

© Copyright 2025

William Kumler

Insights from Automated and Untargeted Marine Microbial Metabolomics

William Kumler

A dissertation

submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Anitra E. Ingalls, Chair

E. Virginia Armbrust

Randelle M. Bundy

Program Authorized to Offer Degree:

Oceanography

University of Washington

Abstract

Insights from Automated and Untargeted Marine Microbial Metabolomics

William Kumler

Chair of the Supervisory Committee:

Anitra E. Ingalls

Oceanography

Metabolomics is the study of small molecules that are the product of cellular metabolism. Most commonly done with mass spectrometry (MS), this field involves the characterization and quantification of the many small, bioactive molecules that serve as building blocks for life. In the marine environment, this gives us a powerful lens with which to view the forms of material and energy as well as the fluxes they undergo in a dynamic ecosystem dominated by microbial life. This dissertation applies metabolomics to the marine ecosystem to determine how the marketplace of metabolites is an integral part of the ocean's productivity. It also includes novel tools for metabolomics that were developed alongside these environmental analyses to streamline access to MS data because the existing frameworks were often fragile black boxes lacking reproducibility and scaling poorly across larger file sizes and sample sets. In Chapter 2, I developed an R package that reads the open-source mzML and mzXML file types into a tidy in-memory object for easy extraction and visualization of data subsets such as ion chromatograms. In Chapter 3 I extend this framework into the storage of MS data in on-disk databases to avoid

the major memory constraints. I then contrast those databases with several alternative MS data storage methods and find that the simpler, on-disk methods result in performance improvements of 10-100x in query speed without much size penalty. In Chapter 4 I use this tool to improve the peak-picking performance of one of the most common untargeted mass-spectrometry tools and show that metabolites from the North Pacific Subtropical Gyre have a wide variety of patterns with depth in the ocean but most follow biomass. Chapter 5 expands upon this dataset by investigating how mesoscale eddy features affect the abundance of those same metabolites, with clear trends in metabolite abundance mirroring those in the biogeochemistry and community. Finally, in Chapter 6 I use isotope tracing to investigate how the open ocean euphotic community uses various forms of nitrogen. I determine that ammonia is highly bioavailable at the surface while nitrate's use is restricted to a subset of organisms. Organic nitrogen, on the other hand, is used slowly and steadily with equal utility at the surface and base of the euphotic zone. In sum, this thesis contains novel insights into marine metabolites with unprecedented detail through the use of new tools and provides a framework for future investigation into marine microbial ecosystems.

Table of contents

Acknowledgements.....	3
Chapter 1: Introduction.....	11
1.1 Marine microbes	11
1.2 Metabolites and metabolomics	12
1.3 Automated and untargeted liquid-chromatography mass spectrometry	15
1.4 Overview of projects.....	17
Chapter 2: Tidy Data Neatly Resolves Mass-Spectrometry's Ragged Arrays	19
2.1 Abstract.....	19
2.2 Introduction.....	19
2.3 Why is it difficult to access mass-spectrometry data?	20
2.4 Why should it be easier to access mass-spectrometry data?.....	22
2.5 Why isn't it already easier to access mass-spectrometry data?	23
2.6 The RaMS package	26
2.7 Summary	37
2.8 Acknowledgements.....	37
Chapter 3: Databases Are an Effective and Efficient Method for Storage and Access of Mass-Spectrometry Data	38
3.1 Abstract	38
3.2 Introduction.....	38
3.3 Experimental section.....	41
3.4 Results.....	45
3.5 Discussion	59
3.6 Conclusion	66
3.7 Acknowledgements.....	67
3.8 Data availability	67
3.9 Supplemental figures	68
Chapter 4: Picky with Peakpicking: Assessing Chromatographic Peak Quality with Simple Metrics in Metabolomics	72
4.1 Abstract.....	72
4.2 Background	73
4.3 Results.....	76
4.4 Discussion	90
4.5 Conclusions.....	97

4.6 Methods.....	98
4.7 Abbreviations	110
4.8 Declarations	111
4.9 Supplemental Figures.....	113
4.10 Supplemental Tables	116
Chapter 5: Metabolites Reflect Variability Introduced by Mesoscale Eddies in the North Pacific Subtropical Gyre	118
5.1 Abstract	118
5.2 Introduction.....	119
5.3 Materials and Methods.....	121
5.4 Results	128
5.5 Discussion	142
5.6 Conclusion	149
5.7 Acknowledgements.....	150
5.8 Data availability	151
5.9 Supplemental Figures.....	152
5.10 Supplemental Tables	157
Chapter 6: Nitrogen's Form Determines its Fate in the North Pacific Subtropical Gyre	159
6.1 Abstract	159
6.2 Introduction.....	160
6.3 Results.....	162
6.4 Discussion	182
6.5 Conclusion	188
6.6 Methods.....	189
6.7 Acknowledgements.....	197
6.8 Supplemental Figures.....	199
6.9 Supplemental Tables	203
Chapter 7: Conclusions	204
Bibliography	207
Appendix: Speedy Quality Assurance via Lasso Labeling for Untargeted Mass-Spectrometry Data	225
7.1 Summary	225
7.2 Statement of need.....	225
7.3 Acknowledgements.....	229

Acknowledgements

This Ph.D. would not have been possible without an enormous number of people. Much like a marine microbial ecosystem, I have been nourished and transformed by the many individuals with whom I've been fortunate to share a lab, a university, and a life.

My deepest gratitude is of course to my phenomenal advisor Anitra Ingalls. Thank you for letting a little data science nerd join your lab and wreak havoc upon all your instruments and data. With your guidance and support I have thrived as a graduate student and grown into more than I ever dreamed. I am incredibly grateful for the many lessons you've shared and I will keep your wisdom in mind wherever I go. You cultivate the ideal environment for bold ideas and I never felt afraid to show up in your office or talk with you about an idea that I had at 10 PM the night before.

Thank you also to Laura Carlson, who is the other reason that I was able to thrive in the lab as much as I did. Your willingness to put my harebrained schemes to the test and figure out how to run weird samples on our instruments or pack things off to Hawai'i at the last minute meant the world to me. I have loved watching Noah grow at a rate far exceeding my thesis and I'm only sorry the instruments never seemed to like me as much as I like them. Josh Sacks, I owe you everything for going boldly ahead of me and figuring things out first so I could follow in your footsteps. I am deeply grateful for your companionship and your constant checking in on me, and I only hope that there are many more conversations about science and our social lives in the future. I cannot wait to see what kind of a lab you bring into the world as a professor yourself someday soon. Susan Garcia, you have been an enormous source of joy with your laughter and my life has been better since you moved into our office. Thank you for reminding me to be

human and making sure that I am being kind to myself even (especially!) when it's by unleashing Chanel upon me as soon as you get to the fifth floor.

The Ingalls lab environment is also not complete without a whole community of other folks who have helped carry me along. Katherine Heal and Angie Boysen were a huge reason I joined this lab and I am grateful every day for it. Your thoughtful guidance and willingness to sit with a baby mass spectrometrists was irreplaceable. Thank you for paving the way for my own work in this lab and serving as a source of inspiration daily. Frank Xavier Ferrer-González, you brought a biological and dissolved metabolite perspective to the lab that I did not realize I had been missing. Your hours spent teaching me biology and commenting on my presentations always broadened my horizons and made me think harder about how to connect with the wider scientific community. I was also lucky to join Regina Lionheart in OSB 506 on day one of graduate school and you kept me sane during a time I thought I was going to start speaking binary. Thank you for being willing to let me interrupt you and talk through particularly complicated snarls of code and for offering advice when I got stuck. Raisha Rahman and Claudia Luthy also deserve recognition for listening to me rant and rave about mass spectrometry and metabolomics, as do Iris Kern and Natalie Kellogg. I am glad I didn't scare you off with the hours I've spent sketching chemical structures and figures on the whiteboard. Finally, a huge amount of work in this lab is done by an army of undergraduates who do the wet chemistry that I am ill-equipped to do myself. Lindsay Turner, Leland Wood, and Alec Meyers welcomed me into the lab and showed me that lab work isn't always something to avoid. Raafay Ahmed and Amy Wang, your patience with my mentoring was admirable and I am proud of what we accomplished together. Everetta Rasyid, Anna Finch, Natalie Kledzik, and Andrew Margolis - your presence at lab meetings and the laughs we have shared remind me to always look on the

bright side even when things aren't going as planned. I am lucky to have been able to lean on and learn from every one of these Ingalls Lab members during my time here.

I am also deeply grateful for the support I received from the broader University of Washington and School of Oceanography community. My committee members Ginger Armbrust, Randie Bundy, and Mark Scheuerell offered detailed guidance and thoughtful encouragement without which I would undoubtedly have gone astray. The Armbrust lab as a whole was a frequent inspiration for my own work and the kindness of the Bundy lab members will be remembered forever. Mark, our long days of planning the Aquatic Sciences Open House between Oceanography and Fisheries got me out of my office and I have always appreciated the way you've led us with lighthearted sincerity.

My graduate school cohort was my first family in Seattle and our adventures kept me going during unprecedented times. Thank you Haila, Zoe, Maleen, Mary Margaret, Jacob, and Kitty for the kindness and support you offered me. The UW Oceanography bowling team also deserves a shoutout for always reminding me that there's more to graduate school than the work. Katy, Rita, Treasure, and everyone else who joined over the years always made my week better. Similarly, the Students Explore Aquatic Sciences (SEAS) team was the reason I came into campus during some tough days. The entire team's eagerness, empathy, and passion for outreach made me feel much less alone in trying to share science with the world.

The graduate students are kept together by our untiring shepherds in the front office with Michelle, Su, Kittie, and Taylor leading the charge. Thank you for handling so much behind the scenes and making my journey that much smoother and going out of your way to fix my mistakes. Likewise, the hardworking staff around the Ocean Sciences Building have been staples of my late nights and early mornings. Thank you for making me eager to come to work every

day. The team up at eScience deserves a special shout-out for making my budding data scientist self feel less adrift in a sea of data. I was incredibly lucky to be paired with Bryna Hazelton during through the Winter Incubator program and your guidance then and afterward forever altered my understanding of what's possible in research. Dave Beck, Naomi Alterman, and Valentina Staneva were also always ready with a kind word and gentle encouragement that kept me going. Finally, I want to acknowledge the mentors and inspirations that I found before I came to UW. Bethanie Edwards, you were the first person to show me what mass-spectrometry could do, taught me how to sample seawater for lipids and metabolomics (and did my first Winkler titration with me!), and were always a friendly face at intimidating conferences. I have enjoyed watching your lab at Berkeley continue to grow. Mimi Koehl took me on as an undergraduate minion early on in my science career and was a fantastic advisor and boss. I enjoyed working in your lab and watching our larvae and protists dance under the microscopes and I still use your crystal ball method to challenge my scientific justifications. Pete Marsden, you gave me the skills I use daily as a marine organic chemist and set an inspiringly high standard for teaching in an interesting and engaging way. Jim Bishop is the reason I came to graduate school as an oceanographer and was the first person to take me to sea. Thank you for letting me into your lab meetings, your cruises, and your home.

I am also deeply grateful for the financial support that made my Ph.D. possible. The Achievement Rewards for College Scientists (ARCS) took a risk on a recent college graduate to made it much easier for me to choose the University of Washington. I still have the pin you gave us on our very first meeting and participating in the many events always reassured me that this is somewhere I belong. All of my work was done with funding from the Simons Collaboration on Ocean Processes and Ecology (SCOPE) and it felt many times like an extended family to the

Ingalls Lab. The University of Hawai'i team took me under their wing many times during cruises and I learned so much from every one of you. Dave Karl, Sonya Dyhrman, Ed DeLong, and Ben Van Mooy were always willing to talk with me about my science and challenged me to connect it to other datasets and collaborate with other labs. Angel White and Matt Church were fantastic chief scientists during my fieldwork and their support before, during, and after was critical for getting my science done. Nick Hawco, Daniel Muratore, and Benedetto Barone also deserve recognition as early career scientists who I look up to as role models of early career scientists. Of course, the SCOPE Ops team Tim Burrell, Ryan Tabata, and Brandon Brenes practically did half of the science for me by collecting the critical environmental data during cruises that I built my metabolomics on and were always willing to lend a spare hand at the CTD. Likewise, my science would not have been possible without the captain and crew of the R/V *Kilo Moana* who safely got us to and from shore while facilitating as much science as possible along the way. Finally, to my friends and colleagues during the PERISCOPE project - an enormous thank you for making me feel at home doing field work. Our fearless champion Emily Seelen led a fantastic team and Emily Townsend, Kevan Merrow, Cat Odendahl, and Phil Kong made me feel like part of a well-honed oceanography machine that worked hard and played hard. I have been incredibly lucky to have such a fantastic network of collaborators and friends to rely upon during these past years.

An often overlooked part of research, especially in data science, is the enormous community of software engineers and data scientists who build tools that made my work possible. I know how much it means to have your work acknowledged so I am shamelessly taking this opportunity to recognize those who made my work possible and my coding life easier (Colin A. Smith et al. 2006; Tautenhahn et al. 2008; Barrett et al. 2006; Wickham et al. 2019;

Xie 2015; Sievert 2020; Chang et al. 2024; Oksanen et al. 2001; Robinson et al. 2014; Richardson 2019; Raasveldt and Mühleisen 2019; Wickham et al. 2015; Temple Lang 2000; Neuwirth 2002; Garnier et al. 2023; Van Den Brand 2021; Wilke and Wiernik 2020; Pedersen 2019; Solymos and Zawadzki 2010; Bengtsson 2021; Vaughan and Dancho 2018). The xcms team in particular (Johannes Rainer, Steffen Neumann, and Philippine Louail) were fantastic to work with virtually and I am grateful for your kindness to someone half a world away. Folks at RStudio (now Posit) especially set examples for what it means to do data science thoughtfully and transparently. Hadley Wickham, Jenny Bryan, and Yihue Xie - thank you for being fantastic role models and inspiring my own growth within the R universe. My day-to-day work was also made enormously simpler by various tools I found scattered across the web and who I gratefully used nearly every day: <https://www.chemcalc.org/mf-finder> and <https://www.sisweb.com/referenc/source/exactmas.htm>. Similarly, several online databases were a huge source of information and made a world of difference to a novice in the untargeted world (Colin A. Smith et al. 2005; Wishart et al. 2022; Aron et al. 2020; Haug et al. 2019; Sud et al. 2016; Kanehisa 2000). Your often invisible work is deeply appreciated and inspirational to a daydreaming data scientist.

Of course, not all my time in grad school was spent on campus or in the field. Thank you to the Stone Gardens (now Edgeworks Seattle) climbing community for letting me be antisocial and burn off excess energy at the end of a day. The Seattle Public Libraries were a regular escape for me as well and kept me feeling inspired to read and write even when I was feeling burned out on words. The Seattle Aquarium was another source of joy during my time here and the evening events team made it incredibly easy to volunteer. They provided willing victims for my long raves about chitons and kept reminding me why ocean science is so important. Finally, Seattle

itself rests on the unceded ancestral lands and waters of the Coast Salish Peoples, past and present. They have been stewards of this land since time immemorial and I am grateful for their ongoing leadership and care.

I consider my close network of dear friends to be my second brains and better halves, so you too deserve credit for having made it through this Ph.D. with me. Shawnee Traylor, I have no idea what I did to deserve running into you during our grad school visits but your steadfast friendship and incredible generosity in time and support has meant the world to me. Kaylyn Torkelson, you have always been my fiercest defender and care more about me than I do myself. Thank you for making sure I socialize with others, daydream boldly about what I can do, and stay honest with myself about what I'm going through. Hannah Dawson and Rachel Liu, you were delightful coworkers but more importantly incredibly close friends who inspired me to get outside and spend time with people even when I wanted to spend the day under a rock. The days you dropped into my office were always the better ones and I'm glad we have only gotten closer. Wave Caldwell, Seattle was much brighter when you were here and I miss our days with Mona in your old apartment but I'm so glad you're thriving in Colorado now and I cannot wait to come visit next. Similarly, Kate Miller has always reminded me that a little bit of manic energy goes a long way and our mutual chaos has snapped me out of many an unhappy day. Finally Sarra Ghezzaz, who always has stories to share and is one of the most thoughtful listeners I know. I am lucky to have picked up each of you during my life and cannot wait for more adventures together.

My older friends have also seen the many ups and downs I've experienced and cheered me through the entire grad school journey. Sylvia Targ has always helped me find the beauty, utility, and joy in the ocean and I have cherished the opportunity to watch you discover yourself

as well. The times I spent crashing on your couch in the Bay were some of the highlights of these last few years. Thank you too to Colby Gekko who picks up the phone even when I call without any warning and talks me out of bad decisions and into good ones even as I lay waste to your productivity with Stardew Valley. Allison Herbert, our mutual love of cat memes and unhinged Lord of the Rings references often helped me start my day with a smile if not a fully-fledged cackle. Natasha Castellon and Rachel Lai, you both have seen me grow so much since our days in Freeborn and I am deeply grateful to you for sticking with me throughout. Our check-ins always came at crucial times and I am excited to continue growing together during another set of big changes in our lives coming up. And Este Gonzalez is one of my oldest friends and has seen it all without judgment or conviction. Let's get Rosa Maria's sometime soon in celebration!

My family, of course, has been with me since day one not only of grad school but of my entire life. My parents Mark and Brigid always encouraged questions and will always inspire me to listen well, think carefully, and act boldly. My grandmothers reminded me to call and take a breather occasionally and kept me on track with their questions about graduation and what comes next. My brother Ben and my sister Maggie were also important parts of my grad school journey and our phone call chats or random Portland trips always brought me enormous joy. I would not be a chemist if we hadn't spent so many hours at the dinner table discussing the state of matter of whipped cream or the heat capacity of a potato. You supported me when I was disheartened and gave me space when I needed it and I wouldn't trade any of you for the world.

And you, dear reader, I would like to acknowledge if you too are wading into the wide world of marine microbial metabolomics. I hope that you enjoy the science as much as I have and that you continue to build our rapidly-growing field with careful ingenuity and accessible reusability. You are in for a treat.

Chapter 1: Introduction

1.1 Marine microbes

Marine carbon fixation happens at an incredible rate. In the blink of an eye (~100 milliseconds), the ocean converts a blue whale's mass of atmospheric carbon into biomass and has performed this continuously for at least the last two billion years (Falkowski 1994; Ligrone 2019). Most of this is performed by single-celled organisms too small to see with the naked eye known as microbes (Falkowski 1994; Falkowski et al. 2008). The process by which they transform air and nutrients into food is the base of the marine food web and regulates Earth's climate, with many fates available to the fixed carbon. A large fraction of this particulate matter will be transformed back into CO₂ via respiration within the surface ocean, either by the phytoplankton themselves or the rest of the food chain. A smaller fraction makes it out of the euphotic zone via the biological pump and is sequestered for hundreds to thousands of years, while an even smaller fraction survives to the seafloor and can be sequestered for millennia in marine sediments (Iversen 2023; Siegel et al. 2023).

The pathway a particular atom of carbon travels is determined by the structure of the molecule it composes and the environment in which it's found. Highly labile compounds such as sugars and amino acids can be converted almost instantaneously back into CO₂, while ultra-refractory compounds can persist for thousands of years (Moran, Ferrer-González, et al. 2022). Our understanding of the marine environment's biogeochemistry and community composition has vastly expanded in the past few decades thanks to the establishment of long-term ecological time series and advances in genetic tools, while our characterization of organic carbon lags far behind (Moran, Kujawinski, et al. 2022; Longnecker et al. 2024). Determining the molecular

composition of marine carbon and its fluxes through the environment is therefore paramount in improving our ability to accurately model the microbial marine ecosystem (Jones et al. 2024).

1.2 Metabolites and metabolomics

Metabolites are defined simply as the products of cellular metabolism, but this uncomplicated definition belies the dizzying complexity of microbial processes. While technically all biologically produced molecules could fall within this category, the conventional usage refers to the small (<1000 Dalton) organic molecules that act as currencies within the cell while excluding macromolecules such as proteins and lipids. Metabolites are often the reaction intermediates and building blocks of larger molecules but have several important roles of their own, including nutrient and energy storage (Van Mooy et al. 2009; Becker et al. 2018; Mojzeš et al. 2020), antioxidation (Narainsamy et al. 2016), osmotic balance (Yancey et al. 1982; Yancey 2005), buoyancy (Yancey 2005; Boyd and Gradmann 2002), and cell signaling (both beneficial and antagonistic interactions) Thukral et al. (2023). There are likely hundreds of thousands of individual molecules composing the metabolome in the environment, making their comprehensive analysis challenging (Schrimpe-Rutledge et al. 2016).

Nonetheless, metabolomics attempts to do so. The study of “all” small molecules in the cell is a rapidly growing field with over 10,000 publications in 2024 and recently eclipsed all other “omics” fields of study according to a topic search in Web of Science (Figure 1.1, Patti et al. (2012), Edwards (2023)). These publications span a massive swath of disciplines, with contributions from medicine, polymer chemistry, astronomy, and oceanography. This interdisciplinary nature has resulted in the construction of expansive databases linking organisms’ genetic potential to their realized state (Bauermeister et al. 2022; Kanehisa 2000; Karp et al. 2019).

