USING METAPROTEOMES AND MODELS TO QUANTIFY CELLULAR TRADE-OFFS IN PHYTOPLANKTON

by

J. Scott. P. McCain

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
December 2021

To my loving family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Phytoplankton fuel biogeochemical processes in the ocean and are key players influencing the global climate. All biogeochemical processes mediated by microbes are ultimately underpinned by gene expression. In this thesis, I aim to connect gene expression to biogeochemically important cellular processes in Southern Ocean phytoplankton. First, I identify and model pervasive biases in metaproteomic analyses and develop methods for overcoming them, leading to more robust inferences (Chapter 2 and Chapter 3). In Chapter 2, I develop a computational model for predicting cofragmentation bias and use this model to study how cofragmentation impacts inferences in metaproteomics. In Chapter 3, I delineate 'proteomic traits' across microbial taxa in an Antarctic phytoplankton bloom, connecting differences in gene expression patterns to ecological strategies. I also highlight the importance of database choice and quantify its implications for metaproteomic conclusions. In Chapter 4, I develop a proteomic allocation model to quantify trade-offs associated with iron and manganese bioavailability, and reframe micronutrient-controlled growth in the ocean as a function of cellular costs and constraints. This model offers a novel framework for leveraging metaproteomic data to learn about cellular processes in phytoplankton and for inferring taxon-specific rates and biogeochemical metrics. A key unknown in this model is the various antioxidant systems used by phytoplankton. I, therefore, review various antioxidant mechanisms and synthesize their contributions to cellular elemental stoichiometry in phytoplankton (Chapter 5). Finally, in Chapter 6, I use metaproteomics to determine environmental controls on ribosomal mass fraction across two taxonomic groups in the Amundsen Sea Polynya.

ACKNOWLEDGEMENTS

It has truly been a privilege to be a PhD student. I feel incredibly lucky to have been a part of the Dalhousie Biology Department, the Institute for Comparative Genomics (ICG), and the Bertrand Lab.

I am extremely thankful to Erin Bertrand: for your mentorship over the past five years, for your thoughtful guidance both in science and in life, and for making the lab such a supportive place to work. I'm particularly thankful for the diverse experiences you encouraged me to embrace throughout my degree: a cruise in Antarctica, a foray into building bioinformatics tools, working with an instrument as expensive as a house. Your early lesson 'if you're not feeling stupid, you're not challenging yourself', has and will stick with me. I would also like to thank my committee members: Rob Beiko and Julie LaRoche, for your thoughtful questions, ideas, and encouragement. Chris Algar, thank you for pushing me to know my fundamentals and challenging every assumption. I will always remember our modelling and beer meetings! To the TOSST Program Directors, Markus Kienast and Doug Wallace, your program completely changed my research direction and had an enormous impact on my experience. Thank you for tolerating my (sometimes) rebellious nature and for the wonderful chats about biogeochemistry.

The Biology Department at Dalhousie has been a special place to work, and I'm very thankful to the people I've interacted with throughout both my MSc and PhD. Thank you to Heike Lotze for introducing me to science and how it can intersect with art. Thank you to Chris Taggart for teaching me to tackle problems critically. Thank you to Sophia Stone, for entertaining my random questions about plant immune systems and for always being a force for graduate students. Thank you to Alastair Simpson for so many laughs (e.g. British raisin cake), the trivia-that-would-never-come, and for constant education about protists. Thank you to Joe Bielawski for the philosophical chats. Thank you to Aileen, Julie, Carolyn and Chris for your help throughout my time at Dalhousie, specifically in navigating the labyrinth of administration.

ICG is a shining gem in Dalhousie, and I will sincerely miss the community of smart and genuinely nice people. Thank you to Ed Susko for your patient advice-turned-collaboration. Thank you to Andrew Roger and John Archibald for journal club discussions.

Thank you to Ford Doolittle for provocative discussions about Gaia, and for being a model of clear thinking.

I've also been fortunate to experience different institutions throughout my degree. Thank you to Andrew Cogswell at the Bedford Institute of Oceanography for R-chats and insights into government science; thank you to Alex Cohen at the Mass Spectrometry Core Facility for the many wonderful conversations and lessons about mass spectrometry. I'm particularly grateful to Alessandro Tagliabue for hosting me in his laboratory for four months at University of Liverpool: I learned so much in that short period, from iron chemistry to modelling. Thank you to Eric Achterberg at GEOMAR in Germany for pushing my writing in particular. Thank you to Rob Middag for an unforgetable experience in the Amundsen Sea, as well as the staff and crew aboard the RV Araon. Thank you to the staff and crew aboard the RV Hudson for my first cruise experience.

I am forever thankful to my fellow graduate students at Dalhousie. Bertrand Lab members, you were all a dream to work with. I'm so thankful to be both friends and colleagues with all of you. Elden Rowland, thank you for your mass spectrometry wizardry, hilarious musings, and calm and playful approach to science. Loay Jabre, thank you for journeying through the manganese and iron and Southern Ocean world together – someday let's tackle aluminium? Cat Bannon and Megan Roberts, thank you for always lighting up the room and for your incredible thoughtfulness. Catalina, Insa, Lena, Miao, Maria, Kira, Gianpaolo, Scott – you've all been so wonderful to work with! Yana – you were always around for a late night beer and bagel in the LSC. I already miss seeing the crazy critters you've been discovering. Jockel – thanks for being an iron buddy and a gym buddy! Taylor – debugging code and venturing across Newfoundland, so many memories that will always remind me of the last year in graduate school! Lisette, Cait, Andrea, Joce, Brent, Angela, Gordon, Daryl, Georgia, Izzy, Liam, Benia, and many more, graduate school would not have been the same without you. To all the TOSSTies, I learned so much alongside you.

To my family, thank you for all the support over the years. Mom, thank you for your *beaming* support since my very first time in a lab (and for learning all these terms as well, starting with 'optical signatures'!). Dad, you helped me through so much of graduate school with your unwavering confidence in my abilities, I'm so grateful. Jonathan, Lauren, Hilary, Hannah, and the rest of my family: thank you for always lending an ear and for providing sage advice on all matters. To the Youssef and Aly families, thank you to the many games of Catan and trips to Golden Bakery to put things into perspective.

Noor Youssef, thank you for your curious, thoughtful, and critical mind. You are truly a light. Thank you for venturing with me through the cloud that is both science and life.

LIST OF ABBREVIATIONS USED

2D-LC	Two-Dimensional Liquid Chromatography
ABC	Approximate Bayesian Computation *or* ATP Binding Cassette
AsA	Ascorbate
ATP	Adenosine Triphosphate
CAT	Catalase
CCP	Cytochrome c Peroxidase
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
DMSP	Dimethylsulfoniopropionate
DTT	Dithiothreitol
EC	Enzyme Commission
EDTA	Ethylenediaminetetraacetic Acid
FeL	Ligand Bound Iron
GLM	Generalized Linear Model
GO	Gene Ontology
GPR	Gaussian Process Regression
GPx	Glutathione Peroxidase
GSH	Reduced Glutathione
IAM	Iodoacetamide
ICP-MS	Inductively Coupled Plasma Mass Spectrometry
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
KOG	EuKaryotic Orthologous Groups
KL	Kullback Leibler
LC	Liquid Chromatography
m/z	Mass-to-Charge Ratio
MAG	Metagenome Assembled Genome
MCL	Markov Clustering
MetE	B12-Independent Methionine Synthase

MS1	Precursor Ion Mass Spectrometry Scan
MS2	Fragment Ion Mass Spectrometry Scan
MS	Mass Spectrometer
NADPH	Nicotinamide Adenine Dinucleotide Phosphate
NRAMP	Natural Resistance-Associated Macrophage Protein
ODE	Ordinary Differential Equation
ORF	Open Reading Frame
PAR	Photosynthetically Active Radiation
Prx	Peroxiredoxin
PSI	Photosystem I
PSII	Photosystem II
PSM	Peptide-Spectrum Match
PSU	Photosystem Unit
RBF	Radial Basis Function
RMF	Ribosomal Mass Fraction
ROS	Reactive Oxygen Species
SDS	Sodium Dodecyl Sulfate
SLSQP	Sequential Least Squares Quadratic Programming
SOD	Superoxide Dismutase
SVM	Support Vector Machine
TIC	Total Ion Current

CHAPTER 1

INTRODUCTION

Microbes are critical players in the transformation of elements in the ocean, forming the heart of marine biogeochemical cycles. Many biogeochemical processes are fundamentally powered by gene expression: nitrogen fixation is mediated by the enzyme nitrogenase; iron (Fe) uptake is a function of the production of Fe transporter proteins; and carbon fixation occurs via the protein RuBisCO. Yet, this connection (from gene expression to microbial behaviours) has rarely been represented in models of biogeochemical cycles. This is due to both technological challenges and a historical focus on the emergent outcomes of gene expression that are biogeochemically relevant.

For an oceanographer attempting to model the fluxes and transformations of elements in the ocean, perhaps it is sufficient to only focus on outcomes of gene expression (e.g. elemental stoichiometric ratios, carbon fixation rates, resource-growth relationships, etc.). These emergent outcomes are highly important, for example the ratio of nitrogen to phosphorus in phytoplankton 'protoplasm' controls the eventual supply of these elements to the surface ocean, therefore influencing global primary production (Redfield, 1958). Many biogeochemical models predict these quantities using phenomenological equations with inputs of environmental variables like light, temperature, and nutrient concentrations (not so different from the equations in Riley, 1946). In doing so, they bypass the connection between cellular behaviours and gene expression. There is nothing inherently wrong with using these phenomenological relationships, but are these 'conventional' predictions (i.e. those ignoring gene-level processes) 'good enough'?

Some biogeochemical models have sufficient predictive abilities for their intended uses, while others are failing. I will only briefly highlight the successes of biogeochemical

modelling (but in general it is remarkable that a complex system can be simplified in this way). The first global biogeochemical model, developed by Bacastow and Maier-Reimer (1991), was able to predict the broad distributions of phosphate and oxygen in different regions. Just over a decade later, Aumont et al. (2003) predicted general patterns in primary productivity across ocean basins using a global biogeochemical model. Using an ecological approach, Follows et al. (2007) showed that a simple trait-based model can recapitulate biogeographic trends in phytoplankton functional types. Some biogeochemical models built from thermodynamic principles can make accurate predictions, particularly in marine sediments (Vallino, 2010; Vallino and Algar, 2016; Vallino and Huber, 2018; Algar and Vallino, 2014; Hardison et al., 2015), and with few free parameters. Anecdotally, predictions that are more reliant on geochemistry tend to be better than those more reliant on biological parameterizations. For example, Xue et al. (2016) modelled pCO_2 in the Gulf of Mexico and found good correspondence between observations and predictions. However, this is partly due to constraints; there are fewer biological observations to constrain models compared to chemical and physical observations (Fennel et al., 2019).

There are many challenges with modern biogeochemical models. For example, even simple interactions between nutrients (e.g. colimitation) are rarely included in many biogeochemical models. Interactions are typically represented using Liebig's Law of the Minimum, which predicts abrupt changes in resource limitation, i.e. 'tipping-point' behaviour (Equation 1.1). In this case, growth-resource functions are the minimum of a set of Michaelis-Menten dependencies, for example:

$$\mu = min(\frac{R_1}{R_1 + K_1}, \frac{R_2}{R_2 + K_2}) \tag{1.1}$$

where μ is the growth rate, R_1 and R_2 represent different resources, and K_1 and K_2 are different resource-specific half saturation constants. However, tipping points have not been realized in empirical observations (e.g. Hillebrand et al., 2020), and there is evidence from gene expression data that microbes do not experience stress from only one nutrient at a time (Saito et al., 2014). For certain biogeochemical cycles, for example the Fe cycle, models have extremely divergent predictions (Tagliabue et al., 2016). Tagliabue et al. (2020) showed that parameterizations of the biology can have dramatic, cascading effects on predictions. In a striking example for the nitrogen cycle, Wrightson and Tagliabue (2020) showed that there is large variation in the magnitude and direction of nitrogen

fixation predictions in nine earth system models. In general, much of the uncertainty that underpins variation in predictions across models is related to how microbes are represented. Perhaps, as a field, we should be including more mechanistic details of microbial growth to improve these representations. Specifically, we should look further into how gene expression directly impacts biogeochemical cycling, to ultimately build biogeochemical models that are more flexible and have better predictions. Encouragingly, studies that have bridged the gap from genes to biogeochemical processes have offered new ways to flexibly model complex processes (e.g. Reed et al., 2014; Coles et al., 2017; Haas et al., 2021).

From another perspective, there is a wealth of gene expression data from complex marine microbial communities (e.g. Cohen et al., 2021). Yet, these data are rarely used to quantify cellular processes. Connecting gene expression measurements to growth for marine microbes can be used for dual purposes, in 1) creating better biogeochemical models and 2) quantifying cellular processes in environmentally important microbes.

How does gene expression impact cellular-level outcomes like growth rates, elemental stoichiometry, or elemental quotas? Everything an organism does is because of gene expression. Beginning in the 1950s, Schaechter and others identified that the number of ribosomes per cell increases linearly with growth rate (Schaechter, Maaløe and Kjeldgaard, 1958). More recently, Scott et al. (2010) used phenomenological equations to relate ribosomal mass fraction to growth rate, demonstrating the interdependence of growth and gene expression. Others have taken a machine learning approach to predict growth rate from gene expression profiles (Wytock and Motter, 2018). The relationship between elemental quotas and gene expression has received less attention. Consider the relationship between iron quotas and transporter proteins: Fe transporters can be used to increase Fe uptake rate, therefore increasing incorporation of free Fe into proteins, and ultimately increasing cellular Fe quotas. More directly, Saito et al. (2011) estimated micronutrient quotas by measuring protein abundance profiles in *Crocosphaera watsonii*. Two main challenges for including gene expression in models of biogeochemical processes are that 1) we do not always understand *how* gene-level processes directly connect to cellular level processes, and 2) quantitative relationships between gene expression and cellular processes have rarely been established, particularly for non-model organisms.

Gene expression can be broadly defined as 'the appearance in a phenotype of a characteristic attributed to a particular gene' (Oxford Languages). Which method is optimal

for measuring gene expression? This partially depends on what the target phenotype is, which also depends on the question and the system. In humans, it might be a disease state. For microbes, a common target phenotype is the collection of reaction rates occurring in the cell (e.g. rates of uptake, translation, etc.). One might argue that in some cases, even detecting a gene can be used as a proxy for gene expression (for example when said gene is constitutively expressed). However, in most cases it would be beneficial to more directly measure the processes that underpin a given phenotype: quantifying mRNA (transcriptomics), protein (proteomics), or metabolites (metabolomics).

Transcriptomics has several advantages: it is comparatively inexpensive to achieve a very high depth of sequencing, identifying lowly abundant transcripts. However, mRNA transcripts do not directly act on the phenotype (i.e. they do not mediate chemical transformations). Furthermore, mRNA abundance only weakly correlates with protein abundance (on a transcript-to-protein basis; Liu, Beyer and Aebersold, 2016; Bender et al., 2017). At the process-level however, transcript allocation has a high correlation with proteomic allocation (0.87 Pearson correlation coefficient using GO-slim processes; Yu et al., 2020). Variants of transcriptomics, like ribosomal profiling, offer quantification accuracy surpassing that of untargeted proteomics by estimating protein synthesis rates (Li et al., 2014b; Mori et al., 2021). Conceptual advances in transcriptomics are fundamentally those that get *closer* to measuring proteins, so why not measure proteins more directly?

Proteomics can be used to measure the abundance of a collection of proteins, the molecular machines that underpin many phenotypes. For example, many proteins mediate specific chemical reactions. Proteomics therefore has major theoretical advantages for measuring the outcomes of gene expression – it is conceptually close to reaction rates in a cell because protein abundance can influence reaction rates directly. Further, proteins make up a dominant fraction of cellular mass compared to other biomolecules (Liefer et al., 2019), also imposing important constraints on growth because of protein synthesis (Molenaar et al., 2009). Despite these advantages, proteomics typically cannot achieve the characterization depth that transcriptomics offers. (Note that this is partially because transcript abundance has a lower dynamic range than protein abundance; Yu et al. (2020).)

Metabolomics is one step closer, in that the collection of metabolites are essentially the proximate end products of gene expression. However, there are three major disadvantages for metabolomics. 1) Metabolites can have very different chemical properties, making

relative quantification *across* metabolites daunting, and preventing a global picture of metabolite abundance (at least using untargeted metabolomics). 2) Metabolites do not have sequence-specificity, so it is sometimes impossible to attribute a metabolite to a particular gene (unlike proteomics and transcriptomics). This is a major hindrance to applying metabolomics to characterize gene expression in a microbial community. 3) Many end-products of protein-transformations are not possible to measure using metabolomics, for example proteins that transform reactive oxygen species. Overall, proteomics is an ideal approach for characterizing gene expression (and therefore metaproteomics for studying assemblages of microbes). In practice, however, metaproteomics has serious challenges that prevent it from being a widespread tool.

In this thesis, I address current problems in metaproteomics, with the ultimate goal of connecting gene expression of marine microbes to biogeochemically-relevant emergent properties. In particular, I consider how the essential micronutrients Fe and manganese (Mn) influence phytoplankton gene expression. Below, I give some background on metaproteomic methods, the challenges in this field, and trace metal biogeochemistry to contextualize the thesis.

1.1 Metaproteomics Overview: A Method for Observing Gene Expression *In Situ*

Metaproteomics, the identification and quantification of proteins from different species, is a powerful tool for learning about microbial behaviours. The term 'metaproteomics' was first coined in 2004 (Rodríguez-Valera, 2004), and Wilmes and Bond (2004) were the first to characterize a metaproteome of activated sludge. Since then there have been many advances. Here I will first describe a general, modern methodological approach for many metaproteomics experiments, and then I will discuss overarching challenges.

Microbes are collected from their environment and promptly stored at -80 ° C. Proteins present in the microbes are then extracted, and then digested into peptides. This type of proteomics is referred to as 'bottom-up' proteomics, and is done because peptides ionize more easily than whole proteins. This complex mixture of peptides is then 'simplified' prior to mass spectrometry using liquid chromatographic separation, which typically separates peptides based on their hydrophobicity (i.e. with reversed-phase chromatography). Once the peptides leave the chromatographic separation column, they are ionized using a 'soft

ionization' approach, such that the peptide molecules mostly stay intact.

The mass spectrometer (MS) then measures the quantity of a given ion, and data from the MS can be used to infer the likely identity of an ion. Most of the work presented in this thesis uses a discovery-based approach (also known as 'shotgun', or 'untargeted' proteomics), specifically using a data-dependent acquisition strategy (DDA). In DDA experiments, whole ions (i.e. peptides) are selected for fragmentation based on their intensity relative to other ions. During this stage (the MS1 or 'precursor ion' scan) the mass-to-charge (*m/z*) ratio is determined for ions eluting from the column. Once an ion is selected, it is then fragmented and then the *m/z* for subsequent fragments are measured (known as the MS2 scan). This scan of peptide fragments' *m/z* is called the peptide mass spectra, and can be used to identify the amino acid sequence of the intact peptide (described below).

Peptides can also be quantified using several different methods. One common method is called 'spectral counting', which sums all MS2 spectra corresponding with a certain amino acid sequence. Another method, which is mostly used in this thesis, is called 'ion intensity integration'. The intensity of an ion is proportional to the number of molecules of that ion. In this second approach, the intensity of the peak in the MS1 scan is used to infer the abundance of a certain ion. For all quantification methods, the abundance of a peptide is normalized to some metric of total peptide abundance (e.g. the sum over all spectral counts, or the sum of peptide intensities).

Peptide sequences are identified in several ways as well, in this thesis I mostly used 'database searching'. In database searching, the user provides a list of *potential* proteins, in the form of amino acid sequences (gene sequences can also be used). From this, theoretical mass spectra of peptides are generated *in silico*, and this large set of theoretical mass spectra are compared to observed mass spectra. There are a multitude of ways to score and match mass spectra, but at the heart of all these methods is simply a distance metric that searches for the theoretical spectra with the lowest distance. After correcting for a user-defined false discovery rate, a set of peptide-spectrum matches (PSMs) are produced. At this stage, there are many different approaches used in metaproteomics to go from the set of PSMs to biological inference (e.g. MetaGOmics, MetaproteomeAnalyzer; Riffle et al., 2018; Muth and Renard, 2017), which largely depend on the biological question (note that there is contention in metaproteomics over whether biological inferences should

be peptide- or protein-centric).

1.2 Current Challenges in Metaproteomics

Metaproteomics has improved immensely over the last decade: Wilmes and Bond (2004) confidently identified 3 proteins in their activated sludge! As a field, it has had a similar progression like other high-throughput 'omics fields. For example, the Human Genome Project's first goal was to *sequence* a human genome, and only after that has the technology been used to *learn* about human genetics. For 'meta'-omics, there is an additional challenge. Laboratory and computational tools used for meta-omics typically begin as the same, or derived from, those from single-organism 'omics. Metaproteomics is no exception. A lot of challenges the field of metaproteomics faces are because some of the basic assumptions that are obviously true for single-organism proteomics, are probably not true for metaproteomics. These challenges can be categorized into either laboratory-based or bioinformatic. In this thesis I focus on the latter, so I will briefly overview these challenges.

Some assumptions for single-organism proteomics are: observed spectra are found in the database of potential peptides, the sample complexity does not shift significantly across samples, and proteins all come from the same organism. Many of these assumptions have not been rigorously assessed or are obviously broken in metaproteomics. Yet, they are foundational, and could bias all observations in perverse ways. For example, Bergauer et al. (2017) compare protein expression across very different ocean regions, and observe an increase with depth of a specific protein group. But, their database mostly comes from one region. Do they observe an increase in protein expression because protein expression is actually increasing, as the authors suggest? One plausible alternative explanation for their observation is that the database used was increasingly poorly matched, which would artificially inflate their normalization factor. It is clear that simple issues like these must be studied. Throughout this thesis I use previously published data, simulation models, and my own data to quantify and evaluate these types of biases.

1.3 Trace Metal Biogeochemistry

Trace metals limit primary production in large swathes of the ocean, with broad consequences for carbon sequestration and fisheries (Tagliabue et al., 2020). Since the 1990s,

Fe has been recognized as an important limiting nutrient (Martin, Gordon and Fitzwater, 1991). The biogeochemical cycling of Fe in the ocean has several notable characteristics. First, dissolved Fe is present in various chemical species: as inorganic free Fe (Fe'), and bound to ligands (FeL) of varying strength. FeL is typically further classified into various forms depending on the binding strength of the ligand. The chemical speciation of Fe is complex, and also includes various colloidal and particle forms (see Tagliabue et al., 2017). Fe is an essential micronutrient for photosynthetic phytoplankton, used for a variety of cellular processes: photosynthesis, reactive oxygen species metabolism, nitrogen metabolism, etc. (Raven, Evans and Korb, 1999). Various ocean regions have been deemed Fe-limited, and in this thesis I particularly focus on the Southern Ocean. The Southern Ocean, or the ocean wrapping around Antarctica, is also strongly influenced by light availability. Other less-studied micronutrients, like manganese, have more recently come to light as playing an important role in limiting primary production (Browning et al., 2021; Wu et al., 2019; Middag et al., 2013).

Manganese is required for all photosynthesis in the ocean because it is a cofactor in Photosystem II (PSII). Since the 1990s, researchers have explored the impact of Mn on primary production in the Southern Ocean, with varied observations (Martin, Fitzwater and Gordon, 1990; Buma et al., 1991). More recently, there have been clear responses of Mn addition to Southern Ocean water in the form of bottle incubation experiments (Wu et al., 2019; Browning et al., 2021). These experiments point to Mn as a relatively understudied micronutrient. Yet, it does appear to play a secondary role compared to Fe, given the variability in response and from seminal calculations by Raven (1990). Colloquially, one Antarctic researcher referred to Mn bottle incubation experiments saying: 'it's not that we haven't looked for a response, we did and just didn't find anything'. This highlights one of the challenges in comparing Mn and Fe responses in the ocean: bottle incubation experiments that do not show any differences are not always published. So it is challenging to compare the relative extent of Mn versus Fe limitation in the Southern Ocean using bottle incubations from the literature alone.

Some researchers have also hypothesized that Fe and Mn interactively influence growth. Peers and Price (2004) hypothesize that under low Fe, Mn requirements increase. The mechanism of this interaction is further described and studied in Chapter 4. From an oceanographic perspective, given that there is evidence both Mn and Fe influence

phytoplankton growth *in situ*, it is important to predict the combined effect on Southern Ocean productivity. Furthermore, it is completely unknown how Mn might change in the changing ocean and the impacts Mn limitation might have on longer time-scales (e.g. Keith Moore et al., 2018).

1.4 Structure of Thesis

In this thesis I aim to quantify cellular processes that underpin phytoplankton growth and metabolism by combining metaproteomic observations with models. Specifically, I focused on Fe and Mn controls on phytoplankton growth. In Chapters 2 and 3, I tackle several methodological challenges associated with metaproteomics. In Chapter 2, I develop a computational model to predict the effects of cofragmentation in metaproteomics. I then use this model to study the general effects of cofragmentation, which informed later data analysis choices in all subsequent thesis chapters. In Chapter 3, I use metaproteomics to identify and quantify proteomic traits across diverse marine microbes in an Antarctic phytoplankton bloom. In this chapter I also examined how database choice, normalization, and protein group inference can be leveraged for metaproteomics with eukaryotic organisms. In Chapter 4, I developed a proteomic allocation model of a diatom to study the interaction between Fe and Mn, using the metaproteomic observations to inform the model parameters. This novel connection with metaproteomic data also enabled inferences about *in situ*, taxon-specific growth rates (and other biogeochemically important characteristics). In Chapter 5, I reviewed a key unknown identified in studying the interaction between Fe and Mn: antioxidants in phytoplankton. In this chapter, I synthesized how antioxidants influence phytoplankton elemental stoichiometry. In Chapter 6, I identify the environmental controls on the ribosomal mass fraction across *Fragilariopsis* and *Phaeocystis* spp., connecting protein synthesis to growth using metaproteomic observations from the Amundsen Sea Polynya. Finally, I conclude the thesis with an overarching view for using gene expression data to learn about biogeochemical processes.

CHAPTER 2

PREDICTION AND CONSEQUENCES OF COFRAGMENTATION FOR METAPROTEOMICS

This work was published previously in *Journal of Proteome Research* (McCain and Bertrand, 2019).

2.1 Abstract

Metaproteomics can provide critical information about biological systems, but peptides are found within a complex background of other peptides. This complex background can change across samples, in some cases drastically. Cofragmentation, the co-elution of peptides with similar mass to charge ratios, is one factor that influences which peptides are identified in an LC-MS/MS experiment: it is dependent on the nature and complexity of this dynamic background. Metaproteomics applications are particularly susceptible to cofragmentation-induced bias; they have vast protein sequence diversity and the abundance of those proteins can span many orders of magnitude. We have developed a mechanistic model that determines the number of potentially cofragmenting peptides in a given sample (called *cobia*, https://github.com/bertrand-lab/cobia). We then used previously published datasets to validate our model, showing that the resulting peptide-specific score reflects the cofragmentation 'risk' of peptides. Using an Antarctic sea ice edge metatranscriptome case study, we found that more rare taxonomic and functional groups are associated with higher cofragmentation bias. We also demonstrate how cofragmentation scores can be used to guide the selection of protein- or peptide-based biomarkers. We illustrate potential

consequences of cofragmentation for multiple metaproteomic approaches, and suggest practical paths forward to cope with cofragmentation-induced bias.

2.2 Introduction

Metaproteomics is a powerful tool for examining microbial community function *in situ*. However, microbial communities are metabolically and phylogenetically diverse, and contain protein amounts ranging many orders of magnitude (Koziol et al., 2013; Zubarev, 2013). This diversity and dynamic range lead to immense and variable sample complexity, a fundamental challenge of the 'proteomic pipeline' (i.e. from sample to biological inference) and a central issue in each pipeline stage. High sample complexity influences sample preparation, liquid chromatography (LC), mass spectrometry (MS), and bioinformatic analyses (Schneider and Riedel, 2010; Muth, Renard and Martens, 2016; Heyer et al., 2017; Schiebenhoefer et al., 2019). Determining how different aspects of metaproteomic characterization impact our biological conclusions is important for interpreting metaproteomic results, and more broadly for understanding microbial community metabolism.

One of the most significant challenges associated with high sample complexity is when multiple peptides of similar mass to charge ratio (*m/z*) elute from the chromatographic separation and are introduced into the mass spectrometer simultaneously. Co-eluting and similar *m/z* peptides interfere with charge state assignment, and thus may not be selected for fragmentation during data dependent acquisition experiments. If selection and fragmentation do occur, these cofragmenting peptides generate mass spectra that are typically of low quality, leading to decreased probability of detection (Houel et al., 2010). (Herein, we describe peptides of similar *m/z* and elution times as 'cofragmenting'.) In addition, a significant portion of peptides are susceptible to cofragmentation because of their low ion intensity (Michalski, Cox and Mann, 2011). Decreased probability of detection may also lead to inaccurate peptide quantification, either via fewer peptide spectrum matches (i.e. for spectral counting), or as interference with feature detection (i.e. for ion intensity integration). Bioinformatic methods have been developed to deconvolute mixtures of mass spectra (Wang, Bourne and Bandeira, 2011; Zhang et al., 2014; Wang, Bourne and Bandeira, 2014; Dorfer et al., 2018), to estimate the extent chimaeric spectra influence proteomics (Houel et al., 2010), and to predict unique peptide transitions for targeted proteomics (Röst, Malström and Aebersold, 2012). Yet, we are not aware of any

computational approaches that predict cofragmentation bias for a given peptide or protein.

While it is clear that cofragmentation has consequences for peptide identification in metaproteomics, predicting cofragmentation-induced bias is inherently challenging. This is because there are many processes that lead to successfully identifying a peptide, and the effects of these processes are difficult to differentiate. For example, it is difficult to distinguish between a peptide that is not observed because cofragmentation reduces MS2 spectral quality, a peptide that is subject to ion suppression during ionization from another co-eluting peptide, or a peptide that is simply not in the sample.

Our main objective is to determine how cofragmentation biases metaproteomics, and to develop a score that reflects peptide-specific cofragmentation risk. Thus, we (1) developed a mechanistic model to predict cofragmentation, (2) examined whether cofragmentation risk was an important driver of peptide detection, (3) assessed potential consequences of cofragmentation at both coarse and fine-scale taxonomic and functional levels, and (4) suggest practical steps forward to address this issue in metaproteomics. To do so, we developed a computational tool to simulate each of the steps of the proteomic pipeline – from sample preparation to LC-MS/MS – to predict peptides that are most at-risk of cofragmentation. The input of our simulation is the 'potential' metaproteome (i.e. a metatranscriptome or metagenome). Thus, we hypothesize that cofragmentation can be approximated by simply counting the number of potentially cofragmenting ions. We validate this approach by examining how predicted cofragmentation is associated with the probability of detection in five datasets from three studies, of coupled metagenomic, metatranscriptomic, and metaproteomic datasets. We also use three single-organism proteomes to examine the influence of cofragmentation on less complex samples.

2.3 Methods

We developed a computational model to simulate each aspect of the proteomic pipeline (called 'cobia' for **co**fragmentation **bia**s). There are three stages of the approach all of which are *in silico*: (1) protein digestion and peptide modification, (2) liquid-chromatography elution time prediction, and (3) tandem mass spectrometry. There are two required inputs, (1) predicted protein sequences (in a *.fasta* file format) and (2) liquid chromatography and mass spectrometry parameters (Fig. 2.1). The entire model can be used from the command line as a series of modular command line functions, and is written in Python 2.7

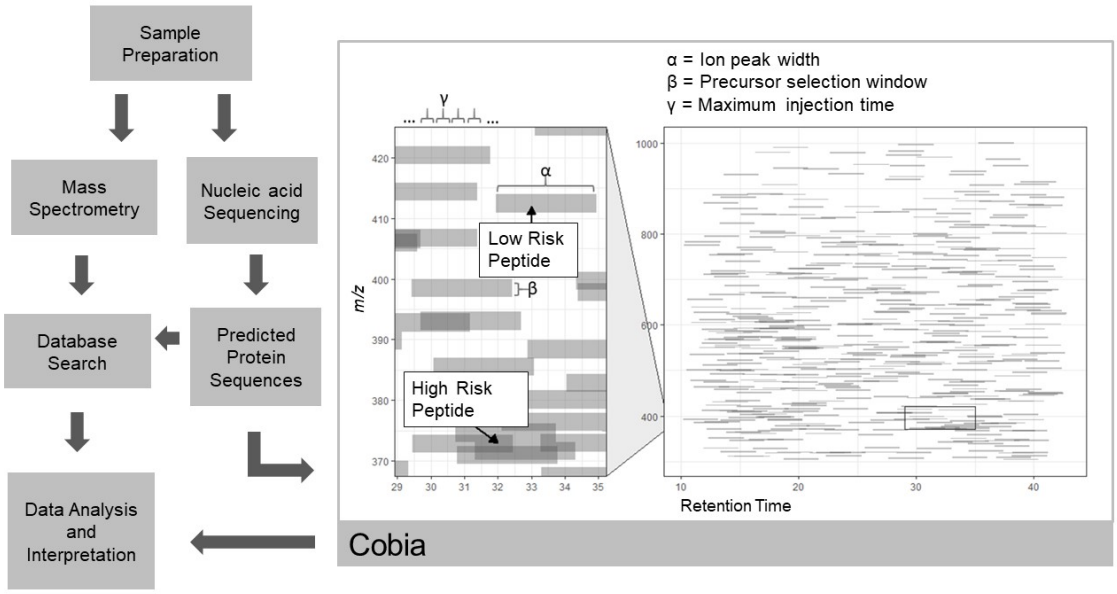

Figure 2.1: Conceptual schematic of cofragmentation bias prediction. The left side indicates where *cobia* fits in a proteomics pipeline and what the main required input is (a predicted protein sequence file). The right side indicates how *cobia* works conceptually. Peptides are shown as grey bars in a retention time and *m/z* map. Three key parameters required for input are shown: α, representing the ion peak width (in minutes, constant for all peptides); β, the precursor selection window (*m/z*); and γ, the maximum injection time, a discretized bin representing an ion packet injection.

(for installation and instructions see https://github.com/bertrand-lab/cobia). All code is open source and available under an MIT license.

2.3.1 Protein Digestion and Peptide Modification

A predicted protein sequence file either from a metatranscriptome or metagenome is the first input. Proteins are digested *in silico* using trypsin, assuming complete digestion. Peptides under five amino acids are removed at this stage to improve computational speed, as they are typically uninformative for protein inference. Peptides with unknown amino acids ('X' or '*') or selenocysteine ('U') were removed from the simulation, as the liquid chromatography prediction could not accommodate these (described below), and they were a very small proportion of peptide sequences. Peptides with uncertain amino acids ('Z' for glutamic acid or glutamine, and 'B' for aspartic acid or asparagine) were converted to one amino acid deterministically (glutamine and asparagine, respectively). We also applied

fixed modifications for methionine and cysteine (oxidation and carbamidomethylation, respectively); cysteine carbamidomethylation is a typical consequence of alkylation and methionine oxidation is due to oxygen exposure during sample preparation. All peptide modifications done *in silico* used the *pyteomics* Python module (Goloborodko et al., 2013).

2.3.2 Liquid Chromatography Elution time Prediction

We used two different retention time prediction approaches for different datasets, (1) a thermodynamic model of peptide behaviour (BioLCCC; Gorshkov et al., 2006; Perlova et al., 2010), and (2) a machine learning model trained on retention times of observed peptides (RTModel; Pfeifer et al., 2007). Our choice of retention time prediction approach depended on the type of liquid chromatography used. For example, BioLCCC cannot predict elution times of a two-dimensional separation. Different retention time prediction approaches can be easily used to replace these, as future retention time prediction methods become more accurate.

For the first approach (BioLCCC), retention times are predicted solely using liquid chromatography column characteristics. These column characteristics are set up in our simulation as a separate input file requiring the following: column length, column diameter, column pore size, solvent concentrations, gradient parameterization for non-linear gradients, and flow rate.

The second approach (RTModel) is fundamentally different, in that a support vector machine (SVM) is trained on observed peptides from a mass spectrometry experiment. The challenge with this approach is the explicit use of a biased training set. Therefore, we expect peptides that are unlikely to be observed via mass spectrometry (for example, because of poor ionization efficiency), to also have biased predictions. We used three different SVM kernels: a sequence-specific SVM kernel (Pfeifer et al., 2007) ('OLIGO' in RTModel), a radial basis function (RBF), and a linear kernel.

2.3.3 Tandem Mass Spectrometry simulation

Peptides are ionized with electrospray ionization and ion charge state depends on the peptide sequence. In our model, peptides are assigned a charge state of two, except if they contain a histidine or a lysine/arginine followed by a proline, where we then assign a charge state of three. We also only consider peptides between 50–2000 *m/z*, replicating typical limits of precursor ion scanning.

The injection and scanning mechanism of mass spectrometers lends itself to a simple way of simulating MS behaviour. In an Orbitrap VelosPRO (Thermo Fisher Scientific), ions accumulate in the C-trap and subsequently are injected into the Orbitrap. To simulate this, we can discretize retention times based on this accumulation-and-injection method (Fig. 2.1). Based on MS/MS ion injection times and recommended settings for an Orbitrap Elite (Kalli et al., 2013), we used a time of 500ms for binning all peptide elution times (γ, the 'maximum injection bin', in Fig. 2.1), and a precursor selection window of 3 *m/z* (β in Fig 2.1). The number of cofragmenting peptides is calculated by summing all peptides within the same 'injection bin' (i.e. similar elution times) and within the same precursor selection window.

We also simulated the behaviour of a peptide eluting off a column over a given period of time, the 'ion peak width' (Fig. 2.1). This value will change as a function of the chromatography parameters, and represents the continuous nature of peptide elution. This is a user-defined parameter, and can either be examined from raw data, or, approximated using a simple linear model of mean ion peak width as a function of gradient and column length (Hsieh et al., 2012).

An important component of this simulation is the explicit ignorance of ion intensity. We hypothesize that the number of potentially cofragmenting ions, as determined from the predicted protein sequence input file and the predicted retention times, relates to the probability of peptide identification. Ignoring ion intensity is necessary, as it is not possible to assign intensities to all peptides *a priori*. Those peptides that could be assigned intensities would be a biased subset; they are identified peptides, so were not impacted by cofragmentation. In addition, we do not simulate automatic gain control, instead simulating a constant ion injection time.

We compute a 'cofragmentation score', which is the average number of cofragmenting ions across all injection bins that a given peptide is present in (Fig. 2.1). Note that the sum of unique peptides potentially cofragmenting with a given peptide is typically much higher than the score, due to the overlap of cofragmenting peptide ion peaks across different portions of the target ion peak.

2.3.4 Computational Approach

We employed two methods to improve run time. First, we used sparse sampling of injection bins. Ion peak width is a much longer time interval than injection bin (i.e. $\alpha > \gamma$), so if we

only calculate number of cofragmenting ions in every *n*th injection bin, we can approximate the exhaustive sampling of every injection bin. We found that given a fixed 'maximum injection time' (500ms), if each ion is sampled at least 14 times (i.e. the ratio of *n*th sampling : ion peak width = 14), cofragmentation scores maintain high similarity to the exhaustive approach (>0.99 coefficient of determination of predicted cofragmentation with sparse vs. exhaustive sampling). We also implemented a parallel-computing option, specified in the input parameter file. Either a global or targeted approach can be designated in the input parameter file; a global approach predicts cofragmentation scores for every peptide, while a targeted approach predicts cofragmentation scores for a subset of peptides.

2.3.5 Model Validation and Datasets: Metaproteomics

To validate our model, we assessed whether the cofragmentation score could explain variation in the presence or absence of a peptide observed with mass spectrometry. In total, we used five different datasets from three different studies that paired metagenomic or metatranscriptomic sequencing with MS-based metaproteomics (Table 2.1, proteomic data retrieved from PRIDE; Vizcaíno et al., 2013; Perez-Riverol et al., 2019). We used these metagenomes and metatranscriptome as the predicted protein sequence input file for *cobia*. We then use an additional validation approach on a subset of two of these metaproteomic datasets.

The first study characterized the metaproteome of diseased oak trees to examine microbial gene expression *in situ* (Broberg et al., 2018). The authors paired metaproteomic work with sample-specific metagenomes and metatranscriptomes. The authors performed a 2D-LC peptide separation, with offline, high pH reversed-phase chromatography as the first dimension, collected in four fractions. For our LC prediction, we only used the characteristics of the second dimension, as the majority of peptides were collected within one fraction (Supplementary Table S33 from Broberg et al., 2018)). We used BioLCCC (Gorshkov et al., 2006) for peptide retention time prediction, replicating their non-linear gradient. We used the metagenome and metatranscriptome of one diseased sample (sample A4) separately as predicted protein sequence input files to *cobia*. Sample-specific metagenomic and metatranscriptomic protein identifications were obtained from Supplementary Table S36 (additional file 39; Broberg et al., 2018), and Swissprot feature lists were converted to protein sequences. Finally, we used peptide identifications from the Supplementary Table (S33) for additional validation, as described below.

The second study used metaproteomics as a method to determine biomass contributions from different microbial taxa (Kleiner et al., 2017). Here we used two different MS experiments which studied a mock microbial community (Kleiner et al., 2017). These experiments used different chromatography run times (260 and 460 minutes; Run 1 and 4, sample C4, PRIDE Project PXD006118). We identified peptides by repeating the mass spectra preprocessing (baseline removal, Savitzky-Golay filtering, and peak picking) and database searching (with MSGF+ and OpenMS; Kim and Pevzner, 2014; Röst et al., 2016; Weisser et al., 2013). The metagenomic database used for their peptide identifications was used as our predicted protein sequence input file for our LC RTModel model training to produce cofragmentation scores (also retrieved from PRIDE Project PXD006118).

Lastly, we used data from a study that paired metagenomics and metaproteomics of a fungal ant garden (Aylward et al., 2012). The authors isolated the bacterial fraction for metagenomics and we used the resulting dataset as our predicted protein sequence file. So, similar to the oak tree study, the predicted protein sequence file (aggregated metagenomes IMG/M taxon subject ID 2029527004, 2029527005, 2029527006) contains only a subset of the sequences expected to be in the mass spectrometry sample. While the authors used two separate mass spectrometry approaches, we used data only from their one-dimensional LC separation and used RTModel for peptide retention time predictions. We trained RTModel from peptide observations across all samples with the same LC methods (identified using Orbitrap). We used all peptide identifications from their supplementary material (Supplementary Dataset 5; Aylward et al., 2012) for subsequent validation.

For each of these studies, we determined cofragmentation scores for all tryptic peptides using the metagenome or metatranscriptome-derived predicted protein sequence file. We then examined which predicted peptides were actually observed with mass spectrometry, and assessed whether low cofragmentation scores are associated with increased probability of peptide detection by MS. To assess the explanatory power of our cofragmentation score, we used a generalized linear model (GLM) with a binomial error distribution and a logit link function (Nelder and Wedderburn, 1972), with cofragmentation score as the only explanatory variable. Note that for some of the datasets, peptides were found by the authors that we did not consider (e.g. peptides with missed trypsin cleavages; Fig. 2.2). To compare the influence of cofragmentation scores across validation datasets, we needed to account for variation in sequencing depth, as higher sequencing depth would lead to higher

cofragmentation scores overall. So, we also used GLMs with a scaled cofragmentation score as an explanatory variable (transforming all scores from 0–100), instead of the raw score.

We wanted to further test the explanatory value our cofragmentation score: perhaps there are other characteristics of peptides that better explain presence/absence? For example, larger peptides would have a higher probability of identification, simply because there is more sequence variation in longer peptides. Thus, we tested if *m/z* and retention time also explain variation in presence-absence, in addition to our cofragmentation score. We repeated the above GLMs, except with three explanatory variables – cofragmentation score, *m/z*, and retention time.

In addition to the approach above, we further examined data collected by Broberg et al. (2018). Peptide identifications are used to infer proteins present in a sample (i.e. protein inference). The collection of proteins inferred to be present in a sample can be used to determine a set of 'inferred peptides' by digesting these proteins *in silico*. Only a subset of these inferred peptides was detected; perhaps the other portion was not detected because of cofragmentation. Thus, we expect that the identified peptides should have lower cofragmentation scores than non-identified peptides. Further, we hypothesized that cofragmentation scores have more explanatory power when we only consider this smaller pool of 'inferred peptides', compared to when using the entire predicted protein sequence file. As above, we used a GLM with presence/absence of a peptide as the response variable and cofragmentation score as the only explanatory variable.

2.3.6 Model Validation and Datasets: Single-Organism Proteomics

After examining whether cofragmentation scores could explain variation in the presence and absence of peptides in highly complex, metaproteomic samples, we wanted to determine whether these scores had the same explanatory power in less complex samples. We used three datasets of single-organism proteomes to determine the explanatory power of our cofragmentation scores. For each single-organism proteome we examined, we 1) trained a support vector machine (RTModel) using observed peptides and retention times with the sequence-specific SVM kernel 'OLIGO', 2) used the protein sequence database from each study to predict peptide-specific cofragmentation scores (their databases were the predicted protein sequence file for *cobia*), and 3) assessed whether higher cofragmentation scores were associated with decreased probability of identifying a peptide.

The first dataset we used was an examination of human prostate cancer biomarkers in urine (Davalieva et al., 2018). We used observed peptides and retention times (PRIDE Project PXD008407), and trained a retention time model using RTModel. We used the corresponding human protein-coding genome that was these authors' database for the predicted protein sequence input file to *cobia*.

The second dataset we used looked at *Escherichia coli* across growth conditions (Schmidt et al., 2016). We used observed peptides and retention times reported in *mzid* files (PRIDE Project PXD000498). These authors used different chromatographic separations, so we only included those that were observed with 1D separation. We used the *E. coli* protein coding genome from UniProt, used by these authors as a protein sequence database, for the predicted protein sequence input file.

The third single organism dataset we used examined the influence of space flight on the mouse liver proteome (Anselm, Novikova and Zgoda, 2017, hereafter we refer to this as 'Space Mouse'). We used all observed peptides and associated retention times (PRIDE Project PXD005102) to train the retention time model (RTModel), and the mouse protein coding genome from UniProt as the predicted protein sequence input file.

We used a similar approach as above to assess the explanatory power of our cofragmentation score for single-organism proteomes. Again, we used a GLM with a binomial error distribution and a logit link function (Nelder and Wedderburn, 1972), with cofragmentation score as the only explanatory variable.

2.3.7 Comparing Metaproteomic and Single-Organism Mass Spectrometry Data

We examined mass spectrometry data associated with the metaproteomic and single-organism proteomic data, to determine if there were differences in sample complexity evident in the raw data. One way to compare sample complexity is to count the number of peaks detected in the MS1 scan, as the number of MS1 peaks is proportional to the number of peptides in a sample. We examined a subset of the studies above (where raw data were available), and computed the average number of MS1 peaks per MS1 scan using pyOpenMS (Röst et al., 2014). When necessary, we converted raw mass spectrometry files into mzML files using ThermoRawFileParser (Hulstaert et al., 2020).

2.3.8 Assessing Cofragmentation Bias at Coarse Functional and Taxonomic Levels

We hypothesized that coarse taxonomic or protein functional groups would be differentially impacted by cofragmentation bias. For example, highly conserved proteins may be comprised of similar peptides and therefore may cofragment. We used a metatranscriptome from the Antarctic sea ice edge as an input, simulating typical levels of sample complexity in surface seawater. This metatranscriptome has previously been used to examine micronutrient colimitation of phytoplankton and the microbial interactions that underpin that colimitation (Bertrand et al., 2015). Note that we do not have corresponding mass spectra for these samples.

In order to assess potential taxonomic and functional biases, we compared cofragmentation scores across groups. We examined the distribution of cofragmentation scores at the taxonomic group level to test for biases. We also examined the distribution of cofragmentation scores in different EuKaryotic Orthologous Groups (KOG) classes to examine biases in protein functional groups. For both the taxonomic groups and KOG classes, we selected peptides that uniquely correspond to a given grouping. We then aggregated these peptides and present cofragmentation risk at the protein level using the minimum peptide cofragmentation score per protein. If there were stronger bias in a given taxonomic or functional grouping, we would observe a distribution of protein cofragmentation scores within that group that is different from the global distribution of cofragmentation scores for the whole dataset. We employed the Kullback-Leibler (KL) divergence to quantify the difference between the score distribution in a given taxonomic group or KOG class and the global score distribution. High KL divergence values mean that two probability distributions are dissimilar. Therefore, if a taxonomic or functional protein group is differentially impacted by cofragmentation scores (i.e. they exhibit cofragmentation bias), we would observe high KL divergence when comparing this grouping to the global distribution of cofragmentation scores. We then calculated bootstrapped confidence intervals for each grouping. We randomly sampled proteins in the dataset n times, where n is the number of assigned protein sequences or open reading frames (ORFs) in a given grouping, and with each sample KL divergence was recalculated. We recalculated KL divergence with 1000 bootstrapped samples to construct confidence intervals.

2.3.9 Examining the Influence of Cofragmentation on Biomarker Peptides in Complex Communities

We employed a case study to identify how cofragmentation might influence conclusions of a metaproteomic study that examines biomarkers. Protein biomarkers can be used as indicators of nutrient-stress in the ocean (Saito et al., 2014). These biomarker proteins can be powerfully profiled in a taxon-specific manner, with the aim of identifying nutritional status or metabolic activity of specific members of the community. Given that this approach requires interpretation of a limited number of peptides within a complex mixture, this approach may be particularly susceptible to problems arising from cofragmentation bias. As an example, we examined vitamin B12-independent methionine synthase (MetE), a protein that has been used as a robust vitamin B12 starvation indicator (Bertrand et al., 2013). We limited our peptide choice to just those uniquely identifying MetE derived from *Fragilariopsis cylindrus*, a dominant phytoplankton in the Southern Ocean.

We employed the same Antarctic metatranscriptomic dataset as above, and determined peptide-specific cofragmentation scores for the peptides of interest. We used the retention time prediction model settings applied as above (Broberg et al., 2018). *Fragilariopsis cylindrus*-specific peptides were identified as previously described (Bertrand et al., 2015) and verified using NCBI non-redundant BLASTP, to ensure that the peptides were present in protein coding genes from the published *F. cylindrus* genome (Mock et al., 2017). We only searched for tryptic peptides, and limited the selection to peptides with sequence-modified *m/z* from 300-2000 assuming a charge state of two (as above). Lastly, we used CONSeQuence to score peptide detectability based on physicochemical properties rather than susceptibility to cofragmentation (Eyers et al., 2011), in order to consider additional variables important for biomarker peptide choice.

2.4 Results

2.4.1 Validation: Metaproteomics

We found that higher cofragmentation scores were associated with a lower probability of observing a peptide in four of five metaproteomic datasets (Fig. 2.2). From these datasets,

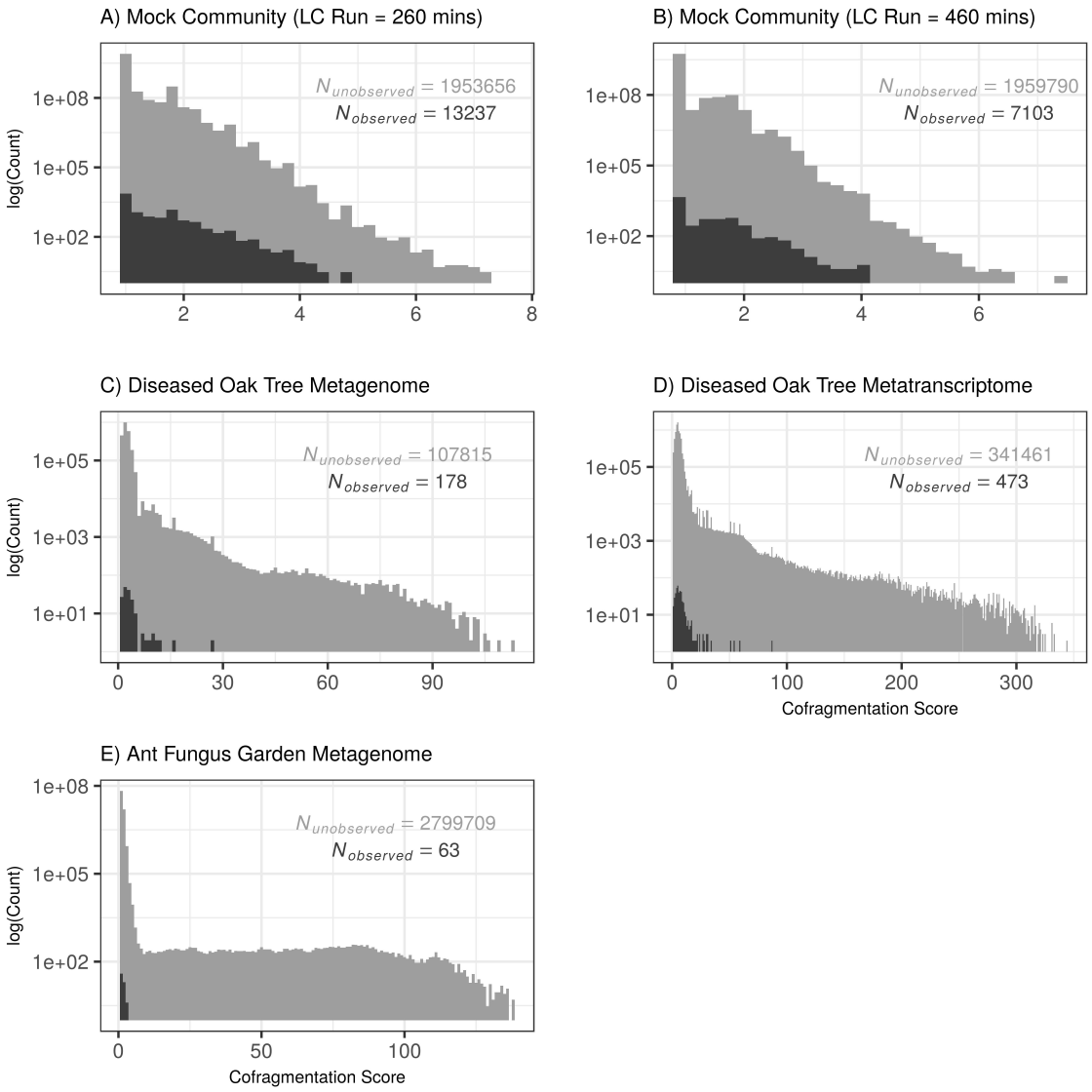

Figure 2.2: Histograms of cofragmentation scores for peptides indicate that higher cofragmentation scores are associated with a lower probability of observing a peptide. In light grey, the distribution of all peptides' cofragmentation scores is shown. In dark grey, the cofragmentation score of observed peptides in the paired metaproteomics experiment is shown. Note panels have different axis scales, and the y-axis is log-transformed (base 10). Different histogram binning represents variation in x-axis scales.

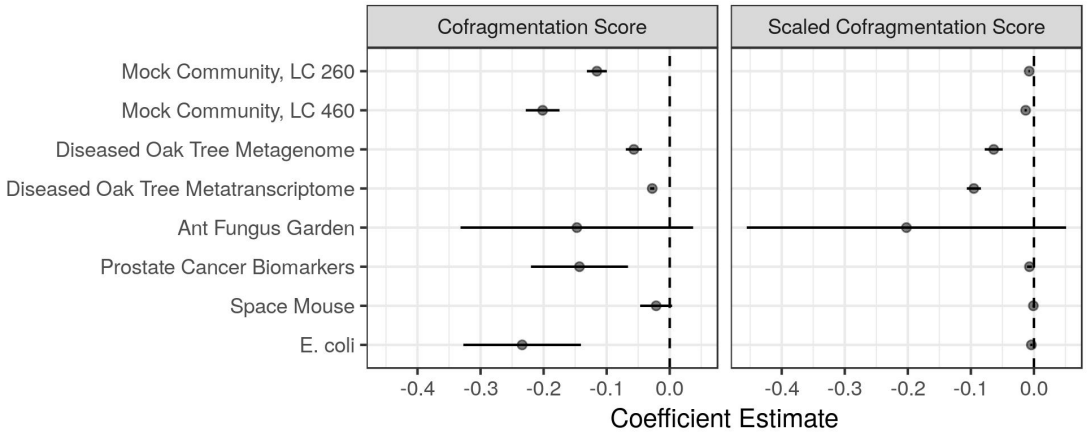

Figure 2.3: Coefficient estimates of generalized linear models for each dataset, describing the effect of cofragmentation score on the probability that a peptide is observed in the mass spectrometry experiment. Left panel shows coefficient estimates for absolute cofragmentation scores used for prediction, right panel shows coefficient estimates calculated when using cofragmentation scores are first scaled from 0–100 (to compare across datasets).

coefficient estimates were all negative (-0.028 – -0.20, Fig. 2.3), indicating that higher cofragmentation scores are associated with lower probability of identification. However, the overall probability of observing a random peptide from all potential peptides was low, given the immense number of potential peptides in a sample (Supplementary Table 2.1). The range in coefficient estimates is also in part due to the range in cofragmentation score. For example, in the diseased oak tree metatranscriptome, cofragmentation scores ranged from 1–300, and the coefficient estimate was -0.02 (Fig. 2.3, left panel). To compare across datasets, we then scaled cofragmentation scores to range from 0–100 and refit the GLMs. We found that scaling accounts for variation in absolute cofragmentation score due to different amounts of sequencing (different diseased Oak Tree coefficient estimates became similar).

Scaled coefficient estimates from the mock community datasets were lower than other metaproteomic datasets, suggesting a lower overall influence of cofragmentation on peptide observability (Fig. 2.3, right panel). The lower overall cofragmentation risk is also reflected in the range of raw cofragmentation scores (mock community scores ranged from 0–8). Further, the choice of retention time prediction method plays a large role in the range of cofragmentation scores.

In the ant fungus garden metaproteomic dataset, cofragmentation scores did not explain variation in presence/absence of a peptide (Fig. 2.2, Fig. 2.3). Notably, we

had low statistical power for this dataset due to the low total number of tryptic peptides without missed cleavages identified (63 peptides, of the 242 identified for the Orbitrap mass spectrometry experiments).

We also sought to examine other peptide characteristics that may more easily explain peptide identification: *m/z* and retention time. For both the 260 and 460 minute LC mock community datasets, adding in these two other explanatory variables increased the coefficient estimate by approximately 50% (Supplementary Fig. 2.7). For all other datasets, adding these additional explanatory variables did not result in significant changes to the cofragmentation score coefficient estimate (Supplementary Fig. 2.7). While including additional explanatory variables decreased the explanatory power of cofragmentation score, this score continued to explain significant variation in peptide presence/absence.

For our second validation approach that examined just 'inferred peptides', we found that higher cofragmentation scores were also associated with decreased probability of detecting a peptide (Fig. 2.4). However, cofragmentation scores had only slightly higher explanatory power (i.e. similarly negative coefficient estimates with overlapping confidence intervals) when using just the inferred peptides rather than the entire set of potential peptides (Fig. 2.4, bottom).

Qualitatively, the distribution of cofragmentation scores differed between retention time prediction methods (Supplementary Fig. 2.8), which is likely related to the distribution of retention times. For the mock community datasets and the ant fungus garden, we used RTModel and RTPredict, a support vector machine trained on the specific mass spectrometry experiment. The distribution of retention times for these studies more closely resembled a Gaussian distribution, while the distribution of retention times for datasets with peptides predicted from BioLCCC has a mass of observations at the beginning and end of the retention time distribution (Supplementary Fig. 2.8). The SVM kernel choice did not appear to have a large effect on predicted retention times or cofragmentation coefficient estimates; except for the ant fungus garden (Supplementary Fig. 2.7). In this dataset, the number of observed peptides was small, leading to very different retention time predictions across SVM kernels (Fig. 2.7).

2.4.2 Validation: Single-Organism Proteomics

In contrast to the metaproteomes, higher cofragmentation scores were not significantly associated with a lower probability of identification for two of the three single-organism

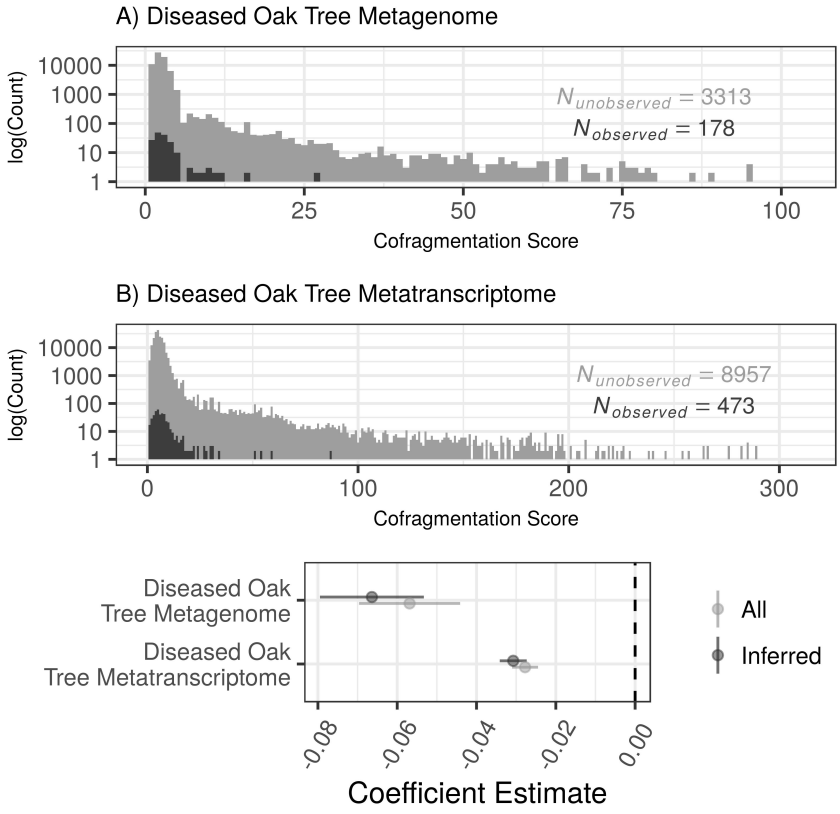

Figure 2.4: Histograms of cofragmentation scores for peptides indicate that higher cofragmentation scores are associated with a lower probability of observing a peptide, using a smaller subset of potential 'inferred' peptides (see Methods). 'Inferred peptides' are all peptides produced by proteins that were detected via observation of one or more peptides. We hypothesized that by using this more accurate representation of peptides present in the sample, cofragmentation scores would have stronger explanatory power. Coefficient estimates of generalized linear models (bottom panel) show that cofragmentation scores for inferred peptides are similar to coefficients from GLMs with all potential peptides.

proteomes we examined (Fig. 2.3, 2.5). Further, the range of cofragmentation scores observed were lower than those in the metaproteomes, indicating a lower overall sample complexity (Fig. 2.5). As above, we added two additional explanatory variables to the GLMs to see if other peptide characteristics could better explain peptide presence/absence. We did not find that cofragmentation scores explanatory value significantly changed with these additional explanatory variables (Fig. 2.7).

2.4.3 Comparing Metaproteomic and Single-Organism Mass Spectrometry Data

Overall we observed higher sample complexity in the metaproteomic mass spectrometry data compared with the single-organism data (Table 2.1). The average number of MS1 peaks per MS1 scan was 1–2 orders of magnitude higher for the metaproteomic datasets compared to the single organism datasets (Table 2.1).

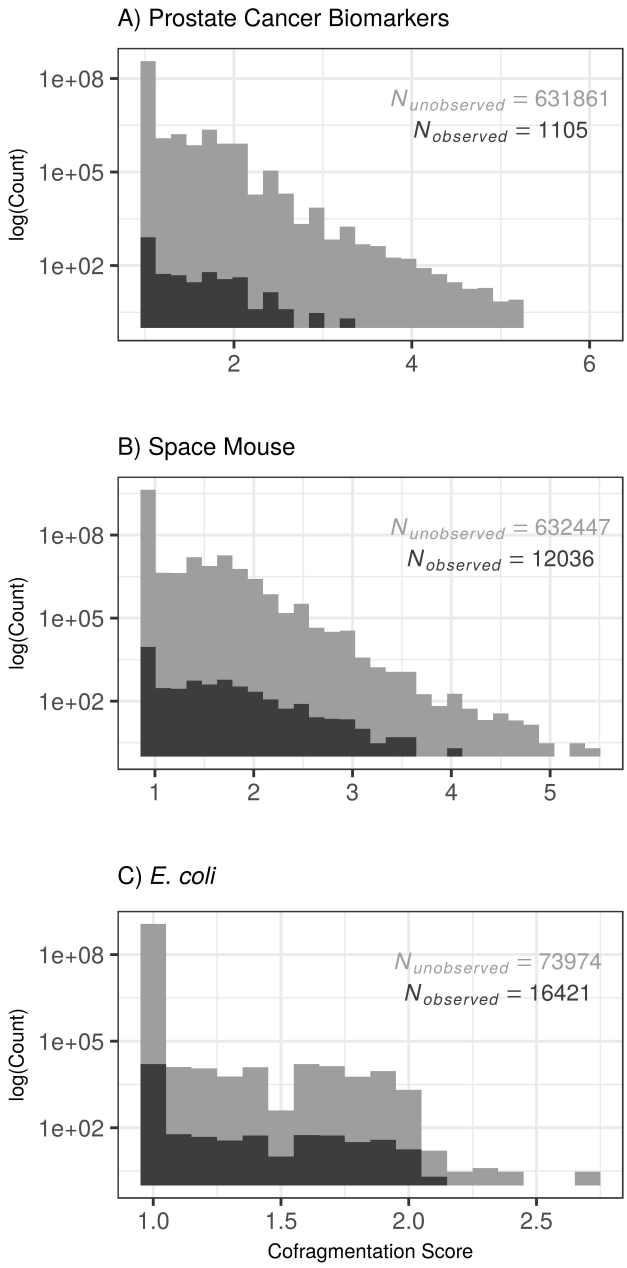

Figure 2.5: Histograms of cofragmentation scores for peptides indicate that higher cofragmentation scores are not significantly associated with a lower probability of observing a peptide for single organism proteomes, except for the *E. coli* dataset. In light grey, the distribution of all peptides' cofragmentation scores is shown. In dark grey, the cofragmentation score of observed peptides in the paired proteomics experiment is shown. Note panels have different axis scales, and the y-axis is log-transformed (base 10). Different histogram binning represents variation in x-axis scales.

Table 2.1: Summary of different metaproteomic and single-organism proteomic datasets we used for *cobia* validation. MS1 spectral characteristics are also presented, calculated from the raw mass spectrometry data. NA = Not available from primary data source. *Datasets had multiple mass spectrometry runs, we calculated the average number of MS1 scans and average sum of MS1 peaks across mass spectrometry runs.

Study	Metaproteome	Retention Time Prediction Method	Mean MS1 Peaks per MS1 Scan	Additional Notes
Mock Community (LC Run = 260 minutes)	Yes	RTModel	3.18E+04	
Mock Community (LC Run = 460 minutes)	Yes	RTModel	2.58E+04	
Diseased Oak Tree Metagenome	Yes	BioLCCC	NA	Offline, 2D-LC separation, with first dimension separated into four fractions.
Diseased Oak Tree Metatranscriptome	Yes	BioLCCC	NA	
Ant Fungus Garden*	Yes	RTModel	3.75E+04	Isolated DNA for bacterial fraction of sample.
Prostate Cancer Biomarkers	No	RTModel	NA	
Space Mouse*	No	RTModel	4.66E+02	Examined only the mouse liver.
*E. coli**	No	RTModel	2.04E+03	Examined many different growth conditions of *E. coli*.

2.4.4 Assessing Cofragmentation Bias at Coarse Functional and Taxonomic Levels

At a coarse level, we observed similar distributions of cofragmentation scores at the protein level across broad taxonomic and functional groupings in our Antarctic sea ice edge metatranscriptome case study. (Figures S2.3, S2.4 display actual distribution comparisons between all taxonomic and functional groups and the overall group distribution). However, we did observe a general trend showing that groupings with fewer protein members had higher KL divergence values (Fig. 2.6), suggesting that they tended to exhibit more cofragmentation bias relative to the full dataset. For example, viruses had 394 protein coding open reading frames (ORFs) assigned to them; this taxonomic group was dissimilar to the global distribution of cofragmentation scores (i.e. the highest KL divergence values). Therefore, in these samples, viruses would be the most differentially impacted by cofragmentation bias. These results suggest that taxonomic or functional rarity confers more cofragmentation bias.

2.4.4.1 Examining the Influence of Cofragmentation on Biomarker Peptides in Complex Communities

We used the same Antarctic metatranscriptome and searched for peptides that would uniquely identify MetE derived from the diatom *F. cylindrus*, in order to (1) illustrate a potential workflow for selecting peptides from environmental samples for robust, in-depth interpretation within metaproteomic datasets and (2) show potential consequences of cofragmentation-induced bias for interpreting taxon-specific protein expression patterns in mixed communities. Overall, we found twelve candidate peptides that would uniquely identify MetE derived from *F. cylindrus*. We additionally examined which of these peptides would be likely detectable via mass spectrometry by using peptide CONSeQuence scores (Eyers et al., 2011), i.e. how likely are they to sufficiently ionize? We found that eight of these twelve peptides would be likely detectable using mass spectrometry (Table 2.2, CONSeQuence score greater than 2). For each of these peptides, there are multiple potential peptides that are of similar elution time and similar mass (i.e. their ion peaks overlap and they are within the same precursor selection window). The cofragmentation scores ranged from 13.55–256.43 (Table 2.2).

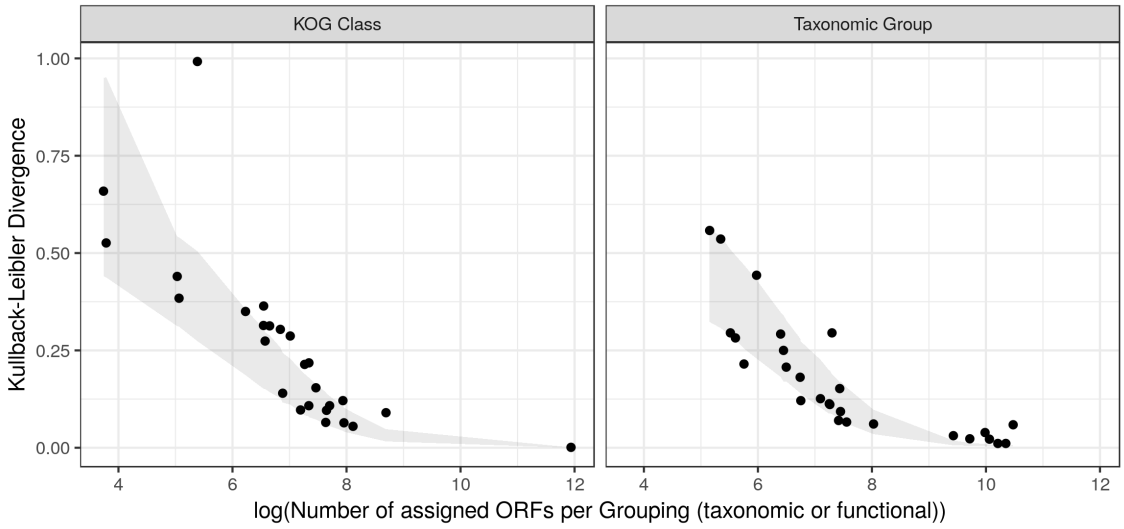

Figure 2.6: Taxonomic and functional groups with fewer assigned open reading frames (ORFs) have greater cofragmentation bias. Kullback-Leibler divergence was calculated by comparing the distribution of all cofragmentation scores calculated from an Antarctic metatranscriptome, with the distribution of cofragmentation scores for each functional (KOG class; left) and taxonomic (right) grouping. Grey area represents 5% and 95% bootstrapped confidence intervals (see Methods).

Table 2.2: Peptides derived from *Fragilariopsis cylindrus* vitamin B12 independent methionine synthase. Cofragmentation scores associated with each peptide, calculated from a metatranscriptomic dataset, are shown with the CONSeQuence score, representing peptide detectability.

Peptide Sequence	Cofragmentation Score	CONSeQuence Score
AVIYGPVTIIR	24.28	3
FAHLDAGIDR	33.69	3
QAYPSI	43.98	1
FALLAELIPIYQK	256.43	3
WFTTNYHYLPSEVDTK	28.49	4
FQTATLGLSR	35.37	3
LIQDLSDMGVK	35.06	3
GVDGATALGLK	39.26	2
HSTFAQTEGSIDVQR	33.83	4
AQAVEELGWSLQLADDK	36.37	4
FVGADK	195.14	0
LLPLYK	13.55	0

2.5 Discussion

We have shown that cofragmentation influences which peptides are observed in metaproteomics, and developed a computational model to calculate peptide-specific scores representing the risk of cofragmentation. We validated this model on multiple datasets, showing that higher cofragmentation scores are associated with decreased probability of identifying a peptide. Our results suggest that metaproteomic samples are influenced more by cofragmentation compared to single organism proteomics. Further, we found that functional or taxonomic rarity is associated with greater cofragmentation risk, such that biological conclusions drawn based on relatively few peptides are more susceptible to cofragmentation bias.

