDF) of the calculated metrics. An empirical cumulative distribution function is a non-parametric estimator of the underlying CDF of a random variable. In this case, the random variable is the set of calculated values for the AutoTuner- and IPO-specific features. CDFs for each metric were significantly different from one another (KS-test, Area: $p < 10\text{-}14$; CWT: $p < 10\text{-}8$, $n = 540$), similar to positive ion mode data.

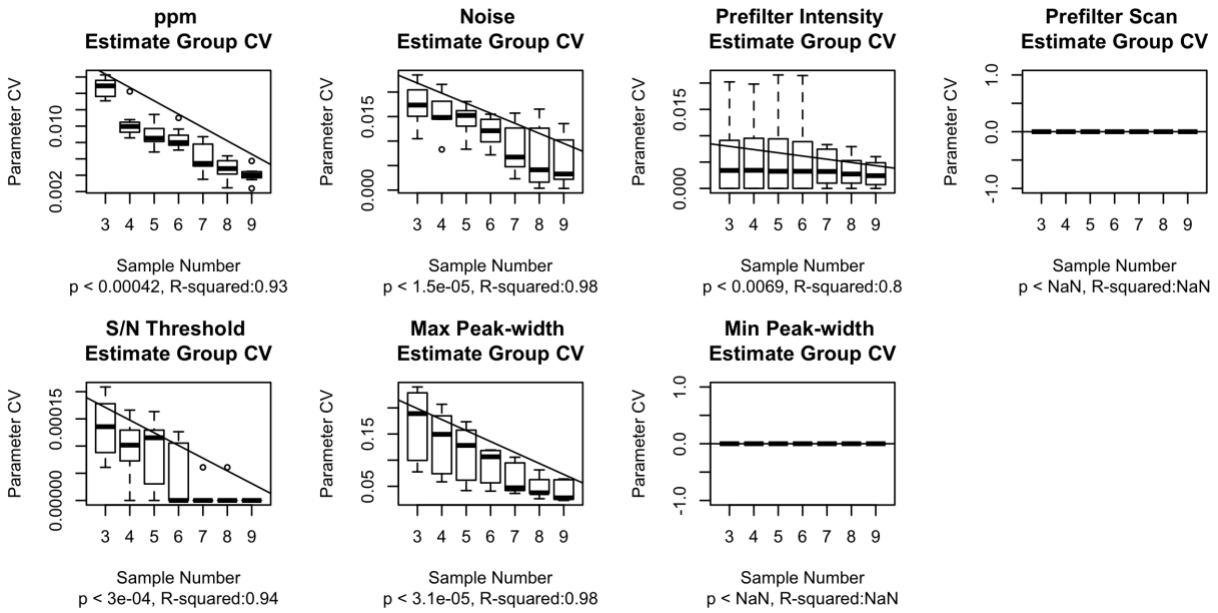


Figure S6. The coefficient of variation (CV) for groups of parameters estimated in the Monte Carlo analysis on negative mode community data. Each plot denotes the calculated CV values for each unique parameter. The x-axis describes the number of samples used to generate estimates, while the y-axis describes the CV of the estimates from each group of 11 randomly selected samples. P-value and R^2 statistics are derived from linear regressions of data (n = 49). (NaN = not a number).

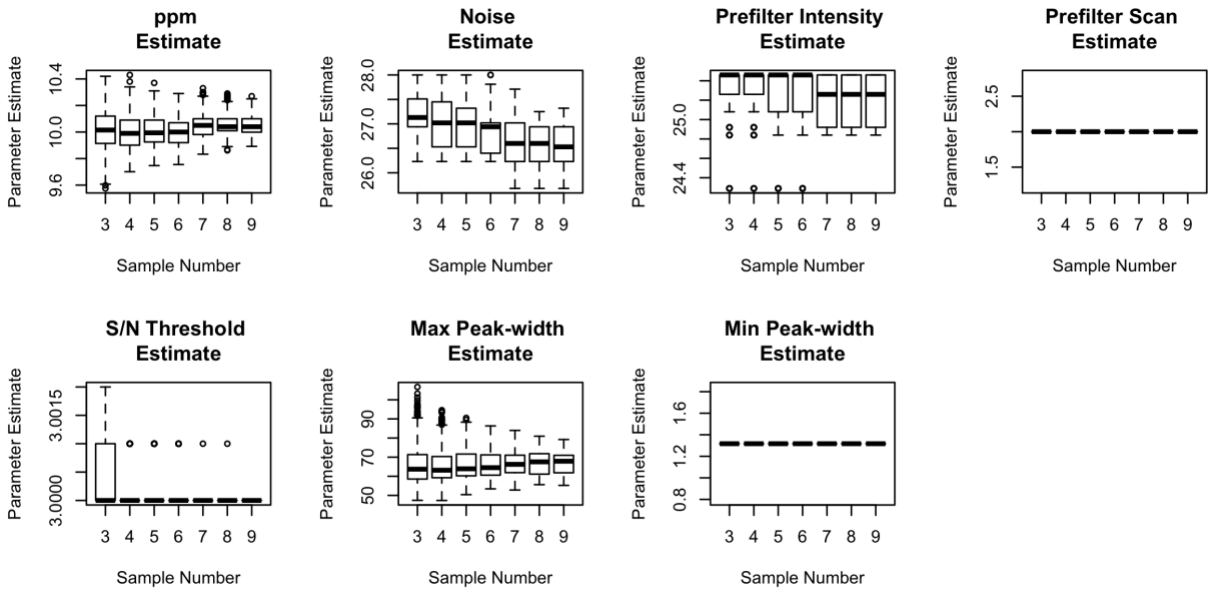


Figure S7. The parameters estimated in the Monte Carlo analysis on negative mode community data. Each plot denotes the calculated parameter estimate values for each unique parameter across 385 runs of AutoTuner. The x-axis describes the number of samples used to generate estimates, while the y-axis portrays the determined 55 parameter estimates within each n-sample subset (n = 3-9).

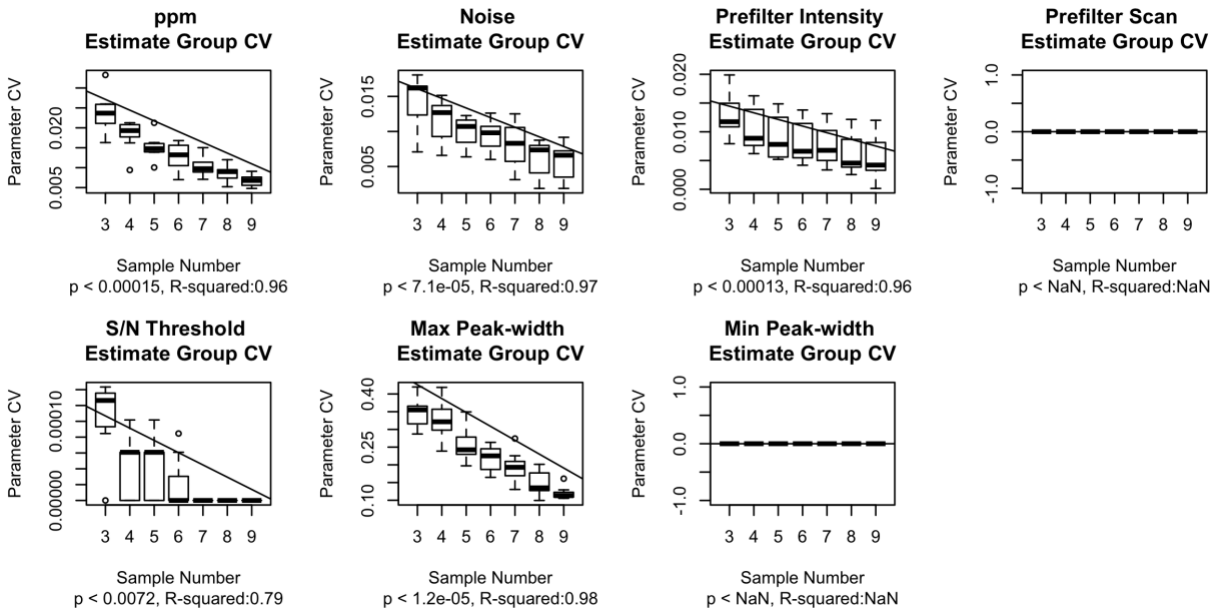


Figure S8. The coefficient of variation (CV) for groups of parameters estimated in the Monte Carlo analysis on positive mode community data. Each plot denotes the calculated CV values for each unique parameter. The x-axis describes the number of samples used to generate estimates, while the y-axis describes the CV of the estimates from each group of 11 randomly selected samples. P-value and R^2 statistics are derived from linear regressions of data (n = 49). (NaN = not a number).

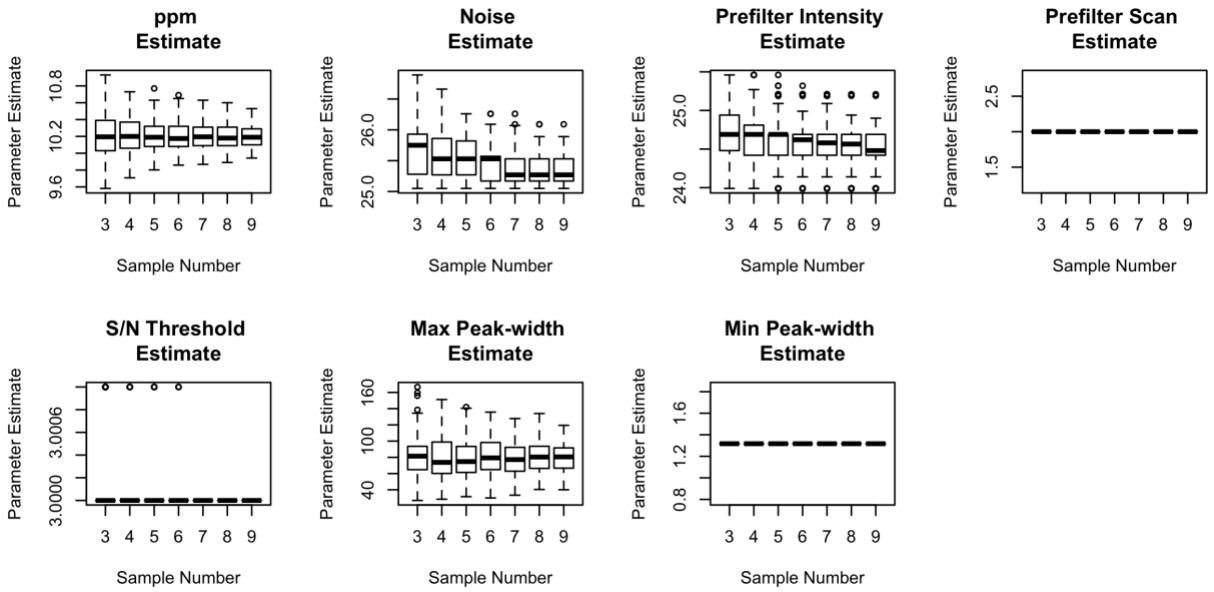


Figure S9. The parameters estimated in the Monte Carlo analysis on positive mode community data. Each plot denotes the calculated parameter estimate values for each unique parameter across 385 runs of AutoTuner. The x-axis describes the number of samples used to generate estimates, while the y-axis portrays the determined 55 parameter estimates within each n-sample subset (n = 3-9).

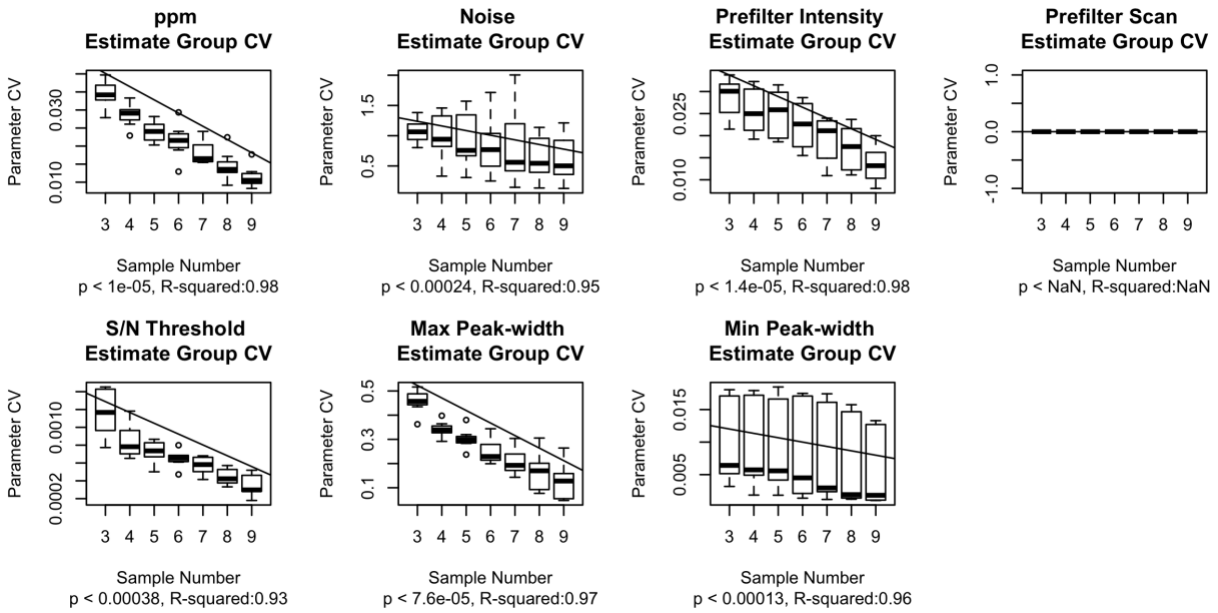


Figure S10. The coefficient of variation (CV) for groups of parameters estimated in the Monte Carlo analysis on negative mode culture data. Each plot denotes the calculated CV values for each unique parameter. The x-axis describes the number of samples used to generate estimates, while the y-axis describes the CV of the estimates from each group of 11 randomly selected samples. P-value and R^2 statistics are derived from linear regressions of data (n = 49). (NaN = not a number).

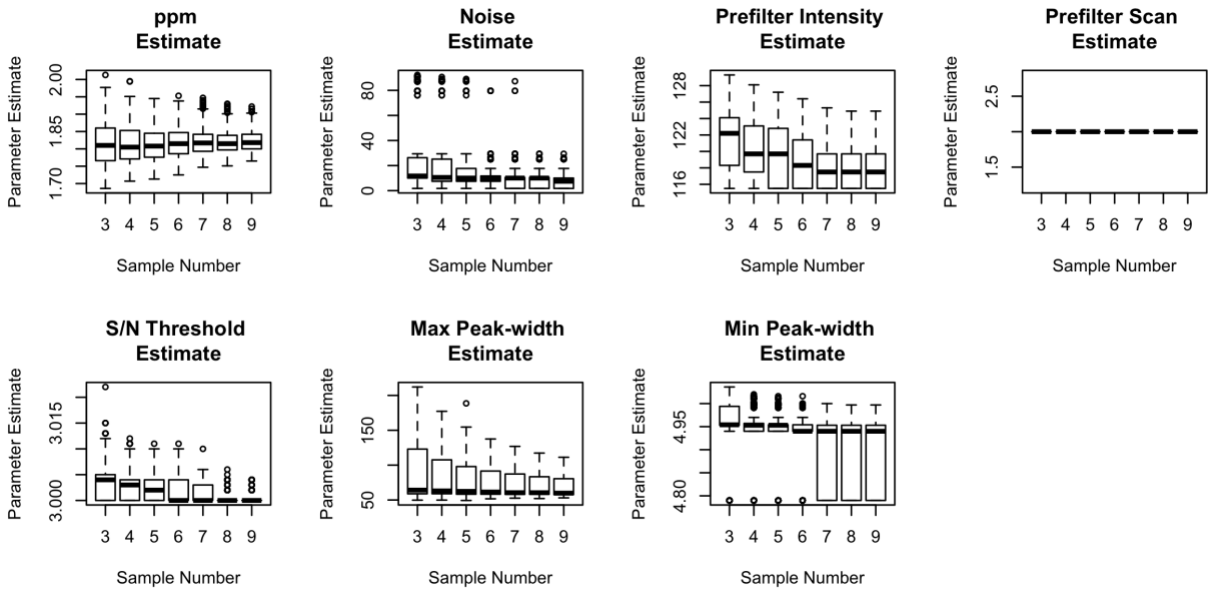


Figure S11. The parameters estimated in the Monte Carlo analysis on negative mode culture data. Each plot denotes the calculated parameter estimate values for each unique parameter across 385 runs of AutoTuner. The x-axis describes the number of samples used to generate estimates, while the y-axis portrays the determined 55 parameter estimates within each n-sample subset (n = 3-9).

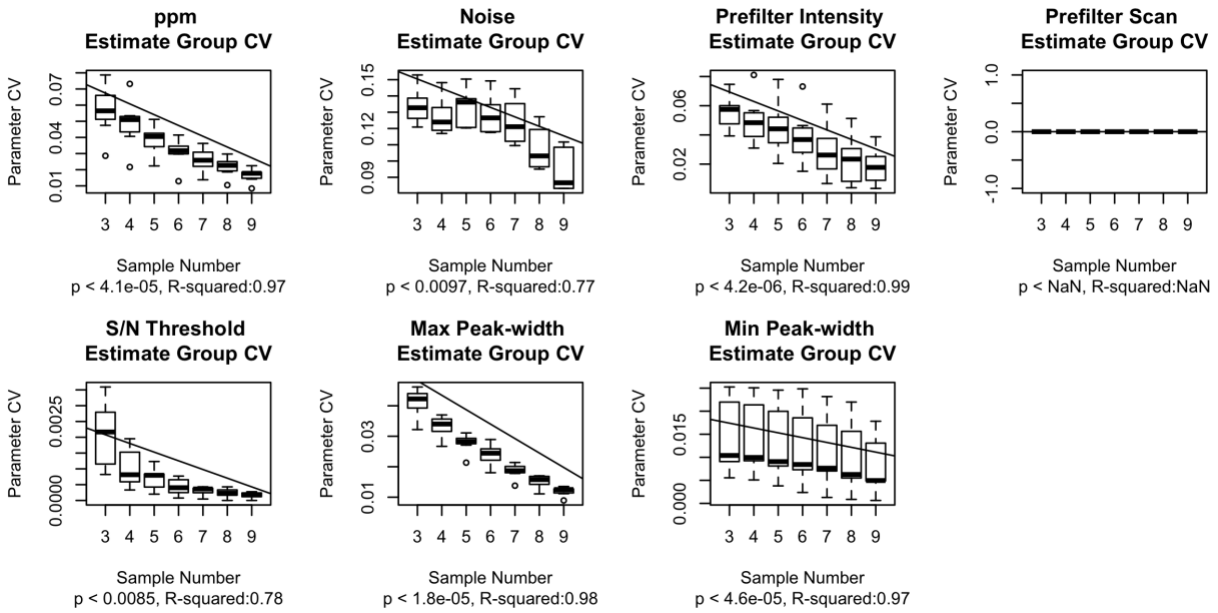


Figure S12. The coefficient of variation (CV) for groups of parameters estimated in the Monte Carlo analysis on positive mode culture data. Each plot denotes the calculated values for each unique parameter. The x-axis describes the number of samples used to generate estimates, while the y-axis describes the CV of the estimates from each group of 11 randomly selected samples. P-value and R^2 statistics are derived from linear regressions of data (n = 49). (NaN = not a number).

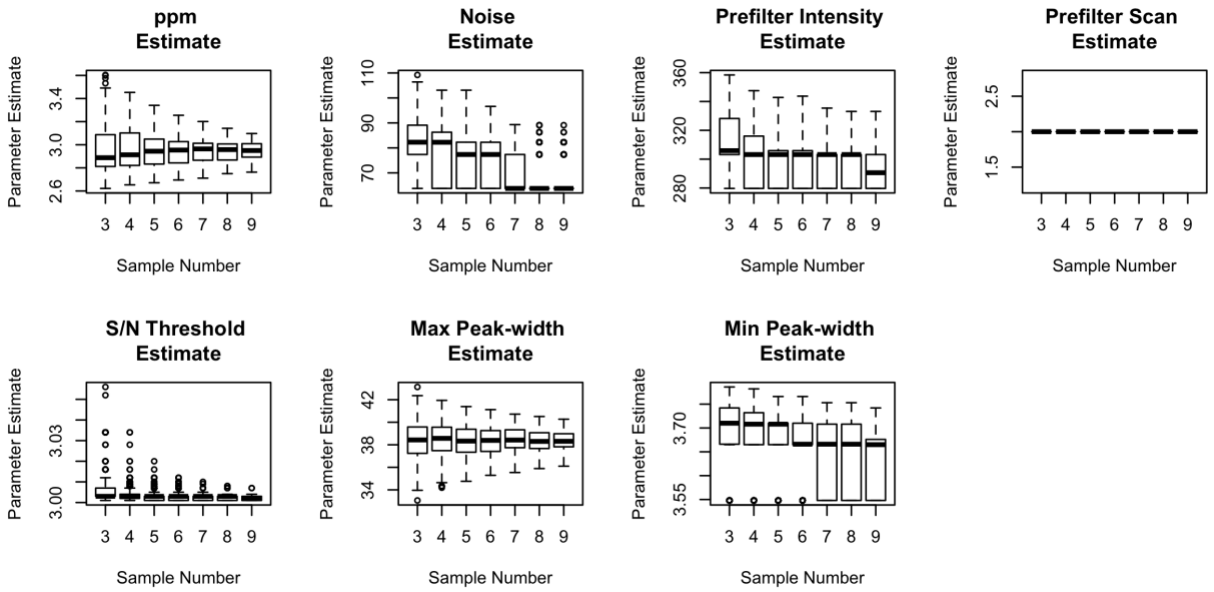


Figure S13. The parameters estimated in the Monte Carlo analysis on positive mode culture data. Each plot denotes the calculated parameter estimate values for each unique parameter across 385 runs of AutoTuner. The x-axis describes the number of samples used to generate estimates, while the y-axis portrays the determined 55 parameter estimates within each n-sample subset (n = 3-9).

2.8.2 Supplementary Tables

Table S1. Standards used to validate AutoTuner accuracy. These compounds are common targets of metabolism and are commonly detected within untargeted metabolomics experiments. Compounds detected in both ionization modes are separated by "|" in the order they were presented in the "Ionization Mode" column.

Compound	Ionization Mode	In AutoTuner	In IPO	m/z	Retention Time (s)
3-methyl-2-oxopentanoic acid	NEG	TRUE	TRUE	129.061|NA	237.7
3-methyl-2-oxobutanoic acid	NEG	TRUE	TRUE	115.05|NA	149.2
4-aminobenzoic acid	POS	TRUE	TRUE	NA|138.043	202.2
4-hydroxybenzoic acid	NEG	TRUE	TRUE	137.028|NA	235.4
4-methyl-2-oxopentanoic acid	NEG	TRUE	TRUE	129.061|NA	255.3
adenosine 5'-monophosphate (5'AMP)	NEG|POS	TRUE|TRUE	FALSE|TRUE	346.039|348.054	53.0
adenosine 3'-monophosphate (3'AMP)	NEG|POS	TRUE|TRUE	FALSE|TRUE	346.039|348.054	59.5
6-phosphogluconic acid	NEG	TRUE	TRUE	275.002|NA	32.9
acetyl taurine	NEG	TRUE	TRUE	166.017|NA	43.6
adenine	NEG|POS	TRUE|TRUE	TRUE|TRUE	134.053|136.063	52.4
adenosine	POS	TRUE	TRUE	NA|268.091	110.6
alpha-ketoglutaric acid	NEG	TRUE	TRUE	145.039|NA	53.0
4-amino-5-aminomethyl-2-	POS	TRUE	TRUE	NA|139.1	27.1

methylpyrimidine (AmMP)					
arginine	POS	TRUE	FALSE	NA|175.103	30.1
aspartic acid	NEG|POS	TRUE|TRUE	FALSE|TRUE	132.025|134.055	31.2
biotin	NEG|POS	TRUE|TRUE	FALSE|FALSE	243.069|245.073	266.2
caffeine	POS	TRUE	TRUE	NA|195.069	248.1
citric acid	NEG	TRUE	TRUE	191.005|NA	59.5
cytosine	POS	TRUE	TRUE	NA|112.054	34.2
desthiobiotin	NEG|POS	TRUE|TRUE	TRUE|TRUE	213.109|215.117	285.7
glucosamine phosphate	NEG	TRUE	FALSE	258.026|NA	30.6
pantothenic acid	NEG|POS	TRUE|TRUE	FALSE|TRUE	218.094|220.108	186.6
ribose 5-phosphate	NEG	TRUE	TRUE	229.006|NA	32.3
3-phosphoglyceric acid	NEG	TRUE	TRUE	184.986|NA	35.9
diacetylchitobiose	POS	TRUE	TRUE	NA|425.132	44.2
dihydroxy acetone phosphate	NEG	TRUE	TRUE	168.984|NA	32.3
dimethylsulfoniopropionate (DMSP)	POS	TRUE	TRUE	NA|135.052	31.9
ectoine	POS	TRUE	TRUE	NA|143.14	38.3
folic acid	NEG|POS	TRUE|TRUE	TRUE|TRUE	440.101|442.121	230.3
fosfomycin	NEG	TRUE	TRUE	137.011|NA	37.7
fumarate	NEG	TRUE	TRUE	115.007|NA	71.0
gamma-aminobutyric acid (GABA)	POS	TRUE	TRUE	NA|104.087	32.4
glucose 6-phosphate	NEG	TRUE	TRUE	259.004|NA	31.2
glutamic acid	NEG	TRUE	FALSE	146.047|NA	32.9
glutamine	POS	TRUE	FALSE	NA|147.073	31.3
glycine betaine	POS	TRUE	TRUE	NA|118.08	34.8

glyphosate	NEG	TRUE	TRUE	168.061|NA	31.7
guanine	POS	TRUE	TRUE	NA|152.064	53.0
guanosine	NEG|POS	TRUE|TRUE	TRUE|TRUE	282.06|284.099	137.6
4-methyl-5-thiazoleethanol (HET)	POS	TRUE	TRUE	NA|144.057	178.1
(4-amino-2-methyl-5-pyrimidinyl)methanol (HMP)	POS	TRUE	TRUE	NA|140.084	46.5
c 3-acetic acid	POS	TRUE	TRUE	NA|176.065	318.6
inosine	NEG	TRUE	TRUE	267.061|NA	138.5
inosine 5'-monophosphate	NEG|POS	TRUE|TRUE	TRUE|TRUE	347.022|349.037	57.1
isethionic acid	NEG	TRUE	TRUE	125.055|NA	34.1
citrulline	POS	TRUE	TRUE	NA|176.089	33.0
glutathione	POS	TRUE	TRUE	NA|308.053	77.7
glutathione oxidized	POS	TRUE	TRUE	NA|613.161	77.7
isoleucine	POS	TRUE	TRUE	NA|132.092	87.3
kynurenine	POS	TRUE	TRUE	NA|209.12	159.9
leucine	POS	TRUE	TRUE	NA|132.091	82.9
phenylalanine	POS	TRUE	TRUE	NA|166.079	166.3
tryptophan	POS	TRUE	TRUE	NA|205.084	213.9
tyrosine	POS	TRUE	TRUE	NA|182.105	81.5
methionine	POS	TRUE	FALSE	NA|150.052	57.1
5'methylthioadenosine (MTA)	POS	TRUE	TRUE	NA|298.081	209.8
muramic acid	NEG	TRUE	TRUE	250.086|NA	40.0
N-acetyl d-glucosamine	POS	TRUE	TRUE	NA|222.077	37.2
N-acetyl l-glutamic acid	NEG	TRUE	TRUE	188.054|NA	70.1

N-acetylmuramic acid	NEG|POS	TRUE|TRUE	TRUE|TRUE	292.085|294.121	109.6
ß-nicotinamide adenine dinucleotide (NAD)	NEG|POS	TRUE|TRUE	FALSE|FALSE	662.041|664.078	57.1
ß-nicotinamide adenine dinucleotide phosphate (NADP)	NEG	TRUE	TRUE	742.011|NA	53.9
ornithine	POS	TRUE	TRUE	NA|133.098	27.7
orotic acid	NEG	TRUE	TRUE	155.004|NA	50.1
phosphoenolpyruvate	NEG	TRUE	TRUE	166.970|NA	37.7
proline	POS	TRUE	TRUE	NA|116.076	32.2
pyridoxine	POS	TRUE	TRUE	NA|170.079	61.2
riboflavin	POS	TRUE	FALSE	NA|377.100	262.4
S-(1,2-dicarboxyethyl)glutathione	POS	TRUE	TRUE	NA|424.121	65.9
S-(5'-adenosyl) -L-homocysteine (SAH)	NEG|POS	TRUE|TRUE	TRUE|TRUE	383.054|385.062	78.7
S-adenosyl-l-methionine (SAM)	POS	TRUE	FALSE	NA|399.200	31.3
serine	POS	TRUE	FALSE	NA|106.052	30.7
sn-glycerol 3-phosphate	NEG|POS	TRUE|TRUE	TRUE|TRUE	170.999|173.004	32.3
succinic acid	NEG	TRUE	TRUE	117.022|NA	76.6
syringic acid	NEG	TRUE	TRUE	197.030|NA	266.2
taurine	NEG	TRUE	FALSE	124.012|NA	43.6
thiamine monophosphate	POS	FALSE	FALSE	NA|345.060	NA
threonine	POS	TRUE	TRUE	NA|120.069	31.9
thymidine	NEG	TRUE	TRUE	241.074|NA	173.8

triacetylchitotriose	POS	TRUE	TRUE	NA|628.269	53.0
uracil	POS	TRUE	TRUE	NA|113.051	114.0
uridine 5'-monophosphate	POS	TRUE	TRUE	NA|325.031	51.2
valine	POS	TRUE	TRUE	NA|118.091	34.8
xanthine	NEG|POS	TRUE|TRUE	TRUE|TRUE	151.017|153.045	161.0
xanthosine	NEG|POS	TRUE|TRUE	FALSE|TRUE	283.053|285.084	161.0

Table S2. Parameters used to process data. We rounded the values returned by AutoTuner and IPO at the tenths place. Each column aside from the "Dataset" and "Method" represent XCMS parameters described in Table 1. The community dataset is not mentioned here, as no comparison between IPO- and AutoTuner-parametrized feature tables was performed. The same standard set of parameters were used for density grouping and loess spline retention time correction. XCMS function syntax is described in parentheses. For the first run of density grouping (group.density): group difference =10, minfrac = 0, minsamp = 1, mzwid = 0.001. For the second run of density grouping after retention time correction (group.density):, group difference = 5, minfrac = 0.5, minsamp = 1, mzwid = 0.001. For loess spline retention time correction (retcor.peakgroups): span = 0.5.