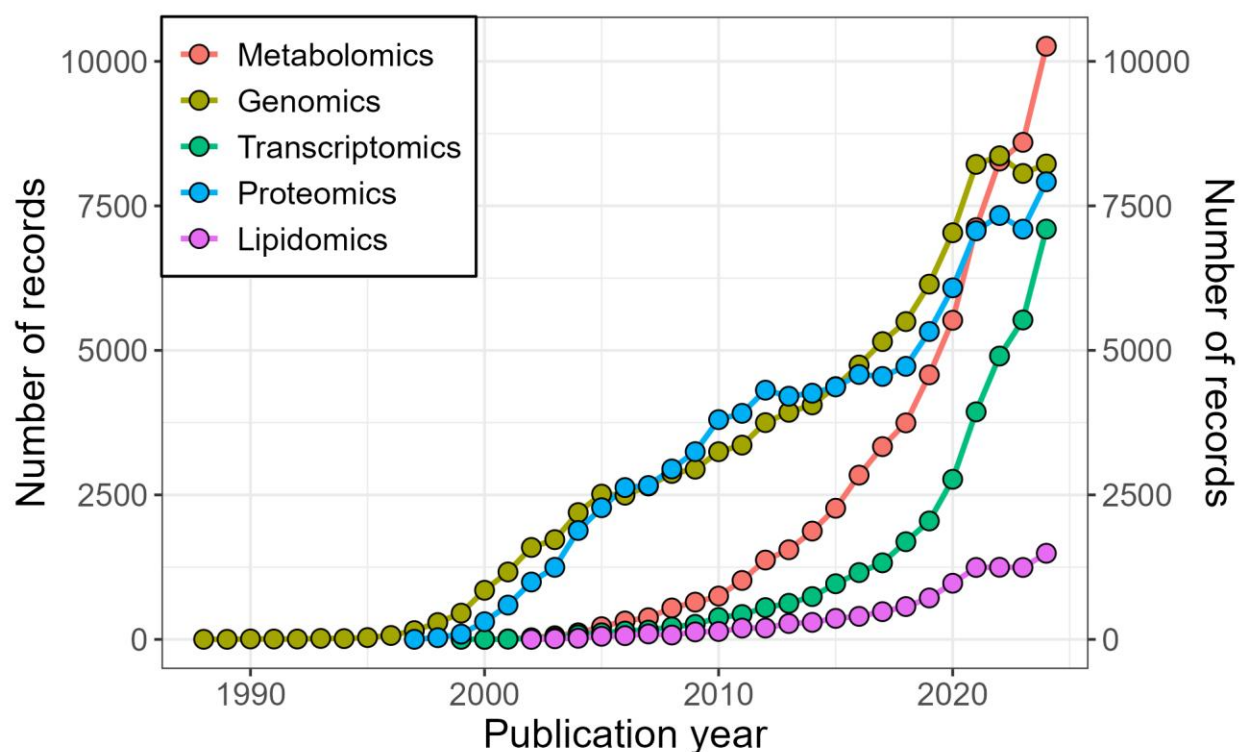


Figure 1.1: Number of publications indexed by Web of Science yearly since 1988 across different 'omics disciplines. Data were generated by searching the term in the legend as a topic and tabulated as a bar chart using the WOS Analyze Results option for Publication Years. All data rows were exported to CSV and plotted here using R's ggplot2 library.

Quantifying all small molecules in the cell is challenging for many reasons. First, metabolites span a wide range of chemical properties that cannot all be extracted simultaneously or separated on the same type of chromatography (Kido Soule et al. 2015; Cajka and Fiehn 2016; Gika et al. 2019). Second, their wide range of roles in the cell mean that annotating signals is more difficult than proteomics or lipidomics because their building blocks are not shared (Schrimpe-Rutledge et al. 2016). Third, the diversity and novelty of many compounds makes pure standards often unavailable, let alone isotopically-labeled versions necessary for the construction of the gold-standard multipoint internal calibration curve (Patti et al. 2012; Cajka and Fiehn 2016).

The problems listed above are exacerbated in marine microbial metabolomics. Primarily this is due to their incredibly low concentrations in both the particulate and dissolved phases, with typical values in the picomolar to nanomolar range (Heal et al. 2021; Sacks et al. 2022; Moran, Kujawinski, et al. 2022; Longnecker et al. 2024). An additional problem is the way the salty matrix of seawater behaves similarly to many metabolites during chemical analysis but numerically dominates their abundance by 10^5 to 10^{10} molecules per liter. (Boysen et al. 2018; Longnecker et al. 2024). In contrast to other metabolomics specialties where the organism of interest is well studied and genetically documented, environmental metabolomics struggles with a lack of genetic representation and less than 5% of the genetic diversity in the ocean has been captured by reference genomes (DeLong 2005; Salazar and Sunagawa 2017). Certainly fewer than 5% of the organisms in the ocean have been cultured in the lab and their metabolites documented, though work to improve this is underway (Heal et al. 2021; Durham et al. 2022; Kujawinski et al. 2023). Finally, the general inaccessibility of the open ocean results in chronic undersampling and significantly reduced sample sizes relative to land-based metabolomics, resulting in low-power analyses that are only able to detect the largest signals (Karl and Church 2017).

Despite these challenges, marine microbial metabolomics shows significant promise for characterizing the composition of seawater and the organisms that live within it. Metabolites have been used to describe the latitudinal variation in marine particles (Heal et al. 2021; Johnson et al. 2023, 2020), the response of the microbial community to nutrient and vitamin availability (Sañudo-Wilhelmy et al. 2014; Heal et al. 2017; Bertrand et al. 2015; Wilson et al. 2019; Dawson et al. 2020), and the response of phytoplankton to changes in temperature and salinity (Dawson et al. 2023) as well as their response over the diel cycle (Muratore et al. 2022; Boysen

et al. 2021). Additionally, recent work on metabolites dissolved in seawater has begun to unlock the vast diversity of organic carbon and nitrogen in the ocean (Sacks et al. 2022; Widner et al. 2021; Johnson et al. 2017). All of these efforts have implications for the way the smallest molecules in the ocean affect its ability to cycle energy and matter through the globe.

1.3 Automated and untargeted liquid-chromatography mass spectrometry

Mass spectrometry (MS) is the dominant analytical platform in metabolomics (Cajka and Fiehn 2016; Gika et al. 2019). Commonly, this technique is paired with chromatographic separation to allow isomers to be quantified independently and to provide additional information about the chemicals' nature. The disadvantage of this pairing is that the signal must then be integrated in retention time to provide an accurate reconstruction of the original quantity. With noisy signals such as those produced by hydrophilic interaction columns (HILIC, Buszewski and Noga (2012)) and compounds near the limit of detection, this becomes a challenge. The conventional solution is manual integration, in which a mass-spectrometrists manually reviews the extracted chromatograms and determines the start and end of chromatographic peak for integration, often via graphical user interface (GUI). However, this method is time consuming (scaling with the number of compounds and the number of samples) and cannot be guaranteed to be reproducible. This has led to the use of software for automatic peak detection and integration.

Automatic peakpicking and annotation software has been developed in parallel for the better part of two decades by both open-source and commercial endeavors (Colin A. Smith et al. 2006; Tautenhahn et al. 2008; Heuckeroth et al. 2024; Schmid et al. 2023; Tsugawa et al. 2015; Rafiei and Sleno 2015; Coble and Fraga 2014; Hohrenk et al. 2020). The focus of these tools is typically on untargeted metabolomics (including proteomics and lipidomics), which uses a data-driven approach to compound detection rather than approaching the dataset with a list of

anticipated compounds (Gika et al. 2019). This approach is particularly useful for marine microbial metabolomics, where many compounds are yet to be discovered and the additional features detected produce more powerful statistics to compensate for small sample sizes. The untargeted method also comes with significant drawbacks, with imperfect integrations by the peakpickers, multiple signals due to adducts and isotopes, and low-confidence annotations still requiring extensive manual review (Myers et al. 2017a). While untargeted analysis is traditionally associated with hypothesis generation because of its compound discovery capability (Giera et al. 2022; Thukral et al. 2023), it is perfectly qualified for testing of a well-formed hypothesis as well.

I highlight here the distinction between untargeted MS and automated MS because they are orthogonal philosophies often conflated. It is entirely possible (and often desired!) to have an automatic targeted workflow where specific compounds of interest are quantified with the speed and reproducibility of an algorithm without expanding the analysis to unknowns. Similarly, it is possible to perform untargeted metabolomics with traditionally targeted tools such as Proteowizard's Skyline (Adams et al. 2020) or even Microsoft Excel as long as the data is used to drive discovery instead of a priori knowledge about the expected compounds. For example, one could imagine an Automated Data Analysis Pipeline (ADAP) type algorithm (Myers et al. 2017b) that recursively extracts the largest intensities in a file and nearby m/z values for manual integration. These "alternate" MS methods (targeted automation and manual untargeted) are underutilized simply because the tools for their use have not yet been implemented or documented sufficiently.

1.4 Overview of projects

This thesis presents a body of work spanning data science and oceanography. In the first chapter, I discuss how MS data can be enormously simplified by converting it into a “tidy” format in the sense of Wickham (2014). This allows for the rapid exploration and reproducible analysis that I use in the rest of the thesis. Chapter 3 logically extends this framework into proper database systems which mitigates Chapter 2’s major problems with memory usage. I additionally compare multiple database systems with emphasis on modern column-oriented and online analytical processing methods that show particular promise. The particular strength of these methods is their ability to look at data *across* files rather than within a single one, something that I heavily leverage in later chapters.

Chapter 4 demonstrates the utility of allowing for rapid raw data access by showing how novel peak metrics calculated from the raw data can significantly reduce the rate of false positives in existing peakpicking software. This “cleaned” data set then shows interesting differences between marine microbial samples taken from different depths that were not apparent in the original. In the appendix, I also illustrate how raw data itself can be treated as a multidimensional array with the largest “signals” being those of high-quality peaks, allowing dimensionality reduction techniques to group MS features for rapid quality annotation.

Chapters 5 and 6 are applications of the above philosophy to oceanographic data collected from the North Pacific Subtropical Gyre (NPSG) near Station ALOHA. The NPSG is the largest biome on the planet and, like most of the surface ocean, is limited by the bioavailability of nitrogen despite large standing stocks of dinitrogen gas and DON in addition to the constant upwelling of nitrate from the deep (Moore et al. 2013; Karl and Church 2017). Since nitrogen limits the amount of carbon fixation and export possible, understanding the forms and fluxes of nitrogen-containing molecules and the organisms they compose directly affects our

ability to predict marine carbon cycling. As the majority of the nitrogen flux is through small, polar molecules (Moran et al. 2016; Moran, Kujawinski, et al. 2022), metabolomics is particularly well suited to describing and quantifying these elemental cycles.

Chapter 5 documents an exploratory metabolomics dataset collected in the NPSG across two sets of mesoscale eddy features of opposing polarity. In many ways this chapter felt like a return to the Challenger era of observational oceanography which required the use of complex ecological statistics to unravel the impacts of sea level anomaly on the ocean's metabolome and reported several compounds for the first time in the open ocean. Chapter 6, in contrast, was a deeply-nested experimental framework using short-term incubations with isotopically-labeled nitrogen substrates to test specific hypotheses about microbial nutrient acquisition and use. In both cases, the rapid and intuitive exploration of select chromatograms as well as access to the raw data was key for constructing a confident and coherent narrative of the microbial role in ocean biogeochemistry.

Chapter 2: Tidy Data Neatly Resolves Mass-Spectrometry's Ragged Arrays

2.1 Abstract¹

Mass spectrometry (MS) is a powerful tool for measuring biomolecules, but the data produced is often difficult to handle computationally because it is stored as a ragged array. In R, this format is typically encoded in complex S4 objects built around environments, requiring an extensive background in R to perform even simple tasks. However, the adoption of tidy data (Wickham 2014) provides an alternate data structure that is highly intuitive and works neatly with base R functions and common packages, as well as other programming languages. Here, we discuss the current state of R-based MS data processing, the convenience and challenges of integrating tidy data techniques into MS data processing, and present RaMS, a package that produces tidy representations of MS data.

2.2 Introduction

Mass-spectrometry (MS) is a powerful tool for identifying and quantifying molecules in laboratory and environmental samples. It has grown enormously over recent decades and has been responsible for countless advances in chemical and biological fields. It is often paired with liquid chromatography (LC) to separate compounds by retention time and improve detection limits. The large quantity of data produced by increasingly rapid and sensitive instruments has facilitated the adoption of computational methods that use algorithms to detect, identify, and quantify molecular signatures.

¹ This chapter was published as Kumler, W. and Ingalls, A.E. 2022. "Tidy Data Neatly Resolves Mass-Spectrometry's Ragged Arrays." *R Journal* 14 (3): 193-202. <https://doi.org/10.32614/RJ-2022-050>

Many mass-spectrometrists have some exposure to programming, often in R, and this familiarity is expected to increase in the future as computational methods continue to become more popular and available. However, these researchers typically focus on results and the conclusions that can be drawn from them rather than the arcane details of any particular language or package. This produces a demand for simple data formats that can be quickly and easily understood by even a novice programmer. One such representation is the “tidy” data format, which is rapidly growing in popularity among R users for its consistent syntax and large library of supporting packages (Wickham 2014). By formatting MS data tidily, the barrier to entry for novice programmers is dramatically reduced, as `tidyverse` functions learned elsewhere will function identically on MS data.

This article begins by reviewing the current theory and implementation of MS data handling, as driven by three major questions. First, why is it difficult to access and interpret MS data? Second, why should it be easier to do this? Finally, why don’t current algorithms make it trivial to do this? In the latter portion of this article, we introduce a new package, called R-based access to Mass Spectrometry data (RaMS) that provides tidy access to MS data and will facilitate future analysis and visualization.

2.3 Why is it difficult to access mass-spectrometry data?

Mass spectrometers produce data in the form of ragged (also sometimes called “jagged”) arrays. These data structures contain an unequal number of columns per row because any number of ion masses (m/z ratios) may be observed at a given time point. This data is typically managed in a list-of-lists format, with a list of time points each containing a list of the ions observed and their abundances. While this is an effective way to preserve the data structure as it was produced by the instrument, it is less helpful when performing analysis. Typically, analysis (both manual and

computational) iterates over m/z windows rather than time. The main focus is the extracted ion chromatogram (EIC) which represents all time points for a given mass, and the spectrum of masses obtained at a given time point is less useful during the preliminary review and initial discovery phases. This nested syntax, often itself contained within S4 objects and encoded as an environment, makes it difficult to extract EICs quickly and intuitively.

Even so, “difficult” is a relative assessment. Veteran R programmers have little difficulty writing elegant code that embraces these ragged arrays and the list-of-lists syntax. Indeed, the dominant MS processing package in R, MSnbase currently uses the S4 object system to great effect. However, MS experts are rarely also R experts and have a working familiarity with R rather than a comprehensive background in computer science. This working knowledge typically includes creating plots, subsetting data, and manipulating simple objects but does not extend to the nuances of the S4 object system or methods for rewriting package code. Thus, a package capable of converting these complex data structures into a familiar format appears to be very much in demand.

Finally, it should be noted that existing MS data processing packages are designed to be holistic pipelines which accept raw data and output definitive results. There is very little room for a user’s customization beyond the provided function arguments despite the enormous variability in MS setups, usage, and data quality. It is often challenging to access intermediate objects as a way to debug unexpected results, and published code is rarely easy to edit safely due to poor documentation and unit test coverage. These issues are compounded by the agglomerative nature of R packages that build extensively upon other R packages; the popular *xcms* processing package has over a hundred dependencies installed from across CRAN and Bioconductor, with further functionality provided by unregulated code from GitHub and

SourceForge. When combined with additional issues from C++ compilers, versioning, and operating system discrepancies, MS data analysis becomes very much a “black box” with functioning pipelines treated as fragile rather than simple, robust, and reproducible.

2.4 Why should it be easier to access mass-spectrometry data?

Mass-spectrometry data is fundamentally simple. In LC-MS full-scan mode, each data point has three coordinates corresponding to the time, molecular mass, and intensity dimensions. Even the more complex fragmentation data requires only a single additional dimension, fragment mass.

While this ignores the large quantity of critical metadata associated with each file that must also be stored somewhere, a core part of MS research is driven by the data alone. In this preliminary stage of analysis, metadata is less relevant than simple exploratory questions about which molecules can be detected and preliminary assessments of data quality. This exploratory phase is driven by rapid, ad hoc discovery and hypothesis testing that typically requires visualizing chromatograms and the raw data to assess quality: this appears to be one of the reasons why R and its built-in plotting ability is so popular for MS analysis (Gatto et al. 2021). These queries should be trivial to implement, even for beginning R users, but current data storage methods make them difficult and often time-consuming. Currently, the easiest questions to answer about MS data are metadata-based queries about the instrument that the analyst is usually already able to answer. This is an artifact of information storage in most raw data files, with metadata available readily at the top level and measurements buried deep within.

Raw MS data is typically converted from vendor-specific formats into open-source versions that can be parsed without proprietary software. The modern standard is the mzML document, which has been designed to combine the best aspects of precursor standards in a single universal format (Deutsch 2010). These XML documents have well-defined schema built

around a controlled vocabulary to enable consistent parsing. Most critically, the development of the modern mzML format established accession numbers for each attribute which (according to the specification document) should never change. This stability means that the data can be accessed robustly with any XML parser. Older formats, such as mzXML, are currently deprecated and will not undergo further development, making them equally stable.

Finally, simple data formats make it easier to work within existing frameworks rather than developing exclusive functions. Tidy data interacts neatly with the entire `tidyverse` thanks to its shared design philosophy and it's simple to upgrade basic data frames to `data.tables` for improved access speed. More crucially, however, simple formats make it possible to port MS data to other languages and interfaces. It is straightforward to convert an R data frame to Python's `pandas` version via the `reticulate` package, encode it as a SQL database, or export it as a CSV file to be viewed in Excel or other familiar GUIs. The same cannot be said for R's environments and S4 objects. This connectivity ensures that the best tools possible can be applied to a problem, rather than the subset available in a given package or programming language. Simplifying access to and working storage of MS data is a critical step for the further development of fast, accurate algorithms for the detection and quantification of molecules across many areas of science.

2.5 Why isn't it already easier to access mass-spectrometry data?

Of course, there are challenges that make simplification difficult and a trade-off must be made between speed, storage, and sanity. Tidy data favors code readability and intuitiveness over computational efficiency: for example, a list-of-lists model is more memory efficient than the proposed rectangular data structure because each time point is stored once rather than repeated in each row. When multiple files are analyzed simultaneously, tidy data also requires that the

filename be repeated similarly, resulting in essentially a doubling of object size in the computer memory. Given that most MS experiments involve tens or hundreds of large files, this is a major concern and current packages handle memory carefully, either reading from disk only what is needed or running files in batches. There are several ways to resolve this problem within the tidy data model as well. During the exploration phase, it is rarely necessary to load all data from files simultaneously, but viewing some portion of the data is still critically important for quality control. With the tidy model, it's not required to import all the data in a single comprehensive step. Instead, quality control files or pooled samples can be viewed as representative of the whole run and rarely challenge memory requirements. Additionally, tidy data makes it easy to subset only the masses of interest for targeted analyses, and the remainder of the data can be discarded from memory. For the final comprehensive analysis, it is much simpler to encode MS data into an external database for access via SQL or other query language when formatted tidily than it is to wrangle current implementations into some accessible object that can handle project sizes larger than the computer's memory.

Theoretically, the ideal data structure for MS data processing speed would invert the current list-of-lists schema by constructing a list of unique m/z values, each containing the time points at which that mass ratio was observed and the corresponding intensity. However, this method is complicated by the instrumental error inherent in measuring molecular masses. The same molecule may be measured to have a slightly different mass at each time point, and “binning” these masses together across all time points for a single consensus value risks incorporating nearby masses together even at hypothetical sub-ppm mass accuracy (Kind and Fiehn 2006). Instead, m/z values are continuous rather than discrete, making it difficult to encode the data in this way. A tidy framework resolves part of this issue by storing the time and m/z

values in columns that can be indexed by a binary search, such as the one implemented by `data.table`. This allows for rapid subsetting by both time and m/z . Finally, it is worth noting that computers have rapidly grown faster and larger while human intuition has not grown as quickly. This indicates that concerns with processing time and memory will lessen over time and that in the long run, sanity should be prioritized over speed and storage.

There are other reasons that a tidy approach has not yet been implemented for MS data. MS files include large amounts of metadata which should not be discarded, but are challenging to encode efficiently in a rectangular format. A proper tidy approach requires that a separate table be constructed to hold this per-file metadata, with a key such as file name that permits joining the metadata back to the original information. Compared to the monolithic S4 objects constructed by traditional workflows, managing multiple tables may be unappealing. S4 objects also excel at recording each process that is performed on the data, and a specific “processes” slot is found in some objects to record exactly this. However, with the emergence of code sharing and open-source projects it becomes less critical that the data itself records the process because the source code is available.

Finally, a significant history exists for today’s methods. MSnbase, the first widely-used R package designed to process MS data, implemented S4 objects as a way to hold entire MS experiments in memory, and dependent packages extend this MSnExp object in various ways rather than discarding it entirely. This development history and connected network of packages is incredibly useful and represents an extensive process of innovation and refinement. We would like to emphasize that the concerns raised here and the package introduced below are not designed to critique or replace this significant effort. Instead, our goal is to function alongside prior work as a way to enable rapid, interactive, and preliminary exploration. Following initial

investigation, we recommend using the existing pipelines and extensive package network to establish a reproducible, scripted process of MS data analysis.

2.6 The RaMS package

The RaMS package implements in R a set of methods used to parse open-source mass-spectrometry documents into the R-friendly data frame format. Functions in the package accept file names and the type of data requested as arguments and return rectangular data objects stored in R's memory. This data can then be processed and visualized immediately using base R functions such as `plot` and `subset`, passed to additional packages such as `ggplot2` and `data.table`, or exported to language-agnostic formats such as CSV files or SQL databases.

2.6.1 Installation

The RaMS package can be installed in two ways, the release version from CRAN:

```
install.packages("RaMS")
```

or the development version from GitHub:

```
# install.packages("remotes")
remotes::install_github("wkumler/RaMS")
```

2.6.2 Input arguments

RaMS is simple and intuitive, requiring the memorization of a single new function `grabMSdata` with the following usage:

```
grabMSdata(files)
```

Where `files` is a vector of file paths to mzML or mzXML documents, which can be located on the user's computer, a network drive, FTP site, or even at a URL on the Internet. Further parameters are documented below in Table 2.1:

Table 2.1: Parameters accepted by the `grabMSdata` function.

Parameter	Description
<code>grab_what</code>	Specifies the information to extract from the mzML or mzXML file. Can

Parameter	Description
	currently accept any combination of “MS1”, “MS2”, “EIC”, “EIC_MS2”, “metadata”, and “everything” (the default).
verbosity	Controls progress messages sent to the console at three different levels: no output, loading bar and total time elapsed, and detailed timing information for each file.
mz	Used when grab_what includes “EIC” or “EIC_MS2”. This argument should be a vector of the m/z ratios interesting to the user, if the whole file is too large to load into memory at once or only a few masses are of interest.
ppm	Used alongside the mz argument to provide a parts-per-million error window associated with the instrument on which the data was collected.
rtrange	A length-two numeric vector with start and end times of interest. Often only a subset of the LC run is of interest, and providing this argument limits the data extracted to those between the provided bounds.