Cofragmentation is a challenging phenomenon to predict because it is one of many factors that influence peptide observability. Despite this, we found that the average number of potentially cofragmenting peptides (i.e. the cofragmentation score) is predictive of cofragmentation risk in metaproteomics. Our testing datasets also demonstrated that, in order to generate the most informative cofragmentation scores, the predicted protein sequence input file should accurately reflect the proteins contained in the actual sample. This was not the case for our validation dataset from the ant fungal garden, where it appears that the majority of protein injected into the mass spectrometer was not microbial while the predicted protein sequence input file was almost entirely of microbial origin. Notably, of all the metaproteome samples interrogated, cofragmentation scores for this dataset had the least explanatory power for peptide observability. Therefore, we anticipate our approach to have optimal predictions with sample-specific sequencing where the nucleic acid sequencing targets the same organisms that are included in the proteomic profiling. Our analysis of 'inferred peptides' did not support our hypothesis that cofragmentation scores would have substantially increased explanatory power with this subset of peptides. However, we anticipate that improvements in retention time prediction would improve cofragmentation scoring (Moruz and Käll, 2016). To that end, we have built *cobia* in a modular way, and demonstrate the use of different retention time predictors (BioLCCC and RTModel).

Single-organism proteomics appears to be less influenced by cofragmentation compared with metaproteomics. Our cofragmentation scores were not significantly associated with observed peptides in two of the three single organism proteomics studies we tested.

Our examination of the raw mass spectrometry data further suggested that metaproteomic samples are much more complex (when assessed using MS1 peaks). While further analysis of single organism proteomes is needed, particularly to compare the impact of cofragmentation in single-celled and multicellular organisms, our results suggest that cofragmentation is of elevated concern in metaproteomic experiments.

Our Antarctic metatranscriptome case study suggested that cofragmentation does not differentially bias biological conclusions that are based on coarse taxonomic and functional groupings when those taxonomic and functional groupings are assessed through the detection of large numbers of peptides. As the grouping becomes more granular or represented by smaller numbers of peptides, for example, examining taxon-specific protein biomarkers or looking at specific microbial strains, cofragmentation bias can play a larger role. This was demonstrated at the coarse level, where we showed that there was greater cofragmentation bias associated with smaller taxonomic and functional groupings (fewer open reading frames per grouping). Methods and types of analyses that aggregate peptides into larger groups, or examine more coarse aspects of metaproteomes, are more robust to cofragmentation bias. For example, we would not expect cofragmentation to significantly influence conclusions from Morris et al. (2010), who examined the distribution of transporter proteins at a coarse level across an oceanographic gradient. Aggregation of peptides into larger groups ('protein group inference', i.e. as 'MetaProteins' in MetaProteomeAnalyzer (Muth, Renard and Martens, 2016), or into Gene Ontology groups in MetaGOmics (Riffle et al., 2018)), are also likely to be robust to cofragmentation bias across diverse samples.

Rare taxa or protein functional groups are more susceptible to cofragmentation bias. Taxon-specific biomarkers represent an extreme level of rarity (i.e. expression levels of very few peptides are interpreted). We have highlighted one example of this to illustrate potential consequences of cofragmentation examining MetE derived from *F. cylindrus* to determine vitamin B12 nutritional status in this biogeochemically important diatom. Peptides derived from different organisms may disrupt the identification and quantification of these candidate peptides. If we examine only the most detectable peptides in terms of ionization efficiency (i.e. CONSeQuence score of 4), a subset of peptides are less susceptible to cofragmentation bias based on our cofragmentation scores. Our model could thus be used to guide peptide choice for in-depth biological interpretation within metaproteomics

experiments. We would therefore choose WFTTNYHYLPSEVDTK and AVIYGPVTIIR peptides as robust markers of *F. cylindrus* MetE protein rather than FALLAELIPIYQK and FVGADK peptides due to their lower risk of cofragmentation bias. Cofragmenting peptides may influence biological conclusions; for example, a bacterial-derived metal ion ATP binding cassette (ABC) transporter-derived peptide is of similar *m/z* and elution time to the candidate biomarker peptide AQAVEELGWSLQLADDK. Such proteins are often more highly expressed under low metal availability conditions (Bertrand et al., 2015). Thus, different indicators of nutrient starvation (i.e. a metal ABC transporter and MetE, for iron and vitamin B12 respectively) would mutually influence observability and therefore be confounded. For example, an increase in the bacterial metal ABC transporter might alter quantification estimates of MetE through prevention of precursor ion selection or interference with spectral counts. Accurately characterizing nutrient starvation indicators is important for quantifying and modelling carbon fixation and microbial interactions in the ocean. However in this case, peptide biomarkers might be obscured, potentially leading to spurious conclusions about nutrient stress and limitations of primary productivity. Cofragmenting peptides may interfere with peptide identification, but they also may hinder accurate quantification within a discovery-based approach (or even a targeted approach if there are conserved transitions between the peptides). With this example we illustrate one potential consequence of cofragmentation in metaproteomics, and also show how our model could guide peptide selection or weighting in metaproteomics.

It is also clear that cofragmentation could influence biological conclusions in other fields. For example, in a study of the human microbiome, 61% of bacterial 'species' were only inferred from one or two peptides (Zhang et al., 2018). Further, strain-specific peptides were used to examine microbiome compositional differences between patients with and without Crohn's disease (Zhang et al., 2018). Human neutrophil peptides have also been used to describe inflammatory bowel disease in humans in complex microbial community samples (Li et al., 2016). Our results suggest that all of these scenarios are susceptible to cofragmentation bias across samples; this could have consequences for inferring relationships between microbial communities and human disease states. Lastly, as we have shown that rarity is associated with greater cofragmentation bias, we caution researchers studying the 'rare microbiome' (Lynch and Neufeld, 2015; Jousset et al., 2017) using metaproteomics. Even though proteotypic and taxon-specific peptides are susceptible

to cofragmentation bias, they can be effective tools for asking particular questions. We do not doubt the importance of rare organisms in both human health and environmental microbiology – on the contrary, they can and do contribute critical functions and diagnostic potential. Therefore, the use of techniques susceptible to cofragmentation bias is still warranted and predicting this bias, as we do here, is necessary.

Several complementary approaches to ours have been previously developed. The most relevant and advanced approach is used in SRMCollider, a method for targeted proteomics that considers the uniqueness of a peptide transition, given a particular anticipated proteomic background (Röst, Malström and Aebersold, 2012). SRMCollider, however, would be a more appropriate choice for targeted proteomics peptide and transition selection compared with *cobia*, which was built for modelling discovery-based mass spectrometry and additionally offers explicit prediction of cofragmentation bias at the peptide level. MS simulation tools are becoming increasingly close to replicating peptide fragmentation and ion elution profiles (Bielow et al., 2011; Noyce et al., 2013; Goldfarb, Wang and Major, 2016), which may be used to improve cofragmentation scores. Adapting these aforementioned computational tools may improve the prediction of cofragmentation-induced bias in metaproteomics.

Discovery mass spectrometry has vastly improved in recent years, and, future applications may be more robust to cofragmentation. For example, decreasing precursor ion selection windows may reduce this bias. However, reducing this window results in a loss of sensitivity, and in practice a range of 2–4 *m/z* is optimal (Kalli et al., 2013). Data-independent acquisition (DIA) strategies embrace cofragmentation of many precursor ions, subsequently disentangling the chimaeric fragmentation spectra (Gillet et al., 2012; Chapman, Goodlett and Masselon, 2014). While we are currently unaware of any published uses of DIA in metaproteomics, we anticipate that our model predictions would be useful for such studies. To apply *cobia* to DIA, the only adjustment required would be increasing the 'precursor selection window' (Fig. 2.1). Another area of rapid advancement in mass spectrometry bioinformatics is *de novo* peptide sequencing, which is also particularly susceptible to 'mixture' spectra (Gorshkov et al., 2016). Improvements in liquid chromatography separation (i.e. using longer separation times or orthogonal separations) can reduce the total amount of cofragmentation. In particular, using orthogonal chromatographic separation techniques would lead to variations in elution times and

therefore produce a different set of cofragmentation scores. Similarly, additional mass spectrometry approaches for ion filtering, such as ion mobility mass spectrometry, would also provide an additional separation. Despite improvements in mass spectrometry and bioinformatics, we anticipate peptide cofragmentation to be a persistent issue.

2.5.1 Recommendations and Conclusions

There are several ways of handling cofragmentation bias in metaproteomics, and several use-cases of *cobia*. At a coarse level, cofragmentation bias is unlikely to influence biological conclusions for abundant taxa or functional groupings. At a more fine taxonomic or functional level, more caution should be used to interpret peptide presence or spectral counts. Choosing peptides as indicators for specific taxa and metabolic processes could incorporate the prediction of cofragmentation scores using our model, as illustrated in the MetE *F. cylindrus* example. Or, cofragmentation scores could be employed to weight evidence of peptide quantifications in a high-throughput MS experiment. To that end, we include a series of functions for programmatically predicting cofragmentation scores for metaproteomics (for instructions see www.github.com/bertrand-lab/cobia). If a peptide is not selected for fragmentation, it would still be detected in the MS1 scan. So, pragmatically, peptides of high importance as diagnostics could be manually examined in the *m/z*-to-retention time map. Future work will incorporate an automated examination of the *m/z* to retention time map to determine if a corresponding MS1 peak was present in the sample. Regardless of the method chosen for handling cofragmentation-induced bias, we encourage researchers using metaproteomics to be cognizant of this issue, particularly while probing individual taxa in complex samples using limited numbers of peptides.

We also suggest that peptides with a cofragmentation score above a given threshold should not be used to infer protein abundance across diverse samples. However, this threshold will vary depending on the question being asked and the researchers' tolerance for cofragmentation risk. A simple way to determine a reasonable threshold is to: 1) examine cofragmentation scores of peptides observed in mass spectrometry experiment, 2) determine the 95% percentile of scores, and 3) assume that peptides not detected that are above this cutoff may not have been detected because of cofragmentation. This has consequences for using peptides to determine trends across large numbers of samples and for imputing missing data, as it would perhaps be inappropriate to impute missing values for peptides with high cofragmentation scores above a set cutoff.

Lastly, we suggest that low abundance taxa should be removed when comparing community composition across diverse samples. Or, that particular care should be taken to identify and remove from analyses peptides that are particularly susceptible to cofragmentation bias. This has particular importance for researchers studying the rare microbiome using metaproteomics, and for comparing community composition and metabolism.

As the field of comparative metaproteomics grows, more and more dissimilar samples will be compared. Our study has shown that complexity in the *m/z*-retention time landscape, resulting in peptide cofragmentation, may influence biological conclusions. Broadly, we believe accounting for this dynamic landscape in metaproteomics is among the most significant (but surmountable) technical challenges facing the field.

2.6 Author Contributions

J.S.P.M. and E.M.B. conceived the study. J.S.P.M. wrote the code and conducted all analyses. J.S.P.M. wrote the paper with input from E.M.B.

2.7 Supplementary Information

2.7.1 Supplementary Table

Estimate	SE	z value	*p* Value	Dataset Name	Variable
-5.06239	0.02378	-212.882	0	Mock Community, LC 260, OLIGO kernel	(Intercept)
-0.15495	0.015304	-10.1244	4.30E-24	Mock Community, LC 260, OLIGO kernel	Mean Cofrag. Score
-5.44449	0.040223	-135.358	0	Mock Community, LC 260, Additional Explanatory, OLIGO kernel	(Intercept)
-0.07921	0.016316	-4.85486	1.20E-06	Mock Community, LC 260, Additional Explanatory, OLIGO kernel	Mean Cofrag. Score
-1.19E-06	2.07E-06	-0.57654	0.564254	Mock Community, LC 260, Additional Explanatory, OLIGO kernel	Retention Time
0.000403	2.62E-05	15.37643	2.36E-53	Mock Community, LC 260, Additional Explanatory, OLIGO kernel	*m/z*
-5.21734	0.011518	-452.969	0	Mock Community, LC 260, Rescaled, OLIGO kernel	(Intercept)
-0.01195	0.001181	-10.1244	4.30E-24	Mock Community, LC 260, Rescaled, OLIGO kernel	Rescaled Cofrag. Score
-5.10448	0.024323	-209.862	0	Mock Community, LC 260, Linear kernel	(Intercept)
-0.13071	0.01622	-8.05842	7.73E-16	Mock Community, LC 260, Linear kernel	Mean Cofrag. Score
-5.55233	0.037593	-147.695	0	Mock Community, LC 260, Additional Explanatory, Linear kernel	(Intercept)
-0.04866	0.01686	-2.88595	0.003902	Mock Community, LC 260, Additional Explanatory, Linear kernel	Mean Cofrag. Score
6.28E-06	1.81E-06	3.464356	0.000532	Mock Community, LC 260, Additional Explanatory, Linear kernel	Retention Time
0.00042	2.60E-05	16.14871	1.16E-58	Mock Community, LC 260, Additional Explanatory, Linear kernel	*m/z*
-5.23518	0.011338	-461.72	0	Mock Community, LC 260, Rescaled, linear kernel	(Intercept)
-0.01046	0.001298	-8.05842	7.73E-16	Mock Community, LC 260, Rescaled, linear kernel	Rescaled Cofrag. Score
-5.04146	0.024445	-206.234	0	Mock Community, LC 260, RBF kernel	(Intercept)
-0.17431	0.016288	-10.7014	1.00E-26	Mock Community, LC 260, RBF kernel	Mean Cofrag. Score
-5.4349	0.039198	-138.653	0	Mock Community, LC 260, Additional Explanatory, RBF kernel	(Intercept)
-0.09536	0.017087	-5.58082	2.39E-08	Mock Community, LC 260, Additional Explanatory, RBF kernel	Mean Cofrag. Score
8.69E-07	1.95E-06	0.446791	0.655026	Mock Community, LC 260, Additional Explanatory, RBF kernel	Retention Time
0.000398	2.61E-05	15.26819	1.25E-52	Mock Community, LC 260, Additional Explanatory, RBF kernel	*m/z*
-5.21577	0.011374	-458.578	0	Mock Community, LC 260, Rescaled, RBF kernel	(Intercept)
-0.01693	0.001582	-10.7014	1.00E-26	Mock Community, LC 260, Rescaled, RBF kernel	Rescaled Cofrag. Score
-5.5788	0.037126	-150.268	0	Mock Community, LC 460, OLIGO kernel	(Intercept)
-0.25774	0.027045	-9.53014	1.57E-21	Mock Community, LC 460, OLIGO kernel	Mean Cofrag. Score
-6.14109	0.054093	-113.528	0	Mock Community, LC 460, Additional Explanatory, OLIGO kernel	(Intercept)

Estimate	SE	z value	*p* Value	Dataset Name	Variable
-0.12736	0.027878	-4.56833	4.92E-06	Mock Community, LC 460, Additional Explanatory, OLIGO kernel	Mean Cofrag. Score
-1.98E-06	1.17E-06	-1.69076	0.090883	Mock Community, LC 460, Additional Explanatory, OLIGO kernel	Retention Time
0.000588	3.41E-05	17.21351	2.10E-66	Mock Community, LC 460, Additional Explanatory, OLIGO kernel	*m/z*
-5.83655	0.015134	-385.667	0	Mock Community, LC 460, Rescaled, OLIGO kernel	(Intercept)
-0.01901	0.001995	-9.53014	1.57E-21	Mock Community, LC 460, Rescaled, OLIGO kernel	Rescaled Cofrag. Score
-5.60371	0.035221	-159.101	0	Mock Community, LC 460, Linear kernel	(Intercept)
-0.23358	0.024923	-9.37194	7.12E-21	Mock Community, LC 460, Linear kernel	Mean Cofrag. Score
-6.19954	0.053106	-116.738	0	Mock Community, LC 460, Additional Explanatory, Linear kernel	(Intercept)
-0.1092	0.025687	-4.25136	2.12E-05	Mock Community, LC 460, Additional Explanatory, Linear kernel	Mean Cofrag. Score
8.74E-07	1.13E-06	0.775986	0.437757	Mock Community, LC 460, Additional Explanatory, Linear kernel	Retention Time
0.000587	3.42E-05	17.18159	3.65E-66	Mock Community, LC 460, Additional Explanatory, Linear kernel	*m/z*
-5.83728	0.015175	-384.665	0	Mock Community, LC 460, Rescaled, linear kernel	(Intercept)
-0.01693	0.001807	-9.37194	7.12E-21	Mock Community, LC 460, Rescaled, linear kernel	Rescaled Cofrag. Score
-5.58552	0.036042	-154.974	0	Mock Community, LC 460, RBF kernel	(Intercept)
-0.24867	0.025762	-9.65252	4.80E-22	Mock Community, LC 460, RBF kernel	Mean Cofrag. Score
-6.15499	0.052805	-116.56	0	Mock Community, LC 460, Additional Explanatory, RBF kernel	(Intercept)
-0.1196	0.026578	-4.4999	6.80E-06	Mock Community, LC 460, Additional Explanatory, RBF kernel	Mean Cofrag. Score
-1.30E-06	1.22E-06	-1.07001	0.284613	Mock Community, LC 460, Additional Explanatory, RBF kernel	Retention Time
0.000587	3.42E-05	17.13311	8.40E-66	Mock Community, LC 460, Additional Explanatory, RBF kernel	*m/z*
-5.83418	0.015197	-383.905	0	Mock Community, LC 460, Rescaled, RBF kernel	(Intercept)
-0.01851	0.001918	-9.65252	4.80E-22	Mock Community, LC 460, Rescaled, RBF kernel	Rescaled Cofrag. Score
-6.03304	0.098171	-61.4544	0	Diseased Oak Tree Metagenome	(Intercept)
-0.05688	0.012751	-4.46105	8.16E-06	Diseased Oak Tree Metagenome	Mean Cofrag. Score
-5.46366	0.18841	-28.9988	6.82E-185	Diseased Oak Tree Metagenome, Additional Explanatory	(Intercept)
-0.07014	0.013637	-5.14319	2.70E-07	Diseased Oak Tree Metagenome, Additional Explanatory	Mean Cofrag. Score
-0.01069	0.00257	-4.16104	3.17E-05	Diseased Oak Tree Metagenome, Additional Explanatory	Retention Time
0.000225	0.000342	0.659215	0.509758	Diseased Oak Tree Metagenome, Additional Explanatory	*m/z*
-6.08992	0.090473	-67.3118	0	Diseased Oak Tree Metagenome, Rescaled	(Intercept)
-0.06378	0.014297	-4.46105	8.16E-06	Diseased Oak Tree Metagenome, Rescaled	Rescaled Cofrag. Score
-6.13683	0.058372	-105.133	0	Diseased Oak Tree Metatranscriptome	(Intercept)
-0.02775	0.003271	-8.48501	2.16E-17	Diseased Oak Tree Metatranscriptome	Mean Cofrag. Score
-5.50598	0.121935	-45.1552	0	Diseased Oak Tree Metatranscriptome, Additional Explanatory	(Intercept)
-0.03439	0.003645	-9.43518	3.90E-21	Diseased Oak Tree Metatranscriptome, Additional Explanatory	Mean Cofrag. Score
-0.01056	0.001563	-6.75869	1.39E-11	Diseased Oak Tree Metatranscriptome, Additional Explanatory	Retention Time
0.000127	0.000207	0.611851	0.540636	Diseased Oak Tree Metatranscriptome, Additional Explanatory	*m/z*
-6.16459	0.056419	-109.265	0	Diseased Oak Tree Metatranscriptome, Rescaled	(Intercept)
-0.09527	0.011228	-8.48501	2.16E-17	Diseased Oak Tree Metatranscriptome, Rescaled	Rescaled Cofrag. Score
-10.4756	0.294744	-35.5414	1.13E-276	Ant Fungus Garden, OLIGO kernel	(Intercept)
-0.14729	0.184471	-0.79843	0.424622	Ant Fungus Garden, OLIGO kernel	Mean Cofrag. Score
-10.8235	0.451033	-23.9971	2.98E-127	Ant Fungus Garden, OLIGO kernel, Additional Explanatory	(Intercept)
-0.08738	0.154869	-0.56419	0.572624	Ant Fungus Garden, OLIGO kernel, Additional Explanatory	Mean Cofrag. Score
-7.11E-05	9.54E-05	-0.74516	0.456175	Ant Fungus Garden, OLIGO kernel, Additional Explanatory	Retention Time
0.000569	0.000326	1.74458	0.081058	Ant Fungus Garden, OLIGO kernel, Additional Explanatory	*m/z*
-10.6229	0.150318	-70.6695	0	Ant Fungus Garden, OLIGO Kernel, Rescaled	(Intercept)
-0.2022	0.253252	-0.79843	0.424622	Ant Fungus Garden, OLIGO Kernel, Rescaled	Rescaled Cofrag. Score
-10.6299	0.322634	-32.9474	4.60E-238	Ant Fungus Garden, Linear kernel	(Intercept)
-0.05142	0.213677	-0.24065	0.809826	Ant Fungus Garden, Linear kernel	Mean Cofrag. Score
-11.2002	0.49889	-22.4501	1.28E-111	Ant Fungus Garden, Linear kernel, Additional Explanatory	(Intercept)
0.078957	0.221474	0.356508	0.72146	Ant Fungus Garden, Linear kernel, Additional Explanatory	Mean Cofrag. Score
-4.55E-05	8.29E-05	-0.54914	0.582906	Ant Fungus Garden, Linear kernel, Additional Explanatory	Retention Time
0.000662	0.000334	1.979459	0.047764	Ant Fungus Garden, Linear kernel, Additional Explanatory	*m/z*
-10.6814	0.151057	-70.7107	0	Ant Fungus Garden, Linear Kernel, Rescaled	(Intercept)
-0.00331	0.013736	-0.24065	0.809826	Ant Fungus Garden, Linear Kernel, Rescaled	Rescaled Cofrag. Score
-10.1329	0.343927	-29.4625	8.72E-191	Ant Fungus Garden, RBF kernel	(Intercept)
-0.41029	0.242038	-1.69513	0.09005	Ant Fungus Garden, RBF kernel	Mean Cofrag. Score
-10.4334	0.530273	-19.6756	3.49E-86	Ant Fungus Garden, RBF kernel, Additional Explanatory	(Intercept)
-0.32493	0.251594	-1.29149	0.196534	Ant Fungus Garden, RBF kernel, Additional Explanatory	Mean Cofrag. Score
-7.77E-05	9.17E-05	-0.84727	0.396844	Ant Fungus Garden, RBF kernel, Additional Explanatory	Retention Time
0.000468	0.000334	1.401402	0.161094	Ant Fungus Garden, RBF kernel, Additional Explanatory	*m/z*
-10.5432	0.148163	-71.1596	0	Ant Fungus Garden, RBF Kernel, Rescaled	(Intercept)
-0.02989	0.017634	-1.69513	0.09005	Ant Fungus Garden, RBF Kernel, Rescaled	Rescaled Cofrag. Score
-6.55858	0.097396	-67.3391	0	Prostate Cancer Biomarkers, OLIGO kernel	(Intercept)
-0.07111	0.072623	-0.97912	0.32752	Prostate Cancer Biomarkers, OLIGO kernel	Mean Cofrag. Score
-6.83305	0.152543	-44.7943	0	Prostate Cancer Biomarkers, OLIGO kernel, Additional Explanatory	(Intercept)
-0.01151	0.075737	-0.15192	0.87925	Prostate Cancer Biomarkers, OLIGO kernel, Additional Explanatory	Mean Cofrag. Score
1.10E-05	2.46E-05	0.445544	0.655927	Prostate Cancer Biomarkers, OLIGO kernel, Additional Explanatory	Retention Time

Estimate	SE	z value	*p* Value	Dataset Name	Variable
0.000263	9.45E-05	2.786246	0.005332	Prostate Cancer Biomarkers, OLIGO kernel, Additional Explanatory	*m/z*
-6.62969	0.038291	-173.139	0	Prostate Cancer Biomarkers, OLIGO kernel, Rescaled	(Intercept)
-0.0034	0.00347	-0.97912	0.32752	Prostate Cancer Biomarkers, OLIGO kernel, Rescaled	Rescaled Cofrag. Score
-4.21048	0.032596	-129.173	0	Space Mouse, OLIGO kernel	(Intercept)
-0.0503	0.025878	-1.94374	0.051927	Space Mouse, OLIGO kernel	Mean Cofrag. Score
-4.25968	0.045884	-92.8355	0	Space Mouse, OLIGO kernel, Additional Explanatory	(Intercept)
-0.0395	0.026913	-1.46773	0.142177	Space Mouse, OLIGO kernel, Additional Explanatory	Mean Cofrag. Score
2.92E-05	4.00E-06	7.300709	2.86E-13	Space Mouse, OLIGO kernel, Additional Explanatory	Retention Time
-9.78E-05	3.04E-05	-3.21472	0.001306	Space Mouse, OLIGO kernel, Additional Explanatory	*m/z*
-4.26078	0.011374	-374.612	0	Space Mouse, OLIGO kernel, Rescaled	(Intercept)
-0.00239	0.001229	-1.94374	0.051927	Space Mouse, OLIGO kernel, Rescaled	Rescaled Cofrag. Score
-1.85061	0.08546	-21.6548	5.47E-104	E. coli, OLIGO kernel	(Intercept)
-0.02915	0.083418	-0.34947	0.726736	E. coli, OLIGO kernel	Mean Cofrag. Score
-1.77988	0.089999	-19.7765	4.74E-87	E. coli, OLIGO kernel, Additional Explanatory	(Intercept)
-0.04868	0.083829	-0.58074	0.561417	E. coli, OLIGO kernel, Additional Explanatory	Mean Cofrag. Score
-3.26E-06	2.39E-06	-1.36247	0.17305	E. coli, OLIGO kernel, Additional Explanatory	Retention Time
-5.52E-05	2.67E-05	-2.06542	0.038883	E. coli, OLIGO kernel, Additional Explanatory	*m/z*
-1.87977	0.009436	-199.218	0	E. coli, OLIGO kernel, Rescaled	(Intercept)
-0.00052	0.001483	-0.34947	0.726736	E. coli, OLIGO kernel, Rescaled	Rescaled Cofrag. Score

Table 2.3: Generalized linear model output from different datasets, support vector machine kernels, explanatory variable sets, and re-scaled cofragmentation scores.

2.7.2 Supplementary Figures

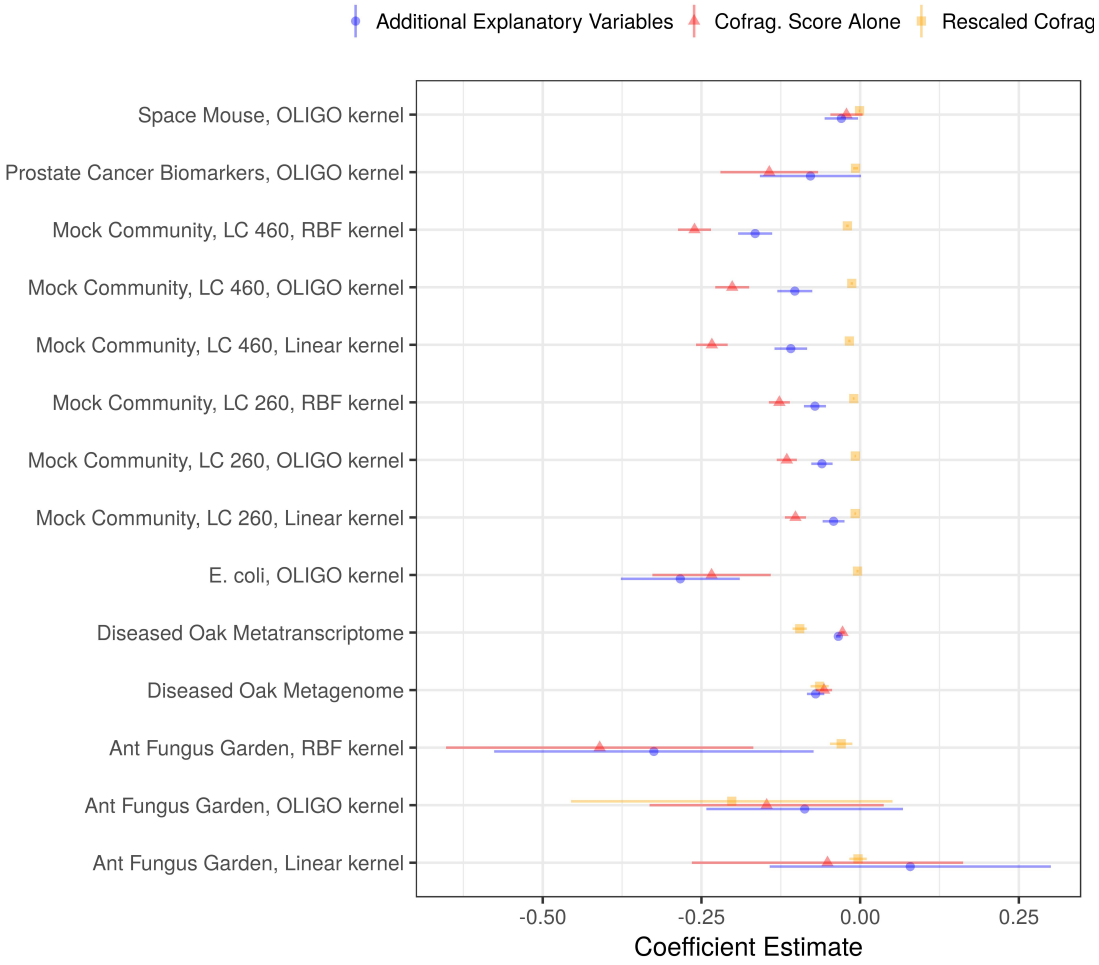

Figure 2.7: Coefficient estimates of generalized linear models describing the effect of cofragmentation score on the probability that a peptide is observed in the mass spectrometry experiment. Coefficients are shown for different retention time prediction methods (y-axis, OLIGO: sequence specific kernel; RBF: radial basis function; Linear; linear kernel); either using RTModel and RTPredict (mock communities and ant fungus garden) or using BioLCCC (diseased oak metagenome and metatranscriptome). For each retention time prediction method, three coefficients are shown: cofragmentation score coefficient as the only explanatory variable (red), cofragmentation score coefficient where scores are first scaled from 0–100 (orange), and cofragmentation score coefficient where other explanatory variable terms are included in the model structure (*m/z* and retention time; blue).

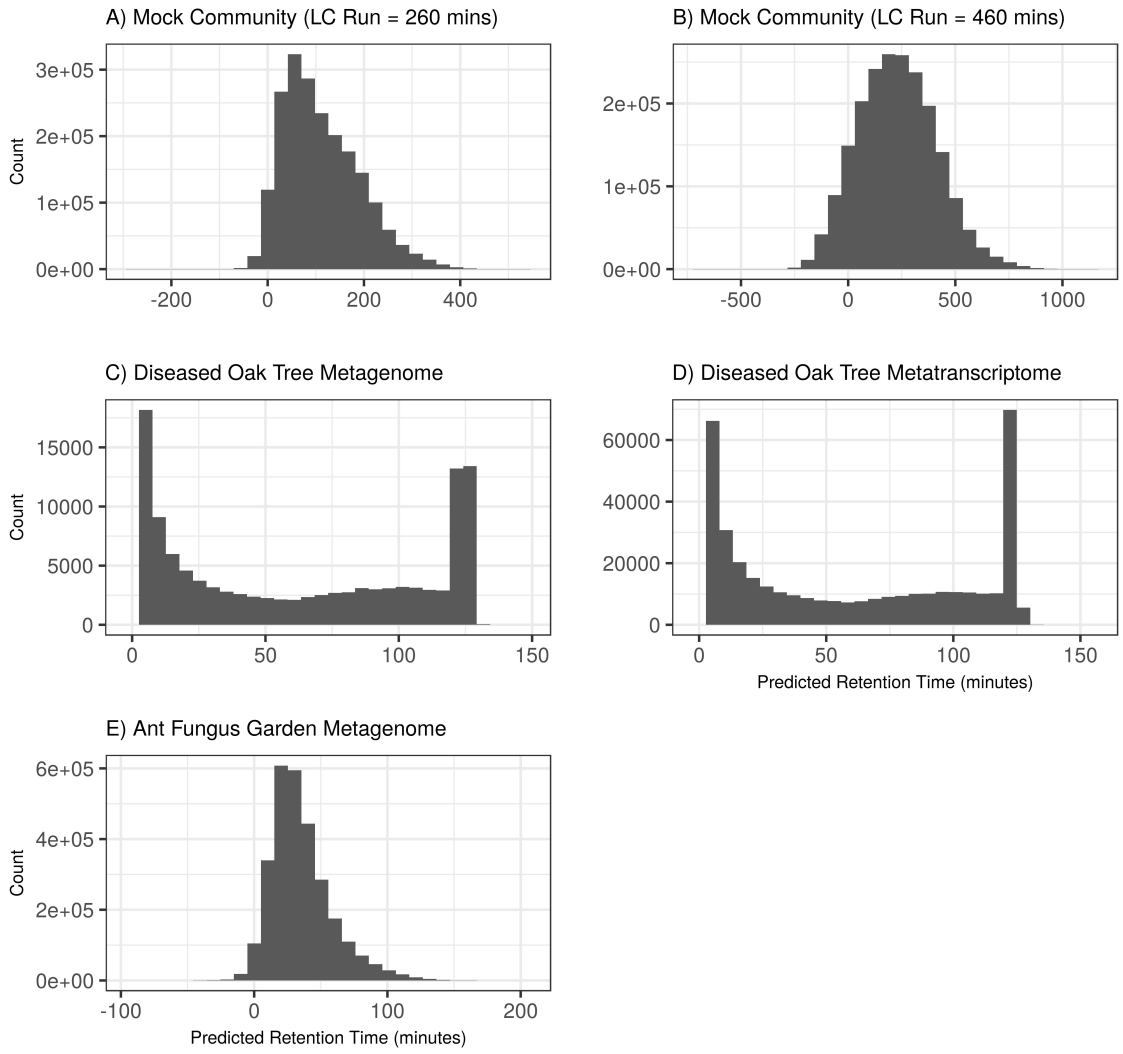

Figure 2.8: Histograms of predicted retention times for all peptides within a database. For panels A, B, and E, RTModel and RTPredict were used and trained on data from the associated mass spectrometry experiments (OLIGO kernel presented here). For panels C and D, BioLCCC was used to predict retention times based only on LC column characteristics.

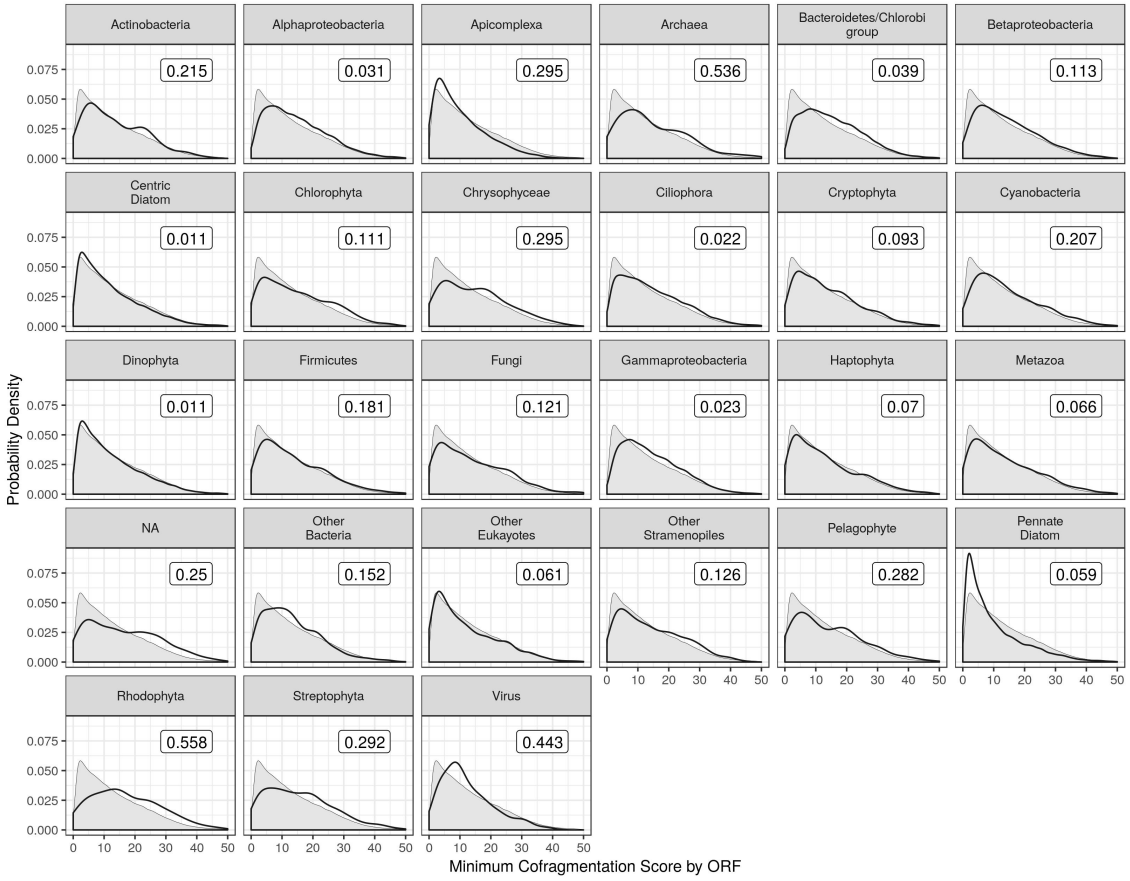

Figure 2.9: Distributions of minimum cofragmentation score by open reading frame from an Antarctic metatranscriptome. Each panel shows a cofragmentation score distribution for a given taxonomic group, with the grey background in each panel showing the cofragmentation score of all groups. Numbers within each panel show the Kullback-Liebler divergence value, representing the dissimilarity between the taxonomic group distribution and the overall distribution.

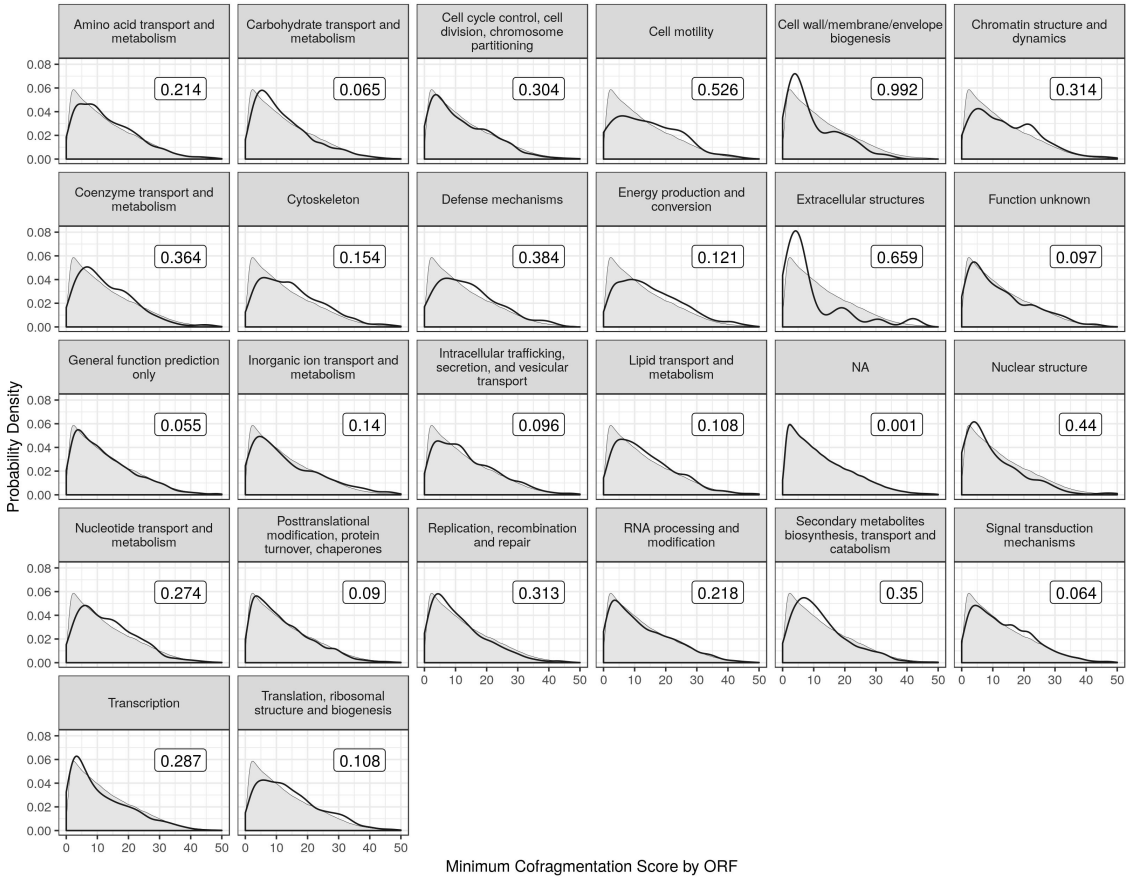

Figure 2.10: Distributions of minimum cofragmentation score by open reading frame from an Antarctic metatranscriptome. Each panel shows a cofragmentation score distribution for a given protein functional group (KOG class), with the grey background in each panel showing the cofragmentation score of all groups. Numbers within each panel show the Kullback-Liebler divergence value, representing the dissimilarity between the KOG class distribution and the overall distribution.

CHAPTER 3

PROTEOMIC TRAITS VARY ACROSS TAXA IN A COASTAL ANTARCTIC PHYTOPLANKTON BLOOM

This work was published previously in *The ISME Journal* (McCain, Allen and Bertrand, 2021).

3.1 Abstract

Production and use of proteins is under strong selection in microbes, but it's unclear how proteome-level traits relate to ecological strategies. We identified and quantified proteomic traits of eukaryotic microbes and bacteria through an Antarctic phytoplankton bloom using *in situ* metaproteomics. Different taxa, rather than different environmental conditions, formed distinct clusters based on their ribosomal and photosynthetic proteomic proportions, and we propose that these characteristics relate to ecological differences. We defined and used a proteomic proxy for regulatory cost, which showed that SAR11 had the lowest regulatory cost of any taxa we observed at our summertime Southern Ocean study site. Haptophytes had lower regulatory cost than diatoms, which may underpin haptophyte-to-diatom bloom progression in the Ross Sea. We were able to make these proteomic trait inferences by assessing various sources of bias in metaproteomics, providing practical recommendations for researchers in the field. We have quantified several proteomic traits (ribosomal and photosynthetic proteomic proportions, regulatory cost) in eukaryotic and bacterial taxa, which can then be incorporated into trait-based models of microbial communities that reflect resource allocation strategies.

3.2 Introduction

Microbes are constantly faced with an optimization problem: which proteins should be produced, when, and how many? The solutions to this problem dictate metabolic rates, cell stoichiometry, and taxonomic distribution (Reimers et al., 2017; Toseland et al., 2013; Twining and Baines, 2013; Saito et al., 2014; Morris et al., 2010). Yet, it's unclear what these solutions actually are in terms of proteome composition, and if different microbes have arrived at different solutions. Microbes are typically compared based on their unique repertoires of potential proteins (e.g. Kashtan et al., 2014; Giovannoni, Cameron Thrash and Temperton, 2014; Braakman, Follows and Chisholm, 2017), but taxa have shared proteins as well – are these shared proteins produced in similar amounts? Or, do taxa produce distinct amounts under identical conditions? Diverse taxa produce proteins in strikingly similar ratios within some pathways (Lalanne et al., 2018), but is stoichiometry conserved between pathways? The answers to these questions will direct future efforts for modelling microbial communities. Perhaps microbes can be represented as collections of genes (Reed et al., 2014; Coles et al., 2017), or, perhaps variation in proteome composition will shed light on the underpinnings of their ecological strategies and biogeochemical contributions.

Ecological strategies are ultimately tied to cellular functions and thus gene expression (Scott et al., 2010), and models can experimentally test hypotheses to evaluate such connections. Material models (i.e. cultures) have clearly demonstrated that selection acts strongly on protein production (Dekel and Alon, 2005; Parker et al., 2020; O'Malley and Parke, 2018). While powerful, these approaches are limited to only a few culturable organisms, which can overlook core differences found in less-studied organisms (e.g. Johnson et al., 2020). Computational models have also characterized trade-offs and metabolic behaviours in microbes (e.g. Molenaar et al., 2009; Faizi et al., 2018; Jahn et al., 2018). While models are critical from a reductionist perspective, characterization and prediction of microbial activity in their environments remains a central research goal.

Observing and measuring gene expression in microbes *in situ* can also link resource allocation to ecological strategies (e.g. Hu et al., 2018; Gifford et al., 2013; Alexander et al., 2015b,a; Morris et al., 2010; Sowell et al., 2009). For example, diatom and haptophyte transcriptional dynamics reflect their distinct growth strategies, inferred using metatranscriptomics (Alexander et al., 2015b,a). Metaproteomics has similarly identified

increased abundance of transporter proteins across an oceanographic gradient of decreasing nutrients (Morris et al., 2010). Both of these meta-omic approaches can quantify *in situ* resource allocation, but proteins cost more to produce and therefore better reflect resource allocation (Russell and Cook, 1995). To our knowledge, metaproteomics has not been used to quantify variation in resource allocation strategies across microbial groups.

Our objective was to identify and quantify proteomic "traits" for various eukaryotes and bacteria, by examining microbial proteome composition through a four-week time series at the Antarctic sea ice edge. We define a proteomic trait as a characteristic of an organism at the proteome-level, that includes both the abundance and identity of a protein (or group of proteins), and is connected to organismal fitness or performance (McGill et al., 2006). Metaproteomics is confronted by several methodological issues and biases, which we rigorously assess in order to characterize these proteomes. We subsequently provide practical recommendations for researchers using metaproteomics to examine microbial resource allocation. Our analyses suggest examining "coarse-grained" proteomes provides a host of conceptual and technical advantages (coarse-grained defined as a grouping of functionally or taxonomically (Phylum, Class, Order) related proteins). Next we use this approach to connect proteomic resource allocation to the ecology of these plankton. Lastly, we suggest that characterizing coarse-grained proteomes may be useful for assessing nutrient deficiency in the ocean.

3.3 Methods

3.3.1 Field Sampling

We collected samples once per week over four weeks at the Antarctic sea ice edge, in McMurdo Sound, Antarctica (December 28, 2014 "GOS-927"; January 6 "GOS-930", 15 "GOS-933", and 22 "GOS-935", 2015; as previously described in Wu et al., 2019). Sea water (150–250 L) was pumped sequentially through three filters of decreasing size (3.0, 0.8, and 0.1 μm, 293 mm Supor filters). Separate filter sets were acquired for metagenomic, metatranscriptomic, and metaproteomic analyses, over the course of $\sim$3 h, each week (36 filters in total). Filters for nucleic acid analyses were preserved with a sucrose-based buffer (20 mM EDTA, 400 mM NaCl, 0.75 M sucrose, 50 mM Tris-HCl, pH 8.0) with RNAlater (Life Technologies, Inc.). Filters for protein analysis were preserved in the same sucrose-based buffer but without RNAlater. Filters were flash frozen in liquid nitrogen in

the field and subsequently stored at -80 °C until processed in the laboratory.

3.3.2 Metagenomic and Metatranscriptomic Sequencing

We used metagenomics and metatranscriptomics to obtain reference databases of potential proteins for metaproteomics. We additionally used a database assembled from a similarly processed metatranscriptomic incubation experiment (Jabre et al., 2021), conducted with source water from the January 15, 2015 time point (these samples were collected on a 0.2 μm Sterivex filter and processed as previously described).

For samples from the GOS-927, GOS-930, GOS-933 and GOS-935 filters, RNA was purified from a DNA and RNA mixture (Rusch et al., 2007). 2 μg of the DNA and RNA mixture was treated with 1 μl of DNase (2 U/μl; Turbo DNase, TURBO DNase, ThermoFisher Scientific), followed by processing with an RNA Clean and Concentrator kit (Zymo Research). An Agilent TapeStation 2200 was used to observe and verify the quality of RNA. 200 ng of total RNA was used as input for rRNA removal using Ribo-Zero (Illumina) with a mixture of plant, bacterial, and human/mouse/rat Removal Solution in a ratio of 2:1:1. An Agilent TapeStation 2200 was used to subsequently observe and verify the quality of rRNA removal from total RNA. rRNA-deplete total RNA was used for cDNA synthesis with the Ovation RNA-Seq System V2 (TECAN, Redwood City, USA). DNA was extracted for metagenomics from the field samples (GOS-927, GOS-930, GOS-933 and GOS-935) according to Rusch et al. (2007). RNase digestion was performed with 10 μl of RNase A (20 mg/ml) and 6.8 μl of RNase T1 (1000 U/μl), which were added to 2 μg of genomic DNA and RNA mixture in a total volume of 100 μl, followed by 1 hour incubation at 37 °C and subsequent ethanol precipitation in -20 °C overnight.

Samples of double stranded cDNA and DNA were fragmented using a Covaries E210 system with the target size of 400 bp. 100 ng of fragmented cDNA or DNA was used as input into the Ovation Ultralow System V2 (TECAN, Redwood City, USA), following the manufacturer's protocol. Ampure XP beads (Beckman Coulter) were used for final library purification. Library quality was analyzed on a 2200 TapeStation System with Agilent High Sensitivity DNA 1000 ScreenTape System (Agilent Technologies, Santa Clara, CA, USA). 12 DNA and 18 cDNA libraries were combined into two pools with concentration 4.93 ng/µl and 4.85 ng/µl respectively. Resulting library pools were subjected to 1 lane of 150 bp paired-end HiSeq 4000 sequencing (Illumina). Prior to sequencing, each library was spiked with 1% PhiX (Illumina) control library. Each lane of sequencing resulted in

between 106,000 Mbp and 111,000 Mbp total and 6,900 Mbp – 12,000 Mbp and 4,800 Mbp - 6,900 Mbp for individual DNA or cDNA libraries respectively.

3.3.3 Metagenomic and Metatranscriptomic Bioinformatics

Metagenomic and metatranscriptomic data were annotated with the same pipelines. Briefly, adapter and primer sequences were filtered out from the paired reads, and then reads were quality trimmed to Phred33. rRNA reads were identified and removed with riboPicker (Schmieder, Lim and Edwards, 2011). We then assembled reads into transcript contigs using CLC Assembly Cell, and then we used FragGeneScan to predict open reading frames (ORFs; Rho, Tang and Ye, 2010). ORFs were functionally annotated using Hidden Markov models and blastp against PhyloDB (Bertrand et al., 2015). Annotations which had low mapping coverage were filtered out (less than 50 reads total over all samples), as were proteins with no blastp hits and no known domains. For each ORF, we assigned a taxonomic affiliation based on Lineage Probability Index taxonomy (Podell and Gaasterland, 2007; Bertrand et al., 2015). Taxa were assigned using two different reference databases: NCBI nt and PhyloDB (Bertrand et al., 2015). Unless otherwise specified, we used taxonomic assignments from PhyloDB, because of the good representation of diverse marine microbial taxa.

ORFs were clustered by sequence similarity using Markov Clustering (MCL; Enright, Van Dongen and Ouzounis, 2002). Sequences were assigned MCL clusters by first running blastp for all sequences against each other, where the query was the same as the database. The MCL algorithm was subsequently used with the input as the matrix of E-values from the blastp output, with default parameters for the MCL clustering. MCL clusters were then assigned consensus annotations based on KEGG, KO, KOG, KOG class, Pfam, TIGRfam, EC, GO, annotation enrichment (Jabre et al., 2021; Bertrand et al., 2015; Kanehisa and Goto, 2000; Kanehisa et al., 2016; Tatusov et al., 2003; Mistry et al., 2021; Haft, Selengut and White, 2003). Proteins were assigned to coarse-grained protein pools (ribosomal and photosynthetic proteins) based on these annotations. For assignment, we used a greedy approach, such that a protein was assigned a coarse-grained pool if at least one of these annotation descriptions matched our search strings (we also manually examined the coarse grains to ensure there were no peptides that mapped to multiple coarse-grained pools). For photosynthetic proteins, we included light harvesting proteins, chlorophyll a-b binding proteins, photosystems, plastocyanin, and flavodoxin. For ribosomal proteins, we just

included the term "ribosom*" (where the * represents a wildcard character), and excluded proteins responsible for ribosomal synthesis.

3.3.4 Sample Preparation and LC-MS/MS

We extracted proteins from the samples by first performing a buffer exchange from the sucrose-buffer to an SDS-based extraction buffer, after which proteins were extracted from each filter individually (as previously described Wu et al., 2019). After extraction and acetone-based precipitation, we prepared samples for liquid chromatography tandem mass spectrometry (LC-MS/MS). Precipitated protein was first resuspended in urea (100 μL, 8 M), after which we measured the protein concentration in each sample (Pierce BCA Protein Assay Kit). We then reduced, alkylated, and enzymatically digested the proteins: first with 10 μL of 0.5 M dithiothreitol for reduction (incubated at 60 °C for 30 minutes), then with 20 μL of 0.7 M iodoacetamide (in the dark for 30 minutes), diluted with ammonium bicarbonate (50 mM), and finally digested with trypsin (1:50 trypsin:sample protein). Samples were then acidified and desalted using C-18 columns (described in detail in McCain et al., 2021).

To characterize each metaproteomic sample, we employed one-dimensional liquid chromatography coupled to the mass spectrometer (VelosPRO Orbitrap, Thermo Scientific, San Jose, California, USA; detailed in McCain et al., 2021). For each injection, protein concentrations were equivalent across sample weeks, but different across filter sizes. We had higher amounts of protein on the largest filter size (3.0 μm) and less on the smaller filters, so we performed three replicate injections per 3.0 μm filter sample, and two replicate filter injections for 0.8 and 0.1 μm filters. We used a non-linear LC gradient totaling 125 minutes. For separation, peptides eluted through a 75 μm by 30 cm column (New Objective, Woburn, MA), which was self-packed with 4 μm, 90 A, Proteo C18 material (Phenomenex, Torrance, CA), and the LC separation was conducted with a Dionex Ultimate 3000 UHPLC (Thermo Scientific, San Jose, CA).

3.3.5 LC-MS/MS Bioinformatics – Database Searching, Configuration, and Quantification

Metaproteomics requires a database of potential protein sequences to match observed mass spectra with known peptides. Because we had sample-specific metagenome and metatranscriptome sequencing for each metaproteomic sample, we assessed various database

configurations, including those that we predict would be suboptimal, to examine potential options for future metaproteomics researchers. We used five different configurations, described below. In each case, we appended a database of common contaminants (Global Proteome Machine Organization common Repository of Adventitious Proteins). We evaluated the performance of different database configurations based on the number of peptides identified (using a peptide false discovery rate of 1%).

In order to make these databases (Table 3.1), we performed three separate assemblies on 1) the metagenomic reads (from samples GOS-927, GOS-930, GOS-933 and GOS-935), 2) metatranscriptomic reads (from samples GOS-927, GOS-930, GOS-933 and GOS-935) and 3) metatranscriptomic reads from a concurrent metatranscriptomic experiment, started at the location where GOS-933 was taken (Jabre et al., 2021). Database configurations were created by subsetting from these assemblies. The first configuration was "one-sample database", constructed to represent the scenario where only one sample was used for metagenomic and metatranscriptomic sequencing (we chose the first sampling week). Specifically, this was done by subsetting and including ORFs from the metagenomic and metatranscriptomic assemblies if reads from this time point were present in that sample (reads mapped as in Jabre et al., 2021), and then removing redundant protein sequences (P. Wilmarth, fasta utilities). The second configuration was the "sample-specific database", where each metaproteomic sample had one corresponding database (prepared from both metagenome and metatranscriptome sequencing completed at the same sampling site), also done by subsetting ORFs from the metagenomic and metatranscriptomic assemblies as described above. The third configuration was pooling databases across size fractions – such that all metagenomic and metatranscriptomic sequences across the same filter sizes (e.g. 3.0 μm) were combined. ORFs were subsetted from the metagenomic and metatranscriptomic assemblies as above. The fourth and fifth configurations were from the concurrent metatranscriptomic experiment (Jabre et al., 2021). The fourth configuration ("metatranscriptome experiment (T0)") was the metatranscriptome of the *in situ* microbial community (i.e. at the beginning of the experiment). This database was created by subsetting from the "metatranscriptome experiment (all)" assembly. Finally, the fifth configuration was the metatranscriptome of all experimental treatments pooled together (two iron levels, three temperatures; "metatranscriptome experiment (all)"). The overlap between databases (potential tryptic peptides) in different samples is presented graphically

Database Configuration	Filter Size	Number of Protein Sequences in the Database*
One-Sample Database	0.1	664521
One-Sample Database	0.8	642132
One-Sample Database	3	334394
Sample-Specific Databases	0.1	713153*
Sample-Specific Databases	0.8	633756*
Sample-Specific Databases	3	440990*
Pooled-across-sizes Databases	0.1	836620
Pooled-across-sizes Databases	0.8	855430
Pooled-across-sizes Databases	3	723057
Metatranscriptome Experiment (T = 0)	0.2	443681
Metatranscriptome Experiment (all)	0.2	2185747

Table 3.1: Characteristics of the five different database configurations we used for metaproteomic database searches. For the "One-Sample Database", the first time point was used, and all samples were matched according to filter sizes. For the "Sample-Specific Databases", each database was matched with the corresponding metaproteomic sample. For the "Pooled-Across-Sizes Databases", databases were pooled across every time point and matched according to filter size. For these aforementioned databases, the metagenomic and metatranscriptomic protein coding sequences were pooled. For the "Metatranscriptome Experiment (T = 0)", only the first sampling point from the metatranscriptome experiment was included. For the "Metatranscriptome Experiment (all)" configuration, all protein coding sequences were included from the treatment outcomes as well as the T = 0. *Averages are presented for Sample Specific Databases

in Supplementary Figs. 3.5, 3.6, 3.7.

After matching mass spectra with peptide sequences for each database configuration (MSGF+ with OpenMS, with a 1% False Discovery Rate at the peptide level; Kim et al., 2014; Röst et al., 2016), we used MS1 ion intensities to quantify peptides. Specifically, we used the FeatureFinderIdentification approach, which cross-maps identified peptides from one mass spectrometry experiment to unidentified features in another experiment – increasing the number of peptide quantifications (Weisser and Choudhary, 2017). This approach requires a set of experiments to be grouped together (i.e. which samples should use this cross-mapping?). We grouped samples based on their filter sizes (including those samples that are replicate injections). First, mass spectrometry runs within each group were aligned using MapAlignerIdentification (Weisser et al., 2013), and then FeatureFinderIdentification was used for obtaining peptide quantities.

After peptides were identified and quantified, we mapped them to proteins or MCL clusters of proteins, which have corresponding functional annotations (KEGG, KO, KOG, Pfams, TIGRFAM; Jabre et al., 2021; Bertrand et al., 2015; Kanehisa and Goto, 2000; Kanehisa et al., 2016; Tatusov et al., 2003; Mistry et al., 2021; Haft, Selengut and White,

2003). Functional annotations were used in three separate analyses. 1) Exploring the overall functional changes in microbial community metabolism, we mapped peptides to MCL clusters – groups of proteins with similar sequences. These clusters have consensus annotations based on the annotations of proteins found within the clusters (described in detail in Jabre et al., 2021). For this section, we only used peptides that uniquely map to MCL clusters. 2) We restricted the second analysis to two protein groups: ribosomal and photosynthetic proteins. For this analysis, we mapped peptides to one of these protein groups if at least one annotation mapped to the protein group (via string matching with keywords). This approach is "greedy" because does not exclude peptides if they also correspond with other functional groupings, but this is necessary because of the difficulties in comparing various annotation formats. 3) The last analysis for functional annotations was for targeted proteins, and we only mapped functions to peptides where the peptides uniquely identify a specific protein (e.g. plastocyanin).

Code for the database setup and configuration, database searching, and peptide quantification is open source (https://github.com/bertrand-lab/ross-sea-meta-omics).

3.3.6 LC-MS/MS Bioinformatics – Normalization

Normalization is an important aspect of metaproteomics: it influences all inferred peptide abundances. Typically, the abundance of a peptide is normalized by the sum of all identified peptide abundances. We use the term normalization factor for the inferred sum of peptide abundances. Note that the apparent abundance of observed peptides is dependent on the database chosen. In theory, if fewer peptides are observed because of a poorly-matching database, this will decrease the normalization factor, and those peptides that are observed will appear to increase in abundance. It is not known how much this influences peptide quantification in metaproteomics.

For each database configuration, we separately calculated normalization factors. We then correlated the sum of observed peptide abundances with each other. To get a database-independent normalization factor, we used the sum of total ion current (TIC) for each mass spectrometry experiment (using pyopenms; Röst et al., 2014), and also examined the correlation with database-dependent normalization factors. If normalization factors are highly correlated with each other, that would indicate database choice does not impact peptide quantification. Using TIC for normalization may have drawbacks, particularly if there are differences in contamination, or amounts of non-peptide ions across samples.

3.3.7 Defining Proteomic Mass Fraction

Protein abundance can be calculated in two ways: 1) the number of copies of a protein (independent of a proteins' mass), or 2) the total mass of the protein copies (the sum of peptides). We refer to the latter as a proteomic mass fraction. For example, to calculate a diatom-specific, ribosomal mass fraction, we sum all peptide abundances that are diatom- and ribosome-specific, and divide by the sum of peptide abundances that are diatom-specific. Note that this is slightly different to other methods, like the Normalized Spectral Abundance Factor, which normalizes for total protein mass (via protein length; Zybailov et al., 2006).

3.3.8 Combining Estimates across Filter Sizes

Organisms should separate according to their sizes when using sequential filtration with decreasing filter pore sizes. In practice, however, organisms can break because of pressure during filtration, and protein is typically present for large phytoplankton on the smallest filter size and vice-versa. We used a simple method for combining observations across filter sizes, weighted by the number of observations per filter. We begin with the abundance of a given peptide, which was only considered present if it was observed across all injections of the same sample. We calculated the sum of observed peptide intensities (i.e. the normalization factor), and divided all peptide abundances by this normalization factor. Normalized peptide abundances are then averaged across replicate injections. If we are estimating the ribosomal mass fraction of the diatom proteome, we first normalize the diatom-specific peptide intensities as a proportion of diatom biomass (i.e. divide all diatom-specific peptides by the sum of all diatom-specific peptides). We then summed all diatom-normalized peptide intensities that are unique to both diatoms and ribosomal proteins, which would give us the ribosomal proportion of the diatom proteome. Yet, we typically would obtain multiple estimates of, for example, ribosomal mass fraction of diatoms, on different filters. We combined the three values by multiplying each by a coefficient that represents a weight for each observation (specific to a filter size). These coefficients sum to one, and are calculated by summing the total number of peptides observed at a time point for a filter, and dividing by the total number of peptides observed across filters (but within each time point). For example, if we observed 100 peptides that are diatom- and ribosome-specific, and 90 of these peptides were on the 3.0 μm filter and only 10 were on the 0.8 μm filter, we would multiply the 3.0 μm filter estimate by 0.9

and the 0.8 μm filter by 0.1. This method uses all available information about proteome composition across different filter sizes (similar to Dupont et al., 2015).

When we estimate the proteomic mass fraction of a given protein pool, we do not need to adjust for the total protein on each filter. This is because this measurement is independent of total protein. However, for merging estimates of total relative abundance of different organisms across filters, we needed to additionally weight the abundance estimate by the amount of protein on each filter. Therefore, in addition to the weighting scheme described above, we multiplied taxon abundance estimates by the total protein on each filter divided by the total protein across filters on a given day.

3.3.9 LC-MS/MS Simulation

We used simulations of metaproteomes and LC-MS/MS to 1) quantify biases associated with inferring coarse-grained proteomes from metaproteomes, and 2) to mitigate these biases in our inferences. Specifically, we asked the question: how does sequence diversity impact quantification of coarse-grained proteomes from metaproteomes? Consider a three organism microbial community. If two organisms are extremely similar, there will be very few peptides that can uniquely map to those organisms, resulting in underestimated abundance. The third organism would also be underestimated, but to a lesser degree, unless it had a completely unique set of peptides. A similar outcome is anticipated with differences in sequence diversity across protein groups, such that highly conserved protein groups will be underestimated.

Our mass spectrometry simulations offer a unique perspective on this issue: we know the "true" metaproteome, and we can compare this with an "inferred" metaproteome. We simulated variable numbers of taxonomic groups, each with different protein pools of variable sequence diversity. From this simulated metaproteome, we then simulated LC-MS/MS-like sampling of peptides. Complete details of the mass spectrometry simulation are available in McCain and Bertrand (2019) and the supplementary materials. The only difference between this model and that presented in McCain and Bertrand (2019) is here we include dynamic exclusion. The ultimate outcomes from these simulations were 1) identifying which circumstances lead to biased inferences about proteomic composition, and 2) determining the underpinnings of these biases.

3.3.10 Cofragmentation Bias Scores for Peptides

We recently developed a computational model ("cobia") that predicts a peptides' risk for interference by sample complexity (more specifically, by cofragmentation of multiple peptides; McCain and Bertrand, 2019). This study showed that coarse-grained taxonomic and functional groupings are more robust to bias, and that this model can also be used to estimate bias. We ran cobia with the sample-specific databases, which produces a "cofragmentation score" – a measure of risk of being subject to cofragmentation bias. Specifically, the retention time prediction method used was RTPredict (Pfeifer et al., 2007) with an "OLIGO" kernel for the support vector machine. The parameters for the model were: 0.008333 (maximum injection time); 3 (precursor selection window); 1.44 (ion peak width); and 5 (degree of sparse sampling). Code for running this analysis, as well as the corresponding input parameter file, is found at https://github.com/bertrand-lab/ross-sea-meta-omics.

3.3.11 Description of Previously Published Datasets Analyzed

We leveraged several previously published datasets to compare our metaproteomic results. Specifically, we used proteomic data of phytoplankton cultures of *Phaeocystis antarctica* and *Thalassiosira pseudonana* (Wu et al., 2019; Nunn et al., 2013), and of cultures of *Escherichia coli* under 22 different culture conditions (Schmidt et al., 2016). Coarse-grained proteomic estimates were also compared with previously published targeted metaproteomic data (Wu et al., 2019).

3.4 Results and Discussion

We characterized proteomic traits of eukaryotic and bacterial taxa at the Antarctic sea ice edge. To do so, we have leveraged a combination of sample-specific nucleic acid sequencing and metaproteomics, assessing various assumptions and challenges with metaproteomics. Below, we first discuss our methodological results, and then we examine observations of different proteomic traits across microbial taxonomic groups. Finally, we touch on using coarse-grained protein pools for measuring nutrient stress in the ocean.

3.4.1 Database Choice Influences Peptide Identifications and Quantification

The sequence databases from the metatranscriptome experiment conducted on our third sampling week (January 15, 2015) outperformed sample-specific databases and other configurations (in terms of number of peptide spectrum matches, Supplementary Fig. 3.8; Table 3.1). Specifically, we identified 14 455 unique peptides using the "metatranscriptome experiment T0" database, while 8022 unique peptides were identified with the "sample-specific database" (Supplementary Fig. 3.8). We identified a core set of 5127 peptides, regardless of the database chosen (Supplementary Fig. 3.8). The database pooled across time points identified more peptides than the "sample-specific database", similar to previous work (Tanca et al., 2016). The metatranscriptomic experiment (both "metatranscriptomic experiment (T0)" and "metatranscriptomic experiment (all)") were more valuable in identifying larger, primarily eukaryotic organisms (Supplementary Fig. 3.8, 3.9, 3.10, 3.11). Overall, the two metatranscriptomic experiment databases performed similarly in terms of number of identified peptides. All subsequent analyses use the identified peptides from the "metatranscriptome experiment (all)" database. Importantly, a difference between the metatranscriptomes of sample-specific filters and the metatranscriptomic experiment databases was sequencing depth (Supplementary Table 3.1). This difference likely influenced the metatranscriptomic read assembly, improving the assembly of eukaryotic-protein sequences and therefore creating a better database (i.e. in terms of peptides identified). Note that databases were constructed after assembly, and then subsetted to create individual databases (Methods). Overall, deep metatranscriptomic sequencing appears to be a promising avenue for metaproteomics with tailored databases (Supplementary Table 3.1).

Database choice influenced peptide quantification due to normalization. We quantified this by correlating sample-specific normalization factors with each other and with the total ion current (i.e. a database-independent normalization, Supplementary Figs. 3.12, 3.13, 3.14). Examining the correlation between the best- and worst performing databases, there was a range of R2 values, from 83–99% (Supplementary Figs. 3.12, 3.13, 3.14). If we consider a peptide observed in a mass spectrometry experiment with an intensity value of 100, we expect variation in the inferred value to range from 92 to 108, reflecting variation in the normalization factor of 16%. This has significant consequences for comparative

metaproteomics: consider two samples, one with a perfectly matched database and a second that uses the same database, but is poorly matched. Using standard methods, the peptides identified in the second sample will appear to increase in abundance, even if the abundance is constant. We anticipate that database choice would similarly affect quantification when other mass spectrometry methods are employed (e.g. labelled untargeted metaproteomics, or experiments using data-independent acquisition). Note that our worst performing database was still well-matched to the community, so for researchers studying very distinct communities it is vital to address this issue by using database-independent normalization, or ensuring bias across samples is minimal. We provided methods for doing both.

One simple alternative is to use a database-independent metric of total peptide abundance: total MS1 ion current (TIC). We found that TIC is well correlated with the total peptide abundance inferred from the best performing database (with correlation coefficients 0.98, 0.95, and 0.91 for the 3.0, 0.8, and 0.1 μm filter sizes, Supplementary Fig. 3.12, 3.13, 3.14, respectively). This result has two consequences: 1) it suggested that TIC may be a viable alternative for normalization in comparative metaproteomics. 2) It validated the use of our best-performing database, as we identified most of the abundant peptides in our sample. Given that these two approaches were highly correlated, we used the "metatranscriptome experiment (all)" database for all subsequent analyses.

3.4.2 Taxonomic and Functional Composition Shifted through the Season at the Antarctic Sea Ice Edge

Taxonomic abundance shifted through the season at the Antarctic sea ice edge (Fig. 3.1a). The microbial community was dominated by *Phaeocystis antarctica* (Haptophyta) early in the season, with diatoms increasing in relative abundance later (predominantly *Fragilariopsis* sp. and *Pseudonitzschia* sp.). The phytoplankton bloom progression and high dinoflagellate biomass contribution were both consistent with previous observations in the Ross Sea (Smith et al., 2013; Andreoli et al., 1995). Bacterial taxa had relatively lower protein biomass and more consistent relative biomass values through time compared with eukaryotic taxa. Of the bacterial taxa we observed, Rhodobacterales was the most abundant group, with abundances being mostly stable though the season.

We identified shifts in protein abundance by mapping peptides to de novo protein clusters (irrespective of taxonomic assignments) – including protein clusters with no known function. Earlier in the season there was a high relative abundance of Chlorophyll

A-B binding proteins and ATP synthase alpha/beta family proteins (Fig. 3.1b), which is anticipated because of the higher levels of dissolved iron (Wu et al., 2019). Demonstrating the importance of de novo protein group assignment, the most abundant protein group in our entire dataset had no functional annotations (Fig. 3.1b, Unknown Protein Cluster 2818, mostly belonging to Ciliates). Further examination of a representative protein sequence within this cluster found no functionally similar proteins within the NCBI non-redundant database. We suggest that these unknown, highly abundant proteins should be targets for functional characterization.

3.4.3 Eukaryotic and Bacterial Taxa have Taxon-Specific Proteomic Allocation Strategies

We quantified two simple proteomic traits of microbes: the ribosomal protein mass fraction and the photosynthetic protein mass fraction (using a combined estimate across filter sizes, Supplementary Fig. 3.15). Eukaryotic taxa formed unique clusters based on these two traits, with more variation across taxa than across time points (Fig. 3.2a). For example, haptophytes had relatively high proportions of both ribosomal and photosynthetic protein fractions. Examining the five most abundant bacterial taxa, we also observed distinct proteomic compositions, with Gammaproteobacteria exhibiting the highest ribosomal protein mass fraction (Fig. 3.2b).

Before examining the underpinnings of these proteomic traits, we first scrutinized these inferences using mass spectrometry simulations and additional data sources. Our analysis was limited to coarse-grained protein functions and taxa, which is robust to bias arising from variable sample complexity (McCain and Bertrand, 2019). Our mass spectrometry simulations suggested that low sequence diversity in taxa or protein groups can lead to underestimation (Supplementary Fig. 3.16), but this bias is mitigated by examining abundant proteins or taxa. Identifying $\sim$50 peptides or more is evidence that there is sufficient sequence diversity in a protein group to avoid this type of underestimation (Supplementary Fig. 3.16). We identified greater than 50 taxon-specific peptides for each protein group, indicating that these observations are not subjected to significant biases arising from sequence diversity. We therefore restricted our analyses to taxa and protein groups that are relatively abundant. Note that for dinoflagellates, we observed relatively few peptides in the photosynthetic proteomic mass fraction, so our observations are likely underestimating the true value (Supplementary Discussion, Supplementary Fig. 3.16).