Dataset	Method	*Maximum Peak-width*	*Minimum Peak-width*	*ppm*	*Noise*	*Prefilter Intensity*	*Scan Count*	*S/N Threshold*
Pos Standards	IPO	26.0	12.0	6.2	250.0	100.0	3.6	10
Pos Standards	AutoTuner	29.3	5.7	4.0	436.8	1421.3	2.0	6
Pos Culture	IPO	48.0	18.6	5.3	470	100.0	2.5	7
Pos Culture	AutoTuner	38.3	3.6	3.0	66.7	292.0	2.0	3
Neg Standards	IPO	26.0	12.0	6.2	250.0	100.0	3.6	10
Neg Standards	AutoTuner	29.3	5.7	4.0	436.8	1421.3	2.0	6
Neg Culture	IPO	60.0	27.4	4.7	121.0	100.0	4.0	9
Neg Culture	AutoTuner	66.9	4.9	1.8	7.8	117.5	2.0	3

Table S3. Feature count from each dataset during the different stages of quality assurance processing of culture data. The initial feature count was reduced after processing to remove blanks ('post blank)', features found in only one replicate ('post reproducibility), isotopologues and adducts ('post isotopes', and 'post adducts', respectively), and features with a CV greater than 0.4 in the pooled samples ('post CV').

Ionization Mode	Algorithm	Initial Feature Count	Post Blank	Post Reproducibility	Post Isotopes	Post Adducts	Post CV
Negative	IPO	40422	37903	8225	7695	4324	4226
Negative	AutoTuner	22599	17640	2921	2805	1444	1363
Positive	IPO	28794	28042	5907	5591	3628	3520
Positive	AutoTuner	13731	12451	2099	2012	1225	1143

Table S4. Counts of total detected features with MS/MS within figures 3 and S5 Venn diagrams.

Ionization Mode	AutoTuner MS^2 Count	IPO MS^2 Count	Intersect MS^2 Count
Positive	122	686	477
Negative	115	448	197

Table S5. Standard parameters used within centWave algorithm and their number of possible combinations. We cite these values in our discussion of speed improvements gained via AutoTuner relative to traditional parameter sweeping approaches dependent on optimization functions.

Parameter	Type	Possible Choices	Checked by AutoTuner
ppm	Continuous	Infinite	Yes
S/N Threshold	Continuous	Infinite	Yes
Scan count	Continuous	Infinite	Yes
Noise	Continuous	Infinite	Yes
Prefilter intensity	Continuous	Infinite	Yes
Minimum Peak-width	Continuous	Infinite	Yes
Maximum Peak-width	Continuous	Infinite	Yes
mzDiff	Continuous	Infinite	No
Fit gauss	Boolean	2	No
Mz center function	Discrete	4	No
Integrate	Discrete	2	No

Table S6. Number of unique features observed after processing data with unique mzDiff values. Columns two and three denote the mzDiff values used during pairwise comparisons of feature tables. Missing Count column represents the number of features observed outside the intersect of both feature tables. Feature tables were generated from 8 negative ion mode community data samples.

Missing Count	mzDiff value of First Feature Table	mzDiff value of Second Feature Table
0	-0.001	-0.002
0	-0.002	-0.003
0	-0.003	-0.004
0	-0.004	-0.005
0	-0.005	-0.006
0	-0.006	-0.007
0	-0.007	-0.008

Chapter 3:

Harmful Algal Bloom-Forming Organism Responds to Nutrient Stress Distinctly From Well-Studied Phytoplankton

3.1 Introduction

The scarcity of nitrogen (N) and phosphorus (P) limits primary production across aquatic ecosystems [10, 100] by regulating the growth and structure of phytoplankton communities [11, 101, 102]. These communities consist of an extremely diverse group of phylogenetically and physiologically distinct phytoplankton [17, 23]. The abundance of individual phytoplankton groups can vary widely over space and time [103], due in large part to their group-specific ecological strategies for managing resource limitation [18, 19]. These strategies vary from reallocation of intracellular nutrients [35, 45, 54], to reduction of nutrient quotas needed for growth [55, 56], and increased production of dissolved inorganic nutrient transporters [24, 25, 45].

The impact of distinct nutrient response strategies is most evident when considering harmful algal blooms (HABs). HABs may occur when a set of physiological traits allow a single phytoplankton group to prosper over its neighbors [13]. Such blooms have increased in frequency over the past 40 years with climate change and eutrophication [4], causing hundreds of millions of dollars in economic damages to fisheries and public health [7]. Many blooms have been linked to resource availability [5], and as a result, knowledge of nutrient response mechanisms and their associated trade-offs towards fitness holds tremendous promise in managing HABs [104]. The mechanics of how nutrients influence metabolic response strategies have been identified in only a few well-studied phytoplankton (*e.g.*, diatoms [57-59], coccolithophores [37, 60, 61]), leaving significant gaps in our understanding of nutrient responses in HAB-forming groups.

It is unclear whether metabolism data from well-studied phytoplankton reflects that of less well-studied phytoplankton. Although core metabolic functional redundancies occur in all primary producers [105], physiological studies suggest that phytoplankton groups vary widely in their capabilities beyond carbon processing [6]. For example, elemental stoichiometry studies observe that phytoplankton groups differ in their intracellular macromolecular pool composition [15], and transcriptome studies show that phytoplankton respond differently to environmental perturbations [19]. Therefore, direct evaluations of less well-studied phytoplankton metabolism, specifically under conditions of acute shortages of essential

nutrients (stress), are needed to evaluate the extent to which metabolic knowledge from well-studied phytoplankton can describe other groups.

One factor driving the knowledge discrepancy between well-studied and less well-studied phytoplankton groups is the paucity of fully sequenced eukaryotic genomes for most phytoplankton. These resources are required to build genome-scale models and evaluate systems level changes to stress [38, 106, 107]. Due to this constraint, investigators have used 'omics techniques like transcriptomics or metabolomics to characterize metabolic responses of phytoplankton to nutrient stress [57-61, 108-110]. While each of these methods offers substantial insights on physiological differences among phytoplankton [18], they fail to capture the systems level changes when used in isolation [37]. For example, metabolomic approaches provide evidence of biochemical reactions, but predicting the mechanism driving the activity is challenging. By contrast, transcriptomic techniques reveal pathway level changes, yet such changes may be inhibited by post-translational regulatory processes beyond the scope of the data. Applying metabolomic and transcriptomic methods in tandem circumvents these issues. Although computational challenges have typically limited these multi-'omics efforts to targeted analyses of specific pathways rather than systems level changes [45], combining 'omics techniques holds tremendous promise for understanding the diversity of phytoplankton responses to N- and P-stress.

In this study, we examined N- and P-stress metabolism using a combination of metabolomics and transcriptomics data for the HAB-forming raphidophyte, *Heterosigma akashiwo*. *H. akashiwo* populations are distributed ubiquitously within coastal subtropical environments [8, 50], and their blooms have caused significant economic losses [51]. Both N- and P-stress are known to be important drivers of *H. akashiwo* blooms [52, 53]. Our findings provide a mechanistic understanding of *H. akashiwo* stress response and suggest that a broader understanding of less well-studied phytoplankton metabolism is necessary to understand how phytoplankton communities will adapt to a changing ocean.

3.2 Materials and Methods:

3.2.1 Culture Maintenance:

We cultured *H. akashiwo* strain CCMP 2393 (isolated from Rehoboth Bay, Delaware, USA) in L1 medium (N:P 24, 882 µM $NaNO_3$, 36.2 µM NaH_2PO_4) made with autoclaved 0.2-µm filtered seawater from Vineyard Sound, MA. Cultures were not axenic, but were uni-algal and uni-eukaryotic. We grew each culture with light intensity of 100 µmol quanta m^{-2} s^{-1} of photosynthetically active radiation (400-700 nm) during a 14h:10h light:dark cycle at 18°C.

3.2.2 Experimental Design:

We used entrainment cultures to initiate experimental cultures to decrease carryover of nutrients from stock cultures and promote acclimation to the experimental conditions. We grew single entrainment cultures (~100 mL) for all organisms for three days in modified L1 medium (base seawater as above) under the following nitrogen and phosphorus conditions: replete (N:P = 16, 576 µM $NaNO_3$ and 36.2 µM NaH_2PO_4), N-stress (N:P = 0.1, 5 µM $NaNO_3$ and 36.2 µM NaH_2PO_4), and P-stress (N:P = 2880, 576 µM $NaNO_3$ and 0.2 µM NaH_2PO_4). We kept entrainment cultures at 18 °C on a 14:10 light:dark cycle (as above) with gentle rotation (75 rpm). We initiated experiments by inoculating triplicate 1-L flasks (containing 0.3 L media) for each treatment with 30 mL of the corresponding entrainment culture. We maintained experimental flasks under the same conditions as the entrainment cultures.

We monitored growth in each flask by *in vivo* chlorophyll fluorescence on a Turner Designs Aquafluor handheld fluorometer with paired cell counts. We preserved cell count samples in 2% (final concentration) acid Lugol's solution and we determined cell concentrations by microscopy for all time points except for those of the inoculum. For that, we report relative fluorescent units. We took cell concentration measurements at the same time each day (during the middle of the light phase) to avoid diel changes in metabolite synthesis. We harvested replete cultures in exponential phase, and harvested N-stress and P-stress cultures once growth rates and cell yields were reduced relative to the replete control, in agreement with the definition of nutrient stress rather than deficiency [111]. Specifically, we harvested treatments for metabolomics analysis on day 3, when we observed significant differences ($p < 0.001$,

Tukey-HSD Test, n = 3) in cell counts between stressed and replete cultures (Figure S1). We filtered cells (300 mL) of each replicate in each treatment onto combusted 47 mm GF/Fs (Whatman) using combusted glass filtration funnels under low vacuum pressure (never exceeding 5 mm Hg) to collect particulate metabolite samples. We flash-froze filters in cryovials in liquid nitrogen and stored them at −80°C until extraction.

3.2.3 Filter Extractions:

We split each filter in half for separate extractions for targeted and untargeted metabolomic analyses. The extraction procedure for untargeted and targeted samples were identical with the exception of the final solid phase extraction (see below). For both types of analyses, we first cut each filter half into six roughly equivalent pieces and placed them into an 8-mL amber glass vial. We extracted metabolites from filters using 1 mL of cold 40:40:20 acetonitrile: methanol:water + 0.1 M formic acid similar to previous work [112, 113]. We then added 25 µL of 1 µg/mL deuterated standard mix (d_3-glutamic acid, d_4-4-hydroxybenzoic acid, and d_5-taurocholate) as extraction recovery standards. We sonicated the solvent-filter mixture for 10 minutes to lyse the cells, and transferred the solvent into a microcentrifuge tube. We rinsed the filters with three 200-µL aliquots of extraction solvent to capture any remaining organic matter. We centrifuged the combined extracts at 20,000 × *g* for 5 minutes, and transferred the supernatant into clean 8-mL amber glass vials with care to leave behind any filter or cellular debris. We neutralized the extracts with 25.6 µL of 6 M ammonium hydroxide in water and dried them down to near dryness in a vacufuge. We reconstituted dried samples for targeted analysis in 200 µL 95:5 water:acetonitrile solution plus 2.5 µL of 5 µg/mL deuterated biotin injection standard.

For the untargeted analysis, a PPL extraction step is necessary to reduce high salt concentrations that can block the ion transfer tube within the mass spectrometer used for untargeted data [114]. We reconstituted these samples with 500 µL 0.1 M HCl to lower the pH to 2 and ran these samples through 100 mg/1 mL Agilent Bond Elut PPL cartridges. We pre-conditioned the cartridge with one cartridge-volume of 100% methanol and passed acidified untargeted samples through the cartridge at a flow rate below 40 mL min^{-1}. We rinsed the cartridges with one cartridge-volume of 0.01 M HCl, dried them down for 5 minutes, and eluted

the metabolites with one cartridge-volume of methanol. We dried untargeted samples again to near dryness and reconstituted them with 247.5 µL of 95:5 water:acetonitrile plus 2.5 µL of 5 µg/mL deuterated biotin injection standard. We combined 45 µL aliquots from each sample to create a pooled sample.

3.2.4 Liquid Chromatography and Mass Spectrometry:

We analyzed metabolite samples for untargeted analyses by high-performance liquid chromatography (HPLC, Micro AS autosampler and Surveyor MS Pump Plus, Thermo Scientific) coupled via electrospray ionization (ESI) to a hybrid linear ion trap- Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer (7T LTQ FT Ultra, Thermo Scientific). We separated metabolites on a Synergi Fusion reverse phase C_{18} column (4 µm, 2.0 x 150 mm, Phenomenex), equipped with a guard column and precolumn filter, and maintained at 35°C. We eluted the column with (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile at a flow rate of 0.25 mL min^{-1}. We held the column at 5% B for 2 min, ramped to 65% B over 18 min, quickly ramped to 100% B over 5 min, held at 100% B for 7 min and then equilibrated at 5% B for 8 min prior to the next injection (total run time = 40 min). We separately injected 20 µL of sample onto the HPLC column individually for positive and negative ion mode analyses. We externally calibrated the mass spectrometer just prior to analysis in positive and negative ion modes using the manufacturer's solutions. We optimized the capillary temperature and ESI voltage at 330°C and 4.2 kV in positive mode and at 365°C and 3.8 kV in negative mode. We maintained sheath gas, auxiliary gas, and sweep gas flow rates at 35, 5, and 2, respectively (arbitrary units) for both polarities. We collected MS and data dependent MS/MS scans as follows: (1) a full MS scan in the FT-ICR analyzer from 100-1000 *m/z*, with mass resolving power set to 100,000 (defined at *m/z* 400); and (2) collision-induced dissociation fragmentation scans (MS/MS) in the linear ion trap for the four most abundant ions in each full scan. We collected MS/MS spectra under dynamic exclusion with an exclusion time of 20 seconds. At the start of each batch, we injected the pooled sample multiple times to condition the column with the sample matrix and to stabilize peak retention times. We also analyzed the pooled sample every nine samples for quality assurance.

We analyzed targeted samples by ultrahigh-performance liquid chromatography (UHPLC, Accela Open Autosampler and Accela 1250 Pump, Thermo Scientific) coupled via heated electrospray ionization (H-ESI) to a triple quadrupole mass spectrometer (TSQ Vantage, Thermo Scientific) operated under selected reaction monitoring (SRM) mode. We set the spray voltage at 4000 V (positive mode) and 3200 V (negative mode). We set source gases at 55 (sheath) and 20 (aux gas), heated capillary temperature at 375 °C, and the vaporizer temperature at 400 °C. We performed chromatographic separation on a Waters Acquity HSS T3 column (2.1 × 100 mm, 1.8 μm) equipped with a Vanguard pre-column and maintained at 40 °C. We eluted the column with (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile at a flow rate of 0.5 mL min^{-1}. The gradient started at 1% B for 1 min, ramped to 15% B from 1-3 min, ramped to 50% from 3-6 min, ramped to 95% B from 6-9 min, held until 10 min, and ramped to 1% B from 10-10.2 min, with final re-equilibration at 1% B (total gradient time = 12 min). We made separate autosampler injections of 5 μL for positive and negative ion modes.

3.2.5 Standard Optimization:

We obtained authentic standards at the highest grade available from Sigma Aldrich for compounds outside of our existing targeted method [115]. We injected standards at concentrations of 1 μg/mL in Milli-Q water to optimize selected reaction monitoring (SRM) conditions (s-lens, collision energy, product ions). We monitored at least two SRM transitions (precursor-product ion pairs) for quantification and confirmation of each target compound, derived from optimization protocols that maximize analyte signal-to-noise. We determined the chromatographic retention time of each compound with standards dissolved in Milli-Q.

3.2.6 Data Processing:

We converted untargeted data files from proprietary Thermo RAW into mzML format using msConvert [91]. We processed these files using XCMS and AutoTuner [71, 92, 116] to generate a spreadsheet of features. We define features as chromatographic peaks with unique mass-to-charge (*m/z*) and retention time values, with relative abundances determined by their area. We subjected processed data to quality-control filtering by removing possible contaminants and non-reproducible features as described previously [117]. Briefly, we removed features within

blanks, features with a coefficient of variation higher than 0.2 within pooled samples, and features with low reproducibility across factor groups. We report feature intensities normalized by the cell counts from day 3.

We used MAVEN to integrate compound peak areas within targeted data [118]. We used an in-house MATLAB script to apply quality-control filtering and to quantify peak areas using a standard curve of 4 to 10 points. We retained metabolites for this analysis if the peak included a confirm ion, and the metabolite was present within two of three biological replicates for each treatment. We further culled the list by correcting for metabolite presence in procedural blanks.

3.2.7 Statistics and Data Analysis:

We used ANalysis Of VAriance (ANOVA) hypothesis testing to identify significantly different untargeted mass spectral features, targeted compound abundances, and growth time points. We identified significant pairwise-comparisons using Tukey's honestly significant difference test (Tukey-HSD Test). We used linear models to identify significant trends between Hessa et al. [119] and Wimley and White [120] amino acid hydrophobicity scales and feature retention time. We applied Benjamini-Hochberg corrections to control for type 1 error following all tests, and considered any *p-value* equal to or less than 0.05 to be significant.

We putatively annotated features to Kyoto Encyclopedia of Genes and Genomes (KEGG) compounds and tetrapeptides if feature masses were within 2.5 ppm error of the expected ion masses [121]. Tetrapeptide masses represent the set of all possible masses of any four amino acids linked together by a peptide bond. We used mummichog to match features to KEGG compounds [122]. Whenever possible, we matched MS/MS spectra with *in silico* modeled MS/MS spectra of known compounds from MetFrag to increase strength of annotation [123]. Due to prohibitive costs required to confirm all features with authentic standards, we focused on features pertinent to pathways involved in central metabolism and intracellular scavenging. See Note S1 for a description of the general trends within both the targeted and untargeted metabolomics data.

We obtained previously-published transcriptome data (see Table S1 from reference 17) from *H. akashiwo* strain CCMP 2393 grown under identical conditions to our study. We did not

collect our metabolite data at the same time as the transcriptomic data. Hence, any signal appearing across both datasets is considered to be highly biologically robust. We combined the expression of all reads mapping to a single KEGG ortholog to find the net expression of the putatively identified KEGG orthologs prior to analysis of the data. All transcriptomic comparisons were between data from stress and replete cultures [124]. We used Analysis of Sequence Counts (ASC) to identify transcripts with a posterior *p-value* (post-*p*) > 0.95 for a $\log_2$ fold change greater than 2 or less than -2. This value reflects the likelihood that the fold change is real based on the distribution of all transcripts across samples. ASC is an empirical Bayes method that estimates the prior distribution by modeling biological variability using the data itself, rather than imposing a negative binomial distribution. ASC has been shown to perform similarly to, though more conservatively than, other differential expression analyses implemented on data sets with and without replicates [124]. We considered individual transcripts satisfying either of these criteria to be significantly more or less abundant. When analyzing groups of transcripts together across nutrient treatments, we first normalized individual genes by the mean expression of that gene to remove baseline differences across genes. We applied the Wilcoxon-Test to check if a group of transcripts was significantly more or less abundant under a given stress condition relative to the control. We corrected all *p-values* for multiple comparisons using Bonferroni method. We considered any *p-value* equal to or less than 0.05 to be significant.

3.3. Results and Discussion:

Raphidophytes are ubiquitous in estuarine and coastal systems worldwide and their blooms cause severe damage to fisheries and local ecosystems [50]. Yet, they are understudied relative to other marine phytoplankton groups like diatoms and coccolithophores. Here we used a combined metabolomic and transcriptomic approach to build a conceptual model of how the raphidophyte *H. akashiwo* remodels its metabolism under N- and P-stress, and compare that to other phytoplankton taxa.

3.3.1 P-stressed cells catabolize lipids for sugar synthesis.

Central carbon metabolism is the biochemical hub connecting intracellular macromolecular pools and its net flux drives differences in intracellular nutrient stoichiometry [125, 126]. To understand the direction of this pathway under P-stress, we first sought to determine whether glycolysis or gluconeogenesis was taking place. These pathways interconvert sugars and organic acids, and are distinguished by a few non-reversible reactions. Hence, we evaluated the differential expression of transcripts unique to each pathway (Figure 1a). Within P-stressed cells, two out of three gluconeogenic transcripts were significantly more abundant (post-p > 0.95, ASC, $\log_2$(FC) > 2), while one of the three glycolytic transcripts was significantly more abundant (post-p > 0.95, ASC, $\log_2$(FC) > 2), suggesting that P-stressed cells may be transcriptionally supporting gluconeogenesis over glycolysis. The $\log_2$ fold change between the mean expression of gluconeogenic and glycolytic exclusive transcripts was 2.4, consistent with this hypothesis (Figure S2). To determine if increased gluconeogenic transcription resulted in a concomitant increase of pathway activity, we measured the activity associated with a downstream product of gluconeogenesis, trehalose [127]. Prior studies show trehalose enrichments within diatom *P. tricornutum* when fed with glycerol [128]. Glycerol enters central carbon metabolism as glycerol-3-phosphate, and must traverse through gluconeogenesis to become trehalose. Within P-stressed cells, trehalose concentrations were significantly (p < 0.05, Tukey-HSD Test, n = 3) enriched (Figure 1b). Additionally, we observed that transcripts for trehalose biosynthesis enzymes trehalose-6-phosphate synthase and trehalose-6-phosphatase had a $\log_2$ fold changes of 0.7 and 0.71, respectively. We next sought to confirm whether the combined signals from our transcriptomic analysis of gluconeogenesis and trehalose enrichment supported a net occurrence of gluconeogenesis. The overall gene expression of these genes was significantly higher within P-stressed cells ($p < 0.05$, Wilcoxon-Test, n = 5). Based on these findings, we conclude that P-stress favors gluconeogenesis in *H. akashiwo*.

Gluconeogenesis requires a flux of reduced carbon from the mitochondria. For reduced carbon to leave the mitochondria, it must avoid tricarboxylic acid (TCA) cycle oxidation via the glyoxylate shunt (GS) [129]. To determine whether GS or oxidation was more prevalent within P-stressed cells, we evaluated the differential expression of transcripts driving the bifurcation

between the pathways; isocitrate lyase (IL) for GS and isocitrate dehydrogenase (ICD) for oxidation. IL had a $\log_2$ fold change of 0.74 while ICD had a fold change of -0.73. Based on these trends, we hypothesize that P-stress transcriptionally increases the relative abundance of GS activity over oxidation (Figure 1a).

TCA cycle oxidation supports oxidative phosphorylation, which can lead to the production of reactive oxygen species (ROS) such as superoxide [130]. Changes in ROS concentrations can lead to oxidative stress. Hence, we sought to estimate the relative amount of oxidative stress within the cells to evaluate the hypothesis of decreased mitochondrial carbon oxidation under P-stress. We first calculated the ratio of glutathione (GSSG) to reduced glutathione (GSH), a measure of cellular oxidative stress [131]. This ratio was significantly ($p < 0.05$, Tukey-HSD, n = 3) decreased in P-stressed cells relative to replete cells, supporting our hypothesis (Figure S3b). Next, we evaluated the transcriptional patterns of genes responsible for ROS production during oxidative phosphorylation, those of Complex I and III [130]. We observed that overall expression of these genes was significantly ($p < 0.05$, Wilcoxon-Test, n = 26) depleted within P-stressed cells (Figure S3c), in agreement with the GSSG/GSH ratio. Future studies could test this result with ROS production measurements via fluorescent staining techniques [132]. Our combined metabolite and transcriptomic trends imply that under P-stress, *H. akashiwo* reduces carbon oxidation, relative to replete conditions. This decrease may support an enhanced flux through the GS under P-stress. Completion of isotope tracer experiments would confirm this hypothesis.

Mitochondrial carbon bound for gluconeogenesis may originate from the catabolism of triacylglyceride (TAG) lipids [126]. To be catabolized, TAGs must be solubilized by cholic acid derivatives [133]. Hence, we quantified taurocholate, a cholic acid derivative. P-stressed cells had nearly significant ($p < 0.07$, Tukey-HSD Test, n = 3) elevated taurocholate concentrations (Figure 1b). We hypothesize that the observed elevated concentrations may be due to increases in TAG catabolism. To test this idea, we evaluated the expression of TAG lipase enzyme TGL4. TAG lipase enzymes are required to mobilize TAGs stored within lipid droplets [134]. TGL4 was significantly more abundant (post-$p > 0.95$, ASC, $\log_2(FC) > 2$) within P-stressed cells. We next sought to confirm whether downstream TAG catabolism enzymes were also enriched under P-

stressed cells. For this, we evaluated the differential expression of transcripts from the carnitine shuttle (CS) and cytosolic lipid elongation pathways. The carnitine shuttle is considered the rate limiting step of lipid catabolism [135] and is regulated to oppose lipid elongation to avoid futile cycles [126]. We observed that one of three carnitine shuttle transcripts were significantly more abundant (post-$p > 0.95$, ASC, $\log_2(FC) > 2$) within P-stressed cells. By contrast, one of three lipid elongation enzymes was significantly less abundant (post-$p > 0.95$, ASC, $\log_2(FC) > 2$) within P-stressed cells. These results suggest that TAG catabolism is being upregulated by P-stressed cells (Figure 1a). Indeed, the combined metabolite abundance and transcription patterns from both datasets were significantly ($p < 0.05$, Wilcoxon-Test, n = 6) enriched within P-stressed cells, supporting this hypothesis. We sought to confirm evidence of carnitine shuttle activity by quantifying carnitine but its concentration was not significantly different in P-stressed cells than replete ones (Figure 1b). One possible explanation is that in preparation for TAG degradation, P-stressed cells accumulate acyl-carnitines [136]. These molecules were not detected by our analytical method and future small molecule quantification experiments are needed to confirm this hypothesis. Additional targeted lipidomic analysis measuring TAG concentrations would serve to validate our hypothesized trends.

These concerted pathway level changes suggest that under P-stress, *H. akashiwo* drives gluconeogenesis via the catabolism of TAG carbon (Figure 1c). The system-level dynamics are most similar to those of P-stressed diatoms, as these organisms are hypothesized to upregulate gluconeogenesis and the carnitine shuttle under P-stress [57, 58]. By contrast, P-stressed metabolism in the coccolithophore, *Emiliania huxleyi,* appears to be quite different. Prior studies reported that this organism upregulates gluconeogenesis, glycolysis, and TAG synthesis in tandem [37, 60]. To our knowledge, this is the first report of P-stress driven glyoxylate shuttle transcription and trehalose enrichment among phytoplankton groups. These distinguishing features may underlie physiological differences exhibited between *H. akashiwo* and other closely related phytoplankton.

3.3.2 N-stressed cells increase respiration and store excess carbon as lipids.

Like our analysis of P-stressed cells, we sought to determine the direction of central carbon metabolism in N-stressed cells. Again, we evaluated whether glycolysis or gluconeogenesis was upregulated by examining the differential expression of pathway-specific transcripts. The results were equivocal, with one of three gluconeogenic transcripts significantly more abundant (post-p > 0.95, ASC, $\log_2$(FC) > 2) within N-stressed cells, and one of three glycolytic transcripts significantly more abundant (post-p > 0.95, ASC, $\log_2$(FC) > 2) within N-stressed cells, suggesting either pathway may be taking place (Figure 1a). It is possible that our calculation integrates gene expression from versions of the pathway that are localized to the cytosol and the plastid, which may be differentially regulated under N-stress [57]. Constraining the cellular location is not possible due to the lack of a fully sequenced genome for *H. akashiwo*. In order to differentiate the activity between pathways, we calculated a $\log_2$ fold change of -1.2 between the averaged expression of gluconeogenic relative to glycolytic exclusive transcripts, suggesting that gluconeogenic transcripts are less abundant than glycolytic ones within N-stressed cells (Figure S2). In addition, trehalose concentrations were not significantly enriched (Figure 1b), pointing to lower gluconeogenic activity in N-stressed cells. Based on these combined findings, we hypothesize that N-stressed cells favor glycolysis over gluconeogenesis.

Increased glycolytic carbon flux would foster greater TCA cycle oxidation. To determine if oxidation was preferentially upregulated in N-stressed cells, we examined the differential expression and the $\log_2$ fold change of isocitrate lyase (IL) to isocitrate dehydrogenase (ICD). We observed that ICD was significantly more abundant (post-p > 0.95, ASC, $\log_2$(FC) > 2) under N-stress (Figure 1a) and the ratio of IL to ICD had a $\log_2$ fold change of -0.98, supporting the hypothesis that N-stressed cells increase carbon oxidation (Figure S2). We then evaluated the differential expression of transcripts downstream of ICD-catalyzed oxidation reactions. We observed that three out of five transcripts were significantly more abundant (post-p > 0.95, ASC, $\log_2$(FC) > 2), supporting our hypothesis of increased oxidation (Figure 1a). To evaluate if transcriptional upregulation patterns resulted in a tandem increase in TCA cycle activity, we quantified citrate because cytosolic citrate accumulation is a common signature of upregulated TCA cycle flux [126]. We observed that citrate concentrations were significantly higher (p <

0.05, Tukey-HSD Test, n = 3) under N-stress, consistent with our hypothesis of increased TCA cycling (Figure 1b). To determine if N-stressed cells experienced elevated oxidative stress from TCA cycling, we evaluated glutathione recycling and Complex I and III genes of the electron transport chain. Although the ratio of glutathione (GSSG) to reduced glutathione (GSH) was not significantly different under N-stress, the concentrations of both GSSG and GSH were significantly depleted ($p < 0.05$, Tukey-HSD Test, n = 3) (Figure 1b) and transcripts driving glutathione recycling were significantly more abundant ($p < 0.05$, Wilcoxon-Test, n = 6) (Fig S3a). Both GSSG and GSH are N-rich molecules; hence their depletion may be due to N reallocation. The increased transcription of glutathione recycling enzymes may suggest that N-stressed cells overcome the decrease in reduced glutathione concentrations by increasing its recycling rates. In line with this observation, we observed that transcripts annotated for genes of Complex I and III of the electron transport chain involved in oxygen radical synthesis were significantly enriched under N-stress ($p < 0.001$, Wilcoxon-Test, n = 26) (Figure S3c). Together, these findings support the hypothesis that N-stressed cells upregulate TCA cycle carbon oxidation.

Excess cytosolic citrate may be converted to TAG lipids to limit the repression of glycolysis [126]. One prior study shows that N-stressed *H. akashiwo* cells are enriched with TAG lipids [137]. In our dataset, concentrations of taurocholate, a TAG mobilization marker, were significantly higher ($p < 0.05$, Tukey-HSD Test, n = 3) within N-stressed cells, suggesting elevated TAG mobilization under N-stress (Fig 1b). To determine if mobilization supported the storage of newly synthesized TAGs, we quantified the precursor to TAG lipid backbone glycerol, glycerol-3-phosphate. This compound was significantly enriched ($p < 0.05$, Tukey-HSD Test, n = 3) in N-stressed cells (Figure 1b). We next investigated changes in transcription of TAG synthesis genes. Specifically, we evaluated the transcriptome changes of the carnitine shuttle and enzymes initiating lipid elongation. While none of these transcripts were significantly more or less abundant, the $\log_2$ fold change of the averaged expression of carnitine shuttle and lipid elongation exclusive transcripts showed a value of -0.38, suggesting that elongation was slightly upregulated relative to the carnitine shuttle (Figure S2). Although we did not measure TAGs

within this study, our results suggest elevated TAG production due to enrichment of precursors and depletion of TAG catabolism transcripts.