2.6.3 Usage

Extracting data with `grabMSdata` returns a list of tables, each named after one of the parameters requested. A `grab_what` argument of "MS1" will return a list with a single entry, the MS¹ (i.e. full-scan data) for all of the files:

```
msfile <- system.file("extdata", "LB12HL_AB.mzML.gz", package = "RaMS")
msdata <- grabMSdata(files = msfile, grab_what="MS1")
head(msdata$MS1)
```

Table 2.2: Tidy format of RaMS output showing columns of MS¹ data, with columns for retention time (rt), mass-to-charge ratio (mz), intensity (int) and name of the source file (filename). Note that this is a subset - the actual object contains 8,500 entries.

rt	mz	int	filename
4.009	104.0710	1297755.000	LB12HL_AB.mzML.gz
4.009	104.1075	140668.125	LB12HL_AB.mzML.gz
4.009	112.0509	67452.859	LB12HL_AB.mzML.gz
4.009	116.0708	114022.531	LB12HL_AB.mzML.gz
4.009	118.0865	11141859.000	LB12HL_AB.mzML.gz
4.009	119.0837	9636.127	LB12HL_AB.mzML.gz

This table is already tidied, ready to be processed and visualized with common base R or

tidyverse operations. For example, it's often useful to view the maximum intensity observed at

each time point: this is known as a base peak chromatogram or BPC. Below are two examples of calculating and plotting a BPC using base R and the tidyverse.

```
# Base R
BPC <- tapply(msdata$MS1$int, msdata$MS1$rt, max)
plot(names(BPC), BPC, type="l")
```

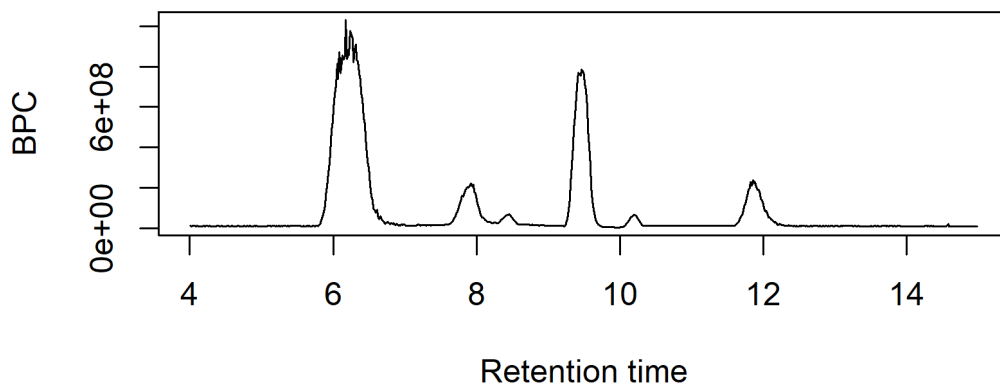


Figure 2.1: A simple chromatogram plotted using base R. This plot shows the retention time of all compounds in a sample plotted against the maximum intensity at each timepoint. Base graphics were used so the plot is fully customizable with normal graphics options.

```
# Tidyverse
library(tidyverse)
BPC <- msdata$MS1 %>%
  group_by(rt) %>%
  summarize(BPC_int=max(int))
ggplot(BPC) + geom_line(aes(x=rt, y=BPC_int))
```

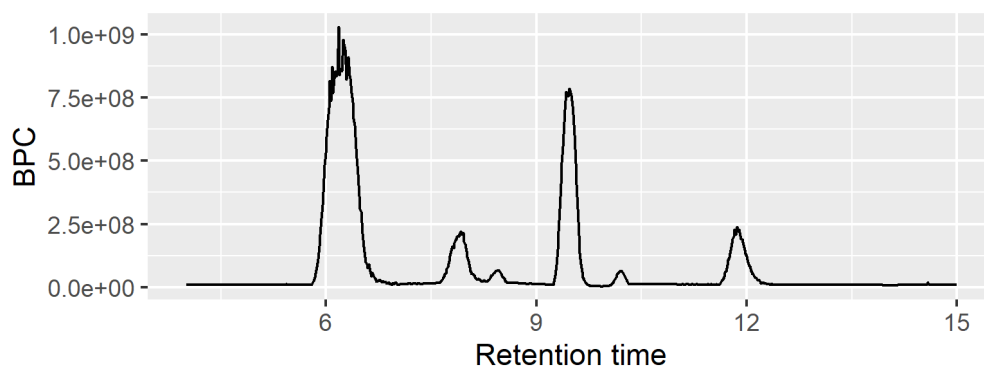


Figure 2.2: A simple chromatogram plotted using the ggplot2 package. This plot shows the same data as Figure 1 of retention time by maximum intensity across compounds but uses ggplot2 syntax and defaults.

Importantly, note that the creation of these plots required no special knowledge of the S3 or S4 systems and the plots themselves are completely customizable. While similar packages provide methods for plotting output, it is rarely obvious what exactly is being plotted and how to customize those plots because the data is stored in environments and accessed with custom code. RaMS was written with the beginning R user in mind, and its design philosophy attempts to preserve the most intuitive code possible.

RaMS uses `data.table` internally to enhance speed, but this also allows for more intuitive subsetting in mass-spectrometry data. With `data.table`, operations are nearly as easy to write in R as they are to write in natural language, leveraging the user's intuition and decreasing the barrier to entry for non-coder MS experts. For example, a typical request for MS data might be written in natural language as

“All MS¹ data points with m/z values between an upper and lower bound, from start time to end time.”

This request can be written in R almost verbatim thanks to `data.table`'s intuitive indexing and `%between%` function:

```
msdata$MS1[mz %between% c(upper_bound, lower_bound) &  
  rt %between% c(start_time, end_time)]
```

Most importantly, this syntax doesn't require the mass-spectrometrists to have an understanding of how the data is stored internally. Current implementations use S4 objects with slots such as “chromatograms” and “spectra” or derivatives of these, despite their inconsistent usage across the field and unclear internal structure. (R. Smith et al. 2015)

RaMS enhances the intuitive nature of `data.table`'s requests slightly by providing the `pmppm` function, short for “plus or minus parts-per-million (ppm)”. Masses measured on a mass-spectrometer have a certain degree of inherent deviation from the true mass of a molecule, and

the size of this error is a fundamental property of the instrument used. This means that mass-spectrometrists are often interested in not only the data points at an exact mass, but also those within the ppm error range. MS data exploration often makes requests for data in natural language like:

“All MS¹ data points with m/z values within the instrument’s ppm error of a certain molecule’s mass”

Which can again be expressed in R quite simply as:

```
msdata$MS1[mz %between% ppmppm(molecule_mass, ppm_error)]
```

2.6.4 Internals

Fundamentally, RaMS can be considered an XML parser optimized for mzML and mzXML documents. The rigorous specification and detailed documentation make it possible for a generic XML parser to efficiently extract the document data. In R, the `xml2` package provides modern parsing capabilities and is efficient in both speed and memory usage by calling C’s `libxml2` library, making it an attractive choice for this processing step. Much of RaMS’s internal code consists of a library of XPath expressions used to access specific nodes and extract the (often compressed) values. Table 2.3 below provides several examples of XPath expressions used to extract various parameters from the mzML internals:

Table 2.3: A few example parameters extracted from the mzML file and the corresponding XPath expression used to extract it.

Parameter of interest	mzML XPath expression
Fragmentation level	<code>//spectrum/cvParam[@name="ms level"]</code>
Retention time	<code>//scanList/scan/cvParam[@name="scan start time"]</code>
m/z values	<code>//binaryDataArrayList/binaryDataArray[1]/binary</code>
Intensity values	<code>//binaryDataArrayList/binaryDataArray[2]/binary</code>
Polarity (for positive mode)	<code>//spectrum/cvParam[@accession="MS:1000130"]</code>

These sample expressions illustrate the controlled vocabulary of the mzML parameters (the `cvParam` elements above) and the remarkable stability of the specification that permits optimization. While the “polarity” parameter for positive mode is the only one above that is specified via its accession number (“MS:1000130”), it’s worth noting that the other parameters also have unique accession number attributes that could be used but instead have been foregone in favor of readability.

MS data files are often highly compressed and the m/z and intensity data is typically encoded as base 64 floating point arrays. MS data extracted from the binary data array must then first be decoded from base64 to binary using the `base64enc` package, then decompressed if necessary using R’s `memDecompress` function, and finally cast to double-precision floating point values via base R’s `readBin`.

After the data has been extracted from the XML document, `RaMS` uses the `data.table` package to provide fast aggregation and returns `data.table` objects to the user. This is also the step which converts the data from a ragged array format into a tidy format, and neatly illustrates the strength of tidy data. Rather than continuing to store the data as a list-of-lists and preserving the nested data structure, this step creates separate columns for retention time (`rt`) and m/z (`mz`) values. This allows the user to perform rapid binary searches on both the retention time and m/z columns and can greatly accelerate the extraction of individual masses of interest, as is often the goal when analyzing MS data.

2.6.5 Comparison to similar packages

While many packages exist to process MS data within R, very few can be found that actually read the raw data into the R environment. The dominant package by far is `MSnbase`, which describes itself as providing “infrastructure for manipulation, processing and visualization of

mass spectrometry and proteomics data”, and is thus very similar to RaMS. MSnbase itself calls the Bioconductor package mzR to provide the C++ backend used to parse the raw XML data. Other packages include readMzXmlData and MALDIquantForeign, both developed by Sebastian Gibb and hosted on CRAN. One additional package to note is the caMassClass package that no longer exists on CRAN but code from which can be found in the CorrectOverloadedPeaks package and only parses the deprecated mzXML format. Finally, the Spectra package is under active development by the RforMassSpectrometry initiative and represents a useful comparison for other cutting-edge frameworks that will be expanded in the future (Rainer et al. 2022). However, all of these packages preserve the list-of-lists format and none produce naturally tidy representations.

This section illustrates how RaMS compares to MSnbase as the current dominant processing package and Spectra as the next iteration of MS processing. MSnbase has undergone constant revision since its inception in 2010, while Spectra has been under development since 2020. The most recent version of MSnbase as of this writing was announced in 2020 and focuses on the new “on-disk” infrastructure that loads data into memory only when needed. This new infrastructure and the legacy storage mode released in the first version of MSnbase provide useful comparisons for RaMS in terms of memory usage and speed and the Spectra package will provide a useful future-oriented comparison. As noted above, however, RaMS has different goals from either of these packages. RaMS is optimized for raw data visualization and rapid data exploration while MSnbase and Spectra are designed to provide a solid foundation for more streamlined data processing and these packages all can work neatly in concert rather than replacing each other.

To compare the different methods, ten MS files were chosen from the MassIVE dataset MSV000080030 to mimic the large-experiment processing of (Gatto et al. 2021). Methods were compared in terms of memory usage, time required to load the data into R's working memory, and the time required to subset an EIC and plot the data. Due to the differences in method optimization, we expected MSnbase to be significantly faster when loading the data, RaMS to be significantly faster during subsetting and plotting, and MSnbase to have the smallest memory footprint. The Spectra package's capabilities were less well known in advance but should represent a consistent improvement over MSnbase. These expectations were well-validated by the results shown in Figure 2.3.

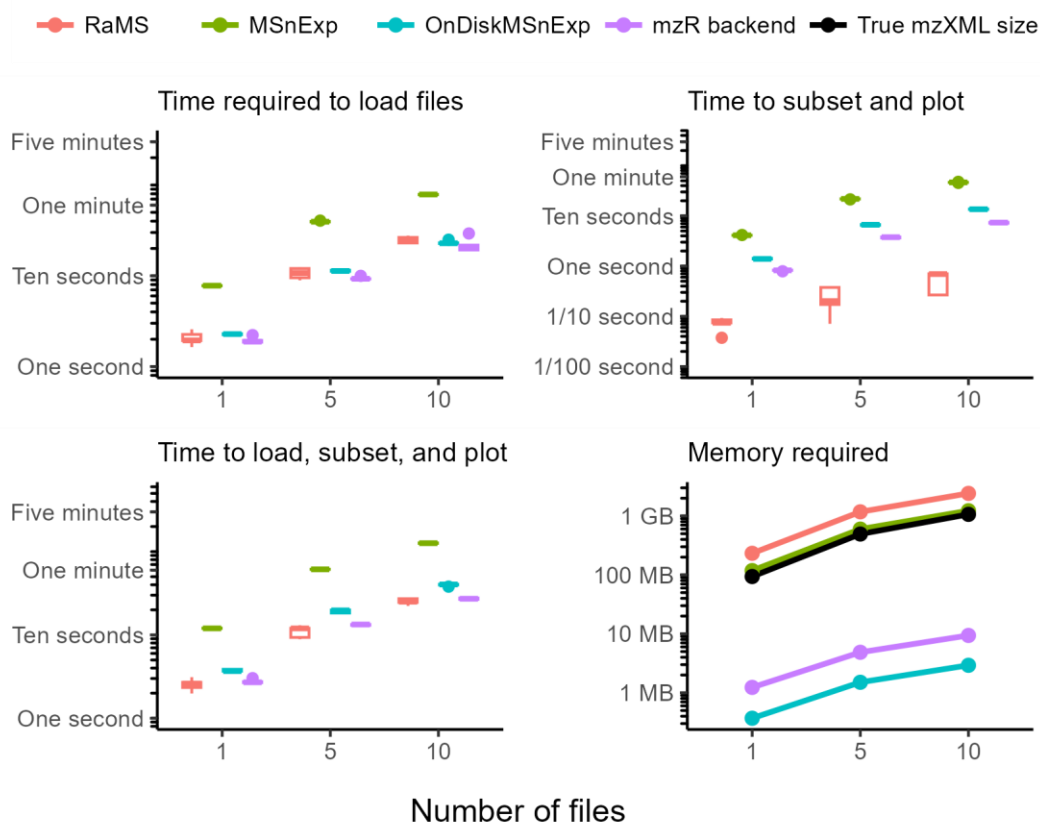


Figure 2.3: Time and memory required by RaMS compared to the MSnbase and Spectra methods across 1, 5, and 10 mzXML files. The top-left plot shows the time required to load the mzXMLs into memory (RaMS and MSnExp) or construct pointers (OnDiskMSnExp, Spectra's mzR backend) with the MSnExp object taking approximately an order of magnitude longer than the

other methods. The top-right plot shows the time required to subset the data by m/z to a single chromatogram and plot that subset after the object has already been created. The RaMS package performs this approximately an order of magnitude faster than the other packages and the Spectra package is second-fastest, with RaMS taking less than a second for up to 10 mzXMLs and the Spectra package taking between one and ten seconds depending on the number of files to be subset. The bottom-left plot shows a combination of the two plots above by timing each package as it performs the full object construction, subsets to a single chromatogram, and plots it with RaMS again the fastest among the packages. The bottom-right plot shows the memory required for each package across different numbers of files as well as the size of the original mzXML documents as a benchmark. Both RaMS and the MSnExp objects occupied more space in RAM than the original file size (RaMS occupying approximately 2x as much memory, MSnExp closer to 1.1x), while the OnDiskMSnExp and mzR backend were consistently two orders of magnitude smaller. Times were obtained by the microbenchmark package and object sizes were obtained with pryr. Note the log-scaled y-axes.

RaMS performed better than expected on the data load-time metric, taking approximately the same amount of time as the new on-disk MSnbase backend and the Spectra package and significantly less than the old in-memory method. This was surprising because while RaMS is performing the physical I/O process essentially equivalent to the creation of the MSnExp, both the OnDiskMSnExp method and the Spectra object instead create a system of pointers to the data and don't actually read the data into memory. However, the new backend begins to perform better as the number of files increases and proportional improvements are expected with even larger file quantities. The Spectra package, as expected, shows consistent improvements over both MSnbase backends.

For the subsetting and plotting metric, our expectation that RaMS would be the fastest method was validated by times approximately two orders of magnitude smaller than those obtained by MSnbase (note the log scale used in the figure). These results also validated earlier results demonstrating the superiority of the new on-disk method (Gatto et al. 2021) and the improvements in the new Spectra package. The sub-second subset and plot times of RaMS are so much smaller than the other timings recorded in this trial that RaMS essentially has a single fixed cost associated with the initial data import, making it ideal for the exploratory phase of data

analysis where files are loaded once and then multiple chromatograms may be extracted and reviewed. This design also aligns with the user's expected workflow in which data import is accepted as a time-consuming task, but subsequent analysis should be relatively seamless and instantaneous.

The greatly reduced subsetting and plotting time required by RaMS and the observation that file load times and data plotting times were approximately equal for MSnbase led to the creation of the bottom-left graph in Figure 2.3. This follow-up analysis highlights that the slightly increased file load time of RaMS combined with the very short subsetting and plotting phase is actually less than the total time required by MSnbase and Spectra to read, subset, and plot, establishing RaMS as the fastest option even if the end goal is to extract a single chromatogram. This follow-up also demonstrates the largest improvements of the new MSnbase on-disk method over the old one and the clearest improvements in Spectra.

As expected, this speed comes at a cost. RaMS has a larger memory footprint than even the old in-memory MSnExp object. While all three objects grew approximately linearly with the number of files processed, the RaMS object was approximately 2 times larger than the in-memory MSnbase object and several orders of magnitude larger than the new, on-disk version. This was expected because RaMS stores retention time and filename information redundantly in the tidy format while the list-of-lists method only stores that information once. In fact, the RaMS object size was larger than the uncompressed mzXML files themselves! However, this trade-off can be minimized through the use of RaMS's vectorized `grab_what = "EIC"` and `grab_what = "EIC_MS2"` functions that can extract a vector of masses of interest and discard the remainder of the data to free up memory for analyses where the specific ions of interest are known beforehand. The general lesson from this analysis seems to be that if the memory is available and

a quick and intuitive interaction is desired, RaMS is now the top contender. For other purposes, MSnbase or Spectra remain the obvious choices depending on expected workflow.

2.6.6 Broader interactions

RaMS is intentionally simple. By encoding MS data in a rectangular, long data format, RaMS facilitates not only R-specific development but contributes to MS analysis across languages and platforms. At the most basic level, subsets of interest can be exported as CSV files for use in any language that can read this ubiquitous format. Even users with zero programming background are familiar with Excel and other spreadsheet GUIs, so this method of export and data-sharing improves transparency by allowing anyone to open the raw data corresponding to compounds of interest.

The list-of-tables format that RaMS returns was inspired by traditional relational databases, and this provides a slightly more complex method of storing data with several advantages over CSV export. The dominant convenience of relational databases is that they can grow almost indefinitely, rather than being limited by computer memory. While existing packages perform admirably when operating on files that fit into RAM, there are few good solutions for the MS experiments that can exceed hundreds of gigabytes in size. Both batching and subset analysis face issues with systematic inter-sample variation rarely controlled for across subsets. Additionally, an external relational database can be easily appended with additional files as experiments continue to be performed, rather than demanding that all samples be run before any analysis can begin. RaMS output can be easily written to SQL databases using existing packages such as DBI and RSQLite:

```
library(DBI)
db <- dbConnect(RSQLite::SQLite(), "msdata.sqlite")
dbWriteTable(db, "MS1", msdata$MS1)
dbListTables(db)
```

```
dbGetQuery(db, "SELECT * FROM MS1 LIMIT 3")  
dbDisconnect(db)
```

Finally, with `reticulate`, R data frames can be directly coerced into Pandas DataFrames. This allows for an unprecedented degree of interaction between R and Python for MS data analysis, reducing the need for parallel development in both languages and allowing the optimal functions to be used at each step rather than the limited selection that have already been implemented in R or Python. As MS data exploration and analysis continues to grow increasingly machine-learning heavy, allowing R to interact elegantly with Python enables the best of R's extensive MS analysis history with Python's powerful interfaces to deep learning frameworks such as TensorFlow and Pytorch.

2.7 Summary

In this paper, we discussed the current paradigm of MS data analysis in R and identify an area where tidy data techniques significantly improve user experience and support increased interaction with other packages and software. We also present `RaMS` as a package that fills this gap by presenting MS data to the R user in a tidy format that can be instantly queried and plotted.

2.8 Acknowledgements

We are grateful to members of the Ingalls Lab and other labs at the University of Washington who gave invaluable feedback on early versions of this package and the philosophy behind it. Katherine Heal and Laura Carlson generated the data used in the demo files and were early adopters, and Angie Boysen and Josh Sacks provided crucial testing and application of the package. We also thank both anonymous reviewers for their insightful commentary and suggestions that improved both the manuscript and the CRAN package. This work was supported by grants from the Simons Foundation (329108, 385428, and 426570, A.E.I.).

Chapter 3: Databases Are an Effective and Efficient Method for Storage and Access of Mass-Spectrometry Data

3.1 Abstract

Current mass spectrometry (MS) data formats lack accessibility, interoperability, and performance. This study evaluates 10 recent MS file formats and readers across several exploratory MS analysis metrics and compares them to a simple database representation implemented in SQLite, DuckDB, and Parquet. We found that most existing formats severely lack the documentation required for adoption and that no existing format offers a balanced combination of speed, storage space, and simplicity. In contrast, our data storage schema improved data discovery and extraction by multiple orders of magnitude with minimal overhead. We argue that these database systems offer a performant and transparent way to store MS data for exploratory analysis while reducing technical debt and allowing mass spectrometrists to leverage recent advances in data science as our own computational complexity continues to grow.

3.2 Introduction

Mass spectrometry (MS) still lacks a performant data access format. The mzML file type (Martens et al. 2011), a result of over a decade of interlaboratory collaboration and workshopping, struggles to provide rapid computational access to the m/z and intensity pairs. This is the crucial component in nearly all mass spectrometry analysis, but mzML's text-based XML format requires time-consuming decompression performed one scan at a time. This is largely due to its preservation of the scan as the unit of transaction while the field moves increasingly away from single-scan analysis (Hannes L. Röst et al. 2014; Ting et al. 2015).

Alternative file formats aimed at improving data access are proposed nearly every year. These include direct improvements to the mzML format with indexing (Hannes L. Röst et al. 2015) and better internal encoding of the data (Bhamber et al. 2021), HDF5-based alternatives (Bhamber et al. 2021; Wilhelm et al. 2012; Bilbao et al. 2023; Tully 2020; Askenazi et al. 2017), relational databases (Shah et al. 2010; Bouyssié et al. 2015; Handy et al. 2017; Yang et al. 2022; Beagley et al. 2009), or fully custom alternatives (Römpf et al. 2011; Lu et al. 2022). Fundamentally, these alternatives exchange ease of use for access speed and/or size on disk with clever compression algorithms and modern data structures that move away from the human-readable format of the mzML. These optimized formats are inherently more difficult to understand and usually lack comprehensive documentation or examples (particularly across programming languages) making it difficult for new users to enjoy their benefits or extend their functionality. This steep learning curve, coupled with a lack of support in conversion tools such as Proteowizard's msconvert (Chambers et al. 2012), has prevented widespread adoption of these new formats despite their clear computational advantages. Such formats are also fragile in the sense that without community support, their continued development depends entirely on the original developers and easily become deprecated (as is the case with YAFMS, Shaduf, and mz5, all of whom have links in their papers that currently redirect to missing webpages). A simple, speedy, and small MS data format remains very much in demand.