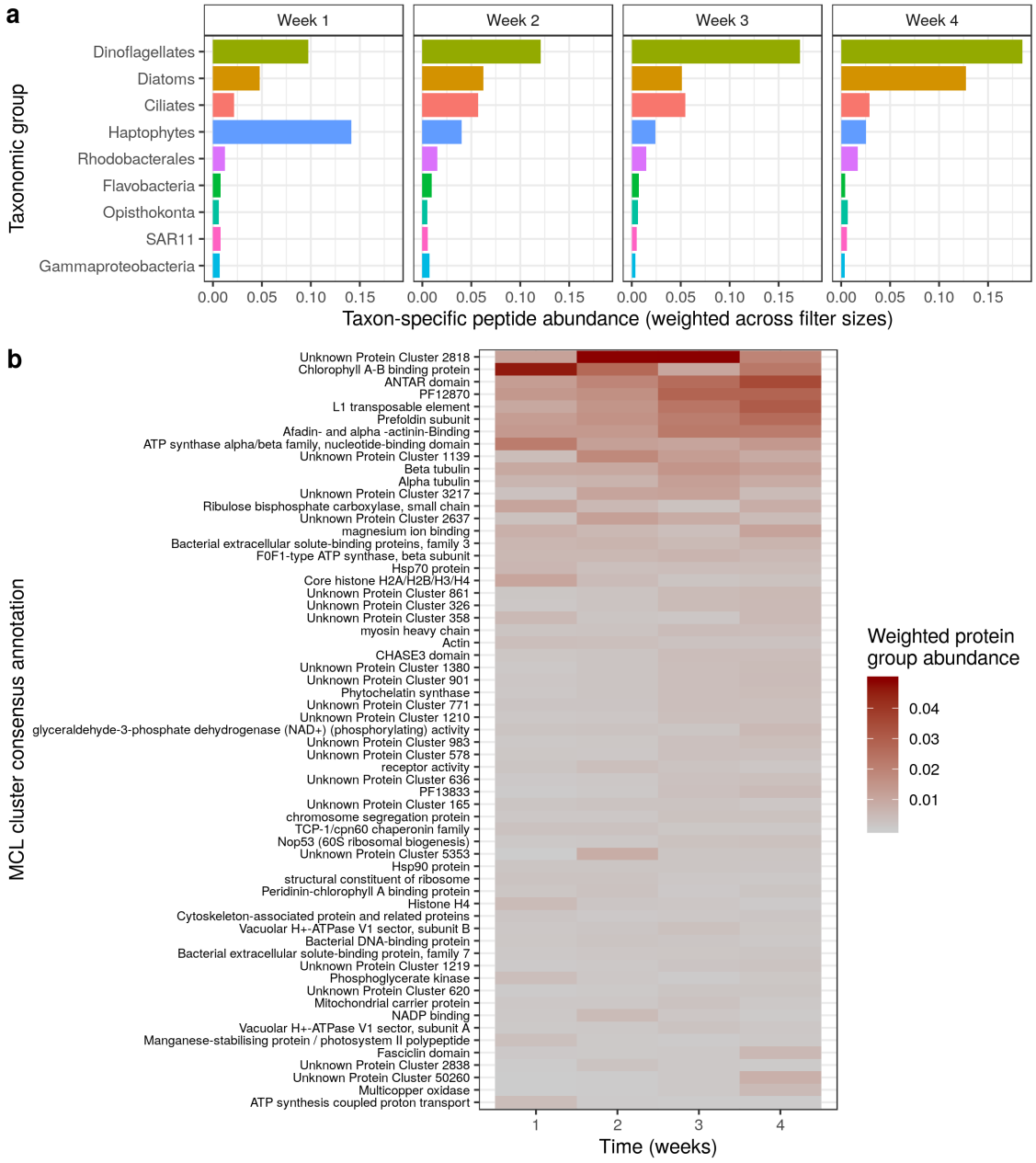

Figure 3.1: **a**, Measurements of relative change in protein biomass identified a taxonomic shift at the Antarctic sea ice edge. Protein biomass is calculated as the sum of taxon-specific peptide intensities, weighted by the protein mass per filter for each sampling time. **b**, relative change in protein functional clusters shows that unknown protein clusters contribute greatly to *in situ* protein biomass, and also identifies a functional shift across weeks.

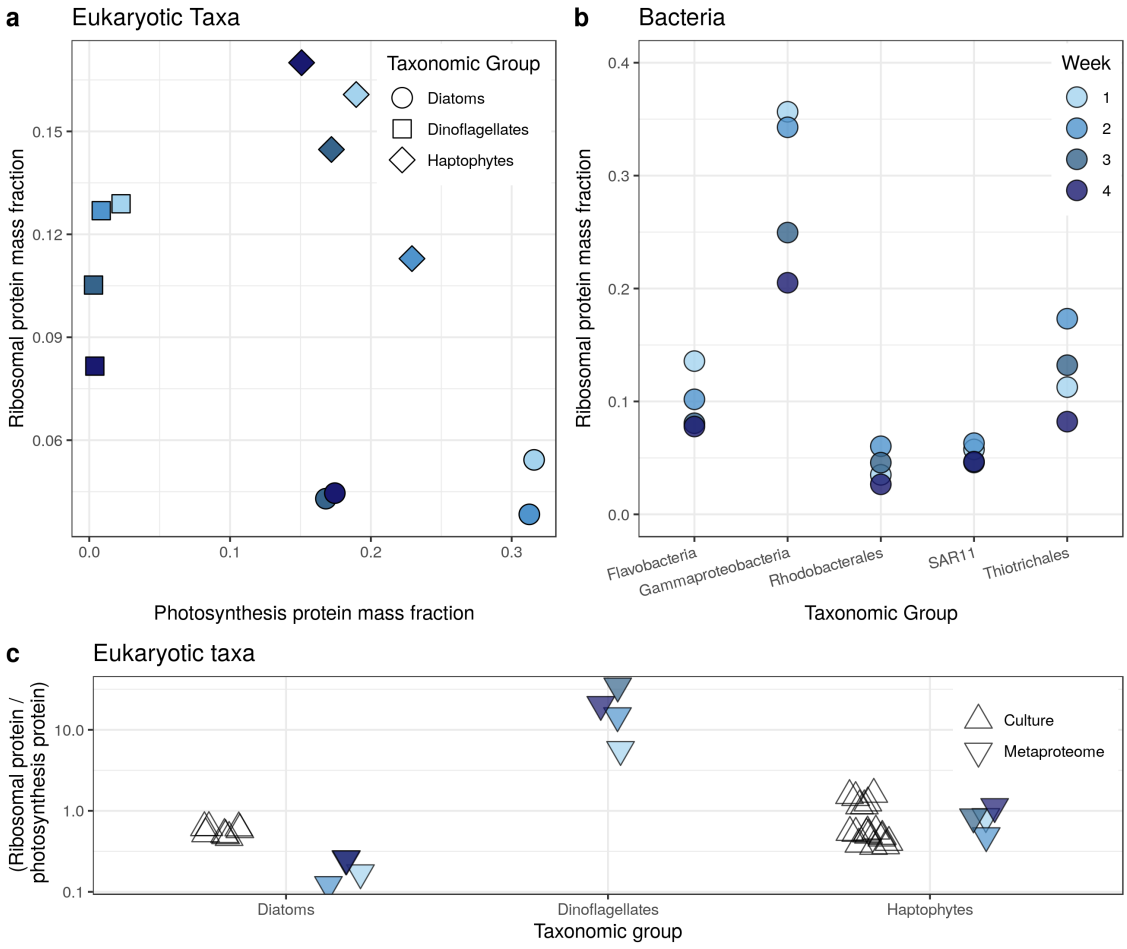

Figure 3.2: **a**, photosynthetic protein mass fraction and ribosomal protein mass fraction (normalized by total amount of a given taxon at a each time point) identifies clear taxonomic subgroupings. **b**, examining bacteria only shows variation in ribosomal mass fraction across groups. Note that Thiotrichales are an order within Gammaproteobacteria, so Gammaproteobacteria here refers to all non-Thiotrichales Gammaproteobacteria. **c**, Ratios of ribosomal protein mass fraction to photosynthetic protein mass fraction derived from metaproteomic observations and compared with phytoplankton proteomes observed in culture (*Phaeocystis antarctica*, *Thalassiosira pseudonana* Wu et al., 2019; Nunn et al., 2013).

Despite this underestimation, the true value is probably quite low (discussed below).

We provided two additional estimates of ribosomal and photosynthetic protein mass fraction from cultured phytoplankton (Fig. 3.2c). Metaproteomics can underestimate taxon-specific protein mass when taxonomically uninformative peptides are not used. For example, we might identify a highly conserved peptide produced by a diatom, but are unable to map it to diatoms because it also corresponds to other taxa, and this peptide would be excluded from the quantification of diatoms. Therefore, we compared the ratios of ribosomal to photosynthetic protein mass fraction from the metaproteomic observations to cultured diatoms and haptophytes. Ratios were similar in cultures compared to populations sampled *in situ* (Fig. 3.2c), despite such culturing experiments occurring under different environmental conditions. Trends observed in the proportion of transcripts mapped to ribosomal proteins in different groups of bacteria also mirrored our estimates of ribosomal protein mass fraction (high for Gammaproteobacteria and low for SAR11; Gifford et al., 2013).

We examined coarse-grained taxonomic groups. It is possible that within these coarse groupings, different taxa included in these groupings employ different allocation strategies. We therefore sought to determine whether taxonomic sub-groupings displayed similar expression patterns. This issue is challenging to assess, because as subgroupings are further examined, there is increased susceptibility to several biases (as outlined above). We therefore examined one subgrouping, diatoms, that contained two dominant species: *Fragilariopsis* sp. and *Pseudonitzschia* sp. The taxonomic assignments for these two diatoms were from the NCBI nt database. We observed similar proteome estimates for both ribosomal and photosynthetic proteins amongst both these subgroups of diatoms (Supplementary Fig. 3.17), suggesting they are functionally similar based on these proteomic traits. However, we cannot exclude the possibility that for other taxonomic groups the trends observed are due to a diversity of underlying microbial strategies. Yet at this coarse taxonomic level, we concluded that different microbial taxa exhibited distinct coarse-grained proteomes.

We now turn to the ecological relevance of these protein expression patterns. Protein synthesis is the primary energy sink in cells (Russell and Cook, 1995), and photosynthesis or respiration is the primary energy source in cells. Why do dinoflagellates have relatively low photosynthetic protein mass fractions? This taxonomic group is typically mixotrophic

or heterotrophic (Jeong et al., 2010), which would require larger investment in respiratory proteins for energy production. Haptophytes and diatoms had similar amounts of photosynthetic proteins, but very different amounts of ribosomal proteins (Fig. 3.2), so there was no direct trade-off between producing ribosomal versus photosynthetic machinery (i.e. they do not form Pareto front; Sheftel et al., 2013; Hart et al., 2015). Gammaproteobacteria had the highest ribosomal mass fraction within the observed bacterial taxa, and haptophytes had higher ribosomal mass fractions compared to diatoms. Gammaproteobacteria ribosomal mass fraction decreased through the season, perhaps corresponding with a decreased growth rate as micronutrients are depleted by the phytoplankton bloom.

What are the ecological implications of having more ribosomes? If we assume constant translation rate per translational apparatus (but see Dethlefsen and Schmidt, 2007), taxa then had different total protein synthesis output. Growth rate is directly related to total protein synthesis output, because protein comprises a large portion of cell mass. To have a faster growth rate, microbes' need to increase protein synthesis (see Scott et al., 2010, for derivation and assumptions). We hypothesize that high total protein synthesis output (via high ribosomes) is more advantageous under high nutrient regimes, as it would allow an elevated growth rate. Indeed, haptophytes and Gammaproteobacteria were more abundant earlier in the season (which had higher concentrations of dissolved Fe and Mn; Wu et al., 2019). Another interpretation is that these early-abundant taxa are better suited to a dynamic environment. Perhaps these early-abundant taxa (Gammaproteobacteria, haptophytes) increased investment in ribosomes as a form of bet hedging, which enables a faster growth rate in a dynamic environment (Mori et al., 2017).

3.4.4 Environment-Independent Proteomic Fraction Varies across Taxa

What is the cost of responding quickly to a dynamic environment? We hypothesized that there is a regulatory cost for producing proteins that are optimal for a set of environmental conditions. Constitutive protein production does not incur this regulatory cost at the risk of being mismatched to environmental conditions. If the proteome is mostly constant across conditions, this indicates a low regulatory cost, and vice versa. We propose a proteomic trait that reflects regulatory cost: the proteomic fraction that is environment-independent. This proteomic trait is quantifiable using metaproteomics, and due to the dynamic nature of the ocean, is likely an important selective force for marine microbes.

We classified peptides that are relatively constant across different environmental conditions, and then summed their average intensities to get an environment-independent peptide mass fraction (Fig 3.3b and c). Note that 1) peptide intensities were first normalized by total taxon-specific peptide intensity (they therefore sum to one for each taxon), and 2) estimates of environment-independent peptide mass fraction were combined across filter sizes. Using previously published proteomic data from replicate cultures of *E. coli* under identical conditions, we chose a cut-off point distinguishing environment-dependent versus -independent peptides (represented with vertical lines Fig. 3.3; Supplementary Fig. 3.18 Schmidt et al., 2016). This cut-off point was calculated by examining the distribution of protein-level coefficients of variation for each *E. coli* culture condition, determining the third quartile, and then taking the mean across all culture conditions (Schmidt et al., 2016). We then can determine the proportion of the proteome that is environment-independent and -dependent (using the mean abundance value per peptide). There are potential biases in this novel method. We address the impact of these biases using published data and by making comparisons with other estimates of regulatory costs across taxa from previously published work (see Supplementary Discussion, Supplementary Fig. 3.18, 3.19).

SAR11 had the highest environment-independent peptide mass fraction across all eukaryotic and bacterial taxa we examined (3.3a and b, Supplementary Fig. 3.20), consistent with previous work suggesting SAR11 has reduced regulatory investment (Giovannoni, 2017). Within eukaryotes, dinoflagellates exhibited the highest environment-independent peptide mass fraction, and dinoflagellates in other oceanic regions also exhibited lower regulatory cost (Alexander et al., 2015b; Hu et al., 2018).

Diatoms had a lower environment-independent proteomic fraction compared with haptophytes, suggesting they have higher regulatory costs. Recall the previous result that diatoms had a lower proportion of ribosomes compared with haptophytes (but similar proportions of photosynthetic proteins; Fig. 3.2a). We speculate that two proteomic traits comprise a trade-off for these two taxa: higher total protein synthesis via more ribosomes (i.e. leading to fast growth under high nutrient conditions), but at a cost of being less able to dynamically regulate their proteomes. This suggests that in a high nutrient environment (that is also dynamic), dynamically responding to the environment is not the optimal strategy. Instead, a better strategy is constitutively expressing proteins that are favourable for rapid growth (e.g. high ribosomal production in haptophytes).

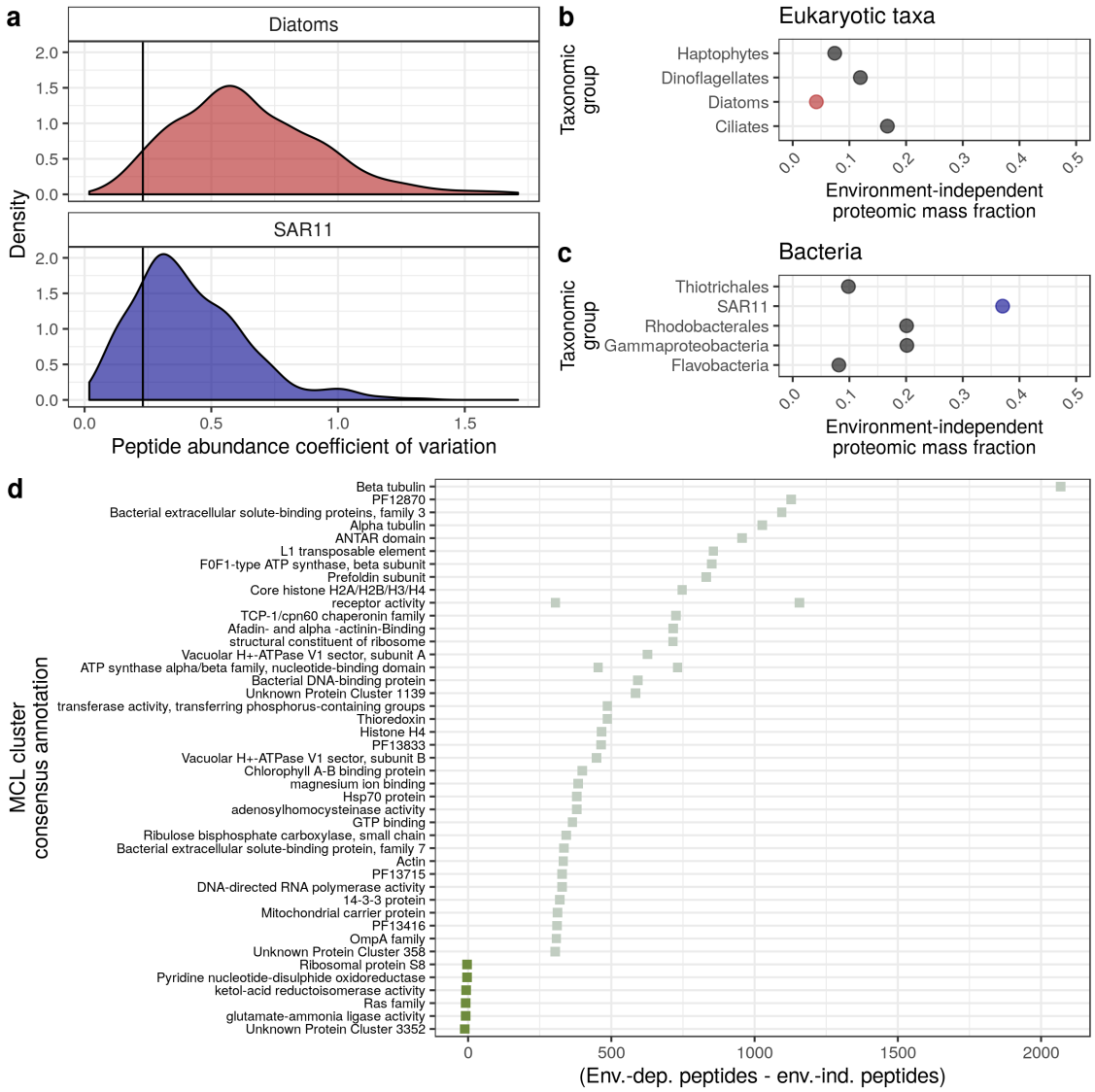

Figure 3.3: **a**, the distribution of peptide-specific coefficients of variation can be used to identify if a peptide is significantly changing across environmental conditions. Peptide abundance is first divided by the total taxon-specific peptide intensity. Diatoms and SAR11 represent two extremes within this dataset – diatoms have a highly variable proteome while SAR11 has a relatively constant protein expression. The cutoff point was chosen using replicate cultures of *E. coli* protein expression (vertical line Schmidt et al., 2016). **b**, **c**, after classifying peptides by their coefficients of variation, we categorized peptides as independent of their environment and those that are not. Points represent the sum of peptide intensities that are environment independent across eukaryotic and bacterial taxa. **d**, a comparison of protein functional clusters that have peptides classified mostly as environment-dependent or environment-independent. The values plotted are the number of environment-dependent peptides minus the number of environment-independent peptides. Note that some MCL clusters had the same consensus annotation, and therefore sometimes have multiple corresponding points. Positive values indicate that more peptides observed within this protein cluster were dependent on their environment (light green), while negative values indicate more peptides were identified as environment-independent (dark green).

The lower ribosomal mass fraction observed in diatoms would limit their growth in higher micronutrient environments but make them more successful in lower micronutrient environments. Ross Sea phytoplankton blooms typically progress from haptophyte- to diatom-dominated, as micronutrients stocks (e.g. Fe and Mn) transition from replete to deplete (Smith et al., 2013; Mangoni et al., 2017; Noble et al., 2013; Peloquin and Smith, 2007). There is also evidence that *Phaeocystis* has a higher Fe requirement (Sedwick et al., 2007), which may be related to these proteomic traits. We posit that differences in regulatory cost and ribosomal mass fraction between diatoms and haptophytes may help explain their ecological succession.

Are some protein functions more often categorized as environment-independent or environment-dependent? Highlighting some examples, the actin protein cluster was often classified as environment-dependent (Fig. 3.3c). Actin is involved in endocytosis, and inorganic Fe uptake occurs via an endocytotic mechanism (with phytotransferrin; McQuaid et al., 2018). Perhaps variable expression of actin is related to the amount of bioavailable Fe, and previously published proteomic experiments also showed that actin was differentially expressed due to Fe (Bertrand et al., 2012; Cohen et al., 2018). ATP synthase-peptides and chlorophyll A-B binding protein-peptides were also mostly classified as environment dependent, likely reflecting higher primary production earlier in the season (Fig. 3.3c). In contrast, the ketol-acid reductoisomerase protein cluster (involved in branched-chain amino acid synthesis) was mostly classified as environment-independent. It is unclear what the mechanistic basis for constitutive expression of this protein might be, but several proteomic studies of diatoms also suggest similar expression across conditions (Nunn et al., 2013; Bertrand et al., 2012; Cohen et al., 2018). Using this extensible approach to identify constitutively expressed proteins across a wide array of taxa would shed light on these mechanisms. With vastly more metaproteomic data being generated (e.g. Cohen et al., 2021), identifying constitutively expressed proteins across diverse taxa would help answer the question: what are the features of constitutively expressed proteins? For example, perhaps there are certain protein functional groupings that are often constitutively expressed.

3.4.5 Coarse-Grained Proteomes can Assess Nutrient Stress

Proteomics is also used in marine microbiology to assess stress corresponding to a deficient nutrient (e.g. Saito et al., 2014). For example, expression of the protein plastocyanin

may reflect Fe deficiency, because plastocyanin does not contain Fe and performs a similar function as the Fe-containing protein cytochrome c (Strzepek and Harrisson, 2004). Biomarkers of physiological stress are increasingly nuanced (e.g. Wu et al., 2019), sometimes taxon specific, and can require targeted mass spectrometry approaches. Coarse-grained approaches may be a complementary method for assessing stress or nutrient deficiency. We compared using coarse-grained proteomes with single-protein biomarkers. Previous bottle incubation work and targeted metaproteomics showed that there was a transition to Fe- and Mn-stress at this sampling location in the Ross Sea, so we focus on Fe-stress indicators (Wu et al., 2019). We first solely examined the photosynthetic protein mass fraction compared to the mass fraction of peptides assigned to the plastocyanin, for diatoms and haptophytes (Fig. 3.4a and b). This approach is biased by variable complexity across samples (McCain and Bertrand, 2019), but we predicted the degree of bias with a quantitative metric (the "cofragmentation score"). This score reflects the expected number of peptides with similar *m/z* and retention times. Overall, there were relatively few potential cofragmenting peptides ($\approx$3), indicating low bias (peptides with high bias can have upwards to 300 cofragmenting peptides, for example; McCain and Bertrand, 2019). We observed a negative relationship between the photosynthetic protein mass fraction and the plastocyanin mass fraction (note that these two variables are not independent, as plastocyanin is considered as part of the photosynthetic mass fraction). We also examined *Phaeocystis antarctica*-specific peptides measured with previously published targeted mass spectrometry, and identified a negative correlation between the abundance values of plastocyanin and the coarse-grained estimates of photosynthetic proteins (Fig. 3.4c). We conducted this analysis as a proof-of-concept for using coarse-grained proteomes to assess nutrient deficiency, as coarse-grained proteomes are amenable for untargeted metaproteomic analyses. These preliminary analyses suggest that coarse-grained proteome composition may be a useful tool for assessing nutrient deficiency. More analyses are required to assess the robustness of this relationship, and also to assess if coarse-grained proteomic signatures are nutrient specific (i.e. would a coarse-grained marker be able to distinguish between Fe and Mn stress?).

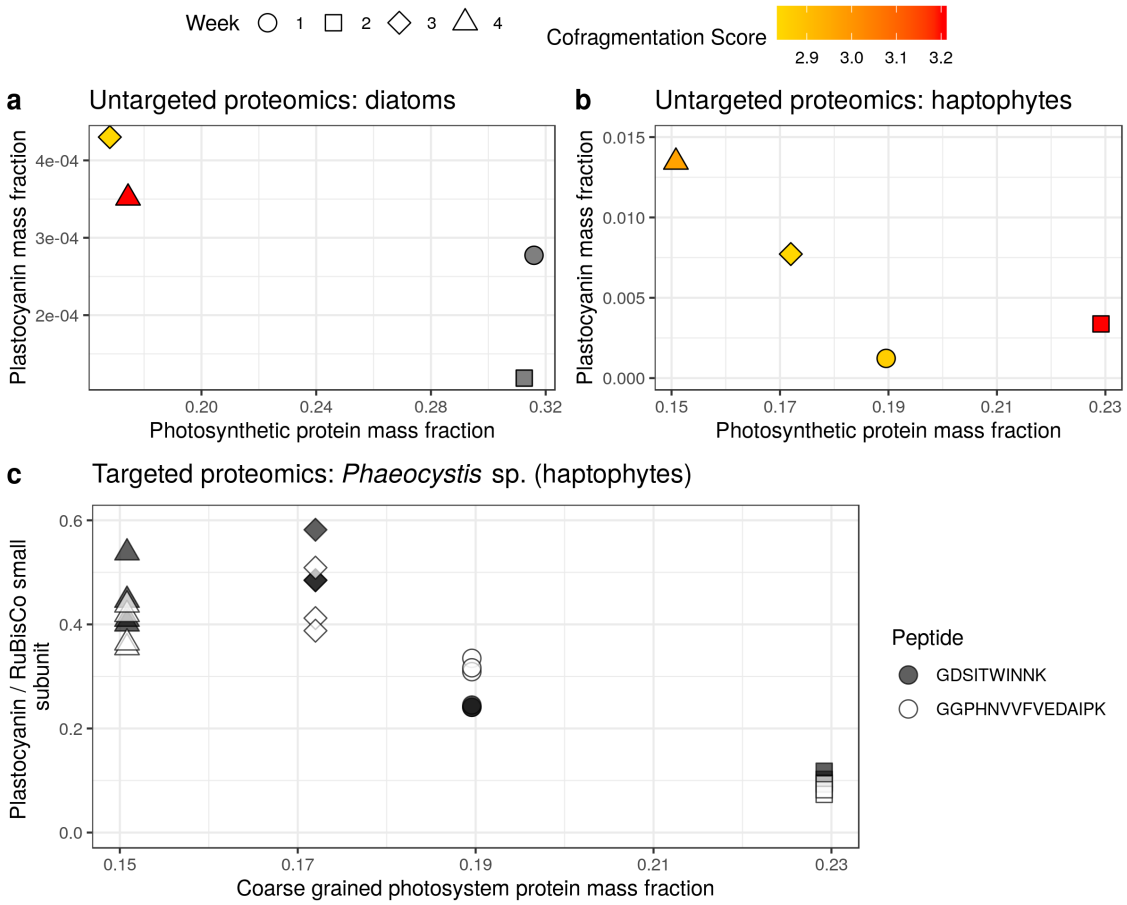

Figure 3.4: **a**-**b**, comparison of the single-protein biomarker plastocyanin with the photosynthetic protein mass fraction for diatoms and haptophytes (using discovery proteomics). Points are coloured with their corresponding, sample-specific cofragmentation score (the number of potentially cofragmenting peptides). Cofragmentation scores were calculated using the sample-specific nucleic acid sequencing, and points coloured in grey correspond to peptides that were identified and quantified with the "Metatranscriptome experiment (all)" database, but were not present in the sample-specific databases. **c**, comparison of the single-protein biomarker (using targeted proteomics) plastocyanin with the photosynthetic protein mass fraction for haptophytes. Two peptides for plastocyanin are shown, and each point represents one technical replicate measurement. *Phaeocystis* plastocyanin abundance is normalized to *Phaeocystis* RuBisCO small subunit abundance, where we used the mean of two taxon-specific peptides (AKPNFYVK and QIQYALNK) to calculate RuBisCO abundance (Wu et al., 2019).

3.5 Conclusion

We conclude that different microbial taxa have distinct coarse-grained proteomic composition, and this composition is more similar across taxa than across environmental conditions. The stoichiometry of proteins within pathways is conserved (Lalanne et al., 2018) – but our results show that this is not the case across pathways. Variation in pathway-to-pathway stoichiometry may indeed underpin ecological strategies, in addition to differing gene repertoires. Connecting *in situ* proteomes to ecological strategies will delineate proteomic traits, which can then be adopted into a trait-based approach for modelling microbial communities. Genomic trait-based approaches have successfully explained large-scale biogeochemical processes (Reed et al., 2014; Coles et al., 2017), but they first had to identify genes that are metabolically important. Therefore, identifying and quantifying proteomic trait variation across taxa will connect protein production to ecological strategies, and ultimately enable modelling of microbial communities by representing proteomic traits and trade-offs in large scale models (e.g. as in Follows et al., 2007).

3.6 Data Availability

The metagenomics and metatranscriptomics data reported here have been deposited in the NCBI sequence read archive (BioProject accession no. PRJNA074702; BioSample accession nos. SAMN18057468-SAMN18057479 (metagenomics) and BioSample accession nos. SAMN18057480-SAMN18057497 (metatranscriptomics). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD022995 (Perez-Riverol et al., 2019). All other data products (the cobia analysis output, formatted databases, peptide abundances for each database search, targeted proteomics data, culture proteomics data, metaproteomic simulation output) are available in Dryad at doi:10.5061/dryad.vt4b8gtrz.

3.7 Author Contributions

J.S.P.M., E.M.B., and A.E.A. conceived the study. J.S.P.M. wrote the code, conducted analyses and metaproteomic lab work, with input from E.M.B. and A.E.A. E.M.B. and A.E.A. collected the Antarctic samples and conducted metagenomic and metatranscriptomic lab

and computational work. J.S.P.M. wrote the paper with input from E.M.B. and A.E.A.

3.8 Supplementary Information

3.8.1 Supplementary Methods

We simulated metaproteomes *in silico* to examine biases arising from inferring taxon-specific proteomes. The primary challenge of inferring taxon-specific coarse-grained proteomes is that not all coarse-grained pools (groups of proteins performing some functional role) are equally identifiable. This also extends to taxa – some taxa are more closely related, and therefore have fewer unique peptides. For example, some coarse-grained pools are easily mapped to a given taxon while others have very few taxon-specific peptides.

To address these expected biases, we created in silico metaproteomic datasets (generative model), and sampled the data similar to how a mass spectrometer would (sampling model). We then compared the sampled data to the known dataset and evaluated which conditions biases would arise.

3.8.1.1 Generative Model

We generate p unique peptides, assigned to k coarse grained pools, belonging to an organism j. We simulate peptides rather than proteins, as peptides are injected into a mass spectrometer with bottom-up mass spectrometry. To simulate different levels of sequence diversity present across protein pools, we generate k sequence 'banks' of different sizes. Peptide sequences banks are created by randomly sampling from all amino acids, generating a sequence ranging in length from 5–15 amino acids per peptide. An organism-specific peptide profile is created, which randomly samples from each 'sequence bank'. So a smaller 'sequence bank' would represent a coarse grained protein pool with low sequence diversity, and vice-versa.

We then assign abundances to each peptide. Peptide abundance is generated using a random sample from a gamma distribution with the shape parameter of 0.15 and the scale parameter of 10. We chose this distribution as it is similar to the distribution of peptides observed in single-organism proteomics (specifically it has overdispersion, non-zero values only, and is continuous). We then multiply each peptide abundance by a taxonomic abundance unique to each taxon j, and by the abundance within a given coarse grained pool k. Both the taxonomic abundance and the coarse-grained pool abundance

values are similarly drawn from a gamma distribution, except with a shape value of 1. For example, to calculate peptide abundance we first draw a value for an organism abundance (e.g. 100) and multiply that by a value drawn for a coarse-grained pool abundance (e.g. 5). Lastly, we generate a value for all peptides from within this organism and coarse-grained pool (e.g. 2), and multiply these three values. In this case, the intensity of the peptide would be 1000. Once the 'true' dataset is generated, we then filter this dataset to create an 'observed' dataset, because peptides that are the same from the 'true' dataset should be summed. From this observed dataset, we calculate peptide mass and assume a peptide charge state of 2.

3.8.1.2 Sampling Model

Mass spectrometers sample and fragment peptides for identification, and there is some stochasticity in this sampling process, particularly when using data-dependent acquisition (DDA). Using DDA, peptides are sampled according to their intensity. We subsample our 'observed dataset' using a simplistic model of a mass spectrometer. Our model assumes a constant ion peak width, and randomly assigns elution times to peptides from a uniform distribution. A similar version of this model has been extensively validated (McCain and Bertrand, 2019), but the key difference here is including dynamic exclusion and top-N sampling.

We describe sampling model algorithmically below (Algorithm 1). We begin by sorting and then binning elution times for all peptides (steps 1–2). We then loop through every *n*th elution time bin, where n represents the number of ions selected for Top n DDA (step 4). So with more ions selected for 'fragmentation' the mass spectrometer would have less time to scan intact peptides, as is true for instruments that move between scanning MS1 in an Orbitrap and fragmenting peptides in a linear ion trap. Then, if a peptide is on the dynamic exclusion list and it has been on the list for longer than the dynamic exclusion time, it is removed from the dynamic exclusion list (step 4–5). All of the m/z windows belonging to a peptide on the dynamic exclusion list are then blocked for sampling (steps 6–7). Of the remaining peptides, we select the top n in terms of abundance, and assume that these peptides are identified (step 8). The final step is adding the identified peptides to the dynamic exclusion list (step 9), which prevents those m/z regions from being

subsequently sampled for a short period of time (the dynamic exclusion time).

Algorithm 1: Mass spectrometry sampling model.

Result: Sampling peptides generated from the observed dataset similar to how a mass spectrometer would sample.

1. Sort peptides by elution time;
2. Bin peptides by elution times;
3. **for** $Elution\ Time\ Bin$ **do**
    4. **if** $Retention\ Time_{Peptide_{j,k}} > Elution\ Time\ Bin + Dynamic\ Exclusion\ Time$ **then**
        5. Remove $Peptide_{j,k}$ from $Dynamic\ Exclusion\ List$;
    end
    6. **if** $m/z_{Peptide_{j,k}} \in Dynamic\ Exclusion\ List$ **then**
        7. Remove $Peptide_{j,k}$;
    end
    8. Select Top n Peptides by Abundance;
    9. **for** $Peptide_{j,k} \in Top\ n$ **do**
        10. Add to $Dynamic\ Exclusion\ List$;
    end
end

3.8.1.3 Model Parameters

We generated 15 datasets with the following characteristics. Each dataset contained 30 distinct taxa with four coarse-grained protein groups of varying diversity. As above, diversity is modeled using varying sizes of sequence 'banks' (we used sizes of 15000, 50000, 100000, 250000, and 500000). From each protein group (represented by these different sequence banks of peptides), each organism has 2000 peptides, which are randomly drawn from these sequence banks. Retention times are assigned from a uniform distribution ranging from 0–90 minutes.

The maximum injection time, which is used as the width of the elution time bin, is 500 ms (or 0.00833 minutes), following from McCain and Bertrand (2019). We assign a constant ion peak width of 0.5 minutes, independent of ion intensity. We use a Top n of 12 ions. Our precursor selection window is set to 3 m/z and our dynamic exclusion time span is set to 0.5 minutes.

3.8.2 Supplementary Discussion

3.8.2.1 Underestimation of Coarse-Grained Protein Groups

Our simulations showed that abundant protein groups have good estimates (close to the 1:1 line, Supplementary Fig. 3.16), while low abundance protein groups tend to be underestimated. Each point in Supplementary Fig. 3.16 represents an estimate of taxon-specific coarse-grained protein pool, with the "true" value compared with the "observed" value. This relationship of underestimation with abundance is because of the data-dependent acquisition sampling method that mass spectrometers use. Data-dependent acquisition specifically targets the highly abundant peptides, so lower abundance groups tend to get sampled less. The method we (and others) typically use is to sum the peptide intensities to obtain an abundance estimate. With fewer peptides quantified, the sum will be lower (Supplementary Fig. 3.16). Note that sequence diversity can also influence these estimates (represented with blue colour gradient, Supplementary Fig. 3.16), but only until there is extremely low diversity (darkest colour), corresponding with only a few peptides identified and mapped to a taxon. Note that we are considering the proteomic mass fraction, and quantified this using peptide intensities. Protein quantification typically adjusts for the length per protein, but if this adjustment was not made, it would be equivalent to how we are calculating the proteomic mass fraction.

3.8.2.2 Sequence Diversity

Our conclusions about the ribosomal and photosynthetic proteomic mass fractions, as well as the environment-independent proteomic mass fraction, are potentially influenced by varying degrees of biodiversity within each taxonomic group. Yet, we restricted our analyses to these taxonomic groups due to the robustness of estimation with higher numbers of peptides (above simulations, and McCain and Bertrand, 2019). Further, this level of taxonomic resolution is typically used to compare ecological strategies across marine microbes, so we reasoned it would be useful to introduce these proteomic traits at the same level (e.g. Alexander et al., 2015b).

Here we outline the challenges in comparing taxonomic groups with varying biodiversity within each group, focusing on the environment-independent mass fraction proteomic 'trait'. Biodiversity could influence the environment-independent mass fraction in several ways, depending on the exact meaning of 'biodiversity' in this context. The source of this variation could be due comparisons between taxonomic groups with varying levels

of biodiversity (in terms of sequence diversity), or it could be due to a shift in community composition within taxonomic groups across samples (for example from one diatom species to another). These different mechanisms lead to different potential problems. For example, if community composition is constant across time, but one grouping is more biodiverse than another, our estimates could be interpreted as an average across subgroups (note this is not necessarily the case). But if there are significant shifts in community composition, then this might correspond with an apparent increase in peptide variability that arises from the change in community composition rather than changes in protein expression.

How could varying degrees of diversity be adjusted for? Simply correcting for total peptide diversity (in terms of numbers of peptides unique for a taxonomic group) is an obvious first step. Consider, however, the relationship between the total number of unique peptides for a species with a high regulatory cost (many regulatory proteins). There would be a causal connection between the number of unique peptides and the exact trait we are examining – regulatory cost – so 'adjusting' for peptide diversity would not be appropriate.

Another approach to assess variable biodiversity across taxa is to examine finer taxonomic resolution, and then estimate and compare the environment-independent proteomic mass fraction at that finer resolution with our original, coarse resolution estimates. This is problematic for two reasons: 1) peptides used for a finer taxonomic resolution are unlikely to be a random subsample, and certain protein functions are most likely enriched. If these protein functions are more or less likely to be constitutively expressed, estimates will not be comparable across taxonomic resolution. 2) Subsampling in mass spectrometry is explicitly biased towards highly abundant peptides. Peptides that are more abundant tend to have lower coefficients of variation (Supplementary Fig. 3.18). So, a subsample will systematically bias the environment-independent mass fraction upwards. We have outlined some of the principal challenges associated with using metaproteomics to estimate this proteomic trait, and future work is needed to address these issues. However, we think that this trait is still worth examining, because it likely underpins key aspects of ecological variability (e.g. as examined theoretically and experimentally in *E. coli*; Mori et al., 2017).

3.8.2.3 Abundance-Noise Relationship

Another potential bias in studying the environment-independent protein mass fraction is that less abundant proteins have more variation across identical conditions, as the mean

protein coefficient of variation is negatively correlated with mean protein abundance in cultures (Supplementary Fig. 3.18; Schmidt et al., 2016). So, identifying more peptides would increase the average coefficient of variation. However, we did not observe a negative correlation between the peptide-specific coefficient of variation and the mean peptide abundance (Supplementary Fig. 3.19), suggesting that this bias does not influence our estimated environment-independent peptide mass fraction.

3.8.3 Supplementary Table

	Metagenome		Metatranscriptome		TFG Metatranscriptome	
Assembly Characteristic	Assembly	Assembly ORF	Assembly	Assembly ORF	Assembly	Assembly ORF
# contigs (>= 0 bp)	12444606	13213851	174643	354900	1315493	2265230
# contigs (>= 1000 bp)	310023	179467	42232	5741	233450	26705
# contigs (>= 5000 bp)	16653	640	1102	0	1445	0
# contigs (>= 10000 bp)	4455	63	94	0	106	0
# contigs (>= 25000 bp)	623	2	0	0	3	0
# contigs (>= 50000 bp)	83	0	0	0	0	0
Total length (>= 0 bp)	4221407857	3189663338	146545284	117067830	937746248	689493243
Total length (>= 1000 bp)	664849561	260551638	78561781	6572541	359162487	29719929
Total length (>= 5000 bp)	160639633	4582950	7593719	0	9769226	0
Total length (>= 10000 bp)	78447587	865530	1173063	0	1394315	0
Total length (>= 25000 bp)	23756887	59832	0	0	84630	0
Total length (>= 50000 bp)	6112124	0	0	0	0	0
# contigs (>=500 bp)	1177759	827970	104174	69636	795370	384770
Largest contig	224823	32961	23227	4809	33111	3999
Total length	1241647307	698185269	122244748	51909138	754213377	276614610
GC (%)	40.66	40.57	42.5	42.89	43.32	45.72
N50	1083	846	1296	777	969	732
N75	681	636	825	621	710	600
L50	267661	275978	26841	27555	251690	153883
L75	638760	516114	56699	46284	480396	258428
# N's per 100 kbp	19.73	6.06	132.07	6.03	91.56	4.23

Table 3.2: Sequencing and assembly characteristics for the three assemblies (one metagenomic and two metatranscriptomic) used for databases of potential proteins for searching mass spectra. The first four columns correspond to the metagenomic and metatranscriptomic sequencing conducted on the GOS-927, GOS-930, GOS-933 and GOS-935 filters (see Methods). The last two columns correspond to the metatranscriptomic experiment described in the Methods. All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs). N50 is the length for which the collection of all contigs of that length or longer covers at least half (50%) the total base content of the Assembly. It serves as a median value for assessing whether the Assembly is balanced towards longer contigs (higher N50) or shorter contigs (lower N50). N75 is used for the same purpose but is the length is set at 75% of total base content instead of 50%. L50 is the number of contigs equal to or longer than the N50 length. In other words, L50, is the minimal number of contigs that contain half the total base content of the Assembly. L75 is used for the same purpose in reference to the N75 length. Columns labelled 'Assembly' refer to assembled contigs, and columns labelled 'Assembly ORF' refer to predicted ORFs from contigs.

3.8.4 Supplementary Figures

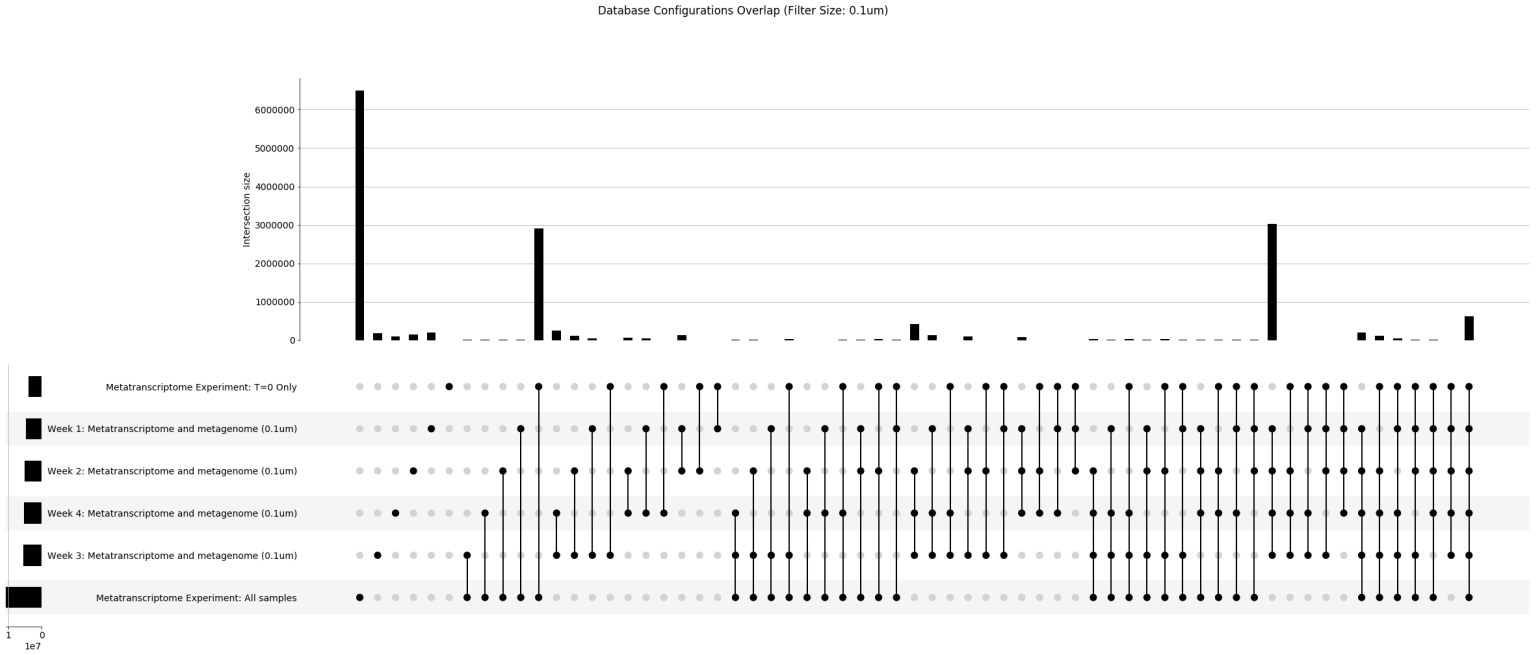

Figure 3.5: Representation of the overlap between different database configurations and the number of tryptic peptides within each. Bar graphs on top represent the number of peptides identified with a given set of sequence groups (i.e. overlapping databases). The set of overlapping sequence groups is represented below with points and lines. For example, the first column shows that the metatranscriptome experiment (all samples) contained by far the greatest number of tryptic peptides, and that the metatranscriptome experiment (T=0) had no unique tryptic peptides (because it is a subset of the former). The side bar plot, next to the database configuration name, is the total number of peptides within each sequence group database (note the scale of the numbers are between 0 and $1{\times}10^7$). In this figure, only the smallest filter size is shown (0.1 μm).

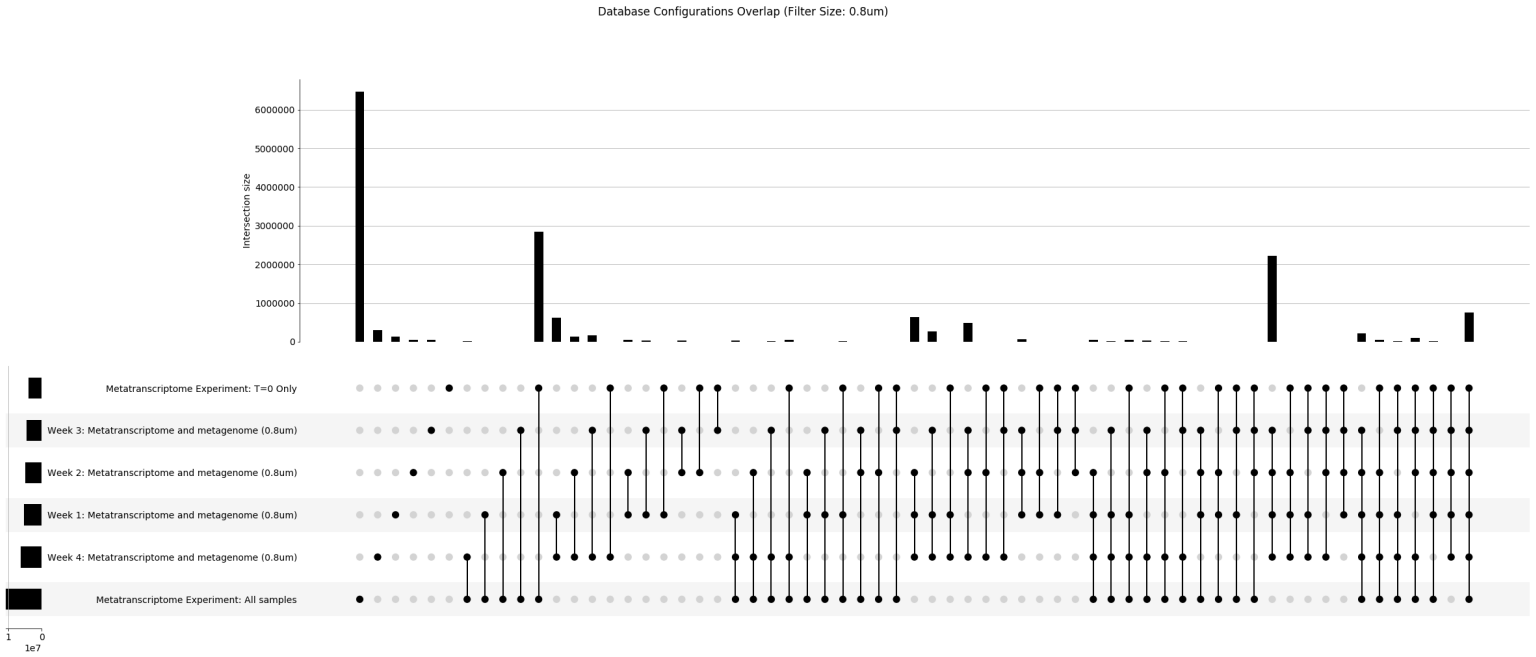

Figure 3.6: Representation of the overlap between different database configurations and the number of tryptic peptides within each. Bar graphs on top represent the number of peptides identified with a given set of sequence groups (i.e. overlapping databases). The set of overlapping sequence groups is represented below with points and lines. The side bar plot, next to the database configuration name, is the total number of peptides within each sequence group database (note the scale of the numbers are between 0 and $1{\times}10^7$). In this figure, only the medium filter size is shown (0.8 μm).

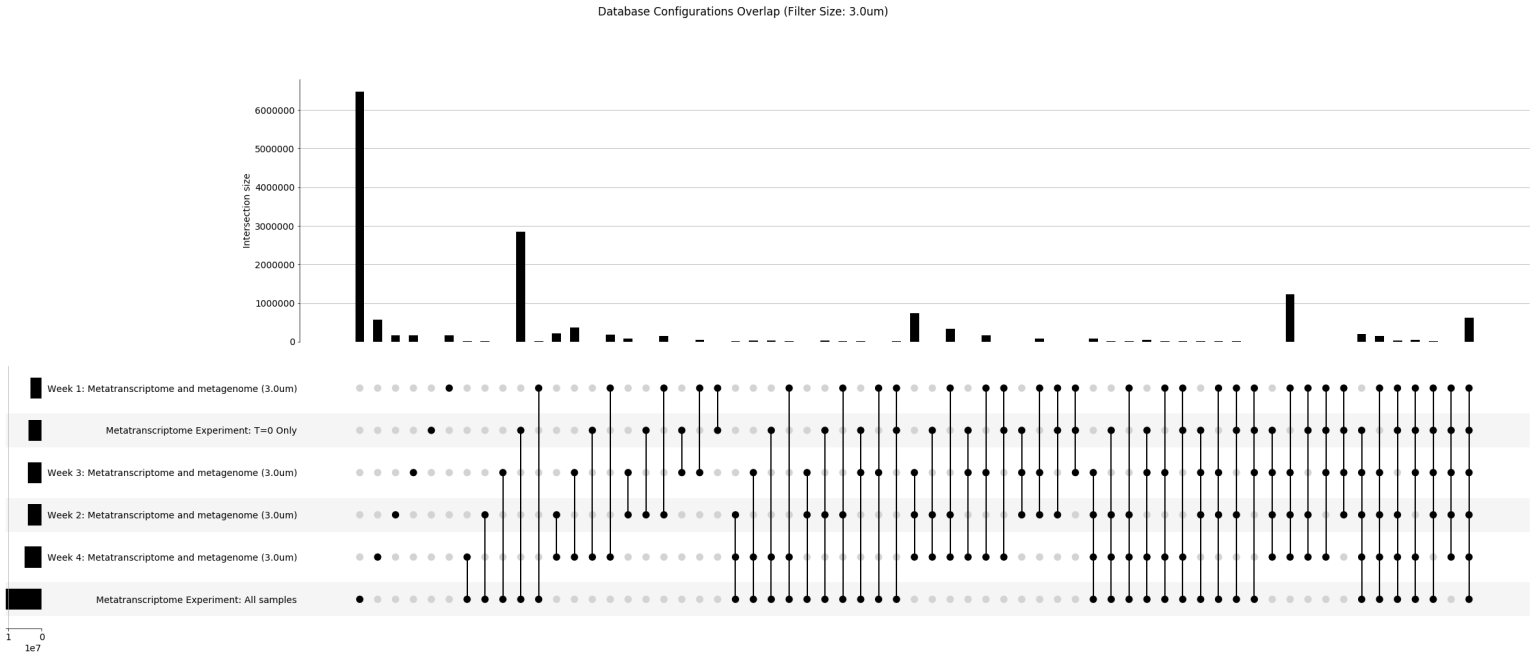

Figure 3.7: Representation of the overlap between different database configurations and the number of tryptic peptides within each. Bar graphs on top represent the number of peptides identified with a given set of sequence groups (i.e. overlapping databases). The set of overlapping sequence groups is represented below with points and lines. The side bar plot, next to the database configuration name, is the total number of peptides within each sequence group database (note the scale of the numbers are between 0 and $1{\times}10^7$). In this figure, only the largest filter size is shown (3.0 μm).

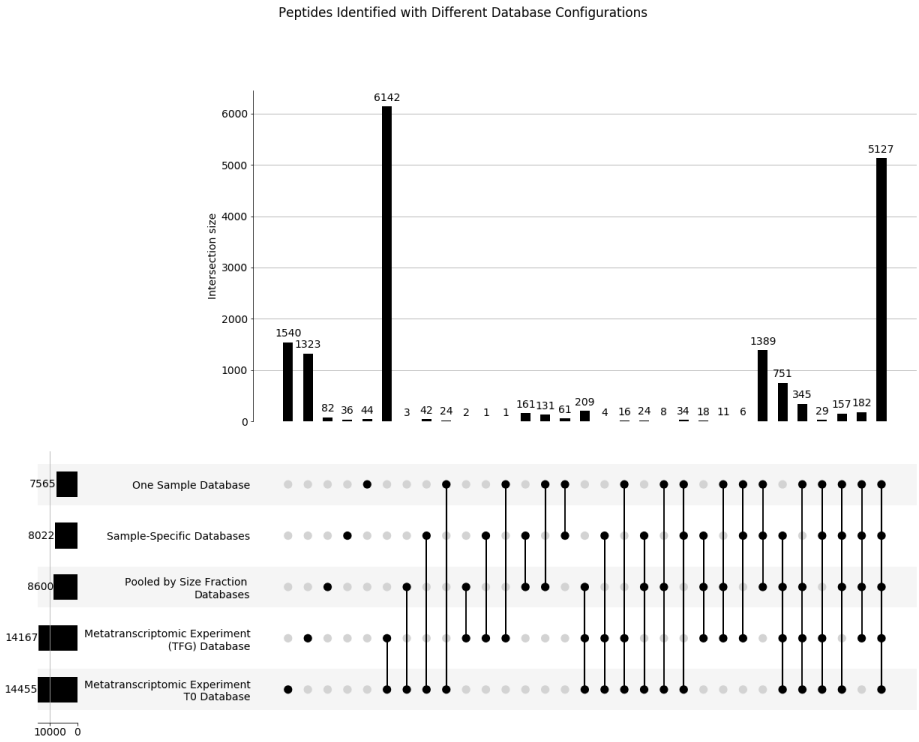

Figure 3.8: Representation of the overlap between different database configurations and the number of peptides identified with each. Bar graphs on top (with numbers above) represent the number of peptides identified with a given set of databases (i.e. overlapping databases). The set of overlapping databases is represented below with points and lines. For example, the first column on the left represents peptides uniquely identified using the database 'Metatranscriptome Experiment T0', where 1540 peptides were uniquely identified. The side bar plot, next to the database configuration name, is the total number of peptides identified using each database. In this figure, all filter sizes are summed together.

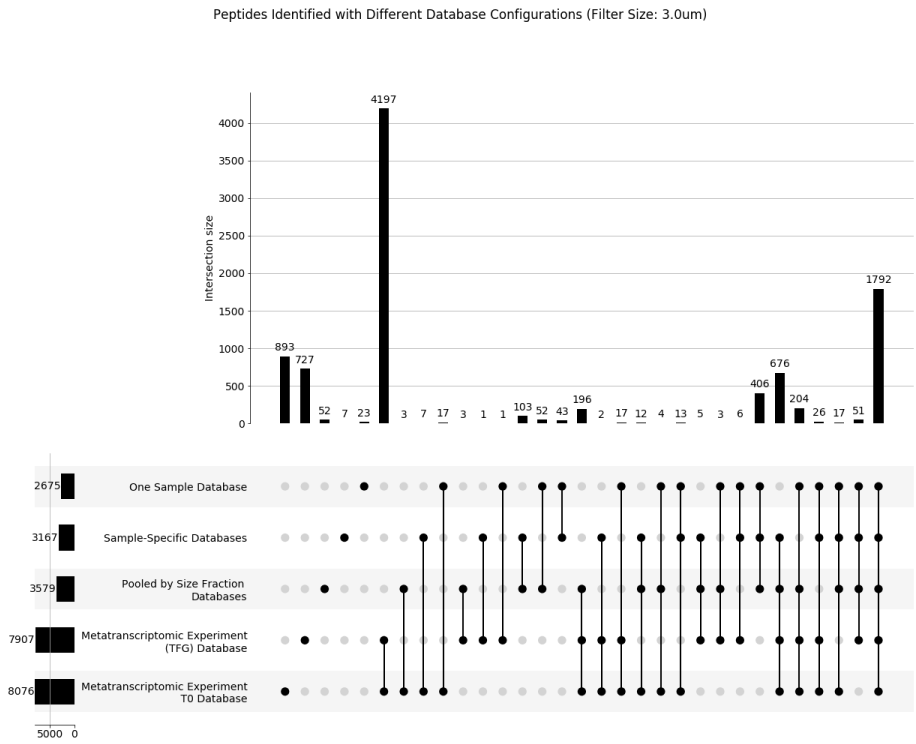

Figure 3.9: Representation of the overlap between different database configurations and the number of peptides identified with each. Bar graphs on top (with numbers above) represent the number of peptides identified with a given set of databases (i.e. overlapping databases). The set of overlapping databases is represented below with points and lines. The side bar plot represents the total number of peptides identified using each database. In this figure, only the largest filter size is shown (3.0 μm).

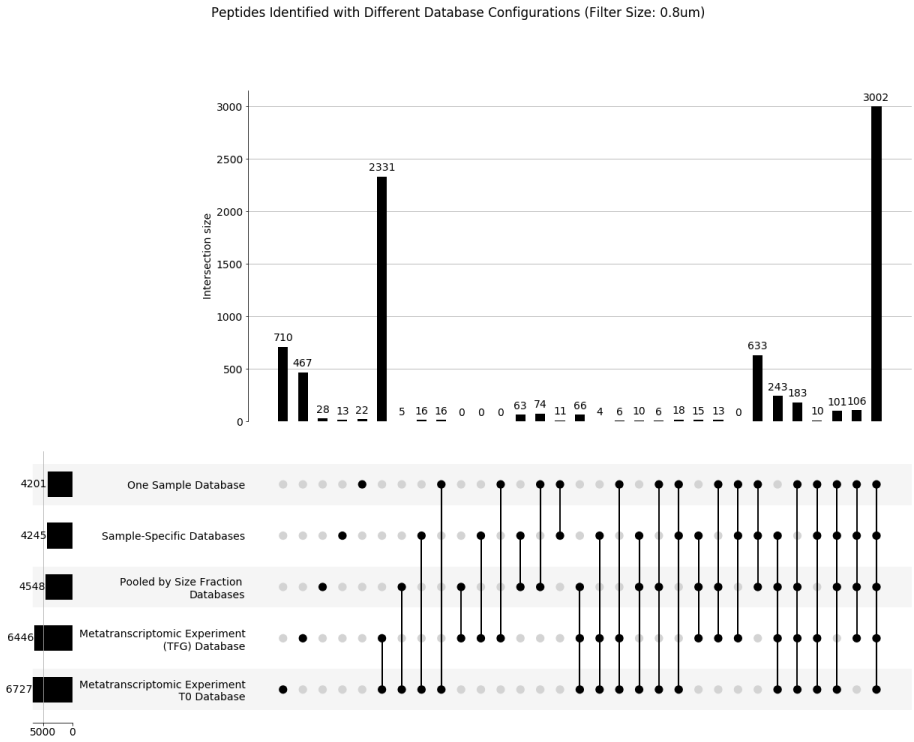

Figure 3.10: Representation of the overlap between different database configurations and the number of peptides identified with each. Bar graphs on top (with numbers above) represent the number of peptides identified with a given set of databases (i.e. overlapping databases). The set of overlapping databases is represented below with points and lines. The side bar plot represents the total number of peptides identified using each database. In this figure, only the middle filter size is shown (0.8 μm).

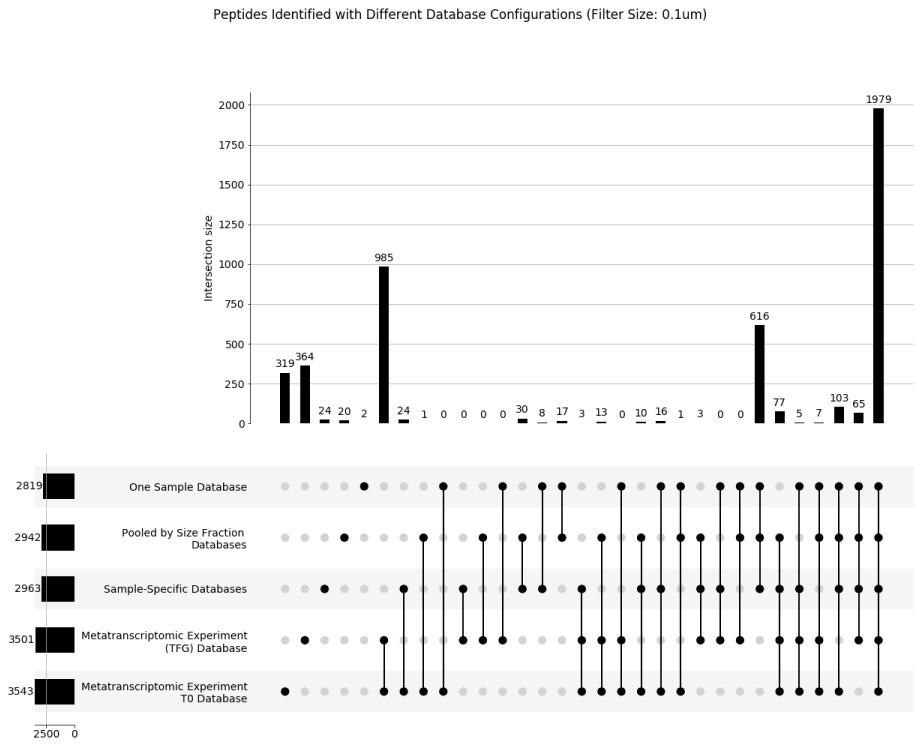

Figure 3.11: Representation of the overlap between different database configurations and the number of peptides identified with each. Bar graphs on top (with numbers above) represent the number of peptides identified with a given set of databases (i.e. overlapping databases). The set of overlapping databases is represented below with points and lines. The side bar plot represents the total number of peptides identified using each database. In this figure, only the smallest filter size is shown (0.1 μm).

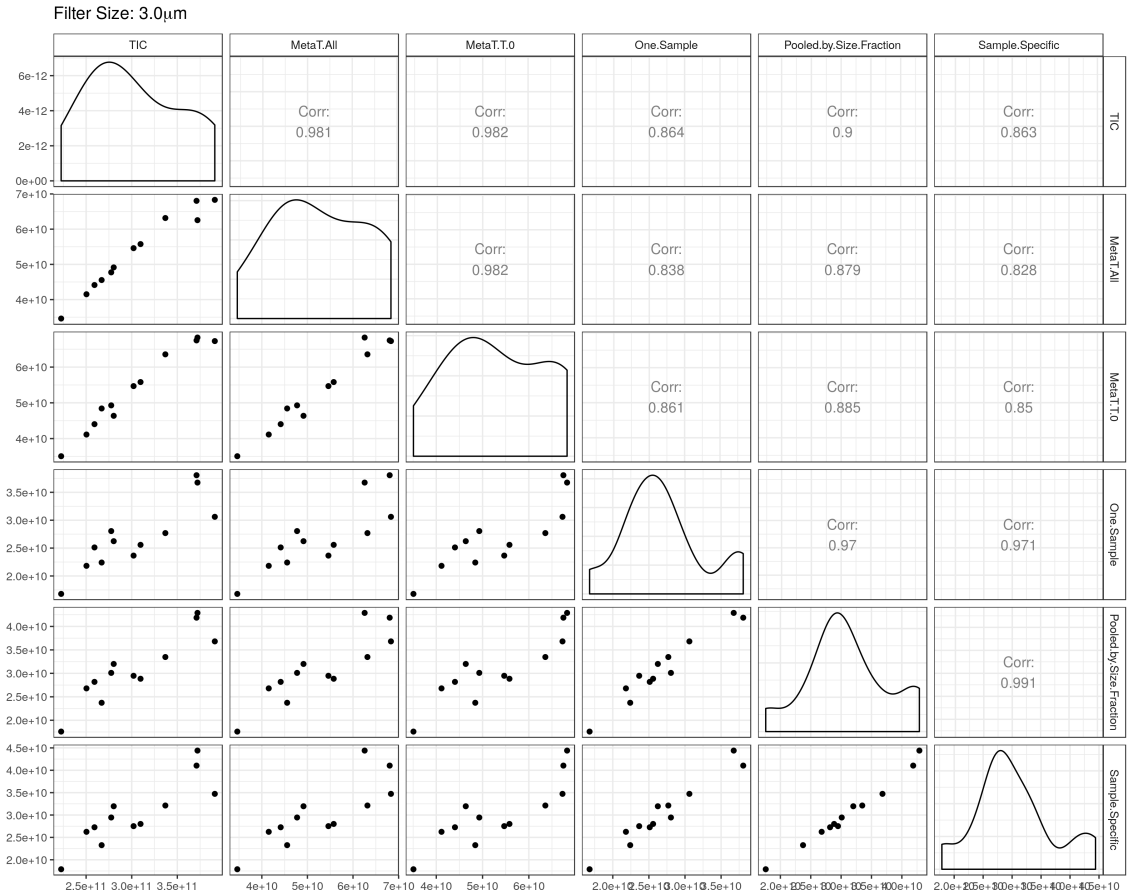

Figure 3.12: The sum of peptide intensities (i.e. normalization factors) are against each other for different database configurations, as well as against total ion current (TIC). Points represent different mass spectrometry experiments. Correlation values (coefficient of determination) are represented in corresponding locations. Only the largest filter size is shown here (3.0 μm).

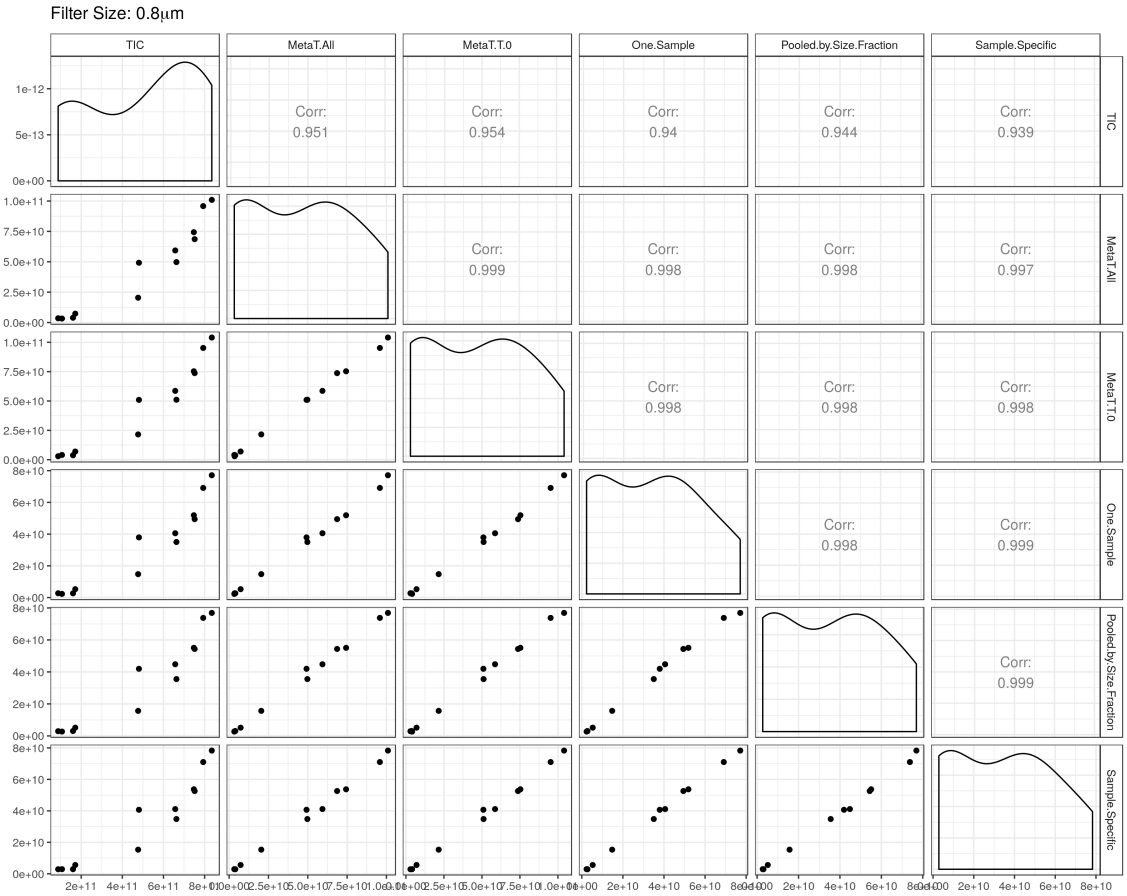

Figure 3.13: The sum of peptide intensities (i.e. normalization factors) are against each other for different database configurations, as well as against total ion current (TIC). Points represent different mass spectrometry experiments. Correlation values (coefficient of determination) are represented in corresponding locations. Only the middle filter size is represented here (0.8 μm).

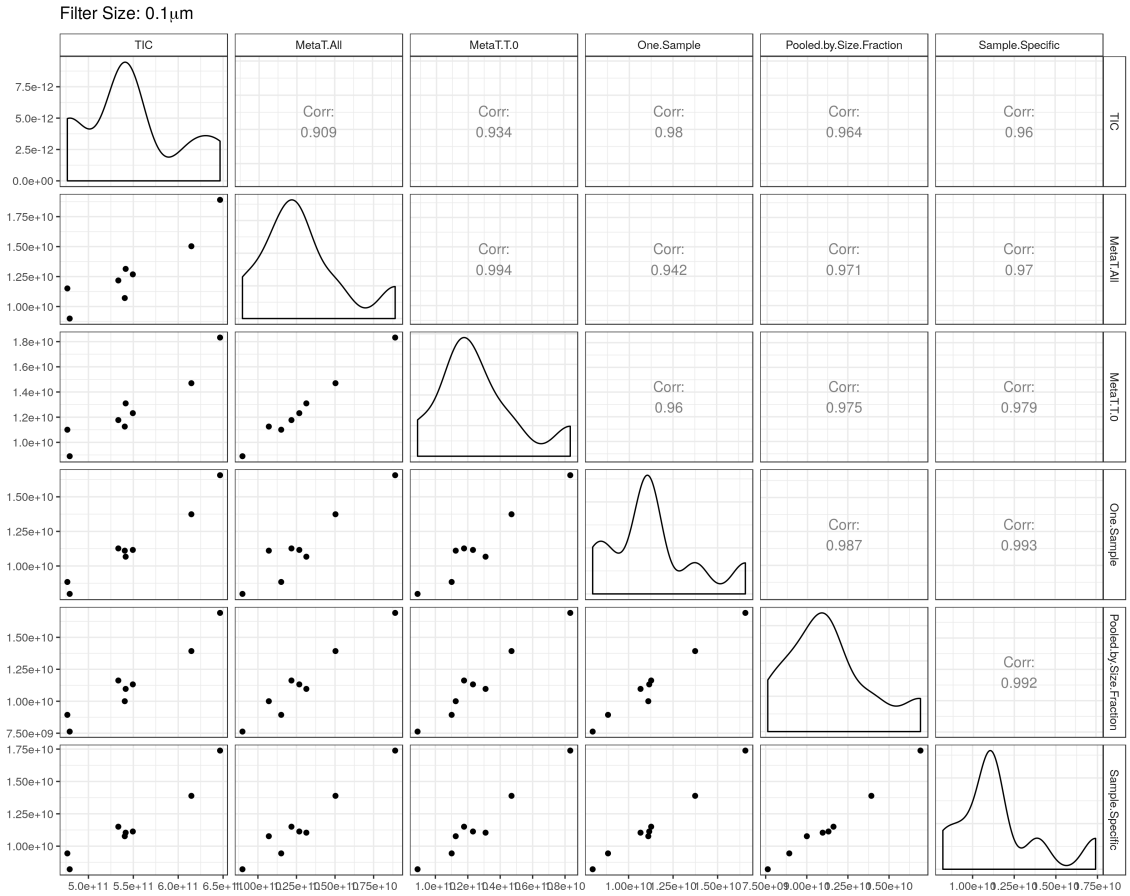

Figure 3.14: The sum of peptide intensities (i.e. normalization factors) are against each other for different database configurations, as well as against total ion current (TIC). Points represent different mass spectrometry experiments. Correlation values (coefficient of determination) are represented in corresponding locations. Only the smallest filter size is represented here (0.1 μm).