These concerted pathway-level changes suggest that N-stress drives TAG synthesis via an increase in glycolysis (Figure 1c). Our findings match previously-published isotope tracer experiments on N-stressed *H. akashiwo* [138]. Similar observations of increased glycolysis, TCA cycling, and TAG synthesis have been reported for both N-stressed diatoms and haploid *E. huxleyi* [59, 61, 139, 140]. Unlike several diatoms and diploid *E. huxleyi, H. akashiwo* enriches citrate [37, 141], which can allosterically inhibit glycolysis or increase TAG synthesis rates [126]. This citrate enrichment and its role in regulation may distinguish phytoplankton like *H. akashiwo* from others.

N-stress has been reported to be linked to the initiation of the diel migration employed by *H. akashiwo* during blooms [53, 142]. The diel migration is hypothesized to help *H. akashiwo* acquire dissolved nutrients below the thermocline [50]. Sinking rates are linked to TAG accumulation and the resulting increase in cellular specific gravity [143-145]. Hence, TAG synthesis *in situ* may be linked to an increased in glycolysis and TCA cycle oxidation as presented here for N-stressed cells. The system level dynamics described here should be considered when building models of *H. akashiwo* blooms.

3.3.3 Nutrient stress responses use central metabolism in opposite ways

H. akashiwo appears to use central carbon metabolism in two distinct ways to overcome N- and P-stress, in contrast to haploid *E. huxleyi*, which responds to N- and P-stress similarly [37, 60]. Our hypothesis of stress-specific metabolic dynamics is based on the enrichment and proposed sources of central carbon metabolism storage molecules, trehalose and TAGs. These molecules vary drastically in their potential to contribute towards future biomass and other cellular functions. Trehalose can enter glycolysis after one reaction (cleavage of the disaccharide bond) [146], and may be quickly redirected towards the synthesis of nucleic and amino acids [125]. In contrast, TAGs must be converted into sugars via gluconeogenesis before this is possible. However, TAGs may be broken down into acetyl-CoA directly during beta-oxidation, thus providing far higher amounts of ATP per carbon than trehalose which must

traverse glycolysis prior to its arrival to the TCA cycle (see Note S2). It is critical to consider these trade-offs when building models to describe *H. akashiwo* ecosystem processes, as accumulation of either would result in distinct impacts on fitness [104]. As TAGs are far more carbon-rich than trehalose, the differences in orientation of central carbon metabolism may also explain why N-stressed cells had a greater measured C:N (14.37) ratio than that (8.95) of P-stressed cells [25]. These metabolic nuances and their hypothesized physiological consequences underscore the importance of understanding the metabolism of more less well-studied phytoplankton to develop models to characterize bloom and nutrient cycling dynamics [32]. These findings would not have been possible without the tandem analysis of transcriptomic and metabolomic data [37]. Hence, our approach may serve investigators of raphidophytes and other less well-studied phytoplankton in similar ways.

3.3.4 Intracellular recycling is pervasive under N-stress

Phytoplankton, including *H. akashiwo*, employ various strategies in response to nutrient stress [24, 35, 45, 54-56]. A prior transcriptomic-based study suggested that under N- and P-stress, *H. akashiwo* increases extracellular inorganic nutrient scavenging transporters, and upregulates urea cycle driven N-recycling and pigment catabolism [25]. However, transcriptome studies are insufficient to confirm these processes in most regards without metabolite data as indicators of the biological cascade initiated by transcription. For example, prior work with *H. akashiwo* was unable to identify which amino acids drive the urea cycle [25]. Here, we evaluated our data to uncover possible intracellular macromolecular recycling processes.

Previously, it was shown that urea cycle transcripts are significantly more abundant within N-stressed *H. akashiwo* cells [25]. Indeed, our data also show that urea cycle intermediates are enriched within N-stressed cells, confirming these trends (see Note S2 and Fig S4). However, the source of these compounds is unclear and could include either extracellular uptake or the degradation of endogenous protein. Prior studies show that *H. akashiwo* favors assimilation of inorganic N over organic N relative to sympatric phytoplankton [147], suggesting that urea cycling is supported from an endogenous source of amino acids. Hence, we hypothesized that *H. akashiwo* may sustain increased urea cycling by degrading endogenous proteins.

One endogenous protein degradation mechanism is proteasome-mediated enzyme degradation (PMED) [148]. PMED is an ATP-dependent process where enzymes are tagged with ubiquitin, shuttled into the proteasome, and released as peptides between 4 and 20 amino acids in length (Figure 2a). We observed that putatively-identified tetrapeptides from untargeted data were significantly depleted ($p < 10^{-10}$, Kruskal-Wallis Test, n = 597) under both P- and N-stress, suggesting differential activity of PMED in both stress cultures (see Note S3). These trends may be explained by either the deactivation of the proteasome under stress or the increased of removal of residual tetrapeptides by peptidase enzymes. To evaluate between these two possible scenarios, we gathered transcripts involved in enzyme ubiquitin tagging (ubiquination), proteasome biosynthesis, and downstream peptide cleavage (peptidases). We observed that transcripts for eight of ten subprocesses within PMED were significantly more abundant ($p < 0.05$, Wilcoxon-Test) in N-stressed cells, while three of ten were significantly more abundant ($p < 0.05$, Wilcoxon-Test) and one was significantly less abundant ($p < 0.05$, Wilcoxon-Test) in P-stressed cells (Figure 2b). These results suggest that trends in untargeted data within P-stressed cells may be due to a decrease in proteasome activity, while trends in N-stressed cells may be due to increased breakdown of peptides. To test whether this transcriptional upregulation corresponded with an increase in peptide degradation activity, we quantified hydroxyproline. Hydroxyproline is synthesized via post-translational modifications of proteinaceous proline through the activity of enzyme prolyl 4-hydrolase [149], hence its cytosolic concentration serves as evidence of protein degradation within diatoms [36]. Hydroxyproline concentrations were significantly enriched ($p < 0.05$, Tukey-HSD Test, n = 3) in N-stressed cells exclusively, supporting our prior hypothesis. Our paired transcriptomic and metabolomic datasets suggest that *H. akashiwo* upregulates proteasome enzyme degradation to overcome N-stress. To our knowledge, this is the first indicator of PMED as an N-stress mitigation strategy in phytoplankton. PMED is highly specific, hence it may contribute to an observed proteome-level depletion of N-rich proteins, as observed in the N-stressed green alga *Chlamydomonas reinhardtii* [150].

In addition to amino acid recycling, we sought to constrain whether nucleotide recycling enabled a reallocation of N within N-stressed *H. akashiwo*. To check this hypothesis, we first

quantified 17 distinct intermediates within nucleic acid metabolism (Figure 3a). Surprisingly, we observed an enrichment of nucleic acid bases and nucleosides and a significant depletion of nucleotide monophosphates (NMPs) (Figure 3a). Indeed, five of eleven measured nucleosides or nucleobases were significantly enriched ($p < 0.05$, Tukey-HSD, n = 3) under N-stress. Similar metabolite trends were reported in N-stressed yeast due to the autophagy-mediated breakdown of ribosomes and other nucleic acids [151]. To investigate if this mechanism could support metabolite enrichments, we gathered the gene expression of 21 DNA and/or RNA degradation enzymes (Figure 3b). These transcripts were significantly more abundant within N-stressed cells ($p < 10^{-10}$, Wilcoxon-Test, n = 21), suggesting that autophagic breakdown may drive our observed metabolite trends. Unfortunately, none of our annotated transcripts corresponded to reactions driving interconversion reactions between distinct purines or pyrimidines. This may be due to challenges assigning putative gene homologs within this pathway, as studies of this pathway in yeast faced similar obstacles [151].

To determine if nucleic acid recycling within *H. akashiwo* produced a downstream response similar to yeast, we evaluated the activity of the pentose phosphate pathway (PPP). In yeast, carbon scavenged from NMPs leads to both an enrichment of PPP metabolite ribose-5-phosphate and an increase in non-oxidative PPP activity [151]. We observed that ribose-5-phosphate concentrations were significantly higher ($p < 0.05$, Tukey-HSD, n = 3) under N-stress (Figure 3a), similar to yeast. However, when we gathered transcripts for non-oxidative PPP reactions, we observed that three of four non-oxidative PPP transcripts were significantly less abundant (post-$p > 0.95$, ASC, $\log_2(FC) > 2$) within N-stressed cells, contrary to our expectation (Figure S5). This result suggests that ribose-5-phosphate derived from nucleic acids degradation may support an function other than non-oxidative PPP activity within *H. akashiwo*.

One possible alternative is the biosynthesis of aromatic amino acids. These compounds all originate from erythrose-4-phosphate, a downstream product of ribose-5-phosphate [125]. We evaluated the activity of the aromatic amino acid biosynthesis pathway (Figure 3a and Figure S5) and observed that two of ten transcripts were significantly more abundant (post-$p >$ 0.95, ASC, $\log_2(FC) > 2$) and that transcription of enzymes within the pathway was enriched, although not significantly ($p < 10^{-7}$, Wilcoxon-Test, n = 14), in N-stressed cells. Additionally, all

aromatic amino acid biosynthesis metabolites with the exception of phenylalanine were significantly enriched ($p < 0.05$, Tukey-HSD Test, n = 3) in N-stressed cells (Figure 3a). Taken together, these observations suggest that ribose scavenged from NMP degradation drives aromatic amino acid biosynthesis (Figure 3c). Within plants and algae, aromatic amino acids serve a variety of processes due to their diverse functions ranging from electron carriers to natural products and chemical signals [152]. For example, tryptophan secreted by the diatom *Thalassiosira pseudonana* was shown to support mutualistic growth with a sympatric microbe [21]. *H. akashiwo* may rely on a similar strategy to overcome N-stress. Studies on diatoms have noted that nutrient stress increases PPP activity [58]. Nucleoside enrichment patterns have been observed in other marine phytoplankton, as inosine was also enriched in N-stressed *Prochlorococcus* [153]. However, to our knowledge, we provide the first evidence of either the connection between the PPP and nucleic acid degradation or its connection to aromatic amino acid biosynthesis.

Intracellular recycling appears to be a critical N-stress response strategy within *H. akashiwo*. Recycling may support *H. akashiwo* life history and behavior changes such as diel migration in addition to supporting sustained growth under resource limiting situations. Modeling of *H. akashiwo* will need to consider intracellular recycling and its impact on fitness for predicting behavior and bloom dynamics. The extent to which these processes uniquely define the niche of *H. akashiwo* or other raphidophytes, relative to other phytoplankton is uncertain and underscores the need for additional multi-'omics studies across a range of phytoplankton lineages.

3.3.5 Phytoplankton nutrient stress biomarkers reveal stress status within *H. akashiwo*

Our data allowed us to evaluate the efficacy of proposed phytoplankton N-stress (glutamine: glutamate) and P-stress (adenosine monophosphate (AMP): adenosine) diagnostics within *H. akashiwo* [45, 148]. We observed that N-stressed cells had significantly lower ($p < 0.05$, Tukey-HSD Test, n = 3) glutamine-to-glutamate ratios relative to the other treatments, while the P-stressed cells had significantly lower ($p < 0.05$, Tukey-HSD Test, n = 3) AMP-to-adenosine ratios (Figure 4). To our knowledge, this is the first evidence of the utility of these stress diagnostics in any raphidophyte. Measurements of these ratios in field raphidophyte

populations would serve as an important new approach for identifying whether a population is experiencing N- or P-stress, and defining the resource controls on bloom dynamics *in situ*.

3.6 Conclusions

This work employs a multi-omics approach to explore the impact of N- and P-stress on the HAB-forming raphidophyte, *H. akashiwo*. We characterized the stress-mediated system-level changes within central carbon metabolism and observed that intracellular recycling of macromolecules is pervasive under stress (Figure 5). Under N- and P-stress, *H. akashiwo* showed similar central carbon metabolism acclimation patterns as other well studied phytoplankton under nutrient stress. However, fine scale enrichment of distinct molecules distinguished its metabolic shifts from more frequently-studied phytoplankton (Figure 5). Identifying these differences would not have been possible without a multi-'omics approach. Evidence of novel intracellular recycling pathways could support sustained growth of *H. akashiwo* under conditions of low N and may be a mechanism underpinning niche segregation among competitors more reliant on N uptake. Taken together, these insights suggest that nutrient stress has distinct physiological impacts on *H. akashiwo* relative to diatoms. Our results suggest metabolism data from well-studied phytoplankton does not capture the nuances of less well-studied phytoplankton completely. Hence, more characterizations of metabolic stress responses within other phytoplankton are critical to accurately understand phytoplankton community composition and function in a changing ocean [32].

3.7 Figures

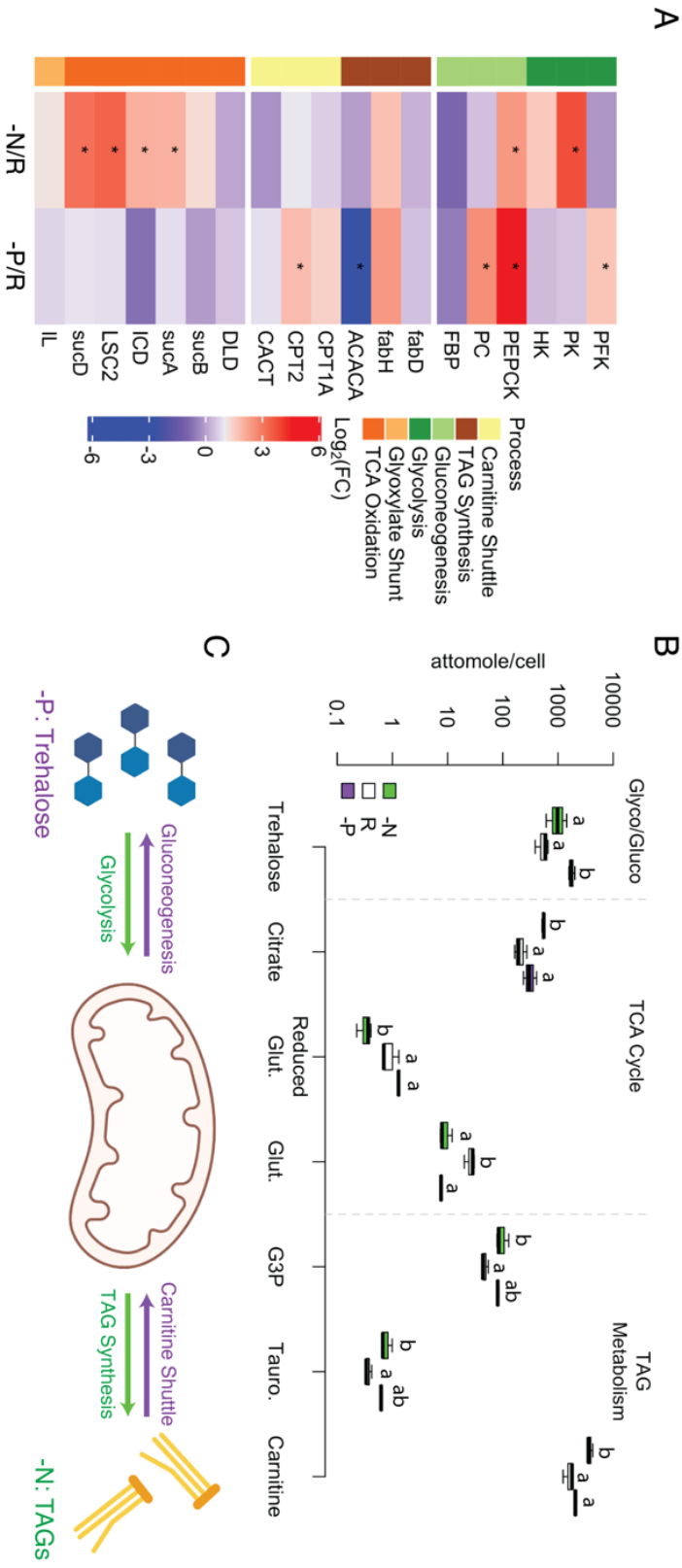

A
PFK
PK
HK
PEPCK
PC
FBP
fabD
fabH
ACACA
CPT1A
CPT2
CACT
DLD
sucB
sucA
ICD
LSC2
sucD
IL
-N/R
-P/R
Process
Carnitine Shuttle
TAG Synthesis
Gluconeogenesis
Glycolysis
Glyoxylate Shunt
TCA Oxidation
Log2(FC)
6
3
0
-3
-6
B
attomole/cell
10000
1000
100
10
1
0.1
-N
R
-P
Glyco/Gluco
TCA Cycle
TAG Metabolism
Trehalose
Citrate
Reduced Glut.
Glut.
G3P
Tauro.
Carnitine
C
-P: Trehalose
Gluconeogenesis
Glycolysis
Carnitine Shuttle
TAG Synthesis
-N: TAGs

Figure 1: Central carbon metabolism stress response. A) Transcriptomic data on irreversible steps of gluconeogenesis, glycolysis, TCA oxidation, glyoxylate shunt, carnitine shuttle (CS), and triacylglyceride (TAG) lipid synthesis. Data are from N-stress or P-stress normalized to the condition in the replete cells (-N/R and -P/R, respectively). * - denotes significance defined as a $\log_2(FC) > 2$ and ASC post-$p > 0.95$. B) Metabolomics data supporting hypothesized pathway activity. We present data of individual metabolites in the order of N-stressed (-N, green), nutrient replete (R, white), and P-stressed (-P, purple) treatments. Boxes with distinct letters above them are significantly different as defined by pairwise Tukey-HSD test $p < 0.05$. C) Hypothesized fluxes of central carbon metabolism under N-stress (green) and P-stress (purple). Transcript abbreviations: PFK – phosphofructokinase, PK – pyruvate kinase, HK – hexose kinase, PEPCK – phosphoenolpyruvate carboxykinase, PC – pyruvate carboxylase, FBP – fructo-bisphosphate phosphatase, fabD – malonyl carrier protein transacylase, fabH – oxoacyl carrier protein synthase III, fabH – oxoacyl carrier protein synthase III, ACACA – acyl-CoA carboxylase, CPT1A – carnitine O-palmitoyltransferase 1, CPT2 – carnitine O-palmitoyltransferase 2, CACT – mitochondrial carnitine/acylcarnitine transporter, DLD – dihydrolipoamide dehydrogenase, sucB – 2-oxoglutarate dehydrogenase complex (dihydrolipoamide dehydrogenase), sucA – 2-oxoglutarate dehydrogenase complex (E1 component), ICD – isocitrate dehydrogenase, LSC2 – succinyl-CoA synthetase (beta subunit), sucD – succinyl-CoA synthetase (alpha subunit), IL – isocitrate lyase, FC – fold change. Metabolomic abbreviations: Glut. – glutathione, G3P – glycerol-3-phosphate, Tauro. – taurocholate.

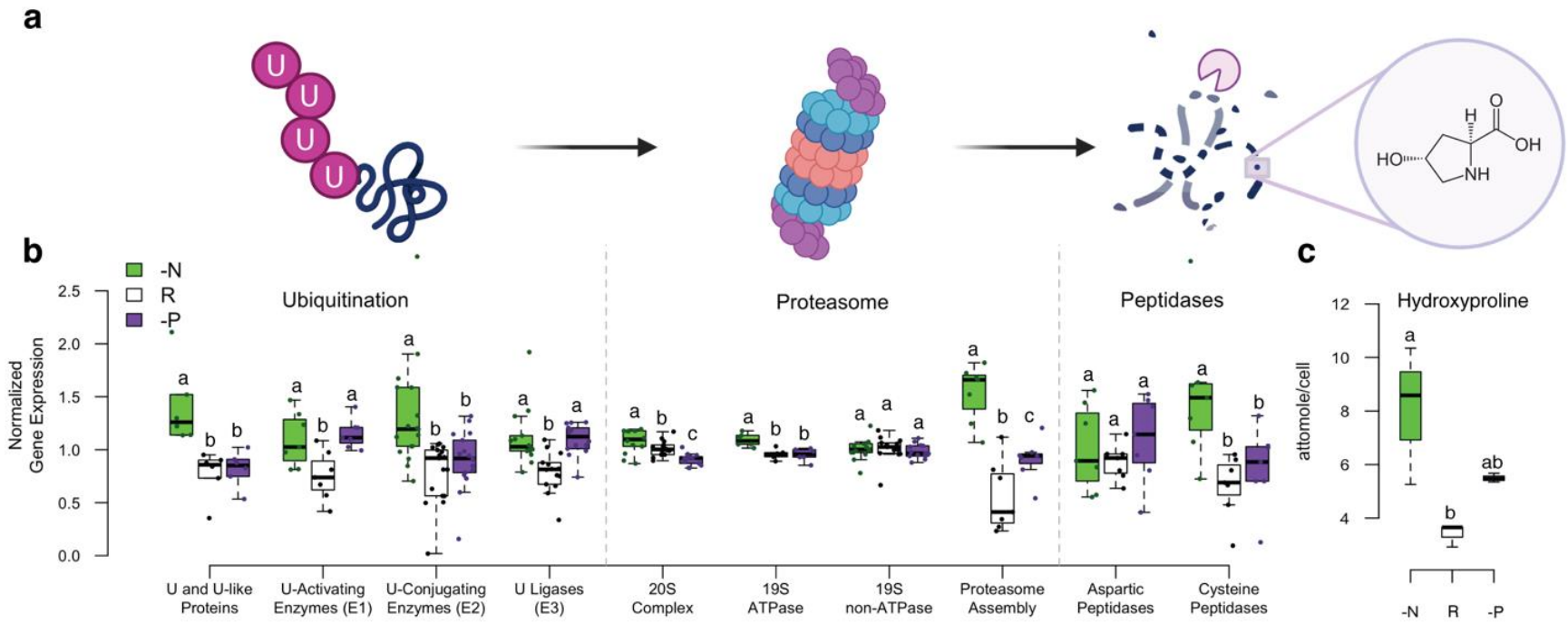

a
U
U
U
U
HO
NH
OH
H
O
b
-N
R
-P
Normalized Gene Expression
2.5
2.0
1.5
1.0
0.5
0.0
Ubiquitination
Proteasome
Peptidases
U and U-like Proteins
U-Activating Enzymes (E1)
U-Conjugating Enzymes (E2)
U Ligases (E3)
20S Complex
19S ATPase
19S non-ATPase
Proteasome Assembly
Aspartic Peptidases
Cysteine Peptidases
c
Hydroxyproline
attomole/cell
12
10
8
6
4
-N
R
-P

Figure 2: Proteasome mediated enzyme degradation. A) Schematic overview of proteasome enzyme degradation process. Enzymes tagged with ubiquitin (U) are brought into the proteasome to be broken down into peptides between 4 and 20 amino acids in length. Peptides are then broken down by cytosolic peptidases. B) Normalized gene expression values for distinct processes related to enzyme ubiquitination, the proteasome, and cytosolic peptidases. Genes were normalized by the mean expression of enzymes across all treatments. Boxes with distinct letters above them are significantly different as defined by pairwise Wilcoxon Test and Bonferroni correction $p < 0.05$. C) Hydroxyproline concentrations, an enzyme degradation by-product. Boxes with distinct letters above them are significantly different as defined by pairwise Tukey-HSD test $p < 0.05$. Abbreviations: -N - N-stressed, R - nutrient replete, -P - P-stressed, U – ubiquitin.

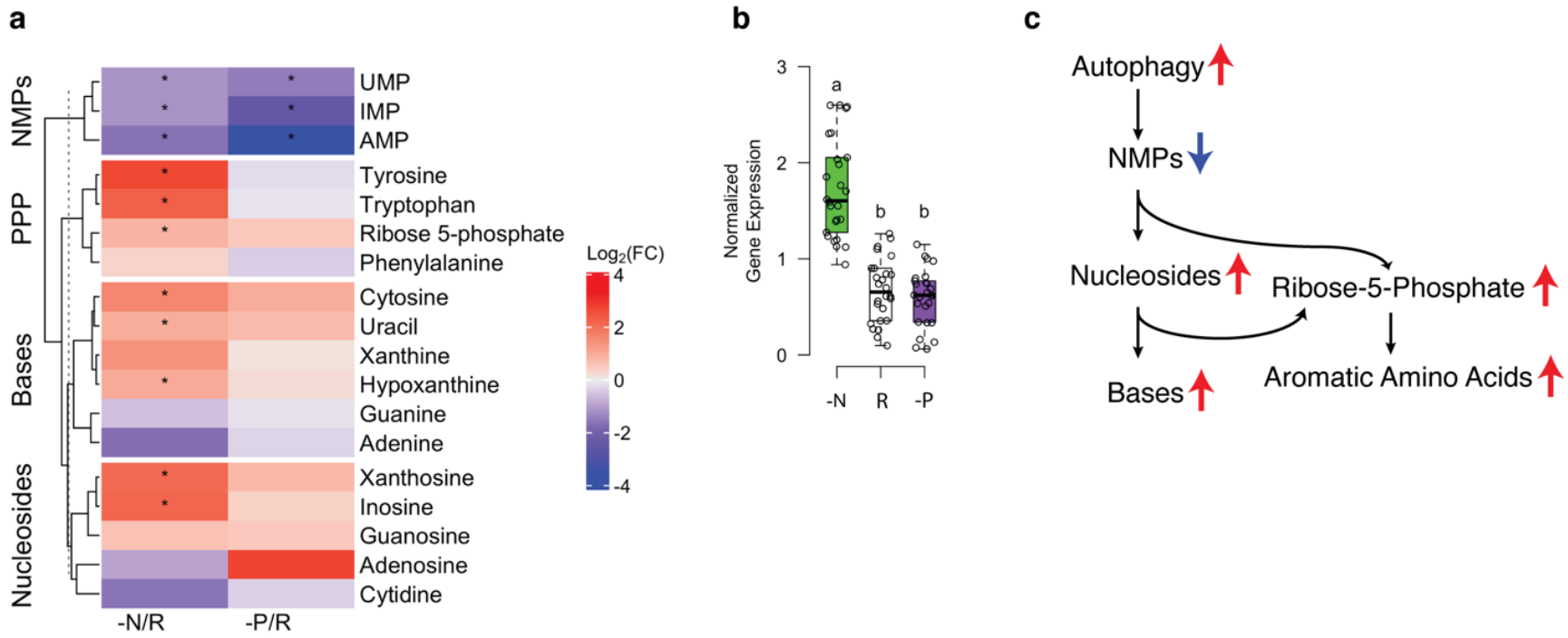

a
NMPs
PPP
Bases
Nucleosides
UMP
IMP
AMP
Tyrosine
Tryptophan
Ribose 5-phosphate
Phenylalanine
Cytosine
Uracil
Xanthine
Hypoxanthine
Guanine
Adenine
Xanthosine
Inosine
Guanosine
Adenosine
Cytidine
-N/R
-P/R
Log2(FC)
4
2
0
-2
-4
b
Normalized Gene Expression
3
2
1
0
a
b
b
-N
R
-P
c
Autophagy
NMPs
Nucleosides
Ribose-5-Phosphate
Bases
Aromatic Amino Acids

Figure 3: Stress induced nucleic acid scavenging. A) Targeted metabolite data of nucleoside monophosphates (NMPs), nucleic acid bases (Bases), nucleosides, and pentose phosphate pathway and aromatic amino acids (PPP). * denotes significant enrichment or depletion of compound concentration ($p < 0.05$, Tukey-HSD Test, $n = 3$). B) Normalized gene expression for RNA and DNA cleavage enzymes. Boxes with distinct letters above them are significantly different as defined by pairwise Wilcoxon Test and Bonferroni correction $p < 0.05$. C) Proposed pathway dynamics within N-stressed cells. Colored arrows adjacent to names indicate the net enrichment/depletion status of compounds or processes described by our data. Abbreviations: -N - N-stressed, R - nutrient replete, -P - P-stressed, UMP – uridine monophosphate, IMP – inosine monophosphate, AMP – adenosine monophosphate.

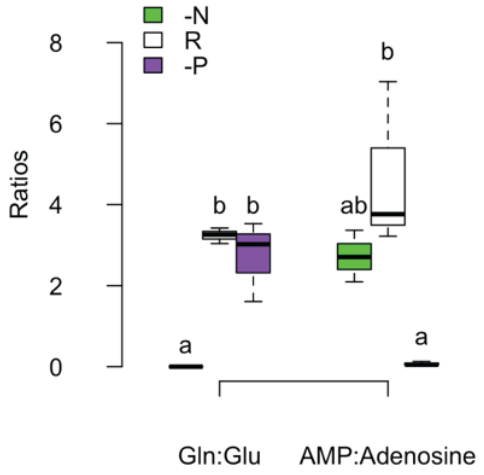
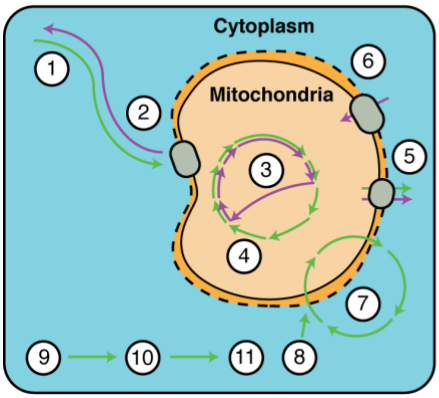


Figure 4: Diagnostic N-stress and P-stress ratios in H. akashiwo. Ratios were calculated using measured concentrations of each metabolite. Boxes with distinct letters above them are significantly different as defined by pairwise Tukey-HSD test $p < 0.05$. Abbreviations: Gln – glutamine, Glu – glutamic acid, AMP – adenosine monophosphate.



		H. akashiwo		Diatoms		*E. huxleyi*	
ID	**Biochemical Process**	**-N**	**-P**	**-N**	**-P**	**-N**	**-P**
1	Glycolysis	+	-	+	-	+/-	+/-
2	Gluconeogenesis	-	+	-	+	+/-	+/-
3	Glyoxylate Shunt	-	+	+	?	?	?
4	Oxidative TCA	+	-	+	+	-	-
5	Citrate Export	+	+	-	+	+	+
6	Carnitine Shuttle	-	+	?	+	+	?
7	Urea Cycle	+	+	+	?	+	+
8	Proteasome	+	+/-	?	?	?	?
9	Nucleic Acid Degradaton	+	+/-	?	?	?	?
10	PPP	+	+/-	+	+	?	?
11	AAA Biosynthesis	+	+/-	?	?	+	?