Relational databases are not new for MS workflows (see references above) and compete predominantly with HDF5-based methods. Both of these systems are widely used for big data and can be applied to MS data in a plethora of ways, leading to the proliferation of implementations we see today. Both backends provide excellent universality, larger-than-memory support, and rapid access to data, but HDF5-based systems excel at self-description and

hierarchical structures (Askenazi et al. 2017) while the relational database model is optimized for multi-table queries using a consistent syntax (Codd 1970). Relational databases are increasingly seen in MS workflows for both raw and processed data, with SQLite backends now supported in the popular peakpicking software xcms (Colin A. Smith et al. 2006) via the Spectra package (Rainer et al. 2022) (though in-memory and HDF5 options are also supported) and on MetabolomicsWorkbench (Sud et al. 2016) while the development of MassQL (Jarmusch et al. 2022) demonstrates the increasing comfort that MS analysts have with the adoption of SQL.

Relational databases also have several distinct advantages over hierarchical or text-based systems, particularly in performing searches for subsets of data via indices. Importantly, this indexing differs from the byte-offset indexes that already exist in the indexed mzML and HDF5 formats because the search for a particular subset cannot be done efficiently with a byte-offset index when the m/z data is encoded (whether by numpress, zlib, or just base64), though access to a particular scan can be incredibly rapid. Additionally, data from multiple samples can be stored together in a single database table to permit queries of all dataset samples to be performed without looping through each file in turn. This differs from existing formats like mzDB, MzTree, and mzMD and thereby avoids the associated computational overhead and query complexity.

SQL databases also allow mass spectrometrists to access the continual improvements and long-term stability produced by the industries who specialize in these. While HDF5 is a common scientific data format, databases are constantly under development by industry titans deeply invested in their maintenance and optimization. Online analytical processing (OLAP) methods are particularly well suited for MS data given their optimization for read speed under the assumption of infrequent transactions, making modern systems such as DuckDB (Raasveldt and

Mühleisen 2019) or Apache’s Parquet formats highly appealing while preserving the familiar file-based serverless approach.

Our previous work showed how the ragged arrays of MS data can be converted into a tidy database table in memory (Kumler and Ingalls 2022) and we now logically extend that method into proper database storage on disk. Here, we test the hypothesis that a “vanilla” implementation of a relational database which exposes the raw m/z and intensity pairs is an intuitive and performant way of storing MS data for exploratory analysis, visualization, and quality control. We compare the time and space required to extract a representative data subset under six conditions and perform these tests on multiple databases as well as mzML and other MS data formats. Our specific questions were:

1. Is there a simple database schema that enables exploratory MS queries with basic SQL statements?
2. How expensive in time and (disk) space is it to access MS data in SQLite, DuckDB, and Parquet formats?
3. How does this cost compare to more complex MS storage formats that have been previously proposed?

3.3 Experimental section

We chose to focus on liquid-chromatography mass-spectrometry (LC-MS) data given its widespread use and fundamentally simple raw data structure as tuples consisting of retention time, m/z , and intensity, though the tidy framework here can be extended easily to other MS data (Supplemental Figure 3.5). We performed a literature search for mass-spectrometry data formats that have been published in the last 15 years and attempted to find or construct parsers for each format in Python, a popular high-level interpreted language. Each parser was written to perform three common exploratory data analysis operations on full-scan data and three common operations on MS/MS fragmentation data. Full scan queries consisted of 1) single scan extraction by scan number, 2) retention time range extraction of all scans within a specified retention time

range, and 3) chromatogram extraction, which collects the ions within a specified parts-per-million (PPM) error of a known mass. These queries generally correspond to the methods used in Bouyssié et al. (2015), which performed similar tests benchmarking the mzDB format against mz5 and an mzML parser. Note that the chromatogram extraction does not extract a precompiled chromatogram of the sort commonly found at the end of mzML files or as a result of SIM/PRM analysis but instead refers to sifting through the raw data for data tuples with an m/z value between specified bounds. MS/MS queries involved extracting three relevant subsets, consisting of 1) a single scan extraction by scan number similar to that of the full scan, 2) extraction of all the fragments associated with a precursor m/z within a given PPM, and 3) extraction of all fragments with m/z values within a given PPM.

We explored the available documentation on PyPI and Github for each mass spectrometry data format and either identified existing functions and packages that would perform the above queries or wrote our own functions if necessary.

3.3.1 Mass-spectrometry files and software used

We browsed Metabolights (Haug et al. 2019) for suitable LC-MS datasets, looking for studies that included 100+ gigabytes of data from both full scan and MS/MS analysis. We were also restricted to the Thermo Scientific .raw file format, as it was the most widely supported by alternative MS storage methods. We also excluded polarity-switching data as it is unclear whether all converters would be able to separate scans based on polarity.

Files were downloaded as .raw. mzML, mz5, and mzMLb were all natively supported by Proteowizard's `msconvert` software (version 3.0.25009) while MZA (v1.24.11.16) and mzDB (v0.9.10_build20170802) had separate extensions to this executable enabling their own conversion. MzTree and mzMD were converted via their GUI which did not have release or

versioning information available but were downloaded from Github (<https://github.com/optimusmoose/MZTree> and <https://github.com/ym9837/mzMD-java>, respectively) and built via Maven (v3.9.9) for Java (v21.0.6). SQL databases were built using Python 3.11.11 with SQLite (v3.48.0) via the Python sqlite3 package (v2.6.0), DuckDB via the duckdb package (v1.1.3), and Parquet files via the pyarrow package (v19.0.0).

mzML access was done with Python's pyteomics package (v4.7.5), the pymzml package (v2.5.10), and the pyopenms package (3.0.0.dev20230306). MZA files were accessed via the mzapy library (v1.8.dev4 from the no_full_mz_array branch on Github) and via custom code built around the h5py package (3.12.1). Custom parsers were required for mzDB, mz5, MzTree, and mzMD.

3.3.2 Database schema

The “vanilla” database style proposed here abandons a 1:1 representation of the original vendor-specific file. This decision was made after discussion with a wide variety of experts, all of whom preserved the original MS files even after conversion to another file type, indicating that a highly-performant addition is more important than direct replacement. Here, we map MS concepts (retention time, drift time, spatial coordinate, m/z , intensity, etc.) directly to database fields to make downstream processing as intuitive as possible. Metadata is stored separately in `file_info` and `scan_info` tables that are linked by filename and scan number (Figure 3.1). We do not force compression of any of these fields because we find that decoding compressed data is both a slow and unintuitive step, though automatic compression is supplied by the DuckDB and Parquet file types.



Database
.sqlite
.duckdb

Table: file_info			
filename	n_scans	timestamp	Etc.
Source file	Scan count	Timestamp	...
Smp_A	1445	2021-08-23 04:30	...
Smp_B	1489	2021-08-23 05:25	...
...additional samples...			
Smp_Y	1388	2021-08-23 22:15	...
Smp_Z	1390	2021-08-23 23:10	...

Table: scan_info					
filename	scan_num	ms_level	rt	TIC	Etc.
Source file	Scan number	MS level	Retention time	Ion current	...
Smp_A	1	1	0.10	2331809	...
Smp_A	2	2	0.12	820427	...
...many additional scans...					
Smp_Z	1389	2	22.43	9555703	...
Smp_Z	1390	2	22.45	1892338	...

Table: MS1				
filename	scan_num	rt	mz	int
Source file	Scan number	Retention time	m/z ratio	Intensity
Smp_A	1	0.10	60.0452	6618
Smp_A	1	0.10	60.0532	2657
...millions of additional entries...				
Smp_Z	1385	22.35	60.0456	158084
Smp_Z	1385	22.35	60.0531	4673

Table: MS2						
filename	scan_num	prescan	rt	fragmz	premsz	int
Source file	Scan number	Precursor scan	Retention time	Fragment m/z	Precursor m/z	Intensity
Smp_A	2	1	0.12	51.0238	241.0894	36104
Smp_A	2	1	0.12	53.0394	241.0894	243165
...millions of additional entries...						
Smp_Z	1390	1385	22.45	52.0186	185.1932	28371
Smp_Z	1390	1385	22.45	57.0923	185.1932	129604

Ion chromatogram extraction: `SELECT * FROM MS1 WHERE mz BETWEEN min AND max`

Retention time range subset: `SELECT * FROM MS1 WHERE rt BETWEEN min AND max`

Fragment search: `SELECT * FROM MS2 WHERE fragmz BETWEEN min AND max`

Precursor search: `SELECT * FROM MS2 WHERE premsz BETWEEN min AND max`

Figure 3.1: Database schema for an example MS/MS dataset showing the organization of mass-spectrometry data into tables. Fields of interest are easily queryable with simple SQL commands as shown in the table at bottom.

3.3.3 Time and space testing

We randomly sampled a single file out of the full dataset for comparison across metrics and file formats (20220923_LEAP-POS_QC04). We sampled 100 random scan numbers using SQLite's `ORDER BY RANDOM()` function and pulled out the largest ions for chromatogram extraction using a 10 ppm mass range window. Retention time ranges were the highest-intensity retention time of each ion chromatogram plus or minus one minute. Similarly, the largest fragments by intensity were used for the MS/MS metrics with a 10 ppm mass range exclusion window.

Timing was performed via Python's `timeit` library and the `timeit.repeat` function, with the various file formats as the innermost loop to ensure bias over time was distributed equally among function calls. File sizes were estimated using Python's `os` library with `os.path.getsize`. We did not exhaustively monitor memory usage, though our heuristic exploration of the timing scripts did not ever indicate that memory was a constraint.

Timing data was obtained on an Intel Xeon CPU with two X5650 (@2.67 GHz) processors and 24 total cores running Windows 10 Pro (64-bit version). 96 gigabytes of RAM (DDR3 @ 1333 MHz) were available and a solid-state drive was used for disk storage.

3.4 Results

We settled on a large dataset of gut microbiota LC-MS files published in Portlock et al. (2025) and available on Metabolights under accession number MTBLS10066.

3.4.1 All existing MS data formats demand a high level of domain knowledge

We were able to obtain or write parsers for seven different existing mass spectrometry (MS) data formats: `mzML`, `mzMLb`, `mz5`, `mzDB`, `MZA`, `MzTree`, and `mzMD`. Multiple Python packages exist for the `mzML` data format so we used each of the three dominant packages (`pyteomics`, `pyOpenMS`, and `pymzml`) and compared their timing results as well. We failed to produce parsers for the `YAFMS` and `Shaduf` file types due to complete deprecation (links to these no longer exist), the `toffee` file type due to its application solely to time-of-flight (TOF) data-independent acquisition (DIA) data, the `Aird` file type due to its current deprecation in Python and C#, and the `UIMF` format due to a complete lack of interface documentation.

3.4.1.1 File conversion support varied enormously

Conversion from the initial Thermo `.raw` file type to the open-source `.mzML` format was seamlessly performed by Proteowizard's `msconvert` library. Similarly, Proteowizard support for the `.mz5` and `.mzMLb` file types made their conversion trivial.

mzDB and MZA both had extensive documentation, providing self-contained extensions to `msconvert` for ease of conversion. However, both converters provide limited coverage, with mzDB missing support for Waters and Agilent .d files while MZA currently lacks support for AB Sciex .wiff and Bruker .baf files. Both converters are only available via binary executable (.exe), restricting their use to Windows platforms. Additionally, both parsers appear to be unable to separate scans from a polarity-switching experiment or support any of the other filters available natively in `msconvert`, as additional arguments passed to the executable throw errors instead of being passed along to the original software.

MzTree and its derivative, mzMD, provided significantly less documentation about the conversion process than the other file types. This documentation consisted solely of the README available in the associated Github repositories and their installation and deployment required rebuilding the Java applet, of which the bare-bones instructions make several assumptions about the user's PATH environmental variable. In the case of mzMD, no documentation for installation and build was provided and this instead needed to be deduced from MzTree. Additionally, we ran into issues with hardware acceleration once the GUI was launched that required extensive debugging. The GUI conversion, however, is straightforward once the app is correctly compiled and launched, albeit requiring a manual entry of a single file at a time with no apparent batch processing available.

The Aird file type was straightforward to convert on Windows via the executable available on Github (v6.0.0) but was not available for other operating systems, much like MZA and mzDB. The Python package designed to allow an interface to the file type has been deprecated and we were unable to install or use it and were unable to reverse-engineer the file type sufficiently to compare it here. The UIMF file type from the Pacific Northwest National Lab

(PNNL) provided documentation exclusively in the form of C# commands and did not supply instructions for file conversion, making it unclear what input formats were supported. The toffee format provided no documentation for conversion from other formats and was restricted to time-of-flight (TOF) data independent acquisition (DIA) MS data. Thus, we were unable to directly compare any of these three file types to the others.

3.4.1.2 Universal lack of support for the six relevant queries

Despite the relative simplicity and relevance of our queries, none of the available mass spectrometry (MS) formats had existing functions or examples of all six queries. The mzML file type had the most extensive coverage but documentation and prebuilt functionality was still sparse. The pyteomics package provides four “combined examples” that focus on the spectrum visualization and annotation common to proteomics research but provide minimal guidance about chromatogram or retention time range extraction. Pyteomics also provides native support for the mzMLb file type and was the only one of the three Python packages to do so, deserving praise for the minimal disruption that mzMLb files placed on existing pipelines if they were to switch from mzML to mzMLb. The pyopenms package provides similarly extensive documentation for proteomic and scan-based analysis but again lacks information about subsetting in the retention time direction, though the existence of an undocumented parser (`get2DPeakDataLong`) provides a simple way to do this for MS¹ data. Additionally, pyOpenMS required installing an old version of the package (3.0.0), Python itself (3.11) and the numpy package (<2.0) due to more recent builds requiring AVX support which was unavailable on our hardware. Pymzml is intentionally a lightweight parser focused exclusively on reading mzML files but does not supply any functions for the queries other than scan extraction by number and the “Spectrum and Chromatogram” documentation module was empty at the time of writing (February 2025).

mz5's documentation was sparse, especially for one of the earliest mzML formats with support from Proteowizard. The original paper (Wilhelm et al. 2012) contains links to a website (<https://software.steenlab.org/mz5>) which currently returns an HTTP error 500. A Python library (pymz5) exists but requires an old version of Python (2.7 or 3.2), has not been updated in 12 years, and is predominately a simple fork of h5py (Collette et al. 2017) with three mz5-specific commits on top. Most problematically, we were unable to determine how mz5 stores precursor m/z ratios, making the fragment and precursor searches impossible. This was largely due to the variable-length nested compound structures mz5 that are not supported in all APIs, e.g. Java (Bhamber et al. 2021).

mzDB access was hamstrung by several issues, primarily the outdated repository that implies Python and R support via a port from Rust but was unavailable at the time of development, though we are grateful for the responsive developer who notified us that this implementation was not feature-complete. This required that we deduce the SQLite BLOB type compression format from scratch when writing a parser and spend extensive time reading through the documentation to determine how best to link the various tables provided in the mzDB file. Scan metadata in this file type is stored as raw XML strings, producing the worst of both worlds in requiring both SQLite knowledge in their extraction and XML processing to obtain the necessary information. Additionally, mzDB seems to dump all MS/MS data into a single bounding box, meaning that we were unable to use the scheme to avoid parsing every MS/MS spectrum when performing precursor and fragment searches.

MZA provides a complementary Python package, mzapy, for access to MZA files. Here again we ran into several issues with its installation and use stemming largely from the deployed package requiring TOF bins for parsing, though a separate Github branch provides a workaround

and the rapid developer response was appreciated. The `mzapy` package provides a clear example of chromatogram extraction as well as a method for retention time range extraction, though there exists no clear function for the extraction of a single spectrum by scan number despite the internal file structure being highly optimized for this purpose. `mzapy` also provides good support for ion mobility extraction but fails to index MS/MS information or provide any clear way to extract fragments by m/z or precursor.

MzTree and mzMD provide a slightly strange interface to MS data, requiring a separate Java server that can then be queried via an HTTP API. For users without prior knowledge of HTTP request methods or exposure to programming APIs, the README is entirely unhelpful because it simply documents the API's endpoints and provides no complete query strings as examples to guide the user. This combination of GUI server and command-line HTTP request inverts the typical paradigm of GUI for exploration and command line for construction to convoluted effect, though the structure of the data returned by the server is impressively simple. More problematically for this analysis, the API provides no apparent way to access MS/MS data or query the files by scan number, with only RT and m/z bounds controlling the subset of data extracted. Finally, the GUI provides no way to open multiple files simultaneously or iterate through files programmatically and instead requiring point-and-click interaction with the GUI each time a file is opened or closed, preventing us from making reasonable comparisons in tests requiring multiple files.

3.4.2 SQL-based parsers were simple to write and use

We then used custom code to convert the mzML files into SQLite and DuckDB databases using a simple schema for full scan (MS^1) and MS/MS (MS^2) data. The MS^1 table consisted exclusively of fields for filename, scan index, retention time, m/z ratio, and intensity. The MS^2

table consisted of the same fields except that the m/z column was separated into precursor and fragment m/z . Although we did not extend these databases to include the metadata associated with each file and scan, the logical framework could be easily extended in future work and the metadata typically represents a small fraction of the total space within the file, allowing us to make reasonable comparisons about file size between the databases and the metadata-rich other file types. We also converted each file's MS¹ and MS² table into Parquet representations for comparison using the same field/column schema.

We found that the documentation for SQLite, DuckDB, and Parquet file formats in Python far exceeded the documentation available for any mzML parser. This is unsurprising given that these file formats are used widely outside of MS research and are developed and maintained by dedicated teams. Additionally, the use of a consistent SQL syntax for table creation and insertion meant that the same code could be used to write to both SQLite and DuckDB, as well as any other databases supported in Python. The use of packages such as SQLAlchemy could be used to additionally streamline this process to any additional database by simply swapping in a new database engine.

Querying the MS¹ and MS² tables was also very straightforward. After establishing a connection to the database, the six queries could be asked using nearly human-readable SQL syntax. Requesting the thousandth MS¹ scan by number consisted simply of `SELECT * FROM MS1 WHERE id = 1000` passed along to the `pandas.read_sql_query` function. More complicated queries such as retention time range (`SELECT * FROM MS1 WHERE rt BETWEEN 6 AND 8`) and a precursor mass search (`SELECT * FROM MS2 WHERE premz BETWEEN 118.086 AND 118.087`) were similarly intuitive.

3.4.3 Time and space requirements for a single DDA file across formats

3.4.3.1 *Spectrum extraction*

The simplest and most abundantly documented query was the extraction of a single spectrum. In many ways, this is the fundamental unit of mass spectrometry and thus many formats are highly optimized for its extraction into manipulable data (Figure 3.2A and 3.2D). Here, we found that the mzML and mzMLb file types were consistently the slowest to parse and required multiple seconds, likely highlighting inefficiencies in the `pyteomics` package used to parse both file types. `pyopenms` also struggled to open and extract a specific scan, requiring several seconds due in large part to the expensive initiation function, after which requests were orders of magnitude faster (Supplemental Figure 3.6). It is also worth noting that while both of these packages provided rapid extraction of a *random* spectrum, a significant overhead was introduced by needing to scan through the file to find a *specific* spectrum by scan number. Scans are not always consecutive and no metadata was obviously available that would have allowed using the index directly to a specific scan number.

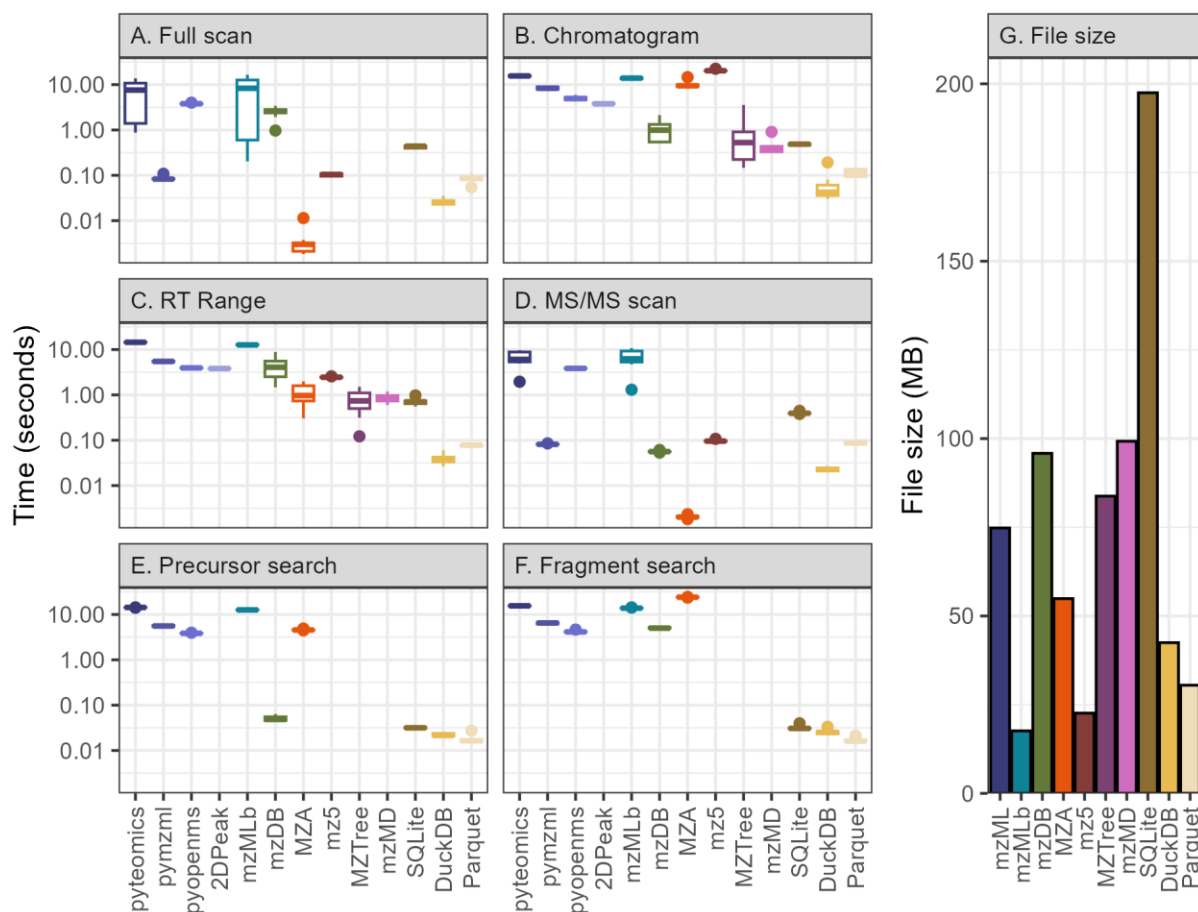


Figure 3.2: Query time for the six data extraction methods and the associated file sizes for all 13 methods explored in this paper. The left six panels show boxplots representing the time required in seconds to extract a full scan spectrum (A), an ion chromatogram (B), all data within a retention time range (C), an MS/MS scan (D), the fragments of a specified precursor (E), and all precursors with a specified fragment (F). The error in the boxplot is composed of timing information for 10 repeated queries, each of a different target scan number, retention time (RT), or m/z . The right panel (G) shows a barplot of the size on disk in megabytes (MB) occupied by each file type.