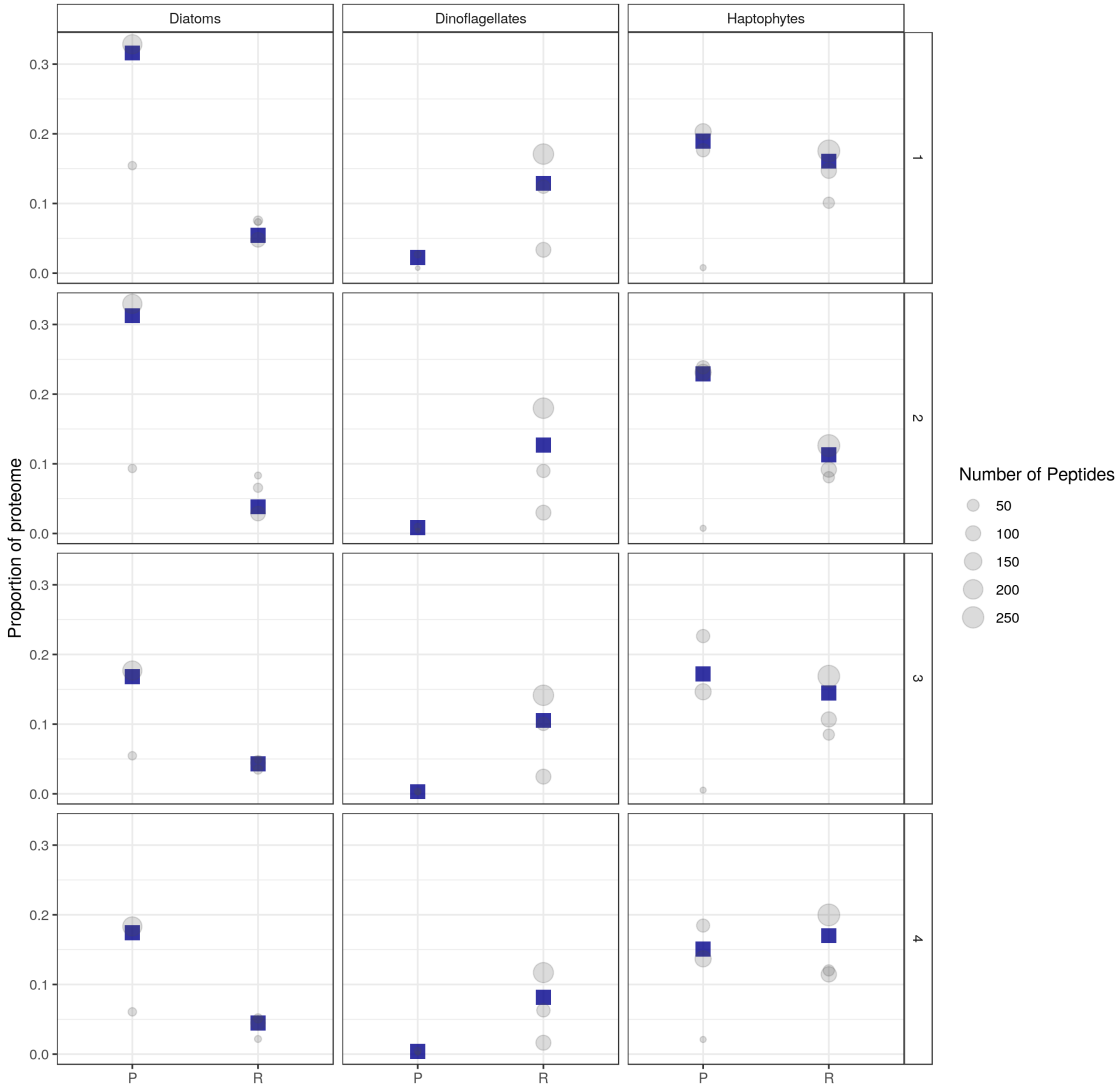

Figure 3.15: Demonstrating how various estimates across filters are collapsed into one estimate, based on the number of peptides identified within each filter size. Grey points represent the different filter sizes, while the size per grey point is the number of peptides observed in that filter, corresponding to a given taxa or a coarse-grained proteomic pool (P is photosynthetic protein pool, R is ribosomal protein pool). Dark blue squares are the weighted estimates. Numbers in the vertical direction (right side) correspond to the different sampling weeks.

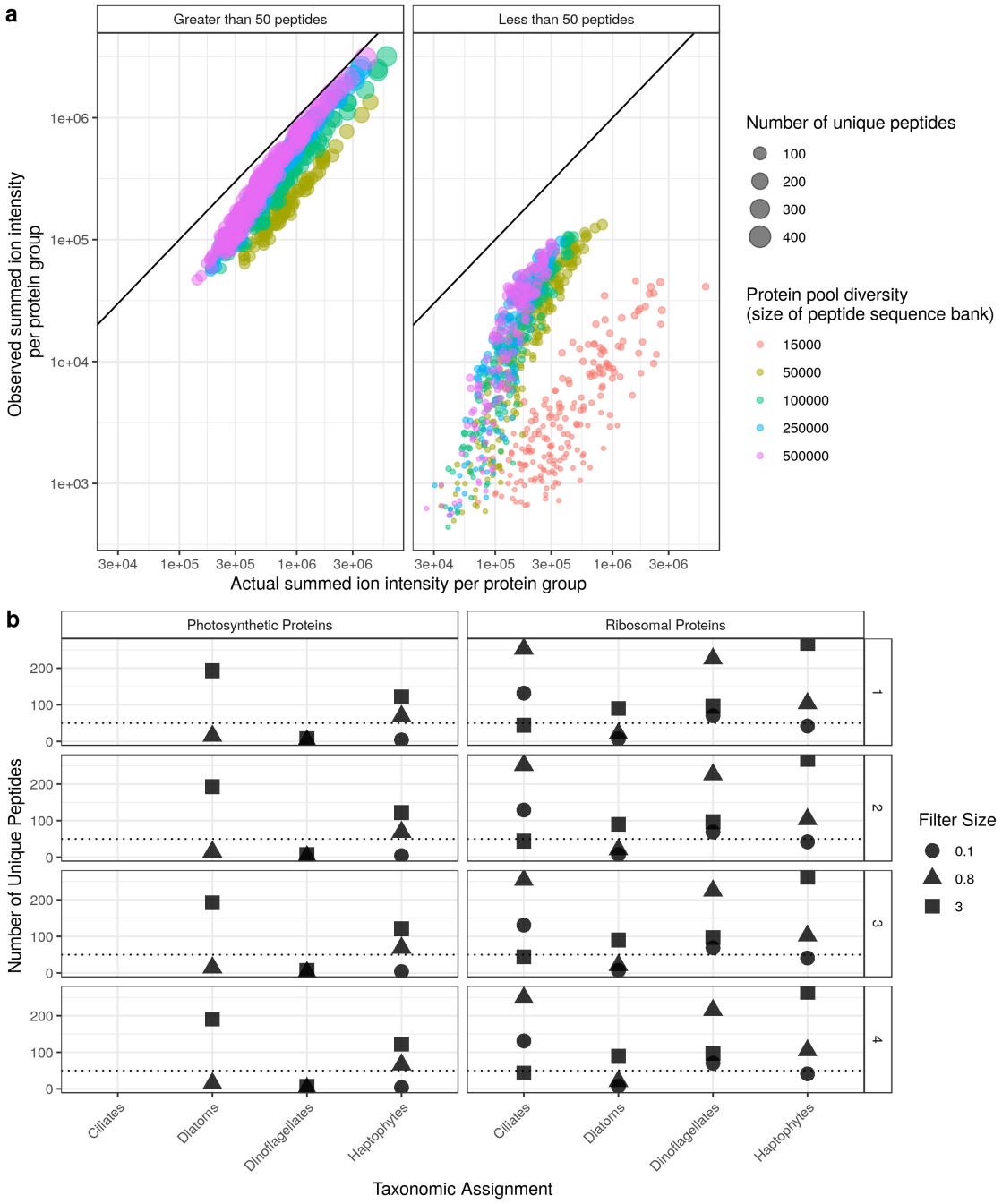

Figure 3.16: Results from the metaproteomic sampling simulation suggest that restricting analyses to protein groups and taxa that are abundant will prevent bias due to different amounts of diversity. **a**, varying degrees of simulated diversity (colour of points), demonstrates that low diversity pools are underestimated (i.e. far away from 1:1 line). Yet, if many peptides are observed (i.e. above 50 peptides), then the estimates are linearly correlated with the 1:1 line, with only slight underestimates due to diversity. **b**, the number of photosynthetic and ribosomal protein specific peptides that are also taxon-specific across different filter sizes (estimates across filter sizes were weighted and merged). At least one filter for each protein pool has greater than 50 peptides, except Dinoflagellate photosynthetic proteins. The dotted horizontal line corresponds with 50 unique peptides. These data suggest that our estimates of ribosomal and photosynthetic protein mass fraction not susceptible to diversity-induced bias.

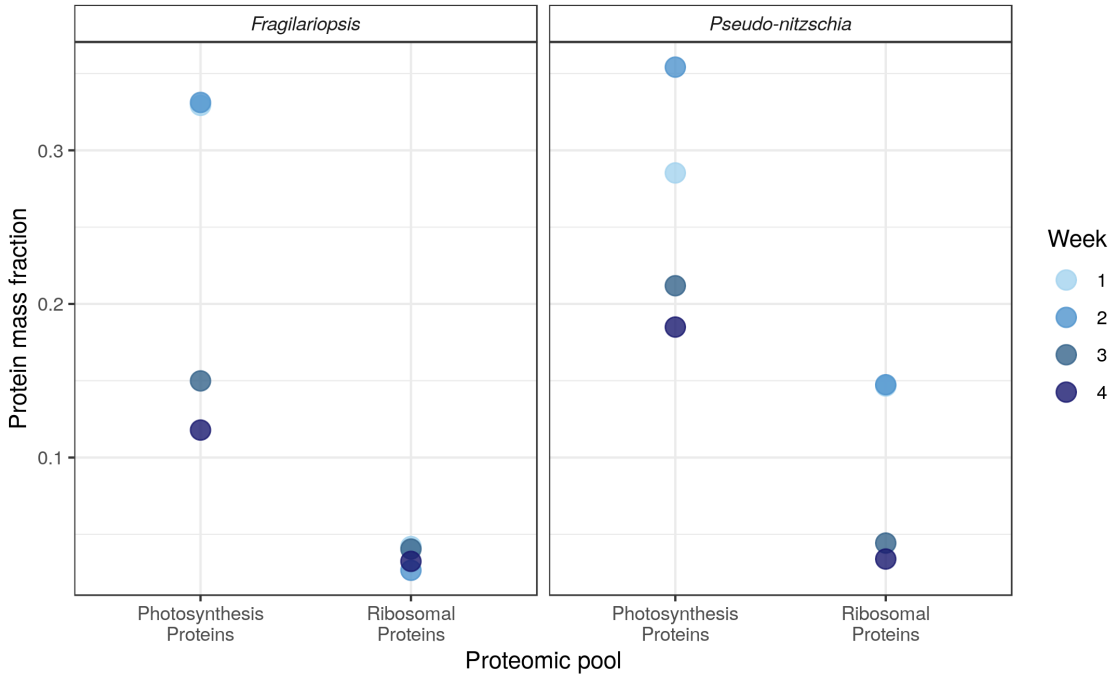

Figure 3.17: Proteomic proportions of two taxonomic groups of diatoms, *Fragilariopsis* sp. and *Pseudo-nitzschia* sp. Ribosomal and photosynthetic proportions were similar across groupings, and also similar to the larger grouping of diatoms.

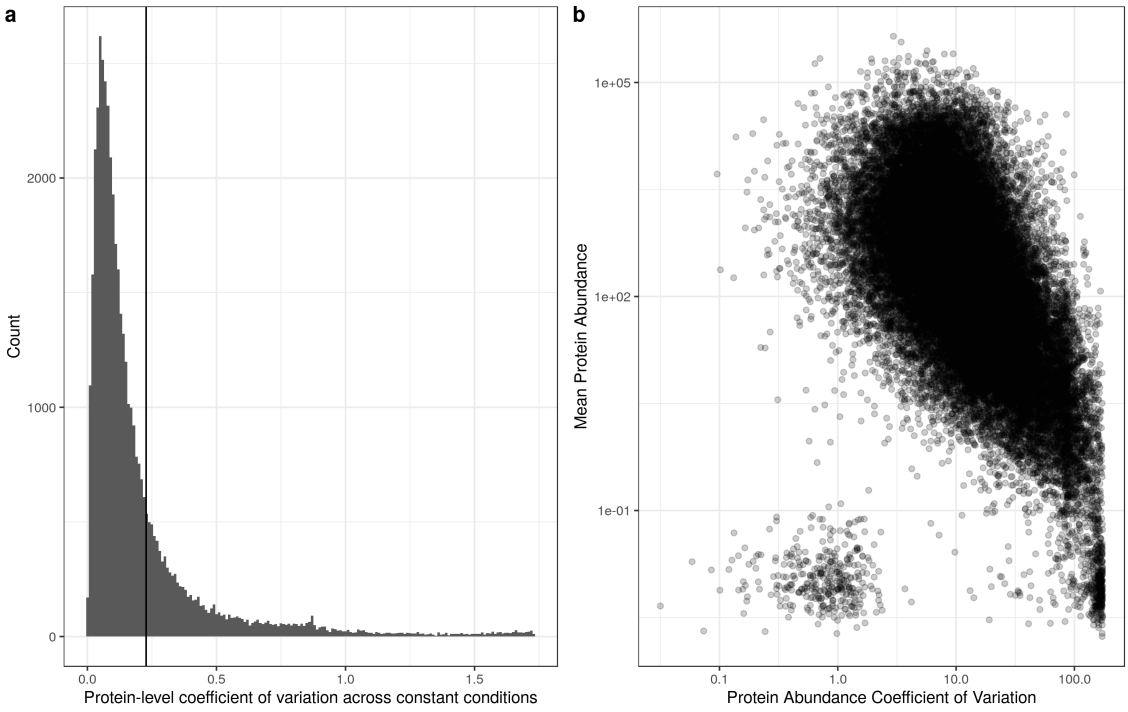

Figure 3.18: Protein-level summary statistics derived from a comprehensive proteomic characterization of *E. coli* (Schmidt et al., 2016). a. the distribution of protein-level coefficients of variation, using all 22 experimental treatments described in Schmidt et al. (2016). The vertical line was calculated by first determining the third quartile of the distribution of coefficients of variation for each condition, and then calculating the mean of these third quartiles. These coefficients of variation should presumably lead to mostly constant protein expression, but there are some proteins that have intrinsic noise in expression levels. We chose an arbitrary cut-off to classify protein expression as constant or not. b. Plotting the relationship between the mean protein abundance across conditions with its coefficient of variation shows a negative correlation between the two summary statistics at the protein level (Spearman's ρ = -0.55).

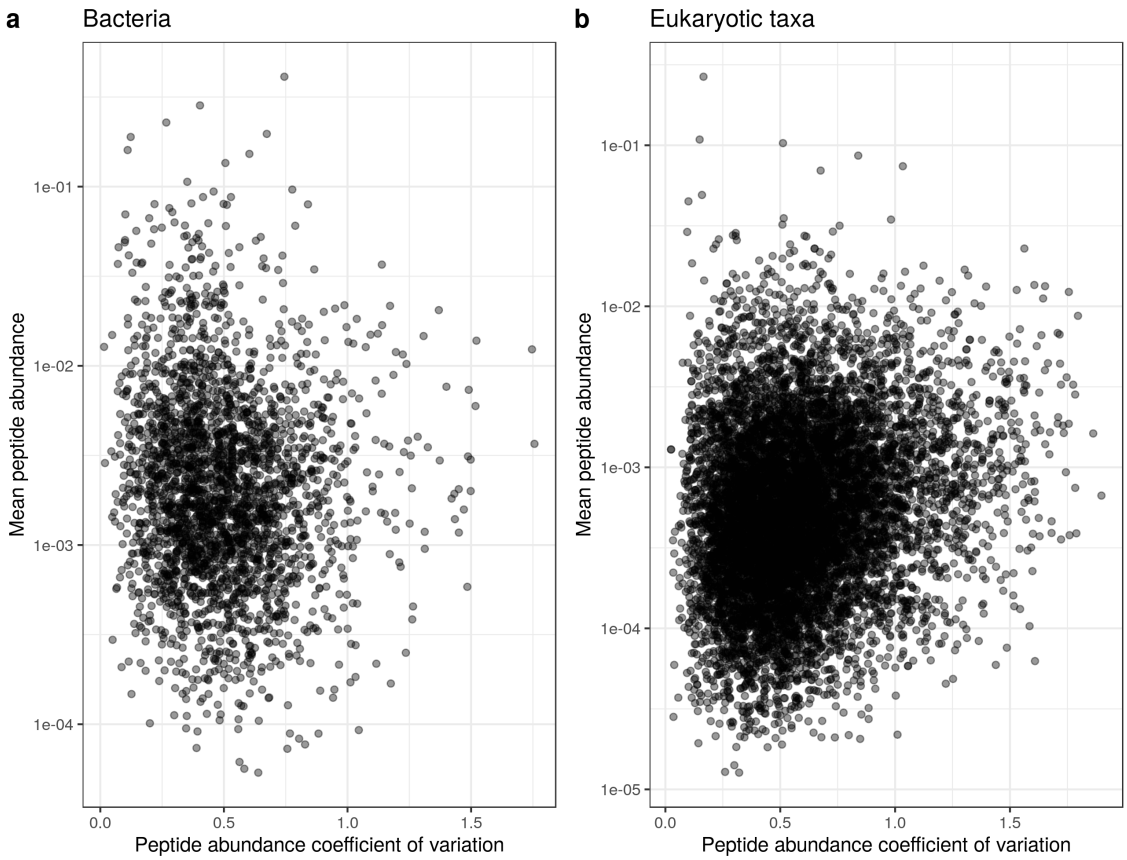

Figure 3.19: Weak relationships between the peptide abundance coefficient of variation and the mean peptide abundance for the prokaryotic and eukaryotic taxa we observed (Spearman's ρ = -0.09 and 0.18, respectively).

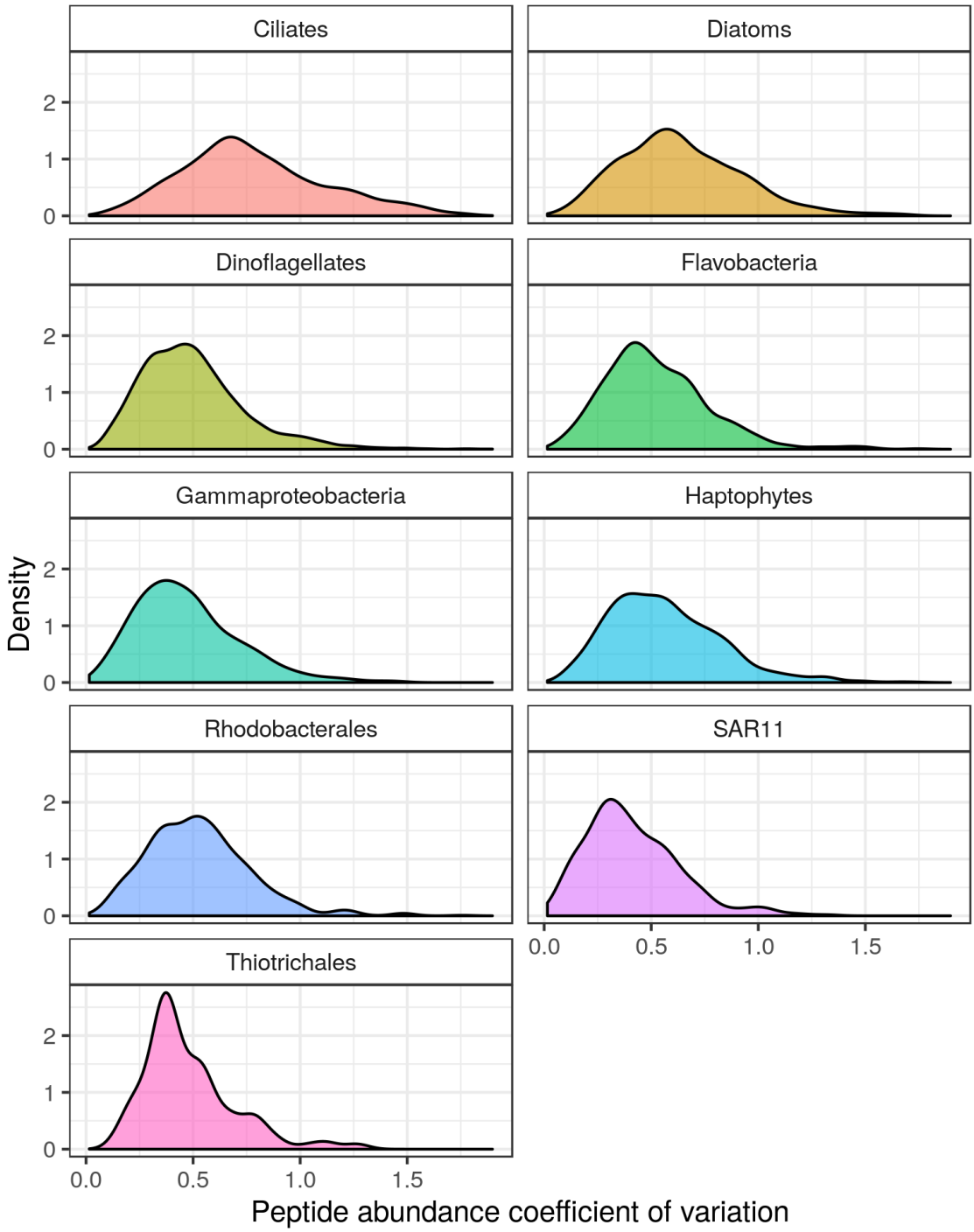

Figure 3.20: Distributions of peptide-specific coefficients of variation for each taxa we examined. In the main manuscript, only SAR11 and diatom distributions are shown. Methods for calculating this distribution are given in the main manuscript.

CHAPTER 4

CELLULAR COSTS UNDERPIN MICRONUTRIENT LIMITATION IN PHYTOPLANKTON

This work was published previously in *Science Advances* (McCain et al., 2021).

4.1 Abstract

Micronutrients control phytoplankton growth in the ocean, influencing carbon export and fisheries. It is currently unclear how micronutrient scarcity affects cellular processes, and how interdependence across micronutrients arises. We show that proximate causes of micronutrient growth limitation and interdependence are governed by cumulative cellular costs of acquiring and using micronutrients. Using a mechanistic proteomic allocation model of a polar diatom focused on iron and manganese, we demonstrate how cellular processes fundamentally underpin micronutrient limitation, and how they interact and compensate for each other to shape cellular elemental stoichiometry and resource interdependence. We coupled our model with metaproteomic and environmental data, yielding a novel approach for estimating biogeochemical metrics including taxon-specific growth rates. Our results show that cumulative cellular costs govern how environmental conditions modify phytoplankton growth.

4.2 Introduction

Marine phytoplankton are responsible for approximately half of global net primary productivity, supporting key ecosystem services (Field et al., 1998). Micronutrients, such as iron, are often depleted in the ocean, limiting phytoplankton growth and therefore impacting fisheries productivity and carbon export globally (Tagliabue et al., 2017, 2020; Assmy et al., 2013). These resources are cofactors for enzymes that catalyze intracellular reactions, and unlike the macronutrients nitrogen and phosphorous, they comprise a negligible fraction of biomass. Cellular micronutrient stoichiometry is highly variable (Twining and Baines, 2013) and elements can conditionally substitute for one another (Saito, Goepfert and Ritt, 2008). Therefore traditional approaches that simply link growth rate to resource scarcity may not apply.

Growth is the emergent outcome of a range of internal cellular processes competing for shared resources (Kafri et al., 2016) that are governed by costs (e.g. number of amino acids per protein or energetic requirements; Dekel and Alon, 2005; Basan et al., 2015; Jahn et al., 2018) and constraints (e.g. limits of protein density in a membrane; Szenk, Dill and de Graff, 2017). Protein synthesis capacity has been identified as a key growth-limiting process in model heterotrophic organisms with various carbon sources (Schaechter, Maaløe and Kjeldgaard, 1958; Scott et al., 2010). Only recently have other non-carbon macronutrients been considered and additional complexities have been revealed (e.g. Li et al., 2018; Kafri et al., 2016). Currently, we lack knowledge regarding which internal processes limit growth under micronutrient deficiency. Further, while we know that multiple nutrients can simultaneously impact growth rate (Browning et al., 2017), the mechanisms by which they interact appear to vary for each nutrient pair (Saito, Goepfert and Ritt, 2008).

The overriding conceptual view in oceanography is not sufficient to mechanistically represent micronutrient limitation and resource interdependence. Currently, external resource scarcity (e.g. bioavailable forms of nitrogen, phosphorus, iron, etc.), relative to fixed requirements, is assumed to control growth and carbon fixation rates (Moore, Doney and Lindsay, 2004; Laufkötter et al., 2015). However, this ignores the role of internal processes in limiting growth. It also prevents general mechanisms of resource interdependence, which may arise because different internal processes compete for shared cellular resources, from being included in large-scale ocean models. While external

resource scarcity is clearly the ultimate cause of limitation, the proximate causes drive the sensitivity to environmental change. For example, temperature-driven changes in ribosomal translation rates might influence cellular nitrogen to phosphorous ratios because ribosomes are a large portion of phosphorous quotas (Toseland et al., 2013). Currently, ocean models used for climate change projections parameterize growth as a simple function of the single most limiting resource (Laufkötter et al., 2015), which introduces substantial uncertainties in a changing environment (Tagliabue et al., 2020). While some phytoplankton models have leveraged quantitative, mechanistic insights into cellular processes (Loladze and Elser, 2011; Toseland et al., 2013; Bonachela et al., 2013; Talmy et al., 2013; Nicholson, Stanley and Doney, 2018; Inomura et al., 2019), none have examined interactions between micronutrients or used *in situ* gene expression data to resolve cellular processes.

In this study, we quantify the proximate costs and constraints associated with micronutrient limitation via a novel coupling of cellular modelling and metaproteomics from the Southern Ocean. By deriving a phenomenological model, we identify key factors controlling interdependence across micronutrients. Finally, we demonstrate a framework for inferring critical biogeochemical metrics, such as growth rates, by coupling *in situ* gene expression and geochemical data with cellular modelling. Taken together, this framework quantifies cellular costs and constraints to examine the mechanistic underpinnings of phytoplankton growth in the ocean.

4.3 Results and Discussion

4.3.1 Estimating Cellular Costs and Constraints with a Diatom Proteomic Allocation Model

We estimated the cellular costs and constraints of micronutrient limitation in phytoplankton by developing a mechanistic, proteomic allocation model for the polar diatom *Fragilariopsis cylindrus* (Mock et al., 2017; Faizi et al., 2018). Our model considers the essential micronutrients iron and manganese, which both influence primary productivity in the Southern Ocean (Assmy et al., 2013; Buma et al., 1991; Browning et al., 2014; Wu et al., 2019), and represents the various processes underlying cellular growth, like photosynthesis and translation (Molenaar et al., 2009; Weiße et al., 2015; Faizi et al., 2018; Zavřel et al., 2019). The model is comprised of several 'coarse-grained' protein pools (i.e. proteins

grouped together with related functions, Fig. 4.1a): iron- and manganese-specific transporters, photosystem units, nitrogen uptake and metabolism (from nitrate to amino acids), and antioxidants (represented here by manganese superoxide dismutase, MnSOD). Each protein pool has an associated cost, which is proportional to the number of amino acids per pool (estimated using the *F. cylindrus* genome; Mock et al., 2017). Ribosomes are assumed to be allocated to maximize the steady-state specific growth rate, and each protein pool, metabolite, and internal free pool of Fe and Mn is described by an ordinary differential equation (ODE). The system of ODEs are connected by various stoichiometric coefficients obtained from the literature, for example Mn atoms per MnSOD or the total number of Fe atoms within all proteins involved in converting nitrate into amino acids. We then integrated the system of ODEs forward in time to obtain steady-state estimates of each state variable, from which we calculate the specific growth rate. In our model we define the specific growth rate as the rate of biosynthesis of amino acids relative to the average protein per cell (Faizi et al., 2018). We used Bayesian optimization to determine the optimal ribosomal allocation under a given set of dissolved Mn (dMn), Fe (dFe), and light conditions (Methods). Iron and Mn interact via oxidative stress, where under low dFe, electrons leak more frequently from electron transport (Niyogi, 1999) thus increasing the requirement for the Mn-containing antioxidant superoxide dismutase (MnSOD Peers and Price, 2004). Under low antioxidant availability, the cell must replace proteins damaged by reactive oxygen species by increasing protein synthesis. Accordingly, the mismatch between superoxide production and its consumption via MnSOD leads to a protein synthesis rate penalty in our model (see Methods).

We then leveraged proteomic and metaproteomic data to estimate three key costs and constraints: (i) internal Fe and Mn protein cost, (ii) available membrane space for transporters (Lis et al., 2015), and (iii) catalytic efficiency of MnSOD (Methods). (i) refers to all proteins required for acquiring, shuttling and storing Fe within the cell (e.g. ferritin; Marchetti et al., 2009), which is dynamic such that Fe protein cost increases with Fe quota (an identical cost is applied for Mn, Supplementary Discussion). (ii) refers to the proportion of membrane space available for metal transporters (Lis et al., 2015; Szenk, Dill and de Graff, 2017), for which we extended a mechanistic nutrient uptake model (Aksnes and Egge, 1991; Aksnes and Cao, 2011; Fiksen, Follows and Aksnes, 2013), accounting for competition for membrane space between iron and manganese transporters.

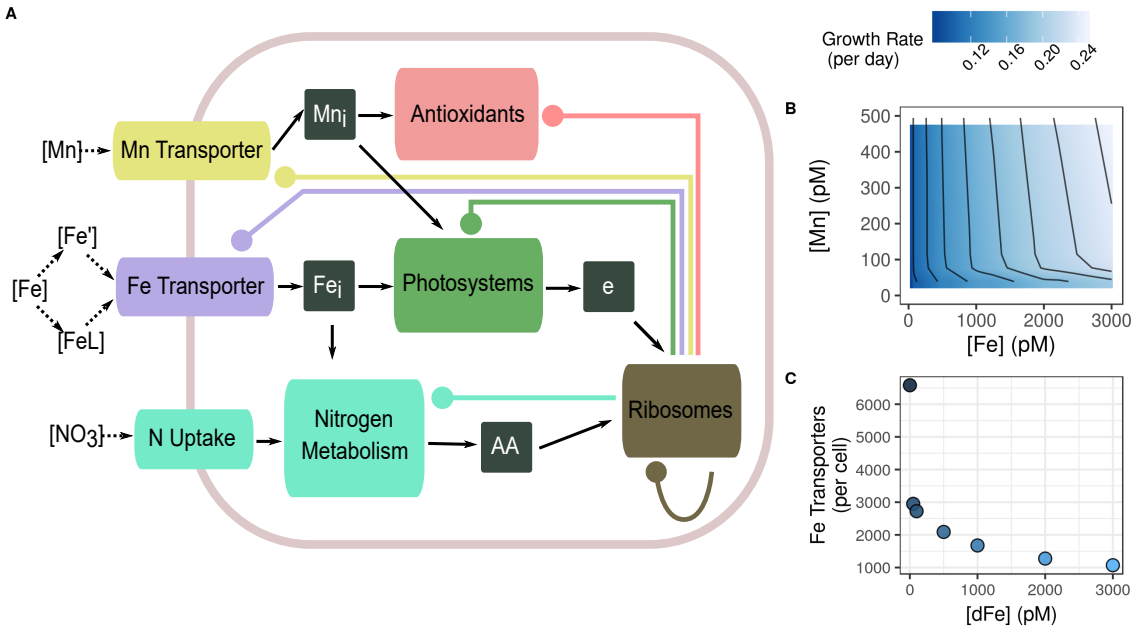

Figure 4.1: A polar diatom-based proteomic allocation model combined with metaproteomic observations reproduces expected cell behaviour. (**A**), schematic of proteomic allocation model. Micronutrients are taken up via nutrient-specific protein transporters (left). Internal pools of Mn and Fe (black boxes) are then accessible for protein synthesis. Photosystems require both Fe and Mn, and are the source of energetic equivalents ('e'; black box), which are then used by protein synthesis, micronutrient uptake and nitrogen metabolism (latter two are not shown with arrows). Protein pools are synthesized via ribosomes and represented with circle-ended lines. All model runs were conducted with nitrate at saturating levels. (**B**), growth rates across a range of Fe and Mn concentrations are quantitatively similar to growth rates in culture (Supplementary Fig. 4.9). (**C**), Fe transporters decrease with increased Fe concentrations (dMn = 500 pM), a commonly observed phenomenon in cultures (Sunda and Huntsman, 1985; Hudson and Morel, 1990).

(iii) represents the effectiveness of a single MnSOD unit.

Model parameters were estimated using Approximate Bayesian Computation (ABC) in a novel combination with diatom proteomes inferred from a metaproteomic time series (Wilkinson, 2013). The metaproteome characterisation coupled peptide mass spectrometry with metatranscriptomics (Jabre et al., 2021) to examine protein expression over time at the Antarctic sea ice edge, where concurrent bottle incubations indicated a transition into micronutrient stress (cobalamin, Mn, Fe; Bertrand et al., 2015; Wu et al., 2019). Coarse-grained diatom protein pool biomass was estimated using the sum of diatom-specific peptide intensities (Supplementary Fig. 4.6; Kleiner et al., 2017). Coarse-graining is necessary to prevent biases in peptide detectability and quantification across complex samples (McCain and Bertrand, 2019). Finally, we combined the inferred diatom proteome observations with two previously published diatom proteomic datasets to estimate each parameter (Methods, Supplementary Fig. 4.7; Cohen et al., 2018; Nunn et al., 2013). We have assessed various forms of biases and developed methods for connecting environmental gene expression data to quantitative models of cellular processes (Methods), providing a path forward to leverage large-scale datasets in this way.

Our model reproduces expected cellular behaviour across a range of dFe and dMn concentrations (Fig. 4.1b, 4.1c; Supplementary Fig. 4.8; using posterior modes for estimated parameters, Supplementary Fig. 4.7). For example, the model quantitatively reproduces growth rates (Jabre and Bertrand, 2020), Mn and Fe cellular quotas (Twining, Baines and Fisher, 2004; Peers and Price, 2004), and dFe uptake rates within observational constraints (McQuaid et al., 2018), despite no prescribed parameterisation or model training on these data types (Supplementary Figs 4.9, 4.10, 4.11, 4.12). We are also able to reproduce the observed increase in transporters under low dFe and dMn (Fig. 4.1c, Supplementary Fig. 4.13; Sunda and Huntsman, 1985; Bonachela et al., 2013), the expected interaction between light and Fe quota (Supplementary Fig. 4.14 and Supplementary Discussion; Sunda and Huntsman, 1997), and the increase in ribosomes with growth rate (Supplementary Fig. Fig. 4.13, Fig. 4.2; Waldron and Lacroute, 1975; Scott et al., 2010). Interestingly, our analysis suggests dMn and dFe interactively influence growth more at high dFe, rather than at low dFe, and a reframing of previous results supports this conclusion (Supplementary Fig. 4.15, Supplementary Discussion). Overall, these results show that our model is able to represent how diatom cells respond in different

environments and is consistent with a variety of empirical observations.

4.3.2 Multiple Internal Processes, Governed by Cellular Costs and Constraints, Control Growth

Internal processes, which are a function of cellular costs, are the proximate causes of limitation. We conducted a set of computational experiments by systematically increasing each model parameter, which allowed us to examine the effect of different internal processes on growth (Fig. 4.2a, Supplementary Figs 4.18, 4.19). As expected, increasing stoichiometric coefficients for micronutrients (e.g. Fe per photosystem unit) had large, negative impacts on growth rates (Fig. 4.2a). However, protein costs (in terms of amino acids), internal rates, and energetic costs also had similarly large impacts (Fig. 4.2a). Moreover, the magnitude by which a given process affected growth differed depending on both dMn and dFe concentrations, illustrating the inadequacy of a simple 'single-resource scarcity' view that underpins many ocean models. Indeed, our results highlight the need to reframe growth as the emergent outcome of internal cellular processes (Fig. 4.2b). This concept is well known in cell systems biology (e.g. Kafri et al., 2016), but it is rarely represented in oceanography (Moore et al., 2013). In our model, growth rate is proportional to the number of biosynthetic pathway units per cell (i.e., all proteins involved in converting nitrate into amino acids; Fig. 4.1a; Faizi et al., 2018), which is, in turn, controlled by (i) available Fe for incorporation as cofactors, (ii) available ribosomes, (iii) sufficient amino acids for protein synthesis, and (iv) sufficient energy (Fig. 4.2c, Supplementary Fig. 4.18). This suite of internal processes simultaneously control growth rate, and the strength of their influence varies under different dFe and dMn concentrations.

The multiplicity of internal processes controlling growth can have significant consequences for cellular stoichiometry and gene expression. For example, under low Mn conditions, synthesis of Mn-containing antioxidants was impeded leading to more oxidative stress (Fig. 4.3a). In our model, the consequence of oxidative stress is damaged proteins. This resulted in increased ribosomes per cell, which maintains total protein synthesis under high oxidative stress (Fig. 4.3a). Ribosomes are a large portion of phytoplankton phosphorus quotas (Elser et al., 1996) and they increase by $\sim$150% as Mn is lowered from 3 nM to 1 nM suggesting that antioxidant allocation and the dynamics of oxidative stress can influence cell macronutrient demands and cellular stoichiometry. This interaction between Fe, Mn and phosphorus around oxidative stress arises because our

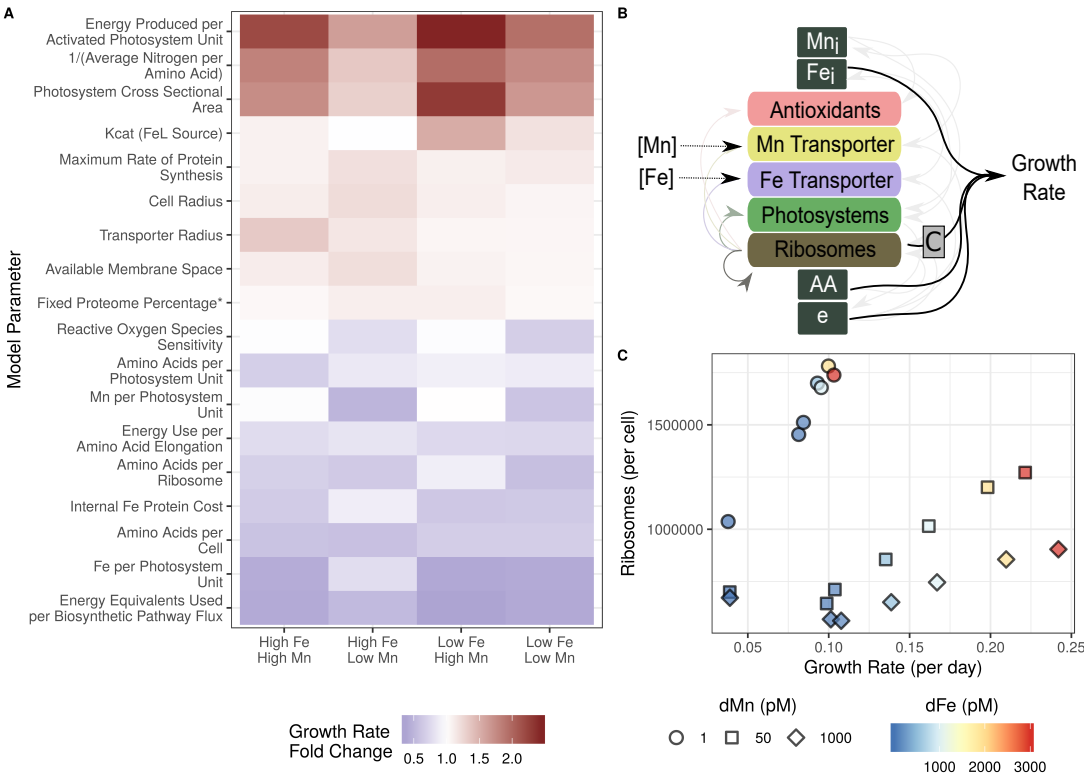

Figure 4.2: Cellular costs and constraints influence growth rate across a range of Fe and Mn concentrations. (**A**), model experiments showing how a five-fold increase in each parameter value influences growth rate, relative to the base model. Note that the parameter 'Fixed Proteome Percentage' is divided by five. (**B**), micronutrient-controlled growth is the outcome of a range of internal processes, simultaneously controlling growth rate (proximate processes controlling growth rate are shown with black arrows). These internal processes are a function of cellular costs and constraints. (**C**), one internal, modelled process directly controlling growth is the number of ribosomes per cell, shown across iron and manganese concentrations.

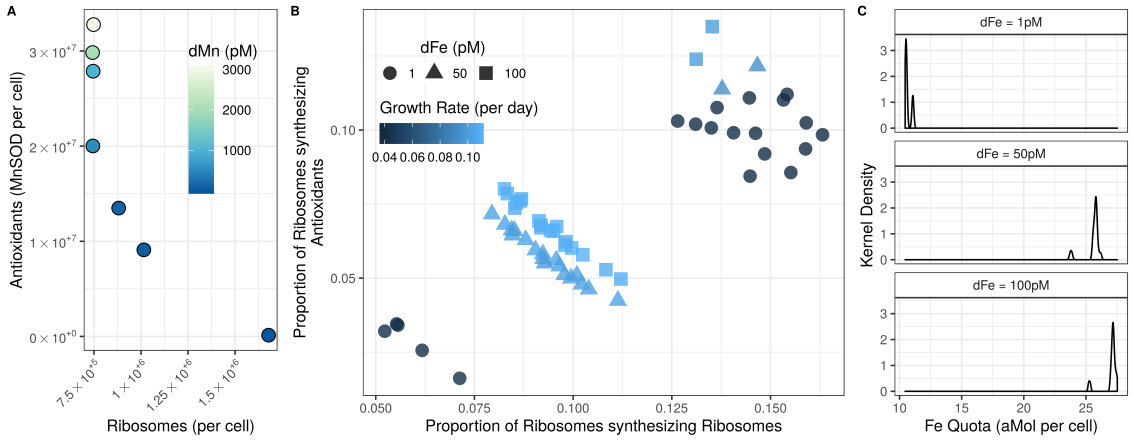

Figure 4.3: Internal processes rearrange to maximize growth rate. (**A**), depletion of dissolved Mn leads to fewer antioxidants. To maintain a sufficient pool of undamaged proteins, the number of ribosomes consequently increased (with constant dFe = 1000 pM). (**B**), examining the distribution of multiple optimization runs revealed a diversity of strategies with similar growth rates (with constant dMn = 1000 pM, n = 20 replicate model runs, and variable dFe displayed as shapes). (**C**), bimodal distributions of total Fe quota per cell, generated from the same optimization runs shown in (**B**), demonstrate another dimension of this antioxidant-allocation strategy (kernel density of distribution shown).

model is able to explicitly represent the internal processes that compensate for each other under micronutrient limitation, which in turn, influences the cellular stoichiometry of Fe, Mn and phosphorus.

Under certain conditions (i.e. low dFe and low dMn, Fig. 4.3b), a diversity of protein allocation strategies to counteract oxidative stress still resulted in similar growth rates in our model. We observed two sources of variation across predictions. First, under low dFe (e.g. at or below 50 pM dFe), there was a trade-off between allocating ribosomes to synthesize ribosomes or antioxidants (Fig. 4.3b). Either approach maintains similar total protein synthesis and growth rates. Second, cells sometimes allocate more ribosomes to Fe transporters and therefore increase the total Fe quota, alleviating electron leakage. This led to a bimodal distribution of Fe quota across these low dFe and dMn conditions (Fig. 4.3c). We predicted a range of strategies with similar growth rates, despite explicitly using an optimization model to explore adaptive hypotheses about protein expression (Parker and Smith, 1990). We speculate that this range of strategies may underlie the diversity of antioxidant systems seen across microbes (Mishra and Imlay, 2012). Furthermore, some variation in microbial metabolic strategies may be due to different configurations of gene expression (with similar cellular costs), yielding similar cellular level outcomes.

4.3.3 Nutrient Interdependence is Influenced by both Nutrient-Specific Costs and Background Costs

We quantified how different cellular processes contribute to interdependence between Mn and Fe, in addition to the explicit interaction via oxidative stress (described above, Methods). Resources such as micronutrients can be considered independent if only a single nutrient controls growth rate and altering the availability of another resource has no impact on the growth rate (in accordance with Liebig's Law of the Minimum). In contrast, interdependence between resources occurs when there are multiple, simultaneously limiting nutrients whose availability affects growth. We used the parameter perturbation experiments conducted at different concentrations of dMn and dFe (as above), and quantified how every parameter influences the strength of interactivity between Fe and Mn (see Methods). Two parameters that exhibited high interactivity were amino acids per ribosome and internal Mn protein cost (Supplementary Fig. 4.19). A higher protein cost per ribosome decreases the growth rate across all conditions, while internal Mn protein cost is only directly related to Mn.

We derived a simple model of an idealized proteome to examine mechanisms of resource interdependence related to these parameters. In this idealized proteome, there are only ribosomes and Mn- and Fe-related proteins (Methods) wherein dFe and dMn control growth by regulating how much of the proteome can be allocated to ribosomes (rather than the micronutrient-specific components). This revealed two mechanisms of interdependence: (i) the global background cost, and (ii) the ratio of Fe and Mn cellular costs. By only increasing the global background cost (analogous to the amino acids per ribosome parameter), interdependence across nutrients is strongly altered by depressing the growth rate across all conditions (Fig. 4.4a-c). In our proteomic allocation model, increasing the amino acids per ribosome parameter led to lower available cellular resources overall, resulting in more interdependence between Fe and Mn. Similar to an ecosystem, when resource availability decreases, competition for this smaller pool of resources increases. In addition to protein synthesis capacity, we hypothesize that this extends to other shared cellular resources (e.g. available membrane space).

Examining the ratio of cellular costs for Mn and Fe showed that maximum interdependence occurs when the cellular costs of Fe and Mn are equal (Fig. 4.4d-f, Methods). For

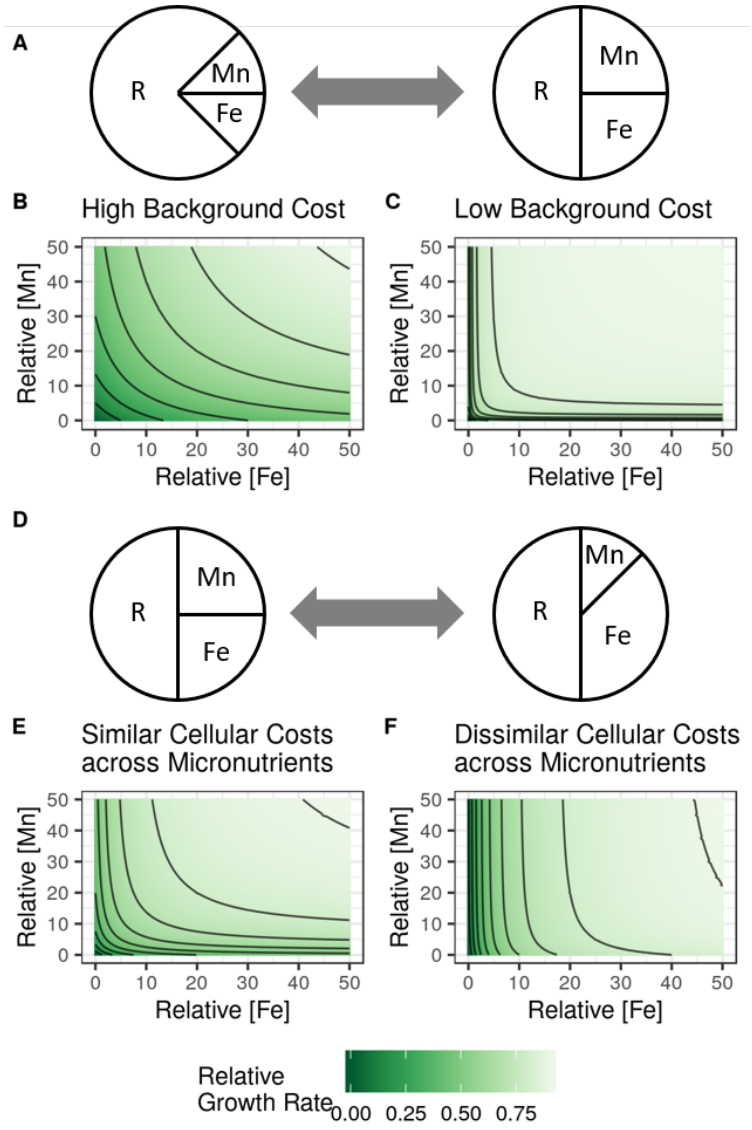

Figure 4.4: Interdependence across micronutrients arises from background cellular costs and the ratio of nutrient-specific costs. (**A**), a phenomenological model of a three-component proteome comprised of ribosomes (R), Mn-related proteins (Mn) and Fe-related proteins (Fe). Growth rate is proportional to dFe and dMn concentrations, and micronutrients influence growth rate by removing potential resources allocated to ribosomes (see Methods for equations). (**B-C**), the background cost of growth (independent of Fe or Mn) can influence the apparent interaction between Mn and Fe (Ψ_{Fe} equal to 0.5, see Methods for equations). Growth rate is lower overall in (**B**) compared with (**C**) because the pie charts represent ribosomal mass fraction (total protein mass in ribosomes), not number of ribosomes, and this corresponds to the parameter perturbation 'amino acids per ribosome' (Supplementary Fig. 4.19). (**D**), the ratio of micronutrient-specific protein costs impacts the apparent interaction between micronutrients (K equal to 5), as shown in (**E-F**). In (**B-C**) and (**E-F**) units are given as relative concentrations, arbitrarily ranging from 0–50.

example, the internal Mn protein cost parameter had a high interaction index (Supplementary Fig. 4.19); when increased, it led to more similar protein costs between Fe and Mn. When cellular costs are similar across resources, then they place similar demands on the pool of shared resources, which consequently increases their interdependence. These two mechanisms provide a tractable means to include interdependence in global ocean models because they suggest that estimating cellular costs for individual nutrients is sufficient to parameterize the overall interaction strength. We speculate that considering relative costs across resources may also apply to other nutrient pairs and help to explain previously observed patterns of interdependencies. For instance, the independent relationship between cobalamin and phosphorus (Droop, 1974) implies large differences in their cellular costs, whereas similar cellular costs between nitrogen and phosphorus may contribute to their interdependence (Harpole et al., 2011).

4.3.4 Inferring *In Situ* Rates and Quotas by Coupling Cellular Modelling with Metaproteomics

While our modeling framework can be combined with proteomic data to estimate the costs and constraints associated with micronutrients, this coupled approach can also be used to predict *in situ* biogeochemical metrics (Fig. 4.5a). In this way, our model is able to quantitatively reproduce growth rates under high and low dFe from a diatom culture (despite no model training on growth rate data, Fig. 4.5b). Using *in situ* dMn and dFe concentrations, and metaproteomes from field samples at the Antarctic sea ice edge, *in situ* diatom-specific growth rates (Fig. 4.5c), Fe cellular quotas (Fig. 4.5d), and Fe uptake rates can be estimated (Fig. 4.5e). These metrics are typically difficult or impossible to measure from *in situ* microbial communities directly, but have important consequences for ocean biogeochemistry and ecosystem services. Our approach connects these rates and quotas directly with resource allocation strategies employed by diatoms, highlighting a decrease in protein allocated to photosynthesis and an increase in protein allocated to iron acquisition in the transition into micronutrient stress (Supplementary Fig. 4.20), resulting in decreased growth rates and iron quotas (Fig. 4.5c-d). These process-based insights are critical for characterizing the role of micronutrients in Southern Ocean phytoplankton bloom progression and fate (Deppeler and Davidson, 2017).

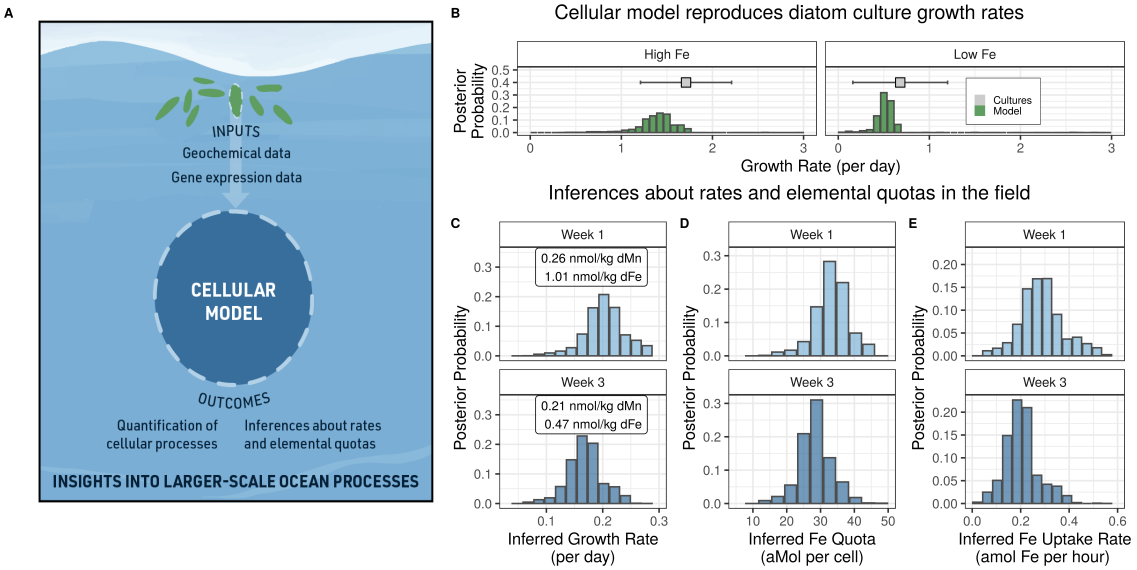

Figure 4.5: By combining cellular modelling with metaproteomic data we inferred *in situ* rates and quotas. (**A**), schematic for combining environmental parameters (e.g. light, dFe), cellular modelling, and metaproteomic observations, to infer rates and elemental quotas. (**B**), we first demonstrated that the proteomic allocation model quantitatively reproduces growth rates from the cultured diatom *Thalassiosira pseudonana* (Nunn et al., 2013) under low and high Fe (culture data do not correspond to a posterior probability, error bars represent the standard deviation across four replicate cultures). (**C-E**), coupling the metaproteome-derived diatom proteome with the cellular model, we can quantitatively infer the growth rates, iron quotas, and iron uptake rates of diatoms in these two time points from a complex microbial community. Week 1 corresponds to higher dFe and dMn, and Week 3 corresponds to lower dFe and dMn (concentrations shown in **C**).

4.3.5 Outlook

We combined mechanistic, proteomic modelling with metaproteomics to estimate the costs and constraints associated with micronutrient-controlled growth in a polar diatom for the first time. Our results highlight the role of cellular costs rather than environmental scarcity in shaping growth, with two key factors: the internal protein cost associated with micronutrient use and the available membrane space for transporters. Identifying the differences in protein cost for vacuolar versus ferritin-based Fe storage, and other micronutrient-associated costs, would further connect ecological strategies with gene expression. Available membrane space has an established temperature dependence and is an important constraint on nutrient uptake kinetics (via membrane saturation; Holton, Blecker and Onorb, 1964; Lis et al., 2015; Held et al., 2020; Casey and Follows, 2020), making it critical to quantify in a changing ocean. Our approach relied on rich *in situ* gene expression datasets to estimate parameters, highlighting a means to quantify cellular costs and constraints.

Parameterizations of phytoplankton growth in global ocean models can have dramatic consequences for projections of ecosystem services in the context of changing upper ocean resource availability (Tagliabue et al., 2020). Embedding a mechanistic representation of resource limitation within global ocean models will leverage rapidly expanding 'omics' datasets to improve predictions of growth responses to environmental change. Developing phenomenological models to represent the outcomes of mechanistic cellular models is a tractable next step. In this way, mechanistic modelling can provide the biological flexibility and realism (e.g. Levine et al., 2016) necessary for predicting potential tipping points in ecosystem services. Mechanistic cellular models, in conjunction with *in situ* gene expression measurements and biogeochemical models, will improve projections of ecosystem services and further characterize the biological underpinnings of nutrient limitation in the changing ocean.

4.4 Materials and Methods

4.4.1 Model Description

We developed a coarse-grained model of intracellular protein allocation in the polar diatom *Fragilariopsis cylindrus* (Mock et al., 2017), extending coarse-grained, kinetic models

previously developed for a range of prokaryotes (Molenaar et al., 2009; Faizi et al., 2018; Weiße et al., 2015; Zavřel et al., 2019). Uniquely, we considered micronutrient controls on proteomic allocation, and applied these principles to a eukaryotic phytoplankton. We used Bayesian optimization to determine the optimal proportion of ribosomes synthesizing different coarse-grained proteomic pools to maximize the steady-state specific growth rate. The cost of producing a given coarse grained pool is a function of the protein length or the sum of protein lengths (in units of amino acids) within a pool. Specifically, the rate of synthesizing one unit of a protein pool is inversely related to the number of amino acids per pool. The units of each intracellular variable (metabolites, proteins, free metal pools) are in molecules per cell. As in Faizi et al. (2018); Zavřel et al. (2019), we used a photosynthetic model (Han, 2001) to parameterize energy production rate, and similarly calculated a biosynthesis specific growth rate. We first provide a high-level overview of the model structure, and then give detailed descriptions of parameterizations.

4.4.1.1 System of Equations

The dynamics of each internal metabolite and protein pool are described using a differential equation, all with growth rate as a loss term (Faizi et al., 2018). The internal free manganese pool (Mn_i) increases with Mn uptake rate (V_{Mn}), and decreases with PSU protein synthesis (photosystem unit, γ_P) at a fixed stoichiometry ($\varphi_{Mn,P}$), and antioxidant protein synthesis (γ_A) at a fixed stoichiometry ($\varphi_{Mn,A}$). We solve this system of equations by integrating them forward in time to a pseudo-steady state (described in more detail in the Supplementary Materials).

$$\frac{\mathrm{d}Mn_i}{\mathrm{d}t} = V_{Mn} - \varphi_{Mn,P}\gamma_P - \varphi_{Mn,A}\gamma_A - \mu[Mn^i] \tag{4.1}$$

The internal free iron pool (Fe_i) is controlled by protein synthesis of Fe-containing protein pools: PSUs and nitrogen metabolism. The fixed stoichiometric coefficient for PSUs is larger for Fe compared with Mn reflecting the higher Fe demand for photosynthesis ($\varphi_{Fe,P}$). Also, the nitrogen metabolism pathway (γ_N) requires a fixed Fe stoichiometry per pathway ($\varphi_{Fe,N}$).

$$\frac{\mathrm{d}Fe_i}{\mathrm{d}t} = V_{Fe} - \varphi_{Fe,P}\gamma_P - \varphi_{Fe,N}\gamma_N - \mu[Fe^i] \tag{4.2}$$

The internal free energy pool e increases with photosynthetic energy production v_e, multiplied by a stoichiometric coefficient that implies a fixed number of ATP and NADPH (reduced form of nicotinamide adenine dinucleotide phosphate) molecules produced per photosynthetic activation (φ_e). There is a small amount of energy required for Mn and Fe uptake (φ_{TMn}, φ_{TFe}), and a large energetic requirement for nitrogen metabolism (φ_N). The total rate of conversion of nitrate into amino acids, V_n, is subsequently used to calculate a biosynthesis-specific growth rate, where V_n is the product of T_{NO_3} and k_{cat,T_N}. Energy is also consumed through protein synthesis. We used the sum of protein synthesis rates for each protein pool multiplied by the amino acids per pool (η_j), and multiplied the entire sum by m_γ, the energetic requirement per amino acid elongation.

$$\frac{\mathrm{d}e}{\mathrm{d}t} = \varphi_e v_{e,adjusted} - \varphi_{TMn}V_{Mn} - \varphi_{TFe}V_{Fe} - \varphi_N V_N - m_\gamma \sum_j \eta_j\gamma_j - \mu[e] \tag{4.3}$$

Amino acids are produced via nitrogen metabolism (V_n); we multiplied this rate by the inverse of the average number of nitrogen atoms within each amino acid. Amino acids are then consumed by protein synthesis and diluted by growth.

$$\frac{\mathrm{d}aa}{\mathrm{d}t} = m_n V_n - \sum_j \eta_j\gamma_j - \mu[aa] \tag{4.4}$$

All protein pools are governed by similar dynamics, such that an increase can only arise from protein synthesis (γ_j) and a decrease from dilution by growth.

$$\frac{\mathrm{d}Protein_j}{\mathrm{d}t} = \gamma_j - \mu[j] \tag{4.5}$$

$$j \in A, P, T_{Mn}, T_{Fe}, T_{NO_3}, R \tag{4.6}$$

4.4.1.2 Internal Protein Cost of Iron and Manganese

We represented the internal cost of iron and manganese by dynamically changing the Fe uptake cost per transporter (n_{TFe}) as a function of dFe uptake rate, scaled by growth rate (i.e. $\frac{V_{Fe}}{\mu}$), and multiplied by a constant coefficient (θ). This approach was similarly applied to Mn uptake and Mn transporter cost. n_{TFe} and n_{TMn} are the uptake and internal management protein costs.

$$n_{TFe} = n_{TFe,unadjusted} + \theta \frac{V_{Fe}}{\mu} \tag{4.7}$$

$$n_{TMn} = n_{TMn,unadjusted} + \theta \frac{V_{Mn}}{\mu} \tag{4.8}$$

4.4.1.3 Nutrient Uptake Kinetics

We modeled nutrient uptake rates of dFe and dMn to include both a variable maximum uptake rate and a diffusion layer (Aksnes and Egge, 1991; Aksnes and Cao, 2011; Fiksen, Follows and Aksnes, 2013; Berg and Purcell, 1977; Zwanzig, 1990). A flexible maximum uptake rate (i.e. V_{max}) has been observed experimentally (Sunda and Huntsman, 1985) and predicted theoretically (Aksnes and Egge, 1991; Bonachela, Raghib and Levin, 2011), and the diffusion layer impacts the total diffusive flux to the cell surface at low bulk substrate concentrations. At high substrate concentrations (i.e. nutrient replete), the nutrient uptake rate approaches the total transporters divided by the 'handling time' (h). Note that handling time (seconds per substrate) is equivalent to the inverse of the maximum turnover rate (k_{cat}) – a commonly measured parameter in enzyme kinetics.

As the substrate concentration decreases (i.e. nutrient deplete), the uptake rate approaches the product of cellular affinity (α) and substrate concentration (S). Affinity is a function of cellular radius, the molecular diffusivity coefficient, and the proportion of cellular area covered by transporters (Aksnes and Cao, 2011). We assumed that Fe and Mn uptake can only be from the dissolved phase, and used a molecular diffusivity coefficient of $0.9 \cdot 10^{-9} m^2 s^{-1}$ for both (Völker and Wolf-Gladrow, 1999). For nitrate, we used a molecular diffusivity coefficient of $1.17 \cdot 10^{-8} m^2 s^{-1}$(Ploug, Stolte and Jørgensen, 1999), with a transporter radius of $1 \cdot 10^{-9} m$ (Aksnes and Egge, 1991).

For modelling multiple nutrient uptake rates simultaneously, we adjusted the nutrient

uptake model above by multiplying the diffusive flux term ($4D\pi r$) by the proportion of surface area covered by other transporters not corresponding to nutrient i (ξ), where D is the diffusivity coefficient, and r is the cellular radius. Given that approximately 50% of a lipid membrane must consist of phospholipids to maintain membrane integrity (Kadner, 1996), and there is a significant requirement for macronutrient transporters, we also restrict the 'available' area for iron and manganese transporters, hypothesizing that a subset of membrane area is available (κ). To model the proportion of membrane space available, we modified the diffusive flux term using the original derivation (Berg and Purcell, 1977). Below, S is the bulk concentration of nutrient i, n_i is the number of transporters for nutrient i, s is the radius of the transporter for nutrient i. Transporters are modelled as circular planes with constant radii on a sphere (Saito, 1968). In addition to the (Aksnes and Cao, 2011) uptake model, we included an additional Michaelis-Menten term of energy dependence.

$$Nutrient\,Uptake\,Rate = V_i = \frac{b}{2a}\left(1 - \sqrt{1 - \frac{4a}{b^2}}\right)\left(\frac{[e]}{K_e + [e]}\right) \tag{4.9}$$

$$b = \frac{1}{\alpha S} + \frac{h}{n_i} \tag{4.10}$$

$$a = \frac{h}{4\pi D r S n_i}\left(1 - \frac{\pi r p}{n_i s}\right) \tag{4.11}$$

$$p = \frac{n_i \pi s^2}{4\pi r^2} \tag{4.12}$$

$$\xi = \left(1 - \sum_{j \neq i} \frac{n_j \pi s^2}{4\kappa \pi r^2}\right) \tag{4.13}$$

$$\xi = \begin{cases} 1 \cdot 10^{-5}, & \text{if } \xi < 0 \\ \xi, & \text{otherwise} \end{cases} \tag{4.14}$$

$$\alpha = 4D\pi r \xi \kappa \frac{n_i s}{n_i s + \kappa \xi \pi r (1 - p)} \tag{4.15}$$

$$i \in [Mn], [Fe] \tag{4.16}$$

4.4.1.4 Iron Speciation

At low concentrations, dFe uptake is bound by physical limits of diffusion to a cell membrane (Hudson and Morel, 1990; Sunda and Huntsman, 1995). Under these conditions, cells are under 'diffusion limitation', as dFe uptake rates are close to the diffusive flux. These studies considered Fe uptake when only Fe' (free, inorganic Fe) was bioavailable and Fe-EDTA is not significantly taken up by eukaryotic phytoplankton (Shaked, Kustka and Morel, 2005). Yet, in the ocean, the majority of dissolved Fe is organically complexed (FeL; Gledhill and Buck, 2012), which is to some extent bioavailable for uptake. We therefore included both sources of iron for uptake. While a large portion of the dFe pool is likely bioavailable, not all dFe species are equally bioavailable (Shaked and Lis, 2012; Shaked et al., 2020). Ligand-bound Fe has maximum uptake rates roughly 3 orders of magnitude lower than Fe' (Shaked, Kustka and Morel, 2005; Shaked and Lis, 2012; McQuaid et al., 2018). We modelled dFe uptake by splitting the dFe pool into subcomponents of Fe' and FeL, and then summing the uptake rates. This formulation assumes that phytoplankton in the ocean are simultaneously under diffusion and 'ligand exchange' limitation (Sunda and Huntsman, 1995). This can be extended to any number of distinct dFe pools with corresponding uptake rate characteristics. Fe' would primarily be controlled by diffusion limitation, and is limited by the chemistry and physics of diffusion to the cell surface, and therefore only affected by α (which is a function of the cell radius, r, diffusivity coefficient of dFe, D, and the proportion of cell surface area covered by transporters). FeL uptake (ligand exchange limitation) is limited by the rate constants of uptake (i.e. the handling time) and the number of transporters.

We first split the dFe pool into Fe' and FeL, by multiplying the bulk concentration of dFe by 2% and 98% respectively (Sunda and Huntsman, 1995). We then use separate kinetic constants, where the maximum turnover rate per transporter of the FeL pool is $k_{cat,Fe'} \cdot 10^{-3}$.

4.4.1.5 Consequences of Reactive Oxygen Species

Reactive oxygen species (ROS) can hamper photosynthesis by negatively impacting protein synthesis (Nishiyama, Allakhverdiev and Murata, 2011). We aimed to capture an overrarching consequence of ROS in phytoplankton cells in this model – damaged proteins. Cells can combat ROS production by producing antioxidants like superoxide dismutase (e.g. Mn/FeSOD), or alternatively manage the consequences by re-synthesizing damaged

proteins. We represented this trade-off in the model by 'leaking' a proportion of energy synthesis from PSUs into superoxide. Superoxide is represented implicitly in the model structure, and not as an internal pool. Superoxide is produced from electrons leaked by photosynthetic energy production and consumed by MnSOD. Excess superoxide that is not consumed by SOD then penalizes the maximum protein synthesis rate, while overinvestment in SOD diverts protein synthesis away from other protein pools. We also model the relationship between electron leakiness and Fe quota (proportion of electrons 'leaked' is ϵ_p below), as previous work suggests the tendency of an electron to be donated to molecular oxygen increases under Fe stress.