- < +/- < +

Figure 5: Conceptual model summarizing the changes in metabolism of *H. akashiwo* under nutrient stress. Each white numbered circle in the model displays a distinct intracellular process responding to stress. The purple and green arrows denote predicted systems-level biochemical activity for P-stressed and N-stressed *H. akashiwo*, respectively. + and – signs indicate increased and decreased activity, respectively, in each phenotype relative to the replete phenotype. Values for diatoms and *E. huxleyi* are based on reported pathway enrichment trends from prior studies *[37, 57-61]*. Boxes with question marks denote biochemical processes without existing data. Diatoms includes data from *T. pseudonana*, *Phaeodactylum tricornutum*, and *Skeletonema costatum*. Abbreviations: PPP – pentose phosphate pathway, AAA – aromatic amino acid.

3.8 Supplementary Material

3.8.1 Supplementary Notes

Note S1 - Overview of Metabolomics Data.

Untargeted Metabolomics: Following data processing and quality control filtering, we obtained 5090 and 1455 unique features under negative and positive ion modes, respectively. From this set, we found 844 and 408 features whose *m/z* values were within 1 ppm error of KEGG compounds under negative and positive ion modes, respectively. We found that 190 and 91 features with KEGG annotations had significantly different ($p < 0.05$, ANOVA, n = 3) intensities between growth treatments under negative and positive ion modes, respectively. Only 7 and 12 significant KEGG annotated features contained MS/MS spectra and metFrag fragmenter scores > 100.

Targeted Metabolomics: After quality control measures, we retained 57 of 69 and 69 of 85 targeted compounds for quantification in negative and positive ion modes, respectively. We found only two compounds with standard curves with R^2 values below 0.9. We evaluated changes in concentrations of targeted compounds based on their known roles in metabolism.

Note S2 - ATP budget per carbon molecule respired via beta oxidation and glycolysis.

Cells contain main pathways that feed carbon into the TCA cycle for respiration. Two of the most common ones are glycolysis and beta oxidation. Glycolysis is responsible for the conversion of sugars into acetyl-CoA, while beta oxidation is responsible for the conversion of fatty acid chains into acetyl-CoA. In addition to utilizing distinct starting materials, these pathways vary in the biosynthetic potential of intermediates and the total possible per-carbon energy. To understand the difference in per-carbon energy, consider the 12 carbon disaccharide trehalose and the 12 carbon saturated fatty acid lauric acid. The following reactions describe the catabolism of trehalose and lauric acid to acetyl-CoA via glycolysis and beta oxidation-respectively. In addition, they highlight the net total units of energy intermediates NADH and UQH_2 formed during respiration via each pathway. Abbreviations: NADH – nicotinamide adenine dinucleotide, UQH_2 - ubiquinol.

Trehalose breakdown:

Trehalose + P_i -> glucose + glucose-6-phosphate

Glucose glycolysis:

glucose + 2[NAD^+] + 2[ADP] + 2[P_i] -> 2[pyruvate] + 2[$NADH^+$] + 2H^+ + 2[ATP] + 2H_2O

Pyruvate conversion to acetyl-CoA:

pyruvate + [NAD^+] + CoA -> Acetyl-CoA + [$NADH^+$] + CO_2 + H^+

Net TCA cycle reaction:

Acetyl-CoA + 3[NAD^+] + UQ + GDP + Pi + 2[H_2O] -> CoA-SH + 3[NADH] + UQH_2 + 3H^+ + GTP + $2CO_2$

Beta oxidation:

Lauric acid-CoA + FAD + NAD + H^+ + H_2O + CoA -> 6[Acetyl-CoA] + 6[UQH_2] + 6[NADH] + 6[H^+] + 6[CoA]

The table below describes the net ATP per reducing/oxidizing energy intermediates.

Molecule	ATP/Molecule
NADH	2.5
UQH_2	1.5

Based on this, we can see that the catabolism of trehalose results in the formation of 20[$NADH^+$] and 4[UQH_2] resulting in a total of 56[ATP] per molecule. By contrast, the catabolism of lauric acid results in the formation of 24[$NADH^+$] and 4[UQH_2] resulting in a total of 69[ATP]. Hence, trehalose catabolism provides 81 percent ATP per carbon more than that of lauric acid.

Note S3 - Metabolomic Evidence of Increased Urea Cycling. *H. akashiwo* increases the relative abundance of transcripts of the urea cycle when stressed by N [25]. We sought to confirm whether previously observed transcriptional changes resulted in a concomitant increase in pathway activity. Hence, we quantified 7 distinct intermediates. We observed that concentrations of arginine, glutamine, and N-acetyl glutamic acid were significantly decreased ($p < 0.05$, Tukey-HSD Test, n = 3), while the concentrations of ornithine and aspartic acid were significantly increased ($p < 0.05$, Tukey-HSD Test, n = 3) exclusively in N-stressed cells (Figure S4A). To understand if these concentration dynamics suggested increased urea cycle flux, we calculated the Global Arginine Bioavailability Ratio (GABR = arginine/[ornithine + citrulline]) [154]. Low GABR values indicate increased flux of arginine through the urea cycle relative to other paths [154]. We observed that GABR values were significantly lower ($p < 0.05$, Tukey-HSD Test, n = 3) in N-stress cells, corroborating a previous transcriptomics study [25] (Figure S4B). Increased urea cycle activity under N-stress was reported in both diatoms and haptophytes [35, 61, 139]. Interestingly, GABR values were also significantly lower in P-stressed cells ($p < 0.05$, Tukey-HSD Test, n = 3) without the transcriptomic enrichment observed in N-stressed cells (Figure S4B). The urea cycle produces TCA intermediate fumarate as a by-product. Fumarate enters the TCA cycle after oxidation, hence it may serve as a carbon source for gluconeogenesis [125]. Prior studies show that P-stress mediated urea cycling varies among phytoplankton. While dinoflagellates also increase urea cycle activity under P-stress [155], *E. huxleyi* was reported to throttle its activity under P-stress [60]. This disagreement speaks to the niche differences experienced by distinct phytoplankton groups. We hypothesize that the increased transcriptional upregulation in N- versus P-stressed *H. akashiwo* may suggest that N-stressed cells have a greater urea cycle flux. Future isotope tracer experiments measuring urea cycle reaction rates would confirm this hypothesis.

Note S4 - Proteasome Mediated Enzyme Degradation (PMED) Evidence from Untargeted Data. We sought to determine if PMED may provide a viable source of amino acids for the urea cycle. For this, we first gathered all untargeted features with *m/z* matches to known tetrapeptides in all treatments. We elected to use tetrapeptides as a biomarker as they are well known by-products of PMED activity [125]. We filtered the total set of 758 putative tetrapeptides into 199 features with *p*-values < 0.05 following hypothesis testing. The intensity of these features was significantly depleted ($p < 10^{-10}$, Wilcoxon-test, n = 597) in stressed cells relative to replete ones (Figure S6), suggesting that PMED may differ significantly between replete and stressed treatments. To increase the strength of annotation of these features, we modeled the retention times of all 199 putative matches against the predicted hydrophobicity values of the tetrapeptides based on amino acid composition. We found that feature retention time was significantly ($p < 10^{-8}$, Linear Model, n = 199) related to predicted hydrophobicity values from two separate scales [119, 120] (Figure S7). Additionally, five of these features contained MS/MS spectra and matched tetrapeptides through *in silico* fragmentation, further supporting our putative identifications. These results suggest there is better than random chance that our features may originate from tetrapeptides.

3.8.2 Supplementary Figures

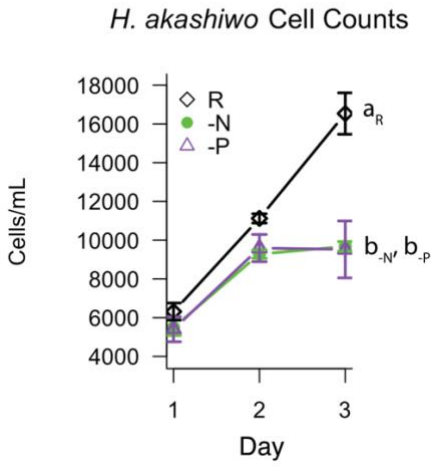


Figure S1: H. akashiwo growth curves. Error bars represent the standard deviation of triplicate cultures. We monitored growth of N-stressed (-N, green circles), P-stressed (-P, open purple triangles), and replete (R, open diamonds) cultures over three days. If two time points share a letter, then Tukey-HSD test between two factors was not significant. Significance is defined as $p < 0.01$. We harvested cells for metabolomics analysis on day 3, similar to previously-published transcriptome work [25].

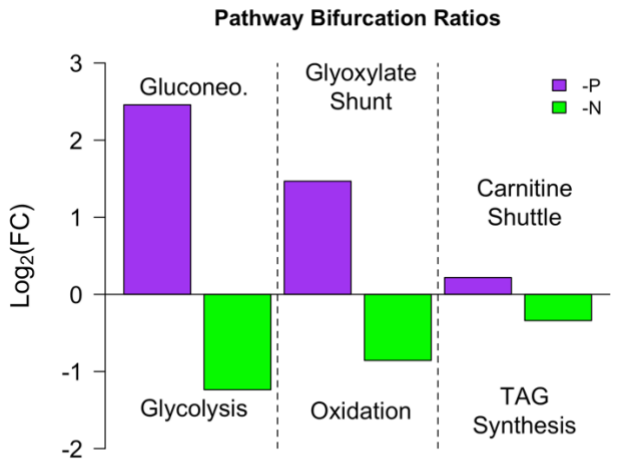


Figure S2: Pathway bifurcation ratios. These ratios represent either the expression of genes known to be inversely regulated (gluconeogenesis vs glycolysis and TAG synthesis vs carnitine shuttle) to avoid futile cycles, or genes driving the immediate bifurcation between the glyoxylate shunt and TCA cycle carbon oxidation. These ratios were calculated by first dividing each enzyme's expression by the enzyme's replete expression. Color denotes data from P-stress (-P, purple) and N-stress (-N, green) data. For inversely related genes, we constructed the ratios using the mean expression of normalized genes of each pathway. For the glyoxylate shunt oxidation comparison, we only considered the expression of two enzymes, isocitrate lyase (glyoxylate shunt) and isocitrate dehydrogenase (oxidation). Abbreviations: FC – fold change, Gluconeo – gluconeogenesis, TAG – triacylglyceride.

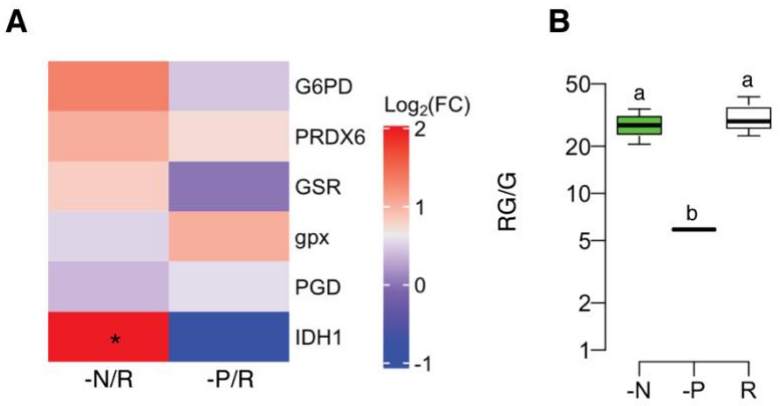


Figure S3: Oxidative stress biomarkers within stressed cells. A) Log_2 fold change of RNA involved in the interconversion of glutathione and reduced glutathione in nutrient stressed cells relative to replete cells. * - denotes significance defined as a $log_2(FC) > 2$ and ASC > 0.95. B) Ratio of reduced glutathione (RG) to glutathione (G) based on the measured concentrations of these compounds. Boxes with distinct letters above them are significantly different as defined by Tukey-HSD test $p < 0.05$. Abbreviations: G6PD – glucose-6-phosphate dehydrogenase, PRDX6 – peroxiredoxin 6, GSR – glutathione reductase, gpx -glutathione peroxidase 2, PGD – phosphogluconate dehydrogenase, IDH1 – isocitrate dehydrogenase.

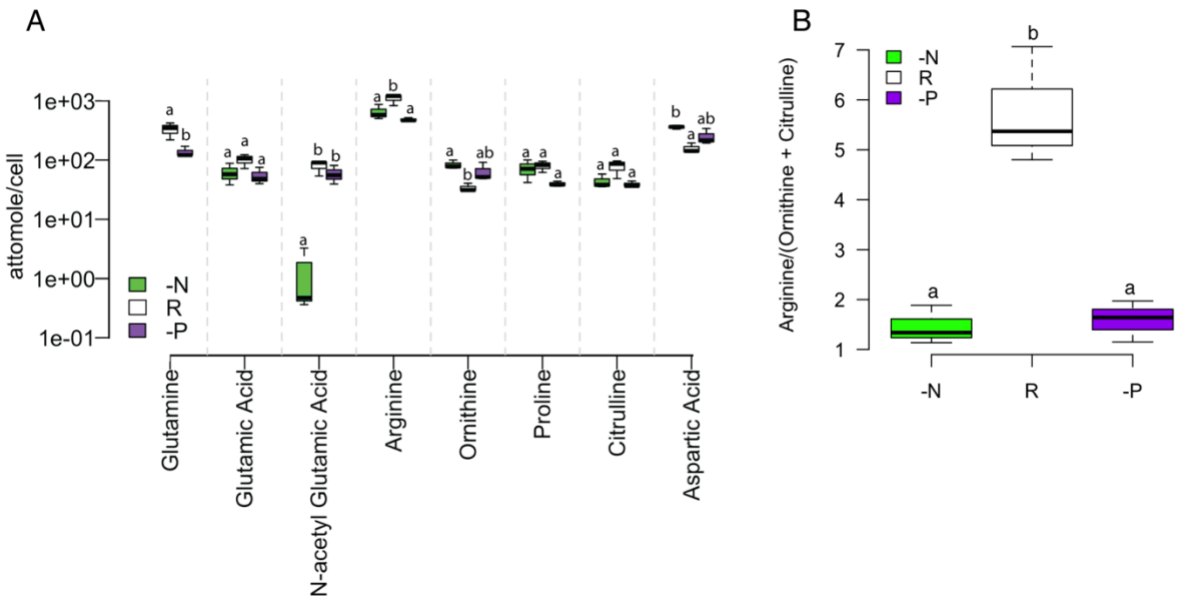


Figure S4: Urea cycle dynamics across experimental factors. A) Concentrations of urea cycle intermediates across experimental treatments. Boxes with distinct letters above them are significantly different as defined by pairwise Tukey-HSD test $p < 0.05$ B) The Global Arginine Bioavailability Ratio (GABR), defined as the ratio of arginine/(ornithine + citrulline) across experimental factors. This ratio provides a measure of urea cycle flux relative to other arginine catabolism pathways; lower values indicate a greater flux through the urea cycle. Boxes with distinct letters above them are significantly different as defined by Tukey-HSD test $p < 0.05$.

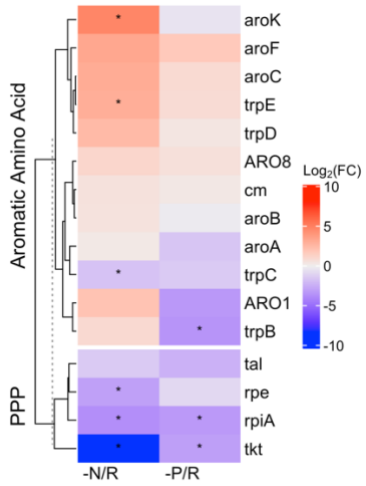


Figure S5: Transcripts related to non-oxidative pentose phosphate pathway (PPP) and aromatic amino acid biosynthesis. * - denotes significance defined as a $\log_2$(FC) > 2 and ASC > 0.95. Abbreviations: aroK – shikimate synthase, aroF – 3-deoxy-7-phosphoheptulonate synthase, aroC – chorismate synthase, trpE – anthranilate synthase, trpD – anthranilate phosphoribosyltransferase, ARO8 – amino acid aminotransferase 1, cm – chorismate mutase, aroB – 3-dihydroquinate synthase, aroA – chorismate mutase, trpC – anthranilate synthase, ARO1 – pentafunctional arom polypeptide, trpB – tryptophan synthase, tal - transaldolase, rpe – ribulose-phosphate 3-epimerase, rpiA – ribose 5-phosphate isomerase, tkt - transketolase.

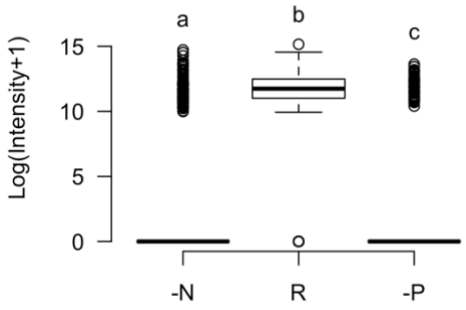


Figure S6: Distribution of intensity values of putative tetrapeptides from untargeted data across nutrient stress treatments. Boxes with distinct letters above them are significantly different as defined by pairwise Wilcoxon test $p < 0.05$.

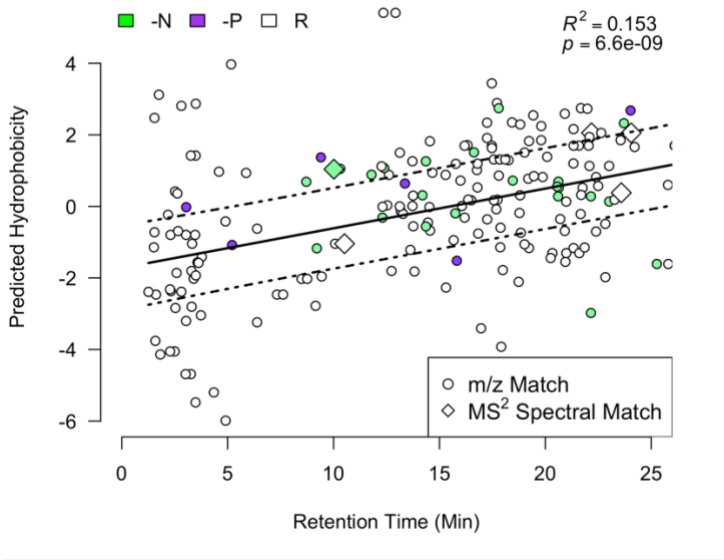


Figure S7: Linear regression between feature retention time and predicted hydrophobicity. The retention times with m/z matches to tetrapeptides show a significant ($p < 10^{-9}$, linear model, n = 119) related predicted hydrophobicity value from the Hessa et al. scale *[119]*. Diamonds represent features with MS/MS predicted to belong to tetrapeptides through in silico fragmentation matching. Colors denote the experimental factor in which the intensity of each feature was maximal. We observed similar trends using the scale from Wimley and White *[120]*.

Chapter 4

Pathway Responses to Nutrient Stress Distinguishes Phytoplankton Groups

4.1 Introduction

Marine phytoplankton are responsible for over 50% of global primary production [2]. Primary production is driven by the external inputs of dissolved inorganic nutrients like phosphorus and nitrogen [9, 10]. Modeling studies predict that the availability of these nutrients will decrease in the future due to the impacts of climate change on marine ecosystems [62], causing decreases in overall global primary production. These studies often regard phytoplankton as a monolithic unit, and do not take into account the distinct evolutionary capacities of these organisms. Hence, the extent to which primary production will be impacted due to climatic changes remains unclear.

Parameterizing models with information describing how phytoplankton communities respond to complex environmental perturbations will reduce the uncertainties surrounding future climate predictions [156]. However, the biological complexity of phytoplankton communities poses as a major obstacle towards the acquisition of this knowledge. Phytoplankton communities are made up of a myriad of phylogenetically distinct organisms [23], each of which has acquired unique adaptations to their environments and applies distinct acclimations towards external disturbances [63]. Adaptations are distinguishing physiological properties of phytoplankton acquired through the course of evolution. Commonly cited examples include differences in cell size [17], changes in membrane composition under various temperature regimes [157], variations in stoichiometric nutrient quotas and macromolecular composition [14], and employment of distinct life history strategies [6]. Acclimations describe behaviors these organisms employ over short time scales to respond to changes in the environment. Examples of acclimation strategies include increased production of inorganic nutrient transporters [24, 25], upregulation of enzymatic recycling [42], and the release of info-chemical signals between community members [158]. Understanding how adaptive and acclimatory strategies distinguish phytoplankton would improve our ability to forecast future phytoplankton community dynamics under changing environmental conditions.

The advent of field-based genomic and transcriptomic techniques has greatly improved our capacity to define and classify inter-phytoplankton adaptations and acclimations. For example, prior studies using these techniques revealed that phytoplankton respond distinctly to

environmental conditions [18, 19, 31], employ various nutritional acquisition strategies [30], and experience varying degrees of stress in different marine environments [159]. Gene-based methods, however, remain limited in their capacity to serve as indicators of ongoing metabolic activity [33]. Orthogonal approaches, such as metabolomics, reveal finger prints of biochemical processes [34], and may describe metabolic activity. The use of metabolomics to understand phytoplankton is burgeoning. Recent studies show they viably describe system-level changes in phytoplankton metabolism as a result of stress [42] and diel dynamics of *in situ* phytoplankton communities [41]. Additionally, due to the paucity of available genomes for most phytoplankton [23], these approaches can help reconstruct pathways that lack complete genomic resolution [42].

Understanding phytoplankton metabolism can reveal mechanisms responsible for driving the unique acclimations of phytoplankton [42]. For example, the raphidophyte *Heterosigma akashiwo* acclimates to nitrogen stress by producing greater quantities of triacylglyceride lipids [137]. These lipids increase the specific gravity of cells, which facilitates diel vertical migratory patterns [160]. This metabolic change fosters a unique physiological response to nitrogen stress that allows *H. akashiwo* to outcompete neighboring species for dissolved inorganic nutrients below the thermocline and sustain their bloom within estuarine environments.

Similarly, metabolism can also reveal unique adaptations that support phytoplankton fitness. Past studies show that unlike plants, diatoms contain a fully functional urea cycle [35]. This pathway was partially acquired from bacterial via horizontal gene transfer, and provides a means of reallocating N and generating additional oxidizable carbon for these organisms. It is hypothesized that this capability primes diatoms to sequester episodic influxes of N, thereby supporting their ecological dominance under environmentally variable conditions [38].

Recent advances in resource allocation modeling [104] make it possible to predict phytoplankton community dynamics in the context of both acclimatory and adaptive behaviors by linking changes in metabolism and physiology. However, they require detailed knowledge of how metabolism drives physiology, which remains limited across phytoplankton groups. A

broader understanding of phytoplankton metabolism and how it responds to external perturbations would provide a foundation for these models.

In this study, we expand our existing knowledge of metabolic adaptations and acclimations of different phytoplankton. To accomplish this aim, we cultured four organisms from distinct abundant and globally important phytoplankton groups under replete, phosphorus-stress, and nitrogen-stress growth conditions. We chose to explore the impact of nitrogen- and phosphorus-stress on metabolism due to their roles as limiting nutrients within marine ecosystems [10]. Additionally, prior studies show phytoplankton groups in low nutrient field populations modulate their transcriptome distinctly [19], suggesting each group employs different metabolic acclimations to these nutrients. In order to evaluate changes in metabolism, we performed untargeted and targeted metabolomic analyses on cells from these groups. We evaluated distinguishing adaptations by comparing the metabolomes across organisms, and acclimations by comparing organism-specific metabolites across stress conditions. To aid in our analysis of the untargeted data in light of a lack of genomes for the taxa, we developed and evaluated a novel Network-based Permutation Test (NEPTune) to identify overrepresented biochemical pathways within each phytoplankton taxon.

4.2 Methods

4.2.1 Culture Maintenance:

We cultured four species of phytoplankton from three globally important phyla (Bacillariophyta, Haptophyta, and Ochrophyta). These species include the cosmopolitan diatom *Chaetoceros affinis* ('diatom') CCMP159 (isolated from Great South Bay, NY, USA, 1958), the haptophytes *Chrysochromulina polylepis* CCMP1757 ('prymnesiophyte') (isolated from the North Sea 1988) and *Gephyrocapsa oceanica* RCC1303 ('coccolithophore') (isolated from Arachon Bay, France, 1999), and the raphidophyte *Heterosigma akashiwo* strain CCMP 2393 ('raphidophyte') (isolated from Rehoboth Bay, Delaware, USA). We cultured all species with the exception of the coccolithophore in modified L1 medium made with autoclaved 0.2-µm filtered seawater from Vineyard Sound, MA under the following culture conditions: replete (576 µM $NaNO_3$, 36.2 µM NaH_2PO_4; N:P = 16), N-stress (5 µM $NaNO_3$ and 36.2 µM NaH_2PO_4; N:P = 0.14), and P-stress (576

µM $NaNO_3$ and 0.2 µM NaH_2PO_4; N:P = 2880). We cultured the coccolithophore under the following conditions: replete (100 µM $NaNO_3$ and 6 µM NaH_2PO_4; N:P = 16.7), N-stress (1 µM $NaNO_3$ and 6 µM NaH_2PO_4; N:P = 0.12), and P-stress (100 µM $NaNO_3$ and no added NaH_2PO_4; N:P = 1000+). The lower nutrient concentrations for the coccolithophore were necessary to ensure consistent calcification. We grew each culture with light intensity of 100 µmol quanta m^{-2} s^{-1} of photosynthetically active radiation (400-700 nm) during a 14h:10h light:dark cycle. We grew the diatom, the coccolithophore, and the raphidophyte at 18° C, and the prymnesiophyte at 15° C to match species-specific preferences. We continuously rotated some cultures on an orbital shaker (the raphidophyte, 75 rpm; the diatom, 100 rpm; the prymnesiophyte, 100 rpm) to maintain optimal growth. Cultures were not axenic, but were uni-algal and uni-eukaryotic.

4.2.2 Experimental Design:

We used entrainment cultures to initiate experimental cultures to decrease carryover of nutrients from stock cultures and promote acclimation to the experimental conditions as described in Harke et al. [24]. No culture was axenic, however each was uni-algal. We grew single entrainment cultures for all organisms with modified L1 medium (base seawater as above) for three days. We inoculated triplicate 2-L bottles (1 L experimental volume) with 25 mL of entrainment culture into the same modified L1 medium at the start of the experiment. We monitored growth in each flask by *in vivo* chlorophyll fluorescence on a Turner Designs Aquafluor handheld fluorometer with paired cell counts (Supplementary Figure 1 and 2). We preserved cell count samples in 2% acid Lugol's solution (final concentration) and we determined cell concentrations by microscopy. We took these measurements at the same time each day (during the middle of the light phase) to avoid diel changes. We harvested treatments for metabolomics analysis when we observed significant differences in cell counts between stressed and replete cultures, in agreement with the definition of nutrient stress rather than deficiency (Supplementary Figure 1 and 2) [111]. We used 47-mm glass fiber filters (nominal pore size = 0.45µm; GF/F, Whatman) and combusted glass filtration funnels under low vacuum pressure (never exceeding 5 mm Hg) to filter cells (300 mL) from each sample. We flash-froze filters in cryovials in liquid nitrogen and stored them at −80°C until extraction.

4.2.3 Filter Extractions:

We split each filter in half for separate extractions for targeted and untargeted metabolomic analyses. The extraction procedure for untargeted and targeted samples are identical with the exception of the final solid phase extraction (see below). For both types of analyses, we cut each filter half into six roughly equivalent pieces and placed them into an 8-mL amber glass vial. We extracted metabolites from filters using 1mL cold 40:40:20 acetonitrile: methanol:water + 0.1 M formic acid similar to previous work [112, 113]. We then added 25 µL of 1 µg/mL deuterated standard mix (d_3-glutamic acid, d_4-4-hydroxybenzoic acid, and d_5-taurocholate) as extraction recovery standards. We sonicated the solvent-filter mixture for 10 minutes to lyse the cells, and transferred the solvent into a microcentrifuge tube. We rinsed the filters with three 200-µL aliquots of extraction solvent to capture any remaining organic matter. We centrifuged the combined extracts at 20,000 × *g* for 5 minutes, and transferred the supernatant into clean 8-mL amber glass vials, taking care to leave behind any scraps of filter or cellular detritus. We neutralized the extracts with 25.6 µL of 6 M ammonium hydroxide and dried them down to near dryness in a vacufuge. We reconstituted dried samples for targeted analysis in 200 µL 95:5 water:acetonitrile solution plus 2.5 µL of 5 µg/mL deuterated biotin injection standard.

For the untargeted analysis, a PPL extraction step is necessary to remove excess salt and prevent ion suppression [113]. We reconstituted these samples with 500 µL 0.01 M HCl to lower the pH to 2-3 and ran these samples through 100 mg/1 mL Agilent Bond Elut PPL cartridges. We pre-conditioned the cartridges with one cartridge-volume of 100% methanol and passed acidified untargeted samples through the cartridge at a flow rate below 40 mL min^{-1}. We rinsed the cartridges with one cartridge-volume of 0.01 M HCl, dried them down for 5 minutes, and eluted the metabolites with 1 cartridge-volume of methanol. We dried untargeted samples again to near dryness and reconstituted them with 247.5 µL of 95:5 water:acetonitrile plus 2.5 µL of 5 µg/mL deuterated biotin injection standard. We combined 45 µL aliquots from each sample to create a pooled sample.

4.2.4 Liquid Chromatography and Mass Spectrometry:

We analyzed metabolite samples for untargeted analyses by high-performance liquid chromatography (HPLC, Micro AS autosampler and Surveyor MS Pump Plus, Thermo Scientific) coupled via electrospray ionization (ESI) to a hybrid linear ion trap- Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer (7T LTQ FT Ultra, Thermo Scientific). We separated metabolites on a Synergi Fusion reverse phase C_{18} column (4 µm, 2.0 x 150 mm, Phenomenex), equipped with a guard column and precolumn filter, and maintained at 35°C. We eluted the column with (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile at a flow rate of 0.25 mL min^{-1}. We held the column at 5% B for 2 min, ramped to 65% B over 18 min, quickly ramped to 100% B over 5 min, held at 100% B for 7 min and then equilibrated at 5% B for 8 min prior to the next injection (total run time = 40 min). We injected 20 µL of sample individually for positive and negative ion mode analyses. We externally calibrated the mass spectrometer just prior to analysis in positive and negative ion modes using the manufacturer's solutions. We optimized the capillary temperature and ESI voltage at 330°C and 4.2 kV, respectively, in positive mode and at 365°C and 3.8 kV, respectively, in negative mode. We maintained sheath gas, auxiliary gas, and sweep gas flow rates at 35, 5, and 2, respectively (arbitrary units) for both polarities. We collected MS and data dependent MS/MS scans as follows: (1) a full MS scan in the FT-ICR analyzer from 100-1000 m/z, with mass resolving power set to 100,000 (defined at m/z 400); and (2) collision-induced dissociation fragmentation scans (MS/MS) in the linear ion trap for the four most abundant ions in each full scan. We collected MS/MS spectra under dynamic exclusion with an exclusion time of 20 seconds. At the start of each batch, we injected the pooled sample multiple times to condition the column with the sample matrix and to stabilize peak retention times. We also analyzed the pooled sample every nine samples for quality assurance.