The pymzml package was able to extract both MS^1 and MS^2 spectra from the mzML file nearly two orders of magnitude faster than the other mzML parsers, largely due to its use of naming the scans by their number and thus avoiding the expensive scan number extraction step. mzDB had the only notable difference between MS^1 and MS^2 scans, performing slightly better than pyteomics and pyopenms methods for MS^1 data and significantly better for MS^2 data,

placing it approximately on par with pymzml in taking about a tenth of a second. The simple database methods (SQLite, DuckDB, and Parquet) also fell in this ~0.1 second range, with SQLite performing most poorly and DuckDB ~10x faster. Finally, both mz5 and MZA were an additional order of magnitude faster than any other method, returning the data within the spectrum in thousandths of a second. This shows the power of the HDF5 file system for data access when its location within the file is known in advance.

3.4.3.2 Chromatogram extraction and subsetting by retention time range

Ion chromatogram extraction and retention time range subsets were a key metric for us, corresponding to essential tasks in chromatographic peakpicking and adduct, isotope, and in-source fragment detection (Figure 3.2B and 3.2C). EIC query times here were universally slower than those for a single spectrum extraction, reflecting the way in which a scan-based file type is sub-optimal for chromatogram extraction because each scan must be parsed to find data within a given m/z range. MZA and mz5 particularly suffered, with this query type entirely negating the advantages of the HDF5 file structure.

MzTree and mzMD are both file types optimized exclusively for chromatogram extraction and performed very well on the EIC metric and were two orders of magnitude faster than those parsing mzMLs, with mzMD surprisingly less performant than the older MzTree file type it was based on. However, we also note that both Java-based applications have a slow initial file load step that must be done through a GUI and therefore could not be counted in the timing comparison, the inclusion of which would likely mitigate any advantage for a single chromatogram extraction. The mzDB file type is also optimized for chromatogram extraction and was an order of magnitude faster than the other existing file types for which all queries could be run (MzTree and mzMD do not provide interfaces for spectrum extraction or MS/MS data).

The SQLite, DuckDB, and Parquet formats were just as speedy as mzMD and MzTree with SQLite taking half a second, Parquet requiring a tenth of a second, and DuckDB reaching query times of hundredths of a second, far outstripping the seconds or even minutes typically expected of this task and resulting in a functionally instantaneous interaction for the user.

Retention time range extraction times were an average of the single-spectrum extraction and the chromatogram extraction times across the board, potentially hinting at a major predictive factor in timing estimation being the total amount of scan parsing required.

3.4.3.3 MS/MS precursor and fragment search

We also investigated the efficacy of the various MS data formats for MS/MS data and found that support for fragmentation data searches was lacking or absent from the documentation and exposed functionality of each of these file types, requiring custom implementations every time. Despite both precursor searches (where all the precursors of a given fragment are found) and fragment searches (where all the fragments of a given precursor are found) representing intuitive and useful methods of MS/MS data processing, these timings were consistently among the slowest of the six query types for the non-database methods (Figure 3.2E and 3.2F).

All existing MS data types required multiple seconds to perform a single fragment search (Figure 3.2F), representing a significant bottleneck for any downstream analysis requiring the data associated with the fragments of a given precursor. The SQL-based parsers, on the other hand, all took fractions of a second and consistently returned the relevant data hundreds of times more quickly than existing methods. The same was true for a precursor search across all methods aside from mzDB (Figure 3.2E), which benefited significantly from constructing a single bounding box for all MS/MS information that requires a single decoding into computer memory, though this strategy will fail for any file with sufficiently large MS/MS data.

3.4.3.4 File sizes

File size is another important constraint on the efficacy of various MS formats. We measured the size on disk of each of the file types and found that they varied by approximately an order of magnitude, with HDF-based file types hovering around one-third the size of the mzML (mzML size = 75 megabytes (MB), mzMLb = 18 MB, mz5 = 23 MB) while mzDB and mzMD were larger (95 MB and 99 MB, respectively). The SQLite object was the largest on disk of all the file types, nearly tripling the mzML's size at 197 MB, while DuckDB improved on MZA at two-thirds of the mzML (42 MB and 55 MB, respectively) and Parquet improved slightly upon that again (30 MB total) with its columnar-based storage format (Figure 3.2G).

However, these comparisons are not perfect because not all files store exactly the same data. MzTree and mzMD appear to entirely lack the MS/MS information in the sample DDA file, representing a potentially significant size reduction that's difficult to estimate though the extraction of the same file via `msconvert` containing only MS¹ scans was 58 MB, a 23% size reduction. The SQLite, DuckDB, and Parquet formats also lack the extensive scan and file metadata that's present in the other file types, though it is difficult to estimate the fraction of disk space allocated for this (and which will depend upon the precise definition of metadata).

3.4.4 Timings for multiple chromatograms

The single-file, single-metric case discussed above and shown in Figure 3.2 is largely a worst-case scenario for many MS data systems that have a slow initial setup step to make downstream analysis faster. To compare these systems more fairly to our database schema, we also tested timings across multiple chromatograms. In each case, this was implemented as a for loop iterating over an increasing number of chromatograms corresponding to the largest intensity ions in the file (Figure 3.3).

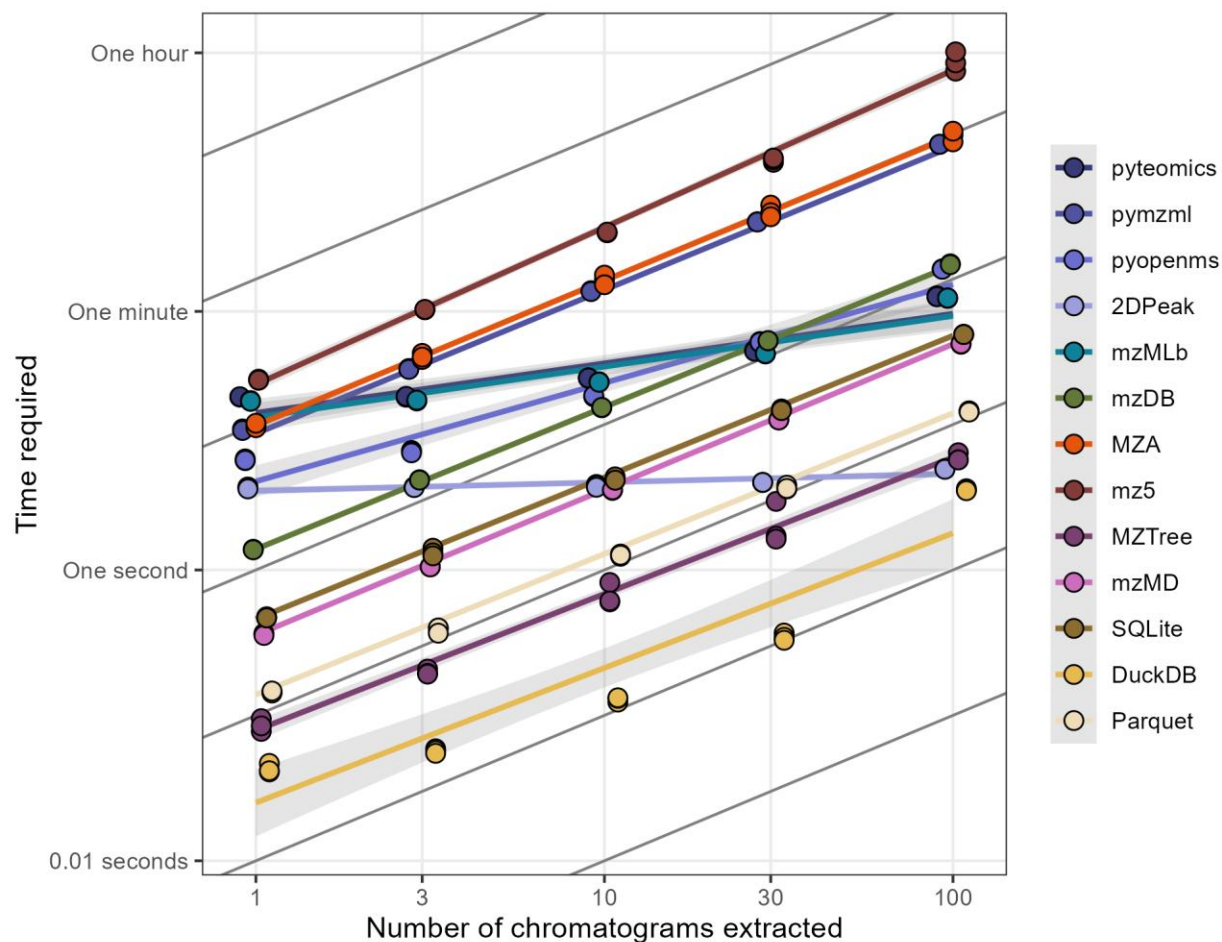


Figure 3.3: Scatter plot of the time required to extract multiple chromatograms using various methods on logarithmic axes. Best-fit linear models have been added for each method are shown behind triplicate timing measurements. Transparent intervals around each best-fit line show a single standard error of the mean. Chromatograms correspond to the largest intensity ions in the file. 1:1 lines have been added in black behind the data for comparison.

Most methods were linear extrapolations of the single chromatogram numbers shown above as expected from a simple for loop, with notable exceptions for pyteomics (and thus the mzMLb format), the 2D peak method of pyopenms, and DuckDB (Figure 3.3). Pyteomics and pyopenms both had significant overhead upon initial load resulting in faster subsequent queries that performed better than predicted from a 1:1 extrapolation, with pyopenms matching SQLite's speed after 10 chromatograms and Parquet's speed after 30. The methods that had a best-fit linear slope less than 1:1 also all had exponential fits, with performance at high chromatogram

number worse than expected from a predicted fit to the timings for 1 and 3 chromatogram extractions (Supplemental Figure 3.7).

We also explored whether database queries could be improved via the use of either a unified query (single SQL statement with multiple OR clauses for each ion's m/z range) or a non-equi join between a peak table with m/z minimum and maximum columns (Supplemental Figure 3.8). SQLite and Parquet performed ~3-5 times faster with the unified query than with the loop method despite the necessity of an additional processing step for the looped query to correctly assign each data point to its original peak information. The opposite was true for DuckDB likely due to its optimized reader, with the unified query consistently outperformed by the non-equi join when 100 chromatograms were extracted.

3.4.5 Database optimization via indices/ordering and multi-file constructions

Databases also provide multiple ways to optimize queries. SQLite allows the construction of indices for a field within a table that then speeds up queries at the cost of additional disk space. Alternatively, DuckDB and Parquet files rely predominantly on the data order when it's written to disk and use their sophisticated row group methodology when subsetting.

We found that SQLite queries benefitted significantly from the construction of an index on the m/z column when extracting chromatograms, improving lookup times by an order of magnitude (dropping from 0.5 seconds to 0.03 seconds, Figure 3.4). However, because the SQLite index is stored on disk alongside the data, this improvement also increased the file size by 24%. Parquet files had the smallest improvement upon data ordering and required 50% more space, likely due to the reordering resulting in worse compression in other columns such as filename or retention time. DuckDB also improved significantly with ordered data but to a

smaller degree and in contrast to SQLite or Parquet sometimes actually decreased in size when ordered.

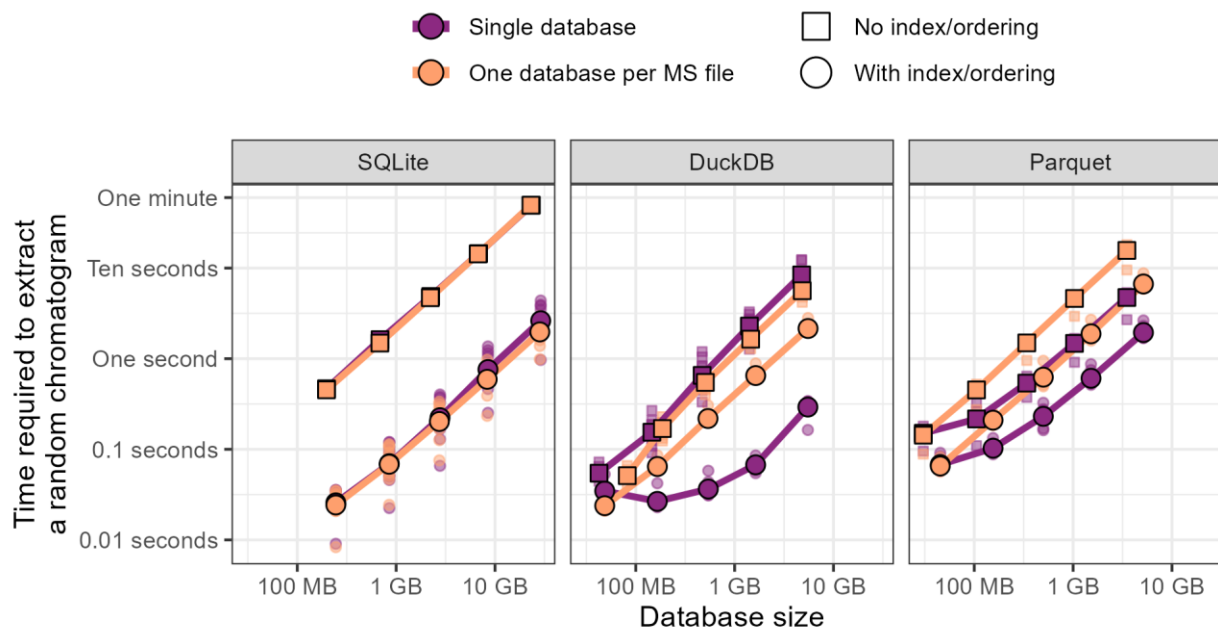


Figure 3.4: Time required to extract an ion chromatogram from multiple files plotted against the size of the data, broken down by the type of database used (SQLite, DuckDB, or Parquet). Points correspond to a random subset of 1, 3, 10, 30, and 100 files, respectively. Colors specify whether the data was stored as a single consolidated database (purple) or with a single database per file (orange) and the shape of the point denotes whether the database was unstructured (squares) or indexed/ordered by m/z (SQLite has indexes, DuckDB and Parquet benefit from ordering). Ten replicates of each query were performed and are shown transparently behind the mean values connected with lines.

These improvements also persisted when multiple files were stored in a single database.

We built databases consisting of between 1 and 100 individual MS files and tested the time required to extract ion chromatograms from each after an index was constructed (Figure 3.4). DuckDB was consistently the fastest ion chromatogram extraction method, with query times around 0.03 seconds for a single file and 1 second for one hundred files. SQLite had much higher variance and slower extraction times with datasets consisting of more than one file, typically an order of magnitude slower than DuckDB, while Parquet fell between the two. Importantly, only

DuckDB had a slope much less than one. This is what would be expected if the database was performing a simple binary search on the index, with an expected time efficiency of $O(\log(\# \text{ of files}))$. However, DuckDB's performance degrades at larger database sizes and approaches a 1:1 slope, possibly due to the time required to read large amounts of data into memory after it's found. We additionally compared these values to the timings obtained from converting each file into its own database and looping over each of those to confirm the linearity of that response (Figure 3.4).

3.5 Discussion

As the gap between data scientist and mass spectrometrists continues to narrow, mass spectrometry (MS) data formats should facilitate this convergence. Instead, MS software remains relatively opaque. Documentation is sparse and data structures are complex, resulting in a landscape that is essentially restricted to the original developer's intent. A particular pain point is the way in which MS data is stored because current methods must make trade-offs between simplicity, size, and speed. When we explored the wide range of MS file readers, we found every method had flaws that interfered with widespread adoption. The mzML file type appears to represent low-hanging fruit, with its XML-based structure that sacrifices speed and size in favor of clarity, but no alternative has yet reached a large audience of active users. Fifteen years of active competition continue to favor the highly explicit format, likely because users are leery of incomprehensible alternatives with no guarantee of continued maintenance.

Certain methods are clear winners for individual use cases but all of the existing formats failed to perform well at the full suite of exploratory data analysis tasks we attempted. Scan extraction is perhaps the most widely used query, especially for proteomics, and as a result has been extensively optimized. Here, MZA was blazingly fast thanks to the decision to index in the

HDF5 file by scan number. The mzML files used here did have a precompiled index that should have made scan extraction highly efficient but this appeared to be mitigated for pyopenms and pyteomics due to their long initialization times, though pymzml performed very well at this task and the other two mzML methods were much faster after initialization (Supplemental Figure 3.6).

However, performant scan-based methods struggled significantly with chromatogram extraction because these axes are inherently orthogonal to each other. mzDB was competitive with the specifically optimized MzTree and mzMD formats, illustrating its success as an axis-agnostic structure. We were especially impressed with the 2DPeak method in pyopenms when extracting multiple chromatograms, as it had essentially a single setup time cost after which any number of chromatograms could be extracted for free and can therefore be highly recommended for visualization applications. Chromatograms (and retention time range queries) are of course only relevant for chromatography-based workflows but this type of analysis has become increasingly popular, making ion extraction increasingly important. None of the existing MS formats we tested performed very well on MS/MS data, despite the growing availability of fragmentation data, though our use-case is oriented more towards exploration instead of comprehensive analysis.

Finally, a complexity penalty must be noted for formats and packages requiring complex installation procedures. Pyopenms appears to be the worst offender here, with its bindings to the OpenMS C++ libraries requiring us to step back to Python 3.11 and numpy 1.26 to successfully access the data on our setup and required direct input from the maintainers. The Java applications for MzTree and, more egregiously mzMD, provided essentially zero documentation on installation consisting of a single README file without an intact HTTP request example and

have not seen updates in years. Similarly, mzDB files were difficult to parse due to its opaque SQLite schema and use of the BLOB encoding type that again lacked examples or documentation outside of Java and had to be deduced iteratively. Of course, these methods eventually provided enough information that they could be parsed unlike Aird, toffee, and UIMF.

3.5.0.1 Timing comparison to existing literature

Novel formats are typically proposed with timing and sizing information, but the inconsistency of what's being queried makes it difficult to directly compare across the literature. However, we mostly obtained results in line with those reported elsewhere where other intercomparisons have been performed and report here the widest set of comparisons between MS formats to our knowledge.

Wilhelm et al. (2012) performed comparisons between mz5 and mzML with results indicating that mz5 was three times faster than mzML parsers. Their values of 0.16 seconds per million m/z /intensity arrays correspond to an estimated query time of 0.13 milliseconds which is faster than our measured 3 milliseconds, though our mz5 file was approximately four times smaller (20% mzML size) instead of half the mzML size.

Bouyssié et al. (2015) also found that mz5 was about 80% smaller than the mzML and on par with their mzDB format. They reported query times of around 30 seconds for a wide (5 Da) ion chromatogram extraction and found that mz5 was about 40 times slower and that mzML was 200 times slower, in contrast to our parser which was 10 times slower for mz5 - possibly due to the large EIC width. They also were able to report mzDB scan queries on par with mz5 which we were able to replicate if the pyopenms or pyteomics libraries were used for full scan queries and pymzml for MS/MS.

MzTree (Handy et al. 2017) compared their SQLite-based system to mzML, mz5, and mzDB and reported high numbers for both EIC and RT range queries from mzML (4-1000

seconds) that were on par with the values we observed here. Their mzDB and mz5 results were unexpectedly comparable to each other at about 0.5 seconds per random EIC query, with the mzDB values equal to ours but the mz5 values much lower than our parser was able to obtain. Our disk size measurements also corresponded well with their size estimate of MzTree at approximately twice the size of the mzML, a surprising result given that our mzML file contained MS/MS information and theirs appearing to be full-scan only. Their mz5 files were much larger (80% mzML size instead of our 20%) and their mzDB much smaller (20% mzML size instead of our 110%).

The mzMD format (Yang et al. 2022) appears to be a thin wrapper around the MzTree format that applies a different philosophy for data subsetting and summarization. They report EIC queries in the 50 millisecond range, very similar to the values we obtained for the mzMD file type. They also estimate file size to be approximately 28 bytes per *m/z*/intensity tuple for a total size of 100MB in our test file which agrees reasonably well with our 72MB actual measurement. Their comparison to MzTree also agrees with our results as they report slightly larger disk usage and slightly better performance.

The mzMLb group (Bhamber et al. 2021) reports only info for spectrum access at approximately 15ms/scan which agrees with our 5-100ms/scan estimates only if the data is loaded ahead of time. They perform extensive comparisons to mzML at varying compression methods and levels but we stuck with the default options of 1024 KB chunk sizes for the mzMLb file and zlib-only compression for the mzML. This resulted in timing values very similar to those they reported when using the pyteomics library for access. They also compared to the mz5 file type and we are able to validate their results of the mzML+zlib occupying significantly more

space, though our mz5 parser outperformed theirs for full scan and MS/MS data by two orders of magnitude in time.

We were also unable to test several other recent and promising formats. Aird does not report full query times for any of the metrics reviewed here, though they claim their StackZDPD algorithm (Wang et al. 2022) can improve decompression speed by three times and that the file size is 54% of the vendor file.(Lu et al. 2022) Similarly, the toffee format for time-of-flight DIA data reports sizes about equal to vendor or 60% of centroid mzML + numpress with query speeds 4 times faster for scans (spectrum-centric, 2 seconds for mzML, 0.5 for toffee) and 100 times faster for chromatograms (peptide-centric, 168 seconds for mzML, 1.8 for toffee).(Tully 2020) The Unified Ion Mobility Format (UIMF) (Beagley et al. 2009) from Pacific Northwest National Laboratory format did not report direct comparisons to any of the available formats and thus we must remain unclear on its performance capabilities.