Oxidative stress can result from a mismatch between ROS consumption rate (via antioxidants) and production rate (electron transport). We modeled ROS consumption rate as the product of the maximum turnover rate of manganese superoxide dismutase (k_{catROS}, MnSOD) and the number of MnSOD copies per cell (A). The rate of ROS production is a proportion (ϵ_p, see below) of energy production (v_e). Higher rates of energy production require increased investment in MnSOD. The ϵ_a parameter represents the efficacy per MnSOD, and is empirically estimated (described below).

$$v_{ROS} = k_{catROS} \cdot [A] \tag{4.17}$$

$$\omega_u = \frac{\epsilon_p v_e - \epsilon_a v_{ROS}}{\epsilon_p v_e + \epsilon_a v_{ROS}} \tag{4.18}$$

$$\omega = \begin{cases} \omega_u & \text{if } \omega_u > 0 \\ 0 & \text{if } \omega_u < 0 \end{cases} \tag{4.19}$$

An imbalance between production of ROS and available MnSOD (ω) decreases the maximum protein synthesis. We represented this phenomenologically by multiplying the protein synthesis rate by a value ranging from 0 – 1 (p_w). A phenomenological variable R_0 is used here with a value of 10.

$$p_\omega = 2\frac{R_0^{-\omega}}{R_0^{-2\omega}+1} \tag{4.20}$$

Electrons that are 'leaked' not only produce ROS, but they decrease energy production (fewer electrons can be used to create ATP or NADPH). We therefore modified the photosynthetic electron production term above by the proportion of leaked electrons:

$$v_{e,adjusted} = (1-\epsilon_p)\cdot v_e \tag{4.21}$$

4.4.1.6 Photosynthetic Energy Production

We used a previously published photosynthetic model (Han, 2001; Faizi et al., 2018). This model assumes a two-state configuration of photosystem units. We obtained an expression for energy production (as in Faizi et al., 2018), by writing this model as a system of two ordinary differential equations, where the inactivated PSUs are synthesized (γ_P), and both inactivated (P^0) and activated PSUs (P^*) are diluted via growth (μ). The rate of PSU activation is v_1 and the rate of switching back to an inactive PSU is v_e.

$$\frac{\mathrm{d}P^0}{\mathrm{d}t} = \gamma_P - v_1 + v_e - \mu\cdot[P^0] \tag{4.22}$$

$$\frac{\mathrm{d}P^*}{\mathrm{d}t} = v_1 - v_e - \mu\cdot[P^*] \tag{4.23}$$

The rate of PSU activation is a function of the absorption cross section (σ), the amount of irradiance (I), and the amount of inactivated PSUs. The rate of conversion from activated to inactivated PSUs is a function of electron turnover rate (τ).

$$v_1 = \sigma\cdot I\cdot[P^0] \tag{4.24}$$

$$v_e = \tau[P^*] \tag{4.25}$$

We can then assume a pseudo-steady state between the inactivated and activated PSUs, and solve for the energy production rate (v_e).

$$v_e = [P] \cdot \tau \cdot \frac{(\sigma \cdot I)}{\sigma \cdot I + \tau + \mu} \tag{4.26}$$

4.4.1.7 Calculating Growth Rate

We calculated growth rate as in Faizi et al. (2018) with some slight modifications. We calculated a biosynthesis-specific growth rate (Faizi et al., 2018; Weiße et al., 2015; Scott et al., 2010), by calculating the rate of biosynthesis relative to the average protein mass per cell. We assumed a fixed average protein mass per cell (M_{Cell}), using data from Supplementary File S1 in Finkel, Follows and Irwin (2016) for the median picograms of protein per cell from *Pseudonitzschia*, which is converted to amino acids per cell. In our model, biosynthesis rate is represented as the conversion of nitrate into amino acids. Total biosynthesis rate (V_n) is equal to the number of biosynthetic pathways multiplied by rate-limiting enzyme maximum turnover rate (see Model Parameterization).

$$V_n = T_{NO_3} \cdot k_{cat,T_N} \tag{4.27}$$

A proportion of the proteome is considered growth rate independent (Hui et al., 2015). We included a fixed proteomic pool (Λ) in our model which represents 'maintenance metabolism' – respiration, lipid biosynthesis, etc. This is modeled by multiplying the total protein per cell by a constant proportion. We assumed 20% of the proteome is growth rate independent (Metzl-Raz et al., 2017), although future research is required to determine this value in eukaryotic phytoplankton.

$$\mu = \frac{V_N \cdot m_n}{M_{Cell} \cdot (1 - \Lambda)} \tag{4.28}$$

4.4.1.8 Relationship between Fe Quota and Electron Leakage

Previous research suggests that the tendency of an electron to be donated to molecular oxygen increases under Fe stress (Niyogi, 1999). We represented this increased 'leakiness' by designating the proportion of electrons leaked to molecular oxygen, ϵ_p, as a function of the total cellular Fe quota. We constrained this from 5% to 30%; using observations

of total Fe to carbon ratios observed in the SOFEX cruise (Twining, Baines and Fisher, 2004), with a range of 5.5–30 μ mol Fe:mol C. By then using carbon-to-volume ratios from (Menden-Deuer and Lessard, 2000), we converted the lower and upper bounds of μmol Fe:mol C to a total cellular quota (Fe atoms per model cell). A linear relationship between ϵ_p and total Fe quota was assumed, when the Fe quota is within these observationally constrained bounds. Below the minimum Fe cell quota, ϵ_p is fixed at 30%, above the maximum Fe cell quota, ϵ_p is fixed at 5%. This corresponds to the following relationship:

$$\epsilon_p = \begin{cases} 0.3 & \text{if } Fe_i \leq 7173653 \\ 0.05 & \text{if } Fe_i \geq 39129014 \\ 3.561E\text{-}1 - 7.823E\text{-}9 \cdot Fe_i & \text{else} \end{cases} \tag{4.29}$$

Protein Synthesis

Protein synthesis connects the internal pools of metabolites and free micronutrients to proteins:

$$Temperature\ Adjusted\ Protein\ Synthesis = \gamma_T = \gamma_{max} \cdot Q_{10}^{\frac{T-20}{10}} \tag{4.30}$$

$$ROS\ Adjusted\ Protein\ Synthesis = \gamma_{ROS} = \gamma_T \cdot p_\omega \tag{4.31}$$

$$Protein\ Synthesis = \gamma_j = \beta_j \frac{\gamma_{ROS}}{\eta_j}[R]\frac{[e]}{K_e + [e]}\frac{[aa]}{K_{aa} + [aa]} \tag{4.32}$$

$$\gamma_N = \gamma_j \frac{[Fe_i]}{K_{Fei} + [Fe_i]} \tag{4.33}$$

$$\gamma_P = \gamma_j \frac{[Fe_i]}{K_{Fei} + [Fe_i]}\frac{[Mn_i]}{K_{Mni} + [Mn_i]} \tag{4.34}$$

$$\gamma_A = \gamma_j \frac{[Mn_i]}{K_{Mni} + [Mn_i]} \tag{4.35}$$

$$j \in A, P, T_{Mn}, T_{Fe}, T_{NO_3}, R \tag{4.36}$$

In the equations above, γ_{max} refers to the maximum protein synthesis rate, which is a function of temperature (degrees Celsius) with a Q_{10} value of 2 (Toseland et al., 2013). We calculate a ROS-adjusted protein synthesis rate, γ_{ROS}, by multiplying the temperature-adjusted protein synthesis rate by p_ω (ranging from 0 – 1). Protein synthesis to protein pool

j (γ_j) is a function of the proportion of ribosomes allocated (β_j), the protein cost (η_j; larger protein pools have a slower rate of synthesizing one unit), the number of ribosomes (*R*), and the availability of energy (*e*) and amino acids (*aa*). Further, those protein pools that have co-factor requirements have an additional Michaelis-Menten term. All half-saturation constants (K_e, K_{aa}, K_{Fei}, K_{Mni}) used for internal metabolites were set to an arbitrarily low value of 10^4 molecules per cell (implying efficient allocation of resources within the cell).

4.4.2 Model Parameterization

We used the BRENDA database to search for kinetics constants. For the protein lengths, we examined the *F. cylindrus* genome (Mock et al., 2017) and searched for protein coding genes with Gene Ontology terms corresponding to our coarse-grained pools. Generally, the protein cost reflected the length of all proteins within a coarse-grained protein pool. Photosynthetic-specific parameters were taken from previously published datasets.

4.4.2.1 Ribosomal Proteins

To estimate the total proteomic cost per ribosome, we used data from the model alga *Chlamydomonas reinhardtii*. In *C. reinhardtii*, 96 proteins were estimated for cytosolic ribosomes (Manuell et al., 2005). These proteins ranged in size from 12-54kDa. Assuming an average size of 33kDa, this converts to a protein cost of 3168kDa (3168000Da), or 28800 amino acids (using the average molecular mass per amino acid, 110Da). We therefore used 28800 amino acids per ribosome as the fixed protein cost.

4.4.2.2 Photosynthetic Proteins

Our protein cost per photosystem unit was taken as 12177 amino acids per PSU (Wollman, Minai and Nechushtai, 1999), assuming a 1:1 architecture of PSII:PSI. We used the reported approximate molecular mass per photosystem unit (1339.5 kDa) and converted that to amino acids using the average molecular mass per amino acid (110 Da).

4.4.2.3 Fe and Mn Transporters

We searched the *F. cylindrus* genome for Gene Ontology term 'iron ion transport' (GO:0006826). We used the sum of unique proteins identified with this search, excluding ferritin, as we explicitly model that protein (see above). We acknowledge that this approach crudely

approximates the protein requirements for Fe uptake, as the exact protein stoichiometry and the specific combination of proteins required is still unclear. We also included the average of the four copies of *FBP1* identified in *F. cylindrus* (Coale et al., 2019). The total cost per transporter for Fe uptake was 4028 amino acids.

Four natural resistance-associated macrophage proteins (NRAMPs) were identified as manganese transporters in the *F. cylindrus* genome (Blaby-Haas and Merchant, 2017). The average protein length per NRAMP was 372 amino acids.

4.4.2.4 Nitrate Uptake and Amino Acid Biosynthesis

We represented the transformation pathway from nitrate to amino acids as the core iron-dependent biosynthetic cellular pathway. This pathway is represented in our model as a single unit with high protein, energetic, and iron costs. We combined the protein lengths of nitrate transporters (NRT2 transporters), nitrate reductase (represented as a homodimer), nitrite reductase, glutamine synthetase, and glutamate synthase, which sums to 5893 amino acids per pathway.

At substrate saturating conditions assuming fixed pathway stoichiometry, the enzyme in a pathway with the lowest maximum turnover rate (k_{cat}) determines the upper bound on pathway flux. We used this 'kinetic bottleneck' approximation to describe the conversion of nitrate into glutamine. For the enzymes described above, we found that glutamine synthetase had the lowest k_{cat} for NH_4 ($2.96sec^{-1}$, Enzyme Commission number 6.3.1.2.; Ishiyama et al., 2006), and we therefore use this value to represent the rate limiting step.

We approximated the energetic requirement for the entire conversion by summing up the ATP and NADPH cofactors required for each step in the synthesis of glutamine from imported nitrate. We accounted for 1 ATP from nitrate uptake, 1 NADPH for nitrate reduction, 1 NADPH for nitrite reduction, 1 ATP for glutamine synthetase, and 1 NADPH for glutamate synthase. Assuming an interconversion ratio of 2.6 ATP to 1 NADPH, the total energetic cost was 9.8 e.

For the Fe requirement in this pathway, we summed up the per-enzyme atoms of Fe. We accounted for 2 Fe atoms in nitrate reductase (one per subunit, but it exists as a homodimer), 5 Fe atoms in nitrite reductase in total (1 siroheme cofactor and 4 in 4Fe-4S cluster), and 3 Fe atoms in glutamate synthase. Thus the total stoichoimetric coefficient for this pathway is 10 Fe atoms ($\varphi_{Fe,N}$).

4.4.2.5 Uptake Rate Kinetic Constants

To obtain kinetic constants for Fe transporters, we leveraged previously published data and methods for inferring maximum uptake rate per transporter. Hudson and Morel (1990) derive a kinetic constant for the maximum turnover rate per transporter, equivalent to the inverse of the handling time, by using pulse chase experiments with labelled Fe. They assume that the whole-cell response of uptake kinetics approximates that of the kinetic constant of the transporter, which, in other words, means that there is no downstream regulation of Fe uptake beyond that of the transporter (i.e. internalization kinetics and saturation). Yet, enzyme kinetics can be regulated at the pathway level (Button, 1998), therefore we challenge the assumption of no downstream regulation from Fe uptake. Indeed, comparing the magnitude of the maximum turnover rate per transporter reported in Hudson and Morel (1990) to other nutrient transport kinetics, but derived differently (Fiksen, Follows and Aksnes, 2013), suggests that using pulse chase experiments to estimate transporter kinetic constants underestimates these constants because of downstream inhibition. However, Hudson and Morel (1990) still provides invaluable measurements of cell-specific uptakes rates that can be used to infer kinetic constants.

We leveraged published uptake rate data (Hudson and Morel, 1990), and recalculated the maximum turnover rate using a method described in Fiksen, Follows and Aksnes (2013), equation 16 in this reference. This resulted in a k_{in} value approximately 3 orders of magnitude higher than inferred in Hudson and Morel (1990), which was much more similar to values estimated for macronutrient transporters (Fiksen, Follows and Aksnes, 2013). We used the following values from Hudson and Morel (1990) to recalculate the handling time: maximum uptake rate (V_{max}) of 180 amol cell^{-1} hour^{-1}; half saturation constant (K_m) of 3.1 nM; diffusion coefficient (D) of $5.4 \cdot 10^{-8}$ m^2 minute^{-1}; a cell radius (r) of $5.6 \cdot 10^{-6}$ m, and transporter size (s) of 1^{-9}m.

4.4.2.6 Protein Synthesis Parameters

We used the translation rate from *Thalassiosira weissflogii* at 20 °C of 1.9 amino acids per ribosome per second (Toseland et al., 2013). Assuming a temperature dependence given by a factor of Q_{10} equal to 2 (Toseland et al., 2013), protein synthesis rate is adjusted in the model according to the input temperature (-1 °C for the metaproteomic conditions). For the energy required per amino acid elongation we used the equivalent of 3 e units (Faizi

et al., 2018).

4.4.2.7 Photosynthetic Parameters

We needed two parameters for the photosynthetic energy production model: the absorption cross section, and the rate of returning from an activated PSU to an inactivated PSU (τ). For the absorption cross section we used a value of 0.01 m^2 μE^{-1} (Strzepek et al., 2012), for the PSU turnover rate we used a value of 6000 $minute^{-1}$ (Strzepek, Boyd and Sunda, 2019).

4.4.3 Culture Diatom Comparison

We used two published diatom datasets that examined how Fe influence the proteome (Nunn et al., 2013; Cohen et al., 2018). The observed protein data from both datasets were manually binned into our corresponding model coarse grains. For Nunn et al. (2013), we used the sum of spectral counts per peptide as an approximation for the mass per protein group. For Cohen et al. (2018), we used the reported Normalized Spectral Abundance Factors (NSAF values) per protein. Note that while both of these datasets used diatoms, the studied diatoms were *Thalassiosira pseudonana* and *Pseudo-nitzschia granii*, and our model is based off of the polar diatom *Fragilariopsis cylindrus*.

To compare the model predictions under these laboratory conditions, we also modified the temperature and light level inputs to the model to reflect the culture conditions. Importantly, the Fe levels in culture were set with EDTA, and most Fe taken up in culture with FeEDTA is inorganic free Fe. Therefore, we changed the Fe speciation input to reflect this, such that there is only a small available FeL pool (1%) while the inorganic free Fe pool was set to 99% of total dFe.

4.4.4 Southern Ocean Mn, Fe, and Light Conditions

4.4.4.1 FISH Data

Surface seawater (approximately 3 m depth) was pumped from a tow FISH into a clean container using a Teflon diaphragm pump (Almatec A15) connected to a clean oil-free air compressor (JunAir) and GEOTRACES cruise JR274 (Achterberg et al., 2001).

Concentrations of trace metals were determined by isotope dilution inductively coupled mass spectrometry (ID-ICP-MS), whilst the mono-isotopic elements Co and Mn were analysed using a standard addition approach followed by ICP-MS detection; all according

to methods described in Rapp et al. (2017). The ICP-MS analyses were conducted following an off-line preconcentration/matrix removal step Rapp et al. (2017) on a WAKO chelate resin column (Kagaya et al., 2009).

4.4.4.2 GEOTRACES Data

We used the GEOTRACES intermediate data product (Schlitzer et al., 2018) to determine average Mn and Fe concentrations within the mixed layer for cruise stations in the Southern Ocean. To calculate the mixed layer depth, we calculated the potential density at 10 m, and determined the depth at which this 10 m potential density is 0.03 kg m^{-3} more dense (de Boyer Montegut et al., 2004). For each station, we used the discrete data product and averaged the Fe and Mn concentrations above the mixed layer depth.

We also calculated the median light level (Photosynthetically Active Radiation, PAR) within the mixed layer. We used monthly climatology of surface PAR and diffuse attenuation coefficient (K_d490) from the Ocean Color database from 2002–2018. The median mixed layer light levels were determined using the surface PAR, K_d490, and mixed layer depth (Behrenfeld et al., 2005):

$$I_g = I_0 \cdot exp(-K_d490 \cdot MLD/2) \tag{4.37}$$

Where MLD is the inferred mixed layer depth and I_0 is the surface irradiance.

4.4.5 Metaproteomic Sampling and LC-MS/MS

We sampled the microbial community at the sea ice edge in McMurdo Sound, Ross Sea at the same location (-77.62S, 165.41E) for four weeks (as described in Wu et al., 2019). We had four sampling dates corresponding to weeks 1 to 4: December 28 2014, January 6, 15, and 22 2015. Large volumes of water (150–250 L) were filtered from 1 m depth at the sea ice edge, and passed through three filters sequentially (3.0, 0.8, and 0.1 μm, each 293 mm Supor filters). Filters with collected biomass were then placed in tubes with a sucrose-based preservative buffer (20 mM EDTA, 400 mM NaCl, 0.75 M sucrose, 50 mM Tris-HCl, pH 8.0) and stored at -80 °C until sample processing. We extracted proteins after buffer exchange into a 3% SDS solution as previously described (Wu et al., 2019).

To prepare samples for LC-MS/MS, the precipitated protein was resuspended in 100 μL 8 M urea, and then we ran a Pierce bicinchoninic acid Protein Assay Kit (Thermo

Fisher Scientific) to quantify the protein concentration in each sample. We then reduced the protein sample using 10 μL of 0.5 M dithiothreitol, and incubated the sample for 30 minutes at 60 °C. Samples were then alkylated using 20 μL 0.7 M iodoacetamide in the dark for 30 minutes, diluted with 50 mM ammonium bicarbonate, and digested with trypsin using a 1:50 trypsin:protein ratio. We then acidified (1.5 μL trifluoroacetic acid (TFA) and 5 μL formic acid added) and desalted samples. We desalted the samples by first conditioning the solid-phase columns with methanol (1 mL), then 50% acetonitrile (ACN) and 0.1% TFA, and then 2x 1 mL of 0.1% TFA. Samples were loaded onto columns that were subsequently washed 5x with 1 mL 0.1% TFA. Finally, peptides were eluted from the columns with 2x 0.6 mL 50% ACN 0.1% TFA, and 1x of 0.6 mL 70% ACN and 0.1% TFA.

We used a one-dimensional liquid chromatography tandem mass spectrometry to characterize the metaproteome. For the largest filter size (3.0 μm) we used three injections per sample, and two injections per sample for the 0.8 and 0.1 μm filters. We ensured that the protein concentration in each urea-resuspended sample was equivalent across sampling weeks and within each filter size. We used a LC gradient from 0 to 10.5 minutes with 0.3 μL per minute flow of 5% solution B, from 10.5 minutes to 60 minutes the flow was 0.25 μL per minute and solution B increased to 25.0%, from 60–90 minutes %B increased to 60%, from 90–97 minutes %B increased to 95%, from 97–102 minutes %B remained at 95%, from 102–105 the flow rate increased to 0.3 μL per minute and %B decreased to 5% for 20 minutes. Solution A is 0.1% formic acid in water, and solution B is 0.1% formic acid in ACN. Peptides were injected onto a 75 μm × 30 cm column (New Objective, Woburn, MA) self-packed with 4 μm, 90 Å, Proteo C18 material (Phenomenex, Torrance, CA), and then online LC was performed using a Dionex Ultimate 3000 UHPLC (Thermo Fisher Scientific, San Jose, CA).

We used a data-dependent acquisition approach with a VelosPRO Orbitrap mass spectrometer (MS; Thermo Fisher Scientific, San Jose, CA) to characterize the metaproteome for each sample. We used an MS method with the following parameters: dynamic exclusion enabled, with an exclusion list of 500 and an exclusion duration of 25 seconds; a *m/z* precursor mass range from 300–2000 *m/z*; and a resolution of 60000. MS2 scans were collected with a TopN method (N = 10), using Collision-Induced Dissociation with a normalized collision energy of 35.0, an isolation width of 2.0 *m/z*, a minimum signal of

30000 required, and a default charge state of 2. Ions with charge states less than 2 were rejected, and those above 2 were not rejected. Lastly, we used polysiloxane as a lock mass.

For a database of potential proteins present, we used a metatranscriptome obtained from a nutrient incubation experiment conducted using water collected during week 2 of protein sampling (Jabre et al., 2021). Prior to database searching we removed all redundant protein sequences (P. Wilmarth, fasta-utilities), and appended the Global Proteome Machine Organization common Repository of Adventitious Proteins database of common laboratory contaminants. We then applied a Savitzky-Golay noise filter, a baseline filter, and applied a high-resolution peak picking approach to centroid the MS data (Weisser et al., 2013). To identify peptides, we conducted a database search with MSGF+ (Kim and Pevzner, 2014). We used a 1% False Discovery Rate at the peptide-spectrum match level. Once we had identified peptides within each MS injection, we quantified these peptides at the MS1 level using the 'FeatureFinderIdentification' approach (Weisser and Choudhary, 2017), where peptides identified in one injection can aid identifying peptides in a different injection without a corresponding MS2 spectra. In this approach, the user must identify a group of samples across which peptides can be cross-mapped. We grouped our samples by filter sizes, with replicate injections also within each group for cross-mapping. Mass spectrometry mzML files within each group were then aligned using MapAlignerIdentification (Weisser et al., 2013), and then we applied FeatureFinderIdentification to obtain peptide-specific MS1 intensities. Once peptides were quantified for each injection, we then obtained a sample-specific peptide quantity, which was the average peptide-specific intensity across injections. We only used this quantity if a given peptide was observed across all injections.

We then mapped peptides to taxa and to protein functions. Peptides were mapped to taxa only if they uniquely correspond to a given taxonomic group. Coarse taxonomic groups (presented at the Phylum level) were chosen, because coarse-graining is robust to various MS-induced biases (McCain and Bertrand, 2019). We suggest that the sum of taxon-specific peptide abundances (MS1 intensities in this case) can be used as a proxy for biomass. To evaluate this approach, we used a previously published, artificially assembled metaproteome (Kleiner et al., 2017). In this dataset, we identified all taxon-specific peptides, and then examined the correlation between the amount of protein used for a taxonomic group and the sum of peptide intensities that correspond to that taxa. We found a

high correlation between the sum of peptide intensities and the total protein (Supplementary Fig. 4.6). Additionally, we examined different mass spectrometry chromatographic methods (data files 'Run1and2_U1.pep.xml' and 'Run4and5_U1.pep.xml' from Kleiner et al., 2017). We show that there is a high correlation between the amount of protein and the sum of peptide intensities across three orders of magnitude (Supplementary Fig. 4.6), and this correlation is higher in the longer chromatographic run.

Mapping peptides to taxon-specific functional groups has additional challenges because there can be multiple functional labels for a given protein, and the functional label can differ based on the annotation used. To address this issue, we used five different functional annotations (KEGG, KO, KOG, Pfams, and TIGRFAM annotations; Kanehisa and Goto, 2000; Kanehisa et al., 2016; Tatusov et al., 2003; Mistry et al., 2021; Haft, Selengut and White, 2003), and mapped coarse-grained functional associations by matching a list of strings, i.e. keywords (which were identified in the construction of the model). In addition, we manually examined the matched proteins to ensure we were not capturing incorrectly mapped proteins to coarse-grains.

4.4.6 Approximate Bayesian Computation for Parameter Estimation

4.4.6.1 Metaproteomic-to-Model Data Comparison

To infer parameters of the model given the proteomic data, we need to determine how similar the observations are to the model predictions. However, there are several challenges associated with comparing the proteomic data with the protein allocation model output. The main challenge with doing a direct comparison of model output (i.e. with protein mass fraction) are the components of the observed proteome that we are not modelling. For example, we do not include DNA synthesis proteins in our cellular model, yet we anticipate this protein mass fraction to vary with growth rate. The consequence of this issue is a poor model fit, which can hamper parameter inference.

We propose a general approach to address this challenge using the ratio of the protein pool abundance from the two conditions observed. By using this ratio, we can still capture the change in protein expression across conditions, but we bypass the issue of the non-modelled proteome. Specifically, we used the ratio of protein group abundance from the low Fe to high Fe condition in the culture diatom proteomes, and the third sampling point to the first sampling point from the metaproteomic time series. This general approach for model-to-metaproteome comparisons might be useful in other contexts, as we anticipate

this issue would be pervasive, because no proteomic allocation model can explicity include all proteins synthesized.

There are also several transformations and considerations required to make comparisons between the model and the observations. The first transformation is to calculate protein mass fraction from the model. The true mass fraction from our model considers the free amino acid pool, yet this pool would not be observed using typical proteomic methods. Thus, we first re-calculate the total observable protein mass from the model. This is done by multiplying all protein abundances by the amino acids per protein pool – for Fe and Mn uptake, this cost is dynamic, so we re-calculate the dynamic cost per transporter and internal machinery first. Once we have re-calculated protein mass, the next consideration is the observed proteins. This is straightforward for all the protein pools except for Fe and Mn uptake and internal cost. This is because the observed proteins for this protein pool can be considered part of the internal or external protein pool (or both). For each of the datasets, we examined the Fe transporters and internal Fe cost proteins and determined if it is appropriate to use the external or internal protein pool from the model as a comparison. We did not observe ferritin in any dataset, the main protein observed for this protein pool was phytotransferrin (ISIP2a). We considered phytotransferrin to be *both* an internal and an external cost, given that the protein is endocytosed (McQuaid et al., 2018). These transformations for each dataset enabled a careful comparison between the data and the observations.

4.4.6.2 Approximate Bayesian Computation for Parameter Inference

We used Approximate Bayesian Computation to draw inferences about the three unconstrained parameters in the model: the efficacy per MnSOD, ϵ_a; the available space on the membrane for Mn and Fe transporters, κ; and the internal Fe and Mn cost coefficient, θ. Note that we assumed that θ is constant for both Mn and Fe, although with additional data we would be able to further discriminate across these costs.

We used ABC to obtain posterior distributions for parameters and predictive distributions for observed data. The stochastic model was combined with our cellular model to allow for errors in approximation. To obtain a posterior distribution for each parameter, we accounted for error in the model and observations. Specifically, our cellular model (f) generates observations (y_i) from a vector of parameters ($\nu_i = (\epsilon_{ai}, \kappa_i, \theta_i)$). We included an error term (e_i), which we assume is normally distributed with a common standard deviation

(h).

$$y_i = f(\nu_i) + e_i \tag{4.38}$$

$$e_i \sim N(0, h^2) \tag{4.39}$$

We treated the standard deviation h as fixed; and estimation of the posterior distribution is described below. The prior for the elements of $\nu = (\epsilon_a, \kappa, \theta)$ were independently uniform, $\epsilon_a \sim U(0.00001, 0.1)$, $\theta \sim U(0.001, 16)$ and $\kappa \sim U(0.001, 0.15)$. For ϵ_a, we drew from a uniform bounded by 0.00001 and 0.1, because initial tests suggested that this range resulted in a Mn:Fe ratio, and a Mn-PSU:Mn-SOD ratio consistent with empirical observations (Nunn et al., 2013). For θ, we drew from a uniform bounded by 0.001 and 16. The upper bound is assuming all internal Fe is stored in ferritin – which would result in a very high internal Fe cost. The lower bound represents an arbitrarily low protein cost. For κ, we used a lower bound of 0.001 and an upper bound of 0.15. We hypothesized that the proportion of membrane space available for Fe and Mn transporters is likely within these bounds – considering that only approximately 50% of the membrane can even have transporter proteins (Kadner, 1996), and there must be a large proportion dedicated to macronutrient transporters.

We used an ABC algorithm to approximate the exact posterior (Fearnhead and Prangle, 2012; Wilkinson, 2013). The approach simulates $\nu_1, \ldots, \nu_B$ from the uniform priors and then generates $y_1 = f(\nu_1), \ldots, y_B = f(\nu_B)$ from the cellular model. For each y_i, a weight $w_i(h) \in \{0, 1\}$ is generated from a Bernoulli distribution $(a_i(h))$ where:

$$a_i(h) = exp[\frac{-||y_i - y_0||^2}{2h^2}] \tag{4.40}$$

For any function $g(\nu)$, its posterior expectation is approximated by:

$$E[g(\nu)|y_0] \approx \frac{\sum_{i=1}^{B} g(\nu_i) w_i(h)}{\sum_{i=1}^{B} w_i(h)} \tag{4.41}$$

We can determine $P(\nu_j \leq t|y_0)$ at each point along the grid, and then convert these

estimates from a cumulative distribution to a probability density. We do so by calculating the height of the kth bin $[t_{k-1}, t_k]$ as:

$$P(\nu_j \leq t_k|y_0) - P(\nu_j \leq t_{k-1}|y_0) \tag{4.42}$$

Intuitively, if the Euclidean distance of a simulated dataset y_i to y_0 is very low, then it is very likely that the parameter vector ν_i would be included in the posterior. This approach gives an approximation of $E[g(\nu)|y_0]$ with a fixed and known h. We then conducted the posterior sampling M times to infer the approximate posterior distribution at fixed intervals (i.e. the posterior as a histogram, with M = 400000). Overall, this method allows for a probabilistic sampling of the posterior, as we transform our deterministic model output to a stochastic model, with the stochasticity coming from the error term (Wilkinson, 2013). Without this step, our posterior variance estimates would solely be a function of the tolerance that we use for inclusion in an approximate posterior (Alahmadi et al., 2020).

After sampling from the prior distribution (182171 samples drawn), we ran the cellular model and generated a set of model outputs for each of three datasets: the metaproteome-derived diatom proteome at two time points with corresponding *in situ* dMn and dFe concentrations, *Thalassiosira pseudonana* proteome under high and low Fe (Nunn et al., 2013), and *Pseudonitzschia granii* diatom proteome under high and low Fe (Cohen et al., 2018). We then compared the model output with each of these datasets, and the success probabilites given above are calculated by combining observations and model predictions across all three datasets. We combined these datasets to estimate the first order effects, however it is possible that the parameters included are environment dependent. For example, the temperature is different across each dataset, yet we assume the membrane space parameter (κ) is from a single distribution. Numerical integration and optimization parameters were adjusted to enable faster sampling of parameter space, specifically, we shortened the integration time to a length of 1x10^6 (with steps of 10). Optimization settings are given in the Supplementary Materials.

4.4.6.3 Estimating Standard Deviation of the Error Term: *h*

The standard deviation of the error term, *h*, is an important parameter for conducting ABC. This error term encompasses error from mass spectrometry, sample processing, and natural

biological variability. We empirically estimated this parameter by using culture replicates from Nunn et al. (2013). We calculated the average standard deviation of the ratio of protein pools across replicates. To do so, we randomly paired biological replicates and determined the sample standard deviation:

$$h = \sqrt{\frac{\sum_{i=1}^{N} (y_i - y)^2}{N - 1}} \tag{4.43}$$

We inferred an average sample standard deviation (across all pairs of biological replicates) of 0.007. However, with such a low standard deviation, our ABC approach was not feasible because the probability of acceptance was so low across all parameter vectors. We therefore increase the value of *h* to a conservative value of 2, likely overestimating the standard deviation of the error distribution.

4.4.7 Model Settings, Parameter Perturbation Experiments, and Interaction Index

We generated model output for a range of dFe and dMn values: 1, 50, 100, 500, 1000, 2000, and 3000 pM in a full factorial combination, with light levels set to 50 $\mu E\, m^{-2} s^{-1}$. For the three unconstrained parameters (described above), we used the modes of their posterior probability distributions for the inferred parameter value. We then conducted 20 replicate model runs for each unique combination of dFe and dMn with the following settings: nitrate at saturating conditions (input nitrate set to arbitrarily high concentration of 1x10^9 nM, note that our 'kinetic bottleneck' approximation is satisfied only under saturating conditions), and an integration time 3x10^6 for the second stage of optimization (with steps of 10, see Supplementary Materials for additional details on optimization settings).

We multiplied every parameter individually by five and examined the change in growth rate (except the 'Fixed Proteome Percentage' parameter, which was divided by five because the base value was 20%). Each perturbation experiment was conducted three replicate times, and the average growth rate of these three was then divided by the base model (i.e. with no parameters altered, also run three times). Four environmental conditions were chosen for parameter perturbation experiments, corresponding to high and low dFe and dMn (all combinations of these conditions). The high dFe and dMn conditions were set to 3000 pM. The low conditions were determined by fitting a Monod-style growth function to

modelled growth rates (Supplementary Figs 4.16, 4.17), and then using the half-saturation constants. For dMn, this corresponded to 1.42 pM, for dFe, this corresponded to 88.9 pM.

We used the parameter perturbation experiments and the following equation to obtain a quantitative metric of how different cellular processes contribute to interdependence between dFe and dMn (μ corresponds to the growth rate):

$$Interaction\,index = min(\mu_{HighFe,LowMn}, \mu_{LowFe,HighMn}) - \mu_{LowFe,LowMn} \tag{4.44}$$

4.4.8 A Phenomenological Model of Nutrient Interdependence

Scott et al. (2010) develop a phenomenological model connecting growth rate with gene expression. We extended a similar framework to micronutrients, and explored interdependence across elemental metabolisms using this framework. Consider a three-component proteomic model, ϕ_R (ribosomal mass fraction), ϕ_{Fe} (Fe-metabolism protein mass fraction), and ϕ_{Mn} (Mn-metabolism protein mass fraction). Scott et al. (2010) suggest that:

$$\mu \propto \phi_R \tag{4.45}$$

Under high micronutrient concentrations, we anticipate that the proteomic mass fraction required to acquire these nutrients decreases, such that ϕ_{Fe} and ϕ_{Mn} are inversely proportional to the amount of dFe and dMn:

$$\phi_{Fe} \propto \frac{1}{dFe} \tag{4.46}$$

$$\phi_{Mn} \propto \frac{1}{dMn} \tag{4.47}$$

It follows that the increased mass fraction required for processing and obtaining Fe and Mn negatively influences the growth rate via decreasing the ribosomal mass fraction (ϕ_R):

$$\mu \propto 1 - \phi_{Fe} - \phi_{Mn} \tag{4.48}$$

If we then assume a saturating function of the proteomic mass fractions ($\phi_{Fe,Mn}$) dependent on the micronutrient concentration:

$$\phi_{Fe} = 1 - \frac{dFe}{dFe + K} \tag{4.49}$$

Where K is the half saturation constant for Fe, and is equivalent (for simplicity) to the half saturation for Mn. We then obtain an expression for the growth rate:

$$\mu \propto 1 - \left(1 - \frac{dFe}{dFe + K}\right) - \left(1 - \frac{dMn}{dMn + K}\right) \tag{4.50}$$

However the expression above requires some proteomic cost weighting factor, otherwise the expression could result in negative growth rates. If we define the proteomic cost weight for Mn and Fe to be ψ_{Mn} and ψ_{Fe}, we obtain:

$$\mu \propto 1 - \psi_{Fe}\left(1 - \frac{dFe}{dFe + K}\right) - \psi_{Mn}\left(1 - \frac{dMn}{dMn + K}\right) \tag{4.51}$$

$$\psi_{Mn} + \psi_{Fe} = 1 \tag{4.52}$$

We found that when protein costs have similar values for Fe and Mn (i.e. ψ_{Fe} is 0.5), the apparent amount of interdependence is greatest. This is demonstrated by looking at the gradient of the growth rate function with respect to dMn and dFe:

$$\nabla\mu = \begin{bmatrix} \frac{d\mu}{d(dFe)} \\ \frac{d\mu}{d(dMn)} \end{bmatrix} = \begin{bmatrix} \frac{K\psi_{Fe}}{(dFe+K)^2} \\ -1\frac{K(\psi_{Fe}-1)}{(dMn+K)^2} \end{bmatrix} \tag{4.53}$$

If we then assume that both dFe and dMn are at concentrations equivalent to their half-maximal growth (i.e. K), the gradient function then simplifies to:

$$\nabla\mu = \begin{bmatrix} \frac{\psi_{Fe}}{4K} \\ -1\frac{(\psi_{Fe}-1)}{4K} \end{bmatrix} \tag{4.54}$$

From this function, we can then show that the direction which corresponds to the most elemental interdependence is when the slope is closest to one, or, that the elements of the gradient with respect to dFe and dMn are equivalent. When we equate the elements of the gradient, evaluated when dFe and dMn are at half maximal growth, we find that the iron cost parameter is equivalent to 0.5 and therefore so is the manganese cost parameter. This phenomenological model suggests that when the ratio of proteomic costs are similar, the extent of elemental interdependence is greatest. Note that 'proteomic' cost in this case can be extended to other cellular costs, for example available membrane space.

4.5 Data Availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials, and all data are publicly available. Model runs, experiments, and output from metaproteomic bioinformatics are deposited with in Dryad and can be found at https://doi.org/10.5061/dryad.xd2547dfs. Model parameters used are provided in a table (Supplementary Table 4.1) in addition to being described in the Methods and Supplementary Information. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD022995 (Perez-Riverol et al., 2019). Code for all metaproteomic analyses is available from https://github.com/bertrand-lab/ross-sea-meta-omics, and code for all other analyses is available from https://github.com/bertrand-lab/mn-fe-allocation.

4.6 Author Contributions

J.S.P.M., A.T., E.P.A., and E.M.B. conceived the study; J.S.P.M. wrote all the code and conducted all analyses and lab work, with input from A.T., E.P.A., E.S., and E.M.B.; E.S. specifically contributed to the design of the ABC analysis; E.P.A. contributed data with Southern Ocean trace metal concentrations; E.M.B. and A.E.A. conducted the Antarctic sample collection and generated the reference metatranscriptome and annotation; J.S.P.M., A.T., E.P.A., and E.M.B. wrote the paper with input from E.S. and A.E.A.

4.7 Supplementary Information

4.7.1 Model Parameters

Table 4.1: Mathematical symbols, names, units, numerical values, and detailed descriptions for all the parameters in the proteomic allocation model described in the main text.

Parameter Symbol	Parameter Name	Units	Value	Description
$\eta_{T_{NO_3}}$	Nitrogen Metabolism Protein Complex Size	$\frac{AminoAcids}{Molecule}$	5893	There are nine NRT2 transporters found in the *F. cylindrus* genome. The average length is 512. There are three proteins in the *F. cylindrus* genome that are associated with the GO term 'Nitrate reductase activity'. After pBLAST, these proteins were further identified as one nitrate reductase and two cytochrome b5. There is one cytochrome b5 domain in nitrate reductase, and nitrate reductase also functions as a homodimer. We take the average length of the two cytochrome b5 proteins (136 and 122 amino acids), added with the nitrate reductase protein length (862), to a amino acid total of 991 (multiplied by two, 1982). There are four enzymes associated with nitrite reductase activity (three with ferredoxin-nitrite reductase and one with NADPH dependent nitrite reductase). We took the average length of all NiR proteins (832 amino acids). There are five proteins in the *F. cylindrus* genome that associate with the GO term 'Glutamine biosynthetic process' (GO Term: 0004356). After pBLAST, two appear to be glutamine synthetase/quinado kinase, two are plastid glutamine synthetase II, and one is glutamine synthetase type. The average protein length is 450. Glutamine is converted to glutamate by glutamate synthase, and we found 10 proteins with 'glutamate synthase activity' via GO terms. These proteins were then searched using pBLAST, and of them, there were four hypothetical proteins, two FMN-linked oxidoreductases, two 'glutamate synthase family proteins', one glutamate synthase (NADPH) large chain protein, and one NADH-glutamate synthase small subunit. We took the average length of the glutamate synthase family proteins and large subunit (1596), and summed it with the small subunit protein (521), to a total of 2117 amino acids. The total amino acid count fo protein complex size is 512 + 1982 + 832 + 450 + 2117 = 5893.
η_R	Ribosome protein complex size	$\frac{AminoAcids}{Molecule}$	28800	96 proteins estimated for cytosolic ribosomes in *Chlamydomonas reinhardtii* (Manuell et al., 2005). These proteins range in size from 12-54kDa. Assuming an average size of 33kDa, this converts to a protein cost of 3168kDa (3168000Da), or 28800 amino acids (assuming an average of 110 Da per amino acid).
$\eta_{T_{Mn,Unadjusted}}$	Manganese uptake protein complex size	$\frac{AminoAcids}{Molecule}$	372	There are four NRAMPs identified in the *F. cylindrus* genome (protein ID 137845, 173050, 172829, 197170) and subsequently checked with BLASTp and confirmed to be divalent metal tranporter 1 or an NRAMP. The average protein size is (422, 314, 398, 355 for the above protein IDs, respectively) is 372.

Table 4.1: Mathematical symbols, names, units, numerical values, and detailed descriptions for all the parameters in the proteomic allocation model described in the main text.

Parameter Symbol	Parameter Name	Units	Value	Description
$\eta_{T_{Fe,Unadjusted}}$	Iron uptake protein complex size	$\frac{AminoAcids}{Molecule}$	4030	GO:0006826 (iron ion transport) was used. Ferritin (protein ID 249610, 201 amino acids), iron permease FTR1 family (PF03239; protein ID 243554, 371 amino acids), ISIP1 (protein ID 241515, 594 amino acids) (ISIPs included because of putative role in iron transport and ubiquitous role in iron response), protein ID 236396 (102 amino acids), periplasmic ABC transporter (GO:0005381 iron ion transmembrane transporter activity, GO:0006827 high-affinity iron ion transmembrane transport; protein ID 193804, 301 amino acids, protein ID 195906, 206 amino acids, protein ID 197108 301 amino acids, protein ID 202747 207 amino acids, protein ID 172711 1733 amino acids, not included because sequence gaps impacts gene model), iron transporter Ferroportin1 (PF06963) protein ID 147624 493 amino acids, ABC transporter periplasmic Fe3+ hydroxamate transport system (protein ID 209416, 348 amino acids), ZIP Zn/Fe transporter (KOG1558, protein ID 184218 296 amino acids, protein ID 268016 179 amino acids), Ferric reductase / NADH/NADPH oxidase and related proteins (Protein ID 232972 824 amino acids, Protein ID 246292 917 amino acids, Protein ID 259423 1321 amino acids, Protein ID 227601 761 amino acids, Protein ID 235878 818 amino acids, Protein ID 252645 599 amino acids, Protein ID 184052 728 amino acids), sideroflexin (protein ID 140510 367 amino acids, Protein ID 216838 290 amino acids, Protein ID 226433 347 amino acids). The sum of all these proteins is 201 + 371 + 594 + 102 + mean(c(301, 206, 301, 207)) + 493 + 348 + mean(c(296, 179)) + mean(c(824, 917, 1321,761, 818, 599, 728)) + mean(c(367, 290, 347)) = 3787. We exclude Ferritin (201 amino acids), as we explicitly include this protein elsewhere. We also include the four copies of *FBP1* identified by Coale et al. (2019), in *F. cylindrus*, which have the average length of 442 (556, 235, 531, 449) thus the final cost is 4029 (rounded to 4030 in the model).
η_P	Photosystem unit size	$\frac{AminoAcids}{Molecule}$	12177	Data from Wollman, Minai and Nechushtai (1999), and we assume a 1:1 ratio of PSII:PSI.
η_A	Manganese superoxide dismutase size	$\frac{AminoAcids}{Molecule}$	227	Protein ID 185706 (232 amino acids), 239458 (223 amino acids). GO:0004784, GO:0006801.
θ	Dynamic Fe uptake cost coefficient	$\frac{AminoAcids}{Fe}$	1.469	Determined using Approximate Bayesian Computation.
θ	Dynamic Mn uptake cost coefficient	$\frac{AminoAcids}{Mn}$	1.469	Determined using Approximate Bayesian Computation.
r	Radius	Metres	3.952 x 10^{-6}	Inferred from Figure 1a in Mock et al. (2017).
s	Transporter complex radius	Metres	1.00E-09	From Berg and Purcell (1977).
M_{Cell}	Amino acids per cell	$\frac{AminoAcids}{Cell}$	1.4E+11	From Finkel, Follows and Irwin (2016), data from Supplementary File S1. The median picograms of protein per cell from *Pseudo-nitzschia* (*Fragilariopsis cylindrus* not in dataset) was 15.53692pg. Converted to grams per cell (1/1e12), then to amino acids per cell.
Λ	Proportion of the proteome that is independent of growth rate	Dimensionless	0.2	Estimated using calculations from Metzl-Raz et al. (2017).

Table 4.1: Mathematical symbols, names, units, numerical values, and detailed descriptions for all the parameters in the proteomic allocation model described in the main text.

Parameter Symbol	Parameter Name	Units	Value	Description
κ	Available space on the membrane for Mn and Fe transporters	Dimensionless	0.1484	Determined using Approximate Bayesian Computation.
m_γ	Energy use per amino acid elongation	$\frac{e}{AminoAcid}$	3	From Faizi et al. (2018).
φ_N	Energy use per NO_3 conversion into amino acids	$\frac{e}{NO_3}$	9.8	1 ATP for import + 1 NADPH for nitrate reduction + 1 NADPH for nitrite reduction + 1 ATP for glutamine synthetase + 1 NADPH for glutamate synthase. Assuming an interconversion ratio of 2.6 ATP to 1 NADPH, the total cost is: 1ATP + 2.6ATP + 2.6ATP + 1ATP + 2.6ATP = 9.8e.
$\varphi_{Fe,P}$	Iron per PSU	$\frac{Fe}{PhotosystemUnit}$	20	From ref. (Raven, 1990).
$\varphi_{Fe,N}$	Iron per N uptake and synthesis	$\frac{Fe}{NitrogenPathway}$	10	2 Fe in nitrate reductase (one per subunit, but it exists as a homodimer), 5 Fe in nitrite reductase in total (1 siroheme cofactor and 4 in 4Fe-4S cluster), 3Fe in glutamate synthase. Total is 10 Fe.
$\varphi_{Mn,A}$	Manganese per Mn-SOD	$\frac{Mn}{MnSOD}$	1	From ref. (Sheng et al., 2014).
$\varphi_{Mn,P}$	Manganese per PSU	$\frac{Mn}{PhotosystemUnit}$	4	From ref. (Raven, 1990).
φ_e	Energy produced per PSU activation	$\frac{e}{PSUActivation}$	8	From Faizi et al. (2018).
$\varphi_{T_{Mn}}$	Energy used for Mn uptake	$\frac{e}{Mn}$	2	Set to an arbitrarily low value to ensure Mn uptake is not possible with zero energy.
$\varphi_{T_{Fe}}$	Energy used for Fe uptake	$\frac{e}{Fe}$	2	Set to an arbitrarily low value to ensure Fe uptake is not possible with zero energy.
m_n	Inverse of nitrogen per amino acids	$\frac{AminoAcid}{Nitrogen}$	6.99E-01	Averaged across all amino acids.
k_{cat,T_N}	Maximum turnover rate of the rate-limiting enzyme in the nitrogen assimilation pathway	$\frac{1}{minute}$	178	At substrate saturating conditions, the enzyme with the lowest maximum turnover rate in a pathway determines the upper bound on flux through the pathway, assuming constant total amount of a enzyme (as it is represented in the model as an entire protein pool). Nitrate reductase is sometimes referred to as the rate limiting step in nitrogen assimilation (k_{cat} = 12 $\frac{1}{second}$ in spinach). Lambeck et al. (2010) estimated the nitrate reductase k_{cat} equal to 20 $\frac{1}{second}$ in *Arabidopsis thaliana*. We note that for glutamine synthetase, there was a lower k_{cat} for $NH_4{}^+$ at 2.96 $\frac{1}{second}$. We therefore used this as the rate limiting step in nitrogen assimilation to amino acids.
γ_{max}	Protein synthesis rate	$\frac{AminoAcids}{Minute \cdot Ribosome}$	1.14E+02	*Thalassiosira weissflogii* translation rate at 20 °C is 1.9 $\frac{AminoAcid}{Ribosome \cdot second}$. Assuming a temperature depence given by a factor of $Q_1 0$ = 2 (Toseland et al., 2013), protein synthesis rate is adjusted in the model. (Protein synthesis rate temperature adjusted = $114 \cdot 2^{(T/10-20/10)}$, where T is temperature and equal to -1 C).

Table 4.1: Mathematical symbols, names, units, numerical values, and detailed descriptions for all the parameters in the proteomic allocation model described in the main text.

Parameter Symbol	Parameter Name	Units	Value	Description
$k_{cat,Fe}$	Maximum turnover rate, Fe transporter	$\frac{1}{Minute}$	352	Discussed in text. Hudson and Morel (1990) values, using Fiksen, Follows and Aksnes (2013) calculation of handling time.
$k_{cat,Mn}$	Maximum turnover rate, Mn transporter	$\frac{1}{Minute}$	352	Kinetic constants for Fe uptake are assumed to be equivalent to Mn transporters, but further work is required for distinct parameterizations across micronutrients.
$k_{cat,ROS}$	Maximum turnover rate, MnSOD	$\frac{1}{Minute}$	10000	Taken from a range of BRENDA Database of kinetic constants for superoxide dismutases and from other publications (lower end was conservatively chosen consistent with argument from (Peers and Price, 2004), as a lesser efficiency would correspond with more demand, and a higher potential interaction between MnSOD production and oxidative stress). Note that this parameter is multiplied in the model by ϵ_a, making it have a more phenomenological interpretation.
$K_{Fe_i,P}$	Half saturation constant for Fe_i to Photosystems	$\frac{Molecules}{Cell}$	10000	Set to an arbitrarily low value.
$K_{Mn_i,P}$	Half saturation constant for Mn_i to Photosystems	$\frac{Molecules}{Cell}$	10000	Set to an arbitrarily low value.
$K_{Mn_i,A}$	Half saturation constant for Mn_i to Antioxidants	$\frac{Molecules}{Cell}$	10000	Set to an arbitrarily low value.
$K_{Fe_i,N}$	Half saturation constant for Fe_i to Tn (N uptake and biosynthesis)	$\frac{Molecules}{Cell}$	10000	Set to an arbitrarily low value.
K_e	Half saturation constant for energy	$\frac{Molecules}{Cell}$	10000	Set to an arbitrarily low value.
ϵ_a	Efficacy per MnSOD	Dimensionless	0.00001	Determined using Approximate Bayesian Computation.
σ	Absorption Cross Section	$\frac{m^2}{uE}$	0.01	From Strzepek et al. (2012).
τ	Activated photosystem turnover rate	$\frac{1}{Minute}$	6000	We obtained the values from the mean shown in Figure 2e for phytoplankton in Strzepek, Boyd and Sunda (2019), of 0.1 ($\frac{1}{ms}$). Converting to minutes, this results in an value of 6000 ($\frac{1}{Minute}$).

Table 4.1: Mathematical symbols, names, units, numerical values, and detailed descriptions for all the parameters in the proteomic allocation model described in the main text.

Parameter Symbol	Parameter Name	Units	Value	Description
R_0	Shape of the ROS mismatch penalty function	Dimensionless	10	Arbitrarily chosen.

4.7.2 Supplementary Methods

4.7.2.1 Optimization

Our model optimizes the set of parameters that describe the proportion of ribosomes translating for a given protein pool ($\vec{\beta}$), and we used the steady-state growth rate as the objective function (Faizi et al., 2018). We used several techniques to improve the accuracy and speed of optimization.

The computational problem can be broken down into two stages. The first stage is solving for the steady state growth rate, and the second stage is determining which set of parameters lead to the optimal growth rate. We chose to set up the first problem as a system of ordinary differential equations (ODEs), and numerically integrate these ODEs to a pseudo-steady state (Faizi et al., 2018). The rates within the ODEs are a function of the proportion of ribosomes translating each proteomic pool ($\vec{\beta}$). So, in order to determine the optimal $\vec{\beta}$ we need integrate the system of ODEs to a steady state to evaluate the growth rate (μ). To perform the numerical integration, we used the python module SciPy *odeint*, which accesses the LSODA algorithm in ODEPACK (Hindmarsh, 1983). The integration time step varied, depending on whether it was used for the ABC analysis or the parameter perturbation experiments. However, the maximum number of internally defined time steps (*mxstep* in *odeint*) was kept at 1E6.

In order to determine the optimal $\vec{\beta}$, we used Sequential Least Squares Quadratic Programming (SLSQP, SciPy), a method used for non-linear, constrained, and bounded optimization problems. However, we found that the minimization was highly start-point dependent due to the nature of the optimization problem. We developed a three-component optimization protocol to approach this problem. In the first component ('the drunkards walk'), we used SLSQP initialized with a random $\vec{\beta}$ with a high error tolerance, performed n times. For all model experiments (parameter perturbations), and for the baseline model, we set this value to n = 20. For running the ABC, we used n = 10, because it is slightly faster. During this stage, the integration time for determining the steady-state growth rate

was set to 5E5 (with a time step of 10). The 'high error' tolerance corresponded with a value of 1E-4 for the *ftol* parameter in SciPy minimize, which is the precision goal for the growth rate value in the stopping criterion. We also set the maximum iterations parameter (*maxiter*) to 200 for the minimize function in SciPy. For all other parameters we used default settings.

The second component is an empirical Bayesian optimization using Gaussian Process Regression, informed from the original set of n random guess. Conceptually, this approach is improving the sampling of parameter space because it is not dependent on randomly generated $\vec{\beta}$, but a guided search. After the first component, we use the SLSQP determined $\vec{\beta}$, paired with their steady state growth rates, to train a Gaussian Process Regression (GPR) model. We use this approach because evaluating our objective function is computationally expensive. The GPR model covariance function is an additive combination of the dot product and White kernel. Once we have trained the GPR model, we generate p random $\vec{\beta_P}$, evaluate all $\vec{\beta_P}$ using the trained GPR model, and then determine which $\vec{\beta_P}$ would have the highest growth rate. For all model experiments and the baseline model runs, we set p = 1000. For the ABC analysis, we set p = 300. We then use the top 20% p values (as ranked by their GPR-predicted μ) as initial starts for the SLSQP approach. After these $\vec{\beta}$ values are evaluated with the objective function, we re-train the GPR model with these additional observations k times. For all model experiments and baseline model runs, we set k = 10, for the ABC analysis we set k = 2.

The third component is a refined optimization with lower error tolerance ('the sober walk'). We begin by taking the top 10% of $\vec{\beta}$, ranked by their growth rate. From this sub-group, we use a k-means clustering of these parameter sets. We then take the j centroids from the k-means clustering, and use these as inputs for the 'sober walk'. This last component uses the k-means clustering centroids as start points for the SLSQP optimization, with a lower error tolerance compared to above. The 'low error' tolerance corresponded with a value of 1E-6 for the *ftol* parameter in SciPy minimize. The optimal $\vec{\beta}$ is the parameter set, from this component, which resulted in the highest steady-state growth rate.

We found that these steps above improved both the computational speed and accuracy of the optimization. Additionally, we used the square-root of the growth rate, which flattens the optimization surface and we found to improve the accuracy of our optimization.

We also used a shorter steady-state time during the 'drunkards walk', which is a good approximation of the steady state growth rate, but it is much more efficient because the time-length of integration is much smaller.

Algorithm 2: Optimization algorithm description. n is the number of initial random parameter guesses. ϵ_f is fixed proteomic fraction. SLSQP: Sequential Least Squares Quadratic Programming, a constrained optimization protocol.

Result: $\vec{\beta_{Opt}}$

for n **do**

  Generate random $\vec{\beta}$ such that $\sum_j \beta_j = 1$;

  High error tolerance SLQSP using $\vec{\beta}$ as initial value and $\sqrt{\mu}$ as objective function;

  Generate initial optimized $\vec{\beta_i}$ and corresponding μ_i;

end

Train a Gaussian Process Regression (GPR) model using all $\vec{\beta_i}$ and corresponding μ_i;

for k **do**

  Generate p, $\vec{\beta_p}$ such that $\sum_j \beta_j = 1$;

  for p **do**

    Predict μ using trained GPR;

  end

  Subset which $\vec{\beta_p}$ had the highest GPR-predicted μ;

  Evaluate these subsetted $\vec{\beta_p}$ using a high error tolerance SLQSP, determine corresponding μ_p;

  Re-train GPR model with new, appended set of $\vec{\beta_p}$ and corresponding μ_p;

end

K-means clustering of top 10%, ranked by μ.

for i **do**

  Low error tolerance SLQSP using K-means cluster centroids as initial values and $\sqrt{\mu}$ as objective function;

end

4.7.3 Supplementary Discussion

4.7.3.1 Model Parameter Posteriors Interpretation

We estimated three key, unconstrained model parameters, each corresponding to different cellular processes: (1) internal Fe protein cost, (2) available membrane space for transporters, and (3) the catalytic efficiency of MnSOD. Below we provide a detailed discussion of their interpretation. The first parameter represents the strength of this internal cost in our model by multiplying the cellular Fe quota by a coefficient, which is interpretable as the amino acids required for managing each Fe atom. If all Fe was bound to ferritin, for example, this coefficient would be the total amino acids per ferritin protein complex divided by the total amount of Fe per ferritin protein complex. This parameter significantly increases the protein cost per transporter.

The second key parameter was the available membrane space for Mn and Fe transporters. Note that our estimate of available membrane space corresponded with the upper-bound of our prior distribution for this parameter – a wider prior distribution may have resulted in a higher value. We chose this upper bound (15% of membrane surface area) because approximately 50% of the membrane can be allocated to proteins to maintain lipid bilayer integrity (Kadner, 1996). There is also a membrane requirement for macronutrient transporters (e.g. phosphate, nitrate, or silicate). We therefore reasoned that Fe and Mn transporters took up a maximum of 15% of the membrane space. Yet, targeted work on membrane protein dynamics, particularly in eukaryotic phytoplankton, is clearly required to obtain a more accurate upper bound for available area for membrane transporters.

The last unconstrained parameter was the catalytic efficiency of MnSOD. Superoxide dismutases are incredibly efficient enzymes (Sheng et al., 2014). If we constructed this model to simply minimize steady-state concentrations of superoxide (assuming a well mixed compartment), only a few copies of MnSOD would be required. Yet, that is not observed in field-based proteomes (our data presented here) or in cultured diatoms (e.g. Nunn et al., 2013). There is also evidence that MnSOD is associated with the chloroplastic membrane directly suggesting some degree of producing this critical protein at levels higher than would be suggested from kinetic-based reasoning (Pilon, Ravet and Tapken, 2011; Ogawa et al., 1995; Regelsberger et al., 2002). In other words, an overproduction of MnSOD would prevent free superoxide from diffusing and interacting with biomolecules. Yet, the *degree* of overproduction is uncertain – this is one way this parameter may be interpreted.

4.7.3.2 Iron-Light Interactions in the Southern Ocean

Recently, some evidence suggesting that Southern Ocean phytoplankton have a unique relationship between Fe and light levels has emerged, which is mediated by low temperatures (Strzepek, Boyd and Sunda, 2019). We predicted an inverse relationship between light levels and cellular Fe quotas, consistent with previous work (Sunda and Huntsman, 1995). It is unsurprising that we do not predict the relationship observed in Strzepek, Boyd and Sunda (2019), as it was not included within our photosynthetic model. Future work is required to address how various temperature-dependent mechanisms (e.g. photosynthetic processes, translation rate, Fe uptake kinetics, membrane saturation, etc.) integrate to influence the complex relationship between Fe, Mn, temperature, and light.

4.7.3.3 Iron and Manganese Interactions

A key mechanism of interdependence between Mn and Fe was included in our model (Peers and Price, 2004). Briefly, under low Fe, electrons leak more from electron transport, increasing the requirement for MnSOD. Therefore, under low Fe, Mn should have a relatively larger impact on growth rate than at high Fe. Despite explicitly including an interaction between Mn and Fe in our model, we found that these two micronutrients influenced growth rate mostly independent of each other.

Antioxidants are produced to counteract ROS production via leaked electron flux. There are two controls on total leaked electron flux: 1) the proportion of electrons leaked and 2) the total electron flux. Low Fe increases the proportion of electrons leaked, but it also decreases the total electron flux. Therefore, while the requirement of MnSOD *per PSU* increases under low Fe, the total requirement for MnSOD decreases. These observations challenge the result that Fe and Mn interact under low Fe – seemingly inconsistent with the observed increase in Mn quota under low Fe (Peers and Price, 2004). However, several lines of evidence from Peers and Price (2004) are actually consistent with our predictions and newer observations. Our model also predicts an increase in Mn quota under low Fe, yet the source of this increased quota is an increased internal free Mn pool, not MnSOD. Increased reactive oxygen species observed under low Fe (Peers and Price, 2004) is consistent with superoxide secreted to increase bioavailability of ligand-bound Fe (Rose, 2012), or with a recently discovered mechanism relating superoxide production with photosynthetic health (Diaz et al., 2014). Our model results suggest that Mn actually has a larger role in influencing growth rate under high Fe, rather than low Fe, and a reframing

of previous growth rate data provides some support for this conclusion (Supplementary Fig. 4.15). Yet, these geochemical conditions are infrequently encountered in the Southern Ocean (Supplementary Fig. 4.8), so it is unclear how much Mn and Fe interact to control phytoplankton growth in the Southern Ocean. Furthermore, it is unclear how to reconcile observations suggesting Mn limits primary productivity (Buma et al., 1991; Browning et al., 2014; Wu et al., 2019; Middag et al., 2011; Sherrell et al., 2015; Browning et al., 2021). Perhaps a more complex, community-interaction is at play, or other organisms (e.g. haptophytes) are contributing to this phenomenon.

4.7.3.4 Proteomic Allocation Model Predictions compared with Field-Based Diatom Proteomes

Our model captures trends in fold change for most protein pools from the two weeks with both proteomic observations and geochemical data. It did not capture the trends in expression of nitrogen metabolism and uptake, and antioxidants. We observed a decrease in photosystem units (PSUs) and ribosomes from Week 1 to Week 3 in both the experimental metaproteomic time series (dotted vertical lines, Supplementary Fig. 4.20) and the model predictions. However, the model did not capture the trend in abundance change for the nitrogen metabolism protein pool. We hypothesize that the lack of correspondence is because our model only considers nitrogen uptake from nitrate, whereas concurrent experiments suggested diatoms were using ammonium (Jabre et al., 2021) which is more Fe efficient (Raven, 1988). The posterior distribution for antioxidant expression did not clearly indicate one direction of expression (i.e. increase or decrease). Thus, more targeted work examining the expression of this key protein pool is required.

4.7.4 Supplementary Figures

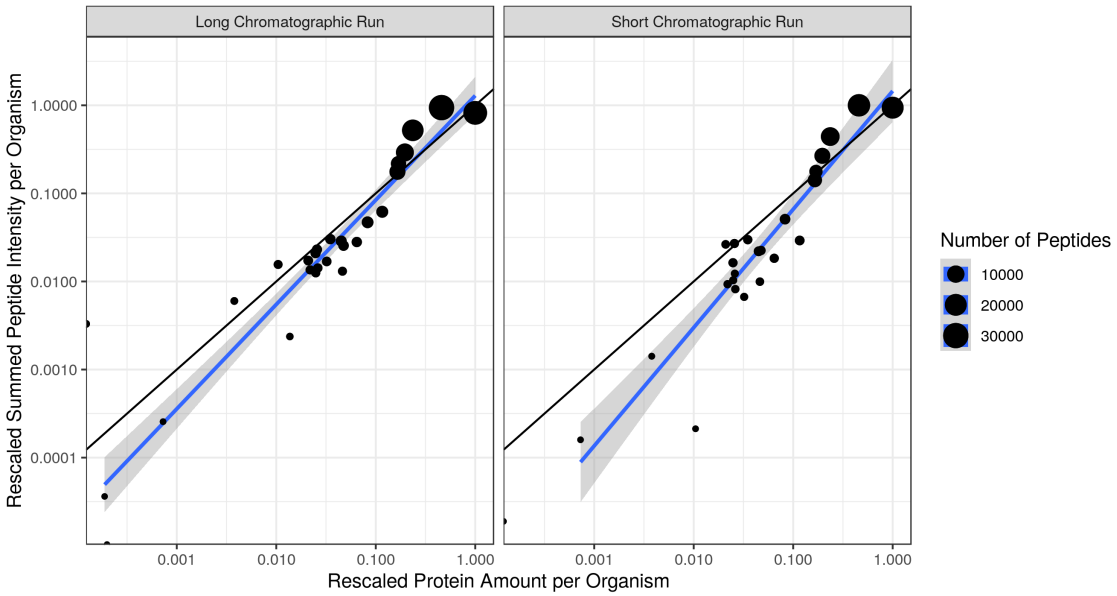

Figure 4.6: Correlation between sum of taxon-specific peptide intensities per taxonomic group and total protein used for each taxa using the Kleiner et al. (2017) artificial metaproteome. The sum of peptide intensities per group is a valid proxy of biomass across both the 'Long' and 'Short' chromatographic runs (right and left panels). Values from both axes are rescaled such that they vary from 0–1. The 1:1 line of y = x is plotted in black, and the blue line represents a linear model. Point size reflects the number of peptides observed for a given taxa.

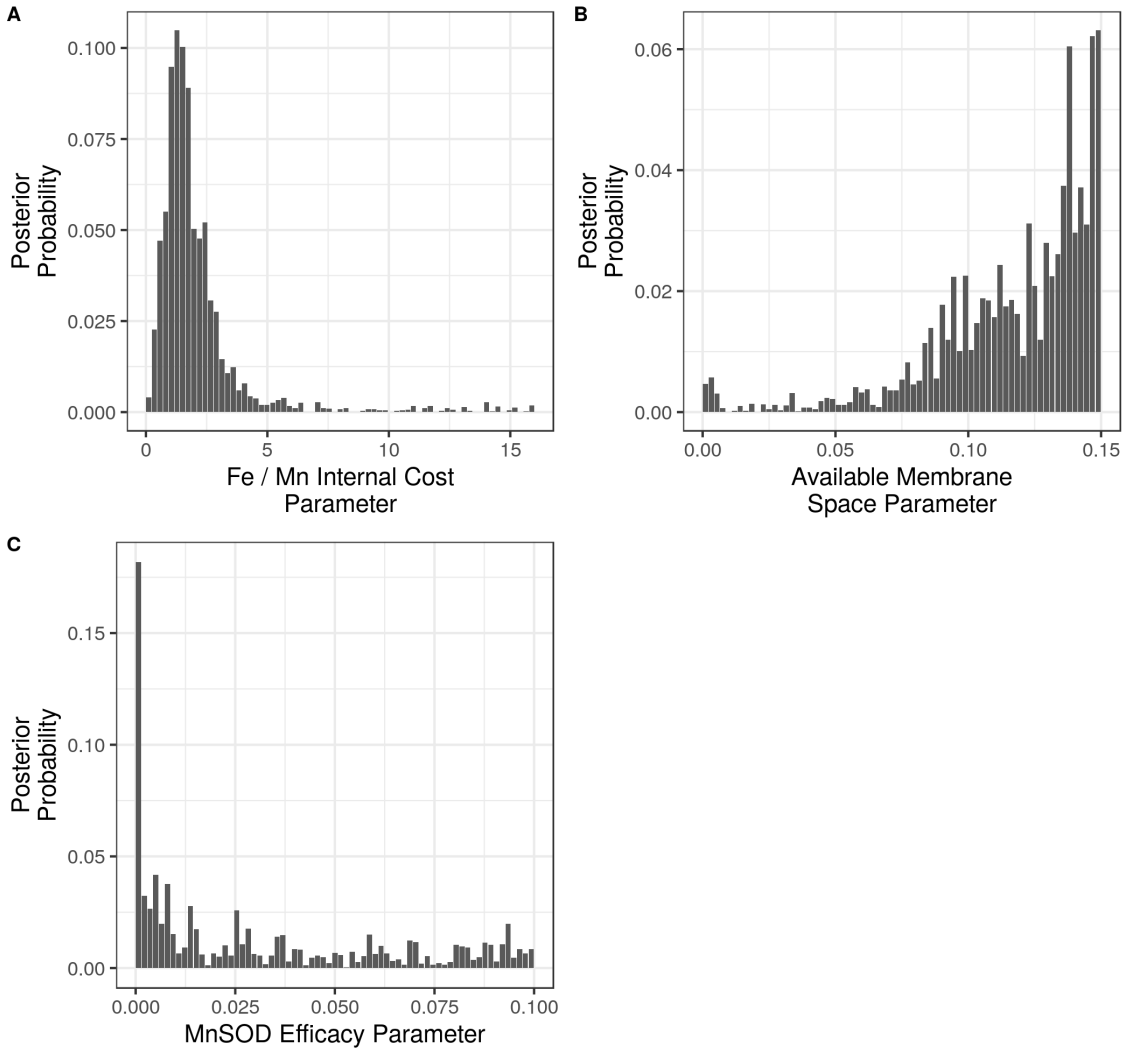

Figure 4.7: Posterior probability distributions for each of the three unconstrained, estimated parameters. Parameters were estimated using Approximate Bayesian Computation (see Methods). The modes of each distribution were used for the inferred parameter value. Each bar represents the posterior probability for a given cellular parameter within that interval (Methods).

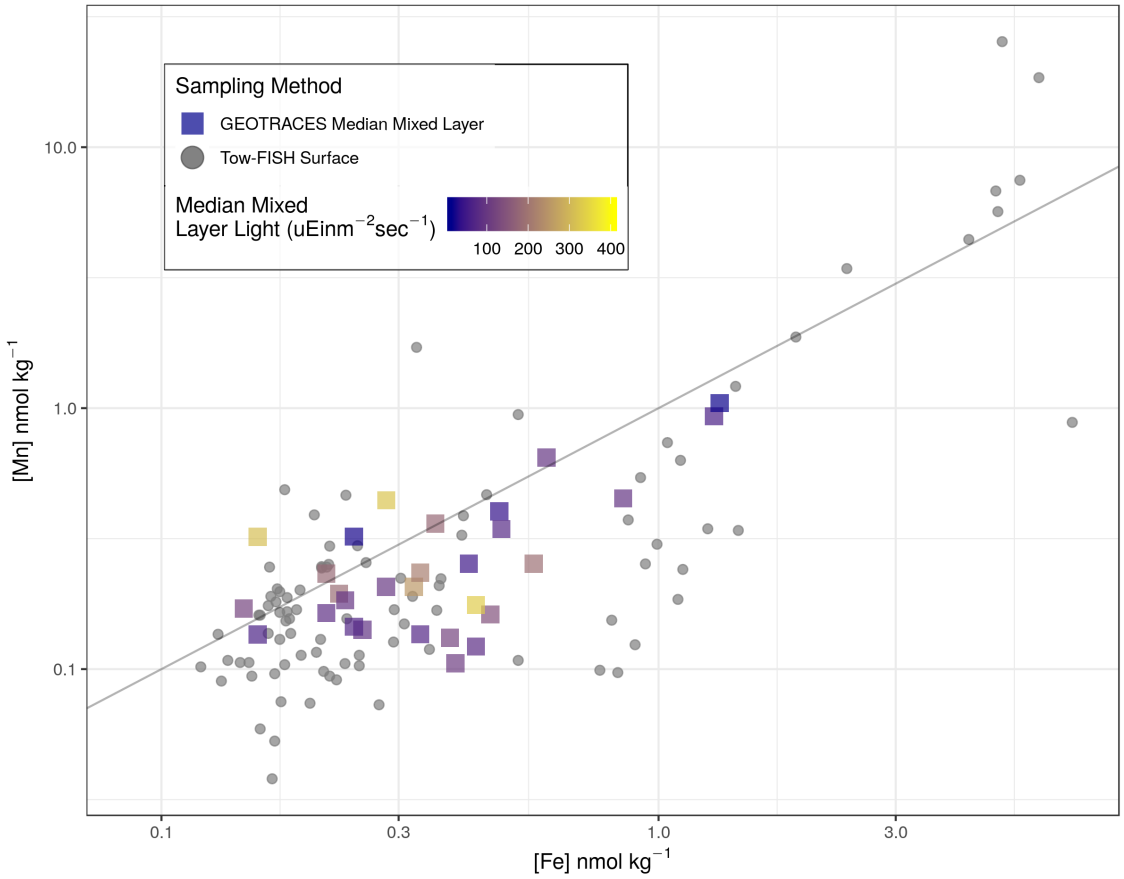

Figure 4.8: Southern Ocean concentrations of dFe and dMn from two data sources (Methods). Circles represent concentrations derived from a surface Tow-FISH on GEOTRACES cruise JR274. Squares are median mixed layer concentrations of dMn and dFe from GEOTRACES cruises in the Southern Ocean, GEOTRACES Intermediate Data Product (Schlitzer et al., 2018). Corresponding light levels were calculated using the Ocean Color database (NASA, 2014).

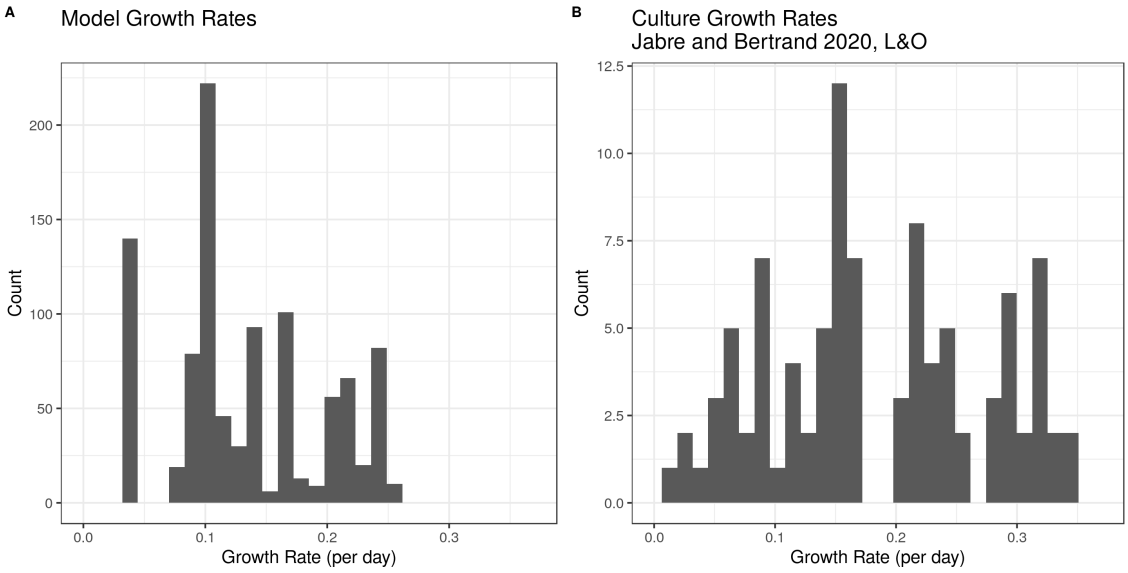

Figure 4.9: Histogram of growth rates from the cellular model (left, 1–3000 pM dFe, 1–3000 pM Mn) are within the same range of observed growth rates of *Fragilariopsis cylindrus* across a range of iron and temperatures (Jabre and Bertrand, 2020).