We analyzed targeted samples by ultrahigh-performance liquid chromatography (UHPLC, Accela Open Autosampler and Accela 1250 Pump, Thermo Scientific) coupled via heated electrospray ionization (H-ESI) to a triple quadrupole mass spectrometer (TSQ Vantage, Thermo Scientific) operated under selected reaction monitoring (SRM) mode. We set the spray voltage at 4000 V (positive mode) and 3200 V (negative mode). We set source gases at 55

(sheath) and 20 (auxiliary gas), heated capillary temperature at 375 °C, and the vaporizer temperature at 400 °C. We performed chromatographic separation on a Waters Acquity HSS T3 column (2.1 × 100 mm, 1.8 μm) equipped with a Vanguard pre-column and maintained at 40 °C. We eluted the column with (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile at a flow rate of 0.5 mL min^{-1}. The gradient started at 1% B for 1 min, ramped to 15% B from 1-3 min, ramped to 50% B from 3-6 min, ramped to 95% B from 6-9 min, held until 10 min, and ramped to 1% B from 10-10.2 min, with final re-equilibration at 1% B (total gradient time = 12 min). We made separate autosampler injections of 5 μL for positive and negative ion modes.

4.2.5 Standard Optimization:

We obtained authentic standards at the highest grade available from Sigma Aldrich and Cayman Chemical for compounds outside of our existing targeted method [113]. Due to prohibitive costs required to confirm all features with authentic standards, we focused on compounds that are both readily commercially available and pertinent to pathways showing significant enrichment within permutation across multiple organisms (**Error! Reference source not found.**). We injected standards at concentrations of 1 μg/mL in Milli-Q water to optimize selected reaction monitoring (SRM) conditions (s-lens, collision energy, product ions). We selected at least two SRM transitions (precursor-product ion pairs) for quantification and confirmation of each target compound. We determined the chromatographic retention time of each compound with standards dissolved in Milli-Q.

4.2.6 Data Processing:

We converted untargeted data files from proprietary Thermo RAW into mzML format using msConvert [91]. We processed these files using XCMS and AutoTuner [71, 92, 116] to generate a matrix of features (see Supplementary Table 2 for processing parameters). We define features as chromatographic peaks with unique mass-to-charge (*m/z*) and retention time values, with relative abundances determined by their area. We subjected processed data to quality control filtering by removing possible contaminants and non-reproducible features as described previously [117]. Briefly, we removed features within blanks, features with a coefficient of

variation higher than 0.2 within pooled samples, and features with low reproducibility across factor groups. We report feature intensities normalized by cell number.

We used MAVEN to integrate compound peak areas within targeted data [118]. We used an in-house MATLAB script to apply quality control-filtering and to quantify peak areas using a standard curve of 4 to 10 points within Milli-Q. We retained metabolites for this analysis if (1) the peak included a confirm ion, and (2) the metabolite was present within two of three biological replicates for each treatment. We further culled the list by correcting for metabolite presence in procedural blanks. We measured the matrix effects of targeted compounds by calculating the relative error between the slope of matrix-matched standard curve and that of a Milli-Q standard curve [161]. Mathematically, this measure is described by the following equation:

$$Matrix\ Effect = \frac{Slope\ of\ matrix\ matched\ curve - Slope\ of\ milli{-}Q\ curve}{Slope\ of\ milli{-}Q\ curve} \times 100 \quad (1)$$

We considered matrix effects between -130 and 130 to be acceptable, in agreement with community standards [161].

4.2.7 Data Analysis:

We used a similar approach with growth curves and targeted metabolomics data to identify significant differences among case-control comparisons. First, we applied ANalysis Of VAriance (ANOVA) hypothesis testing to identify if measures across experiment factors were significantly different from one another. We identified significant pairwise-comparisons using Tukey's honestly significant difference test (Tukey-HSD Test). We then applied the Benjamini-Hochberg correction to control for type-I error. We considered any *p-value* following these tests equal to or less than 0.05 to be significant.

We applied several distinct techniques to analyze the untargeted data. First, we used mummichog to match feature *m/z* values to Kyoto Encyclopedia of Genes and Genomes (KEGG) compounds [121, 122]. We consider any feature with an *m/z* match to be a level 3 annotation

as described by Sumner et al. [162]. Next, to reduce the dimensionality of this data, we applied non-negative matrix factorization (NMF) [163]. For this, we used the ranks of 4 and 3 for NMFs performed on the entire dataset and organism specific data subsets, respectively, after optimizing with cophenetic correlation metrics [164]. We used a basis contribution score threshold of 0.99 to assign features to particular basis vectors. We performed Global Natural Products Molecular Networking (GNPS) to facilitate mass shift analysis [165]. For this, we required that linked nodes have a cosine score of 0.7 and a minimum of 6 matched peaks. We considered all mass differences prescribed for chemical reactions with MetaNetter 2 [166, 167]. We used a chi-squared test to determine if mass shifts were distributed significantly differently across organisms. To increase the strength of annotation of untargeted features, we used METLIN [168] and the METLIN-guided In-Source Annotation (MISA) algorithm [83] to putatively identify compounds. We upgraded the annotation level of features from 3 to 2 if they matched METLIN reference spectra via MISA.

We applied several techniques exclusively to the analysis of targeted data. We used hierarchical clustering to group compounds with similar measured concentrations across organisms. For this, we first averaged organism-specific target concentrations and then standardized these values across organisms using z-score normalization. We applied NMF with a rank of 4 to identify organism-specific compounds [163]. We used a basis contribution score threshold of 0.95 to assign compounds as organism-specific. This measure calculates the relative abundance of a given compound across the condensed columns. We utilized two distinct measures to check for enrichments or depletions of groups of targeted compounds. The first measure was designed to measure the effect size, i.e., the overall magnitude difference of case-control comparisons of targeted compounds. For this, we used a linear discriminant analysis (LDA). Similar to the microbiome feature selection algorithm LEfSe [169], we calculated the effect size of the comparison by using the difference between group centroids determined along the first axis LDA axis. Next, we applied a Wilcoxon-Rank Sum Test to evaluate the statistical significance of a given case-control comparison.

Finally, we applied a Procrustes analysis to compare the structure of targeted and untargeted data. For this, we compared the first two principal components of these datasets and evaluated the significance of this comparison using a permutation test.

4.2.8 Network-based permutation test (NEPTune):

Within metabolism, metabolites are connected to one another via enzymatically catalyzed reactions. The total sum of these combined metabolites and reactions may be described as a network, such as those described by KEGG pathways. We took advantage of the mathematical properties of these networks to construct a NEtwork-based Permutation Test (NEPTune) to identify overrepresented pathways within a dataset. This approach is advantageous over alternative pathway enrichment analyses, as it does not require *a priori* knowledge of the genes present in an organism. Instead, it relies on the proximity between nodes in a network to assign significance. Through this section, we describe the mathematical details of this approach (Figure 1).

We treat each KEGG pathway as an individual undirected network. Within each network, we define the nodes as the compounds and the reactions as links. Given a network, *G*, we first identify all the *reaction paths* within *G*. We define a *reaction path* as the path of length *L* spanning nodes $N = \{N_1, \dots, N_L\}$. Here, N_1 is the node of highest degree in G and N_L is the furthest node away from N_1 mapped by a feature *m/z* value. The score of each reaction path for G is defined by

$$S_{rxn} = \Sigma_{i=1}^{L} z_i \tag{2}$$

Here, $z_i = 1$ if node is mapped by a feature *m/z* value and $z_i = 0$ otherwise. Intuitively, this score reflects the extent to which unique *m/z* values fall into a predefined set of consecutive reactions.

We next want to determine if the S_{rxn} score is higher than we would expect by chance. For this, we first divide G into a subnetwork of radius L, centered at N_1. Then we repeatedly make randomized draws of L many nodes from the subnetwork and calculate the empirical scores for all draws. We create a distribution from the scores of our random draws and use the

z-scores method to compute an empirical *p-value* for the observed *reaction path* score S_{rxn}. During the randomization, we drew 5000 permutations unless limited by the size of the network. We consider any *reaction path* score S_{rxn} to be significant when it has a *p-value* below 0.05 following type-II error correction through the Benjamini-Hochberg method. To confirm that NEPTune recapitulated previously-published overrepresented pathways, we applied the algorithm to an existing dataset of targeted metabolites [170].

It is possible for NEPTune to incorrectly classify a pathway as overrepresented if the features map onto a reaction path that contains many structural isomers. Hence, we sought to evaluate the validity of hypothesized pathway overrepresentation to determine the efficacy of our algorithm. First, we checked whether organism-specific characterized features making up reaction paths were present within matching targeted samples. Next, to confirm that these putatively characterized features show similar trends in abundance to measured targeted compounds, we regressed the normalized intensity of observations from matching samples across both datasets. To evaluate the overall error of this approach across all targeted and untargeted matches, we calculated the empirical cumulative distribution function of all residuals. Lastly, we checked the concordance in retention time between targeted compounds and matching untargeted features by regressing their elution ranks across both datasets.

4.2.9 Network-based permutation test filters:

One challenge with the interpretation of untargeted metabolomics data is that features matched by mass cannot distinguish between structural isomers, i.e., compounds with the same mass but different atomic arrangements. We account for the possible occurrence of structural isomers within our significant *reaction paths* in two ways. First, we consider the probability that nodes making up a *reaction path* do not map to any alternative *reaction path* from a different pathway. Second, we consider the possibility that a given compound match within our data could have structural isomers.

Our first measure, the pathway overlap probability, $P_{overlap}$, provides an estimate of the probability that a reaction path is specific to a given pathway. In our case, we compared KEGG pathways against one another, as these were the biochemical networks utilized by

NEPTune. Thus, it provided us with a means of evaluating how specific a detected reaction path is to a biochemical pathway. We calculate $P_{overlap}$ for each *reaction path* as:

$$P_{overlap} = 1 - \frac{\max(alternative\ pathways)}{L} \quad (3)$$

where *alternative pathways* represent the number of times an *m/z* matched node from a given reaction path matches compounds within other pathways by mass. Each alternative pathway is counted once per node.

Our next measure, path structural isomer coverage, Iso_{path},, describes how well our data covers structural isomers of compounds within significant *reaction paths* that may appear within metabolism. We calculate Iso_{path} as:

$$Iso_{path} = \frac{1}{l}\Sigma_{i=1}^{l} Iso_{score,i} \quad (4)$$

Where l is the number of *m/z* matched compounds within a *reaction path* and Iso_{score} represents the compound structural isomer coverage. Iso_{score} is defined as:

$$Iso_{score} = \frac{n+\frac{n}{m}}{n+\frac{m}{n}} \quad (5)$$

where n represents the number of unique features matching a given node by *m/z* and m represents the total number of structural isomers of the *m/z* value in KEGG. We used a 15 second retention time filter to ensure that adducts were not counted twice towards n. This measure was inspired by the Michaelis-Menten equation for enzyme kinetics, hence as $n \to m$, $Iso_{score} \to 1$. This metric assumes that KEGG is a complete representation of all possible biochemical pathways.

We filtered out any significant *reaction paths* below our thresholds $P_{overlap} > 0.9$ or $Iso_{path} > 0.9$. We chose these thresholds after considering distributions of these metrics across all significant *reaction paths* (Supplementary Figure 3).

4.3 Results and discussion

4.3.1 Taxonomy is the primary driver of variability within untargeted data

In order to understand how phytoplankton stress response varied between organisms, we grew four organisms from distinct phytoplankton groups under replete (R), phosphorus stressed (-P), and nitrogen stressed (-N) conditions. Although cultures varied from one another in total time required to respond to stress, each culture showed a significant drop in relative fluorescent units (Supplementary Figure 1) or cell counts (Supplementary Figure 2) relative to replete growth after 3 to 8 days. This result suggests that the cultures experienced nutrient stress similar to previous results from transcriptomic studies [24, 25, 111].

We initiated our investigation of the stress-response metabolisms of these phytoplankton by performing untargeted metabolomics on each culture. After obtaining these data, we applied several quality assurance filters to our data to distinguish biologically robust features from irreproducible or contaminant ones. In total, we retained 2564 and 4735 total features across all samples within the positive and negative ionization modes, respectively (Supplementary Table 3). These feature counts are on par with prior studies on a marine heterotrophic bacterium [88] and a diatom [171] generated using an identical analytical platform. From these totals, we matched 600 and 440 features within positive and negative ionization mode features to KEGG compounds by mass. We considered these features to have level 3 annotation ('putative characterization') based on Metabolomics Standards Initiative guidelines [162].

We next sought to determine which experimental factor introduced the greatest amount of variability to our putatively characterized features: taxonomy or nutrient stress status. To evaluate this, we ran a non-negative matrix factorization (NMF) on our positive (Figure 2A) and negative mode (Supplementary Figure 4A) datasets. This unsupervised clustering method condensed both datasets into a matrix of organism-specific vectors, suggesting that taxonomy was the primary driver across the dataset. We assigned features with a basis contribution score of .99 or greater to organism-specific NMF columns. This measure describes the relative abundance of a feature within a specific column vector. The number of unique features assigned to each organism ranged between 31 and 258 across both ionization

modes (Supplementary Table 4). This range is similar to that observed within a prior comparative study of the metabolomes of 12 distinct phytoplankton despite the use of different analytical platforms [172]. We observed that 31 and 77 features in negative and positive mode data, respectively, were present within multiple organisms, similar in size to previously reported work evaluating the core metabolome of phytoplankton [173]. Overall, untargeted features were primarily observed within samples belonging to a single organism. These trends are consistent with a prior transcriptomics study where most orthologous groups in these organisms did not overlap across taxa [24]. Hence, taxonomic differences in the metabolome may be transcriptionally conserved.

Prior studies show that nutrient stress status has large impacts on the transcriptomic composition of these organisms [24, 25]. Hence, we next sought to determine the role of nutrient stress status on organism-specific putatively characterized features. We considered only taxonomically-specific putatively characterized features to eliminate variability introduced by taxonomic differences between samples. We next applied distinct NMFs to each subset. Each NMF separated the data into three columns each representing samples from nutrient stress experimental factors in an unsupervised manner (Figure 2B & Supplementary Figure 4B), suggesting that nutrient stress status also contributed to the variability within the dataset. We detected proportionally fewer putatively characterized features with basis contribution scores of .99 or greater from this NMF than the one done on the entire set of putatively characterized features. These differences suggest that subtle differences in metabolite distributions may underly physiological stress. Hence, considering groups of metabolites may be a more powerful measure of physiological stress status than considering individual metabolites.

4.3.2 Metabolic differences explain taxonomic variability

We hypothesized that metabolic differences are driving the taxonomic clustering of features detected via NMF analysis. One possible driver for this trend is that the clustered putatively characterized features originate from distinct biochemical pathways. To evaluate this hypothesis, we sought to perform a pathway enrichment analysis. Currently, several methods exist to identify over-represented pathways within untargeted metabolomics data [122, 174, 175]. Applying these approaches to study organisms without sequenced genomes is

challenging, as they depend on *a priori* knowledge of the organism's biochemical pathway makeup and may be sensitive to missing intermediates due to low abundance or poor chemical ionization. To overcome these challenges, we crafted a novel NEtwork-based Permutation Test (NEPTune) to identify over-represented pathways within the clustered putatively characterized features (Figure 3).

Prior to using NEPTune on our data, we validated its efficacy by testing whether it could reproduce previously published results [170]. This study reported an enrichment in the activity of purine and tyrosine metabolisms based on the observance of three or more targets. Using these data, our permutation test returned *p-values* below 0.05 for these pathways, confirming the previously reported results.

Next, we applied NEPTune to each cluster of putatively characterized features. NEPTune detected a total of 39 significant ($p < 0.05$) over-represented pathways across all four organisms following quality assurance filtering (Figure 3 and Supplementary Table 5). To confirm the validity of these hypothesized over-representations, we sought to strengthen the annotation of features driving the pathway enrichments. To accomplish this goal, we applied the METLIN-guided in-source annotation (MISA) algorithm to check whether our putatively characterized features and coeluted ions matched reference spectra from METLIN. MISA reported that seven of our annotated features had a cosine score above 0.8 to their matching 0 volt MS/MS spectra (Supplementary Table 6). Six of the MISA annotated features belonged to the arachidonic acid metabolism pathway and the other one to retinol metabolism, in support of our annotation predictions. We upgraded the annotation of 8 features with MISA matches from level 3 to level 2 ("putative annotation"). These totals are on par with previous efforts to identify untargeted features on a similar analytical platform [88]. Our ability to generate more MISA annotations was limited by the compounds' abilities to form in-source fragments at 0 volts and the available MS/MS reference spectra within METLIN. The addition of more standards to METLIN could improve on the latter limitation in future studies.

The distribution of predicted pathways highlights the presence of both core and specialized metabolisms across the phytoplankton groups. Several core pathways like purine metabolism, glutathione metabolism and porphyrin and chlorophyll metabolism were observed

across multiple organisms. These observations likely reflect the fundamental physiological roles of these pathways for phytoplankton in processes like DNA synthesis, oxidative stress mitigation, and photosynthesis [125]. Indeed, prior studies found that several compounds belonging to these pathways were observed ubiquitously across many phytoplankton groups [173]. The observations of arachidonic acid metabolism and steroid hormone biosynthesis were unexpected due to their roles in secondary metabolism [158]. Both pathways have been identified in diatoms, and serve to initiate stress response and programmed cell death [158, 176]. To our knowledge, this is the first observation of these pathways within a prymnesiophyte, raphidophyte, or a coccolithophore, and they may serve these organisms to a similar capacity. Alternatively, these enrichments may be due to the presence of structural isomers, as many of these putatively characterized features were assigned from distinct adducts. Seven pathways were observed exclusively within individual organisms. Many of these pathways like retinol, tyrosine metabolism, and terpenoid biosynthesis have well known responses in cell signaling and are important components of secondary metabolism [177, 178]. One alternative explanation for these patterns is that they emerge from the differences in culturing conditions between organisms. Additional targeted metabolomics studies of compounds from these pathways across various taxa would help elucidate between these hypotheses. Future studies evaluating the expression of genes from these pathways within environmental datasets may reveal if they foster unique ecological processes.

In addition to profiling differences of well-characterized biochemical pathways, we evaluated whether distinct suites of chemical reactions characterized the metabolome of each organism. This type of analysis is possible due to advances in Global Natural Products Molecular Networking (GNPS) [165]. GNPS calculates the chemical similarity of features by correlating their MS/MS spectra. Features with correlations above a threshold form links within a network. The mass differences of the linked features may represent a specific chemical relationship that distinguishes two molecules with a common core structure [166]. We found 216 unique mass shifts across our data describing 23 distinct biochemical reactions. The distribution of mass shifts of 8 of these biochemical reactions across the organisms was significantly different ($p < 0.05$, *chi squared test*) from random (Supplementary Table 7), suggesting that several reactions

distinguish these taxa, including but not limited to glycosylation within the raphidophyte and ribosylation in the prymnesiophyte. Perhaps one explanation for these trends is that these reactions occur more frequently in these taxa due to differences in compound concentrations. Regrettably, we were not able to putatively identify any of these compounds as GNPS lacked reference MS/MS spectra to these data. Future enhancements to spectral database GNPS may ameliorate this issue.

4.3.3 Targeted analysis confirms hypothesized compound annotation and pathway enrichments

We analyzed targeted metabolite distributions in the remaining sample filters to confirm the NEPTune generated hypotheses. We measured a total of 135 unique metabolites, of which we detected between 93 and 118 within each organism (Supplementary Table 8). The raphidophyte dataset contained fewer total targets, as these data were analyzed for a prior study [42] and did not contain recently added standards to our existing method. The measured totals are in agreement with recent work on marine microbe *Sulfurimonas denitrificans* [179] and are similar to those measured in phytoplankton *Micromonas pusilla [180].* To further confirm the validity of our results, we benchmarked the measured concentrations of sulfur metabolites against previously published concentrations values of these compounds within similar organisms [181]. Most of the previously published values were an order of magnitude higher than those reported here (Supplementary Table 9). Observed differences may be due to the impact of nutrient stress on sulfur-containing metabolites and differences in culturing between the two experiments. We measured the matrix effects from each organism on our targeted compounds using matrix matched standard curves. Of the total set of compounds, we measured the matrix effect on 71 to 78 of them (Supplementary Table 9 Supplementary Table 8). High baseline concentrations of certain compounds precluded us from calculating additional matrix-matched standard curves. Of these, 44 to 51 of the matrix effects were considered acceptable (< 130 – see equation 1) [161]. Small organic and phosphate-containing compounds were particularly susceptible to matrix effects. Neither of these groups of compounds is chromatographically separated well, which may contribute to our difficulties in determining their matrix effects [182]. New analytical separation techniques may alleviate such issues.

Among the compounds we targeted were adenosine, adenosine monophosphate (AMP), glutamine (Gln), and glutamate (Glu). The ratios of adenosine to AMP and Gln to Glu are diagnostics for P- and N-stress, respectively, within diatoms, prymnesiophytes, and raphidophytes [42, 45, 180]. We calculated these ratios within the organisms in order to corroborate the stress status of these organisms and evaluate the sensitivity of the ratios across phytoplankton (Supplementary Figure 5). We observed that the expected trends occurred across organisms (low Gln to Glu and low adenosine to AMP under N-stress and P-stress, respectively). However, the Gln to Glu ratio was only significantly different within raphidophyte and coccolithophore cultures. The P-stress ratio was significantly higher in all organisms with the exception of the prymnesiophyte. It is possible that these ratios may become more pronounced upon extended nutrient stress [183]. Alternatively, the sensitivity of these ratios may be organism-specific. Additional studies of these ratios over time would clarify between these hypotheses.

To determine the efficacy of the NEPTune hypothesized pathway over-representations, we evaluated putatively characterized features used for predictions with targeted data. We measured several compounds driving the hypothesized pathway over-representation, which allowed us to check a total of 36 distinct putatively characterized features. Of these, 33 were correctly annotated (Figure 4a), raising their annotation to level 1 (“identified”). Two of the mismatches belonged to prostaglandin molecules, which contain a high number of structural isomers relative to other compounds. Of the total annotations, 26 putatively characterized features contained masses matching reported adducts within METLIN reference spectra. We confirmed all of these compounds [168]. This result suggests that restricting NEPTune predictions to annotations matching METLIN adducts increases the veracity of predictions (Figure 4a). Financial constraints and the limited commercial availability of standards limited our ability to confirm additional putatively characterized compounds.

In order to support the remaining NEPTune pathway enrichment hypotheses, we sought to evaluate the similarity between confirmed identifications and their matching putatively characterized untargeted features with two measures to check the concordance in abundance and retention time. We first evaluated whether the normalized intensity values of putatively

characterized features were proportional to their corresponding normalized measured concentrations. For this, we regressed the normalized targeted and matching untargeted abundances. Next, we calculated the residuals of these regressions, and used the distribution of these values to evaluate the homogeneity between matched data (Figure 4b). The empirical cumulative distribution functions for all residuals and those from compounds with METLIN matched adducts were comprised of 234 and 171 total residual measures, respectively. Over 46 percent of all measured residuals were less than 0.1, over 80 percent were less than 0.25, and over 95 percent were less than 0.4 across all measured compounds. Untargeted metabolomics data is subject to high variability due to differences in ionization efficiency across features, and coefficient of variation thresholds of 0.25 are suggested when applying quality control filtering to samples with well-defined sample matrices [94]. The majority of our measured residuals appear to abide by these standards, suggesting that we observed high concordance in abundance between datasets. Next, we assessed the retention time agreement between both datasets. For this, we regressed the retention time ranks of these observations against one another (Figure 4c). Due to the different chromatographic platforms used to generate the targeted and untargeted datasets, direct comparisons of measured retention time values are not possible. Outside of a few outlier points, our analysis revealed that the rank correspondence in the data was highly significant ($p < 0.001$, linear model) and showed a clear linear trend, supporting the retention time coherence between putatively characterized features and their targeted analyses. We conclude that the NEPTune reported pathway enrichments are based on correctly identified compounds, and support the validity of other hypothesized pathway enrichments we were unable to confirm. Taken together, these results suggest that NEPTune is capable of detecting viable pathway over-representations within untargeted metabolomics data.

4.3.4 Sparsity within untargeted data is driven by intracellular concentration differences

The trends from untargeted metabolomics data from this and prior [172, 173] studies revealed that features are primarily phytoplankton specific (Figure 2 and Supplementary Figure). This observation is unexpected considering that the taxa share aspects of core metabolism for central metabolic pathways [24]. Hence, we used our targeted data to reconcile these two

observations. We first determined whether our targeted data shared a similar mathematic structure to our untargeted data. For this, we performed a Procrustes analysis on the first two principal components of both datasets (Figure 5a). We chose to use the first two principal components as they explained a total of 70 and 40 percent of the variance within the targeted and untargeted datasets, respectively. The Procrustes analysis revealed that minimal transformations were required to align the ordinations from each dataset. A permutation test revealed that this degree of similarity was highly significant ($p < 0.001$, permutation test), suggesting that the variability in both datasets is due to a common underlying biological signal, i.e., taxonomic differences.

As the untargeted data was highly sparse (91% of feature table elements were zeroes), we sought to determine if this was also captured by the targeted data. However, only 18 out of 135 of targeted metabolites were present exclusively within a single organism (Supplementary Figure 6a). One possible explanation for this trend is that ion suppression and lower sensitivity of the mass spectrometer used for untargeted analysis artificially imposes this sparsity. To evaluate this hypothesis, we profiled distribution of concentrations of targeted metabolites using two measures, Shannon entropy and normalized concentration range (Figure 5b). The Shannon entropy provides a measure of evenness across concentration measurements, while the normalized concentration would provide a measure of the spread in concentration. We categorized metabolites into one of three domains based on their values of these two measures. The first domain characterizes metabolites with high entropy values and low normalized concentration ranges. The concentrations of these metabolites were similar across all organisms (Figure 5c). The second domain describes metabolites within one standard deviation of the mean of each measure. Metabolites within this domain were generally measured across all organisms, however with differences spanning several orders of magnitude (Figure 5d). The third domain describes metabolites with low entropy and high normalized concentration range. These metabolites are highly enriched exclusively in a single taxon (Figure 5e). 114 of 135 metabolites belonged to the latter two domains, suggesting that large concentration metabolite concentration differences between organisms drive the reported similarity by the Procrustes analysis. These results suggest that sparsity within untargeted data

is primarily caused by differences in metabolite concentrations rather than a lack of biosynthesis of these molecules. The utility of phytoplankton-specific subsets of untargeted features has been suggested as a viable *in situ* indicator of individual taxa [173]. However, our result suggests that these observations may also be an artifact of instrument sensitivity and chromatographic separation rather than biological differences among phytoplankton. Future experiments comparing phytoplankton metabolomes should include both targeted and untargeted data to determine the extent to which sparsity is due to biological differences between cultures.

4.3.5 Targeted metabolite distributions reveal unique physiological and ecological adaptations that distinguish these taxa

Applying NEPTune to the untargeted data suggested that the taxonomically clustered putatively characterized features may describe distinct biochemical processes. We sought to confirm this hypothesis using the targeted metabolite concentrations. We first normalized metabolite concentrations by cell volume [184], and performed a non-negative matrix factorization on all metabolites (Figure 6a). We assigned metabolites with basis contribution scores of 0.95 or greater to individual organisms. We chose to use a slightly more permissive threshold for the targeted data as there were fewer observations than in the untargeted data. We assigned 34 metabolites to individual organisms. The assignment of a metabolite to one organism does not preclude its non-zero measurement within others. Rather, it suggests that its concentration is highest in the organism where it was assigned.

We observed several instances where multiple compounds belonging to a single metabolic pathway were assigned to an individual organism. For example, we observed that sulfur-containing metabolites taurine, isethionic acid, and n-acetyltaurine were all assigned to the diatom (*Chaetoceros affinis*) (Figure 6a). A prior study shows that taurine and isethionic acid are enriched within diatoms relative to other phytoplankton and may be exchanged *in situ* with other marine phytoplankton [181]. N-acetyltaurine was not measured in that study, but may serve a similar role. Further targeted analysis evaluating the covariation of these metabolites within marine communities may help elucidate any connectivity between these compounds.

As a second example, compounds 3-methyl-2-oxopentanoic acid (3m2op) and 4-methyl-2-oxopentanoic acid (4m2op) were assigned exclusively to the raphidophyte samples (Figure 6a). These compounds are intermediates of branched amino acid degradation and biosynthesis [185]. Hence, we sought to evaluate the dynamics of these processes within the raphidophyte (Figure 6b). All measured branched chain amino acids were significantly depleted under stress relative to replete conditions ($p < 0.05$, Tukey-HSD, n = 3). By contrast the concentrations of 3m2op and 4m2op were enriched within stress relative to replete conditions. These dynamics may suggest that branched chain amino acid degradation may increase under stress. Branched chain amino acids have greater hydrophobicity relative to other amino acids [185], and primarily serve as proteinaceous building blocks. During catabolism, 3m2op and 4m2op lack the amino group of their precursors and can readily enter the TCA cycle, thus they may provide the raphidophyte with a carbon source for respiration and increased bioavailable N [125]. Alternatively, they may be enriched as a sink of excess carbon. Analysis of the transcriptome or isotope tracer experiments on the raphidophyte under these types of stress would clarify between these two hypotheses.

A third example involves the co-occurrence of metabolites related to glucosamine within the coccolithophore (Figure 6a). Glucosamine is produced through its catabolism via the degradation of deacylated chitin, chitosan [186]. Through our method, we measured the concentrations of several metabolites involved in chitin anabolism and catabolism, and chitosan catabolism (Figure 6c). Among these compounds, glucosamine-6-phosphate was significantly depleted ($p < 0.05$, Tukey-HSD Test, n = 3) under P-stress. We next evaluated the overall enrichment of metabolites involved in chitin anabolism and catabolism. We observed that under P-stress while chitin catabolism was significantly ($p < 0.05$, Wilcoxon-test, n = 6) enriched under N-stress was significantly depleted ($p < 0.05$, Wilcoxon-test, n = 6). Based on these trends, we hypothesize that N-stressed cells slow the catabolism of chitin in favor of the catabolism of chitosan. Chitin and chitosan represent cell surface acetylated and deacetylated amino sugar chains, respectively [186]. Changes in their relative abundance have been shown to impact physiology. For example, the conversion of chitin to chitosan was shown to regulate the growth of fungi [187]. Hence, these changes may have physiological consequences on the

coccolithophore. Other studies have shown that chitin is readily produced within other phytoplankton, [188]. Despite this, we detected glucosamine exclusively within the coccolithophore. Hence, the coccolithophore may have become more reliant on this strategy over the course of evolution than other marine phytoplankton. Future studies measuring the composition of these two molecules within the cell surface *in situ* would reveal how this process supports the organism's physiology.