3.5.0.2 Fundamental inefficiencies in existing mass-spectrometry formats

We identified several fundamental inefficiencies when writing the parsers. First, scan metadata that was encoded within a scan instead of in a separate unit required looping over each scan to see whether it contained the information requested. Scan number, MS level, and retention time were all necessary bits of information that could be included in a file header or footer to relate the three to the data location within the file and allow index use instead of looping over every scan. Second, needing to decode or parse a compressed m/z /intensity array in each scan introduced an additional overhead that was especially punishing during ion chromatogram extraction and MS/MS search. While the m/z and intensity tuples are an obvious candidate for data compression, this penalty should be of significant concern to engineers. Third, looping over files is inherently slow and introduces additional complexity relative to a single unified database that encodes filename or sample ID as an additional column. A particular strength of the database

system we propose is its inherent support for multi-file systems, while all other methods require looping over files.

The problems above highlight an important distinction between data *access* and data *search* that has been largely overlooked in our opinion. While HDF5 files or scan indexes excel at improving data access, they assume that the location of the data is known in advance and can be skipped to via bitwise offsets. If a search is required, however, this advantage is fully negated because each bit of information must be queried anyway. Finally, we must note that scan number is not inherently a useful bit of information. While we included it in our extraction metrics, it is entirely unclear when the scan number itself would be known in isolation. Additionally, this method often confuses the scan number with the scan's indexes in the data structure. Scan number is not always consecutive (e.g. during polarity switching, multi-experiment samples, or if any filtering is performed during processing), so even if the first or second item in the structure can be queried speedily this is no guarantee that the item will contain the information of interest.

3.5.0.3 Leveraging robust, future-oriented software development with SQL

The proliferation of MS data storage formats and access algorithms illustrates the general dissatisfaction with existing alternatives to the vendor file or mzML. Formats that are faster to query or smaller on disk tend to be significantly more opaque, and those optimized for a particular method often fail to perform well on other metrics. This complexity is generally expected as optimization tends to require more complex data structures and assumptions about its use but it is not required if the complexity is outsourced to a robust and growing framework such as structured query language (SQL).

SQL is widely used for data processing outside of mass-spectrometry, though its adoption is increasing in recent years. Efforts like mzDB, the Pacific Marine Environment Laboratory's UIMF format (Beagley et al. 2009), and the internals of MzTree hint at SQL's suitability for MS

data storage. SQL backends for the next-generation R processing package Spectra now exist (Rainer et al. 2022) and the development of MassQL (Jarmusch et al. 2022) indicates a growing comfort with SQL syntax for downstream processing, though the language itself strives for human readability in simple queries. The searching and subsetting inherent to MS data exploration represent very simple queries in database space, agnostic to high-level programming language and rarely requiring more than a single line of code. Additionally, the extensive documentation that exists across the internet means that large language models such as ChatGPT are easily able to translate queries for those unfamiliar with SQL's syntax.

Just as the original database paper from Codd (1970) argued that the same problems were being solved repeatedly, mass spectrometry data scientists are re-solving problems that have been more elegantly ironed out by dedicated teams in computer science and industry with much more extensive support. By leveraging existing optimizations in SQLite and DuckDB, we were able to create a highly performant system for storage of MS data that does not come with significant trade-offs between data extraction methods.

While SQLite is broadly used and its long history testifies to its continued utility, we can use even more modern database methods to improve further upon its analytical processing capacity. We tested both DuckDB the Apache Parquet data formats (Raasveldt and Mühleisen 2019; Vohra 2016) and found that they both performed better than SQLite in disk usage and query speed. DuckDB in particular is nearly a drop-in replacement for SQLite in many cases that's been extensively optimized for MS-related queries given its online analytical processing (OLAP) structure. DuckDB provides automatic compression algorithms and uses zonemaps to create bounding boxes for each subset of data, bringing together existing optimizations from mz5 (delta encoding), mzDB (bounding boxes), and MzTree (axis-agnostic queries) at zero additional

cost. Importantly, as with all databases, only the subset of interest needs to be written into memory, making the hardware requirements relatively lightweight.

Of course, to claim that existing frameworks should be discarded in favor of a novel method is to ignore decades of discussion. We acknowledge that our use case, that of largely exploratory and quality-control steps, is not a universal need and our lack of perfect metadata preservation in particular indicates that databases should become an auxiliary data structure alongside the vendor files or mzMLs, not substitute for them directly. Ultimately, the design decision for mass spectrometry data format will likely continue to be a point of contention and will result from a variety of factors, most crucially 1) initial vendor type, 2) programming language of the developer, 3) types of MS data included (e.g. full scan only versus MS/MS or metadata requirements), 4) whether the entire file will be processed or only a subset, and 5) how well a file type interfaces with downstream software. We intend to show with this manuscript that there is significant overlap between the goals of organizations much larger than any individual lab and that mass spectrometrists can benefit significantly from co-opting their development.

3.6 Conclusion

We propose that a simple relational database is an intuitive and performant mass spectrometry (MS) data storage format. Tables containing fields that map directly to known MS concepts means that adoption is straightforward and facilitated by the widely-understood structured query language (SQL), reducing the code required to extract subsets of interest to a single line. We show that this structure can also take advantage of regular advancements in computer science by leveraging modern data formats such as DuckDB and Parquet to reduce the disk space required while improving access times by 1-2 orders of magnitude. We hope that widespread adoption of

this format alongside the metadata-heavy vendor and mzML files will reduce the barriers to data access for mass spectrometrists and provide a consistent framework that covers a majority of the exploratory use cases.

3.7 Acknowledgements

We would like to acknowledge the University of Washington's eScience Institute and especially Bryna Hazelton and Dave Beck for their guidance and support during this project. We are also grateful to Theo Portlock and the other authors of their 2025 manuscript for posting their metabolomics data to Metabolights and allowing us to reuse it. Finally, we would like to acknowledge Josh Sacks and other members of the Ingalls Lab for their helpful discussions and for beta-testing many parts of the project.

3.8 Data availability

All data and code are available on the Github repository associated with this project at <https://github.com/wkumler/mzsql> under the `manuscript_things` branch.

3.9 Supplemental figures



Database
.sqlite
.duckdb

Table: features							
feature	name	formula	rt	mz	avg_area	med_snr	med_qscore
Feature	Feature ID	Formula	Retention time	m/z	Average area	Median SNR	Median quality
FT0001	Valine	C5H11NO2	6.5	118.0862	1104948.0	8.79	0.97
FT0002	TMAO	C3H9NO	12.1	76.0760	1729482.0	7.80	0.87
...additional entries...							
FT2311	Unknown	NA	23.7	231.1005	7953192.0	1.72	0.88
FT2312	Unknown	NA	29.5	254.0325	2397.9	3.19	0.24

Table: MS1					
filename	scan_num	x_coord	y_coord	mz	int
Source file	Scan number	X coordinate	Y coordinate	m/z ratio	Intensity
Smp_A	1	1	1	60.0452	6618
Smp_A	1	1	1	60.0532	2657
...millions of additional entries...					
Smp_Z	651501	678	327	60.0456	158084
Smp_Z	651501	678	327	60.0531	4673

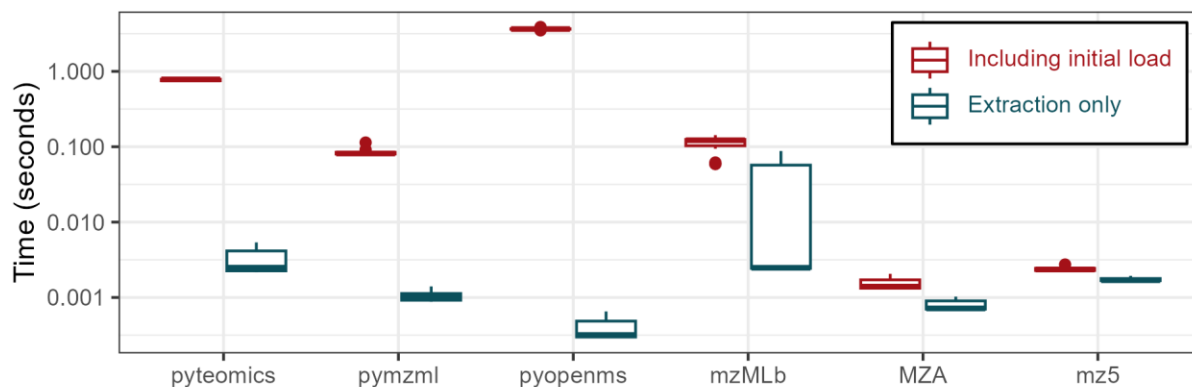
Table: MS1					
filename	scan_num	rt	dt	mz	int
Source file	Scan number	Retention time	Drift time	m/z ratio	Intensity
Smp_A	1	0.10	0.07	60.0452	6618
Smp_A	1	0.10	0.07	60.0532	2657
...billions of additional entries...					
Smp_Z	651501	12.37	6.16	60.0456	158084
Smp_Z	651501	12.37	6.16	60.0531	4673

Table: chrom_peaks							
filename	feature	rt_min	rt_max	mz_min	mz_max	peak_area	qscore
Source file	Feature	Min peak RT	Max peak RT	Min peak m/z	Max peak m/z	Area	Quality
Smp_A	FT0001	6.5	7.6	118.0862	118.0865	1104948.0	0.97
Smp_A	FT0002	12.1	13.0	76.0760	76.0762	1729482.0	0.87
...additional entries...							
Smp_Z	FT2311	23.7	25.0	231.1005	231.1006	7953192.0	0.88
Smp_Z	FT2312	29.5	30.6	254.0325	254.0328	2397.9	0.24

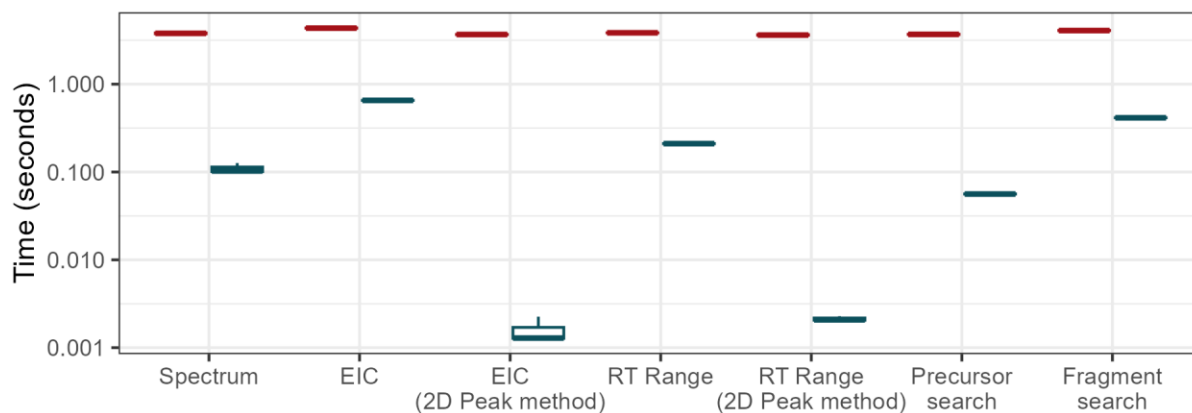
Table: rt_cor			
filename	scan_num	rt	rt_cor
Source file	Scan number	Retention time	Corrected RT
Smp_A	1	0.10	0.1
Smp_A	1	0.10	0.1
...additional entries...			
Smp_Z	23338	12.37	13.5
Smp_Z	23338	12.37	13.5

Supplemental Figure 3.5: Extended database schema for chromatographic peaks, multi-file features, MS^I data with drift time information, MS^I data from imzML files, and retention time correction.

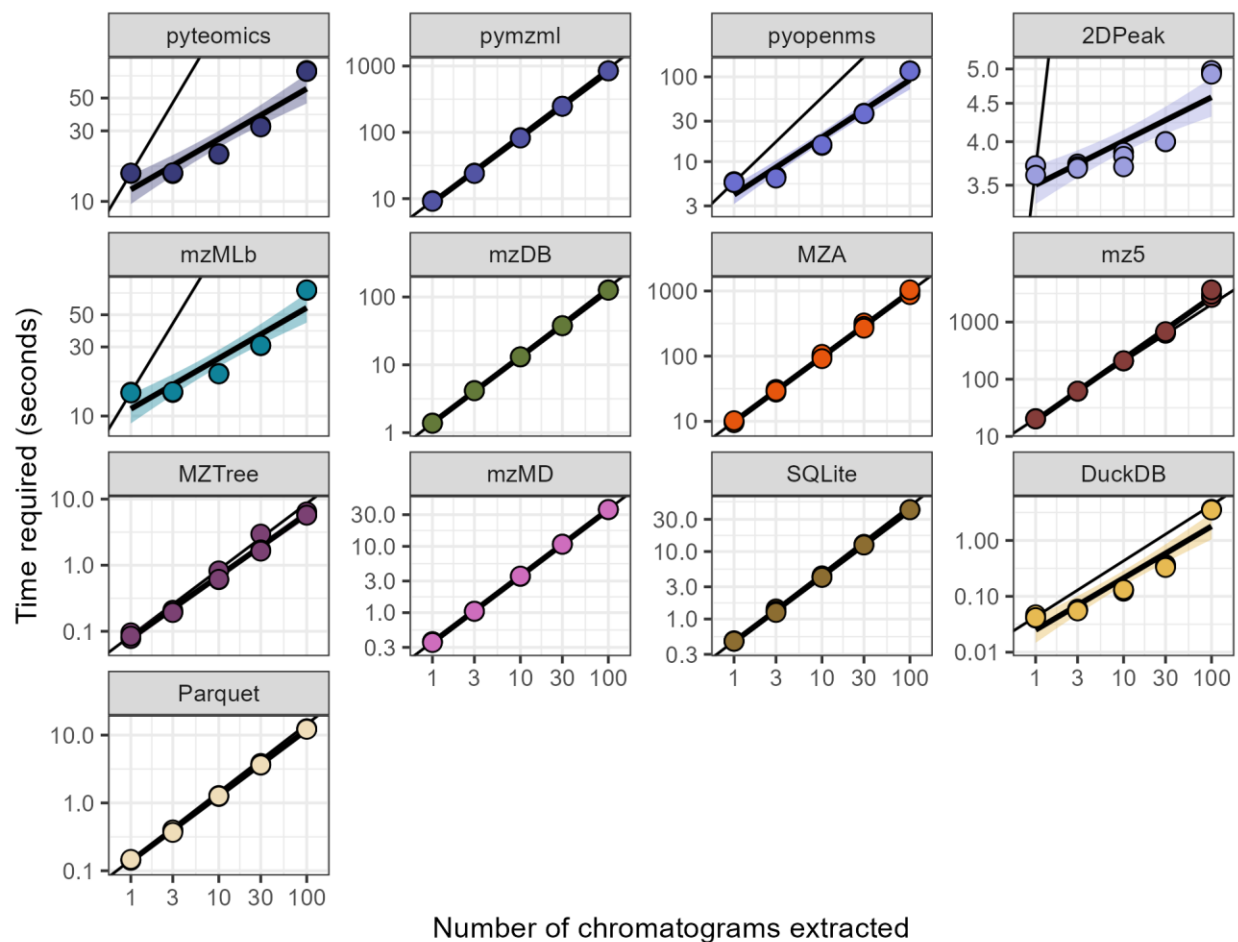
A) Scan extraction with and without pre-loading



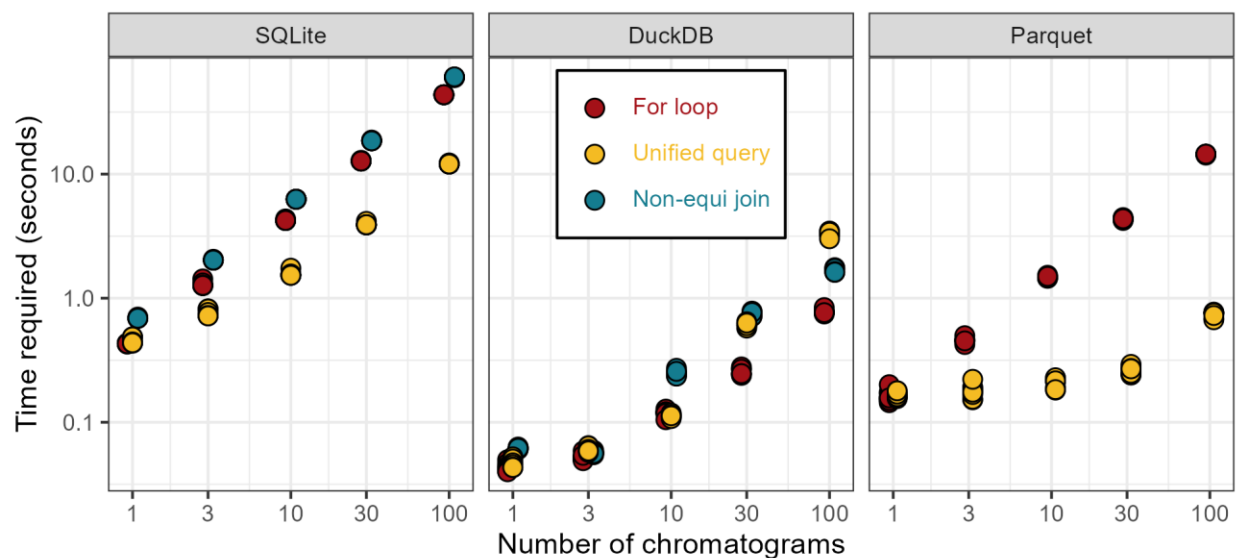
B) pyopenms metric timings with and without pre-loading



Supplemental Figure 3.6: Effect of “pre-loading” the data object prior to executing the query. Top panel shows what scan extraction looks like with and without pre-loading across multiple file types while the bottom panel shows what multiple metric timings look like with and without pre-loading for pyopenms specifically.



Supplemental Figure 3.7: Data from Main Text Figure 3.3 broken down by individual file access type instead of plotted on top of each other so the variance can be seen more clearly. Best-fit (Type I regression) lines have been fit behind the data with a 95% confidence interval shown in light grey. Black 1:1 lines show a predicted fit to the timings for 1 and 3 chromatogram extractions to show non-linear behavior at higher chromatogram extraction number.



Supplemental Figure 3.8: Number of seconds required to extract various numbers of chromatograms via multiple methods. The for loop method simply iterated over each mass of interest and submitted a single mass query each time, the unified query concatenated all masses and submitted a single query once for multiple masses, and the non-equi join method constructed a temporary table consisting of m/z bounds and used a join to extract the data. Parquet files have no joining capability so those values are not shown.

Chapter 4: Picky with Peakpicking: Assessing Chromatographic Peak Quality with Simple Metrics in Metabolomics

4.1 Abstract²

4.1.1 Background

Chromatographic peakpicking continues to represent a significant bottleneck in automated LC-MS workflows. Uncontrolled false discovery rates and the lack of manually-calibrated quality metrics require researchers to visually evaluate individual peaks, requiring large amounts of time and breaking replicability. This problem is exacerbated in noisy environmental datasets and for novel separation methods such as hydrophilic interaction columns in metabolomics, creating a demand for a simple, intuitive, and robust metric of peak quality.

4.1.2 Results

Here, we manually labeled four HILIC oceanographic particulate metabolite datasets to assess the performance of individual peak quality metrics. We used these datasets to construct a predictive model calibrated to the likelihood that visual inspection by an MS expert would include a given mass feature in the downstream analysis. We implemented two novel peak quality metrics, a custom signal-to-noise metric and a test of similarity to a bell curve, both calculated from the raw data in the extracted ion chromatogram, and found that these outperformed existing measurements of peak quality. A simple logistic regression model built on two metrics reduced the fraction of false positives in the analysis from 70-80% down to 1-5% and showed minimal overfitting when applied to novel datasets. We then explored the implications of this quality thresholding on the conclusions obtained by the downstream analysis

² This chapter was published as Kumler, W., Hazelton, B.J., and Ingalls, A.E. 2023. "Picky with peakpicking: assessing chromatographic peak quality with simple metrics in metabolomics." *BMC Bioinformatics* 24 (404). <https://doi.org/10.1186/s12859-023-05533-4>

and found that while only 10% of the variance in the dataset could be explained by depth in the default output from the peakpicker, approximately 40% of the variance was explained when restricted to high-quality peaks alone.

4.1.3 Conclusions

We conclude that the poor performance of peakpicking algorithms significantly reduces the power of both univariate and multivariate statistical analyses to detect environmental differences. We demonstrate that simple models built on intuitive metrics and derived from the raw data are more robust and can outperform more complex models when applied to new data. Finally, we show that in properly curated datasets, depth is a major driver of variability in the marine microbial metabolome and identify several interesting metabolite trends for future investigation.

4.2 Background

Liquid chromatography-mass spectrometry (LC-MS) is a powerful tool for exploring the molecular composition of biological samples. Its rapid sample processing (typically <1 hr run time), low limits of detection (pM-nM range), and ability to characterize novel molecules via fragmentation fingerprints make it a common workhorse for metabolomic research. In the past two decades, data-driven methods have established workflows for untargeted metabolomics but the imperfect performance of the core peakpicking algorithms continue to require manual oversight and curation. This problem has been exacerbated by the increased use of non-traditional chromatography such as hydrophilic interaction which tends to produce noisier peaks (Bajad et al. 2006; Myers et al. 2017b; Gika et al. 2019).

Noisy data and imperfect detection algorithms introduce a tradeoff between false positives (where contamination, background instrument or chemical noise is misclassified as biological signal) and false negatives (where real signals are undetected). Existing algorithms

tend to favor the inclusion of false positives because downstream analyses can always remove erroneous mass features, but false negatives cannot be later recovered (Pirttilä et al. 2022; Gloaguen et al. 2022). However, this approach requires more time from the researcher as they manually evaluate a potentially enormous number of mass features (MFs), a task that scales combinatorially with the number of samples and compounds measured (Myers et al. 2017a). Instead of minimizing false negatives, we believe that emphasis should be placed on allowing the experimenter to set a threshold for the proportion of false positives (the false discovery rate or FDR) and accept that this will inherently add to the number of MFs already lost in the data collection process.

Existing peak-detection software does not provide a clear way to exclude false positives in a consistent and unbiased way. Typical outputs consistent across the different implementations consist of the m/z ratio, retention time, and area for each mass feature, with some additional useful information occasionally provided such as the peak's signal-to-noise ratio or degree of skew (Pirttilä et al. 2022). None of these parameters answer the critical question about the likelihood that a given feature corresponds to a molecule present in the original sample. This parameter is crucial for downstream analysis because it represents the base rate for error propagation and acceptable thresholds should vary widely by the particular project's goals. In an exploratory analysis, any mass feature more than 50% likely to be real is perhaps worth considering, while in a confirmatory study this threshold may need to be above 99% likely to be real. Despite significant effort invested in improving the peakpicking algorithms, very little has been done to quantify the accuracy and precision of their outputs across the wide variety of datasets to which they are applied. The difficulty associated with comprehensively testing peak quality tends to result in the development of complex models overfit to their training data that

perform poorly when facing truly novel datasets, when a simpler model may produce more reliable quality estimates.