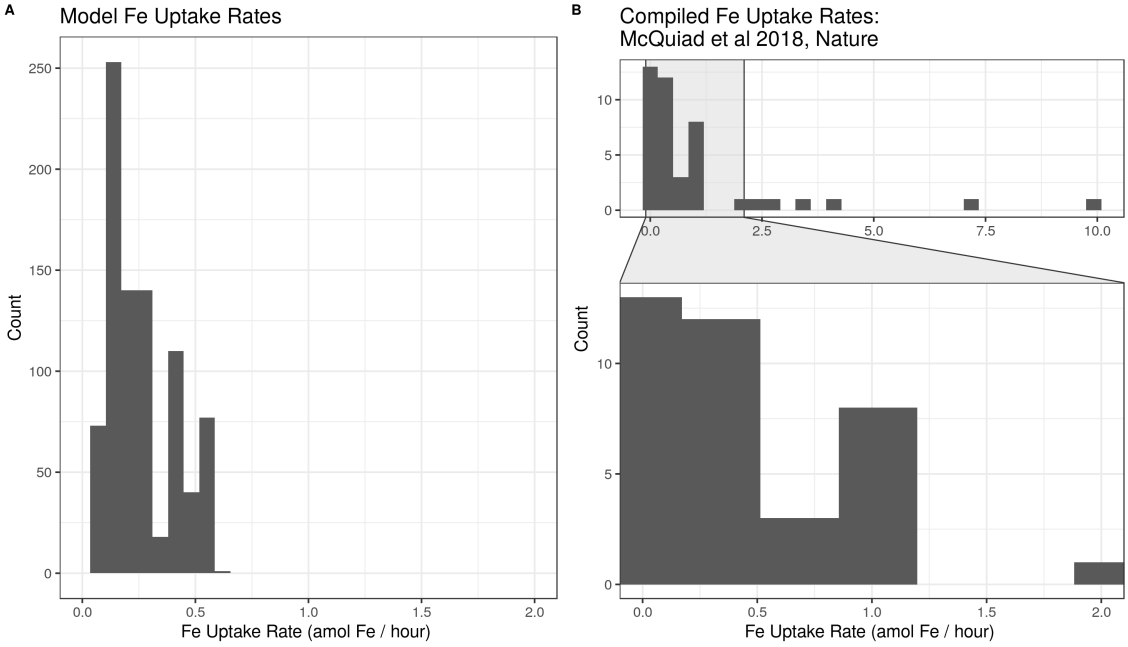

Figure 4.10: Histogram of Fe uptake rates from the cellular model (left, 1–3000 pM dFe, 1–3000 pM Mn) are within the observed range of Fe uptake rates of *Phaeodactylum tricornutum* (McQuaid et al., 2018).

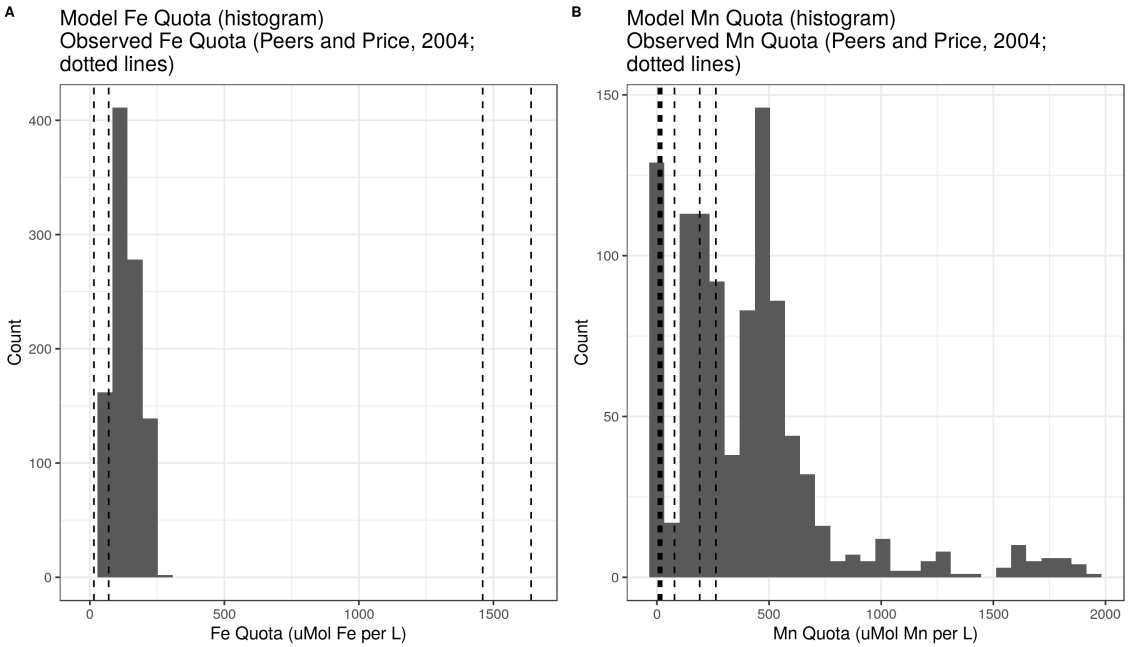

Figure 4.11: Model Fe and Mn quotas (histograms, 1–3000 pM dFe, 1–3000 pM dMn) overlap with the observed range of cellular quotas from *Thalassiosira pseudonana* (dotted lines, Peers and Price, 2004).

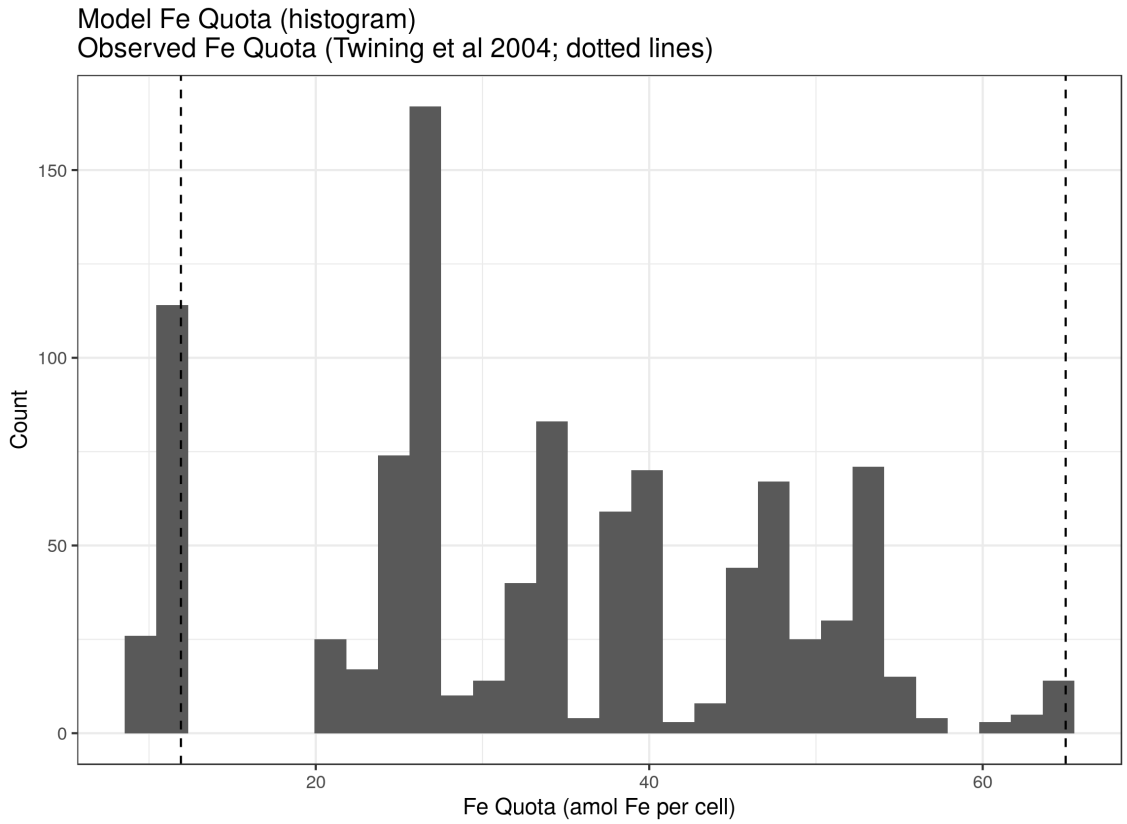

Figure 4.12: Model Fe and Mn quotas (histograms, 1–3000 pM dFe, 1–3000 pM dMn) overlap with the observed range of cellular quotas from diatoms collected on the SOFeX expedition to the Southern Ocean (dotted lines, Twining, Baines and Fisher, 2004).

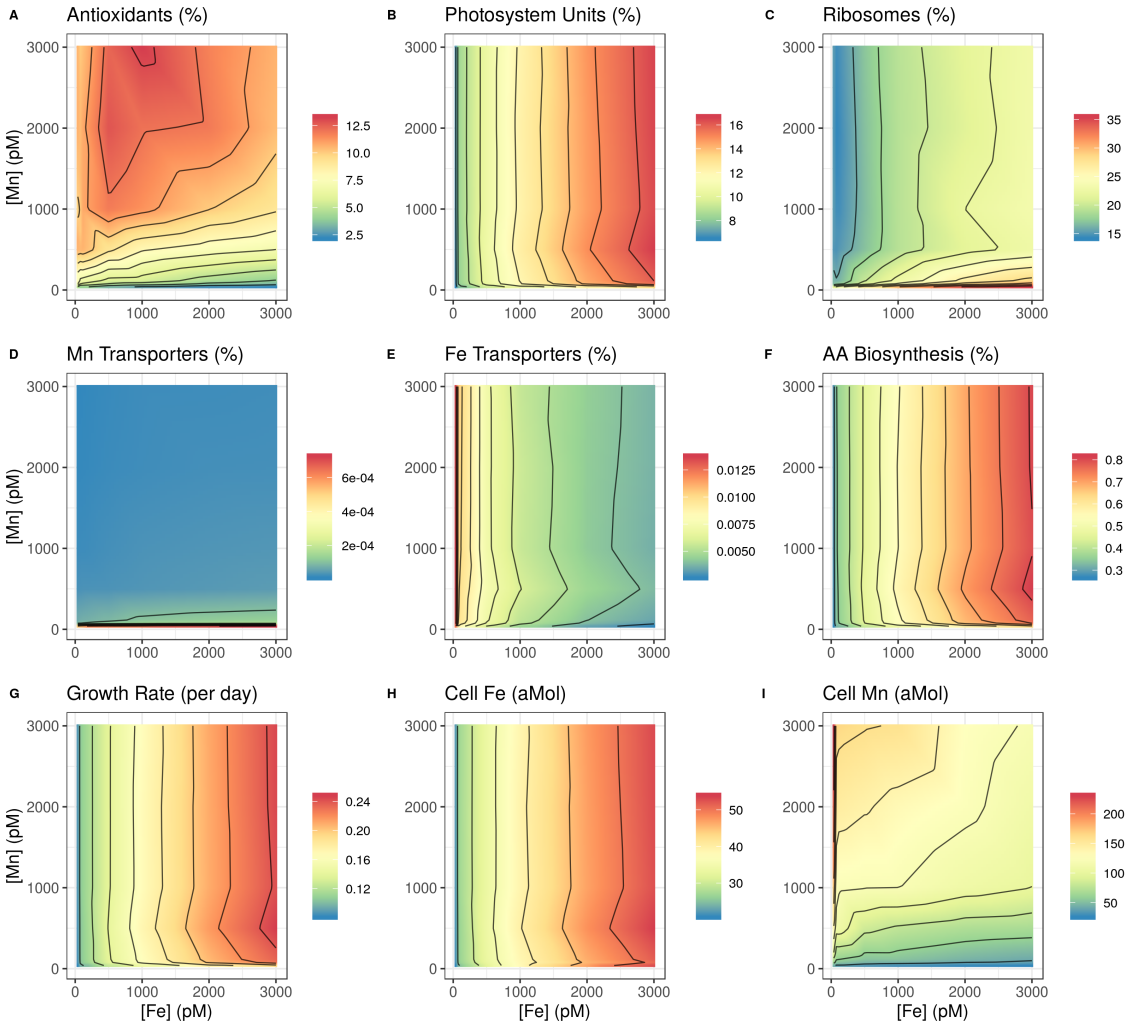

Figure 4.13: Model runs across a wide range of iron and manganese concentrations typically observed in the Southern Ocean. Light levels were 50 $\mu Einm^{-2}s^{-1}$. (**A-F**), proteomic mass fractions for each proteomic pool from the cellular model. (**G**), growth rate across a wide range of Fe and Mn concentrations. (**H-I**), total cellular quotas of Fe and Mn, including the free Mn and Fe pools.

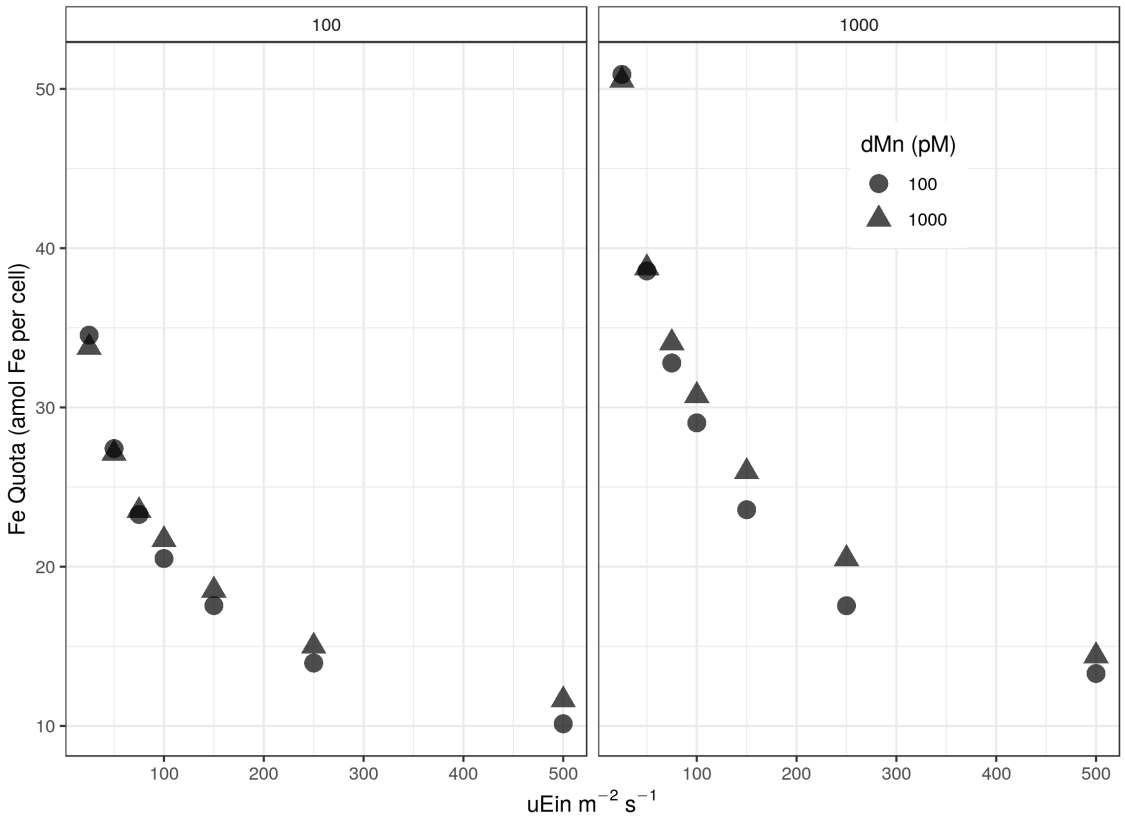

Figure 4.14: Light and cellular Fe quota from the cellular model have an inverse relationship. Two dFe concentrations (100 and 1000 pM) are shown (left and right panels), with two concentrations of dMn (100 and 1000 pM), across a range of light levels.

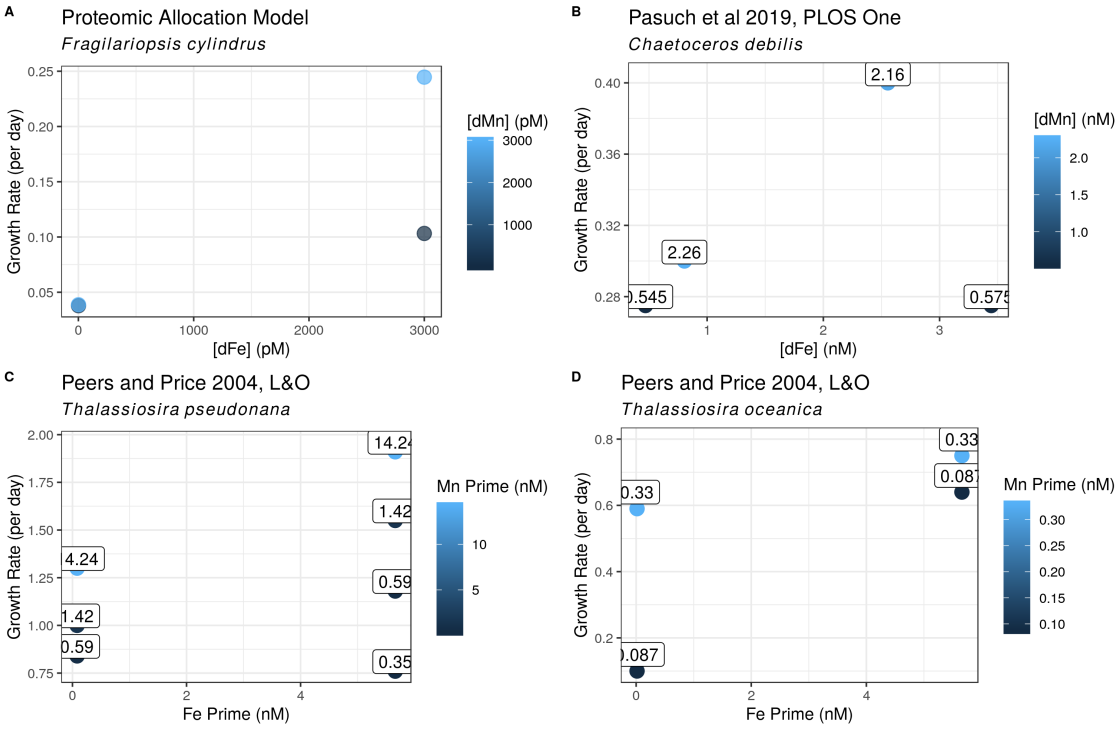

Figure 4.15: Model (**A**) and culture (**B-D**) growth rates under low and high Fe and low and high Mn. Manganese concentrations are given as the colour of the points (and the labels on the points). The proteomic allocation model and two datasets support the conclusion that Mn has a bigger impact on growth rate under high Fe than under low Fe (Peers and Price, 2004; Pausch, Bischof and Trimborn, 2019).

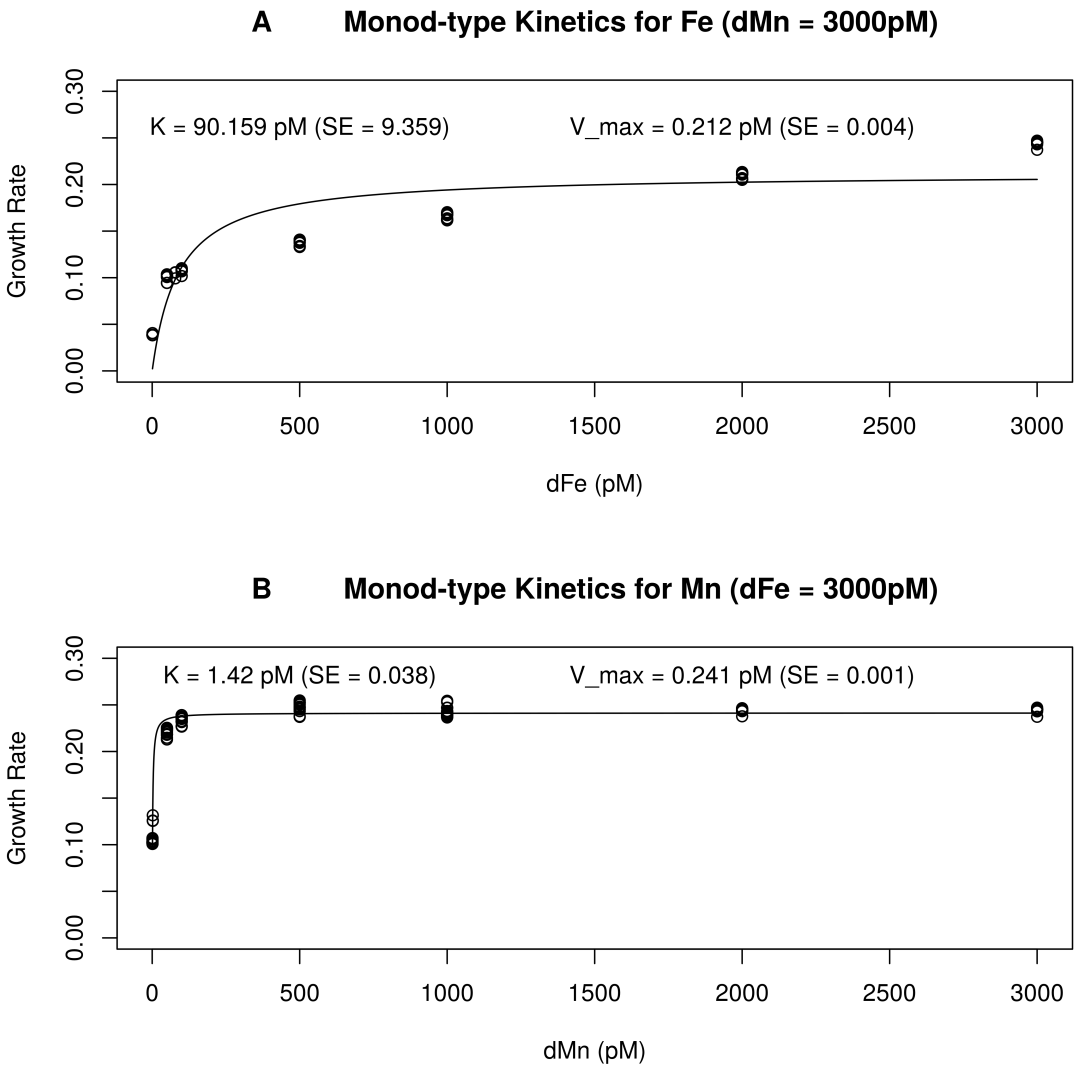

Figure 4.16: Monod growth kinetics for dFe and dMn. Inset parameter values were found using nonlinear-least squares. For both Fe and Mn (top and bottom, respectively), saturating concentrations of the non-varying nutrient were used (saturating concentrations = 3000pM). Monod-type functions fit dMn well, but dFe poorly, but this approach was used to simply choose concentrations for testing parameter changes.

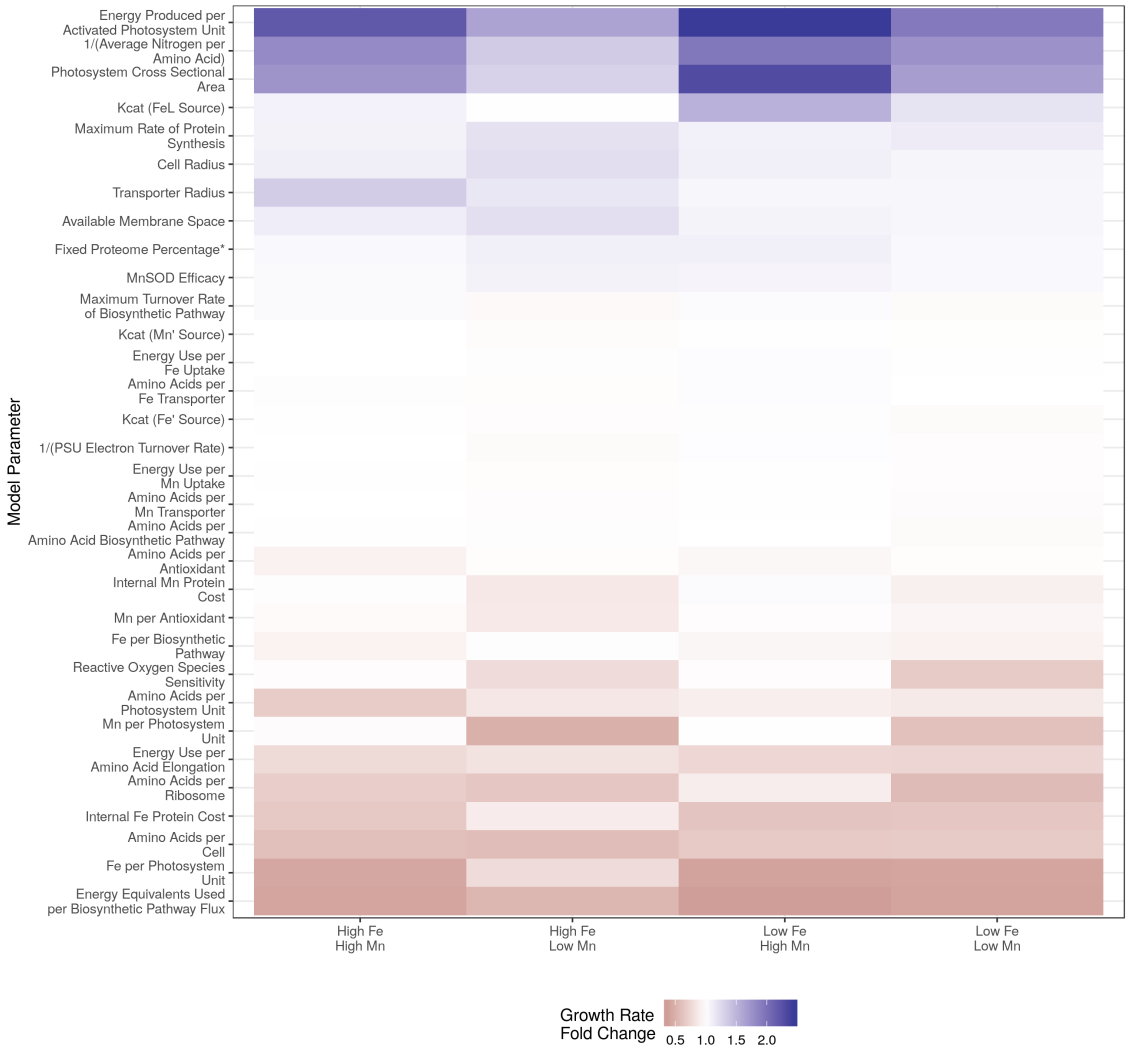

Figure 4.17: Change in growth rate under four different concentrations of dMn and dFe. Concentrations were chosen using Monod-functions, with the 'Low' value as the half saturation constant, and the 'High' value as an arbitrarily high, saturating, concentration (3000 pM). Parameter values were multiplied by a factor of five, and the resulting growth rate after three replicate model runs was then divided by the base model (no parameters altered). *Note that the 'Fixed Proteome Percentage' parameter was divided by five, not multiplied, because the base value is 20%.

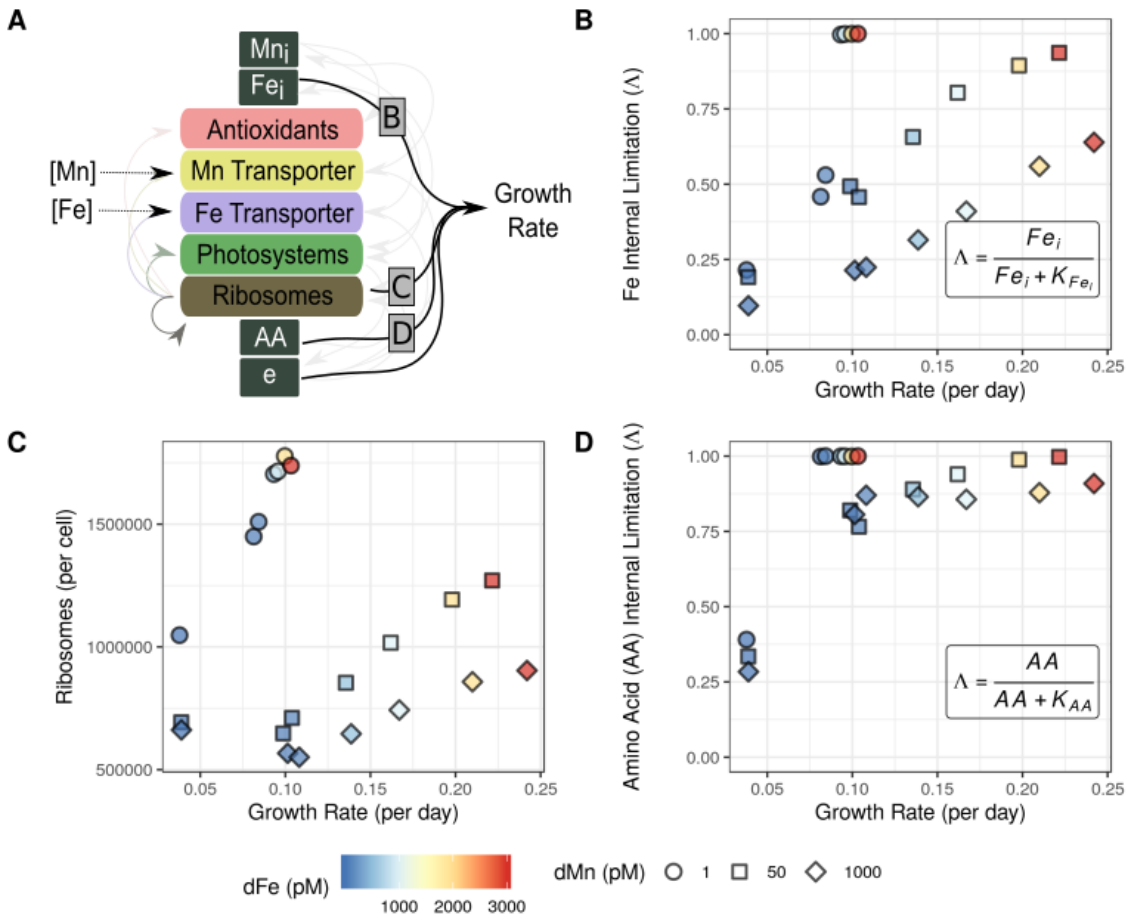

Figure 4.18: Extended version of additional internal cellular processes that are the proximate causes of growth rate. (**A**), the internal, modelled processes directly inhibiting growth vary across iron and manganese concentrations. (**B**), internal limitation proxy of iron varies across the growth rate and external concentration of manganese (dFe ranges from 1 pM to 3 nM, three dissolved Mn levels are shown: 1, 50, and 1000 pM). Internal Fe status (inset equation) influences the synthesis of proteins in the nitrogen metabolism pathway. (**C**), variation in ribosomes per cell with growth rate, (**D**), the total amount of available amino acids (inset equation) impacts the growth rate.

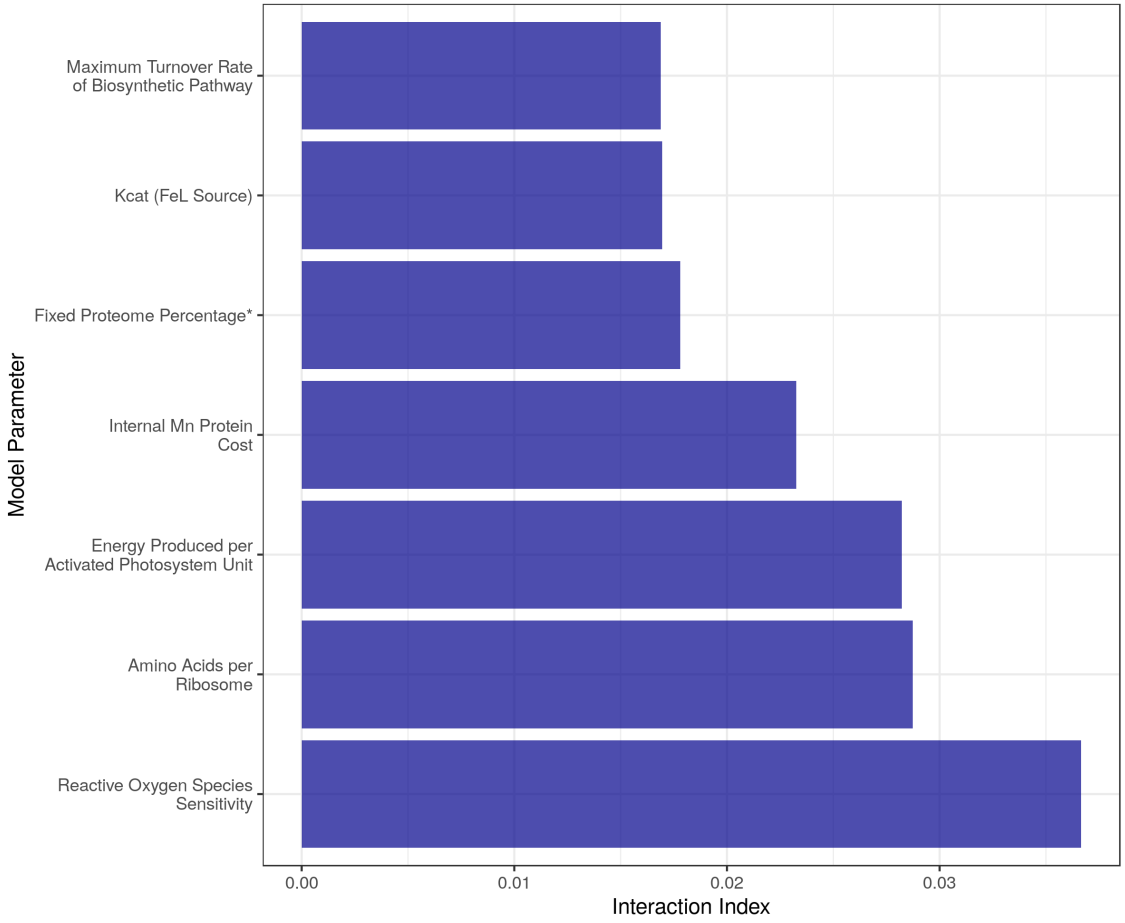

Figure 4.19: A parameter-specific interaction index based on the influence a parameter perturbation had on growth rate under high and low dMn and dFe concentrations (the same conditions in Fig. 4.4a; description of interaction index equation in the Methods). Shown here are the parameters from the proteomic allocation model with the highest interaction index (top 20%).

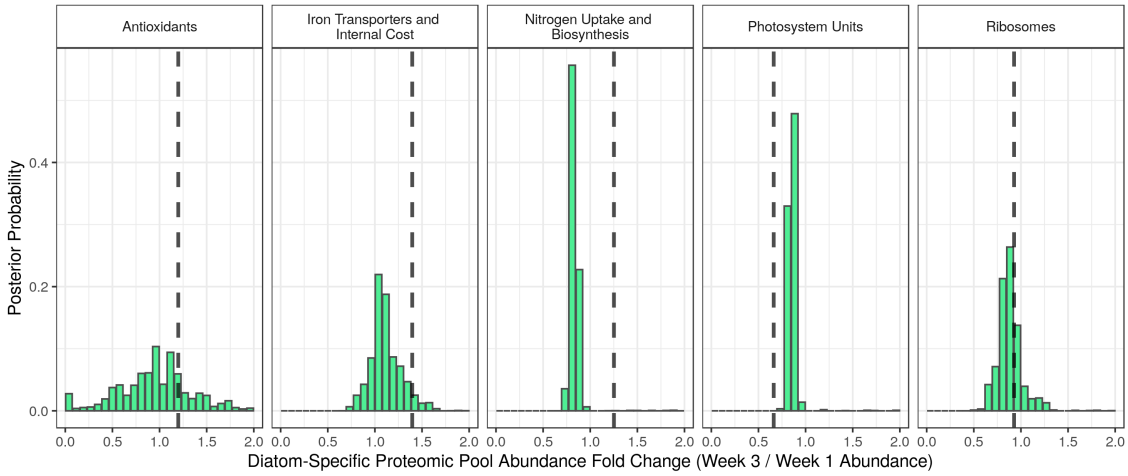

Figure 4.20: Posterior probability distributions of fold change from two weeks of diatom protein expression, inferred from a metaproteome. Week 1 corresponds with higher Fe and Mn, and Week 3 corresponds with lower Fe and Mn. Five of six protein pools are shown, with model posterior probability distributions given as green histograms (Mn transporters were not observed from the metaproteome). The empirical observations are shown as grey, vertical dashed lines.

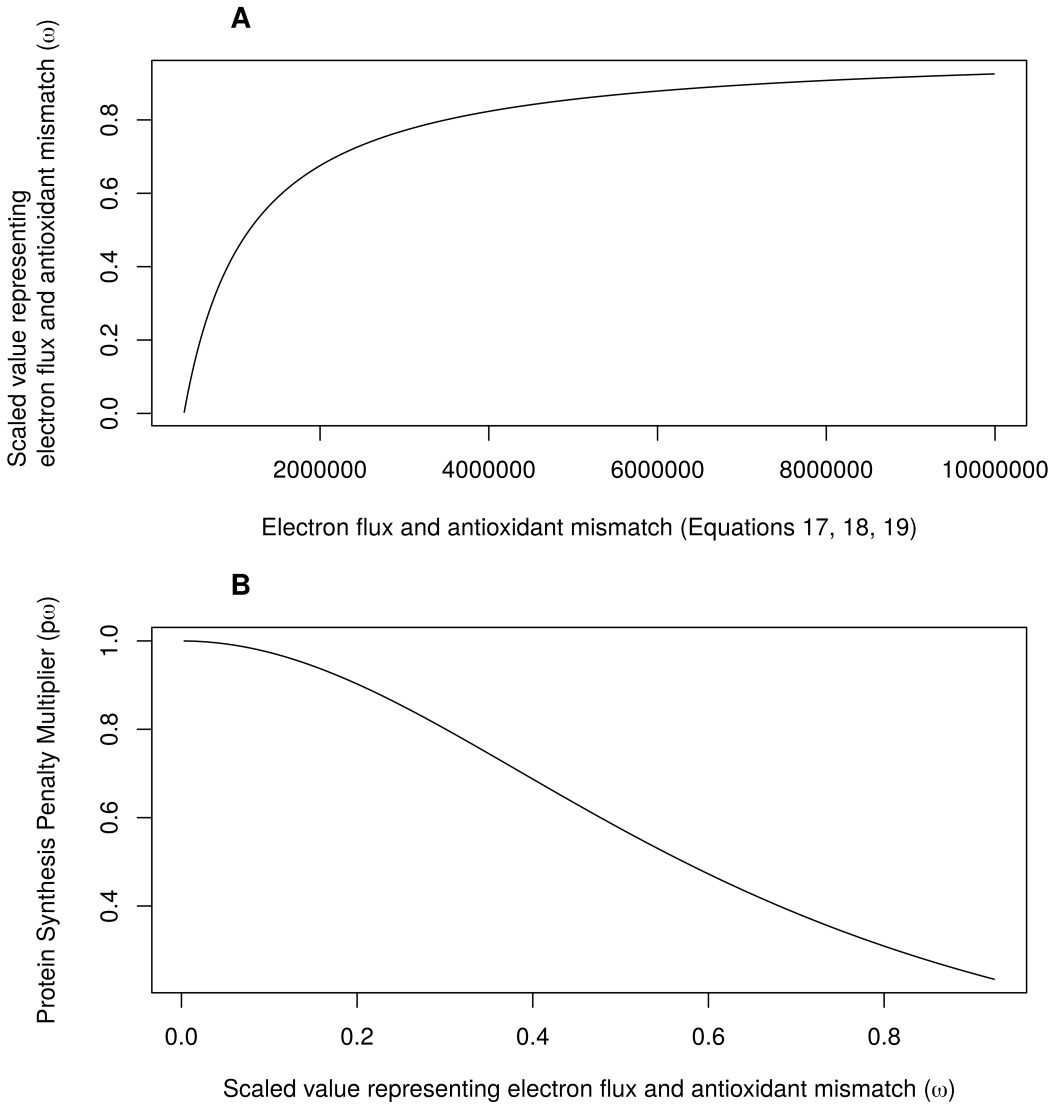

Figure 4.21: Graphical description of transformations for calculating the protein synthesis penalty, specifically described in equations 4.17, 4.18, and 4.19. Panel **A** shows equation 4.17 and 4.18, while panel **B** shows equation 4.19 graphically.

CHAPTER 5

PHYTOPLANKTON ANTIOXIDANT SYSTEMS AND THEIR CONTRIBUTIONS TO CELLULAR ELEMENTAL STOICHIOMETRY

This work was accepted for publication in December, 2021 at *Limnology and Oceanography Letters*.

5.1 Abstract

Oxidative stress plays a role in many aspects of cellular metabolism, and as a result, antioxidants have the potential to impact cellular stoichiometry and biogeochemical cycles. We reviewed how antioxidant systems influence macro- and micronutrient stoichiometry in marine phytoplankton, and identified that antioxidant systems have important implications for micronutrient stoichiometry. By leveraging diatom proteomic data, we empirically estimated the level of micronutrient quota variation that can be attributed to antioxidant systems. Fe-containing antioxidant expression may contribute to 3.3–10 μmol:mol variation in Fe:C, and superoxide dismutases appear to be important contributors to variation in Mn, Ni, Zn, and Cu quotas in phytoplankton. Critical next steps for the study of phytoplankton antioxidant systems are to 1) distinguish between oxidative stress and redox-based gene regulation, and 2) determine how antioxidants influence variation or consistency in micronutrient quotas under various environmental conditions.

5.2 Introduction

Elemental stoichiometry connects individual organisms to earth-scale processes, and underpins the connections between all biogeochemical cycles. Redfield (1958) famously connected ratios of nitrogen and phosphorus in surface phytoplankton to dissolved concentrations in the deep ocean, describing a homeostatic system. Comparatively less focus has been on exactly why nutrients have specific stoichiometric ratios. Early work by Elser et al. (1996) connected biochemical composition to life history, providing a causal link of between life history and stoichiometry (the biochemical foundations of stoichiometry have been further characterized since then, for example in Geider and LaRoche, 2002; Elser et al., 2000; Sterner and Elser, 2002). Expanding on this body of work, Loladze and Elser (2011), suggest that the ratio of nitrogen to phosphorus is $\sim$16 because of fundamental constraints on protein synthesis by phosphorus-rich ribosomes. Which other cellular processes impact deviations from, and consistency with, the Redfield ratio? In addition to macronutrients, micronutrients like iron (Fe) can play a large role in influencing primary productivity and biogeochemical cycling (Martin and Fitzwater, 1987; Tagliabue et al., 2017). Yet, attempts to extend the Redfield ratio to micronutrients has uncovered enormous variability. For example, Sunda and Huntsman (1995) found that Fe:C varies by as much as two orders of magnitude, and much of this variability is likely due to high Fe uptake (often called luxury uptake), rather than biochemical responses (e.g. protein production). Which other cellular mechanisms lead to variability in micronutrient stoichiometry?

Oxidative stress influences many aspects of cellular function and metabolism, and thus has the potential to influence cellular stoichiometry. For example, cells invest significant resources in protecting against oxidative stress, shown by the proportion of proteomes invested in antioxidant enzymes (Müller et al., 2020). These resources also include protein chaperone networks (Santra, Dill and Graff, 2018), protective biomolecules (e.g. glutathione, polyphosphate), and protein synthesis (Nishiyama, Allakhverdiev and Murata, 2011). Numerous components of metabolism are influenced by oxidative stress. For example, glycolysis is controlled by oxidative stress due to peroxide-induced inactivation of the key protein glyceraldehyde-3-phosphate dehydrogenase (Shenton and Grant, 2003). In addition to its influences on cellular stoichiometry, oxidative stress also has large consequences for eco-evolutionary dynamics (Laman Trip and Youk, 2020; Morris et al., 2011), cell signalling (Mittler et al., 2004, 2011; Mittler, 2017; Wood, Poole and Karplus,

2003; Rosenwasser et al., 2014; Fomenko et al., 2011; Petrov and Van Breusegem, 2012), circadian rhythms (Edgar et al., 2012), marine viruses (Sheyn et al., 2016), and marine cell gravitaxis (Carrara et al., 2021).

Our central goal is to examine how antioxidant systems influence cellular stoichiometry. In doing so, we ask: how might oxidative stress contribute to consistency and variation in cellular stoichiometry? We use the term 'contribution' because antioxidants can influence cellular stoichiometry through the processes they mediate, but they themselves form a portion of cellular elemental quotas. We focus on marine phytoplankton, as they are key players in global biogeochemical cycles (Falkowski, Fenchel and Delong, 2008), and there is motivation to move beyond model organisms and explore oxidative stress in diverse environments (Imlay, 2019).

We begin with some definitions and a brief review of conditions that lead to oxidative stress *in situ*, and then we highlight different antioxidant systems present in phototrophic phytoplankton, and their mechanisms. In discussing these mechanisms, we point to specific examples from different research fields that may have oceanographic relevance. In the following section, we ask: under increased oxidative stress, would increased production of a given antioxidant increase or decrease elemental ratios to carbon? Lastly, we assess the magnitude by which different systems could influence phytoplankton stoichiometry using previously published proteomic data.

5.3 What is Oxidative Stress, and Which Conditions Lead to it *In Situ*?

An antioxidant can be defined as "any substance that delays, prevents, or removes oxidative damage to a target molecule" (Halliwell and Gutteridge, 2007). We use the definition from Sies (1991) for oxidative stress, to be associated with "a disturbance in the prooxidant-antioxidant balance in favour of the prooxidant". Many antioxidant systems are directly involved in gene regulation (Mittler et al., 2011; Sies, 2017). In other words, "a disturbance in the prooxidant-antioxidant balance" does not necessarily equate irreversible damage, and this disturbance may directly influence gene expression. Indeed, many ROS (except hydroxyl radicals) have been invoked in some signalling context (singlet oxygen, hydrogen peroxide, and superoxide: Triantaphylidès and Havaux, 2009; Sies, 2017; Case, 2017, respectively). Therefore, the definition of oxidative stress we use is consistent with gene

regulatory functions of antioxidants.

In phytoplankton, oxidative stress is typically experienced when photosynthetic electron transport is in excess of that required for CO_2 fixation and nitrate assimilation (Asada, 2006). *In situ*, oxidative stress may correspond with low CO_2, high light, or low Fe. All of these conditions impact the rate of photosynthetic electron transport, more specifically, they typically increase the *proportion* of electrons leaking from the electron transport chain (producing superoxide). Superoxide is produced via reduction of molecular oxygen mostly at the reducing side of photosystem I (PSI; Asada, 2006). Oxygen can therefore act as a sink of electrons, which otherwise would have been donated to $NADP^+$. Photosynthetic electron transport is such a dominant source of ROS that even predators of photosynthetic cells have unique adaptations to their preys' photosynthetic oxidative stress (Uzuka et al., 2019).The unique reactions of different ROS with biomolecules are described in more detail below.

High exogenous hydrogen peroxide is also a direct source of oxidative stress (Cooper, Saltzman and Zika, 1987; Shaked, Harris and Klein-Kedem, 2010). However, hydrogen peroxide (H_2O_2) concentrations within cells are not identical to those outside of cells (Seaver and Imlay, 2001; Sies, 2017), so it is uncertain how much exogenous H_2O_2 (e.g. via rainfall) actually contributes to oxidative stress. These definitions of antioxidants and oxidative stress are broad, which reflects the broad uses (e.g. signalling, protective, etc.) of various antioxidant molecules.

For Fe specifically, it is unclear if there is more oxidative stress under low or high Fe. Under low Fe, photosynthetic electron transport is restricted, thus making the production of superoxide more likely (Niyogi, 1999). But, the dominant negative consequences of superoxide and H_2O_2 on biomolecules arise mainly through interactions with Fe (Anjem and Imlay, 2012; Imlay, 2013), therefore one might expect more oxidative stress under high Fe. Consistent with high Fe leading to oxidative stress, Anand et al. (2019) observed convergent evolution in several bacteria in the oxidative stress regulator OxyR under a high Fe treatment. Strikingly, Graff van Creveld et al. (2016) showed that chronic Fe starvation leads to more resistance to exogenous H_2O_2 than Fe replete conditions. They also showed that the chronic-Fe starved proteomic profile resembled *in situ* conditions observed using metatranscriptomics from Ocean Station Papa (Marchetti et al., 2011), suggesting that an exogenous ROS-tolerant phenotype under low Fe is the norm in iron-limited ocean

regimes.

5.4 Antioxidants

5.4.1 Enzymatic Consumers

5.4.1.1 Superoxide Dismutases

Superoxide dismutases (SODs) are ubiquitous enzymes with metal cofactors (Miller, 2012; Wolfe-Simon et al., 2005). They are incredibly efficient enzymes, converting superoxide into dioxygen and H_2O_2 (Equation 1) with first order rate constants approaching diffusion-limited rates:

$$2\,O_2\cdot^- + 2\,H^+ \longrightarrow H_2O_2 + O_2 \tag{5.1}$$

In addition to being kinetically fast, they are also broadly distributed throughout organisms on earth and evolved billions of years ago (Case, 2017). There are three families of SODs containing distinct metal cofactors: nickel SODs (NiSOD); copper / zinc SODs (CuZnSODs); and manganese / iron SODs (MnFeSODs) (Miller, 2012).

SODs protect against the deleterious effects of superoxide – but what are the exact effects of superoxide? Interestingly, superoxide reacts with most biomolecules at slow rates (Halliwell and Gutteridge, 2007; Winterbourn and Metodiewa, 1999). The main targets of superoxide are Fe-S clusters and mononuclear Fe enzymes (Imlay, 2013). Gu and Imlay (2013) elegantly showed that superoxide can abstract Fe from mononuclear Fe-containing enzymes *in vitro* (reversibly), which are then replaced by Zn resulting in a non-functional protein. This then requires re-metallating mononuclear Fe enzymes. SODs are therefore central to mitigating superoxide-induced mismetallation. We hypothesize that SODs play an important role in Southern Ocean phytoplankton in particular, where dissolved Zn levels are high (Vance et al., 2017), and Fe and Mn concentrations can be very low as well (Middag et al., 2011). Mismetallation could therefore strongly influence an organism's fitness, particularly given low Mn (Imlay, 2014). The expression of SOD can also lead to increased levels of H_2O_2 (Equation 1; Mittler et al., 2011), which can then have distinct deleterious effects (discussed below). Superoxide also can react directly with H_2O_2, which produces the hydroxyl radical, but the reaction of superoxide with H_2O_2 is

unlikely under physiological conditions (Haber and Weiss, 1932; Wardman and Candeias, 1996; Imlay, 2003).

SODs also have an atypical relationship with temperature, with higher rates of superoxide dismutation under colder temperatures (Perelman, Dubinsky and Martínez, 2006), which could increase requirements for trace metal in rapidly warming polar regions. Differential regulation of superoxide dismutases under various conditions has suggested that superoxide itself is a signalling molecule (Case, 2017). So, regulation of SODs may not solely be due to repression of superoxide levels alone, but rather the modulation of superoxide consumption and H_2O_2 production. Various viruses even encode SODs (e.g. Cao et al., 2002), which may alter the hosts' regulatory program by interfering with redox signalling. It would be beneficial to empirically quantify the drivers of the SOD expression-fitness landscape. In other words, is superoxide mostly a toxic byproduct of metabolism, or is it used for cell signalling? If phytoplankton SODs are mostly being used to prevent superoxide toxicity, there might be increased metal cofactor requirements in a warming ocean.

Antioxidant Category	Antioxidant	Stoichiometry	Median Monomer Length (SD)	Multimer	E. C. Number	N:C	P:C	S:C	Fe:C	(Zn, Cu, Mn, Ni):C
Enzymatic consumers	Mn Superoxide Dismutase	$C_{3.64} : H_{5.67} : O_{1.09} : N_1 : S_{0.03} : Mn_{0.003}$	235.5 (148)	Homodimer or homotetramer, each with 1 Mn atom[a]	1.15.1.1	↕	–	–	–	↑
	Fe Superoxide Dismutase	$C_{3.64} : H_{5.67} : O_{1.09} : N_1 : S_{0.03} : Fe_{0.003}$	235.5 (148)	Homodimer or homotetramer, each with 1 Fe atom[a]	1.15.1.1	↕	–	–	↑	–
	CuZn Superoxide Dismutase	$C_{3.64} : H_{5.67} : O_{1.09} : N_1 : S_{0.03} : Cu_{0.003} Zn_{0.003}$	235.5 (148)	Homodimer, each with 1 Zn and 1 Cu atom[a]	1.15.1.1	↕	–	–	–	↑
	Ni Superoxide Dismutase	$C_{3.64} : H_{5.67} : O_{1.09} : N_1 : S_{0.03} : Ni_{0.003}$	235.5 (148)	Homohexamer, each with 1 Ni atom[a]	1.15.1.1	↕	–	–	–	↑
	Ascorbate Peroxidase	$C_{3.53} : H_{5.58} : O_{1.08} : N_1 : S_{0.02} : Fe_{0.0022}$	336 (134)	Monomer, with 1 Fe atom (in a haem group)[b]	1.11.1.11	↕	–	–	↑	–
	Glutathione peroxidase	$C_{3.72} : H_{5.75} : O_{1.11} : N_1 : S_{0.02}$	197 (163)	Monomer or dimer[c]	1.11.1.9	↕	–	–	–	–
	Catalase	$C_{3.51} : H_{5.61} : N_1 : O_{1.07} : S_{0.03} : Fe_{0.002}$	367 (114)	Homotetramer, each with 1 Fe atom (in a haem group)[d]	1.11.1.6	↕	–	–	↑	–
	Peroxiredoxin	$C_{3.72} : H_{5.9} : O_{1.14} : N_1 : S_{0.03}$	221 (133)	Monomer to Homo-12-mer[e]	1.11.1.15	↕	–	–	–	–
	Cytochrome c peroxidase	$C_{3.44} : H_{5.46} : O_{1.06} : N_1 : S_{0.03} : Fe_{0.0032}$	459.5 (833)	Dimer, 2 Fe atoms per monomer (in a haem group)[f]	1.11.1.5	↕	–	–	↑	–
Non-enzymatic consumers	Manganese phosphate	$Mn_1 : H_1 : P_1 : O_4$	N.A.	N.A.	N.A.	–	–	–	–	↑
	DMSP	$C_5 : H_{10} : O_2 : S_1$	N.A.	N.A.	N.A.	–	–	↑	–	–
	Glutathione	$C_{10} : H_{17} : N_3 : O_6 : S_1$	N.A.	N.A.	N.A.	–	–	↑	–	–
	Ascorbate	$C_6 : H_7 : O_6$	N.A.	N.A.	N.A.	↓	↓	↓	↓	↓
	Tocopherols	$C_{29} : H_{50} : O_2^g$	N.A.	N.A.	N.A.	↓	↓	↓	↓	↓
	Carotenoids	$C_{40} : H_{56}^h$	N.A.	N.A.	N.A.	↓	↓	↓	↓	↓
Protective biomolecules	Chaperones	N.A.	N.A.	N.A.	N.A.	↕	–	–	–	–
	Polyphosphate	$P_1 : O_3$	N.A.	N.A.	N.A.	–	↑	–	–	–
	Ferritin	$C_{3.71} : H_{5.76} : O_{1.15} : N_1 : S_{0.03} : Fe_{0.0-0.083}^i$	230 (N.A.)	Homo-24-mer	1.16.3.1[i]	↕	–	–	–	

–

Table 5.1: Summary table of antioxidant systems in phytoplankton and their stoichiometric composition. Stoichiometric composition of proteins are given with respect to nitrogen. Proteins were subsetted from photosynthetic phytoplankton metagenome assembled genomes (Delmont et al., 2021), and then categorized using their Enzyme Commission (E.C.) numbers. Stoichiometry of macronutrients was determined from the amino acid sequence using pyteomics (Goloborodko et al., 2013). For micronutrient stoichiometry, we examined literature sources and the Protein Data Bank for structural information on ligands. On the right side of the table, we indicate the direction that various antioxidant systems might influence cellular stoichiometry, by asking how increased production of a given antioxidant would change stoichiometric composition (see text for rationale). [a] Sheng et al. (2014), [b] Protein Data Bank structure 1V0H, [c] Navrot et al. (2015), [d] Borges et al. (2014), [e] (Poole and Nelson, 2016), [f] Protein Data Bank structure 2VHD, [g] Formula for alpha-tocopherol is given, [h] (Marchetti et al., 2009); we did not use this E.C. number for assessing protein stoichiometry, but rather this *Pseudo-nitzschia multiseries* specific protein.

5.4.1.2 Catalases and Catalase-Peroxidases

Catalases (CATs) are ancient H_2O_2 metabolizing enzymes that dismutate two molecules of H_2O_2 (which is mostly generated from photosynthesis), with overall reaction stoichiometry (Zamocky, Furtmüller and Obinger, 2008; Zámocký et al., 2012; Tehrani and Moosavi-Movahedi, 2018; Vlasits et al., 2010):

$$2\,\mathrm{H_2O_2} \longrightarrow 2\,\mathrm{H_2O} + \mathrm{O_2} \tag{5.2}$$

Monofunctional catalases use haem to catalyze the above reaction. Peroxidase-catalases and Mn-catalases have similar catalytic mechanisms, except peroxidase-catalases can use an external reductant to reduce the active site (Tehrani and Moosavi-Movahedi, 2018; Vlasits et al., 2010). For peroxidases, the overall stoichiometry follows (where R denotes a reductant):

$$\mathrm{H_2O_2} + 2\,\mathrm{RH} \longrightarrow 2\,\mathrm{H_2O} + 2\,\mathrm{R}\cdot^{+} \tag{5.3}$$

Catalases are extremely efficient enzymes that are not saturated by H_2O_2 within most physiological concentration ranges (Aebi, 1984), and therefore do not display typical Michaelis-Menten kinetics (described further in Tehrani and Moosavi-Movahedi, 2018).

Peroxidase-catalases, however, can rely on an external reductant, and have a lower half saturation coefficient (i.e. a higher affinity; Vlasits et al., 2010). This difference in affinity provides a hypothesis for why bacteria have multiple enzymes that metabolize H_2O_2 (Mishra and Imlay, 2012). Catalases are typically inhibited by light, although inhibition can be protected against by chlorophyll (Feierabend and Germany, 1986).

How does H_2O_2 damage cells? H_2O_2 sluggishly reacts with most biomolecules (Halliwell and Gutteridge, 2007; Imlay, 2013; Winterbourn and Metodiewa, 1999). Its toxicity mainly derives from the interaction with Fe (or other metals with Fenton chemistry like copper), where the highly oxidizing hydroxyl radical is formed via Fenton chemistry. Fenton chemistry refers to the reaction of a reduced form of a metal reacting with an oxidant like H_2O_2, to produce an oxidized metal, a hydroxyl radical, and a hydroxide ion (Wardman and Candeias, 1996). No known enzyme 'metabolizes' hydroxyl radicals, and these ROS react at diffusion-limited rates. Notably, Mn does not have Fenton chemistry (Cheton and Archibald, 1988), and several hypotheses have been put forward suggesting that Mn-containing antioxidants have evolved because they lack the ability to produce hydroxyl radicals via this mechanism (Aguirre and Culotta, 2012).

5.4.1.3 Ascorbate Peroxidases and Glutathione Peroxidases

Catalases and peroxidase-catalases are complemented by several other H_2O_2 metabolizing enzymes. Glutathione peroxidases (GPxs) are thiol-based and ascorbate peroxidases are haem-based enzymes. These two enzyme groups are tied together through a common set of reductants. Ascorbate and glutathione are used as reductants (as in Equation 3). The concentrations of ascorbate and glutathione are critical for the kinetics of these two enzyme groups. For example, insufficient ascorbate will lead to rapid deactivation of ascorbate peroxidase (under 0.1 μM ascorbate; Miyake, Michihata and Asada, 1991). Detailed descriptions of these systems have been previously reviewed (Asada, 2006). The degree of coupling between glutathione-based and ascorbate-based antioxidant systems is complex, but modelling studies showed that coupling is partially dependent on the activity of a key enzyme, monodehydroascorbate radical reductase (Polle, 2001; Tuzet, Rahantaniaina and Noctor, 2019). Overall, there is incredible redundancy between these systems, catalases, and other H_2O_2 metabolizing systems (Tuzet, Rahantaniaina and Noctor, 2019; Mhamdi et al., 2010).

5.4.1.4 Peroxiredoxins

Peroxiredoxins (Prxs) metabolize H_2O_2 as above, they also first reduce H_2O_2 to water, and then are reduced by an external reductant (Karplus, 2015; Perkins et al., 2015). This reductant typically comes in the form of reduced thioredoxin (e.g. in *Synechocystis*; Pérez-Pérez et al., 2009). Peroxiredoxins are unique because they are oxidized at moderate concentrations of H_2O_2; and this oxidation-inactivation has been suggested to mediate barrier-free compartmentalization (Wood, Poole and Karplus, 2003; Perkins et al., 2015). Prxs also display chaperone behaviour (Jang et al., 2004). There is ample evidence that Prxs modulate H_2O_2 concentrations to control gene expression in highly localized sub-cellular regions (Brown et al., 2013; Perkins et al., 2015).

5.4.1.5 Cytochrome c Peroxidases

Cytochrome c peroxidases (CCPs) are haem-containing and they convert H_2O_2 into water using reduced cytochrome. In *Escherichia coli*, CCPs can donate electrons to H_2O_2 to be used as a terminal electron acceptor in respiration, and likely do not metabolize a large fraction of H_2O_2 (Khademian and Imlay, 2017). Jamers et al. (2006) showed that CCP in *Chlamydomonas reinhardtii* is differentially expressed under various copper stressors. Compared to the aforementioned antioxidant systems, the role of CCPs in photosynthetic microbes has received much less attention.

5.4.2 Non-Enzymatic Consumers

Non-enzymatic consumers of ROS are important factors in the defense and modulation of redox status in cells (Noctor, 2006). We should not only consider reaction rates between these compounds and ROS directly, but more importantly, the rate constants of regenerating oxidized compounds. Davies and Holt (2018) concluded this exact issue underpins why dietary antioxidants have failed clinical trials for their antioxidative effect – they require a kinetically fast system for regenerating the oxidized compound (also described in Imlay, 2013). This same argument may be applied to enzymatic antioxidant systems described above that require a reductant (e.g. ascorbate peroxidase).

5.4.2.1 Ascorbate and Glutathione

Ascorbate and glutathione (AsA and GSH) are both small molecule antioxidants. Yet, they are both key players in reducing antioxidative enzymes, and therefore essential components

of different antioxidant systems. Are these two small molecules important in reacting with ROS alone? It seems unlikely, because the first order reaction rate constants of ascorbate and glutathione are several orders of magnitude lower than those of enzymes which directly metabolize H_2O_2 or superoxide (Rahantaniaina et al., 2013).

There are many unknowns regarding the *in situ* role of GSH in particular – even in highly studied systems, the main routes of GSH oxidation are unclear (Rahantaniaina et al., 2013). Several studies have shown intriguing results. For example, GSH can chelate metals like Cu, therefore inhibiting production of hydroxyl radicals via Fenton chemistry (Halliwell and Gutteridge, 2007). Perhaps GSH can modulate Fe-induced oxidative stress in this manner, similar to ferritin. GSH is intertwined in multiple antioxidant enzyme systems (glutathione peroxidase, ascorbate peroxidase, peroxiredoxins). This may explain why GSH displays diurnal variations in concentration in phytoplankton (Dupont et al., 2004), with higher concentrations during the day.

5.4.2.2 Tocopherols and Carotenoids

Tocopherols and carotenoids protect against singlet oxygen, a unique ROS. This ROS is produced from the transfer of energy from a photosensitized chlorophyll to ground-state triplet dioxygen to form highly reactive singlet dioxygen. This transfer of energy changes the electron configuration of oxygen, which then substantially alters its reactivity (Laing, 1989). There are no known enzymes that metabolize singlet oxygen, but tocopherols and carotenoids can protect cells from singlet oxygen through two mechanisms: physical and chemical quenching (Krieger-Liszkay and Trebst, 2006; Ledford and Niyogi, 2005). Physical quenching occurs after the transfer of energy from singlet oxygen to a carotenoid, after which the energy is dissipated as heat. Chemical quenching is simply the reaction of singlet oxygen with either tocopherols or carotenoids (Ramel et al., 2012). After chemical quenching, the oxidized molecule is typically resynthesized (Ramel et al., 2012).

5.4.2.3 Other

Several other compounds have received some attention as antioxidants. For example, the highly studied marine metabolite DMSP displays antioxidant activity (Sunda et al., 2002). Manganous phosphate can also act as a superoxide dismutase (albeit with lower catalytic activity; Barnese et al., 2008). Interestingly, cobalamin (a cobalt-containing micronutrient) can also act as a superoxide dismutase with rates similar to SODs (Suarez-Moreira et al.,

2009). As above, antioxidant activity *in vitro* does not necessarily equate with activity *in vivo*, and it is unclear how these various compounds contribute to antioxidant system capacity in marine phytoplankton. However, these three examples may play important roles in sulfur, manganese, and cobalt quotas in photosynthetic microbes.

5.4.3 Protective Biomolecules

Several other antioxidant biomolecules have evolved that prevent reactions of ROS with target biomolecules, rather than destroying ROS. Protein chaperones, for example, act as protectors of unfolded proteins, which are particularly sensitive to oxidative stress (Santra, Dill and Graff, 2018; Dahl, Gray and Jakob, 2015). Ferritin, a large multi-unit protein that sequesters Fe, can play an important role in preventing interactions between H_2O_2 and Fe (Marchetti et al., 2009). Polyphosphates, which contribute varying amounts to total cell P (Lin et al., 2016) and have roles in phosphate and energy storage, can also protect proteins from ROS (Dahl, Gray and Jakob, 2015). Previous estimates suggest that polyphosphates can comprise of up to 40% of total P (Rhee, 1974; Geider and LaRoche, 2002). The requirement for polyphosphates as a protective antioxidant may specifically contribute to high variation in P quotas (Galbraith et al., 2013).

5.5 Antioxidant Influences on Cellular Stoichiometry

How do antioxidants impact macronutrient stoichiometry? If oxidative stress leads to increases in total protein per unit of cell biomass (via increased enzymatic antioxidants), this would increase N per cell, as protein is a large proportion of cellular N (therefore increasing N:C in cells, Geider and LaRoche, 2002). However, it is also possible that only the proportion of protein in antioxidants is shifted, which would then lead to no change in C:N:P ratios, but could influence metal or sulphur stoichiometry (depending on the composition and function of the antioxidant). This uncertainty is shown in Table 1, where we hypothesize how oxidative stress would influence stoichiometric ratios with carbon (only directions of influence are considered here). For other macronutrients, non-protein antioxidants may influence cell stoichiometry. For example, beta-carotenes might alter C per cell, polyphosphate might alter P per cell, and glutathione might alter S per cell (Table 1). Overall, antioxidants would probably have the largest impacts on micronutrient stoichiometry, because previous work has suggested they are a large fraction of the total

quota (for example MnSOD and NiSOD impacts on Mn and Ni quotas; Wolfe-Simon et al., 2006; Twining and Baines, 2013). The focus from this point on is specifically looking at antioxidant impacts on micronutrient quotas.

We expect that increased oxidative stress would result in increased expression of antioxidants (Table 1). For example, increased oxidative stress resulting from excess light would produce a saturated electron transport chain and increased superoxide production. This superoxide increase would then be met with increased amounts of NiSOD (for example). In this case, oxidative stress would increase the cellular Ni quota. Using a similar logic, we predicted how increased oxidative stress would change total cellular stoichiometry of N, P, S, Fe, Mn, Cu, Zn, and Ni (Table 1). An increase in superoxide production could be met with no change in antioxidant production, but we posit that this would eventually lead to deleterious effects from excess superoxide. The responses of antioxidants to oxidative stress are complex, and here we only aim to make first-order predictions.

For each antioxidant system, we estimated macronutrient stoichiometry from amino acid composition with a large dataset of photosynthetic phytoplankton metagenome assembled genomes (Delmont et al., 2021). Protein sequences (n = 2767) were subsetted using their Enzyme Commission numbers (Table 1), and then stoichiometric composition was empirically estimated, summarized using median values (Table 1, Figure 1; Goloborodko et al., 2013). Notably, there were no large differences in H:N, O:N, C:N, or even S:N ratios across antioxidants. Therefore, the major connections between antioxidant system use and cellular stoichiometry would arise from antioxidants that have unique cofactors. For example, glutathione peroxidases and catalases were quite similar in their macromolecular stoichiometry, but differ because catalase contains Fe (as haem). This analysis also showed that thiol-based antioxidant systems were not enriched in sulphur compared to non-thiol based antioxidant systems (Figure 1, Table 1).

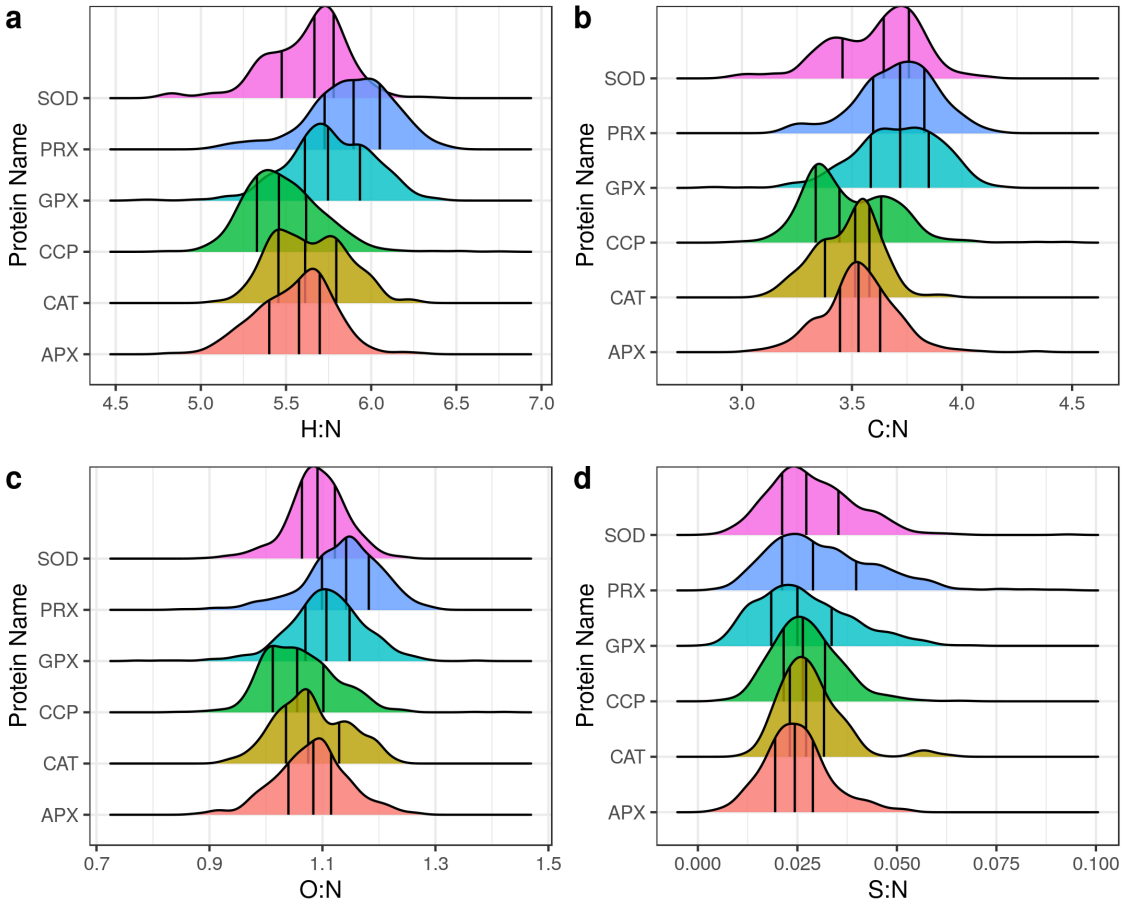

Figure 5.1: Distributions (kernel densities) of stoichiometric ratios for different enzymatic antioxidants. **a**, H:N; **b**, C:N; **c**, O:N; **d**, S:N. Data for protein lengths and stoichiometric composition are from photosynthetic phytoplankton metagenome assembled genomes (Delmont et al., 2021), and were subsetted using Enzyme Commission numbers from EggNogg annotations (Huerta-Cepas et al., 2019). The ticks on the vertical axis correspond with various enzymatic antioxidants: superoxide dismutase (SOD), peroxiredoxin (PRX), glutathione peroxidase (GPX), cytochrome c peroxidase (CCP), catalase (CAT), and ascorbate peroxidase (APX). Vertical lines correspond to the first, second, and third quartiles of the distribution.

5.6 Quantifying Antioxidant Contributions to Cellular Stoichiometry

5.6.1 Methods for Quantifying Antioxidant Contributions to Cellular Stoichiometry

How much variation in micronutrient stoichiometry is due to antioxidants, or more specifically to metal containing antioxidant enzymes? In this section, we use several data sources and some simplifying assumptions to examine the range of stoichiometric contributions antioxidants could have, with a focus on diatoms. Note that the range estimates we produce are not necessarily realized, as antioxidant expression could contribute to either the variation or consistency in micronutrient stoichiometry.