Differences in overflow metabolism may also contribute to the distribution of organism-specific features [189, 190]. Overflow metabolism occurs when a metabolite is exuded from cells due to its increased intracellular production [190]. Exuded metabolites may then trigger distinct physiological responses in neighboring organisms [21]. The data reveals several organism-specific candidates for overflow metabolism: N-acetylserotonin, prostaglandins E2 A2 and D2, and spermidine (Figure 6a). These molecules were assigned to the coccolithophore, diatom, and raphidophyte cultures, respectively. Prior studies show that these molecules possess some cell-to-cell signaling capacity. Both N-acetylserotonin originates from tryptophan metabolism and has reported roles in cell-to-cell signaling within studies of the gut microbiome. Indeed, N-acetylserotonin functions as a neurotransmitter within metazoa [192]. To our knowledge, this is the first report of these molecules within marine phytoplankton, and the possible signaling roles of this compound must be validated with culture screening experiments. Alternatively, targeted searches of these compounds along with their receptor genes within public datasets may reveal examples of the biosynthesis and uptake of these molecules. Recent studies showed that the diatom *Thalassiosira rotula* increases synthesis of prostaglandins at the end of exponential growth phase, resulting in sustained increases of prostaglandins within the extracellular medium [158]. These authors hypothesize that the production of this molecule is designed to elicit cell-to-cell signaling within the population similar to metazoa [158]. Finally, spermidine is a well-studied polyamine and a viable source of nitrogen and carbon for marine microbes [193]. Prior studies hypothesize that dinoflagellate-derived spermidine may support the growth of bacteria such as *Ruegeria pomeroyi* [194]. These data suggest that these organisms may distinguish themselves from one another by employing distinct cell-to-cell signaling strategies. One alternative interpretation to these trends is that the

metabolite data highlighted here originates from bacteria rather than algae given that cultures were not axenic. Additional analyses using orthogonal yet taxonomically resolved data such as transcriptomics would help elucidate whether this was the case.

4.3.6 Responses to nutrient stress may reveal phytoplankton-specific acclimations

Nutrient stress status was a secondary driver of the variability in the untargeted data. We hypothesized this trend would hold within the targeted data. To confirm this hypothesis, we checked the distribution of significantly different metabolites across the organisms (Supplementary Figure b and c). In total, 54 and 58 of 135 metabolites were significantly different ($p < 0.05$, Tukey-HSD Test, n = 3) within N-stressed and P-stressed cells, respectively, relative to replete cells. The majority of significant metabolites under N-stress were only significant in one (33 total) or two phytoplankton taxa (18 total). Similarly, most significant metabolites under P-stress were only significant in one organism (39 total) or two phytoplankton taxa (13 total) (Supplementary Figure b and c). These distribution patterns may highlight specific stress acclimation strategies between the phytoplankton. Alternatively, they could be related to the differential degree of stress experienced by each organism.

We hypothesize that this observed stress response specificity is supported at a pathway level or across structurally similar compounds. Hence, we scored differences in compound distributions between case and control pairs using a measure that combines both statistical significance and the overall difference between groups, or effect size [169]. We applied this approach to compounds grouped into fourteen distinct categories based on their chemical characteristics (e.g., nucleosides, nucleobases) or known metabolic pathways (e.g. tryptophan metabolism) (Figure 7). 37 of the groups were significantly different under stress relative to replete conditions. Several groups showed similar activity across organisms under a single stress condition, suggesting their potential viability as stress-specific biomarkers. For example, nucleotides were significantly depleted across all organisms under P-stress, matching previously published results [45]. Another example is the enrichment of thiamine metabolism compounds under N-stress. All organisms showed an increased effect size and three out of four organisms showed significant enrichments ($p < 0.05$, Wilcoxon-test) of metabolites within this group. Thiamine is an essential cofactor of energy metabolism [195]. Its increase under N-stress may

suggest an increase in carbon respiration. Indeed, each organism with a significant enrichment of thiamine metabolism compound also shared an effect size greater than two for glycolysis compounds. Prior studies show that other diatoms upregulate glycolysis under N-stress to facilitate increased TCA cycling [58, 59, 196]. Hence, increased thiamine metabolism may support these diatom species due to its role in facilitating energy metabolism. Additional experiments with prymnesiophytes and raphidophytes are necessary to evaluate the generality of this response within other members of these groups.

In addition to capturing general stress response mechanisms, these data revealed several examples of phytoplankton-specific stresses. The prymnesiophyte appeared to show the starkest response, as 15 out of 28 molecule groups were enriched under both types of stress. Several of these enrichments were unique for this organism, including the Yang cycle, the nucleobases and chitin metabolism under N-stress. It is unclear whether these trends are due to increases in metabolic activity associated with metabolite groups or due to decreases in the rates of their consumption. Clarifying these two would be possible using fluxomic type approaches [197].

Interestingly, the prymnesiophyte featured the greatest response in effect size and degree of significance for tryptophan metabolism relative to the other organisms. Tryptophan metabolism serves a myriad of distinct roles in physiology through its capacity to form signaling compounds [125, 191, 198]. These signaling compounds have been shown to initiate both mutualistic [21] and antagonistic [22] interactions between phytoplankton and neighboring bacteria. It is possible that the prymnesiophyte relies on these or similar strategies to overcome nutrient stress.

We also observed interesting changes associated with the metabolite groups of the P-stressed coccolithophore that may highlight unique aspects of its metabolism. Under P-stress, this organism showed an exclusive enrichment of deoxynucleosides. These molecules are used as oxidative stress DNA damage biomarkers [199]. A major source of intracellular oxidative stress is energy metabolism [125]. Energy metabolism requires a steady flux of carbon from glycolysis or lipids to persist. Interestingly, glycolysis metabolites were significantly depleted under P-stress ($p < 0.05$, Wilcoxon-test). Perhaps this depletion was caused by an intentional

downregulation of energy metabolism to prevent additional oxidative stress. To support this hypothesis, we evaluated the concentration of citrate. Excess citrate is exported beyond the mitochondria where it allosterically inhibits glycolytic enzymes [125, 197]. The coccolithophore contained the greatest concentration of citrate relative to other phytoplankton. Additionally, its concentration was also significantly ($p < 0.05$, Tukey-HSD Test) higher under P-stress. Similar to the prymnesiophyte, the addition of fluxomic type experiments or genome scale modeling would evaluate this hypothesis.

4.4 Conclusion

Contemporary ocean circulation models predict that under future scenarios, marine primary production may decrease as early as 2090 due to increased scarcity of phosphorus and nitrogen [62]. This study hypothesizes that lack of these nutrients will impact phytoplankton communities in a uniform manner, thereby causing a decrease in community photosynthesis. This assumption does not necessarily account for the biological and chemical complexity of these communities. In this study, we attempted to illustrate the importance of this complexity by profiling the metabolism of four phytoplankton and their responses to nitrogen and phosphorus stress.

To facilitate biological discovery, we crafted and validated a new method for the analysis of untargeted metabolomics data. Our new method, NEPTune, predicted pathway-overrepresentations robustly that aided in the identification of new metabolites. This approach does not depend on knowledge of what genes are present within an organism, hence would serve investigators who currently study organisms without a sequenced genome. Our ability to thoroughly validate NEPTune was constrained to targets we could acquire commercially, and additional studies should be employed to confirm its efficacy.

Through the analysis of the measured metabolite concentrations across the taxonomic groups, we uncovered several examples of group-distinguishing changes in metabolism. These changes reflect systematic differences in metabolite concentrations driven by taxonomy. Their distribution highlights differences in signaling and biochemical pathways between organisms. We hypothesize that these differences underlie the unique physiology and ecology of these organisms and that these patterns may have arisen through the course of evolution. Additional

studies evaluating the genetic make-up of these organisms would help support the latter hypothesis. If these pathway and metabolite patterns are confirmed to be due to adaptations, they may serve as indicators of group-specific physiology and could be parameterized into models of phytoplankton communities. Such models may provide a mechanistic perspective on seasonal successions of phytoplankton driven by the changes in the composition of dissolved organic matter [20], and aid in our understanding of contemporary and future global primary production.

Secondly, by comparing the pathway dynamics within an organism across nutrient stress conditions, we identified several examples of acclimations to stress unique to each organism. These examples include both well-known nutrient stress biomarkers, and ones not previously reported. Though many organisms appeared to contain most of the evaluated acclimatory pathways, the dynamics of these pathways varied significantly between them and by stress conditions. Similar results were observed within transcriptomic studies [19]. Future studies should seek to identify which pathways are most prominent in driving the differences in physiology due to an external perturbation within these groups. This knowledge would provide a mechanistic basis for the study of phytoplankton community stability and help explain the observed differences in stoichiometric nutrient quotas across these organisms.

4.5 Figure

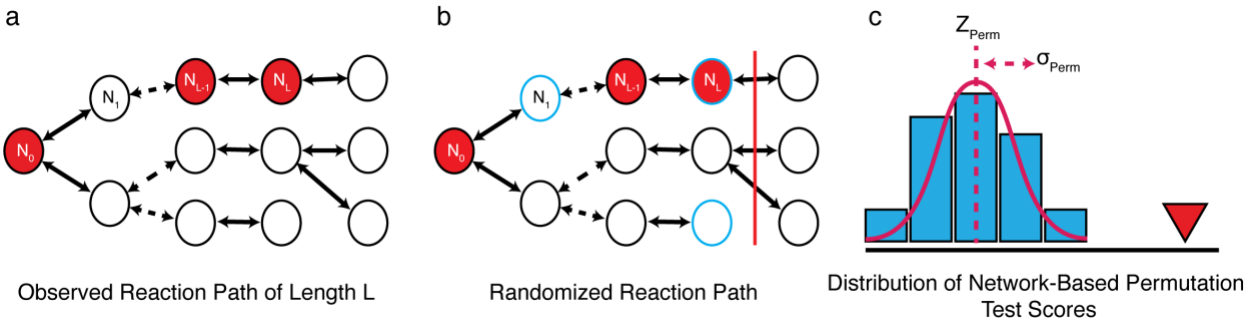


Figure 1: Schematic of network-based permutation test (NEPTune). a) An observed reaction path of length L within a KEGG pathway network. The circles denote KEGG pathway compounds, while arrows depict reactions. Dashed arrows are used to represent an unspecified series of reactions. Labeled nodes are nodes that belong to the reaction path $\{N_0, N_1, \dots, N_{L-1}, N_L\}$, and red nodes are nodes with an m/z match. b) Circles with blue outlines represent random draws from a network of radius L. NEPTune draws L-many nodes from the network randomly each time. c) Scores of empirical linked nodes (red triangle) and permutations from randomized draws (blue histogram). We evaluate the significance of the empirical linked nodes using the properties of the distribution of random draws.

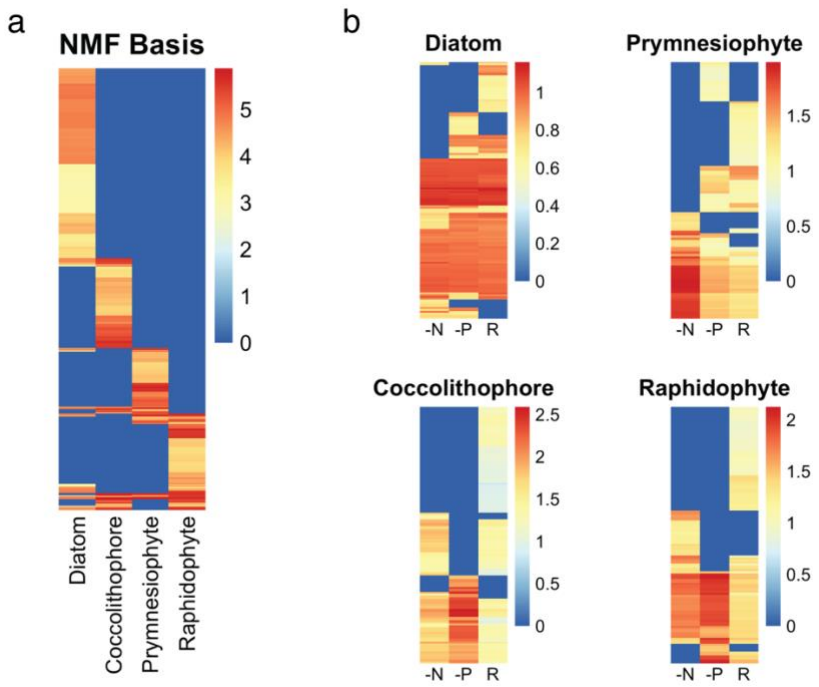


Figure 2: Non-negative Matrix Factorization (NMF) of positive mode untargeted metabolomics features. (a) Each column vector contains weights describing the contribution of features within groups of samples. The NMF assigned sample groups based on taxonomic differences in an unsupervised manner. (b) Taxon-specific features clustered via NMF groups by nutrient stress status. We determined the NMF's rank using the cophenetic correlation measure. The scale is a log-transformed unitless representation of the original data determined following matrix factorization. We used a cophenetic correlation measure to define the rank of these matrices. See Supplementary Figure for NMF of negative mode data. Abbreviations: -N – nitrogen stress, -P – phosphorus stress, and R – replete.

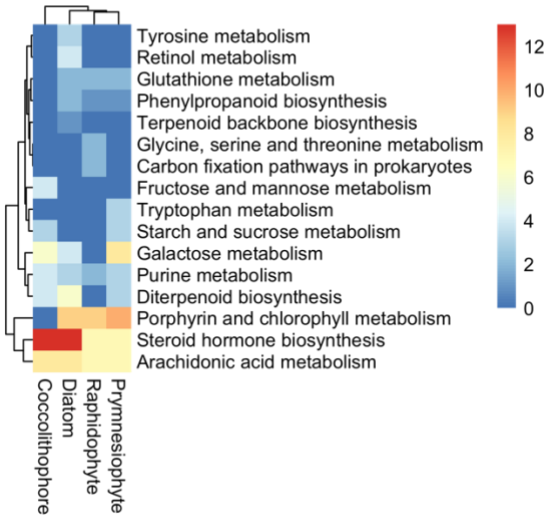


Figure 3: Significant ($p < 0.05$) metabolic pathway over-representations detected by NEPTune. Legend color corresponds to the number of m/z matched to unique network compounds observed within that pathway for a particular organism. Pathways are defined by KEGG networks.

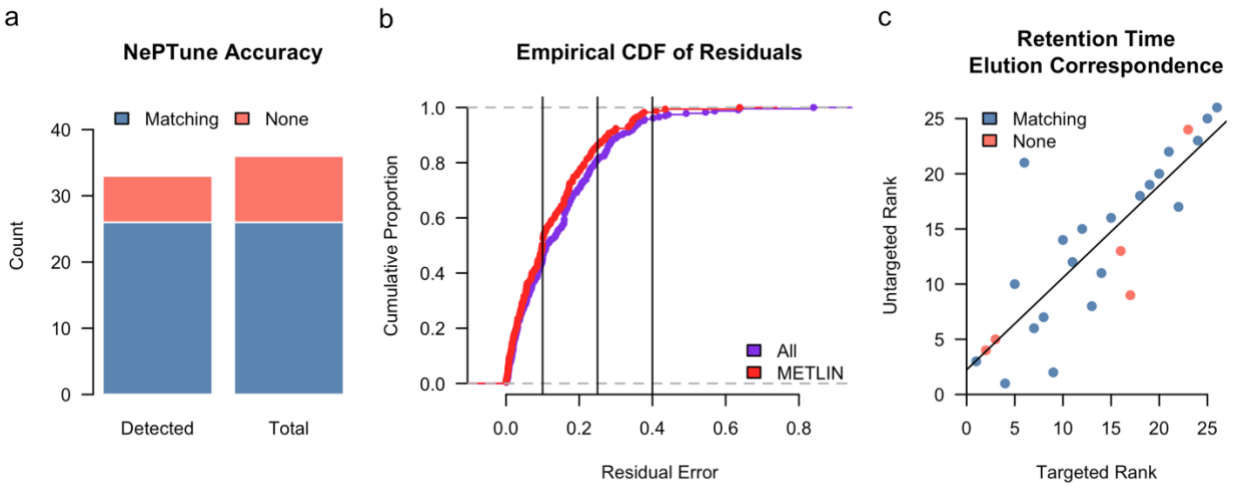


Figure 4: NEPTune Validation. a) Detection of compounds used for NEPTune pathway over-representations. Blue categories (matching) represent annotated features with masses matching METLIN adducts. Orange categories (None) represent untargeted feature annotations without matching METLIN adducts. b) Comparison of intensity dynamics between measured targeted compounds and annotated untargeted features. Concordance was measured using the residuals of a line of best fit between these two data types. Red Empirical Cumulative distribution function (CDF) represents the CDF of residuals of comparisons of annotated untargeted features with masses matching METLIN adducts. Purple represents all residuals. c) Retention time correspondence between annotated untargeted features and matching targets. Retention time ranks calculated using elution order. We used ranks instead of absolute retention time due to the differences in chromatographic platforms between these data.

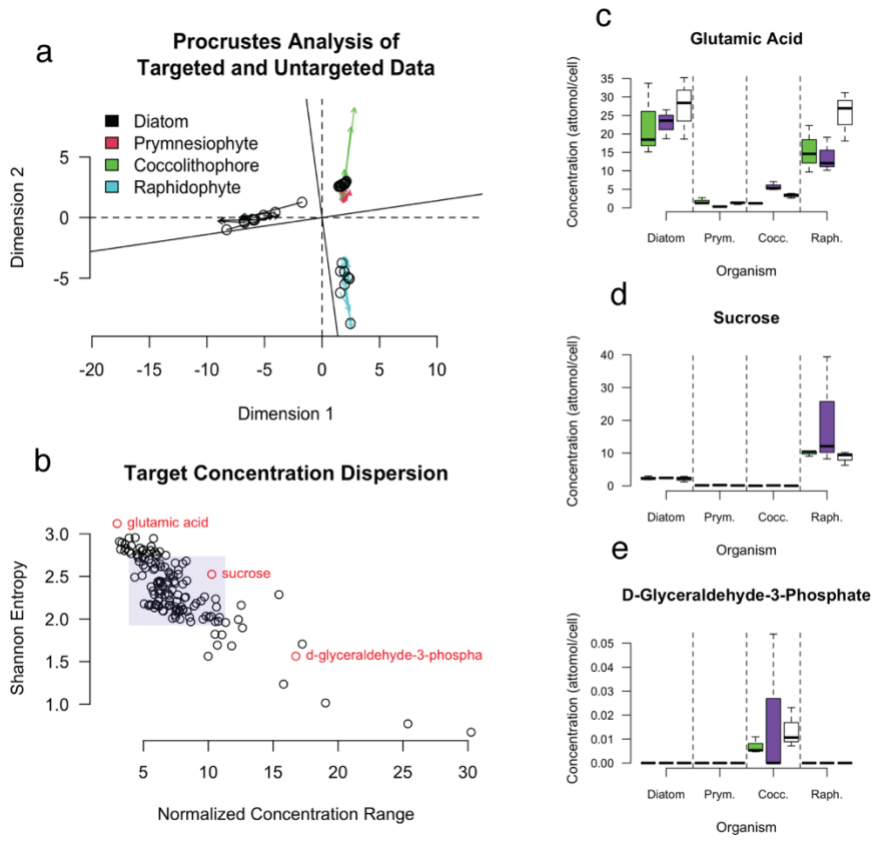


Figure 5: Mathematical structure of measured metabolites. a) Procrustes analysis of untargeted and targeted first two principal components. Circles and arrow heads represent the location in principle component space of untargeted samples and targeted samples, respectively. Length of arrow represents the amount of stretching between both datasets. Difference between solid and dashed axis lines represents the rotation required to align the datasets. First two principal components explained 59 and 40 percent of the variance within targeted and untargeted datasets, respectively. Similarity between both datasets was highly significant, suggesting that taxonomy servers as the primary driver of variability within targeted data (permutation test, $p < 0.001$). b) Distribution of targeted metabolites across inter-organism variability measures. Shannon entropy describes the evenness of data across factors while normalized concentration range describes the overall difference in magnitude. Blue square denotes area on plot one standard deviation from the mean of each marginal distribution. c-e) concentrations of metabolites denoted on b). Their dynamics represents the three types of trends observed across all molecules: c) small differences accros all organisms, d) large concentration

differences among organisms, e) metabolites are primarily detected within a single organism. Abbreviations: Prym. = prymnesiophyte, Cocc. = coccolithophore, Raph. = raphidophyte.

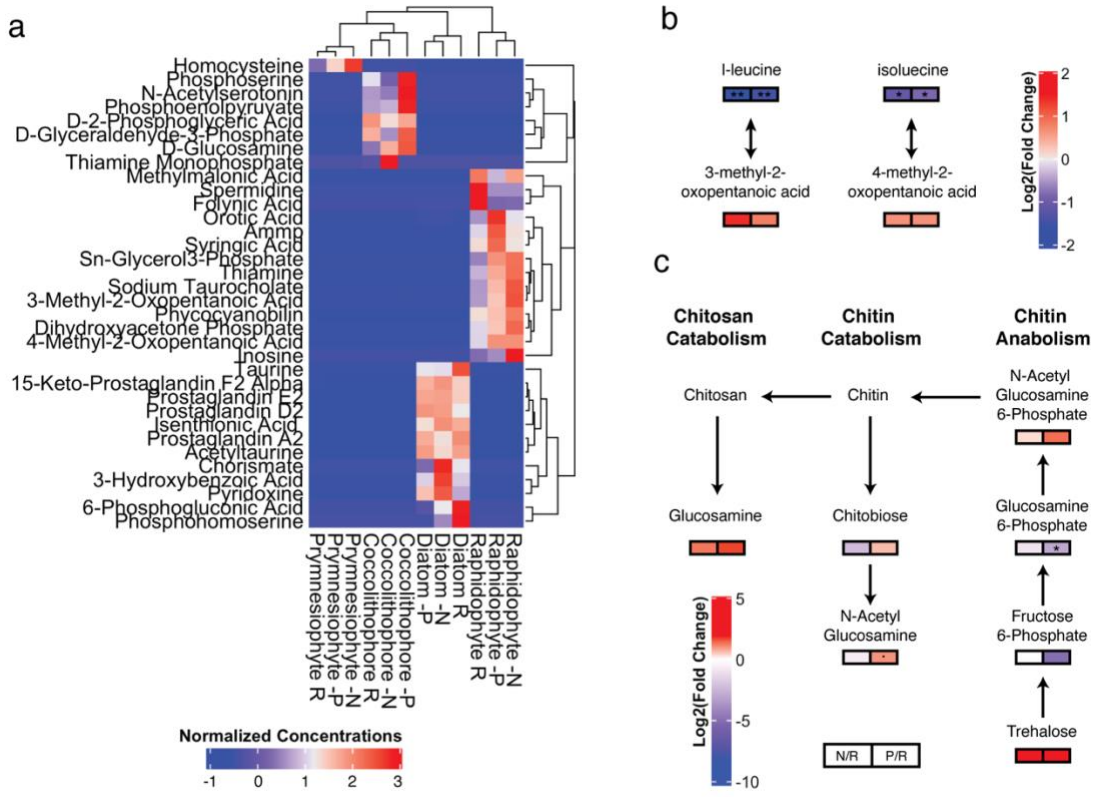


Figure 6: Distribution of taxonomically clustered metabolites and their biochemical processes. a) Distribution of targeted metabolites assigned to a phytoplankton via NMF. Metabolites were assigned to individual organisms if they had a feature contribution score at of at least 0.95. These metabolites are often highly abundant within one organism relative to other. b) Branched chain amino acid metabolism within the raphidophyte, H. akashiwo. Both 3-methyl-2-oxopentanoic acid and 4-methyl-2-oxopentaonic acid were assigned to the raphidophyte. These metabolites represent the degradation products of branched chain amino acids leucine and isoleucine, respectively. c) chitin metabolism within the coccolithophore G. oceanica. Glucosamine and fructose-6-phosphate were assigned to this organism from NMF analysis. The net concentration of Chitin anabolism metabolites was significantly depleted within P-stressed cells relative to replete ($p < 0.05$, Wilcoxon-test, $n = 12$). The net concentration of Chitin catabolism metabolites was significantly depleted within N-stressed cells relative to replete ($p < 0.05$, Wilcoxon-test, $n = 6$). . – $p < 0.1$, * - $p < 0.05$, ** - $p < 0.01$, *** - $p < 0.001$.

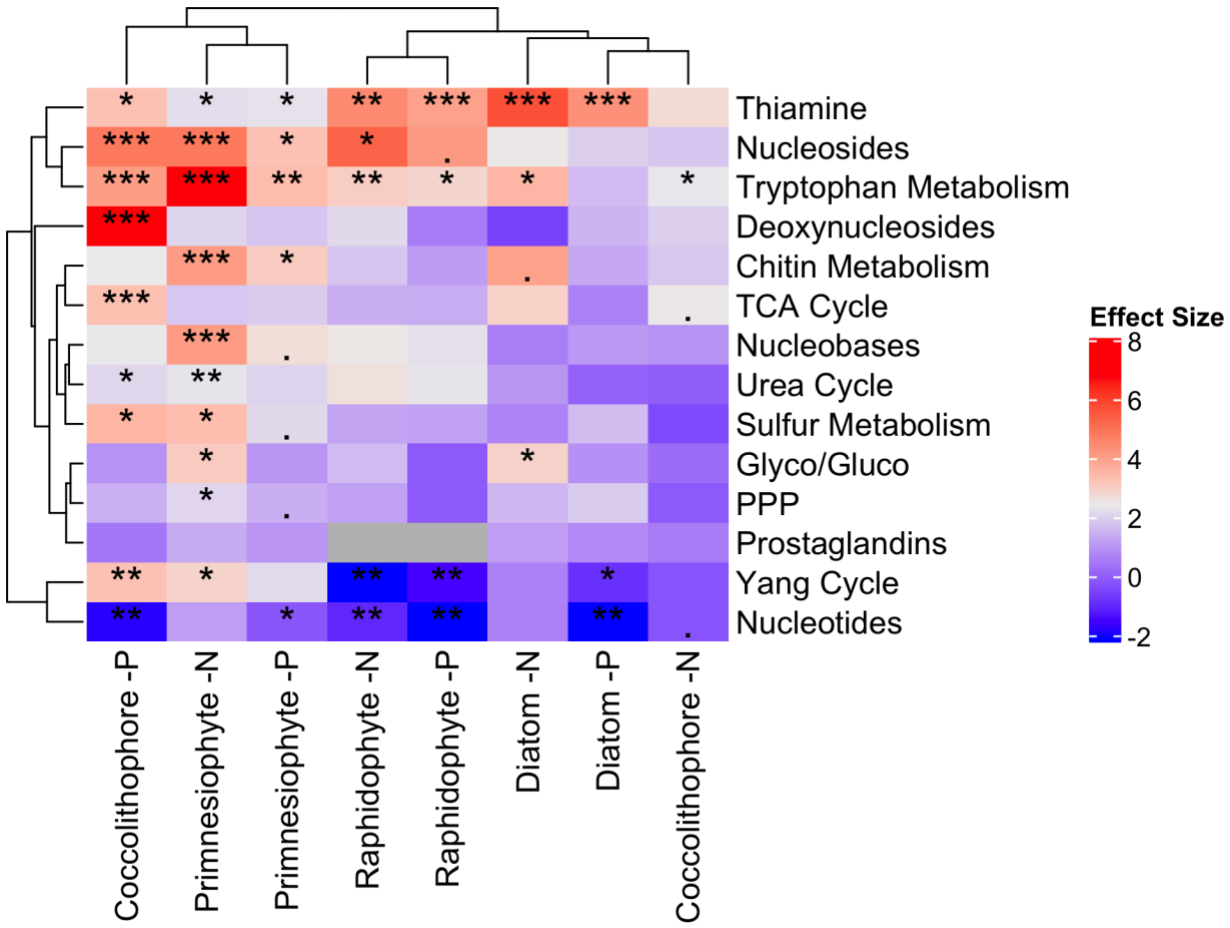
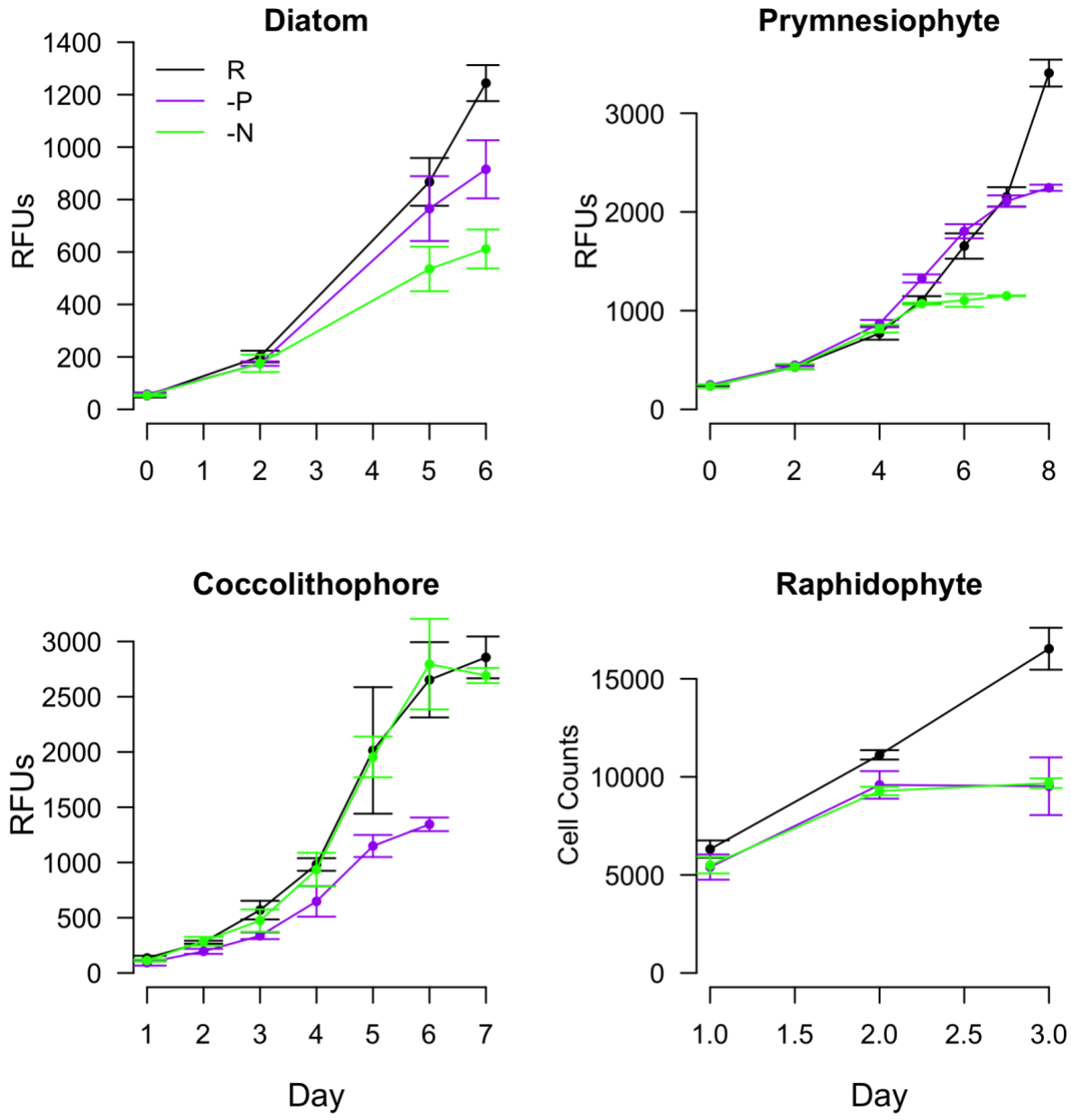


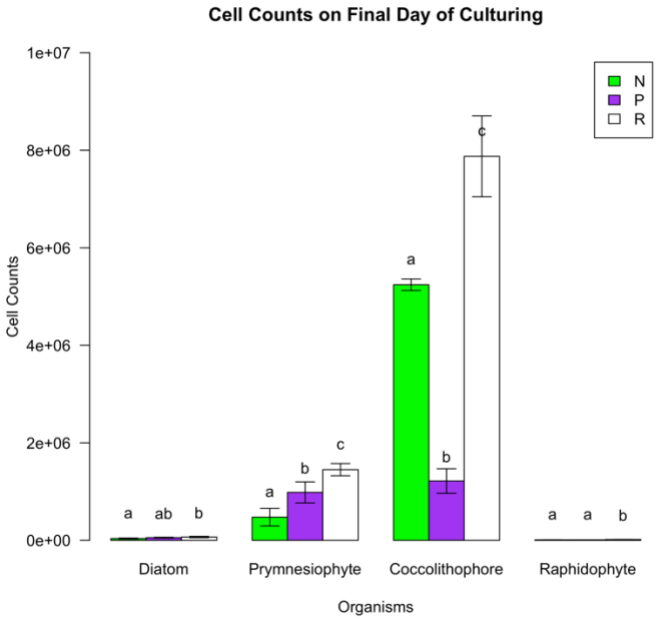
Figure 7: Enrichment of pathways and compound groups under nitrogen and phosphorus stress. Effect size describes the magnitude of change between case-control comparisons, where positive values indicate enrichment and negative ones indicate depletion. . – $p < 0.1$ * - $p < 0.05$, ** - $p < 0.01$, *** - $p < 0.001$. Abbreviations: PPP – pentose phosphate pathway.