A single parameter of MF quality also facilitates downstream analyses in multiple ways. This metric would improve statistical power by reducing the number of effective hypotheses tested and allow researchers to focus effort on features least likely to be noise. Additionally, this parameter could be optimized to improve peakpicking and chromatographic settings independently of the software used and minimize inter-lab variability when scripted to provide consistent, reproducible results independent of the particular expert reviewing its performance. Constructing such a single comprehensive metric calibrated to likelihood is also more effective than multiple independent thresholds because it has meaningful units, does not require estimating the relative power of individual metrics, and allows a good MF to compensate for weak performance in one area with strong performance in other metrics, e.g. as implemented in Pirttilä et al. (2022) and Kantz et al. (2019).

An area particularly ripe for improved tools for metabolomic data analysis is that of the open ocean (Kido Soule et al. 2015). Low compound and high salt concentrations make metabolomics analyses difficult to study in this area but its vast size and the direct effect of its microbial communities on the Earth's biogeochemistry make it critical that we understand the transformation of energy and nutrients on a molecular scale (Boysen et al. 2018). Metabolites are the currency of chemical exchange both intra- and inter-cellularly, serving as building blocks of larger molecules, regulators of osmotic balance and storage of nutrients, as well as important chemical signals on their own. These small molecules serve both as signposts for the complex biological landscape in this highly dynamic region and give a sense of not only who is present

but also what ecological roles they are serving and the niches they fill (Kido Soule et al. 2015; Boysen et al. 2021; Heal et al. 2021).

In this paper, we use open ocean marine metabolite LC-MS samples to develop and test a variety of chromatographic peak metrics. We construct and validate multiple predictive models of MF quality based on metrics both common in the literature and custom implementations we have found useful in our own analysis, with a particular focus on developing a model robust across datasets and avoiding overfitting on training data. This allows us to connect the physical, chemical, and biological measurements taken regularly around the globe to a molecular-scale perspective of particulate organic matter in the ocean by linking the chemical currencies that fuel the planet to the environments in which they are found.

4.3 Results

4.3.1 Dataset characterization

We performed untargeted peakpicking with XCMS on four datasets, two environmental (MESOSCOPE and Falkor) and two culture (Ptime and CultureData), detecting an average of 3,300 mass features (MFs). The fewest (1,495) were detected in the Falkor data and the most were found in the Ptime samples (7,781). In the Falkor and MESOSCOPE datasets that were fully labeled by an MS expert, approximately 70% (69% and 73%, respectively) of the features were given a “Bad” designation, corresponding to noise MFs that the expert would not have included in a downstream analysis. In both, 5% of the MFs were unable to be assigned confidently to either “Good” or “Bad” classes and 10% were identified as appearing only in the standards, leaving only ~15% of the features classified as “Good” (16% and 12%, respectively).

Most metrics had reasonably normal distributions after the scaling and normalization described in Methods. Visually, the most compelling separations between good and bad MFs were observed in our peak shape and novel SNR metrics, with almost complete separation

between good and bad peaks provided by the new peak shape metric alone. Peak width and its standard deviation also showed reasonable separation between good and bad MFs (good MFs tended to have low SDs and larger peak widths). The isotope shape and area correlations also showed good separation (Supplemental Figure 4.9).

4.3.2 Logistic regression performance

According to all three logistic regression models (see Methods), the majority of MFs were estimated to have a less than 1% chance of being good. The full model (containing all evaluated peak metrics) and the XCMS model (built on only those metrics calculated from the XCMS output) both displayed a strongly bimodal distribution, with a large number of MFs also exceeding a 99% chance of being good, while the two-parameter model (consisting of the novel SNR metric and the peak shape correlation metric) had a flatter distribution with fewer high-confidence MF assignments and more intermediate values (Figure 4.1).

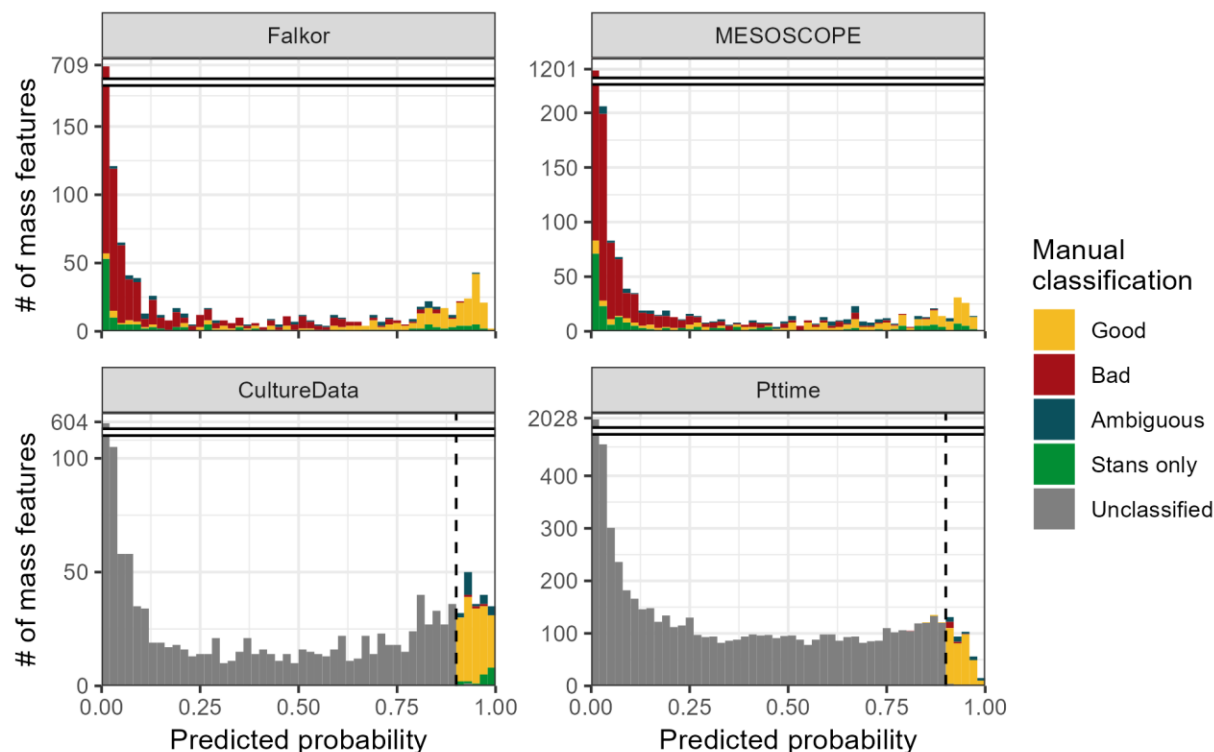


Figure 4.1: Histograms showing the estimated likelihood of a given mass feature being categorized as “Good” according to the two-parameter logistic model trained on the combined fully-labeled Falkor and MESOSCOPE environmental datasets. Colors indicate the category in which each feature was manually assigned by an expert, with “Stans only” referring to a good mass feature that was only visible in the standards run alongside the samples. Culture datasets CultureData and Pptime were manually labeled only for those features with an estimated likelihood above 90% (dashed black vertical line) according to the final model and were otherwise unclassified.

We explored the relative predictive power of the individual parameters using the full model and found that the predictors least likely to be different from zero due to chance were the mean m/z ratio, our novel peak shape correlation metric, and our novel SNR estimate, all with reported p -values $< 10^{-10}$. The value of the novel parameters was then validated using a random forest model that also found them to have the highest importance (Supplemental Table 4.1).

The full model performed very well when tested internally on the same dataset both during 80/20 cross validation and when using the full dataset, with FDR (false discovery rate, defined as the number of false positives divided by the total number of positive predictions)

values in the 5-10% range and 80-90% GFF (% good features found, defined as the number of true positives divided by the total number of features manually classified as good) values implying that a large majority of the good MFs passed the threshold with very little noise included. The XCMS metrics performed slightly worse, with FDR values in the 10-15% range and GFF values closer to 75%. The two-parameter model performed worst when tested internally, with an FDR of about 20% and GFF also around 75% (Figure 4.2). However, when the models were trained on a different dataset than the one they were used to predict classifications for, they all had similar performance with FDRs around 10-25% and GFF around 60-80%. The model trained on MESOSCOPE and tested on Falkor had consistently higher values, indicating that it was favoring more MFs recovered at the cost of a higher FDR, while the reverse was true for the model trained on Falkor and tested on MESOSCOPE (Figure 4.2).

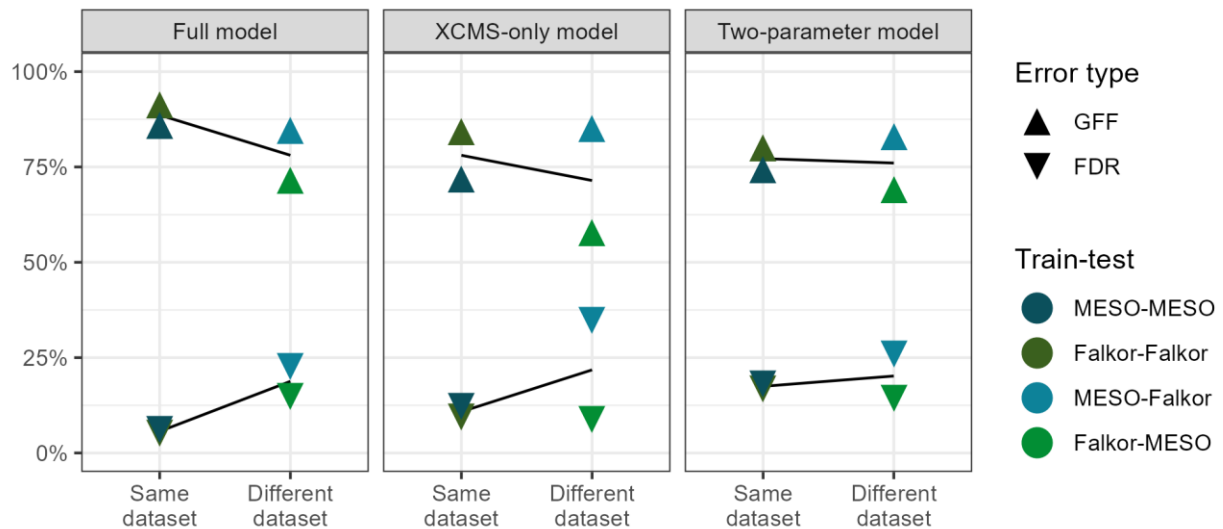


Figure 4.2: False discovery rate (FDR) and fraction of good features found (GFF) plotted across different subsets of model parameters. Lower FDR indicates a smaller fraction of false positives among those mass features the model categorized as “Good” using a threshold of 0.5, and higher GFF indicates a larger fraction of the total good features were found using the same threshold. Points are colored by the model used for training and testing, with internal validation (using the same dataset for training as prediction) in the darker colors on the left and external validation (using a different dataset for training than prediction) in the lighter colors on the right of each panel. Lines of best fit have been estimated and plotted in black behind the data

points, with the steeper slopes found in the full and XCMS-only models indicating overfitting on the training data.

4.3.3 Model stability under different training sets

We found that the predictions made from a Falkor-trained dataset consistently differed from a MESOSCOPE-trained dataset for the full and XCMS-only models. In the raw probability space, the two-parameter models had the highest Pearson correlation coefficient (r) value of 0.996, while the full models and the XCMS-trained models had r values of 0.799 and 0.863, respectively. When compared in ranked space using Spearman's ranked correlation, we found an intensification of this effect, with a higher ρ for the two-parameter model of 0.998 but lower ρ values for the full and XCMS-trained model of 0.725 and 0.804, respectively (Figure 4.3).

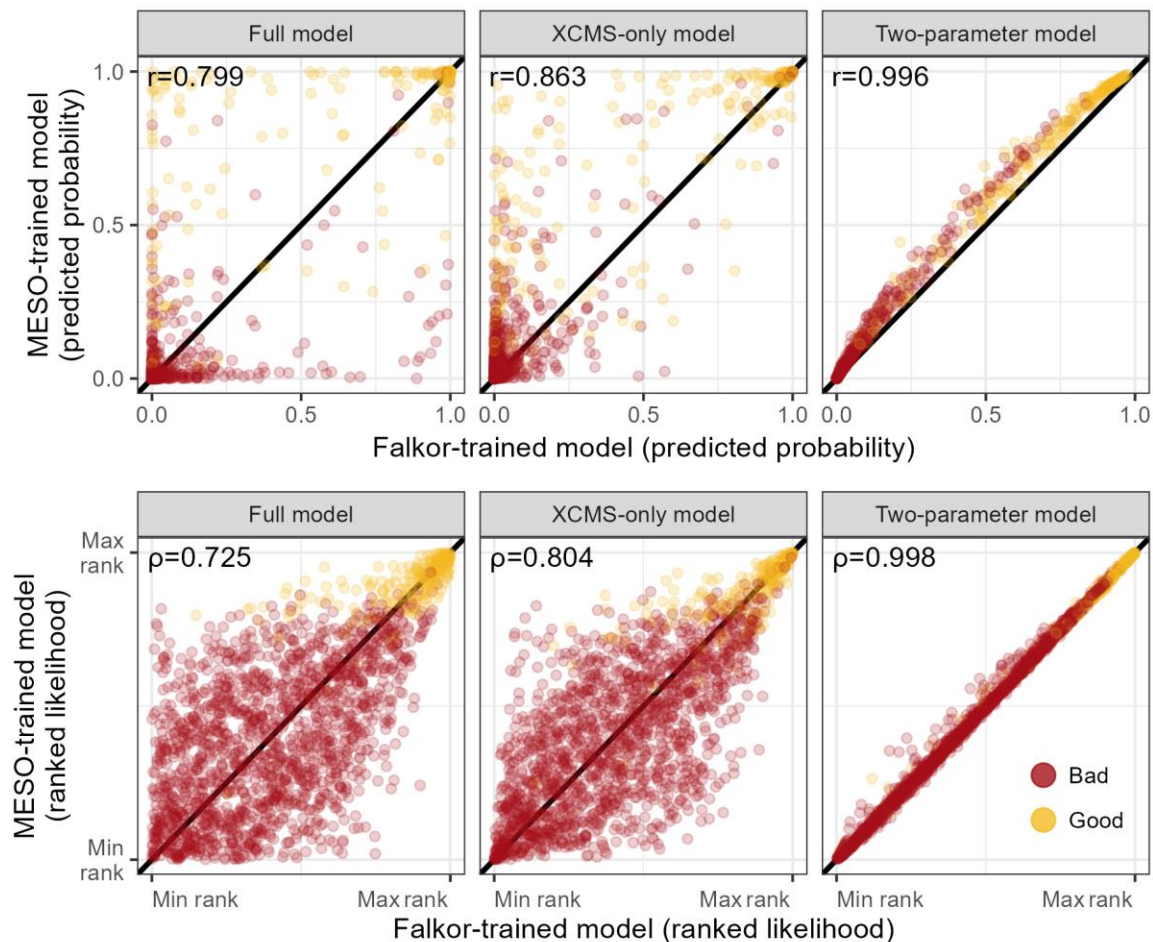


Figure 4.3: Predicted likelihood of a feature being “Good” according to a model trained on the MESOSCOPE dataset vs a model trained on the Falkor dataset. The top row of plots shows the exact likelihood predicted by the logistic model across three different subsets of parameters, while the bottom row shows the estimates ranked from least likely to most likely. Points are colored by their manually-assigned quality according to an expert.

A majority of the time, the estimates from the two models disagreed by more than two times the standard error of the estimate. Some parameters disagreed not only in magnitude but also in sign, with the Falkor-trained full model increasing MF goodness likelihood with larger PPM variation and a wider peak width, while the MESOSCOPE-trained full model had negative estimates for each of these parameters. Notably, the peak shape and novel SNR parameters used in the two-parameter model were among the most robust to training model variation, potentially explaining the consistency described above (Supplemental Figure 4.10).

When testing model stability under a smaller sample size, we found reasonably good convergence in a dataset containing half the mass features with most model parameters falling within two standard errors of the estimate for the XCMS and two-parameter model, while the full model required closer to 80% of the mass features to produce estimates consistent with the original model (Supplemental Figure 4.11).

4.3.4 Regularized regression and random forests perform about the same

None of the penalized regression models significantly improved cross-validated performance between the MESOSCOPE and Falkor datasets when measured by both initial performance and the performance drop when applied across datasets. All three regularized regression models had similar behavior, with ridge regression ($\alpha = 0$) obtaining the lowest rates for both GFF and FDR, while lasso ($\alpha = 1$) obtained higher rates for both and represented a less-stringent false negative acceptance. As expected, the elastic net ($\alpha = 0.5$) fell in between the two (Supplemental Figure 4.12). The random forest model, interestingly, had perfect predictive capacity when tested internally on the training data (FDR=0%, GFF=100%, for both MESOSCOPE and Falkor) but showed a significant drop in improvement when applied across datasets (Supplemental Figure 4.12). In each case, the performance drop when applied to a novel dataset was more extreme than the simple two-parameter model described above.

4.3.5 Performance of a stricter threshold on novel datasets

We settled on a 90% likelihood threshold for application to novel datasets because it struck a balance between the number of MFs we estimated to be necessary for robust testing while still remaining reasonable to manually label. For the CultureData dataset, we obtained 1,790 total mass features, 192 of which had predicted likelihoods above 0.9. Of these, 151 were identified manually as “Good”, 21 were given “Ambiguous” designations, and only 3 were flagged as

“Bad”, with the remaining 17 appearing only in the standards. For the Pptime dataset, 7,781 were obtained with 400 flagged by the model as “Good”. 348 were truly good MFs, 35 were ambiguous, and 17 were “Bad”. No standards were run during this analysis, so there were no features in that category.

With the stricter threshold, we obtained FDR values consistently below 5% even on the novel datasets, with values of 1.0%, 0.0% (truly zero false positives), 2.0%, and 4.6% for Falkor, MESOSCOPE, CultureData, and Pptime respectively (Figure 4.4). Of course, this low error rate meant that we miss out on additional potentially valuable features, with only a fraction of the total good MFs making it past this threshold. In both the Falkor and MESOSCOPE datasets, less than half of the good MFs were labeled as such, with actual values of 39.4% and 26.5%, respectively. Since we did not label the complete dataset for CultureData and Pptime, we cannot accurately calculate the GFF but expect it to be in a similar range (Figure 4.4).

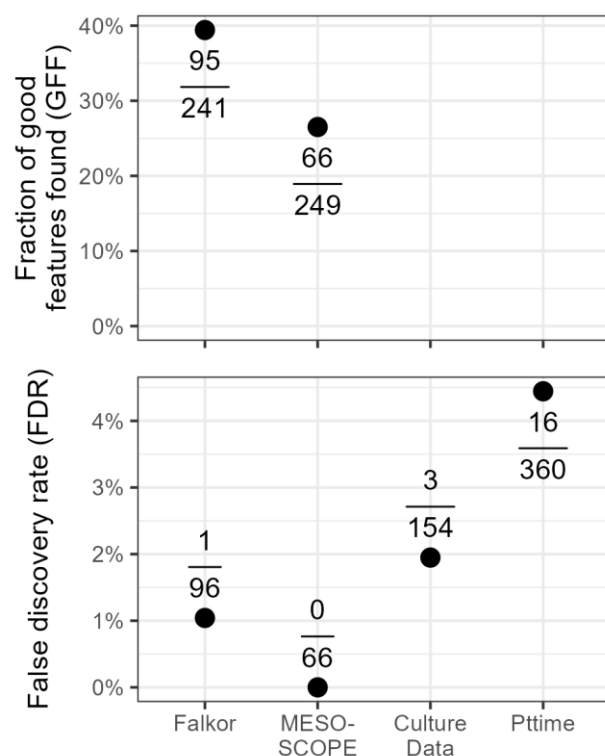


Figure 4.4: False discovery rate and proportion of total good features identified as good by the two-parameter model trained on the combined MESOSCOPE/Falkor dataset and applied to each dataset. A stricter likelihood threshold is used here (0.9) than in Figure 4.2. FDR is calculated by dividing the number of false positives by the total positives produced by the model and GFF is calculated by dividing the number of true positives by the total number of good features as identified manually (only possible for fully-labeled datasets). Points correspond to the calculated percentage and absolute numbers are provided above/below the point.

4.3.6 Implications for biological conclusions

4.3.6.1 Univariate techniques

A majority of the features (1,323 of 1,832 total) in the original, non-thresholded MESOSCOPE dataset had no significant trend with depth, with FDR-controlled Kruskal-Wallis p-values exceeding 0.05 (Figure 4.5). The largest category that did have a trend with depth was the category containing 118 mass features with largest peak areas distributed evenly between the 15 meter and deep chlorophyll maximum (DCM) samples, while the 175 meter samples had significantly smaller areas (15m = DCM > 175m). The similar but statistically distinct categories of 15m > DCM > 175m and DCM > 15m > 175m had 68 and 35 features, respectively, and

together indicate that many molecules are highly abundant throughout the surface ocean down to the DCM layer and decrease in concentration at 175 meters. A surprising number of features were also found to have DCM minima ($DCM < 15m = 175m$, 26 features) or linear increases with depth ($15m < DCM < 175m$, 12 features) given the few environmental parameters that have these trends (Figure 4.5).