To illustrate these calculations (see Figure 2), consider the example: how much does FeSOD expression contribute to Fe:C variation? Fe:C ratios are particularly variable (e.g. Sunda and Huntsman, 1995; Twining et al., 2020), but it's unclear what underpins this variability. FeSOD is a dimeric protein, with each monomer containing one Fe cofactor. This is divided by the number of amino acids per enzyme molecule, converting this to Fe per amino acid. These two parameters (metal cofactor atoms per antioxidant molecule and amino acids per antioxidant molecule) are well constrained using genomic data and data on protein cofactors. We then multiply this value by the proportion of the proteome that is made up by FeSOD (details given below), now with units of Fe (from FeSOD) per total protein. Converting protein to N, we divide by the average number of N atoms per amino acid, and then multiply this value by the ratio of N in protein to N total. We incorporate variation in the ratio of N in protein:total N by sampling from a uniform distribution bounded by 0.5 and 0.85 (Geider and LaRoche, 2002). Lastly, we convert this ratio to Fe:C by multiplying by the Redfield ratio (16N:106C Redfield, 1958), but variation in N:C is incorporated by sampling from a truncated normal distribution (mean = 16, SD = 5, lower bound at 0 and no upper bound), and then adjusting the numerator, assuming a constant denominator.

One key parameter is the proportion of the proteome attributable to a given antioxidant protein. We used two previously published pennate diatom proteomes (*Fragilariopsis cylindrus*, *Phaeodactylum tricornutum*; Kennedy et al., 2019; Müller et al., 2020), and re-analyzed their data to obtain a range of proteomic proportion estimates for each micronutrient-containing enzymatic antioxidant. In brief, we converted mass spectrometry

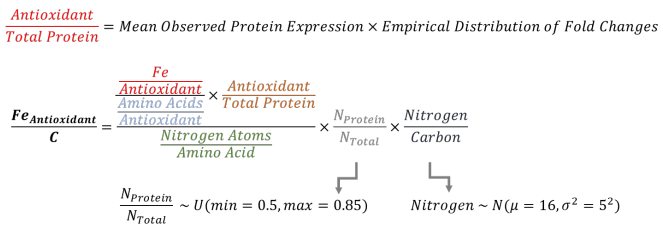

Figure 5.2: Illustrating the equation for obtaining distributions of Fe:C in antioxidants. Each coloured fraction represents a parameter. Fe per antioxidant is the number of Fe atoms per antioxidant, amino acids per antioxidant is the length of the protein. The antioxidant per total protein is empirically estimated from two previously published proteomes, and multiplied by the empirical distribution of fold change expression values. This entire value is converted into Fe:N in protein by dividing by the average number of nitrogen atoms per amino acid, and then converted to Fe:total N by multiplying by a value drawn from a uniform distribution of observed values (Geider and LaRoche, 2002). Lastly, we incorporate variation in N:C by adjusting just the numerator (N) by sampling from a truncated normal distribution with a bound at zero, and then dividing that value by 106.

raw files with ThermoRawFileParser (Hulstaert et al., 2020), appended a database of common contaminants (Global Proteome Machine Organization common Repository of Adventitious Proteins), searched mass spectra against a database of proteins (using published genomes, with MSGF+ and OpenMS; Mock et al., 2017; Bowler et al., 2008; Kim et al., 2014; Röst et al., 2016), and then quantified proteomic mass fraction by summing quantified peptides (quantified at the MS1 level with FeatureFinderIdentification; Weisser and Choudhary, 2017; Weisser et al., 2013). We then obtained the mean expression value across taxa to give a representative proteomic proportion for these pennate diatoms. One disadvantage of averaging over different diatoms is that their repertoire of antioxidants are slightly different. This becomes particularly important for the MnFeSOD family, because our predictions are different if the protein considered contains an Mn or an Fe cofactor. To address this, we show all calculations assuming that the observed MnFeSOD expression value is from a MnSOD or from a FeSOD. Code for all analyses is provided at: https://github.com/bertrand-lab/antiox-review.

Ideally, we would have antioxidant proteomic proportions observed across all realistic environmental conditions, which would give the exact contribution to variation in elemental stoichiometry. These data are currently unavailable (if they did exist, these calculations wouldn't be necessary!). However, we can estimate how most proteins vary using the distribution of fold changes for proteins across different environmental conditions. In

other words, how much might antioxidant expression change across all environmental conditions? We compared protein expression data from high and low Fe treatments in the coastal diatom *Thalassiosira pseudonana* (Nunn et al., 2013), and then examined the distribution of fold-changes. This distribution showed that most proteins ($\sim$75%) change between 2- and 20-fold across different environmental conditions (from 0.05-2 times). In using this distribution of fold changes, we make two key assumptions: 1) High and low Fe treatments with *T. pseudonana* represent typical variation in protein fold changes across taxa and conditions; and 2) antioxidants can be considered as 'average' proteins following this distribution. To assess the former assumption, we re-analyzed the fold-change distribution from an *E. coli* proteomic experiment which examined 22 different growth conditions (ranging from pH, to media, to growth phase; Schmidt et al., 2016). By comparing every condition with every other condition to calculate fold changes, we intriguingly found an almost identical distribution of fold changes. Using protein concentrations inferred from protein synthesis rates in *E. coli* under three conditions also revealed a similar distribution (Li et al., 2014a). Without specific data on antioxidant expression in phytoplankton across environmental gradients, the second assumption is difficult to rigorously assess. Of the antioxidants Nunn et al. (2013) observed varying across Fe concentrations, the median fold change was 1.75 (compared with a median of 0.97 for all proteins).

Conservatively, we moved forward by sampling from the empirical fold-change distributions (Nunn et al., 2013; Li et al., 2014a; Schmidt et al., 2016). So, we included this factor in the Monte Carlo sampling to estimate contributions from expression variability by sampling from each empirical distribution of fold changes (with equal probability of sampling from each of these three datasets). We then multiplied this sampled value by the average expression value across the two diatom taxa, which ends with a final estimate of antioxidant expression variation that contributes to Fe:C ratios in diatoms. We repeated the Monte Carlo sampling $1\text{x}10^6$ times to obtain distributions of metal:C.

5.6.2 Antioxidants Can Contribute Important Variation to Micronutrient:C

5.6.2.1 Fe-containing Antioxidants

Overall, we estimated that Fe-containing antioxidant expression can account for variation in Fe:C of between 3.3–10 (μmol:mol; depending on whether FeSOD or MnSOD is

present; Fig. 2, 2.5-97.5 quantile range). Median values of antioxidant contribution to Fe:C ratios were 0.7–2.2 (again dependent on MnSOD vs. FeSOD; Fig. 5.3). Fe:C can vary even greater than 100 μmol:mol (Twining et al., 2020), but this magnitude of variation is uncommon. Most Fe:C ratios vary around 30 Fe:C (μmol:mol; Twining et al., 2020; Strzepek et al., 2011, 2012; Sunda and Huntsman, 1995, however see Twining et al., (2020) for estimated variation across a wide gradient in the South Pacific Ocean). There are two main conclusions from this analysis: antioxidant contributions to Fe:C may explain some variation (11–33.3%, based on the effective range of 30 Fe:C). But, antioxidant systems are unlikely to explain the enormous variation sometimes observed (for example the $>$100 μmol:mol changes in Fe:C for some taxonomic groups; Twining et al., 2020).

Is this calculated range of antioxidant Fe:C important? In many areas of the ocean, Fe is a limiting resource, and it would seem reasonable to assume that 11–33.3% variation in the cellular Fe:C ratio would significantly affect growth. In low-Fe conditions, Fe-containing antioxidant expression would play a larger role in influencing Fe:C compared to high-Fe conditions. For example, diatoms in low Fe conditions can have Fe:C below 10 μmol:mol, and therefore Fe-containing antioxidants would impart a very significant stoichiometric signal. However, remember that antioxidant systems display extreme redundancy. All of these Fe-containing antioxidants have non-Fe counterparts: for example Mn-, CuZn-, or NiSOD instead of FeSOD, or glutathione peroxidase instead of ascorbate peroxidase. However, it is unclear how interchangeable these enzymes are, particularly given some of them are only expressed in certain subcellular compartments (see section on antioxidant system redundancy). Perhaps organisms retain these different antioxidants to respond to environmental conditions, such that these non-Fe counterparts would be used under low Fe conditions.

Antioxidants are important in mediating the negative effects of a high Fe quota, which would be particularly relevant with dramatic changes in Fe:C (Twining et al., 2020). This magnitude of change is likely due to high uptake of Fe, sometimes referred to as 'luxury' uptake (Twining et al., 2020). However, high uptake of Fe comes with a cost – free Fe can react with H_2O_2 to produce hydroxyl radicals (as discussed above). So high amounts of Fe must be met with either a system for metabolizing H_2O_2 to limit this reaction, or storing Fe to prevent contact with H_2O_2 sources (or both). In this case, antioxidant systems indirectly influence Fe stoichiometry.

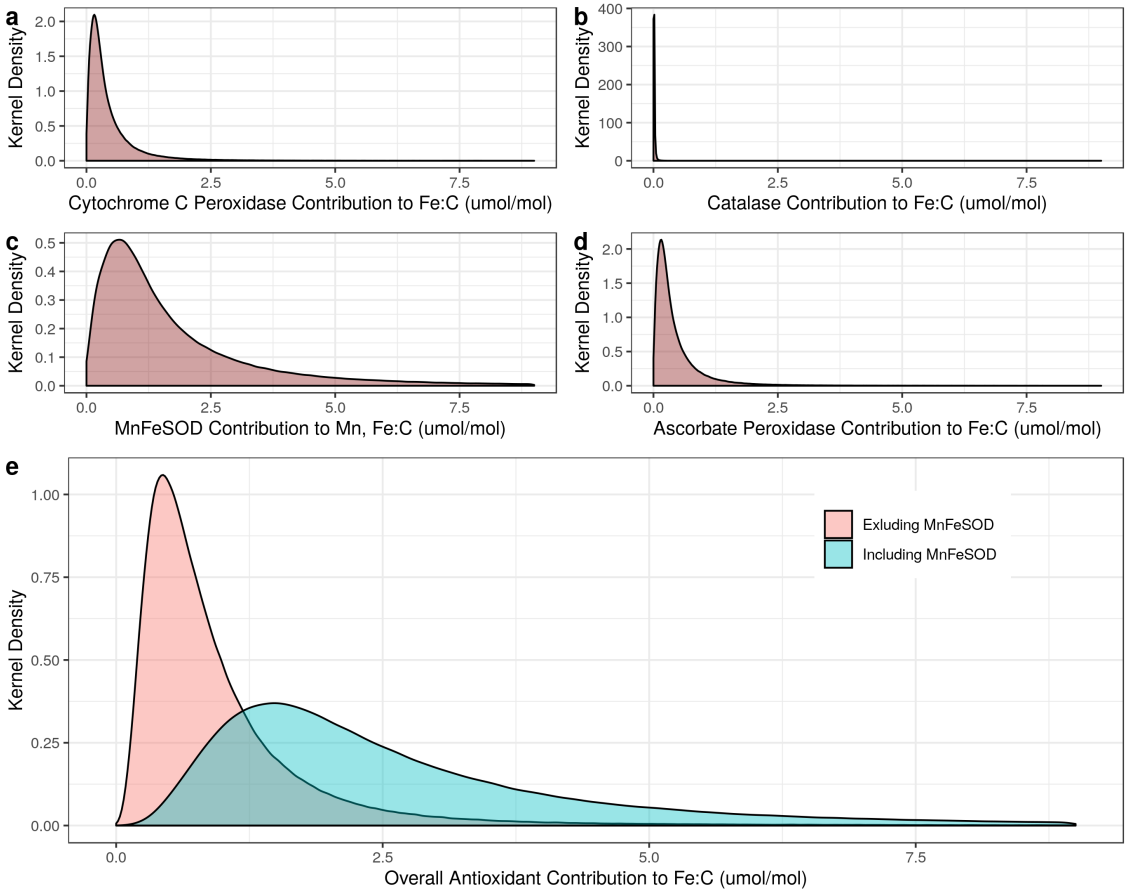

Figure 5.3: The potential contribution of antioxidant expression to Fe:C ratios, showing the kernel density estimates. **a**. cytochrome c peroxidase, **b**. catalase, **c**. MnFeSOD, and **d**. ascorbate peroxidase are all shown individually. **e**. The distribution of each Fe-containing antioxidant is summed. Two distributions are shown: 1) assuming the MnFeSOD is FeSOD, and 2) assuming it is MnSOD. Calculation underpinning the Monte Carlo estimates is shown in Fig. 5.2.

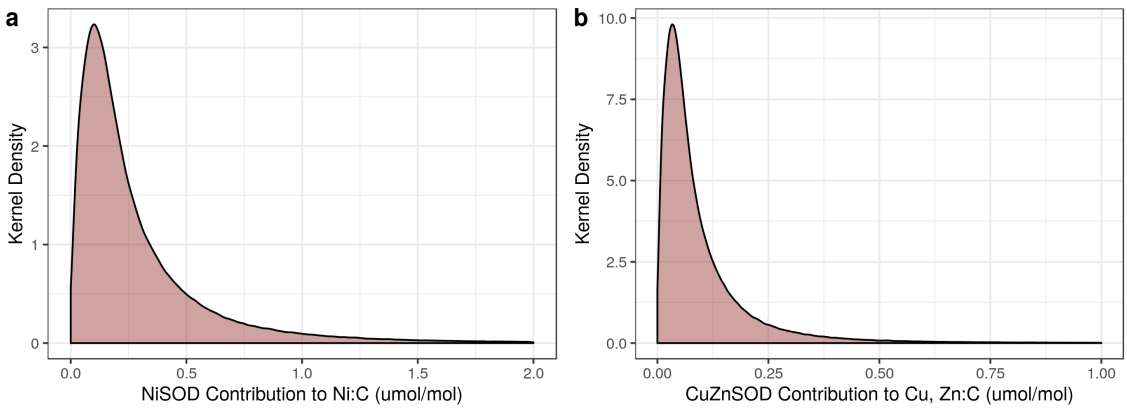

Figure 5.4: The potential contribution of antioxidant expression to Ni:C, Cu:N, and Zn:N ratios, showing the kernel density estimates. **a**. NiSOD, **b**. CuZnSOD. Calculation underpinning the Monte Carlo estimates is shown in Fig. 5.2.

5.6.2.2 Ni, Cu, Zn, and Mn-containing Antioxidants

Other micronutrients (Ni, Cu, Zn, and Mn) play important roles in global biogeochemistry (e.g. Twining et al., 2012; Richon and Tagliabue, 2019), and there is increasing evidence that some limit primary production (e.g. Mn; Buma et al., 1991; Wu et al., 2019; Browning et al., 2021). Comparing across cultures and field observations, Mn and Zn displayed similar amounts of variation across conditions compared to Fe (Twining, Baines and Fisher, 2004). We found that MnSOD contributions are unlikely to exceed 8.7 Mn:C (μmol:mol; 2.5–97.5 quantile range). Considering a range of Mn:C observations (3.4–46.7 μmol:mol for diatoms; Twining, Baines and Fisher, 2004), the variation in MnSOD expression that we calculated could account for $\sim$20% in Mn:C variation. Despite using very different approaches, our calculations complemented Wolfe-Simon et al. (2006) who found that chloroplast MnSOD accounted for 10-20% of cellular Mn. As with Fe, the contribution of MnSOD to Mn:C ratios would be even more important under low Mn. Consider the investment of Mn in MnSOD versus PSII (two dominant components of the Mn cellular quota in diatoms; Peers and Price, 2004; Wolfe-Simon et al., 2006). Under low Mn, we hypothesize that this SOD can be replaced by another SOD with a different metal cofactor, as there is no replacement for the Mn in PSII (see section on antioxidant system redundancy).

Moving to the other micronutrients we considered, NiSOD is likely to play a dominant role in Ni cell stoichiometry in diatoms (Fig. 4; Twining et al., 2012). Twining, Baines and Fisher (2004) observed that the maximum variation in Ni:C within taxa was 1.6 Ni:C

(μmol:mol), and we can attribute up to 1.37 Ni:C (μmol:mol; 2.5–97.5 quantile range) to NiSOD. It is interesting to note that some SODs are membrane associated (Ogawa et al., 1995; Regelsberger et al., 2002), so perhaps the high Ni content of diatom frustules is related to frustule-associated NiSOD (Twining et al., 2012). Another important Ni-dependent metalloenzyme is urease (Boer, Mulrooney and Hausinger, 2014). As with the MnSOD and PSII pair, we hypothesize that under low Ni conditions, NiSOD could be replaced with a different SOD but urease would not be.

5.6.2.3 Antioxidant System Redundancy and the Implications for Cellular Elemental Stoichiometry

Antioxidant systems display a lot of functional redundancy; what are the implications for cellular elemental stoichiometry? If various antioxidants are interchangeable, then environmental scarcity of an element would cause the production of another similar antioxidant that does not contain this scarce element. For example under low Fe, as occurs in much of the ocean, MnSOD might replace FeSOD, therefore implicating both Fe and Mn cellular stoichiometry. This prediction requires two assumptions: 1) 'nutritional coherence' and 2) functional similarity. We define 'nutritional coherence' as a characteristic of protein expression, such that environmental availability of an element would negatively correlate with protein expression, if a given protein uses this element as a cofactor. Are these reasonable assumptions? Page et al. (2012) provide evidence that FeSOD behaves in a nutritionally *in*coherent fashion. They found that FeSOD expression in *Chlamydomonas reinhardtii* increased under low Fe rather than decreased. Functional similarity is somewhat easier to assess. For example, all SODs catalyze the same reaction. However, even SOD isoforms with the same cofactors display unique expression patterns (e.g. Najmuldeen et al., 2019; Gallie and Chen, 2019), suggesting that even though functionally similar proteins can mediate the same reaction, that doesn't mean they actually do *in vivo*. Furthermore, this suggests that protein expression patterns, at least for SODs, are entrenched (Shah, McCandlish and Plotkin, 2015; Lalanne, Parker and Li, 2021) and perhaps less interchangeable than would be anticipated. Overall, the degree to which different antioxidants are functionally interchangeable in phytoplankton is a key unknown. We require empirical observations of antioxidant protein expression to first determine if a given antioxidant behaves in a 'nutritionally coherent' way. Then, we would be able to assess if functionally similar proteins are interchanged under conditions of elemental

scarcity. These types of observations would 1) help determine the relative costs of Mn in PSII versus MnSOD, or Ni in urease versus NiSOD, etc., and 2) determine the elemental stoichiometric consequences of antioxidant system redundancy.

5.6.3 Conclusions and Next Steps

We reviewed the central antioxidant systems present in phytoplankton, and used several approaches to identify and quantify how antioxidant system use may contribute to phytoplankton cell stoichiometry. Throughout, we have discussed various ways that antioxidant systems could influence cell stoichiometry, and concluded that they most likely have the largest impacts on micronutrient quotas.

Our original goal was to outline how antioxidant systems contribute to both variation and consistency in elemental stoichiometry. Using a series of simulations, we quantified how antioxidant systems may contribute to variation in trace metal quotas. A critical next step is quantifying how these systems are behaving *in situ*, to then determine the exact contribution under various environmental conditions. We were unable to assess how antioxidants influence consistency in elemental stoichiometry, and large-scale proteomic characterization of phytoplankton across diverse environmental conditions would achieve this goal. In terms of macronutrient stoichiometry, it is less likely that antioxidant systems play a dominant role, but a notable exception is polyphosphates for P quotas.

This leads us to synthesize two major unknowns and next steps for studying antioxidant systems in phytoplankton:

1. Is differential production of antioxidants a sign of oxidative stress in phytoplankton leading to damaged biomolecules, or of redox-based regulatory mechanisms? By extension, when cells are challenged with H_2O_2, for example, is this mainly inducing irreversible damage or interfering with regulatory networks? Quantifying what underpins protein expression-fitness landscapes is challenging, but promising new tools and techniques may suit these questions (e.g. Parker et al., 2020).

2. What are the environmental controls on specific antioxidants? There are many antioxidant systems in prokaryotes and eukaryotes (e.g. Mishra and Imlay, 2012). Quantifying how these superoxide- and hydrogen-peroxide metabolizing enzymes are produced in tandem may provide more insight into both the selective pressure on

antioxidant production (regulatory or metabolic), as well as their contributions to stoichiometry *in situ*. Also, antioxidants sometimes behave counter-intuitively (e.g. Page et al., 2012), so direct measurements of antioxidants under various environmental conditions is necessary.

Oxidative stress has shaped many facets of life. Describing and quantifying how the mediators of oxidative stress – antioxidants – affect cellular stoichiometry is important for connecting cellular processes to ocean biogeochemistry.

5.7 Data Availability

We provide code for all analyses at: https://github.com/bertrand-lab/antiox-review. All data used here are previously published.

5.8 Author Contributions

J.S.P.M. and E.M.B. conceived of the paper. J.S.P.M. wrote the paper, J.S.P.M. and E.M.B. edited and revised the paper.

CHAPTER 6

EXAMINING THE GROWTH-RIBOSOME ABUNDANCE RELATIONSHIP IN PHYTOPLANKTON UNDER MICRONUTRIENT AND LIGHT CONTROLLED GROWTH IN AN ANTARCTIC POLYNYA

6.1 Abstract

Ribosomes synthesize protein biomass, and variation in the number of ribosomes is thought to underpin variation in growth rate. This mechanistic connection was first identified because of observations connecting growth rate to the ribosomal protein mass fraction. Yet, these observations have largely been made in model heterotrophic organisms. Here we use 42 metaproteomes (3–12 micron size fraction) to characterize the ribosomal mass fraction in photosynthetic phytoplankton in the Amundsen Sea Polynya, Antarctica. The genera *Fragilariopsis* and *Phaeocystis* made up to 45 and 29% of total protein concentrations respectively, so we focused our analyses on these dominant taxa. We first show with previously published data that a simple Approximate Bayesian Computation method can be used to estimate the uncertainty around an observed ribosomal mass fraction. We then coupled the estimated ribosomal mass fraction with paired data on dissolved iron and manganese concentrations, macronutrients, light, and temperature. Ribosomal mass fraction and temperature were inversely related in the diatom genus *Fragilariopsis*,

and comparisons with other data suggest that there is a non-linear relationship between ribosomes and temperature in this genus. We did not observe strong relationships between other environmental variables and ribosomal mass fraction for *Fragilariopsis* spp. and *Phaeocystis* spp. Taken together, our work suggests that in these conditions, either protein turnover or a variable proportion of total translating ribosomes complicates the relationship between growth and ribosomes.

6.2 Introduction

Protein synthesis and growth rate are intimately related; to divide, organisms need to double their protein biomass. To divide *faster*, cells need either 1) more protein synthesizers (ribosomes), 2) increased translation rate per ribosome, 3) decreased protein turnover rate (i.e. decreased protein damage), or 4) increased proportion of actively translating ribosomes. Many researchers have observed a positive linear correlation between ribosomes and growth rate, suggesting that cells typically grow faster by increasing the number of ribosomes. Despite making several assumptions, Scott et al. (2010) showed that a simple model based on this linear relationship can quantitatively predict changes in growth rate under various conditions. Yet, the dominant focus of these studies has been on model, heterotrophic microorganisms. In addition, there has been comparatively little research on the relationship between ribosomes and non-carbon nutrients or other factors like light or temperature (however note Jahn et al., 2018).

In the ocean, nutrients like nitrate, phosphate, iron, and manganese can control phytoplankton growth (Wu et al., 2019; Tagliabue et al., 2017; Moore et al., 2013). When these nutrients limit growth in the ocean, what are the cellular processes that are limiting growth? For example, does total protein synthesis via variable total ribosomes underpin different growth rates in these ocean environments? There is a major gap in our understanding of how ribosomes vary with growth rate in the ocean because of the diversity of resources and organisms. This gap is partially because of difficulties in culturing many taxonomic groups, both in terms of replicating environmental conditions (e.g. extremely low concentrations of trace nutrients) and isolating and maintaining cultures of poorly characterized microbes.

One promising approach for studying gene expression in diverse microbes is to not culture them at all, but rather observe their gene expression *in situ*. Metaproteomics is an appropriate tool for probing microbial gene expression in this way, and has been

used to explore how gene expression across environmental gradients changes for various organisms (e.g. Saito et al., 2014). Yet, untargeted metaproteomics so far has not been used to quantitatively assess proteomic composition in a way that aims to characterize the 'true' proteomic composition. For example, Cohen et al. (2021) studied dinoflagellate gene expression across environmental gradients in the Pacific Ocean using metaproteomics and metatranscriptomics. However, the metric they used to assess dinoflagellate proteome composition can only assess relative changes. To easily connect metaproteomic observations with proteomes from cultured organisms or computational models, we need to quantitatively assess proteomes from metaproteomes.

In this contribution, we use metaproteomics to characterize the ribosomal mass fraction (RMF), in two important photosynthetic marine eukaryotes. We define the ribosomal mass fraction as the proportion of the proteome, by mass, allocated to ribosomal proteins. This quantity is different from relative changes in expression because it provides information about the absolute investment in protein synthesis machinery. All of our metaproteomic observations are paired with environmental data, including measurements of extremely low trace metal concentrations (picomolar level). We focused our analyses on two dominant genera: *Fragilariopsis* and *Phaeocystis*, as they were both dominant in our samples and have been previously included in biogeochemical models of the region (Kwon et al., 2021). We then attempt to explain variation in RMF using these environmental characteristics, observing a negative relationship with temperature and RMF in *Fragilariopsis* spp.

6.3 Methods

6.3.1 Sample Collection

Samples were collected in the Amundsen Sea Polynya, Southern Ocean from December 2017 until February 2018 aboard the icebreaker RV Araon, from 15 different stations (locations with unique latitude and longitude). We used a trace-metal clean sampling system (Titan; De Baar et al., 2008), with a mounted conductivity (salinity), temperature, and depth sensor (CTD; Seabird SBE 911+). After the Titan sampling system was brought aboard the ship it was moved into a cleanroom environment for subsampling. Water was collected and filtration for metaproteomics began between 1 to 2.5 hours after samples were brought aboard the ship. Keeping the water containers on ice packs, we filtered water

using a peristaltic pump through a series of connected polycarbonate filters of decreasing size (12.0, 3.0, and 0.2 μm pore sizes). Filtration was stopped after 1.5 hours or the filters clogged, and subsequently stored at -80 °C until protein extraction. We also collected water for analyzing dissolved and particulate trace metal concentrations at corresponding depths to the metaproteomic samples, which are discussed elsewhere (van Manen, 2021). Water was also collected on a separate rosette sampling system from the Korea Polar Research Institute for analyzing dissolved nitrate and nitrite, and dissolved silicate. The exact depths for these nutrient concentrations did not match identically to the metaproteomic samples, but differed by a median value of 5 m.

6.3.2 Proteomic Sample Preparation

Proteins were extracted from only the 3.0 μm filters frozen in cryovials with the following method. We focused on the 3.0 μm filter size to explore *Fragilariopsis* spp. and *Phaeocystis* spp. non-colony protein expression, which should be predominantly captured on this filter size. Protein extraction buffer (0.1 M Tris/HCl, pH 7.5, 5% glycerol, 5 mM EDTA, 2% SDS) was put into the cryovial, after which it was incubated at 95 °C for 15 minutes. Filters with extraction buffer were then sonicated on ice (15 seconds on, 15 seconds off, 2 minutes total sonication time, 50% amplitude 125 W, Qsonica Sonicator Q125, Newtown, Connecticut, USA). After sonication, we incubated the sample at room temperature for 30 minutes. Extracted protein in buffer was then removed from the cryovial, centrifuged at 15,000 G at room temperature for 30 minutes to pellet cell debris, and the supernatant was removed and stored at -80 °C. We measured the total protein concentration using a BCA assay (Thermo Fisher Scientific, California, USA) at this point to then calculate the total μg protein per volume seawater.

We then reduced and alkylated the extracted protein, and removed the SDS extraction buffer using S-traps (Protifi, Farmingdale, New York, USA). We first prepared solutions of 500 mM dithiothreitol (DTT) and 500 mM iodoacetamide (IAM) in 50 mM ammonium bicarbonate. We then reduced the protein with DTT, bringing up the concentration to 5 mM and incubating at 37 °C for one hour in a Thermomixer (F1.5, Eppendorf, Hamburg, Germany) at 350 RPM. Reduced protein was then cooled to room temperature, and alkylated using IAM, bringing the concentration to three times that of DTT (15 mM). After incubating in the dark for 30 minutes at room temperature, we then quenched the reaction with 5 mM of DTT. We denatured the extracted proteins with 12% phosphoric

acid by bringing it to 1.2% by volume. Our samples were then diluted with S-trap buffer (1:7, sample : S-trap buffer; 90% methanol in 100 mM triethylammonium bicarbonate, acidified to 7.1 pH with phosphoric acid). The sample and S-trap mixture were then loaded onto the S-traps, which were kept on a vacuum manifold but prevented from becoming completely dry. After the sample and S-trap mixture was fully loaded on each unit, we washed the sample with 10x 600 μL of S-trap buffer to remove the SDS. For the first three washes, buffer was left on the S-trap without using the vacuum pump. S-traps with sample loaded were then centrifuged at 4000 XG for 1 minute to remove remaining S-trap buffer. Finally, we digested the protein using trypsin in 50 mM ammonium bicarbonate, with a ratio of 1:25 sample protein:trypsin, and incubated them at 37 °C for 16 hours. Peptides were then eluted from the S-traps with 80 μL of 50 mM ammonium bicarbonate, 80 μL 0.2 % aqueous formic acid, and 80 μL 50% acetonitrile containing 0.2% formic acid. Samples were then dried in a Vacufuge Plus (aqueous vacuum setting (V-AQ), room temperature, Eppendorf, Hamburg, Germany; between 3-4 hours), and then reconstituted in 3% acetonitrile and 0.1% formic acid.

The samples were desalted using 50 mg C18 columns (HyperSep, Thermo Fisher Scientific). Columns were first conditioned with 500 μL methanol, and then 500 μL 50% acetonitrile, 0.1% formic acid. Columns were then equilibrated with two aliquots of 500 μL 0.1% trifluoroacetic acid. We then increased the volume of samples that were previously reconstituted with 3% acetonitrile and 1% formic acid by adding 100 μL of this solution, and then loaded the diluted sample onto the equilibrated column. Samples were then pushed through the column using a syringe, and then washed three times with 1000 μL of 0.1% TFA each time, removing the salt and retaining the peptides on the column. Finally, peptides were eluted into a low binding plastic microcentrifuge tube (Thermo Fisher Scientific, California, USA) with two aliquots of 200 μL of 50% acetonitrile and 0.1% formic acid, and then one aliquot of 70% acetonitrile and 0.1% formic acid. Samples were then dried down using a Vacufuge Plus (Eppendorf, Hamburg, Germany; between 5-6 hours), until only the dried peptides remained.

6.3.3 Liquid Chromatography Mass Spectrometry

We used liquid chromatographic (LC) separation of the complex peptide mixture to reduce the sample complexity prior to injecting into the mass spectrometer. The LC was coupled directly to a Q Exactive hybrid quadrapole-Orbitrap mass spectrometer (Thermo Fisher

Scientific, California, USA), and the entire run lasted 125 minutes, using a non-linear gradient. Solvent A consisted of 0.1% formic acid in water, and solvent B contained 0.1% in acetonitrile. From the start until 15 minutes, the flow rate was 0.3 μl/minute and 5% solvent B. From 15.1 until 90 minutes, the flow rate was 0.25 μ/minute and solvent B increased to 30% linearly. From 90 until 102 minutes, solvent B was increased to 55% linearly. From 102.1 until 106 minutes, solvent B was increased to 95% linearly at a flow rate of 0.3 μ/minute. The flow rate was kept constant at 95% solvent B until 110 minutes. From 111 until 125 minutes, solvent B was decreased to 5%. We used a data-dependent acquisition mass spectrometry approach, specifically with a TopN of value of 8. The MS1 scans were run at 140,000 resolution, with a scan range from 400 to 2000 *m/z* and an automatic gain control target of 3E6. For the MS2 scans, we chose a resolution of 17,500, and automatic gain control target of 1E6, an isolation window of 2 *m/z*, and a scan range of 200 to 2000 *m/z*.

6.3.4 Metaproteomic Bioinformatics

Metaproteomics requires a database of potential proteins to search mass spectra against. We used a custom database of metatranscriptomic sequences from the neighbouring Ross Sea (Jabre et al., 2021), and appended those protein sequences with sequences from metagenome-assembled genomes (MAGs, described below; Delmont et al., 2021). We were interested in the diversity of antioxidant proteins that phytoplankton are using, so we identified all antioxidant-proteins of interest from this large collection of eukaryotic MAGs using Enzyme Commission numbers associated with their sequence annotations. We then reduced the database size by combining protein sequences that are 95% or higher sequence similarity (with CD-HIT, Li and Godzik, 2006). Finally, we appended a database of common contaminants (Global Proteome Machine Organization common Repository of Adventitious Proteins). In total, there were 414498 protein sequences in our database. We then used MSGF+ (Kim et al., 2014) within OpenMS (Röst et al., 2016) with the following settings: fixed cysteine carbamidomethyl, and variable methionine oxidation, N-terminal glutamate to pyroglutamate, deamidated asparagine, and deamidated glutamine. A 1% false discovery rate was applied at the peptide spectrum match level. Raw mass spectrometry files were converted to mzML using ThermoRawFileParser (Hulstaert et al., 2020).

Peptides were quantified at the MS1 level with their corresponding ion intensities

(Weisser et al., 2013; Weisser and Choudhary, 2017). FeatureFinderIdentification is an approach that cross-maps identified MS2 spectra to unidentified features across samples. This approach requires grouping samples, and we conservatively only grouped samples that were technical replicates. Note that for two samples, we were unable to acquire duplicate injections (sample ID number 74 and 197). Peptide abundances were then calculated for each injection by taking the sum of peptide intensities per injection, and normalizing the intensities of each peptide by this sum. This method is a database-dependent normalization, which can lead to problematic inferences (McCain, Allen and Bertrand, 2021). We ensured our quantifications were robust by correlating normalizing factors to total ion current across all samples.

6.3.5 Ribosomal Mass Fraction

Ribosomes are a critical molecular machinery for all microbes, and they are comprised of many different proteins. Our goal is to estimate the ribosomal mass fraction (RMF), defined as:

$$RMF = \frac{\sum_{i \in Ribosomal}^{N} A_i}{\sum_{i}^{N} A_i} \tag{6.1}$$

Where *i* denotes a unique peptide, A_i is the abundance of the i^{th} peptide, and the numerator only considers those peptides that unambiguously correspond to a ribosomal protein. Throughout we only consider peptides that uniquely map to a taxon, and below we only consider the two genera *Fragilariopsis* and *Phaeocystis*. We considered a peptide taxonomically uninformative (and therefore not included) if the peptide amino acid sequence was found in two or more proteins, and those protein sequences were from distinct genera. We specify the ribosomal *mass* fraction because we do not normalize for the length per protein. We also assume that the mass per peptide is relatively constant, which is empirically observed. Previous analyses have shown that adjusting for the mass per peptide has a negligible impact on inferences about proteomic composition.

With high amounts of sampling, the above equation can approximate the true RMF well. However, mass spectrometers sample peptides according to abundance, and this bias can lead to problematic inferences of RMF. More specifically, in data-dependent acquisition experiments, ions are sampled explicitly because of their abundance (relative to

the abundance of other co-eluting ions). Even in data-independent acquisition experiments, there is an implicit bias towards abundant ions because there is a higher probability of identifying abundant peptides.

Consider two scenarios of diatom abundance in an assemblage of organisms, but with no change in RMF. In the first scenario, diatom abundance is relatively high. Because of this, lesser-abundant diatom proteins will be quantified and the denominator will be high. In the second scenario, diatom abundance is relatively low. Because ribosomes tend to be a large proportion of the proteome, they are highly likely to be detected (if diatom proteins are detected at all). However, the lesser-abundant diatom proteins are unlikely to be detected. Therefore, the measured RMF will appear to increase despite no change in the actual RMF. This issue is compounded because it is unclear how to quantify uncertainty for an observed RMF. We therefore require methods for explicitly addressing these types of bias, and to quantify uncertainty around the observed RMF.

We use a simple Approximate Bayesian Computation (ABC) method for inferring the RMF in metaproteomic samples. First, we generate a distribution of peptide intensities from a lognormal distribution with the log mean equal to 13.8, and the log standard deviation equal to 1.45 (generating 10000 peptides). These values were obtained empirically by determining the maximum likelihood estimates of distribution parameter values using peptide intensities from a pure *Pseudomonas denitrificans* culture (Kleiner et al., 2017). We then generate a value from a uniform distribution bounded by 0 and 1, which parameterizes a Bernoulli distribution categorizing all peptide abundance values as either ribosomal or non-ribosomal. Peptides are then sampled with probabilities proportional to their abundances, and an observed RMF is calculated from this sample. The number of samples is equivalent to the empirically observed number of unique peptides sampled from the metaproteome. For example, if we observed only ten peptides for *Fragilariopsis* spp. in a metaproteomic sample, then we would only sample ten times during this procedure. Finally, we calculate the absolute difference between the sampled and observed RMF, and use top 1% closest generated datasets to calculate the 95% credible intervals for the true ribosomal mass fraction. In total, we generate 10000 empirical distributions.

To assess this approach, we used a previously published artificially assembled metaproteome Kleiner et al. (2017). From these data, we specifically used the uneven-protein

per organism metaproteome (the short and long chromatographic runs), and the even-protein per organism metaproteome. Four, pure-culture proteomes were also characterized in Kleiner et al. (2017), which were the exact same cultures used in the metaproteome assembly. From these four pure cultures, we calculated a 'true' RMF.

6.3.6 Statistical Analyses of Southern Ocean data

Every metaproteomic sample had paired environmental data: dissolved iron and manganese (dFe, dMn), temperature, and light. These dissolved metal data were not collected from identical bottles, but bottles were collected at the same water depth on the same CTD cast. We hypothesized that these environmental variables are dominant controls on ribosomal mass fraction, and therefore sought to determine if they explained variation in RMF for two major phytoplankton taxa, *Phaeocystis* spp. and *Fragilariopsis* spp. To do so, we used a weighted linear regression, where the weights were inversely proportional to the 95% credible intervals derived from the ABC method. Specifically, we fit the following model (for each taxon j), where β represents the coefficient value:

$$\begin{aligned} logit(RMF_j) \sim \beta_{0,j} + \beta_{1,j} \cdot Temperature + \\ \beta_{2,j} \cdot dFe + \beta_{3,j} \cdot dMn + \beta_{4,j} \cdot PAR + \beta_{5,j} \cdot [Nitrate + Nitrite] \end{aligned} \tag{6.2}$$

Note that we originally included silicate as an explanatory variable in this regression. However, silicate was highly correlated with dMn and nitrate and nitrite, so we subsequently removed it. The RMF values were transformed using a logit function.

6.4 Results

6.4.1 Estimating the Ribosomal Mass Fraction from Metaproteomes

Our simple ABC method applied to an artificially assembled metaproteome generally captured the 'true' RMF from pure culture proteomes for two of the three species (using data from Kleiner et al., 2017), suggesting that this method can be used for metaproteomic samples under specific circumstances (Fig. 6.1). We overestimated the RMF in all cases for the *Pseudomonas denitrificans* proteome. In this case, the artificially assembled metaproteome included two other species within the same genus: *P. fluorescens* and *P. pseudoalcaligenes*. In previous work we showed using simulations that low diversity can

lead to underestimation of a proteomic quantity from a metaproteome, when using the sum of peptide intensities (McCain, Allen and Bertrand, 2021). This is because low sequence diversity can limit the number of unique peptides that are taxonomically informative. Consistent with these simulations, the denominator of total *P. denitrificans* protein was underestimated, increasing the inferred RMF. In other words, there were comparatively few peptides that could be used to distinguish between the different *Pseudomonas* species. Overall, these results suggested that this simple ABC method cannot be used to estimate RMF at the species level, and provided reasonable estimates of uncertainty. Note that Kleiner et al. (2017) did not include single-taxon proteomes for the other two *Pseudomonas* species, so we cannot evaluate the performance at the genus level.

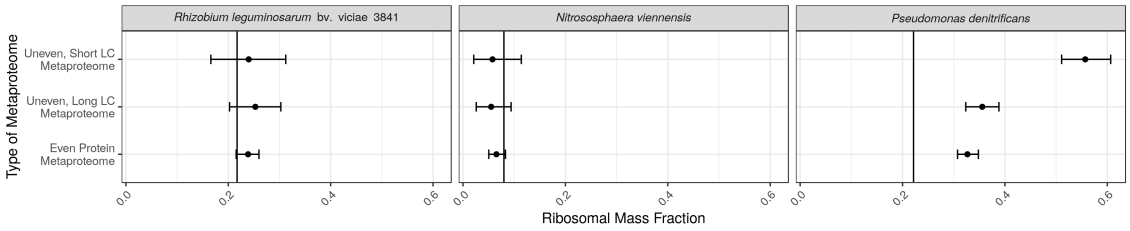

Figure 6.1: Estimating the ribosomal mass fraction (RMF) in three lab-generated metaproteomic samples (vertical axis), across different organisms (different panels) (Kleiner et al., 2017). In total, there were 30 different organisms used, and here we compare three of those 30 that had corresponding pure culture proteomes. The 'true' RMF observed from the pure cultures is shown as vertical lines in each panel. Points represent the median of the approximate posterior distribution, and the error bars correspond to the 95% credible intervals of the approximate posterior distribution.

6.4.2 Amundsen Sea Metaproteome Characterization and Taxonomic Abundance Profiles

In total, we matched 410876 spectra to peptides (peptide-spectrum matches) across all samples. Of these, there were in total 32535 unique peptides. To ensure our database was not differentially performing and therefore affecting our normalization and quantification (McCain, Allen and Bertrand, 2021), we correlated the TIC for each sample with the sum of observed peptide intensities. The Pearson correlation coefficient was 0.93 (Supplementary Fig. 6.6), indicating that we identified the majority of protein across diverse samples and therefore it is unlikely the database choice biased peptide quantification. Another measure of database performance is the number of MS2 spectra identified as a percentage of total MS2 spectra collected per mass spectrometry experiment. We identified on average 25%

of MS2 spectra (SD = 6%), with a minimum of 16% and a maximum of 38%. These values further support that database choice did not bias peptide quantification.

We used these metaproteomic data to examine the vertical distribution of taxon-specific protein in the water column in the Amundsen Sea Polynya (Fig. 6.3). We calculated the proportion of peptide intensities mapped to two taxa, *Fragilariopsis* spp. and *Phaeocystis* spp., by taking the sum of all taxon-specific peptide intensities divided by the sum of all peptides that were taxonomically informative. Note that at this stage we excluded peptides that are taxonomically uninformative to better approximate the proportion of total protein for each taxa. This proportion was then multiplied by the total protein per filter, and then divided by the total volume of seawater filtered, which yielded a taxon-specific protein concentration profile (Fig. 6.3).

These profiles clearly illustrated two different biogeographic regions: *Fragilariopsis* spp. dominated versus *Phaeocystis* spp. dominated. In the center of the Amundsen Sea Polynya and near stations 31–49, *Phaeocystis* spp. protein tended to be more abundant than *Fragilariopsis* spp. protein. Closer to the edge of the polynya and into the marginal sea ice zone (stations 50–52), *Fragilariopsis* spp. tended to be more dominant. In stations 24, 55, and 57, *Fragilariopsis* spp. was much more abundant than *Phaeocystis* spp. Overall, these two genera contributed a large proportion of total protein (Fig. 6.3), even up to 49 and 25% for *Fragilariopsis* and *Phaeocystis* respectively.

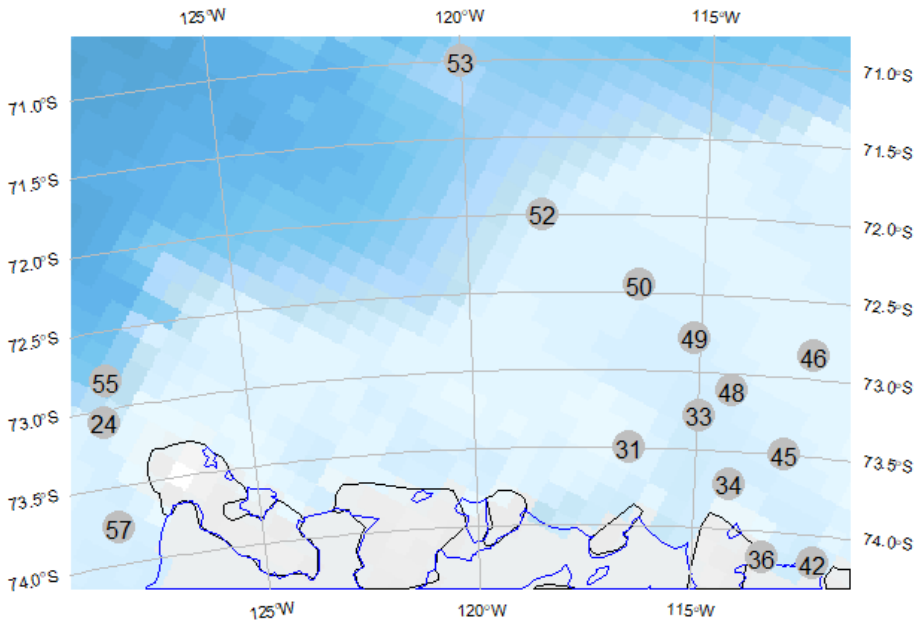

Figure 6.2: Map of the stations that had corresponding metaproteomic samples at various depths. Station numbers are displayed within each grey circle. Background colour represents the bottom bathymetry, and the black lines near the bottom represent the land, while blue lines represent ice. Note that station 53 was in the marginal sea ice zone.

6.4.3 Environmental Correlates of the Ribosomal Mass Fraction

We estimated the ribosomal mass fraction of two taxonomic groups, observing strikingly constant vertical profiles in RMF for both *Fragilariopsis* spp. and *Phaeocystis* spp. (Fig 6.4). These vertical profiles demonstrated reproducibility and consistency of these metaproteomic methods across 42 different samples.

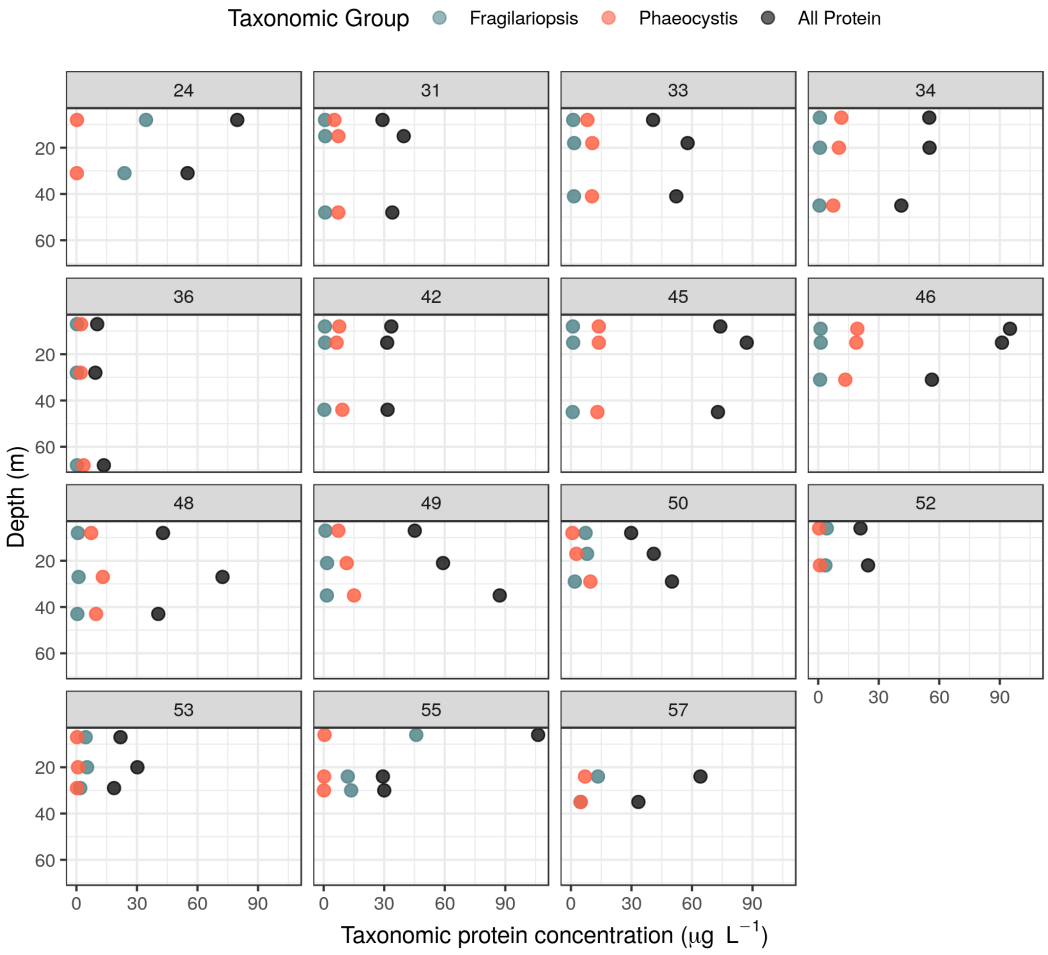

Figure 6.3: Protein concentration with depth attributed to *Fragilariopsis* spp. and *Phaeocystis* spp. across 15 different stations, as well as total protein concentration.

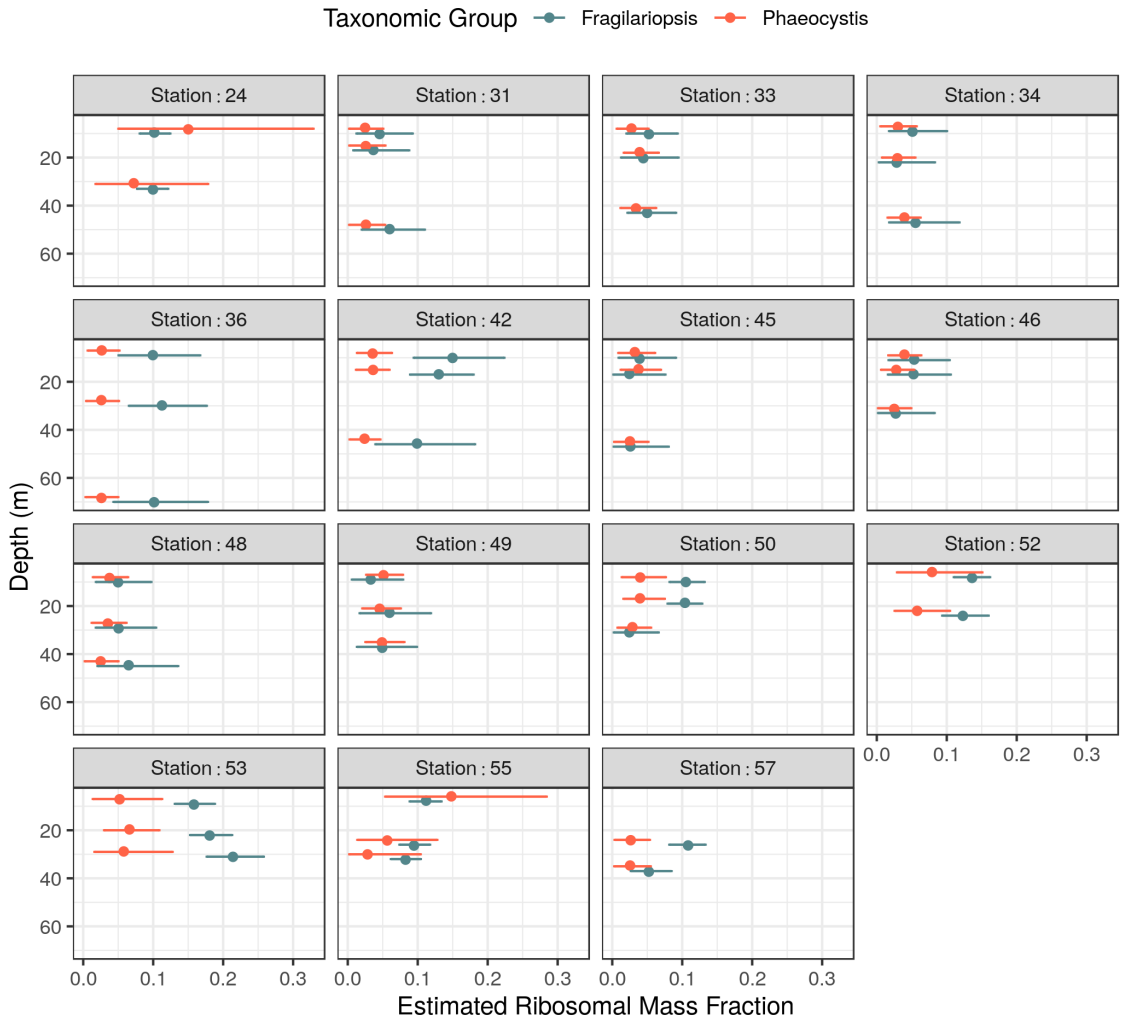

Figure 6.4: Variation of estimated ribosomal mass fraction with depth for *Fragilariopsis* spp. and *Phaeocystis* spp., with corresponding 95% credible intervals. Each panel corresponds to a unique station, with station numbers displayed in Fig. 6.2. Note that the estimates for *Fragilariopsis* spp. are vertically offset for visualization purposes by 2 m.

We sought to explain variation in RMF using the coupled nutrient concentration data (considering dFe, dMn, and nitrate and nitrite), as well as the *in situ* light and temperature values (Fig. 6.5). For *Fragilariopsis* spp., we observed no significant correlation between dissolved trace metals or light, but observed a significant negative correlation between temperature and RMF (Fig. 6.5, Table 6.1). This coefficient means that for a degree increase in temperature (for example, from -2 to -1 C), the RMF decreased 5.3%. (Note that because of the logit transform this decrease is not linear). For *Phaeocystis* spp., we observed a significant negative correlation between dMn and RMF (Fig. 6.5, Table 6.1).

However, the effect size for this coefficient was very low: a 1 nM increase (from 1 nM to 2 nM) in dMn corresponds with a 0.3% decrease in RMF.

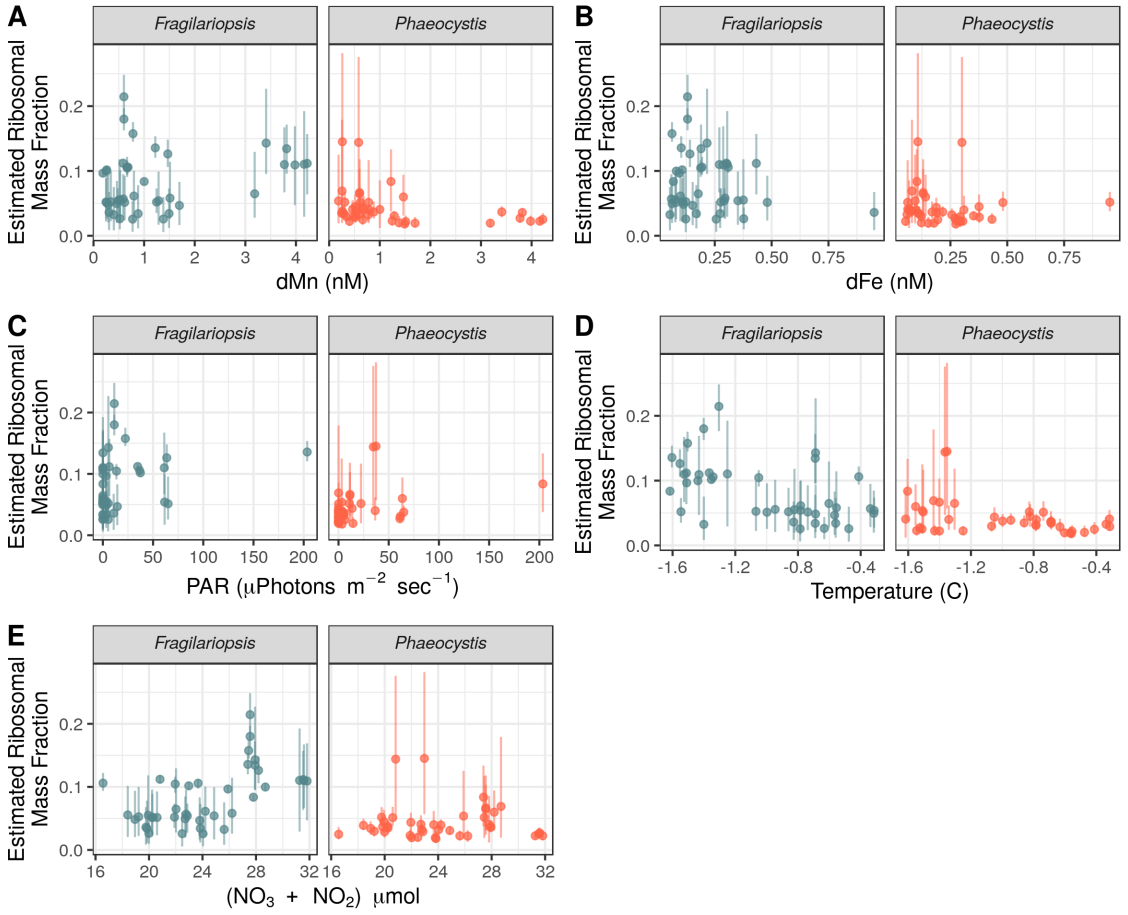

Figure 6.5: Variation of estimated RMF with four environmental variables (dissolved Fe, dissolved Mn, light, and temperature) for both *Fragilariopsis* spp. and *Phaeocystis* spp.

Taxonomic Group	Environmental Variable	Estimate	Standard Error	t value	p value
Fragilariopsis spp.	β_0	-3.26809	0.625875	-5.22162	7.62E-06
	β_1 (Temperature)	-0.63853	0.237113	-2.69292	0.010685
	β_2 (dFe)	-0.51537	0.554276	-0.92981	0.35866
	β_3 (dMn)	0.068376	0.104602	0.653674	0.517475
	β_4 (PAR)	0.002634	0.002054	1.282352	0.207916
	β_5 (NO3 + NO2)	0.003637	0.03088	0.117789	0.90689
Phaeocystis spp.	β_0	-3.75283	0.424584	-8.83883	1.51E-10
	β_1 (Temperature)	-0.11715	0.15407	-0.76038	0.451981
	β_2 (dFe)	0.239665	0.34303	0.698671	0.489247
	β_3 (dMn)	-0.14188	0.060694	-2.33758	0.025086
	β_4 (PAR)	0.003247	0.002152	1.509023	0.14002
	β_5 (NO3 + NO2)	0.012979	0.022131	0.586488	0.561207

Table 6.1: Weighted linear regression coefficient estimates with corresponding standard errors, t values, and p values, for both *Fragilariopsis* spp. and *Phaeocystis* spp. Weights were the inverse of the 95% credible intervals of the estimated RMFs across these two taxonomic groups.

6.5 Discussion

We connected quantitative estimates of the RMF with environmental variables across two taxa. This connection was possible because of a novel approach for calculating uncertainty in RMF from metaproteomic samples. For *Fragilariopsis* spp., we observed a negative relationship between RMF and temperature, and a negative relationship (albeit with a small effect size) between *Phaeocystis* spp. RMF and dMn. Below, we first discuss the strengths and weaknesses of metaproteomics as a tool for probing these relationships, then touch on the observed and expected relationships between trace metal concentration and light with RMF. Finally, we discuss the relationship between RMF and temperature. Note that throughout the following, we assume that at least one of the environmental variables we considered is controlling growth rate. This is a reasonable assumption because light, Fe, and Mn, have been identified as critical limiting resources in various regions around the Southern Ocean and in the Amundsen Sea Polynya (Sherrell et al., 2015; Wu et al., 2019; Kwon et al., 2021; Alderkamp et al., 2015; Oliver et al., 2019; Browning et al., 2021).

Metaproteomics is a promising method for probing gene expression of uncultured microbes in their natural environment, particularly those environments that are difficult or impossible to replicate in the lab. To date, one of the weaknesses with untargeted metaproteomics is that it has mostly been used to evaluate relative trends in gene expression. Untargeted metaproteomics has mostly failed to provide quantitative metrics of proteomic composition that can be easily compared with data from material or computational models (O'Malley and Parke, 2018). We have developed a simple method for estimating one quantitative metric, the proteomic mass fraction, alongside a measure of uncertainty. Comparing the performance of our method with artificially assembled metaproteomes demonstrated where it fails, specifically when considering fine-scale taxonomic resolution. This conclusion reflects our previous research, showing that biases in metaproteomics due to high sample complexity (McCain and Bertrand, 2019) and low diversity (McCain, Allen and Bertrand, 2021) tend to be mitigated by looking at coarse taxonomic or functional groupings. Future work should include the taxonomically uninformative peptides (e.g. Pible et al., 2020), which provide more information but are more challenging to incorporate.

Observations correlating growth and ribosomes per cell motivated the development of phenomenological growth laws in bacteria (Schaechter, Maaløe and Kjeldgaard, 1958; Scott et al., 2010). Here we expand the set of available observations to photosynthetic

marine microbes. In all microbes, growth and total protein synthesis are intimately related. To increase total protein synthesis, organisms can increase the number of total ribosomes, the rate of translation per ribosome, decrease the rate of degradation or protein damage, or increase the proportion of total ribosomes that are actively translating. In general, dFe, dMn, and light (which we expect to control growth rates here) had little explanatory value for the RMF in our samples, indicating that RMF and growth rate do not have a simple linear relationship as observed in some model bacteria (Schaechter, Maaløe and Kjeldgaard, 1958). Our previous modelling work for *Fragilariopsis cylindrus* mirrors these conclusions, particularly because protein turnover is dependent on each of these environmental variables (McCain et al., 2021). This model actually predicts almost no correlation between RMF and growth rate, when considering that dFe and dMn are typically correlated in the Southern Ocean. These environmental variables had little explanatory value for *Phaeocystis* spp. as well. Re-examining previously published *Phaeocystis antarctica* proteomic data across controlled dMn, dFe, and light levels, growth rate and RMF were only weakly correlated (0.39 Pearson correlation coefficient; Wu et al., 2019). *Phaeocystis antarctica* forms large colonies of many individual cells, and perhaps this unique life history complicates the connection between RMF and growth rate. Furthermore, the RMF range here is much lower than previously observed (McCain, Allen and Bertrand, 2021), perhaps because we mostly excluded the colonies by examining only the 3 μm filters. (Note that previously published *Phaeocystis antartica* proteomic data show divergent trends in ribosomal protein expression with increased iron, across two different strains (Supplementary Figure 6.7; Bender et al., 2017).) Overall, variation in RMF was not explained by light, dFe, or dMn: why was there no relationship?

Returning to the four potential connections of growth rate and total protein synthesis rate, we hypothesize that either 1) a variable proportion of ribosomes are actively translating, or 2) the rate of protein degradation is variable across the range of environmental conditions we examined. If either of these are a function of the environmental variables, we would not expect a simple correlation between RMF and growth rate. The first hypothesis suggests that phytoplankton cells have an excess capacity for protein synthesis, and would be consistent with other phytoplankton behaviours like 'luxury uptake' of Fe (Twining et al., 2020). In very slowly growing cultures of *E. coli*, a high proportion (even up to 80%) of ribosomes are inactive (Dai et al., 2016). However, it is currently unknown how

the proportion of translating ribosomes varies under Fe-controlled growth (even in highly studied model organisms). For the second hypothesis regarding protein degradation, it is reasonable to assume that depletion of Fe and Mn, which are critical antioxidant cofactors, might hamper a cell's capacity to avoid oxidative stress. Santra, Dill and Graff (2018) showed that oxidative stress and protein turnover are directly connected via chaperones. If the environmental variable that controls growth also controls protein degradation and turnover, we would not predict a simple relationship between growth and RMF. Overall, these data suggest that the bacterial growth laws exhibited in model organisms do not clearly translate to eukaryotic microbes in the ocean limited by different resources.

Our observation of a negative relationship between RMF and temperature in *Fragilariopsis* spp. suggests a variable rate of translation per ribosome. Translation rate per ribosome increases under increased temperature (Toseland et al., 2013). Others have observed a decrease in total RNA (most of which is ribosomal RNA) with temperature (Toseland et al., 2013), and a decrease in ribosomal protein mRNA with increased temperatures (Jabre et al., 2021), which both directly mirror our observations. Toseland et al. (2013) looked over a 12 °C temperature range and considered three unique temperature treatments, so the exact shape of the relationship between ribosomes and temperature is unclear. Our results suggest that the majority of the change in RMF with temperature occurs between -2 and 0 °C, consistent with a non-linear relationship between RMF and temperature for *Fragilariopsis* spp.

Does increased translation rate, as a function of temperature, explain the increased growth rate that is observed in cultures (Jabre and Bertrand, 2020)? Increasing temperature from 1 to 3 °C, Jabre and Bertrand (2020) observed approximately a doubling of growth rate under low Fe conditions, we therefore assume doubling of growth rate from -2 to 0 °C. Yet, we observed a decrease from $\sim$ 15% to $\sim$ 5% RMF, a decrease of $\sim$ 66%. Using a previously described temperature dependence with a Q10 function (Toseland et al., 2013), the change in translation rate per ribosome would increase only by 15% with this change in temperature. This leaves a paradox, how would protein synthesis capacity increase to support faster growth? We speculate that there is a temperature dependence on the proportion of ribosomes that are actively translating, such that under low temperatures, this proportion is lower. As discussed above, Dai et al. (2016) show that up to 80% of ribosomes can be inactive under low growth rates in *E. coli* (under

carbon substrate-controlled growth). This order-of-magnitude change in active ribosomes would be sufficient to explain the growth rate increase with temperature in *Fragilariopsis cylindrus*.

Coupling metaproteomics with environmental variables offers a natural laboratory to examine gene expression. Here we have used this laboratory, located in the Amundsen Sea Polynya, to study the regulation of ribosomes in photosynthetic eukaryotes. Our results suggest a non-linear relationship between RMF and temperature for *Fragilariopsis* spp., which was only possible to quantify because of the gradient of *in situ* temperatures we observed. Future metaproteomic characterizations will help uncover the diversity of protein synthesis-growth rate relationships in microbes.

6.6 Author Contributions

J.S.P.M. collected samples for metaproteomics, conducted laboratory processing and mass spectrometry with E.R. (Elden Rowland), ran all bioinformatics and analyses, and wrote the manuscript with input from E.M.B.

6.7 Supplementary Figures

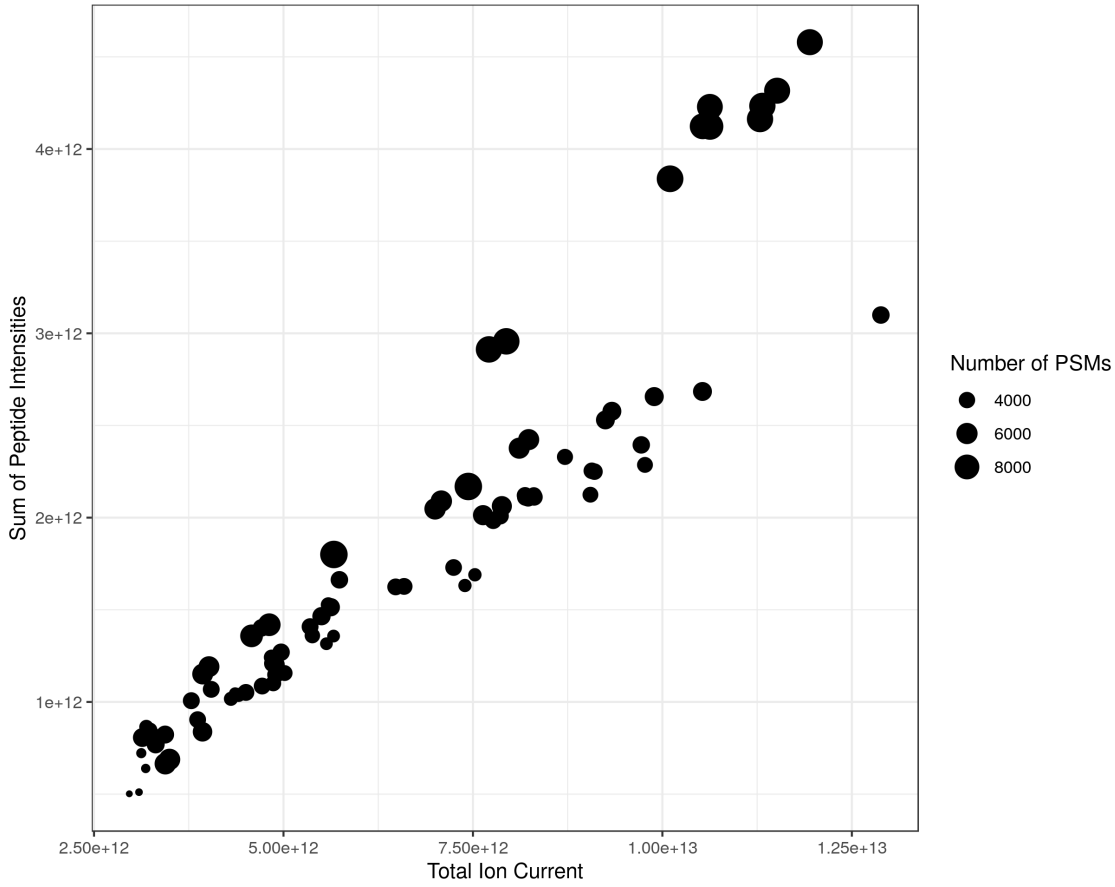

Figure 6.6: High correlation between the total ion current and the sum of peptide intensities per sample indicates that the database choice did not significantly impact peptide quantification.

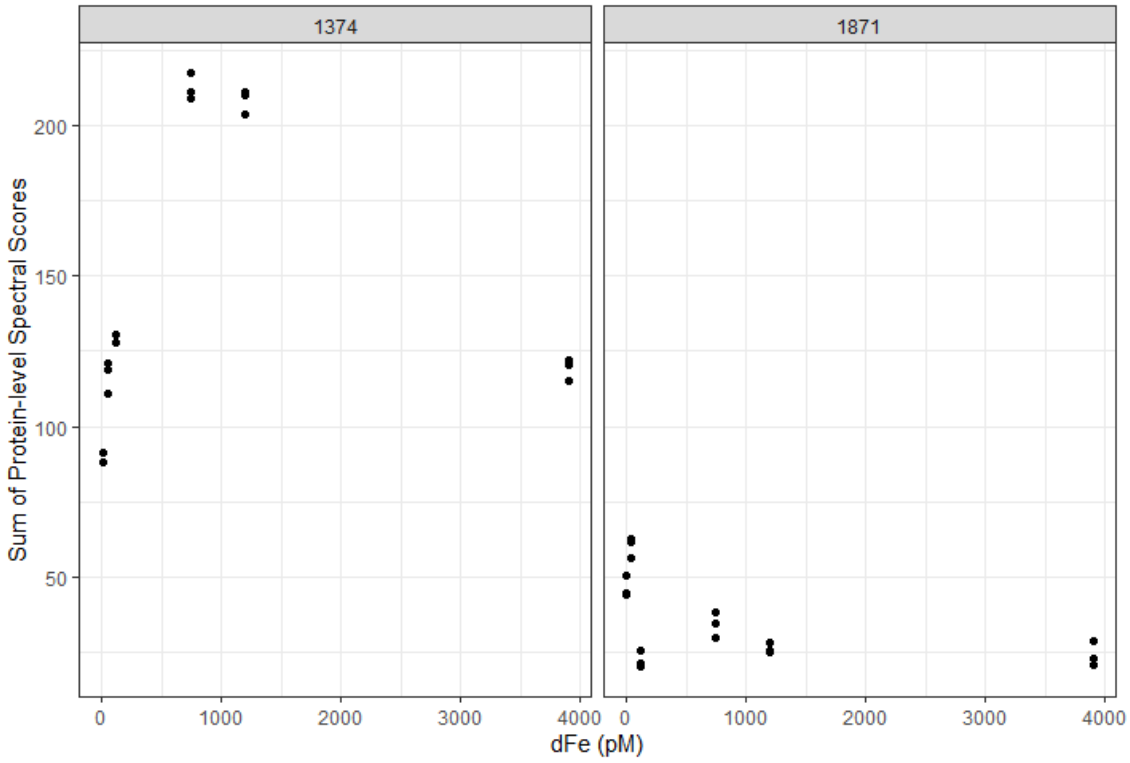

Figure 6.7: Divergent relationships between ribosomes and dissolved iron in two different cultured strains (left and right panels refer to strain numbers) of *Phaeocystis antartica* (Bender et al., 2017).

CHAPTER 7

CONCLUSIONS

7.1 Overview of Contributions

In this thesis I have used metaproteomics and models to examine trade-offs in phytoplankton. While there are five distinct contributions, these can be subcategorized into three themes. The first theme is putting metaproteomics as a tool 'under the microscope', the second theme is using observations and models to explore and test hypotheses about proteomic composition and trade-offs with respect to trace metals. The third theme is interrogating these proposed hypotheses using alternative methods, additional data, and synthesis. Here I will connect these three themes.