4.6 Supplementary Material

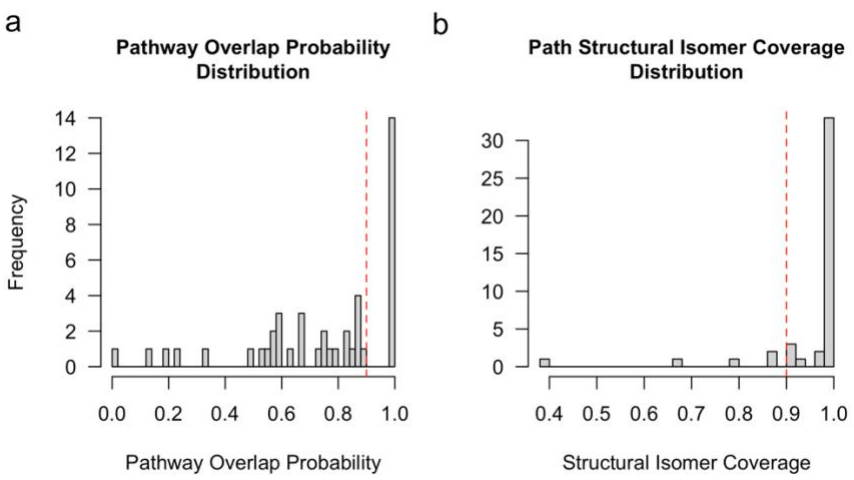
4.6.1 Supplementary Figures



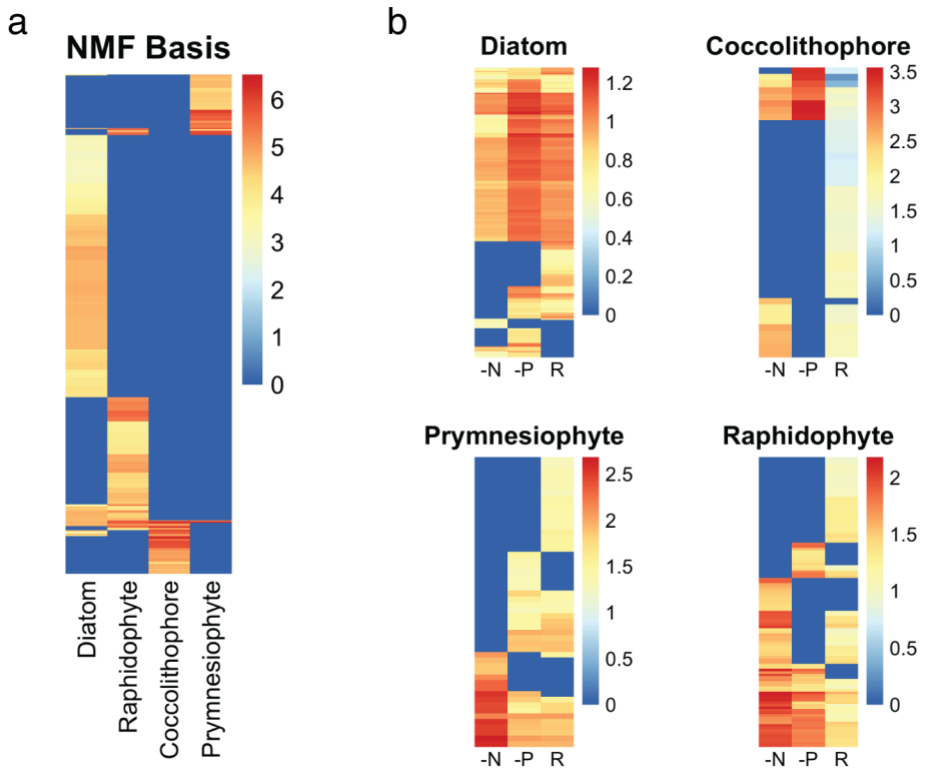
Supplementary Figure 1: Phytoplankton cell counts and relative fluorescent units (RFUs) under replete (R), P-stressed (-P), and N-stressed (-N) conditions. Error bars represent the standard deviation among triplicate measurements. We harvested cultures for metabolomics at the final measured timepoint of each growth curve, when nutrient stressed growth conditions diverged significantly from replete growth (ANOVA, $p < 0.05$, n = 3).



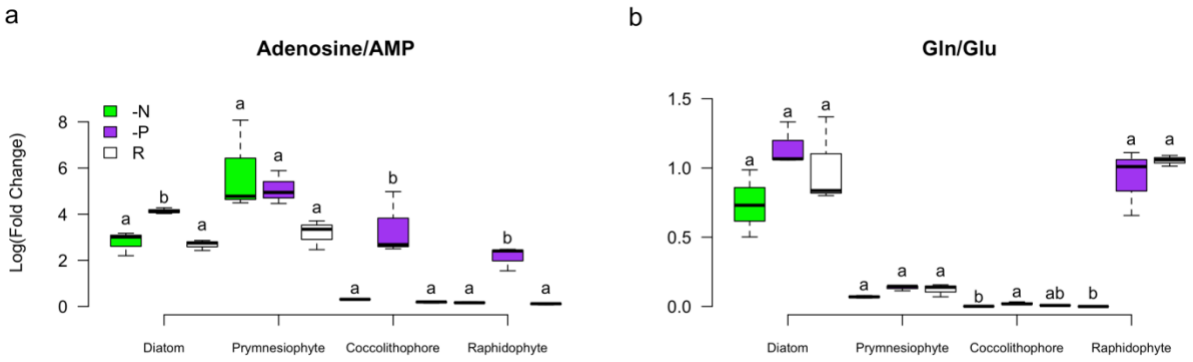
Supplementary Figure 2: Phytoplankton cell counts under replete (R), P-stressed (-P), and N-stressed (-N) conditions at the final time point of sampling. Error bars represent the standard deviation among triplicate measurements. Difference in letters between bars signifies that measurements were significantly from one another (ANOVA, $p < 0.05$, n = 3).



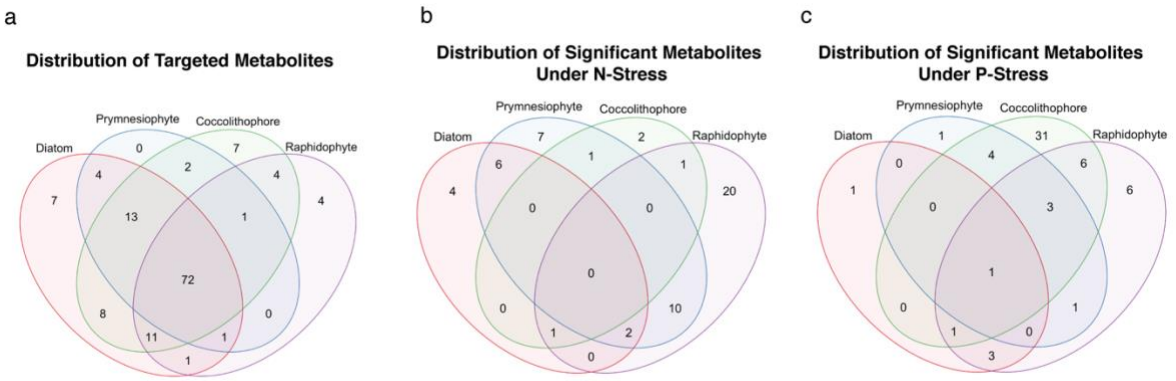
Supplementary Figure 3: Distribution of network-based permutation test filtering metrics of (a) Pathway Overlap Probability and (b) Path Structural Isomer Coverage. We retained paths with scores greater than 0.9 for Pathway Overlap Probability and Path Structural Isomer Coverage, (dashed lines).



Supplementary Figure 4: Non-negative Matrix Factorization (NMF) of negative mode untargeted metabolomics features. (a) Each column vector contains weights describing the contribution of features within groups of samples. The NMF assigned sample groups based on taxonomic differences in an unsupervised manner. (b) Taxon specific features clustered via NMF groups by nutrient stress status. We determined the NMF's rank using the cophenetic correlation. The scale is a log-transformed unitless representation of the original data determined following matrix factorization. We used a cophenetic correlation measure to define the rank of these matrices. Abbreviations: -N – nitrogen stress, -P – phosphorus stress, and R – replete.



Supplementary Figure 5: Measured phosphorus (a) and nitrogen (b) stress ratios. P-stress ratio (a) was formed from the measured concentrations of adenosine and adenosine monophosphate (AMP), while the N-stress ratio was formed from the measured concentrations of glutamine (Gln) and glutamate (Glu). Boxplots sharing a common letter are not significantly different from one another ($p > 0.05$). We only made comparisons across intra-organism stress samples.



Supplementary Figure 6: Distribution of targeted metabolite data across organisms. a) metabolites detected across organisms. Metabolite is counted if its measured concentration is above the limit of detection following QC. Distribution of significantly different metabolites in N-stressed (b) and P-stressed (c) cultures relative to replete.

4.6.2 Supplementary Tables

Supplementary Table 1: SRM transitions of validation compounds predicted by the network-based permutation test.

Target	**Parent**	**Confirm Ion**	**Quantification Ion**	**Retention Time (s)**
Prostaglandin F2 alpha	353.13	309.313	193.249	6.7
Prostaglandin E2	351.13	271.315	203.201	6.82
15-Keto Prostaglandin F2 alpha	333.116	219.209	173.353	6.92
Prostaglandin D2	351.13	271.315	189.243	7
Prostaglandin A2	333.124	271.329	315.309	7.62
N-Acetylserotonin	219.102	115.057	160.081	4.44
Anthranilate	120.059	65.049	92.056	4.88
6-Hydroxymelatonin	249.101	158.057	190.091	4.76

Supplementary Table 2: AutoTuner derived parameters used to process untargeted metabolomics data through XCMS.

Parameters	Negative	Positive
Parts Per Million Mass Error	1.4	1.8
Noise	900	7
Prefilter Intensity	1419	135
Prefilter Scan Count	2	2
Signal-to-Noise Threshold	3	3
Maximum Peak-width	48	108
Minimum Peak-width	4	3

Supplementary Table 3: Feature counts following QC of processed data from both ionization modes. Pre-QC represents the total number of features following XCMS processing. Post Blank represents the total number of features following blank correction. Post CV (coefficient of variation) represents the total number of features after removing high variability features, i.e., those features whose intensity values had coefficients of variation above 0.2 across organism specific pooled samples. In 2 out of 3 samples represents the features occurring in only two out of three experimental replicates.

QC Check	Positive Mode	Negative Mode
Pre-QC	19124	28865
Post Blank	17800	28133
Post CV	11263	18158
In 2 out of 3 samples	2564	4735

Supplementary Table 4: Distribution of organism specific features determined by NMF of untargeted metabolomics data. We used a relative basis contribution threshold of .99 (max 1) to assign features to individual organisms. Features within the column, No Organism, were not considered definitive for any organism.

Ionization Mode	Diatom	Raphidophyte	Coccolithophore	Prymnesiophyte	No Organism
Negative	215	88	31	43	33
Positive	258	81	111	74	76

Supplementary Table 5: Network-based permutation test quality filtering. Total significant paths represent the number of significant reaction paths determined by the network-based permutation test. Pathway overlap probability and structural isomer coverage rows represent the number of reaction paths retained after applying these filters.

Filter	Diatom	Raphidophyte	Prymnesiophyte	Coccolithophore
Total Significant Paths	12	14	11	7
Pathway Overlap Probability	11	11	10	7
Structural Isomer Coverage	11	11	10	7

Supplementary Table 6: Putatively identified compounds using METLIN-guided In-Source fragment Annotation (MISA) algorithm. The Feature Group column denotes the set of coeluting adduct ion species and in-source fragments mapped to available MS/MS spectra in METLIN to make putative identifications.

Compound	Fragments Matched	Total Fragments	Cosine Score	Feature Group
Leukotriene B	2	2	0.817	A
Retinoate	2	5	0.809	B
Leukotriene B	3	4	0.883	C
PGD2	3	6	0.998	D
PGE2	3	5	0.998	D
PGH2	3	6	0.854	D
PGD2	2	6	0.960	E
PGE2	2	5	0.979	E
PGH2	2	6	0.987	E
5(S)-HpETE	2	4	0.880	G
Leukotriene B	2	2	1.000	I
13,14-dihydro-15-ketone PGE2	2	2	0.850	J

Supplementary Table 7: Characteristic mass shifts occurring across organisms. Each row represents a mass shift whose observed distribution was significantly different ($p < 0.05$, chi squared test) from random.

Chemical Formula	Reaction	Diatom	Prymnesiophyte	Coccolithophore	Raphidophyte
C2H3NO	Glycine	23	4	0	0
C3H5NO	Alanine	7	0	0	1
C4H6N2O2	Asparagine	1	9	0	17
C5H8N2O2	Glutamine	1	12	1	12
C6H10O5	monosaccharide (-H2O)	5	5	3	15
C6H11NO	"Isoleucine, Leucine"	5	0	0	0
O	hydroxylation (-H)	17	28	2	5
C5H8O4	D-Ribose (-H2O) (ribosylation)	0	5	0	0

Supplementary Table 8: statistics of targeted data quality assurance. Each column is defined as follows: Targets detected in MAVEN – metabolites identified within MAVEN peak area integration software, Above LOD – metabolites with concentrations above the limit of detection, Passed CV Check – metabolites with a coefficient of variation below 0.4 across pooled samples, Constrained valid standard curve – metabolites with standard curves with R^2 values above 0.8 and 5 or more points in milliQ water, Constrained matched standard curve – metabolites with standard curves with R^2 values above 0.8 and 5 or more points in matrix, Acceptable matrix effect – metabolites with measured matrix effects below 130, Dereplicated counts – total unique metabolites detected.

Organism	Targets Detected in Maven	Above LOD	Passed CV Check	Contained Valid Standard Curve	Matrix Matched Standard Curve	Acceptable Matrix Effect	Dereplicated Counts
Diatom	184	171	165	161	34	5	117
Prymnesiophyte	184	144	135	131	29	5	93
Coccolithophore	189	173	171	163	31	3	118
Raphidophyte	156	140	131	123	NA	NA	94

Supplementary Table 9: Comparison of measured concentrations of sulfur metabolites to prior results [181]. Units – amol/cell.

Compound	Diatom *(chaetoceros affinis*)	Coccolithophore	Prymnesiophyte	Previously Reported Diatom	Previously Reported Haptophyte
DHPS	5-20	1-5	<1-3	92-2146	341-715
Cysteate	<1	<1	<1	1-99	2-3
Isethionate	600-1000	<1	<1	34626-99357	< 1
Taurine	<1-2	< 1	< 1	10-71	< 1

Chapter 5

Conclusions

This thesis begins the journey of linking phytoplankton metabolism to physiology through a series of detailed culture studies. The results serve as a foundation to accomplish similar aims within *in situ* communities. Each individual chapter helps bridge the technical and biological gaps that have precluded these efforts in the past. This knowledge may serve future investigations within the fields of metabolomics data science [200], harmful algal bloom formation [5], phytoplankton community ecology [201], and global carbon cycling [32].

Within the second chapter, I describe an algorithm, AutoTuner. I developed AutoTuner to facilitate the analysis of untargeted metabolomics data. More specifically, AutoTuner was designed to improve on existing data processing capabilities [116]. Prior to this, the existing solutions were computationally demanding, time consuming, and often inaccurate. Through the results of this chapter, I show that AutoTuner is robust, rapid, and trustworthy. This algorithm facilitated the analysis of data presented within chapters 3 and 4. The algorithm could be improved in many ways. For one, prior to running this algorithm, the user must perform a sliding window analysis on peaks within the mass chromatogram of individual samples. More sophisticated signal processing algorithms could automate this function, which would increase the scalability of the algorithm. Secondly, many practitioners of metabolomics include internal standards into their samples to ensure the fidelity of the mass spectrometry measurements and data processing. Perhaps AutoTuner could include a post-processing validation check on these peaks to automatically evaluate the fidelity of processing.

Chapter 3 describes an analysis of metabolite and gene expression data to characterize the metabolism of the harmful algal bloom formatting raphidophyte *Heterosigma akashiwo.* Through this effort, I constructed conceptual models that describe how the metabolism of this organism responds to the acute shortage of nitrogen (N) and phosphorus (P) (nutrient stress). These models were dissimilar to previously reported models of N- and P-stress metabolism of diatoms and coccolithophores. Diatoms and coccolithophores are among the most well-studied phytoplankton. Our research suggests that additional studies evaluating the metabolisms of less-well studied phytoplankton groups would greatly expand on the known suite of stress response mechanisms. Understanding these responses is key to predicting, managing, and concluding harmful algal blooms [6, 13]. The presented results provide several impactful areas

for continued investigation. Firstly, there is strong similarity in the metabolism of *H. akashiwo* under P-stress and while it is within a cyst state. For example, *H. akashiwo* reduces triacylglyceride lipids as a cyst [202] and under P-stress [42]. A similar analysis to the one described within chapter 3 using encysted *H. akashiwo* cells would confirm this hypothesis. Secondly, we uncover several previously unrecognized responses to stress which may provide a fitness advantage to this organism and foster bloom formation. A high resolution metatranscriptomic and metabolomic analysis of an *in situ H. akashiwo* bloom could reveal whether these pathways improve organismal fitness prior, during, and post a bloom. Additionally, such field campaigns would help confirm the viability of the reported nutrient stress biomarkers presented within this chapter. The approach presented here may be readily adapted to study the metabolism of other harmful algal bloom forming phytoplankton, as it does not depend on the availability of a fully sequenced genome.

Finally, within the fourth chapter of this thesis, I reported the distinguishing adaptations and acclimations of four phytoplankton from three phyla. These differences are hypothesized to support the coexistence of these organisms through their unique impacts on physiology [19, 31]. Through this effort, I found several examples of pathways and signaling molecules that appear within single organisms. I hypothesize that their occurrence may be due to organism specific adaptations. Additionally, I reported several examples of pathways that organisms use to acclimate to nutrient stress. Although most organisms contained each evaluated pathway, the response to stress often varied between them. This suggested that each organism performs their own acclimatory strategies to stress. These findings may be followed in several distinct ways to obtain a better understanding of phytoplankton community ecology and marine carbon cycle. First, the dynamics of the pathways we identified may be evaluated during experiments characterizing physiological behavior of unique phytoplankton group. For example, they could be explored for their roles in diatom aggregation [203], raphidophyte swimming [204], or coccolithophore sinking. A secondary avenue of further exploration involves the evaluation of the role of group-specific signaling compounds. Prior studies show that phytoplankton-derived marine signaling compounds may produce a systems level shift within sympatric bacteria [21, 22, 88]. My study revealed several potential candidate molecules for future exploration. Finally,

a third area would be to check the efficacy of group specific stress markers to determine whether they can reveal *in situ* stress status. These results may provide valuable information for modelers predicting how an influx of timely nutrients would impact community structure and carbon sequestration.

Perhaps the most exciting area for continued research following this thesis would be the exploration of the causal factors driving the diatom-specific enrichment of prostaglandins. These molecules have long been studied for their role within various aspects of metazoan physiology from the initiation of inflammation, initiation of distinct disease states, and cell-to-cell signaling [205]. In 2017, a study showed that the diatom *Skeletonema costatum* was capable of producing prostaglandins and contained prostaglandin biosynthesis genes similar to those in metazoa [206]. In subsequent work, these investigators have shown that the extracellular concentrations of these molecules within cultures of diatom *Thalassiosira rotula* peak upon the initiation of stationary growth phase when faced with silica limitation [158]. Additionally, using genomic data mining approaches, they have posited that dinoflagellates are also capable of producing prostaglandins following the annotation of prostaglandin biosynthesis genes from reference transcriptomes [207]. Within this study, we show the first evidence that these molecules are also produced within a coccolithophore and a prymnesiophyte. Additionally, we expand on the range of known diatoms capable of producing prostaglandins, by confirming that diatom *Chaetoceros affinis* is capable of this function.

Interestingly, the concentrations of all measured prostaglandins were far higher in *C. affinis* than other phytoplankton under all growth media states. We hypothesize that these differences suggest that *C. affinis* is more likely to exude these molecules due to over-flow metabolism than other phytoplankton. If this is the case, then prostaglandins may serve diatoms as mediators of cell-to-cell signaling similar to metazoa. Within higher organisms, these molecules are known to trigger a transition in metabolic state, hence they may cause similar effects to whatever aquatic recipient is responding to their presence. Prior work suggests that these changes are driven by stress; hence such a metabolic change may be associated with an ecological strategy.

Several possible follow-up experiments may determine whether prostaglandins trigger a metabolically driven change in physiology and if this process is ecologically relevant. First, there are many distinct prostaglandin molecules. Careful culture experiments may help narrow down the ones which most affect diatom physiology. Following this, observing how physiology changes with the addition of key prostaglandins followed by transcriptomic analyses may reveal the mechanisms driving the changes. An alternative strategy would be to mine existing datasets containing diatoms for either transcriptomic or metabolomic evidence that prostaglandin biosynthesis or availability increases with diatom population size. Such an approach would be more tractable than the experimental route, and could help facilitate targeted experiments. Understanding the impact of prostaglandins on a microbial community will be key to exploring their impact in higher community processes related to the physiology of diatoms.

My choice of prostaglandins as an interesting area of continued research is a little biased by the known role of these molecules in metazoa. Many other molecules and pathways identified here hold similar promise. Another example would be tryptamine, or the entire suite of molecules within tryptophan metabolism. These molecules were all enriched under stress within the prymnesiophyte culture described within chapter four. In general, such molecules feature a variety of functions ranging from growth [152], and cell-to-cell signaling [21], to antagonism [22]. Due to their increase under stress, we hypothesize that they may serve the studied prymnesiophyte in a similar manner. Perhaps the most tractable means of ascertaining this would be to evaluate available metatranscriptomics data that has been mapped to the prymnesiophyte described here. The expression of genes involved in this pathway may be normalized by other genes shown to denote nutrient stress status, like inorganic phosphate transporters [24], to determine within which areas of the ocean these genes are most likely to be engaged. Finally, considering which neighboring microbes correlate with pathway engagement may provide some testable hypotheses for which organism may respond to the increased availability of these molecules.

Together these findings help bridge our understanding of phytoplankton physiology and ecology by evaluating stress response through the lenses of biochemistry and metabolism. These results help advance several of the posited hypotheses during the beginning of this

thesis, either by directly providing knowledge towards their resolution or by providing a foundation to begin asking more targeted questions. In regards to bloom formation, we may consider the pathways level changes detected within *H. akashiwo*. Such changes may be considered while understanding how blooms occur and persist, while the approach may be adapted to study the bloom formation of other species. For coexistence, consider the distinct pathways detected by each phytoplankton described in chapter 4. Their metabolisms were highly dissimilar, along with their responses of stress. Perhaps these stress response pathways may serve as indicators of *in situ* stress status for these organisms. Determining whether an organism is experiencing phosphorus stress would certainly help improve forecasts of community structure to perturbations like the addition of dissolved phosphate from groundwater. This foundation of knowledge may be expanded to understand processes that take place on local and global scales. Moreover – I hope that these findings end up serving diverse stakeholders, thereby improving upon existing problems.

References

1. Sarmiento, J.L. and N. Gruber, *Ocean Biogeochemical Dynamics*. 2006: Princeton University Press.
2. Field, C.B., et al., *Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components.* Science, 1998. **281**(5374): p. 237.
3. Emerson, S. and J. Hedges, *Chemical Oceanography and the Marine Carbon Cycle*. 2008, Cambridge: Cambridge University Press.
4. Wells, M.L., et al., *Harmful algal blooms and climate change: Learning from the past and present to forecast the future.* Harmful Algae, 2015. **49**: p. 68-93.
5. Anderson, D.M., et al., *Harmful algal blooms and eutrophication: Examining linkages from selected coastal regions of the United States.* Harmful Algae, 2008. **8**(1): p. 39-53.
6. Smayda, T.J., *Harmful algal blooms: Their ecophysiology and general relevance to phytoplankton blooms in the sea.* Limnology and Oceanography, 1997. **42**(5): p. 1137-1153.
7. Hoagland, P., et al., *The economic effects of harmful algal blooms in the United States: Estimates, assessment issues, and information needs.* Estuaries, 2002. **25**(4): p. 819-837.
8. Taylor, F.J.R. and R. Haigh, *The ecology of fish-killing blooms of the chloromonad flagellate heterosigma in the strait of georgia and adjacent waters.* Toxic Phytoplankton Blooms in the Sea, 1993: p. 705-710.
9. Tyrrell, T., *The relative influences of nitrogen and phosphorus on oceanic primary production.* Nature, 1999. **400**(6744): p. 525-531.
10. Moore, C.M., et al., *Processes and patterns of oceanic nutrient limitation.* Nature Geoscience, 2013. **6**(9): p. 701-710.
11. Arrigo, K.R., *Marine microorganisms and global nutrient cycles.* Nature, 2005. **437**(7057): p. 349-355.
12. Loureiro, S., et al., *Harmful algal blooms (HABs), dissolved organic matter (DOM), and planktonic microbial community dynamics at a near-shore and a harbour station influenced by upwelling (SW Iberian Peninsula).* Journal of Sea Research, 2011. **65**(4): p. 401-413.
13. Wilson, S.T., et al., *Kīlauea lava fuels phytoplankton bloom in the North Pacific Ocean.* Science, 2019. **365**(6457): p. 1040.
14. Finkel, Z.V., et al., *Phylogenetic Diversity in the Macromolecular Composition of Microalgae.* PLOS ONE, 2016. **11**(5): p. e0155977.
15. Bonachela, J.A., et al., *The role of phytoplankton diversity in the emergent oceanic stoichiometry.* Journal of Plankton Research, 2016. **38**(4): p. 1021-1035.
16. Klausmeier, C.A., et al., *Optimal nitrogen-to-phosphorus stoichiometry of phytoplankton.* Nature, 2004. **429**(6988): p. 171-174.
17. Caron, D.A., et al., *Probing the evolution, ecology and physiology of marine protists using transcriptomics.* Nature Reviews Microbiology, 2017. **15**(1): p. 6-20.
18. Alexander, H., et al., *Metatranscriptome analyses indicate resource partitioning between diatoms in the field.* Proceedings of the National Academy of Sciences, 2015. **112**(17): p. E2182.

19. Alexander, H., et al., *Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean.* Proceedings of the National Academy of Sciences, 2015. **112**(44): p. E5972.
20. Teeling, H., et al., *Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom.* Science, 2012. **336**(6081): p. 608.
21. Amin, S.A., et al., *Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria.* Nature, 2015. **522**(7554): p. 98-101.
22. Segev, E., et al., *Dynamic metabolic exchange governs a marine algal-bacterial interaction.* Elife, 2016. **5**.
23. Keeling, P.J., et al., *The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing.* PLOS Biology, 2014. **12**(6): p. e1001889.
24. Harke, M.J., et al., *Conserved Transcriptional Responses to Nutrient Stress in Bloom-Forming Algae.* Front. Microbiol., 2017. **8**: p. 1279.
25. Haley, S.T., et al., *Transcriptional response of the harmful raphidophyte Heterosigma akashiwo to nitrate and phosphate stress.* Harmful Algae, 2017. **68**: p. 258-270.
26. Dyhrman, S.T., et al., *The transcriptome and proteome of the diatom Thalassiosira pseudonana reveal a diverse phosphorus stress response.* PLoS One, 2012. **7**(3): p. e33768.
27. Walworth, N., et al., *Trichodesmium genome maintains abundant, widespread noncoding DNA in situ, despite oligotrophic lifestyle.* Proceedings of the National Academy of Sciences, 2015. **112**(14): p. 4251.
28. Livolant, F. and Y. Bouligand, *New observations on the twisted arrangement of Dinoflagellate chromosomes.* Chromosoma, 1978. **68**(1): p. 21-44.
29. Hutchinson, G.E., *The Paradox of the Plankton.* The American Naturalist, 1961. **95**(882): p. 137-145.
30. Hu, S.K., et al., *Shifting metabolic priorities among key protistan taxa within and below the euphotic zone.* Environmental Microbiology, 2018. **20**(8): p. 2865-2879.
31. Lampe, R.H., et al., *Strategies among phytoplankton in response to alleviation of nutrient stress in a subtropical gyre.* The ISME Journal, 2019. **13**(12): p. 2984-2997.
32. Worden, A.Z., et al., *Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes.* Science, 2015. **347**(6223): p. 1257594.
33. Kirchman, D.L., *Processes in microbial ecology*. 2018: Oxford University Press.
34. Patti, G.J., O. Yanes, and G. Siuzdak, *Innovation: Metabolomics: the apogee of the omics trilogy.* Nature reviews. Molecular cell biology, 2012. **13**(4): p. 263-269.
35. Allen, A.E., et al., *Evolution and metabolic significance of the urea cycle in photosynthetic diatoms.* Nature, 2011. **473**(7346): p. 203-207.
36. Allen, A.E., et al., *Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation.* Proc. Natl. Acad. Sci. U. S. A., 2008. **105**(30): p. 10438-10443.
37. Wördenweber, R., et al., *Phosphorus and nitrogen starvation reveal life-cycle specific responses in the metabolome of Emiliania huxleyi (Haptophyta).* Limnology and Oceanography, 2018. **63**(1): p. 203-226.

38. Smith, S.R., et al., *Evolution and regulation of nitrogen flux through compartmentalized metabolic networks in a marine diatom.* Nature Communications, 2019. **10**(1): p. 4552.
39. Johnson, W.M., et al., *Insights into the controls on metabolite distributions along a latitudinal transect of the western Atlantic Ocean.* bioRxiv, 2021: p. 2021.03.09.434501.
40. Johnson, W.M., et al., *Metabolite composition of sinking particles differs from surface suspended particles across a latitudinal transect in the South Atlantic.* Limnology and Oceanography, 2020. **65**(1): p. 111-127.
41. Boysen, A.K., et al., *Diel Oscillations of Particulate Metabolites Reflect Synchronized Microbial Activity in the North Pacific Subtropical Gyre.* bioRxiv, 2020: p. 2020.05.09.086173.
42. McLean, C., et al., *Harmful Algal Bloom-Forming Organism Responds to Nutrient Stress Distinctly From Model Phytoplankton.* bioRxiv, 2021: p. 2021.02.08.430350.
43. Zamboni, N., A. Saghatelian, and G.J. Patti, *Defining the metabolome: size, flux, and regulation.* Mol Cell, 2015. **58**(4): p. 699-706.
44. Alvarez, L., et al., *Bacterial secretion of D-arginine controls environmental microbial biodiversity.* The ISME Journal, 2018. **12**(2): p. 438-450.
45. Kujawinski, E.B., et al., *Phosphorus availability regulates intracellular nucleotides in marine eukaryotic phytoplankton.* Limnology and Oceanography Letters, 2017. **2**(4): p. 119-129.
46. White, R.A., et al., *The past, present and future of microbiome analyses.* Nature Protocols, 2016. **11**(11): p. 2049-2053.
47. Ren, S., et al., *Computational and statistical analysis of metabolomics data.* Metabolomics, 2015. **11**(6): p. 1492-1513.
48. Myers, O.D., et al., *Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data.* Anal Chem, 2017. **89**(17): p. 8689-8695.
49. Brodsky, L., et al., *Evaluation of Peak Picking Quality in LC−MS Metabolomics Data.* Analytical Chemistry, 2010. **82**(22): p. 9177-9187.
50. Smayda, T.J., *Ecophysiology and bloom dynamics of Heterosigma akashiwo (Raphidophyceae)*, in *Physiological Ecology of Harmful Algal Blooms*, A.D.C.A.G.M.H. D. M. Anderson, Editor. 1998, Springer: Berlin.
51. Rensel, J.E.J., *Fish kills from harmful alga Heterosigma akashiwo in Puget Sound: Recent blooms and review*. 2007: Arlington, Washington. p. 63.
52. Shikata, T., et al., *Factors influencing the initiation of blooms of the raphidophyte Heterosigma akashiwo and the diatom Skeletonema costatum in a port in Japan.* Limnology and Oceanography, 2008. **53**(6): p. 2503-2518.
53. Ji, N., et al., *Metatranscriptome analysis reveals environmental and diel regulation of a Heterosigma akashiwo (raphidophyceae) bloom.* Environ. Microbiol., 2018. **20**(3): p. 1078-1094.
54. Van Mooy, B.A.S., et al., *Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity.* Nature, 2009. **458**(7234): p. 69-72.
55. Grzymski, J.J. and A.M. Dussaq, *The significance of nitrogen cost minimization in proteomes of marine microorganisms.* The ISME Journal, 2012. **6**(1): p. 71-80.

56. Read, R.W., et al., *Nitrogen cost minimization is promoted by structural changes in the transcriptome of N-deprived Prochlorococcus cells.* The ISME Journal, 2017. **11**(10): p. 2267-2278.
57. Brembu, T., et al., *The effects of phosphorus limitation on carbon metabolism in diatoms.* Philos. Trans. R. Soc. Lond. B Biol. Sci., 2017. **372**(1728).
58. Alipanah, L., et al., *Molecular adaptations to phosphorus deprivation and comparison with nitrogen deprivation responses in the diatom Phaeodactylum tricornutum.* PLOS ONE, 2018. **13**(2): p. e0193335.
59. Alipanah, L., et al., *Whole-cell response to nitrogen deprivation in the diatom Phaeodactylum tricornutum.* Journal of Experimental Botany, 2015. **66**(20): p. 6281-6296.
60. Rokitta, S.D., et al., *P- and N-Depletion Trigger Similar Cellular Responses to Promote Senescence in Eukaryotic Phytoplankton.* Frontiers in Marine Science, 2016. **3**(109).
61. Rokitta, S.D., et al., *Emiliania huxleyi endures N-limitation with an efficient metabolic budgeting and effective ATP synthesis.* BMC Genomics, 2014. **15**: p. 1051.
62. Bopp, L., et al., *Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models*. 2013: Biogeosciences. p. 6225–6245.
63. Moreno, A.R. and A.C. Martiny, *Ecological Stoichiometry of Ocean Plankton.* Annual Review of Marine Science, 2018. **10**(1): p. 43-69.
64. Chong, J., et al., *MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis.* Nucleic Acids Res, 2018. **46**(W1): p. W486-w494.
65. Busch, K.L. *Chemical Noise in Mass Spectrometry*. *Spectroscopy*, 2003. **18** 52 – 55.
66. Lommen, A.
67. Jiang, W., et al., *An Automated Data Analysis Pipeline for GC−TOF−MS Metabonomics Studies.* Journal of Proteome Research, 2010. **9**(11): p. 5974-5981.
68. Röst, H.L., et al., *OpenMS: a flexible open-source software platform for mass spectrometry data analysis.* Nature Methods, 2016. **13**(9): p. 741-748.
69. Samanipour, S., et al., *Self Adjusting Algorithm for the Nontargeted Feature Detection of High Resolution Mass Spectrometry Coupled with Liquid Chromatography Profile Data.* Analytical Chemistry, 2019. **91**(16): p. 10800-10807.
70. Pluskal, T., et al., *MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.* BMC Bioinformatics, 2010. **11**: p. 395.
71. Smith, C.A., et al., *XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification.* Anal. Chem., 2006. **78**(3): p. 779-787.
72. Myers, O.D., et al., *One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks.* Analytical Chemistry, 2017. **89**(17): p. 8696-8703.
73. Li, Z., et al., *Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection.* Anal Chim Acta, 2018. **1029**: p. 50-57.

74. Baran, R., *Untargeted metabolomics suffers from incomplete raw data processing.* Metabolomics, 2017. **13**(9): p. 107.
75. Manier, S.K., A. Keller, and M.R. Meyer, *Automated optimization of XCMS parameters for improved peak picking of liquid chromatography-mass spectrometry data using the coefficient of variation and parameter sweeping for untargeted metabolomics.* Drug Test Anal, 2019. **11**(6): p. 752-761.
76. Eliasson, M., et al.
77. Zheng, H., et al., *Time-Saving Design of Experiment Protocol for Optimization of LC-MS Data Processing in Metabolomic Approaches.* Analytical Chemistry, 2013. **85**(15): p. 7109-7116.
78. Libiseller, G., et al., *IPO: a tool for automated optimization of XCMS parameters.* BMC Bioinformatics, 2015. **16**: p. 118.
79. Makarov, A., et al.
80. Gumustas, M., et al., *UPLC versus HPLC on Drug Analysis: Advantageous, Applications and Their Validation Parameters.* Chromatographia, 2013. **76**(21): p. 1365-1427.
81. Mahieu, N.G. and G.J. Patti, *Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites.* Analytical Chemistry, 2017. **89**(19): p. 10397-10406.
82. Mahieu, N.G., et al., *Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The Mz.unity Algorithm.* Analytical Chemistry, 2016. **88**(18): p. 9037-9046.
83. Domingo-Almenara, X., et al., *Autonomous METLIN-Guided In-source Fragment Annotation for Untargeted Metabolomics.* Analytical Chemistry, 2019. **91**(5): p. 3246-3253.
84. Domingo-Almenara, X., et al., *Annotation: A Computational Solution for Streamlining Metabolomics Analysis.* Analytical Chemistry, 2018. **90**(1): p. 480-489.
85. Stanstrup, J. *XCMS Workshop*. 2017.
86. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2013: Springer Science & Business Media.
87. Tibshirani, R., G. Walther, and T. Hastie, *Estimating the number of clusters in a data set via the gap statistic.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2001. **63**(2): p. 411-423.
88. Johnson, W.M., M.C. Kido Soule, and E.B. Kujawinski, *Evidence for quorum sensing and differential metabolite production by a marine bacterium in response to DMSP.* The ISME Journal, 2016. **10**(9): p. 2304-2316.
89. Kale, N.S., et al., *MetaboLights: An Open-Access Database Repository for Metabolomics Data.* Current Protocols in Bioinformatics, 2016. **53**(1): p. 14.13.1-14.13.18.
90. Casero, D., et al., *Space-type radiation induces multimodal responses in the mouse gut microbiome and metabolome.* Microbiome, 2017. **5**(1): p. 105.
91. Chambers, M.C., et al., *A cross-platform toolkit for mass spectrometry and proteomics.* Nat. Biotechnol., 2012. **30**(10): p. 918-920.
92. Tautenhahn, R., C. Böttcher, and S. Neumann, *Highly sensitive feature detection for high resolution LC/MS.* BMC Bioinformatics, 2008. **9**: p. 504.

93. Kuhl, C., et al., *CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets.* Analytical Chemistry, 2012. **84**(1): p. 283-289.
94. Broadhurst, D., et al., *Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies.* Metabolomics, 2018. **14**(6): p. 72.
95. Tautenhahn, R., et al.
96. Peters, K., et al., *PhenoMeNal: processing and analysis of metabolomics data in the cloud.* Gigascience, 2019. **8**(2).
97. Davidson, R.L., et al., *Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data.* GigaScience, 2016. **5**(1).
98. Giacomoni, F., et al., *Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics.* Bioinformatics, 2015. **31**(9): p. 1493-5.
99. Yang, A., M. Troup, and J.W.K. Ho, *Scalability and Validation of Big Data Bioinformatics Software.* Comput Struct Biotechnol J, 2017. **15**: p. 379-386.
100. Ardyna, M. and K.R. Arrigo, *Phytoplankton dynamics in a changing Arctic Ocean.* Nature Climate Change, 2020. **10**(10): p. 892-903.
101. Harke, M.J., et al., *Nutrient-Controlled Niche Differentiation of Western Lake Erie Cyanobacterial Populations Revealed via Metatranscriptomic Surveys.* Environmental Science & Technology, 2016. **50**(2): p. 604-615.
102. Follows, M.J., et al., *Emergent Biogeography of Microbial Communities in a Model Ocean.* Science, 2007. **315**(5820): p. 1843.
103. de Vargas, C., et al., *Eukaryotic plankton diversity in the sunlit ocean.* Science, 2015. **348**(6237): p. 1261605.
104. Sharma, S. and R. Steuer, *Modelling microbial communities using biochemical resource allocation analysis.* Journal of The Royal Society Interface, 2019. **16**(160): p. 20190474.
105. Louca, S., et al., *Function and functional redundancy in microbial systems.* Nature Ecology & Evolution, 2018. **2**(6): p. 936-943.
106. Thiele, I. and B.Ø. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nat. Protoc., 2010. **5**(1): p. 93-121.
107. Levitan, O., et al., *Remodeling of intermediate metabolism in the diatom Phaeodactylum tricornutum under nitrogen stress.* Proceedings of the National Academy of Sciences, 2014: p. 201419818.
108. Hennon, G.M.M. and S.T. Dyhrman, *Progress and promise of omics for predicting the impacts of climate change on harmful algal blooms.* Harmful Algae, 2019.
109. Wurch, L.L., et al., *Transcriptional Shifts Highlight the Role of Nutrients in Harmful Brown Tide Dynamics.* Front. Microbiol., 2019. **10**: p. 136.
110. Wurch, L.L., et al., *Proteome Changes Driven by Phosphorus Deficiency and Recovery in the Brown Tide-Forming Alga Aureococcus anophagefferens.* PLOS ONE, 2011. **6**(12): p. e28949.
111. MacIntyre, H.L. and J.J. Cullen, *Using cultures to investigate the physiological ecology of microalgae*. Algal culturing techniques, ed. R.A. Andersen. 2005, Amsterdam: Elsevier Academic Press.

112. Rabinowitz, J.D. and E. Kimball, *Acidic acetonitrile for cellular metabolome extraction from Escherichia coli.* Anal. Chem., 2007. **79**(16): p. 6167-6173.
113. Kido Soule, M.C., et al., *Environmental metabolomics: Analytical strategies.* Mar. Chem., 2015. **177**: p. 374-387.
114. Longnecker, K. and E.B. Kujawinski, *Intracellular Metabolites in Marine Microorganisms during an Experiment Evaluating Microbial Mortality.* Metabolites, 2020. **10**(3).
115. Johnson, W.M., M.C. Kido Soule, and E.B. Kujawinski, *Extraction efficiency and quantification of dissolved metabolites in targeted marine metabolomics.* Limnol. Oceanogr. Methods, 2017. **15**(4): p. 417-428.
116. McLean, C. and E.B. Kujawinski, *AutoTuner: High Fidelity and Robust Parameter Selection for Metabolomics Data Processing.* Analytical Chemistry, 2020. **92**(8): p. 5724-5732.
117. Dunn, W.B., et al., *Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry.* Nat. Protoc., 2011. **6**(7): p. 1060-1083.
118. Melamud, E., L. Vastag, and J.D. Rabinowitz, *Metabolomic Analysis and Visualization Engine for LC−MS Data.* Analytical Chemistry, 2010. **82**(23): p. 9818-9826.
119. Hessa, T., et al., *Recognition of transmembrane helices by the endoplasmic reticulum translocon.* Nature, 2005. **433**(7024): p. 377-381.
120. Wimley, W.C. and S.H. White, *Experimentally determined hydrophobicity scale for proteins at membrane interfaces.* Nature Structural Biology, 1996. **3**(10): p. 842-848.
121. Kanehisa, M., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res., 2000. **28**(1): p. 27-30.
122. Li, S., et al., *Predicting Network Activity from High Throughput Metabolomics.* PLoS computational biology, 2013. **9**: p. e1003123.
123. Ruttkies, C., et al., *MetFrag relaunched: incorporating strategies beyond in silico fragmentation.* J. Cheminform., 2016. **8**: p. 3.
124. Wu, Z., et al., *Empirical bayes analysis of sequencing-based transcriptional profiling without replicates.* BMC Bioinformatics, 2010. **11**: p. 564.
125. Voet, D. and J.G. Voet, *Biochemistry, 4th Edition*. 2010.
126. Hue, L. and H. Taegtmeyer, *The Randle cycle revisited: a new head for an old hat.* Am. J. Physiol. Endocrinol. Metab., 2009. **297**(3): p. E578-91.
127. Gupta, R. and S. Laxman, *Steady-state and Flux-based Trehalose Estimation as an Indicator of Carbon Flow from Gluconeogenesis or Glycolysis.* Bio Protoc, 2020. **10**(1): p. e3483.
128. Villanova, V., et al., *Investigating mixotrophic metabolism in the model diatom Phaeodactylum tricornutum.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2017. **372**(1728): p. 20160404.
129. Ahn, S., et al., *Role of Glyoxylate Shunt in Oxidative Stress Response.* J. Biol. Chem., 2016. **291**(22): p. 11928-11938.
130. Orrenius, S., V. Gogvadze, and B. Zhivotovsky, *Mitochondrial oxidative stress: implications for cell death.* Annu Rev Pharmacol Toxicol, 2007. **47**: p. 143-83.
131. Dringen, R., J.M. Gutterer, and J. Hirrlinger, *Glutathione metabolism in brain.* European Journal of Biochemistry, 2000. **267**(16): p. 4912-4916.

132. Yuasa, K., et al., *Nutrient deficiency stimulates the production of superoxide in the noxious red-tide-forming raphidophyte Chattonella antiqua.* Harmful Algae, 2020. **99**: p. 101938.
133. Staels, B. and V.A. Fonseca, *Bile acids and metabolic regulation: mechanisms and clinical responses to bile acid sequestration.* Diabetes care, 2009. **32 Suppl 2**(Suppl 2): p. S237-S245.
134. Athenstaedt, K. and G. Daum, *Tgl4p and Tgl5p, two triacylglycerol lipases of the yeast Saccharomyces cerevisiae are localized to lipid particles.* J Biol Chem, 2005. **280**(45): p. 37301-9.
135. Houten, S.M. and R.J.A. Wanders, *A general introduction to the biochemistry of mitochondrial fatty acid β-oxidation.* Journal of inherited metabolic disease, 2010. **33**(5): p. 469-477.
136. Kong, F., et al., *Lipid catabolism in microalgae.* New Phytol., 2018. **218**(4): p. 1340-1348.
137. Fuentes-Grünewald, C., et al., *Improvement of lipid production in the marine strains Alexandrium minutum and Heterosigma akashiwo by utilizing abiotic parameters.* J. Ind. Microbiol. Biotechnol., 2012. **39**(1): p. 207-216.
138. Takahashi, K. and T. Ikagawa, *Effect of nitrogen starvation on photosynthetic carbon metabolism in Heterosigma akashiwo (Raphidophyceae).* Jpn. J. Phycol., 1988. **36**: p. 212-220.
139. Bender, S., et al., *Transcriptional responses of three model diatoms to nitrate limitation of growth.* Frontiers in Marine Science, 2014. **1**(3).
140. Kim, J., et al., *Effect of cell cycle arrest on intermediate metabolism in the marine diatom.* Proc. Natl. Acad. Sci. U. S. A., 2017. **114**(38): p. E8007-E8016.
141. Bromke, M.A., et al., *Metabolomic Profiling of 13 Diatom Cultures and Their Adaptation to Nitrate-Limited Growth Conditions.* PLOS ONE, 2015. **10**(10): p. e0138965.
142. Ji, N., et al., *Utilization of various forms of nitrogen and expression regulation of transporters in the harmful alga Heterosigma akashiwo (Raphidophyceae).* Harmful Algae, 2020. **92**: p. 101770.
143. Wada, M., A. Miyazaki, and T. Fujii, *On the Mechanisms of Diurnal Vertical Migration Behavior of Heterosigma akashiwo (Raphidophyceae).* Plant and Cell Physiology, 1985. **26**(3): p. 431-436.
144. Wada, M., et al., *Diurnal appearance, fine structure, and chemical composition of fatty particles inHeterosigma akashiwo (Raphidophyceae).* Protoplasma, 1987. **137**(2): p. 134-139.
145. Hatano, S., Y. Hara, and M. Takahashi, *Photoperiod and nutrients on the vertical migratory behavior of a red tide flagellate, Heterosigma akashiwo.* Journal of Japanese Phycology, 1983. **31**: p. 263-269.
146. Elbein, A.D., et al., *New insights on trehalose: a multifunctional molecule.* Glycobiology, 2003. **13**(4): p. 17r-27r.
147. Zhang, H., et al., *Functional Differences in the Blooming Phytoplankton Heterosigma akashiwo and Prorocentrum donghaiense Revealed by Comparative Metaproteomics.* Applied and Environmental Microbiology, 2019. **85**(19): p. e01425-19.

148. Boer, V.M., et al., *Growth-limiting intracellular metabolites in yeast growing under diverse nutrient limitations.* Mol. Biol. Cell, 2010. **21**(1): p. 198-211.
149. Gorres, K.L. and R.T. Raines, *Prolyl 4-hydroxylase.* Critical reviews in biochemistry and molecular biology, 2010. **45**(2): p. 106-124.
150. Schmollinger, S., et al., *Nitrogen-Sparing Mechanisms in Chlamydomonas Affect the Transcriptome, the Proteome, and Photosynthetic Metabolism.* Plant Cell, 2014. **26**(4): p. 1410-1435.
151. Xu, Y.-F., et al., *Nucleotide degradation and ribose salvage in yeast.* Molecular systems biology, 2013. **9**: p. 665-665.
152. Lynch, J.H. and N. Dudareva, *Aromatic Amino Acids: A Complex Network Ripe for Future Exploration.* Trends in Plant Science, 2020. **25**(7): p. 670-681.
153. Szul, M.J., et al., *Carbon Fate and Flux in Prochlorococcus under Nitrogen Limitation.* mSystems, 2019. **4**(1).
154. Tang, W.H.W., et al., *Diminished global arginine bioavailability and increased arginine catabolism as metabolic profile of increased cardiovascular risk.* Journal of the American College of Cardiology, 2009. **53**(22): p. 2061-2067.
155. Shi, X., et al., *Transcriptomic and microRNAomic profiling reveals multi-faceted mechanisms to cope with phosphate stress in a dinoflagellate.* Isme j, 2017. **11**(10): p. 2209-2218.
156. Coleman, M., *Diagnosing nutritional stress in the oceans.* Science, 2021. **372**(6539): p. 239.
157. Liang, Y., et al., *Molecular mechanisms of temperature acclimation and adaptation in marine diatoms.* The ISME Journal, 2019. **13**(10): p. 2415-2425.
158. Di Dato, V., et al., *Variation in prostaglandin metabolism during growth of the diatom Thalassiosira rotula.* Scientific Reports, 2020. **10**(1): p. 5374.
159. Ustick, L.J., et al., *Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation.* Science, 2021. **372**(6539): p. 287.
160. Honjo, T., *Overview on bloom dynamics and physiological ecology of Heterosigma akashiwo*, in *Toxic Phytoplankton Blooms in the Sea.*, T.J. Smayda and Y. Shimizu, Editors. 1993, Elsevier: Amsterdam. p. 33-41.
161. Chawla, S., et al., *Evaluation of Matrix Effects in Multiresidue Analysis of Pesticide Residues in Vegetables and Spices by LC-MS/MS.* Journal of AOAC INTERNATIONAL, 2017. **100**(3): p. 616-623.
162. Sumner, L.W., et al., *Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI).* Metabolomics, 2007. **3**(3): p. 211.
163. Lee, D.D. and H.S. Seung, *Learning the parts of objects by non-negative matrix factorization.* Nature, 1999. **401**(6755): p. 788-791.
164. Brunet, J.-P., et al., *Metagenes and molecular pattern discovery using matrix factorization.* Proceedings of the National Academy of Sciences, 2004. **101**(12): p. 4164.
165. Wang, M., et al., *Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking.* Nature Biotechnology, 2016. **34**(8): p. 828-837.

166. Burgess, K.E.V., et al., *MetaNetter 2: A Cytoscape plugin for ab initio network analysis and metabolite feature classification.* J Chromatogr B Analyt Technol Biomed Life Sci, 2017. **1071**: p. 68-74.
167. Longnecker, K. and E.B. Kujawinski, *Using network analysis to discern compositional patterns in ultrahigh-resolution mass spectrometry data of dissolved organic matter.* Rapid Communications in Mass Spectrometry, 2016. **30**(22): p. 2388-2394.
168. Guijas, C., et al., *METLIN: A Technology Platform for Identifying Knowns and Unknowns.* Analytical Chemistry, 2018. **90**(5): p. 3156-3164.
169. Segata, N., et al., *Metagenomic biomarker discovery and explanation.* Genome biology, 2011. **12**(6): p. R60-R60.
170. Wei, Z., et al., *Metabolomics coupled with pathway analysis characterizes metabolic changes in response to BDE-3 induced reproductive toxicity in mice.* Scientific Reports, 2018. **8**(1): p. 5423.
171. Longnecker, K., M.C. Kido Soule, and E.B. Kujawinski, *Dissolved organic matter produced by Thalassiosira pseudonana.* Marine Chemistry, 2015. **168**: p. 114-123.
172. Marcellin-Gros, R., G. Piganeau, and D. Stien, *Metabolomic Insights into Marine Phytoplankton Diversity.* Marine Drugs, 2020. **18**(2).
173. Heal, K.R., et al., *Marine community metabolomes carry fingerprints of phytoplankton community composition.* bioRxiv, 2020: p. 2020.12.22.424086.
174. Pirhaji, L., et al., *Revealing disease-associated pathways by network integration of untargeted metabolomics.* Nature methods, 2016. **13**(9): p. 770-776.
175. Barupal, D.K. and O. Fiehn, *Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets.* Scientific Reports, 2017. **7**(1): p. 14567.
176. Gallo, C., et al., *Autoinhibitory sterol sulfates mediate programmed cell death in a bloom-forming marine diatom.* Nature Communications, 2017. **8**(1): p. 1292.
177. Torres-Águila, N.P., et al., *Diatom bloom-derived biotoxins cause aberrant development and gene expression in the appendicularian chordate Oikopleura dioica.* Communications Biology, 2018. **1**(1): p. 121.
178. Athanasakoglou, A. and S.C. Kampranis, *Diatom isoprenoids: Advances and biotechnological potential.* Biotechnology Advances, 2019. **37**(8): p. 107417.
179. Götz, F., et al., *Targeted metabolomics reveals proline as a major osmolyte in the chemolithoautotroph Sulfurimonas denitrificans.* MicrobiologyOpen, 2018. **7**(4): p. e00586.
180. CL, F., et al., *A phosphate starvation response gene (psr 1-like) is present and expressed in Micromonas pusilla and other marine algae.* 2021. **86**: p. 29-46.
181. Durham, B.P., et al., *Sulfonate-based networks between eukaryotic phytoplankton and heterotrophic bacteria in the surface ocean.* Nature Microbiology, 2019. **4**(10): p. 1706-1715.
182. Taylor, P.J., *Matrix effects: the Achilles heel of quantitative high-performance liquid chromatography–electrospray–tandem mass spectrometry.* Clinical Biochemistry, 2005. **38**(4): p. 328-334.
183. Flynn, K.J., *The determination of nitrogen status in microalgae.* Marine Ecology Progress Series, 1990. **61**: p. 297-307.

184. Olenina, I., et al., *Biovolumes and size-classes of phytoplankton in the Baltic Sea*, in *Baltic Sea Environment Proceedings*. 2006. p. 1-144.
185. Brosnan, J.T. and M.E. Brosnan, *Branched-chain amino acids: enzyme and substrate regulation.* J Nutr, 2006. **136**(1 Suppl): p. 207s-11s.
186. Muthukrishnan, S., et al., *7 - Chitin Metabolism in Insects*, in *Insect Molecular Biology and Biochemistry*, L.I. Gilbert, Editor. 2012, Academic Press: San Diego. p. 193-235.
187. El Gueddari, N.E., et al., *Developmentally regulated conversion of surface-exposed chitin to chitosan in cell walls of plant pathogenic fungi.* New Phytologist, 2002. **156**(1): p. 103-112.
188. Benner, R. and K. Kaiser, *Abundance of amino sugars and peptidoglycan in marine particulate and dissolved organic matter.* Limnology and Oceanography, 2003. **48**(1): p. 118-128.
189. Ferrer-González, F.X., et al., *Resource partitioning of phytoplankton metabolites that support bacterial heterotrophy.* The ISME Journal, 2021. **15**(3): p. 762-773.
190. Pinu, F.R., et al., *Metabolite secretion in microorganisms: the theory of metabolic overflow put to the test.* Metabolomics, 2018. **14**(4): p. 43.
191. Bhattarai, Y., et al., *Gut Microbiota-Produced Tryptamine Activates an Epithelial G-Protein-Coupled Receptor to Increase Colonic Secretion.* Cell Host Microbe, 2018. **23**(6): p. 775-785.e5.
192. Jang, S.-W., et al., *N-acetylserotonin activates TrkB receptor in a circadian rhythm.* Proceedings of the National Academy of Sciences, 2010. **107**(8): p. 3876.
193. Lu, X., et al., *Metatranscriptomic identification of polyamine-transforming bacterioplankton in the Gulf of Mexico.* Environ Microbiol Rep, 2020. **12**(3): p. 258-266.
194. Landa, M., et al., *Bacterial transcriptome remodeling during sequential co-culture with a marine dinoflagellate and diatom.* The ISME Journal, 2017. **11**(12): p. 2677-2690.
195. Lonsdale, D., *A review of the biochemistry, metabolism and clinical benefits of thiamin(e) and its derivatives.* Evidence-based complementary and alternative medicine : eCAM, 2006. **3**(1): p. 49-59.
196. Hockin, N.L., et al., *The response of diatom central carbon metabolism to nitrogen starvation is different from that of green algae and higher plants.* Plant Physiol., 2012. **158**(1): p. 299-312.
197. Hackett, S.R., et al., *Systems-level analysis of mechanisms regulating yeast metabolic flux.* Science, 2016. **354**(6311): p. aaf2786.
198. Yano, J.M., et al., *Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis.* Cell, 2015. **161**(2): p. 264-76.
199. Patejko, M., et al., *Urinary Nucleosides and Deoxynucleosides.* Adv Clin Chem, 2018. **83**: p. 1-51.
200. Misra, B.B., *New software tools, databases, and resources in metabolomics: updates from 2020.* Metabolomics, 2021. **17**(5): p. 49.
201. Hennon, G.M.M. and S.T. Dyhrman, *Progress and promise of omics for predicting the impacts of climate change on harmful algal blooms.* Harmful Algae, 2020. **91**: p. 101587.
202. Tobin, E.D., et al., *Behavioral and physiological changes during benthic-pelagic transition in the harmful alga, Heterosigma akashiwo: potential for rapid bloom formation.* PLoS One, 2013. **8**(10): p. e76663.

203. Gärdes, A., et al., *Diatom-associated bacteria are required for aggregation of Thalassiosira weissflogii.* The ISME Journal, 2011. **5**(3): p. 436-445.
204. Li, Y. and T.J. Smayda, *Heterosigma akashiwo (Raphidophyceae): On prediction of the week of bloom initiation and maximum during the initial pulse of its bimodal bloom cycle in Narragansett Bay.* Plankton Biol. Ecol., 2000. **47**(2): p. 80-84.
205. Di Costanzo, F., et al., *Prostaglandins in Marine Organisms: A Review.* Marine Drugs, 2019. **17**(7).
206. Di Dato, V., et al., *Animal-like prostaglandins in marine microalgae.* The ISME Journal, 2017. **11**(7): p. 1722-1726.
207. Di Dato, V., A. Ianora, and G. Romano, *Identification of Prostaglandin Pathway in Dinoflagellates by Transcriptome Data Mining.* Marine Drugs, 2020. **18**(2).