There are many basic assumptions in metaproteomics that have not been thoroughly examined. Throughout this thesis I aimed to explicitly address these assumptions. In Chapter 2, I studied a byproduct of high sample complexity, cofragmentation, and explored the theoretical implications of cofragmentation bias. I also developed a computational tool to predict this bias for a given peptide. In Chapter 3, I used simulations to study the effect of sequence diversity in peptide-centric metaproteomics, and evaluated how database configuration can influence both identification and quantification in metaproteomics. In Chapter 6, I used a simple Approximate Bayesian Computation method to estimate the proteomic mass fraction of ribosomes (with uncertainties).

The second theme collapses both Chapters 3 and 4 together. In Chapter 3, I quantified various proteomic 'traits', one of which was the ribosomal mass fraction, hypothesizing that this was related to the haptophyte-to-diatom transition frequently observed in the Ross Sea. In Chapter 4, I developed a proteomic allocation model to study the interaction between Fe and Mn. This lead to the discovery of general mechanisms of interdependence across

resources, a novel method for inferring taxon-specific rates by coupling metaproteomic data with a cellular model, and a reframing of micronutrient-controlled growth in the ocean.

These two chapters were opportunities to put forward quantitative, concrete hypotheses, which were subsequently further interrogated. One of the strengths of mathematical modelling is that it encourages assumptions to be made explicit. For example, I was forced to write a mathematical relationship for the fitness costs associated with producing antioxidants. In developing these hypotheses and mathematical relationships, I also identified unknowns. Specifically, the fitness costs of antioxidants and their impacts on cellular elemental stoichiometry were unclear. Furthermore, the assumptions regarding ribosomal use in the proteomic allocation model had not yet been validated under trace metal-controlled growth.

In the last theme I investigate some of the assumptions and hypotheses put forward, specifically the 1) relationship between antioxidants and cellular elemental stoichiometry (Chapter 5) and 2) the relationship between ribosomes and growth (Chapter 6). Chapter 5 revealed many nuances of antioxidants, mainly stemming from the diversity of antioxidant systems. This synthesis identified several unknowns that should be tested in a material model to ultimately quantify both the fitness costs of antioxidants and their contributions to cellular stoichiometry (O'Malley and Parke, 2018). With this synthesis I also put forward quantitative bounds for the contribution of various metal-containing antioxidants using previously published data. In Chapter 6, I provided evidence that the assumption of 100% ribosomal use is not supported under Fe-, Mn-, and light-controlled growth. Data from Chapter 6 also refuted my earlier hypothesis about trait differences between diatoms and haptophytes (however the differences in filter sizes complicates this comparison). While the discoveries in Chapters 5 and 6 highlight additional complexities, they also provide a way forward.

7.2 Future Directions

7.2.1 Metaproteomics

Metaproteomics as a tool to examine *in situ* gene expression is still in its infancy. I believe the next frontier of metaproteomics will be rooted in two advances, one technological and one conceptual. From the technological frontier, there have been major advances in mass

spectrometry, particularly using data-independent acquisition (DIA) strategies. Once these approaches are more accessible for metaproteomics, I think the amount of information gleaned per sample will inevitably increase. Another advantage for DIA experiments is that the data can effectively be recycled with better algorithms. Also note that there are many 'disruptive' proteomic technologies on the horizon that threaten mass spectrometry as the premiere tool for proteomics (Timp and Timp, 2020).

Conceptual advances in metaproteomics will come when methods and measurements are centered on absolute quantification. Here I use the term 'absolute' to refer to a measurement that is 'viewed or existing independently and not in relation to other things'. For example, in Chapter 6 I attempted to infer the proteomic mass fraction of ribosomes, independent of organismal abundance. This type of quantification is embraced in targeted proteomics, however in a metaproteomic context 'absolute' quantification can still be a non-trivial endeavor. The conceptual shift towards absolute quantification will enable seamless comparisons with cultured organisms and computational models. (Note there is no consensus on which values should be quantified).

Mitigating against biases in metaproteomics is critical to move closer to absolute quantification, as illustrated throughout this thesis. Most of these biases can be simplified with the statement: 'the devil is in the denominator'. For example, the problems associated with normalization and quantification from database dependence (Chapter 3), or the issues in assessing proteomic mass fraction (Chapter 6), all stem from variable denominators. Datasets like those of Kleiner et al. (2017) will be increasingly valuable for the field.

7.2.2 Representations of Growth in the Ocean

Most biogeochemical models include a function with inputs of environmental variables and an output of growth (or something similar). These simplistic representations of biology in biogeochemical models can lead to some challenges (discussed in Chapter 1). Using models that predict both gene expression profiles alongside growth rate would achieve more realistic biological behaviours. Further, this type of model would enable direct comparisons with the large volumes of metaproteomic and metatranscriptomic data that are being generated.

Should proteomic allocation models be embedded in biogeochemical models? Not necessarily. In Chapter 5, I aimed to test specific hypotheses about the interaction between Fe and Mn. Furthermore, it would not be computationally tractable to put this cellular

model into a large oceanographic model. If the aim was to directly connect cellular and biogeochemical modelling, it would only be necessary to represent the phenomenological outcomes of cellular models. For example, a high-order polynomial regression could be used to represent the predictions of a proteomic allocation model, and distill these predictions into a simple equation. Representations of growth in the ocean have not changed much since the seminal work of Riley (1946). At the same time, quantitative systems biology of microbes is burgeoning. It is time for oceanographers to broaden the toolkit of biological representations.

Another prominent use of metaproteomics is diagnosing nutrient stress using biomarkers (e.g. Saito et al., 2014). One advantage of using biomarkers over traditional bottle incubation experiments is that it can be done with high-throughput methods. Yet, one open challenge is: what do we do with this information? One goal might be to use biomarkers to inform large-scale models, thus providing data to constrain the biological representations in biogeochemical models (Fennel et al., 2019). However, to make this connection, there needs to be some representation in biogeochemical models that link biomarker expression to growth rate. Cellular models that represent gene expression would enable this connection.

7.2.3 Of Models and Metaproteomes: A Proposal

To borrow from John Steinbeck's 'Of Mice and Men', I will briefly propose a research program building from ideas presented in this thesis (touching on the theme of dreams throughout Steinbeck's book).

Computational models can be used to represent biological phenomena. Connecting computational modelling with *in situ* gene expression measurements from complex microbial communities provides a path forward to studying organisms that are difficult to culture. But, as was obvious from this thesis, this can lead to clear questions that should be interrogated in another way.

Material modelling (O'Malley and Parke, 2018) is an appropriate tool for asking certain questions: for example, what is the limit on membrane protein density? How does the proportion of actively translating ribosomes vary under iron-controlled growth? Are MnSOD, FeSOD, CuZnSOD, and NiSOD functionally replaceable? Material models have built-in constraints, which can help answer these questions. Yet, it is difficult to imagine asking some of these questions using an organism like *Fragilariopsis cylindrus*

(at least considering the doubling time, but also the genetic tractability; although the latter is being challenged with new methods; Faktorová et al., 2020). I propose to use genetically tractable model organisms, like *E. coli*, to examine specific biological processes (like the ones listed above).

This trifecta of approaches – mathematical modelling, metaproteomics, and material modelling – is apt for studying environmentally important microbes. Mathematical modelling provides a structured approach for connecting hypotheses and integrating processes. These models can uniquely capture characteristics that material models cannot. Coupling mathematical models with *in situ* gene expression measurements, like metaproteomics, can be used to infer taxon-specific rates and quantify certain facets of diverse microbes. Finally, material models can promote discovery of unknown processes, and can be used to examine certain processes when knowledge of a system is very limited. Iteration between these three approaches will help uncover the hidden lives of microbes.

APPENDIX A

Chapter 2 was reprinted with permission from *J. Proteome Res.* 2019, 18, 10, 3555–3566. Copyright 2019 American Chemical Society.

Chapter 3 includes material from: McCain, J.S.P., Allen, A.E., and E.M. Bertrand. Proteomic traits vary across taxa in a coastal Antarctic phytoplankton bloom. *The ISME Journal*. Published 2021.

BIBLIOGRAPHY

Achterberg EP, Holland TW, Bowie AR, Mantoura RFC, Worsfold PJ. 2001. Determination of iron in seawater. Analytica Chimica Acta. 442:1–14.

Aebi H. 1984. Catalase in Vitro. Methods in Enzymology. 105:121–126.

Aguirre JD, Culotta VC. 2012. Battles with iron: Manganese in oxidative stress protection. Journal of Biological Chemistry. 287:13541–13548.

Aksnes DL, Cao FJ. 2011. Inherent and apparent traits in microbial nutrient uptake. Marine Ecology Progress Series. 440:41–51.

Aksnes DL, Egge JK. 1991. A theoretical model for nutrient uptake in phytoplankton. Marine Ecology Progress Series. 70:65–72.

Alahmadi AA, Flegg JA, Cochrane DG, Drovandi CC, Keith JM. 2020. A comparison of approximate versus exact techniques for Bayesian parameter inference in nonlinear ordinary differential equation models. Royal Society Open Science. 7:191315.

Alderkamp AC, van Dijken GL, Lowry KE, et al. (12 co-authors). 2015. Fe availability drives phytoplankton photosynthesis rates during spring bloom in the Amundsen Sea Polynya, Antarctica. Elementa: Science of the Anthropocene. 3:000043.

Alexander H, Jenkins BD, Rynearson TA, Dyhrman ST. 2015a. Metatranscriptome analyses indicate resource partitioning between diatoms in the field. Proceedings of the National Academy of Sciences. 112:E2182–E2190.

Alexander H, Rouco M, Haley ST, Wilson ST, Karl DM, Dyhrman ST. 2015b. Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. Proceedings of the National Academy of Sciences. 112:E5972–E5979.

Algar CK, Vallino JJ. 2014. Predicting microbial nitrate reduction pathways in coastal sediments. Aquatic Microbial Ecology. 71:223–238.

Anand A, Chen K, Yang L, et al. (13 co-authors). 2019. Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. Proceedings of the National Academy of Sciences. 116:25287–25292.

Andreoli C, Tolomio C, Moro I, Radice M, Moschin E, Bellato S. 1995. Diatoms and dinoflagellates in Terra Nova Bay (Ross Sea-Antarctica) during austral summer 1990. Polar Biology. 15:465–475.

Anjem A, Imlay JA. 2012. Mononuclear iron enzymes are primary targets of hydrogen peroxide stress. Journal of Biological Chemistry. 287:15544–15556.

Anselm V, Novikova S, Zgoda V. 2017. Re-adaption on earth after spaceflights affects the mouse liver proteome. International Journal of Molecular Sciences. 18:1–12.

Asada K. 2006. Production and scavenging of reactive oxygen species in chloroplasts and their functions. Plant Physiology. 141:391–396.

Assmy P, Smetacek V, Montresor M, Klaas C, Henjes J, Strass VH. 2013. Thick-shelled, grazer protected diatoms decouple ocean carbon and silicon cycles in the iron-limited Antarctic Circumpolar Current. Proceedings of the National Academy of Sciences. 110:20633–20638.

Aumont O, Maier-Reimer E, Blain S, Monfray P. 2003. An ecosystem model of the global ocean including Fe, Si, P colimitations. Global Biogeochemical Cycles. 17:1–15.

Aylward FO, Burnum KE, Scott JJ, et al. (15 co-authors). 2012. Metagenomic and metaproteomic insights into bacterial communities in leaf-cutter ant fungus gardens. The ISME Journal. 6:1688–1701.

Bacastow R, Maier-Reimer E. 1991. Dissolved organic carbon in modeling oceanic new production. Global Biogeochemical Cycles. 5:71–85.

Barnese K, Gralla EB, Cabelli DE, Valentine JS. 2008. Manganous phosphate acts as a superoxide dismutase. Journal of the American Chemical Society. 130:4604–4606.

Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T. 2015. Overflow metabolism in Escherichia coli results from efficient proteome allocation. Nature. 528:99–104.

Behrenfeld MJ, Boss E, Siegel DA, Shea DM. 2005. Carbon-based ocean productivity and phytoplankton physiology from space. Global Biogeochemical Cycles. 19:1–14.

Bender SJ, Moran DM, Mcilvin MR, Zheng H, Mccrow JP, Badger J, Ditullio GR, Allen AE, Saito MA. 2017. Iron triggers colony formation in Phaeocystis antarctica: connecting molecular mechanisms with iron biogeochemistry. Biogeosciences Discussions. 5194:2017–558.

Berg HC, Purcell EM. 1977. Physics of chemoreception. Biophysical Journal. 20:193–219.

Bergauer K, Fernández-Guerra A, Garcia JA, Sprenger RR, Stepanauskas R, Pachiadaki MG, Jensen ON, Herndl GJ. 2017. Organic matter processing by microbial communities throughout the Atlantic water column as revealed by metaproteomics. Proceedings of the National Academy of Sciences. 115:E400–E408.

Bertrand EM, Allen AE, Dupont CL, Norden-Krichmar TM, Bai J, Valas RE. 2012. Influence of cobalamin scarcity on diatom molecular physiology and identification of a cobalamin acquisition protein. Proceedings of the National Academy of Sciences. 109:1762–1771.

Bertrand EM, McCrow JP, Moustafa A, et al. (13 co-authors). 2015. Phytoplankton-bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge. Proceedings of the National Academy of Sciences. 112:9938–9943.

Bertrand EM, Moran DM, McIlvin MR, Hoffman JM, Allen AE, Saito MA. 2013. Methionine synthase interreplacement in diatom cultures and communities: Implications for the persistence of B12 use by eukaryotic phytoplankton. Limnology and Oceanography. 58:1431–1450.

Bielow C, Aiche S, Andreotti S, Reinert K. 2011. MSSimulator: Simulation of Mass Spectrometry Data. Journal of Proteome Research. 10:2922–2929.

Blaby-Haas CE, Merchant SS. 2017. Regulating cellular trace metal economy in algae. Current Opinion in Plant Biology. 39:88–96.

Boer J, Mulrooney S, Hausinger R. 2014. Nickel-dependent metalloenzymes. Archives of Biochemistry and Biophysics. 15:142–152.

Bonachela JA, Allison SD, Martiny AC, Levin SA. 2013. A model for variable phytoplankton stoichiometry based on cell protein regulation. Biogeosciences. 10:4341–4356.

Bonachela JA, Raghib M, Levin SA. 2011. Dynamic model of flexible phytoplankton nutrient uptake. Proceedings of the National Academy of Sciences. 108:20633–8.

Borges PT, Frazão C, Miranda CS, Carrondo MA, Romão CV. 2014. Structure of the monofunctional heme catalase DR1998 from Deinococcus radiodurans. The FEBS Journal. 281:4138–4150.

Bowler C, Allen AE, Badger JH, et al. (77 co-authors). 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 456:239–244.

Braakman R, Follows MJ, Chisholm SW. 2017. Metabolic evolution and the self-organization of ecosystems. Proceedings of the National Academy of Sciences. 114:E3091–E3100.

Broberg M, Doonan J, Mundt F, Denman S, McDonald JE. 2018. Integrated multi-omic analysis of hostmicrobiota interactions in acute oak decline. Microbiome. 6:1–15.

Brown JD, Day AM, Taylor SR, Tomalin LE, Morgan BA, Veal EA. 2013. A peroxiredoxin promotes H2O2 signaling and oxidative stress resistance by oxidizing a thioredoxin family protein. Cell Reports. 5:1425–1435.

Browning TJ, Achterberg EP, Engel A, Mawji E. 2021. Manganese co-limitation of phytoplankton growth and major nutrient drawdown in the Southern Ocean. Nature Communications. 12:1–9.

Browning TJ, Achterberg EP, Rapp I, Engel A, Bertrand EM, Tagliabue A, Moore CM. 2017. Nutrient co-limitation at the boundary of an oceanic gyre. Nature. 551:242–246.

Browning TJ, Bouman HA, Henderson GM, Mather TA, Pyle DM, Schlosser C, Woodward EMS, Moore CM. 2014. Strong responses of Southern Ocean phytoplankton communities to volcanic ash. Geophysical Research Letters. 41:2851–2857.

Buma AGJ, Baar HJWD, Nolting RF, Bennekom AJV. 1991. Metal enrichment experiments in the Weddell-Scotia Seas: Effects of iron and manganese on various plankton communities. Limnology and Oceanography. 36:1865–1878.

Button DK. 1998. Nutrient uptake by microorganisms according to kinetic parameters from theory as related to cytoarchitecture. Microbiology and Molecular Biology Reviews. 62:636–645.

Cao JX, Teoh ML, Moon M, McFadden G, Evans DH. 2002. Leporipoxvirus Cu-Zn superoxide dismutase homologs inhibit cellular superoxide dismutase, but are not essential for virus replication or virulence. Virology. 296:125–135.

Carrara F, Sengupta A, Behrendt L, Vardi A, Stocker R. 2021. Bistability in oxidative stress response determines the migration behavior of phytoplankton in turbulence. Proceedings of the National Academy of Sciences. 118:e2005944118.

Case AJ. 2017. On the origin of superoxide dismutase: An evolutionary perspective of superoxide-mediated redox signaling. Antioxidants. 6:1–21.

Casey JR, Follows MJ. 2020. A steady-state model of microbial acclimation to substrate limitation. PLoS Computational Biology. 16:1–17.

Chapman JD, Goodlett DR, Masselon CD. 2014. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spectrometry Reviews. 33:452–470.

Cheton PL, Archibald FS. 1988. Manganese complexes and the generation and scavenging of hydroxyl free radicals. Free Radical Biology and Medicine. 5:325–333.

Coale TH, Moosburner M, Horák A, Oborník M, Barbeau KA, Allen AE. 2019. Reduction-dependent siderophore assimilation in a model pennate diatom. Proceedings of the National Academy of Sciences. 116:23609–23617.

Cohen NR, Gong W, Moran DM, McIlvin MR, Saito MA, Marchetti A. 2018. Transcriptomic and proteomic responses of the oceanic diatom Pseudo-nitzschia granii to iron limitation. Environmental Microbiology. 20:3109–3126.

Cohen NR, McIlvin MR, Moran DM, et al. (13 co-authors). 2021. Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. Nature Microbiology. 6:173–186.

Coles VJ, Stukel MR, Brooks MT, et al. (11 co-authors). 2017. Ocean biogeochemistry modeled with emergent trait-based genomics. Science. 1154:1–26.

Cooper WJ, Saltzman ES, Zika RG. 1987. The contribution of rainwater to variability in surface ocean hydrogen peroxide. Journal of Geophysical Research: Oceans. 92:2970–2980.

Dahl JU, Gray MJ, Jakob U. 2015. Protein quality control under oxidative stress conditions. Journal of Molecular Biology. 427:1549–1563.

Dai X, Zhu M, Warren M, Balakrishnan R, Patsalo V, Okano H, Williamson JR, Fredrick K, Wang YP, Hwa T. 2016. Reduction of translating ribosomes enables Escherichia coli to maintain elongation rates during slow growth. Nature Microbiology. 2:1–9.

Davalieva K, Kiprijanovska S, Maleva Kostovska I, Stavridis S, Stankov O, Komina S, Petrusevska G, Polenakovic M. 2018. Comparative Proteomics Analysis of Urine Reveals Down-Regulation of Acute Phase Response Signaling and LXR/RXR Activation Pathways in Prostate Cancer. Proteomes. 6:1.

Davies AM, Holt AG. 2018. Why antioxidant therapies have failed in clinical trials. Journal of Theoretical Biology. 457:1–5.

De Baar HJ, Timmermans KR, Laan P, et al. (11 co-authors). 2008. Titan: A new facility for ultraclean sampling of trace elements and isotopes in the deep oceans in the international Geotraces program. Marine Chemistry. 111:4–21.

de Boyer Montegut C, Madec G, Fischer AS, Lazar A, Iudicone D. 2004. Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. Journal of Geophysical Research. 109:1–20.

Dekel E, Alon U. 2005. Optimality and evolutionary tuning of the expression level of a protein. Nature. 436:588–592.

Delmont TO, Gaia M, Hinsinger DD, et al. (25 co-authors). 2021. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. bioRxiv. p. 10.15.341214.

Deppeler SL, Davidson AT. 2017. Southern Ocean phytoplankton in a changing climate. Frontiers in Marine Science. 4:1–28.

Dethlefsen L, Schmidt TM. 2007. Performance of the translational apparatus varies with the ecological strategies of bacteria. Journal of Bacteriology. 189:3237–3245.

Diaz JM, Hansel C, Voelker B, Mendes C, Andeer P, Zhang T. 2014. Widespread production of extracellular superoxide by heterotrophic bacteria. Science. 1223.

Dorfer V, Maltsev S, Winkler S, Mechtler K. 2018. CharmeRT: Boosting peptide identifications by chimeric spectra identification and retention time prediction. Journal of Proteome Research. 17:2581–2589.

Droop MR. 1974. The nutrient status of algal cells in continuous culture. Journal of the Marine Biological Association of the UK. 9:825–855.

Dupont CL, Goepfert TJ, Lo P, Wei L, Ahner BA. 2004. Diurnal cycling of glutathione in marine phytoplankton: Field and culture studies. Limnology and Oceanography. 49:991–996.

Dupont CL, McCrow JP, Valas R, et al. (15 co-authors). 2015. Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. The ISME Journal. 9:1076–1092.

Edgar RS, Green EW, Zhao Y, et al. (19 co-authors). 2012. Peroxiredoxins are conserved markers of circadian rhythms. Nature. 485:459–464.

Elser JJ, Dobberfuhl DR, MacKay NA, Schampel JH. 1996. Organism size, life history and N:P stoichiometry. BioScience. 46:674–685.

Elser JJ, Sterner RW, Gorokhova E, Fagan WF, Markow TA, Cotner JB, Harrison JF, Hobbie SE, Odell GM, Weider LW. 2000. Biological stoichiometry from genes to ecosystems. Ecology Letters. 3:540–550.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. 30:1575–1584.

Eyers CE, Lawless C, Wedge DC, Lau KW, Gaskell SJ, Hubbard SJ. 2011. CONSeQuence: Prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. Molecular and Cellular Proteomics. 10:1–12.

Faizi M, Zavřel T, Loureiro C, Červený J, Steuer R. 2018. A model of optimal protein allocation during phototrophic growth. BioSystems. 166:26–36.

Faktorová D, Nisbet RER, Fernández Robledo JA, et al. (113 co-authors). 2020. Genetic tool development in marine protists: emerging model organisms for experimental cell biology. Nature Methods. 17:481–494.

Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive earth's biogeochemical cycles. Science. 320:1034–1039.

Fearnhead P, Prangle D. 2012. Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. Journal of the Royal Statistical Society. Series B: Statistical Methodology. 74:419–474.

Feierabend J, Germany W. 1986. Photoinactivation of catalase in vitro and in leaves. Archives of Biochemistry and Biophysics. 251:567–576.

Fennel K, Gehlen M, Brasseur P, et al. (16 co-authors). 2019. Advancing marine biogeochemical and ecosystem reanalyses and forecasts as tools for monitoring and managing ecosystem health. Frontiers in Marine Science. 6:1–9.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science. 281:237–240.

Fiksen Ø, Follows MJ, Aksnes DL. 2013. Trait-based models of nutrient uptake in microbes extend the Michaelis-Menten framework. Limnology and Oceanography. 58:193–202.

Finkel ZV, Follows MJ, Irwin AJ. 2016. Size-scaling of macromolecules and chemical energy content in the eukaryotic microalgae. Journal of Plankton Research. 38:1151–1162.

Follows MJ, Dutkiewicz S, Grant S, Chisholm SW. 2007. Emergent biogeography of microbial communities in a model ocean. Science. 315:1843–1847.

Fomenko DE, Koc A, Agisheva N, et al. (11 co-authors). 2011. Thiol peroxidases mediate specific genome-wide regulation of gene expression in response to hydrogen peroxide. Proceedings of the National Academy of Sciences. 108:2729–2734.

Galbraith ED, Kienast M, Albuquerque AL, et al. (40 co-authors). 2013. The acceleration of oceanic denitrification during deglacial warming. Nature Geoscience. 6:579–584.

Gallie D, Chen Z. 2019. Chloroplast-localized iron superoxide dismutases FSD2 and FSD3 are functionally distinct in Arabidopsis. PLoS ONE. 14:e0220078.

Geider RJ, LaRoche J. 2002. Redfield revisited: Variability of C:N:P in marine microalgae and its biochemical basis. European Journal of Phycology. 37:1–17.

Gifford SM, Sharma S, Booth M, Moran MA. 2013. Expression patterns reveal niche diversification in a marine microbial assemblage. The ISME Journal. 7:281–298.

Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, Bonner R, Aebersold R. 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. Molecular and Cellular Proteomics. 11:O111.016717–O111.016717.

Giovannoni SJ. 2017. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. Annual Review of Marine Science. 9:231–255.

Giovannoni SJ, Cameron Thrash J, Temperton B. 2014. Implications of streamlining theory for microbial ecology. The ISME Journal. 8:1553–1565.

Gledhill M, Buck KN. 2012. The organic complexation of iron in the marine environment: a review. Frontiers in Microbiology. 3:1–17.

Goldfarb D, Wang W, Major MB. 2016. MSAcquisitionSimulator: Data-dependent acquisition simulator for LC-MS shotgun proteomics. Bioinformatics. 32:1269–1271.

Goloborodko AA, Levitsky LI, Ivanov MV, Gorshkov MV. 2013. Pyteomics — a Python framework for exploratory data analysis and rapid software prototyping in proteomics. Journal for the American Society of Mass Spectrometry. 24:301–304.

Gorshkov AV, Tarasova IA, Evreinov VV, Savitski MM, Nielsen ML, Zubarev RA, Gorshkov MV. 2006. Liquid chromatography at critical conditions: comprehensive approach to sequence-dependent retention time prediction. Analytical Chemistry. 78:7770–7777.

Gorshkov V, Hotta SYK, Verano-Braga T, Kjeldsen F. 2016. Peptide de novo sequencing of mixture tandem mass. Proteomics. 16:2470–2479.

Graff van Creveld S, Rosenwasser S, Levin Y, Vardi A. 2016. Chronic iron limitation confers transient resistance to oxidative stress in marine diatoms. Plant Physiology. 172:968–979.

Gu M, Imlay JA. 2013. Superoxide poisons mononuclear iron enzymes by causing mismetallation. Molecular Microbiology. 89:123–134.

Haas S, Robicheau BM, Rakshit S, Tolman J, Algar CK, LaRoche J, Wallace DW. 2021. Physical mixing in coastal waters controls and decouples nitrification via biomass dilution. Proceedings of the National Academy of Sciences. 118:e2004877118.

Haber F, Weiss J. 1932. The catalytic decomposition of hydrogen peroxide by iron salts. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 147:332–351.

Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Research. 31:371–373.

Halliwell B, Gutteridge JM. 2007. Free Radicals in Biology and Medicine. Oxford University Press, fourth edi edition.

Han BP. 2001. Photosynthesis-irradiance response at physiological level: a mechanistic model. Journal of Theoretical Biology. 213:121–127.

Hardison AK, Algar CK, Giblin AE, Rich JJ. 2015. Influence of organic carbon and nitrate loading on partitioning between dissimilatory nitrate reduction to ammonium (DNRA) and N2 production. Geochimica et Cosmochimica Acta. 164:146–160.

Harpole WS, Ngai JT, Cleland EE, et al. (11 co-authors). 2011. Nutrient co-limitation of primary producer communities. Ecology Letters. 14:852–862.

Hart Y, Sheftel H, Hausser J, Szekely P, Ben-Moshe NB, Korem Y, Tendler A, Mayo AE, Alon U. 2015. Inferring biological tasks using Pareto analysis of high-dimensional data. Nature Methods. 12:233–235.

Held NA, Webb EA, McIlvin MM, et al. (11 co-authors). 2020. Co-occurrence of Fe and P stress in natural populations of the marine diazotroph Trichodesmium. Biogeosciences. 17:2537–2551.

Heyer R, Schallert K, Zoun R, Becher B, Saake G, Benndorf D. 2017. Challenges and perspectives of metaproteomic data analysis. Journal of Biotechnology. pp. 24–36.

Hillebrand H, Donohue I, Harpole WS, Hodapp D, Kucera M, Lewandowska AM, Merder J, Montoya JM, Freund JA. 2020. Thresholds for ecological responses to global change do not emerge from empirical data. Nature Ecology and Evolution. 4:1502–1509.

Hindmarsh A. 1983. ODEPACK, a Systematized Collection of ODE Solvers in Scientific Computing. Elsevier.

Holton RW, Blecker HH, Onorb M. 1964. Effect of growth temperature on the fatty acid composition of a blue-green alga. Phytochemistry. 3:595–602.

Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn N, Old W. 2010. Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomic studies. Journal of Proteome Research. 9:4152–4160.

Hsieh EJ, Bereman MS, Durand S, Valaskovic GA, MacCoss MJ. 2012. Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. Journal of the American Society for Mass Spectrometry. 24:148–153.

Hu SK, Liu Z, Alexander H, Campbell V, Connell PE, Dyhrman ST, Heidelberg KB, Caron DA. 2018. Shifting metabolic priorities among key protistan taxa within and below the euphotic zone. Environmental Microbiology. 20:2865–2879.

Hudson RJ, Morel FM. 1990. lron transport in marine phytoplankton: Kinetics of cellular and medium coordination reactions. Limnology and Oceanography. 35:1002–1020.

Huerta-Cepas J, Szklarczyk D, Heller D, et al. (12 co-authors). 2019. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research. 47:D309–D314.

Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, Hwa T, Williamson JR. 2015. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. Molecular Systems Biology. 11:e784–e784.

Hulstaert N, Shofstahl J, Sachsenberg T, Walzer M, Barsnes H, Martens L, Perez-Riverol Y. 2020. ThermoRawFileParser: Modular, scalable, and cross-platform RAW file conversion. Journal of Proteome Research. 19:537–542.

Imlay JA. 2003. Pathways of oxidative damage. Annual Review of Microbiology. 57:395–418.

Imlay JA. 2013. The molecular mechanisms and physiological consequences of oxidative stress: Lessons from a model bacterium. Nature Reviews Microbiology. 11:443–454.

Imlay JA. 2014. The mismetallation of enzymes during oxidative stress. Journal of Biological Chemistry. 289:28121–28128.

Imlay JA. 2019. Where in the world do bacteria experience oxidative stress? Environmental Microbiology. 21:521–530.

Inomura K, Deutsch C, Wilson ST, et al. (11 co-authors). 2019. Quantifying oxygen management and temperature and light dependencies of nitrogen fixation by Crocosphaera watsonii. mSphere. 4:1–16.

Ishiyama K, Inoue E, Yamaya T, Takahashi H. 2006. Gln49 and Ser174 residues play critical roles in determining the catalytic efficiencies of plant glutamine synthetase. Plant and Cell Physiology. 47:299–303.

Jabre L, Bertrand EM. 2020. Interactive effects of iron and temperature on the growth of Fragilariopsis cylindrus. Limnology and Oceanography Letters. 5:363–370.

Jabre LJ, Allen AE, McCain JSP, et al. (11 co-authors). 2021. Molecular underpinnings and biogeochemical consequences of enhanced diatom growth in a warming Southern Ocean. Proceedings of the National Academy of Sciences. 118:1–9.

Jahn M, Vialas V, Karlsen J, Ka L, Uhle M, Hudson EP. 2018. Growth of cyanobacteria is constrained by the abundance of light and carbon assimilation proteins. Cell Reports. pp. 478–486.

Jamers A, Van der Ven K, Moens L, Robbens J, Potters G, Guisez Y, Blust R, De Coen W. 2006. Effect of copper exposure on gene expression profiles in Chlamydomonas reinhardtii based on microarray analysis. Aquatic Toxicology. 80:249–260.

Jang HH, Lee KO, Chi YH, et al. (18 co-authors). 2004. Two enzymes in one: Two yeast peroxiredoxins display oxidative stress-dependent switching from a peroxidase to a molecular chaperone function. Cell. 117:625–635.

Jeong HJ, du Yoo Y, Kim JS, Seong KA, Kang NS, Kim TH. 2010. Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. Ocean Science Journal. 45:65–91.

Johnson GE, Lalanne JB, Peters ML, Li GWW. 2020. Functionally uncoupled transcription–translation in Bacillus subtilis. Nature. 585:124–128.

Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, Küsel K, Rillig MC, Rivett DW. 2017. Where less may be more: how the rare biosphere pulls ecosystems strings. The ISME Journal. 11:853–862.

Kadner R. 1996. Cytoplasmic membrane. In Escherichia coli and Salmonella: Cellular and Molecular Biology, Washington, DC: The American Society for Microbiology. 2nd edition.

Kafri M, Metzl-Raz E, Jona G, Barkai N. 2016. The cost of protein production. Cell Reports. 14:22–31.

Kagaya S, Maeba E, Inoue Y, Kamichatani W, Kajiwara T, Yanai H, Saito M, Tohda K. 2009. A solid phase extraction using a chelate resin immobilizing carboxymethylated pentaethylenehexamine for separation and preconcentration of trace elements in water samples. Talanta. 79:146–152.

Kalli A, Smith GT, Sweredoski MJ, Hess S. 2013. Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: Focus on LTQ-Orbitrap mass analyzers. Journal of Proteome Research. 12:3071–3086.

Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 28:27–30.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. 44:D457–D462.

Karplus PA. 2015. A primer on peroxiredoxin biochemistry. Free Radical Biology and Medicine. 80:183–190.

Kashtan N, Roggensack SE, Rodrigue S, et al. (13 co-authors). 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 344:416–420.

Keith Moore J, Fu W, Primeau F, Britten GL, Lindsay K, Long M, Doney SC, Mahowald N, Hoffman F, Randerson JT. 2018. Sustained climate warming drives declining marine biological productivity. Science. 359:113–1143.

Kennedy F, Martin A, Bowman JP, Wilson R, McMinn A. 2019. Dark metabolism: a molecular insight into how the Antarctic sea-ice diatom Fragilariopsis cylindrus survives long-term darkness. New Phytologist. 223:675–691.

Khademian M, Imlay JA. 2017. Escherichia coli cytochrome c peroxidase is a respiratory oxidase that enables the use of hydrogen peroxide as a terminal electron acceptor. Proceedings of the National Academy of Sciences. 114:E6922–E6931.

Kim JG, Park SJ, Quan ZX, et al. (11 co-authors). 2014. Unveiling abundance and distribution of planktonic Bacteria and Archaea in a polynya in Amundsen Sea, Antarctica. Environmental microbiology. 16:1566–1578.

Kim S, Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nature Communications. 5:1–10.

Kleiner M, Thorson E, Sharp CE, Dong X, Liu D, Li C, Strous M. 2017. Assessing species biomass contributions in microbial communities via metaproteomics. Nature Communications. 8:1–14.

Koziol JA, Griffin NM, Long F, Li Y, Latterich M, Schnitzer JE. 2013. On protein abundance distributions in complex mixtures. Proteome Science. 11:1–9.

Krieger-Liszkay A, Trebst A. 2006. Tocopherol is the scavenger of singlet oxygen produced by the triplet states of chlorophyll in the PSII reaction centre. Journal of Experimental Botany. 57:1677–1684.

Kwon YS, La HS, Jung JY, Lee SH, Kim T, Kang H, Lee S. 2021. Exploring the roles of iron and irradiance in dynamics of diatoms and Phaeocystis in the Amundsen Sea continental shelf water. Journal of Geophysical Research: Oceans. 126:e2020JC016673.

Laing M. 1989. The three forms of molecular oxygen. Journal of Chemical Education. 66:453–455.

Lalanne J, Parker DJ, Li GW. 2021. Spurious regulatory connections dictate the expression-fitness landscape of translation factors. Molecular Systems Biology. 17:1–23.

Lalanne JB, Taggart JC, Guo MS, Herzel L, Schieler A, Li GW. 2018. Evolutionary convergence of pathway-specific enzyme expression stoichiometry. Cell. 173:749–761.e38.

Laman Trip DS, Youk H. 2020. Yeasts collectively extend the limits of habitable temperatures by secreting glutathione. Nature Microbiology. 5:943–954.

Lambeck I, Chi JC, Krizowski S, Mueller S, Mehlmer N, Teige M, Fischer K, Schwarz G. 2010. Kinetic analysis of 14-3-3-inhibited Arabidopsis thaliana nitrate reductase. Biochemistry. 49:8177–8186.

Laufkötter C, Vogt M, Gruber N, et al. (20 co-authors). 2015. Drivers and uncertainties of future global marine primary production in marine ecosystem models. Biogeosciences. 12:6955–6984.

Ledford HK, Niyogi KK. 2005. Singlet oxygen and photo-oxidative stress management in plants and algae. Plant, Cell and Environment. 28:1037–1045.

Levine NM, Zhang K, Longo M, et al. (17 co-authors). 2016. Ecosystem heterogeneity determines the ecological resilience of the Amazon to climate change. Proceedings of the National Academy of Sciences. 113:793–797.

Li GW, Burkhardt D, Gross C, Weissman JS. 2014a. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 157:624–635.

Li HP, Daniel B, Creeley D, et al. (12 co-authors). 2014b. Superoxide production by a manganese-oxidizing bacterium facilitates iodide oxidation. Applied and Environmental Microbiology. 80:2693–2699.

Li SHJ, Li Z, Park JO, King CG, Rabinowitz JD, Wingreen NS, Gitai Z. 2018. Escherichia coli translation strategies differ across carbon, nitrogen and phosphorus limitation conditions. Nature Microbiology. 3:939–947.

Li W, Godzik A. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22:1658–1659.

Li X, LeBlanc J, Elashoff D, et al. (15 co-authors). 2016. Microgeographic proteomic networks of the human colonic mucosa and their association with inflammatory bowel disease. Cellular and Molecular Gastroenterology and Hepatology. 2:567–583.

Liefer JD, Garg A, Fyfe MH, Irwin AJ, Benner I, Brown CM, Follows MJ, Omta AW, Finkel ZV. 2019. The macromolecular basis of phytoplankton C:N:P under nitrogen starvation. Frontiers in Microbiology. 10:1–16.

Lin H, Kuzminov FI, Park J, Lee SH, Falkowski PG, Gorbunov MY. 2016. Phytoplankton: The fate of photons absorbed by phytoplankton in the global ocean. Science. 351:264–267.

Lis H, Shaked Y, Kranzler C, Keren N, Morel FM. 2015. Iron bioavailability to phytoplankton: An empirical approach. The ISME Journal. 9:1003–1013.

Liu Y, Beyer A, Aebersold R. 2016. On the dependency of cellular protein levels on mRNA abundance. Cell. 165:535–550.

Loladze I, Elser JJ. 2011. The origins of the Redfield nitrogen-to-phosphorus ratio are in a homoeostatic protein-to-rRNA ratio. Ecology Letters. 14:244–250.

Lynch MD, Neufeld JD. 2015. Ecology and exploration of the rare biosphere. Nature Reviews Microbiology. 13:217–229.

Mangoni O, Saggiomo V, Bolinesi F, Margiotta F, Budillon G, Cotroneo Y, Misic C, Rivaro P, Saggiomo M. 2017. Phytoplankton blooms during austral summer in the Ross Sea, Antarctica: Driving factors and trophic implications. PLoS ONE. 12:1–23.

Manuell AL, Yamaguchi K, Haynes PA, Milligan RA, Mayfield SP. 2005. Composition and structure of the 80 S ribosome from the green alga Chlamydomonas reinhardtii: 80 S ribosomes are conserved in plants and animals. Journal of Molecular Biology. 351:266–279.

Marchetti A, Parker MS, Moccia LP, Lin EO, Arrieta AL, Ribalet F, Murphy MEP, Maldonado MT, Armbrust EV. 2009. Ferritin is used for iron storage in bloom-forming marine pennate diatoms. Nature. 457:467–470.

Marchetti A, Schruth DM, Durkin CA, Parker MS, Kodner RB, Berthiaume CT, Morales R, Allen AE, Armbrust EV. 2011. Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. Proceedings of the National Academy of Sciences. 109:E317–E325.

Martin J, Fitzwater S, Gordon R. 1990. Iron deficiency limits phytoplankton growth in Antarctic waters. Global Biogeochemical Cycles. 4:5–12.

Martin JH, Fitzwater SE. 1987. Iron deficiency limits phytoplankton growth in the northeast Pacific subarctic. Nature. 3:31–33.

Martin JH, Gordon RM, Fitzwater SE. 1991. The case for iron. Limnology and Oceanography. 36:1793–1802.

McCain JSP, Allen AE, Bertrand EM. 2021. Proteomic traits vary across taxa in a coastal Antarctic phytoplankton bloom. The ISME Journal. pp. 1–11.

McCain JSP, Bertrand EM. 2019. Prediction and consequences of cofragmentation in metaproteomics. Journal of Proteome Research. 18:3555–3566.

McCain JSP, Tagliabue A, Susko E, Achterberg EP, Allen AE, Bertrand EM. 2021. Cellular costs underpin micronutrient limitation in phytoplankton. Science Advances. 7.

McGill BJ, Enquist BJ, Weiher E, Westoby M. 2006. Rebuilding community ecology from functional traits. Trends in Ecology and Evolution. 21:178–85.

McQuaid JB, Kustka AB, Oborník M, et al. (11 co-authors). 2018. Carbonate-sensitive phytotransferrin controls high-affinity iron uptake in diatoms. Nature. 555:534–537.

Menden-Deuer S, Lessard EJ. 2000. Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton. Limnology and Oceanography. 45:569–579.

Metzl-Raz E, Kafri M, Yaakov G, Soifer I, Gurvich Y, Barkai N. 2017. Principles of cellular resource allocation revealed by condition-dependent proteome profiling. eLife. 6:1–21.

Mhamdi A, Queval G, Chaouch S, Vanderauwera S, Van Breusegem F, Noctor G. 2010. Catalase function in plants: A focus on Arabidopsis mutants as stress-mimic models. Journal of Experimental Botany. 61:4197–4220.

Michalski A, Cox J, Mann M. 2011. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. Journal of Proteome Research. 10:1785–1793.

Middag R, de Baar HJ, Laan P, Cai PH, van Ooijen JC. 2011. Dissolved manganese in the Atlantic sector of the Southern Ocean. Deep-Sea Research Part II: Topical Studies in Oceanography. 58:2661–2677.

Middag R, de Baar HJW, Klunder MB, Laan P. 2013. Fluxes of dissolved aluminum and manganese to the Weddell Sea and indications for manganese co-limitation. Limnology and Oceanography. 58:287–300.

Miller AF. 2012. Superoxide dismutases: Ancient enzymes and new insights. FEBS Letters. 586:585–595.

Mishra S, Imlay J. 2012. Why do bacteria use so many enzymes to scavenge hydrogen peroxide? Archives of Biochemistry and Biophysics. 525:145–160.

Mistry J, Chuguransky S, Williams L, et al. (12 co-authors). 2021. Pfam: The protein families database in 2021. Nucleic Acids Research. 49:D412–D419.

Mittler R. 2017. ROS are good. Trends in Plant Science. 22:11–19.

Mittler R, Vanderauwera S, Gollery M, Van Breusegem F. 2004. Reactive oxygen gene network of plants. Trends in Plant Science. 9:490–498.

Mittler R, Vanderauwera S, Suzuki N, Miller G, Tognetti VB, Vandepoele K, Gollery M, Shulaev V, Van Breusegem F. 2011. ROS signaling: The new wave? Trends in Plant Science. 16:300–309.

Miyake C, Michihata F, Asada K. 1991. Scavenging of hydrogen peroxide in prokaryotic and eukaryotic algae: Acquisition of ascorbate peroxidase during the evolution of cyanobacteria. Plant and Cell Physiology. 32:33–43.

Mock T, Robert P, Strauss J, et al. (34 co-authors). 2017. Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature. 541:536–540.

Molenaar D, van Berlo R, de Ridder D, Teusink B. 2009. Shifts in growth strategies reflect tradeoffs in cellular economics. Molecular Systems Biology. 5:1–10.

Moore CM, Mills MM, Arrigo KR, et al. (23 co-authors). 2013. Processes and patterns of oceanic nutrient limitation. Nature Geoscience. 6:701–710.

Moore JK, Doney SC, Lindsay K. 2004. Upper ocean ecosystem dynamics and iron cycling in a global three-dimensional model. Global Biogeochemical Cycles. 18:1–21.

Mori M, Schink S, Erickson DW, Gerland U, Hwa T. 2017. Quantifying the benefit of a proteome reserve in fluctuating environments. Nature Communications. 8:1225.

Mori M, Zhang Z, Esfahani AB, et al. (11 co-authors). 2021. From coarse to fine: The absolute Escherichia coli proteome under diverse growth conditions. Molecular Systems Biology. 17:e9536.

Morris JJ, Johnson ZI, Szul MJ, Keller M, Zinser ER. 2011. Dependence of the cyanobacterium Prochlorococcus on hydrogen peroxide scavenging microbes for growth at the ocean's surface. PLoS ONE. 6.

Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G. 2010. Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. The ISME journal. 4:673–685.

Moruz L, Käll L. 2016. Peptide retention time prediction. Mass Spectrometry Reviews. 36:615–623.

Müller JB, Geyer PE, Colaço AR, et al. (13 co-authors). 2020. The proteome landscape of the kingdoms of life. Nature. 582:592–596.

Muth T, Renard BY. 2017. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? Briefings in Bioinformatics. pp. 1–17.

Muth T, Renard BY, Martens L. 2016. Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. Expert Review of Proteomics. 13:757–769.

Najmuldeen H, Alghamdi R, Alghofaili F, Yesilkaya H. 2019. Functional assessment of microbial superoxide dismutase isozymes suggests a differential role for each isozyme. Free Radical Biology and Medicine. 134:215–228.

NASA. 2014. Goddard Space Flight Center Ocean Biology Processing. Sea-viewing Wide Field-of-view Sensor (SeaWiFS) Ocean Color Data, NASA OB.DAAC, Greenbelt, MD, USA. Accessed 2019-04-02.

Navrot N, Skjoldager N, Bunkenborg J, Svensson B, Hägglund P. 2015. A redox-dependent dimerization switch regulates activity and tolerance for reactive oxygen species of barley seed glutathione peroxidase. Plant Physiology and Biochemistry. 90:58–63.

Nelder AJA, Wedderburn RWM. 1972. Generalized linear models. Journal of the Royal Statistical Society, Series A (General). 135:370–384.

Nicholson DP, Stanley RHR, Doney SC. 2018. A phytoplankton model for the allocation of gross photosynthetic energy including the trade-offs of diazotrophy. Journal of Geophysical Research: Biogeosciences. pp. 1796–1816.

Nishiyama Y, Allakhverdiev SI, Murata N. 2011. Protein synthesis is the primary target of reactive oxygen species in the photoinhibition of photosystem II. Physiologia Plantarum. 142:35–46.

Niyogi KK. 1999. Photoprotection revisited: Genetic and molecular approaches. Annual Review of Plant Biology. 50:333–359.

Noble AE, Moran DM, Allen AE, Saito MA, Islas AMA. 2013. Dissolved and particulate trace metal micronutrients under the McMurdo Sound seasonal sea ice: basal sea ice communities as a capacitor for iron. Frontiers in Chemistry. 1:1–18.

Noctor G. 2006. Metabolic signalling in defence and stress: The central roles of soluble redox couples. Plant, Cell and Environment. 29:409–425.

Noyce AB, Smith R, Dalgleish J, Taylor RM, Erb KC, Okuda N, Prince JT. 2013. Mspire-Simulator: LC-MS shotgun proteomic simulator for creating realistic gold standard data. Journal of Proteome Research. 12:5742–5749.

Nunn BL, Faux JF, Hippmann AA, Maldonado MT, Harvey HR, Goodlett DR, Boyd PW, Strzepek RF. 2013. Diatom proteomics reveals unique acclimation strategies to mitigate Fe limitation. PLoS ONE. 8.

Ogawa K, Kanematsu S, Takabe K, Asada K. 1995. Attachment of CuZn-superoxide dismutase to thylakoid membranes at the site of superoxide generation (PSI) in spinach chloroplasts: Detection by immuno-gold labeling after rapid freezing and substitution method. Plant and Cell Physiology. 36:565–573.

Oliver H, St-laurent P, Sherrell RM, Yager PL. 2019. Modeling iron and light controls on the summer Phaeocystis antarctica bloom in the Amundsen Sea Polynya. Global Biogeochemical Cycles. 33:570–596.

O'Malley MA, Parke EC. 2018. Microbes, mathematics, and models. Studies in History and Philosophy of Science Part A. 72:1–10.

Page MD, Allen MD, Kropat J, Urzica EI, Karpowicz SJ, Hsieh SI, Loo JA, Merchant SS. 2012. Fe sparing and Fe recycling contribute to increased superoxide dismutase capacity in iron-starved Chlamydomonas reinhardtii. Plant Cell. 24:2649–2665.

Parker DJ, Lalanne JB, Kimura S, Johnson GE, Waldor MK, Li GW. 2020. Growth-optimized aminoacyl-tRNA synthetase levels prevent maximal tRNA charging. Cell Systems. 11:121–130.e6.

Parker G, Smith JM. 1990. Optimality theory in evolutionary biology. Nature. 348:27–33.

Pausch F, Bischof K, Trimborn S. 2019. Iron and manganese co-limit growth of the Southern Ocean diatom Chaetoceros debilis. PLoS ONE. 14:1–16.

Peers G, Price NM. 2004. A role for manganese in superoxide dismutases and growth of iron-deficient diatoms. Limnology and Oceanography. 49:1774–1783.

Peloquin JA, Smith WO. 2007. Phytoplankton blooms in the Ross Sea, Antarctica: Interannual variability in magnitude, temporal patterns, and composition. Journal of Geophysical Research: Oceans. 112:1–12.

Perelman A, Dubinsky Z, Martínez R. 2006. Temperature dependence of superoxide dismutase activity in plankton. Journal of Experimental Marine Biology and Ecology. 334:229–235.

Pérez-Pérez ME, Mata-Cabana A, Sánchez-Riego AM, Lindahl M, Florencio FJ. 2009. A comprehensive analysis of the peroxiredoxin reduction system in the cyanobacterium Synechocystis sp. strain PCC 6803 reveals that all five peroxiredoxins are thioredoxin dependent. Journal of Bacteriology. 191:7477–7489.

Perez-Riverol Y, Csordas A, Bai J, et al. (23 co-authors). 2019. The PRIDE database and related tools and resources in 2019: Improving support for quantification data. Nucleic Acids Research. 47:D442–D450.

Perkins A, Nelson KJ, Parsonage D, Poole LB, Karplus PA. 2015. Peroxiredoxins: Guardians against oxidative stress and modulators of peroxide signaling. Trends in Biochemical Sciences. 40:435–445.

Perlova TY, Goloborodko AA, Margolin Y, Pridatchenko ML, Tarasova IA, Gorshkov AV, Moskovets E, Ivanov AR, Gorshkov MV. 2010. Retention time prediction using the model of liquid chromatography of biomacromolecules at critical conditions in LC-MS phosphopeptide analysis. Proteomics. 10:3458–3468.

Petrov VD, Van Breusegem F. 2012. Hydrogen peroxide-a central hub for information flow in plant cells. AoB PLANTS. 12:1–13.

Pfeifer N, Leinenbach A, Huber CG, Kohlbacher O. 2007. Statistical learning of peptide retention behavior in chromatographic separations: A new kernel-based approach for computational proteomics. BMC Bioinformatics. 8:1–14.

Pible O, Allain F, Jouffret V, Culotta K, Miotello G, Armengaud J. 2020. Estimating relative biomasses of organisms in microbiota using "phylopeptidomics". Microbiome. 8:1–13.

Pilon M, Ravet K, Tapken W. 2011. The biogenesis and physiological function of chloroplast superoxide dismutases. Biochimica et Biophysica Acta - Bioenergetics. 1807:989–998.

Ploug H, Stolte W, Jørgensen BB. 1999. Diffusive boundary layers of the colony-forming plankton alga Phaeocystis sp.— implications for nutrient uptake and cellular growth. Limnology and Oceanography. 44:1959–1967.

Podell S, Gaasterland T. 2007. DarkHorse: A method for genome-wide prediction of horizontal gene transfer. Genome Biology. 8.

Polle A. 2001. Dissecting the superoxide dismutase-ascorbate-glutathione-pathway in chloroplasts by metabolic modeling. Computer simulations as a step towards flux analysis. Plant Physiology. 126:445–462.

Poole LB, Nelson KJ. 2016. Distribution and features of the six classes of peroxiredoxins. Molecules and Cells. 39:53–59.

Rahantaniaina MS, Tuzet A, Mhamdi A, Noctor G. 2013. Missing links in understanding redox signaling via thiol/disulfide modulation: How is glutathione oxidized in plants? Frontiers in Plant Science. 4:1–13.

Ramel F, Birtic S, Cuiné S, Triantaphylidès C, Ravanat JL, Havaux M. 2012. Chemical quenching of singlet oxygen by carotenoids in plants. Plant Physiology. 158:1267–1278.

Rapp I, Schlosser C, Rusiecka D, Gledhill M, Achterberg EP. 2017. Automated preconcentration of Fe, Zn, Cu, Ni, Cd, Pb, Co, and Mn in seawater with analysis using high-resolution sector field inductively-coupled plasma mass spectrometry. Analytica Chimica Acta. 976:1–13.

Raven J. 1990. Predictions of Mn and Fe use efficiencies of phototrophic growth as a function of light availability for growth and of C assimilation pathway. New Phytologist. 116:1–18.

Raven JA. 1988. The iron and molybdenum use efficiencies of plant growth with different energy, carbon and nitrogen sources. New Phytologist. 109:279–287.

Raven JA, Evans MCW, Korb RE. 1999. The role of trace metals in photosynthetic electron transport in O-2-evolving organisms. Photosynthesis Research. 60:111–149.

Redfield AC. 1958. The biological control of chemical factors in the environment. American Scientist. 46:205–221.

Reed DC, Algar CK, Huber JA, Dick GJ. 2014. Gene-centric approach to integrating environmental genomics and biogeochemical models. Proceedings of the National Academy of Sciences. 111:1879–84.

Regelsberger G, Atzenhofer W, Rüker F, Peschek GA, Jakopitsch C, Paumann M, Furtmüller PG, Obinger C. 2002. Biochemical characterization of a membrane-bound manganese-containing superoxide dismutase from the cyanobacterium Anabaena PCC 7120. Journal of Biological Chemistry. 277:43615–43622.

Reimers AM, Knoop H, Bockmayr A, Steuer R. 2017. Cellular trade-offs and optimal resource allocation during cyanobacterial diurnal growth. Proceedings of the National Academy of Sciences. p. 201617508.

Rhee G. 1974. Phosphate uptake under nitrate limitation by Scenedesmus sp. and its ecological implications. Journal of Phycology. 10:470–475.

Rho M, Tang H, Ye Y. 2010. FragGeneScan: Predicting genes in short and error-prone reads. Nucleic Acids Research. 38:1–12.

Richon C, Tagliabue A. 2019. Insights into the major processes driving the global distribution of copper in the ocean from a global model. Global Biogeochemical Cycles. 33:1594–1610.

Riffle M, May DH, Timmins-Schiffman E, Mikan MP, Jaschob D, Noble WS, Nunn BL. 2018. MetaGOmics: A Web-Based Tool for Peptide-Centric Functional and Taxonomic Analysis of Metaproteomics Data. Proteomes. 6:1–17.

Riley G. 1946. Factors controlling phytoplankton populations on Georges Bank. Journal of Marine Research. pp. 54–73.

Rodríguez-Valera F. 2004. Environmental genomics, the big picture? FEMS Microbiology Letters. 231:153–158.

Rose AL. 2012. The influence of extracellular superoxide on iron redox chemistry and bioavailability to aquatic microorganisms. Frontiers in Microbiology. 3:1–21.

Rosenwasser S, Van Creveld SG, Schatz D, et al. (19 co-authors). 2014. Mapping the diatom redox-sensitive proteome provides insight into response to nitrogen stress in the marine environment. Proceedings of the National Academy of Sciences. 111:2740–2745.

Röst H, Malström L, Aebersold R. 2012. A computational tool to detect and avoid redundancy in selected reaction monitoring. Molecular and Cellular Proteomics. 11:540–549.

Röst HL, Sachsenberg T, Aiche S, et al. (27 co-authors). 2016. OpenMS: A flexible open-source software platform for mass spectrometry data analysis. Nature Methods. 13:741–748.

Röst HL, Schmitt U, Aebersold R, Malmström L. 2014. pyOpenMS: A Python-based interface to the OpenMS mass-spectrometry algorithm library. Proteomics. 14:74–77.

Rusch DB, Halpern AL, Sutton G, et al. (40 co-authors). 2007. The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biology. 5:0398–0431.

Russell JB, Cook GM. 1995. Energetics of bacterial growth: balance of anabolic and catabolic reactions. Microbiological Reviews. 59:1–15.

Saito MA, Bertrand EM, Dutkiewicz S, Bulygin VV, Moran DM, Monteiro FM, Follows MJ, Valois FW, Waterbury JB. 2011. Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph Crocosphaera watsonii. Proceedings of the National Academy of Sciences. 108:2184–2189.

Saito MA, Goepfert TJ, Ritt JT. 2008. Some thoughts on the concept of colimitation: Three definitions and the importance of bioavailability. Limnology and Oceanography Letters. 53:276–290.

Saito MA, Mcilvin MR, Moran DM, Goepfert TJ, Ditullio GR, Post AF, Lamborg CH. 2014. Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. Science. 345:5–10.

Saito Y. 1968. A theoretical study on the diffusion current at the stationary electrodes of circular and narrow band types. Review of Polarography. 15.

Santra M, Dill KA, Graff AMRD. 2018. How do chaperones protect a cell's proteins from oxidative damage? Cell Systems. 6:1–9.

Schaechter M, Maaløe O, Kjeldgaard NO. 1958. Dependency on Medium and Temperature of Cell Size and Chemical Composition during Balanced Growth of Salmonella typhimurium. Journal of General Microbiology. 19:592–606.

Schiebenhoefer H, Bossche TVD, Fuchs S, Renard BY, Muth T. 2019. Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. Expert Review of Proteomics. 16:1–16.

Schlitzer R, Anderson RF, Dodas EM, et al. (246 co-authors). 2018. The GEOTRACES Intermediate Data Product 2017. Chemical Geology. 493:210–223.

Schmidt A, Kochanowski K, Vedelaar S, Ahrne E, Volkmer B, Callipo L, Knoops K, Bauer M, Aebersold R, Heinemann M. 2016. The quantitative and condition-dependent Escherichia coli proteome. Nature Biotechnology. 34:104–110.

Schmieder R, Lim YW, Edwards R. 2011. Identification and removal of ribosomal RNA sequences from metatranscriptomes. Bioinformatics. 28:433–435.

Schneider T, Riedel K. 2010. Environmental proteomics: Analysis of structure and function of microbial communities. Proteomics. 4:785–798.

Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. 2010. Interdependence of cell growth and gene expression: Origins and consequences. Science. 330:1099–1102.

Seaver LC, Imlay JA. 2001. Hydrogen peroxide fluxes and compartmentalization inside growing Escherichia coli. Journal of Bacteriology. 183:7182–7189.

Sedwick PN, Garcia NS, Riseman SF, Marsay CM, DiTullio GR. 2007. Evidence for high iron requirements of colonial Phaeocystis antarctica at low irradiance. Biogeochemistry. 83:83–97.

Shah P, McCandlish DM, Plotkin JB. 2015. Contingency and entrenchment in protein evolution under purifying selection. Proceedings of the National Academy of Sciences. 112:E3226–E3235.

Shaked Y, Buck KN, Mellett T, Maldonado MT. 2020. Insights into the bioavailability of oceanic dissolved Fe from phytoplankton uptake kinetics. The ISME Journal. 14:1182–1193.

Shaked Y, Harris R, Klein-Kedem N. 2010. Hydrogen peroxide photocycling in the Gulf of Aqaba, Red Sea. Environmental Science and Technology. 44:3238–44.

Shaked Y, Kustka AB, Morel MM. 2005. A general kinetic model for iron acquisition by eukaryotic phytoplankton. Limnology and Oceanography. 50:872–882.

Shaked Y, Lis H. 2012. Disassembling iron availability to phytoplankton. Frontiers in Microbiology. 3:1–26.

Sheftel H, Shoval O, Mayo A, Alon U. 2013. The geometry of the Pareto front in biological phenotype space. Ecology and Evolution. 3:1471–1483.

Sheng Y, Abreu IA, Cabelli DE, Maroney MJ, Miller AF, Teixeira M, Valentine JS. 2014. Superoxide dismutases and superoxide reductases. Chemical Reviews. 114:3854–3918.

Shenton D, Grant CM. 2003. Protein S-thiolation targets glycolysis and protein synthesis in response to oxidative stress in the yeast Saccharomyces cerevisiae. Biochemical Journal. 374:513–519.

Sherrell R, Lagerström M, Forsch K, Stammerjohn S, Yager P. 2015. Dynamics of dissolved iron and other bioactive trace metals (Mn, Ni, Cu, Zn) in the Amundsen Sea Polynya, Antarctica. Elementa: Science of the Anthropocene. 3:1–27.

Sheyn U, Rosenwasser S, Ben-Dor S, Porat Z, Vardi A. 2016. Modulation of host ROS metabolism is essential for viral infection of a bloom-forming coccolithophore in the ocean. The ISME Journal. 10:1742–1754.

Sies H. 1991. Oxidative stress: From basic research to clinical application. American Journal of Medicine. 91:S31–S38.

Sies H. 2017. Hydrogen peroxide as a central redox signaling molecule in physiological oxidative stress: Oxidative eustress. Redox Biology. 11:613–619.

Smith WO, Tozzi S, Long MC, Sedwick PN, Peloquin JA, Dunbar RB, Hutchins DA, Kolber Z, DiTullio GR. 2013. Spatial and temporal variations in variable fluoresence in the Ross Sea (Antarctica): Oceanographic correlates and bloom dynamics. Deep-Sea Research Part I: Oceanographic Research Papers. 79:141–155.

Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, Carlson CA, Smith RD, Giovanonni SJ. 2009. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. The ISME Journal. 3:93–105.

Sterner R, Elser J. 2002. Ecological Stoichiometry. Princeton University Press.

Strzepek R, Harrisson P. 2004. Photosynthetic architecture differs in coastal and oceanic diatoms. Nature. 431:689.

Strzepek RF, Boyd PW, Sunda WG. 2019. Photosynthetic adaptation to low iron, light, and temperature in Southern Ocean phytoplankton. Proceedings of the National Academy of Sciences. 116:4388–4393.

Strzepek RF, Hunter KA, Frew RD, Harrison PJ, Boyd PW. 2012. Iron – light interactions differ in Southern Ocean phytoplankton. Limnology and Oceanography. 57:1182–1200.

Strzepek RF, Maldonado MT, Hunter KA, Frew RD, Boyd PW. 2011. Adaptive strategies by Southern Ocean phytoplankton to lessen iron limitation: Uptake of organically complexed iron and reduced cellular iron requirements. Limnology and Oceanography. 56:1983–2002.

Suarez-Moreira E, Yun J, Birch CS, Williams JHH, Brasch NE. 2009. Vitamin B12 and redox homeostasis: Cob(II)alamin reacts with superoxide at rates approaching superoxide dismutase (SOD). Journal of the American Chemical Society. 131:15078–15079.

Sunda W, Huntsman S. 1985. Regulation of cellular manganese and manganese transport in the unicellular rates alga Chlamydomonas. Limnology and Oceanography. 30:71–80.

Sunda W, Kieber DJ, Kiene RP, Huntsman S. 2002. An antioxidant function for DMSP and DMS in marine algae. Nature. 418:317–320.

Sunda WG, Huntsman SA. 1995. Iron uptake and growth limitation in oceanic and coastal phytoplankton. Marine Chemistry. 50:189–206.

Sunda WG, Huntsman SA. 1997. Interrelated influence of iron, light and cell size on marine phytoplankton growth. Nature. 2051:389–392.

Szenk M, Dill KA, de Graff AM. 2017. Why do fast-growing bacteria enter overflow metabolism? Testing the membrane real estate hypothesis. Cell Systems. 5:95–104.

Tagliabue A, Aumont O, DeAth R, et al. (14 co-authors). 2016. How well do global ocean biogeochemistry models simulate dissolved iron distributions? Global Biogeochemical Cycles. 30:149–174.

Tagliabue A, Barrier N, Du Pontavice H, Kwiatkowski L, Aumont O, Bopp L, Cheung WW, Gascuel D, Maury O. 2020. An iron cycle cascade governs the response of equatorial Pacific ecosystems to climate change. Global Change Biology. 26:6168–6179.

Tagliabue A, Bowie AR, Philip W, Buck KN, Johnson KS, Saito MA. 2017. The integral role of iron in ocean biogeochemistry. Nature. 543:51–59.

Talmy D, Blackford J, Hardman-Mountford NJ, Dumbrell AJ, Geider RJ. 2013. An optimality model of photoadaptation in contrasting aquatic light regimes. Limnology and Oceanography. 58:1802–1818.

Tanca A, Palomba A, Fraumene C, et al. (11 co-authors). 2016. The impact of sequence database choice on metaproteomic results in gut microbiota studies. Microbiome. 4:51.

Tatusov RL, Fedorova ND, Jackson JD, et al. (17 co-authors). 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4:1–14.

Tehrani HS, Moosavi-Movahedi AA. 2018. Catalase and its mysteries. Progress in Biophysics and Molecular Biology. 140:5–12.

Timp W, Timp G. 2020. Beyond mass spectrometry, the next step in proteomics. Science Advances. 6:1–17.

Toseland A, Daines SJ, Clark JR, et al. (11 co-authors). 2013. The impact of temperature on marine phytoplankton resource allocation and metabolism. Nature Climate Change. 3:1–6.

Triantaphylidès C, Havaux M. 2009. Singlet oxygen in plants: production, detoxification and signaling. Trends in Plant Science. 14:219–228.

Tuzet A, Rahantaniaina MS, Noctor G. 2019. Analyzing the function of catalase and the ascorbate-glutathione pathway in H2O2 processing: Insights from an experimentally constrained kinetic model. Antioxidants and Redox Signaling. 30:1238–1268.

Twining BS, Antipova O, Chappell PD, Cohen NR, Jacquot JE, Mann EL, Marchetti A, Ohnemus DC, Rauschenberg S, Tagliabue A. 2020. Taxonomic and nutrient controls on phytoplankton iron quotas in the ocean. Limnology and Oceanography Letters. 6:96–101.

Twining BS, Baines SB. 2013. The trace metal composition of marine phytoplankton. Annual Review of Marine Science. 5:191–215.

Twining BS, Baines SB, Fisher NS. 2004. Element stoichiometries of individual plankton cells collected during the Southern Ocean Iron Experiment (SOFeX). Limnology and Oceanography. 49:2115–2128.

Twining BS, Baines SB, Vogt S, Nelson DM. 2012. Role of diatoms in nickel biogeochemistry in the ocean. Global Biogeochemical Cycles. 26:1–9.

Uzuka A, Kobayashi Y, Onuma R, Hirooka S, Kanesaki Y, Yoshikawa H, Fujiwara T, Miyagishima Sy. 2019. Responses of unicellular predators to cope with the phototoxicity of photosynthetic prey. Nature Communications. 10:1–17.

Vallino JJ. 2010. Ecosystem biogeochemistry considered as a distributed metabolic network ordered by maximum entropy production. Philosophical Transactions of the Royal Society B: Biological Sciences. 365:1417–1427.

Vallino JJ, Algar CK. 2016. The thermodynamics of marine biogeochemical cycles: Lotka revisited. Annual Review of Marine Science. 8.

Vallino JJ, Huber JA. 2018. Using Maximum Entropy Production to Describe Microbial Biogeochemistry Over Time and Space in a Meromictic Pond. Frontiers in Environmental Science. 6.

van Manen M. 2021. Interaction between dissolved and particulate iron and manganese in the Amundsen Sea, Southern Ocean. Ph.D. thesis, Royal Netherlands Institute for Sea Research.

Vance D, Little SH, de Souza GF, Khatiwala S, Lohan MC, Middag R. 2017. Silicon and zinc biogeochemical cycles coupled through the Southern Ocean. Nature Geoscience. 10.

Vizcaíno JA, Côté RG, Csordas A, et al. (18 co-authors). 2013. The Proteomics Identifications (PRIDE) database and associated tools: Status in 2013. Nucleic Acids Research. 41:1063–1069.

Vlasits J, Jakopitsch C, Bernroitner M, Zamocky M, Furtmüller PG, Obinger C. 2010. Mechanisms of catalase activity of heme peroxidases. Archives of Biochemistry and Biophysics. 500:74–81.

Völker C, Wolf-Gladrow DA. 1999. Physical limits on iron uptake mediated by siderophores or surface reductases. Marine Chemistry. 65:227–244.

Waldron C, Lacroute F. 1975. Effect of growth rate on the amounts of ribosomal and transfer ribonucleic acids in yeast. Journal of Bacteriology. 122:855–865.

Wang J, Bourne PE, Bandeira N. 2011. Peptide identification by database search of mixture tandem mass spectra. Molecular and Cellular Proteomics. 10:M111.010017.

Wang J, Bourne PE, Bandeira N. 2014. MixGF: Spectral probabilities for mixture Spectra from more than one peptide. Molecular and Cellular Proteomics. 13:3688–97.

Wardman P, Candeias LP. 1996. Fenton chemistry: An introduction. Radiation Research. 145:523–531.

Weiße AY, Oyarzún DA, Danos V, Swain PS. 2015. Mechanistic links between cellular trade-offs, gene expression, and growth. Proceedings of the National Academy of Sciences. 112:E1038–E1047.

Weisser H, Choudhary JS. 2017. Targeted feature detection for data-dependent shotgun proteomics. Journal of Proteome Research. 16:2964–2974.

Weisser H, Nahnsen S, Grossmann J, et al. (11 co-authors). 2013. An automated pipeline for high-throughput label-free quantitative proteomics. Journal of Proteome Research. 12:1628–1644.

Wilkinson RD. 2013. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. Statistical Applications in Genetics and Molecular Biology. 12:129–141.

Wilmes P, Bond PL. 2004. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environmental Microbiology. 6:911–920.

Winterbourn CC, Metodiewa D. 1999. Reactivity of biologically important thiol compounds with superoxide and hydrogen peroxide. Free Radical Biology and Medicine. 27:322–328.

Wolfe-Simon F, Grzebyk D, Schofield O, Falkowski PG. 2005. The role and evolution of superoxide dismutases in algae. Journal of Phycology. 41:453–465.

Wolfe-Simon F, Starovoytov V, Reinfelder JR, Schofield O, Falkowski PG. 2006. Localization and role of manganese superoxide dismutase in a marine diatom. Plant Physiology. 142:1701–1709.

Wollman FA, Minai L, Nechushtai R. 1999. The biogenesis and assembly of photosynthetic proteins in thylakoid membranes. Biochimica et Biophysica Acta - Bioenergetics. 1411:21–85.

Wood ZA, Poole LB, Karplus PA. 2003. Peroxiredoxin evolution and the regulation of hydrogen peroxide signalling. Science. 300:650–653.

Wrightson L, Tagliabue A. 2020. Quantifying the impact of climate change on marine diazotrophy: Insights from earth system models. Frontiers in Marine Science. 7:1–9.

Wu M, McCain JSP, Rowland E, Middag R, Sandgren M, Allen AE, Bertrand EM. 2019. Manganese and iron deficiency in Southern Ocean Phaeocystis antarctica populations revealed through taxon-specific protein indicators. Nature Communications. 10:3582.

Wytock TP, Motter AE. 2018. Predicting growth rate from gene expression. Proceedings of the National Academy of Sciences. 60208:367–372.

Xue Z, He R, Fennel K, Cai WJ, Lohrenz S, Huang WJ, Tian H, Ren W, Zang Z. 2016. Modeling pCO2 variability in the Gulf of Mexico. Biogeosciences. 13:4359–4377.

Yu R, Campbell K, Pereira R, Björkeroth J, Qi Q, Vorontsov E, Sihlbom C, Nielsen J. 2020. Nitrogen limitation reveals large reserves in metabolic and translational capacities of yeast. Nature Communications. 11:1–12.

Zamocky M, Furtmüller PG, Obinger C. 2008. Evolution of catalases from bacteria to humans. Antioxidants and Redox Signaling. 10:1527–1547.

Zámocký M, Gasselhuber B, Furtmüller PG, Obinger C. 2012. Molecular evolution of hydrogen peroxide degrading enzymes. Archives of Biochemistry and Biophysics. 525:131–144.

Zavřel T, Faizi M, Loureiro C, Poschmann G, Stühler K, Sinetova M, Zorina A, Steuer R, Červený J. 2019. Quantitative insights into the cyanobacterial cell economy. eLife. 8:1–29.

Zhang B, Pirmoradian M, Chernobrovkin A, Zubarev RA. 2014. DeMix workflow for efficient identification of cofragmented peptides in high resolution data-dependent tandem mass spectrometry. Molecular and Cellular Proteomics. 13:3211–3223.

Zhang X, Deeke SA, Ning Z, et al. (14 co-authors). 2018. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. Nature Communications. pp. 1–14.

Zubarev RA. 2013. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics. 13:723–726.

Zwanzig R. 1990. Diffusion-controlled ligand binding to spheres partially covered by receptors: An effective medium treatment. Proceedings of the National Academy of Sciences. 87:5856–5857.

Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. 2006. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. Journal of Proteome Research. 5:2339–2347.