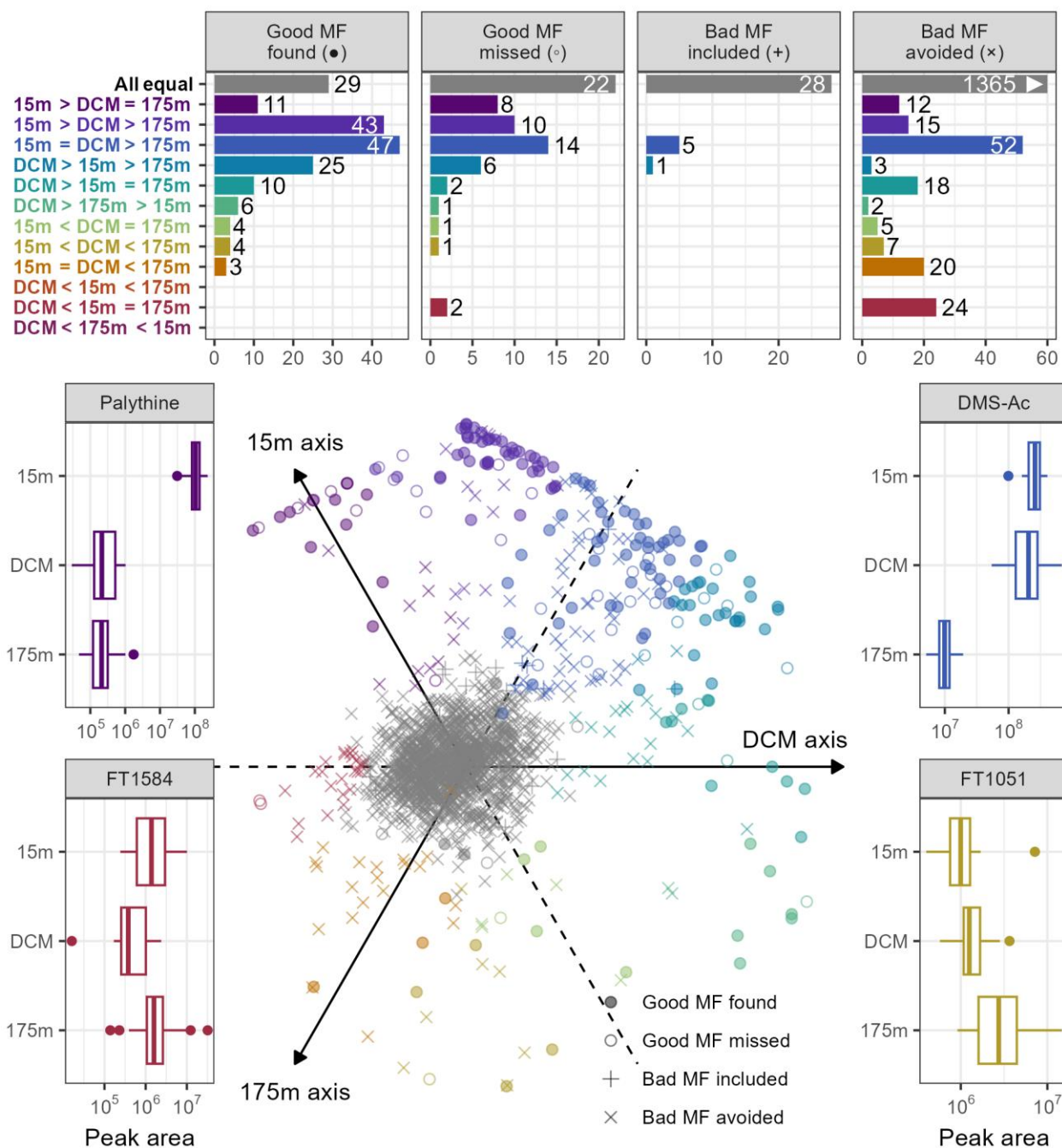


Figure 4.5: Plot of metabolite response to depth and the model classification error distribution. Barplots at the top show the number of MFs (mass features) in each depth response category and are broken down by the classification error types. Compounds were assigned a depth category via Dunn's post-hoc test for significant differences between the sample depths. Good MF found = true positive, good MF missed = false negative, bad MF found = false positive, and bad MF avoided = true negative according to a 0.5 likelihood threshold. Note that the majority of the features in the "bad MF avoided" category fell into the "All equal" depth class for which there was no significant differences between the depths (1365 MFs) but the x-axis has been truncated at 60 for ease of visualization. The boxplots in the bottom illustrate the depth response type for 4

specific categories, with raw peak area plotted on a log scale against the sample depth (DCM = deep chlorophyll maximum, ~110 meters, DMS-Ac = dimethylsulfonioacetate). All MFs are shown in the central bottom plot across three axes using the rank-normalized median value at each depth as the coordinate for that axis. Each mass feature corresponds to a point in the plot, and their position on the plot describes the shape of their depth profile. Compounds aligning with the 15m axis correspond to compounds with most of their abundance found in the surface ocean; points far to the right side correspond to compounds that are found only at the DCM; points found at the bottom of the plot are those compounds that increased more or less linearly with depth.

A different story emerged, however, when the bad MFs were removed from this analysis. Good features were most commonly found to have their highest concentrations at the DCM or the surface, rather than being fixed with respect to depth. Of the 182 good MFs, less than a fifth had no trend with depth (44/249) and a majority had unequivocally lowest values in the 175 meter samples (those with 15m/DCM > 175m, 145 features). The two-parameter model, when applied with a 50% likelihood threshold, also recovered this general feature distribution and classified many of the features with no significant depth signal as likely to be bad.

Additionally, a large number of features manually identified as bad nonetheless had significant differences with depth. This was surprising because we had assumed that bad MFs corresponded to instrument or chemical noise, which we did not expect to have any biological trend. Further investigation of a few randomly selected bad features with a biological difference revealed the reason behind this: most of those investigated were actually tails of other MFs. Integrating just the tail of a peak retains the biological signal of the full peak while still looking visually like instrument noise, thereby introducing pseudoreplication in the feature space.

The model did fail to recover some interesting biological variation, however. Two features of particular interest were those good MFs with a DCM minimum (DCM < 15m = 175m), both of which were missed by the two-parameter model. These features possess an unexpected biological signal that does not track with depth or other common oceanographic

parameters, thereby potentially representing an interesting biomarker that decreases despite an increase in biomass. Further inspection, however, revealed them to be an isotope pair of a brominated compound that actually had the highest peak areas in the blanks, indicating that this is likely a contaminant introduced during sample processing with an m/z ratio of 188.9577 and a retention time around 2 minutes.

4.3.6.2 Multivariate techniques

Multivariate statistic strength also benefitted from the reduced FDR when applying the two-parameter model. For the PERMANOVAs, we found that the proportion of variance explained (R^2) and the pseudo-F statistic increased monotonically with the likelihood threshold used to subset the data (Table 4.1).

In each test, the permutational p-value obtained was less than 0.001, indicating that the differences between samples due to depth were unlikely to be due to chance. However, the pseudo-F was much larger with higher thresholds, scaling from around 8.5 when thresholding at a 1% likelihood to 42 when thresholded at a 90% likelihood (Table 4.1).

Table 4.1: Number of mass features, percent variance explained, pseudo-F statistic, and stress values from performing a permutational MANOVA and 2D non-metric multidimensional scaling (NMDS) on subsets of the full mass feature selection according to variable likelihood thresholds.

Data subset	# of MFs	% Var Expl.	Pseudo-F	2D NMDS stress
All MFs	2086	10.4	5.5	0.169
Threshold 0.01	1129	15.2	8.6	0.191
Threshold 0.1	516	25.0	16.0	0.200
Threshold 0.5	287	34.6	25.3	0.180
Threshold 0.9	75	46.7	42.0	0.097
Only good MFs	249	44.2	38.1	0.142

We also tested the inclusion of all the features identified with XCMS (corresponding to a 0% threshold) and the results when only the manually-identified “Good” features were included. The default XCMS output continued the trend observed above, as expected, with the least

variance explained and the lowest R^2 value. Subsetting for the “Good” MFs only, however, did not actually return the highest F-ratio or R^2 , instead falling between the 50% and 90% thresholds for these two metrics. In large part this is due to the much smaller number of features: 249 features were manually labeled as Good, while only 75 exceeded the 90% likelihood threshold (Table 4.1).

The relative power of identifying only the very best MFs was also illustrated visually with non-metric multidimensional scaling (NMDS) plots (Figure 4.6). In these common exploratory plots, the MFs with likelihoods above 50% strongly separated by depth while lower thresholds disguised the true signal and had higher stress values. Performing an NMDS on the manually-identified Good MFs resulted in output nearly indistinguishable from those of the 90% and 50% thresholds.

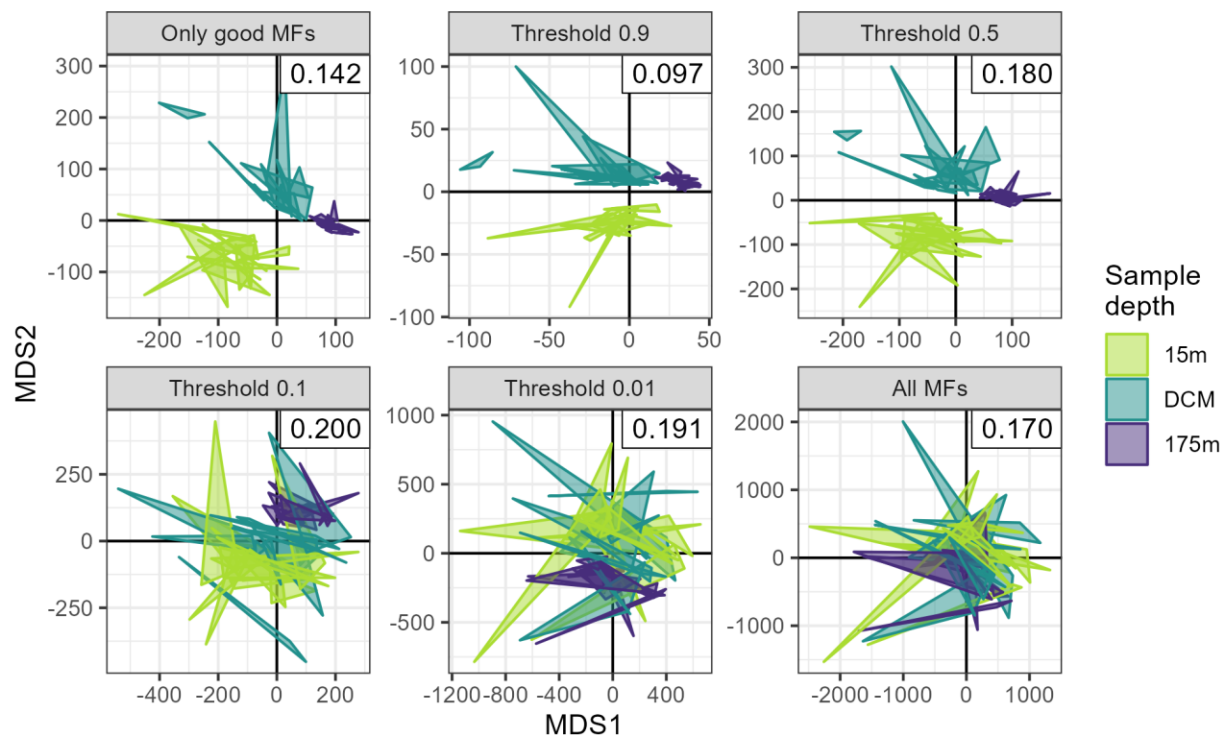


Figure 4.6: 2D non-metric multidimensional scaling (NMDS) plots of metabolite similarity according to sample depth across multiple likelihood thresholds. Triplicate samples are represented by the vertices of the triangles and colored by the depth from which they were

sampled (DCM = deep chlorophyll maximum, ~110m). “Only good MFs” refers to those features manually labeled as “Good”. NMDS stress values are reported in the upper right corner of each plot.

4.4 Discussion

We used two fully-labeled and two partially-labeled HILIC LC-MS datasets to assess the performance of the XCMS algorithm and construct a robust model of peak quality. To measure performance, we used two measurements of success closely tied to intuitive questions about a dataset: the percentage of total good features found (GFF, also known as recall or sensitivity) and the percentage of bad mass features (MFs) included, also known as the false discovery rate (FDR). We decided against using the F_1 score as an overall summary statistic because false negatives and false positives have very different implications in this context and should be treated separately.

One of the major ways in which this manuscript differs from prior work is its focus on summary statistics calculated across multiple files. Most existing peakpicking literature uses the single-file EIC peak as the core bit of training data, but that approach ignores critical information obtained elsewhere in the MS run that can change the judgement made on a single chromatogram (Pirttilä et al. 2022; Guo et al. 2021; Müller et al. 2020). Features that are high quality are typically represented in multiple files, especially in quality control pooled samples, and a feature that only has a peak in a single file is typically regarded as highly suspect, if not removed entirely. Classifying an entire feature at once not only has the advantage of reducing the amount of manual labor by a factor equal to the number of files in the run (typically 10s-100s smaller) but is also a better representation of the judgement made by an MS expert. An exemplary implementation of this multi-file approach in prior work is reported in Kantz et al. (2019), who compared the multi-file summary statistic model to a deep neural network and came to many similar conclusions.

4.4.1 Two-parameter logistic regression model with raw data metrics showed the most reliable performance

We explored several different types of classification models for separating good mass features (MFs) from bad, with a particular focus on quantifying the likelihood of each class rather than just returning the label. We found that a simple two-parameter logistic regression model trained on two novel metrics of peak quality had reasonably good performance on the training set and was highly robust when applied to novel datasets. The logistic regression in particular was favored over the random forest and regularized regression we tried due to their similar performance and increased interpretability (Supplemental Figure 4.12). The random forest model overfit the data particularly strongly, producing perfect performance during training that generalized very poorly even when trained on a small subset of the features. Given our initial goal of producing a highly reliable model rather than the one with the maximum performance, we strongly favored the simple two-parameter approach.

This model outperforms the previously reported logistic regression model in Kantz et al. (2019) and is highly simplified. There, they used a nineteen-parameter multiple logistic regression model and found a maximum performance of 80% GFF and an FDR of 34% on a cross-validated second cohort, similar to our cross-dataset testing. Our final two-parameter model also had an 80% GFF at a 0.5 likelihood threshold, but a significantly lower FDR of ~22.9%. This increased performance is likely due to our use of the metrics recalculated from the raw data, as their metrics were only calculated from the default MzMine2 peak parameters reported: similar to our XCMS-only model. Previous work on a “shape-orientated” algorithm also established the utility of testing the extracted ion chromatogram against a Gaussian shape (Bai et al. 2022). There, the use of a Marr wavelet had GFF values in the 98-100% range but

very high FDR values of 82-91%, representing a very lenient threshold much closer to the XCMS or ADAP defaults.

4.4.1.1 Performance relative to recent deep learning methods

Guo et al. (2021) presented EVA and reported an accuracy of 90-95%, a range inclusive of our accuracy on both the Falkor (92.1%) and MESOSCOPE (94.4%) datasets when using a likelihood threshold of 0.5. However, we note that accuracy alone can be a highly misleading statistic to report when working with unbalanced datasets because very high accuracy can be obtained by simply classifying everything as bad, with a strong incentive to actually *increase* the number of bad MFs initially picked while doing so. This strategy, when applied to our data, returned accuracies in the 80-90% range despite being a useless classifier for downstream analysis.

The class imbalance, with mostly poor quality MFs, is partially why we chose to measure precision and recall separately instead of total accuracy. However, precision and recall can also be ambiguous when the positive class is not specified and the raw confusion matrices are unavailable, thus our very precise use of the FDR and GFF metrics as well as providing the confusion matrices in Supplemental Table 4.2. Melnikov et al. (2020) reported precision and recall in their presentation of *peakonly*, relative to which we obtained higher accuracy (they report 89% accuracy) but worse GFF and FDR (89% and 3%, respectively, relative to our 77.1% GFF and 19.6% FDR overall). However, if we report precision and recall with the positive class set to “Bad”, essentially trying to predict poor-quality MFs instead of good ones, our precision becomes 96.5% and our recall 95.7% due to the strong prior information about most MFs being bad.

Gloaguen et al. (2022) later introduced NeatMS, another CNN, and compared it directly to *peakonly* to claim equivalent or superior performance across a range of dilution factors.

However, they do not report total precision or recall metrics in a comprehensive untargeted way but instead focus only on assessing the model's performance on known chemical standards. They report a percentage of standards found for the peakonly model applied to their data and find that its performance is significantly lower (79.4%) than the recall reported in Melnikov et al. (2020), perhaps indicating that the peakonly model is still overfit.

While the model we present here likely has reduced performance relative to the CNNs, we would argue that its utility is not in maximizing performance but instead in maximizing interpretability and robustness as previously noted by Kantz et al. (2019). In particular, the CNNs provide no way to control the tradeoff between false positives and false negatives and no relative ranking of individual MF quality beyond the broad bins into which they are placed or explanation of relative metric strength for later analyses.

4.4.1.2 Assessing the relative power of individual metrics

Although the deep learning models show promise for peak quality recalibration, many mass-spectrometrists are reluctant to jump fully to their black-box nature. For this reason, we also reported here the relative power of individual parameters in our full model and use the results to dispel several myths about which parameters are useful in distinguishing signal from noise.

The two metrics in the final model were rederived from the raw EIC data because they matched our intuition about what makes an MF look good to an MS expert. These are very simple metrics and therefore fast to calculate, but we expect that more complicated metrics could perform even better. For example, the method of using the data within the peak boundaries for SNR calculation rather than data outside of them is not known to the authors to be implemented elsewhere but could be further improved by more advanced smoothing methods rather than using the residuals directly. Additionally, the calculation of peak shape using a Pearson's correlation to an idealized curve was not expected to be especially powerful given prior research (e.g. Ipsen et

al. (2010)) and that the centWave algorithm essentially uses this information already during the wavelet fitting, but still proved to be a highly informative parameter. This metric could be improved with more careful summary statistics that account for the differences between samples. Currently, the use of the overall median value does a reasonable job at identifying MFs that appear in many samples but performs poorly when detecting MFs that appear in only a few. Also worth noting is that the calculation of any new metrics such as these that rely on access to the raw data require exact specification of the maximum and minimum m/z and retention time for a peak, values that are not always returned by peakpicking algorithms and must be recalculated, as in Kantz et al. (2019). To avoid the additional overhead of recalculation and the possibility of raw data unavailability, we have implemented these metrics during the initial peakpicking step of XCMS in a fork of the GitHub available at <https://github.com/wkumler/xcms> and have submitted a pull request to implement them into XCMS directly.

We were surprised at the poor performance of several other metrics. The isotope information in particular was expected to be a very strong predictor of MF quality given previous work that uses this metric extensively (Libiseller et al. 2015; Treutler and Neumann 2016; El Abiead et al. 2021). We learned that many noise MFs still have reliable isotopes (perhaps unsurprising, given that the noise is in fact often caused by solvents or contaminants that are still chemical in nature) and that many real MFs are simply too small (low-intensity) to have detectable isotope peaks in this kind of dilute environmental sample.

The relative standard deviation (RSD), also called the coefficient of variance, among pooled samples is another parameter that performed surprisingly poorly given its general acceptance as a quality scoring metric. In the full model, neither the traditional calculation of RSD (standard deviation divided by the mean) nor the robust implementation (median absolute

deviation divided by the median) was a significant parameter. This result was also reported by Gloaguen et al. (2022) who noted that while the RSD was typically lower for high-quality features there were many noise MFs with low RSDs as well.

We also showed that the automatically calculated SNR parameter from XCMS is not especially useful in distinguishing signal from noise. After inspecting a selection of MFs that had anomalous values for this metric, we are inclined to agree with Myers et al. (2017a) and conclude that this is often due to insufficient data outside of the peak for a robust calculation of noise level.

Finally, we were surprised to find essentially no predictive power offered by peak area or intensity, with good MFs distributed almost identically to the bad MFs in this space. This cautions strongly against an arbitrarily-decided intensity threshold for winnowing down the number of MFs, in agreement with previous work (Houriet et al. 2022; Barupal et al. 2021). Similarly surprising was the lack of power in the design-of-experiments metrics, although this was less surprising given the number of missing values that were later filled in with an order of magnitude outside the most extreme value (Supplemental Figure 4.9).

4.4.1.3 Model selection and simplification

We settled on the highly reduced model of just two parameters because we found that additional parameters often improved performance on the training set but did not do so significantly for the novel datasets where the application of such a model is actually useful (Figure 4.2). The drastic drop in performance on out-of-sample data was particularly concerning because it creates overconfidence in the true level of noise actually ending up in the final dataset. One important caveat to note is that for the partially-labelled CultureData and Pptime datasets, there exists an uncontrolled degree of experimenter bias because the MS expert responsible for labeling did

know that these MFs were all expected to be good. However, given that we do still see poor-quality MFs in this set indicates that this was not an overwhelming bias.

We also found that this reduced two-parameter model was largely independent of the particular training set used, unlike in the more complex models (Figure 4.3). This was true in both absolute likelihood as well as rank-ordered space, a particularly important distinction when one imagines manually labeling “down” the dataset where the researcher starts viewing the chromatograms associated with the very best features and eventually reaches a point where enough MFs have been reviewed or bad MFs are frequent enough that they decide to stop.

A final benefit to the reduced model is the smaller training set required to reach stability (Supplemental Figure 4.11). This reduced size means that a useful model could be trained using only a fraction of the MFs identified in a sample set and then used to predict the quality for the remainder of the features. This reduction in training set size was not as significant as we expected, however, with several hundred features requiring manual labeling before even the two most stable parameters reached a consensus.

4.4.2 Biological conclusions vary significantly by feature quality

We found that the conclusions obtained from the metabolomic datasets differed in significant ways depending on the quality threshold used to remove bad MFs from the downstream analysis. In the multivariate case, we ran the same analysis of PERMANOVAs and NMDS plots on various subsets of the original XCMS output and found that the effect size of depth was strongly influenced by the threshold chosen. This is unsurprising given that most noise MFs should not have a biological signal to begin with, but is troubling for interpreting analyses where the FDR is not reported or the dataset not manually reviewed because the absence of a notable effect could simply be due to the overwhelming degree of noise in the default output.

In the univariate case, we showed that while noise MFs are predominantly absent of a large biological signal, there are many that still have a significant biological trend. While some of these are inherently due to the likelihood of getting a small p-value with enough attempts despite FDR correction, a larger number of these poor-quality MFs were due to partial integration in which only the tail of a feature was integrated. This essentially duplicates the signal of the original MF in later analyses and should be removed. The real features showed a strong biological trend of high concentration throughout the surface ocean and down through the deep chlorophyll maximum (DCM), with most features equally abundant at 15 meters and this ~110 meter depth feature before dropping off at depth. This pattern tracks well with previous reports of biomass from the same sample site as well as earlier literature (Barone et al. 2022; Heal et al. 2021). Critically, this also highlights the danger of noise MFs when additional normalizations are later applied. Scaling metabolomic data to biomass measurements is a common technique, and yet here it would have caused an enormous number of false positives that would have appeared to be intriguingly enriched below the DCM.

4.5 Conclusions

The large number of mass features due to noise present in metabolomics datasets can be controlled using a simple logistic classification model. We trained such a model on two full-labeled open ocean HILIC datasets and found that the best performing parameters in the model were a custom signal-to-noise metric and a test of similarity to a bell curve. This model showed robustness to overfitting, independence from the training set, and a reduced degree of manual labeling required. With this model, we showed how the distribution of metabolites in the open ocean is strongly affected by depth and categorized molecules according to their depth response.

This distribution reproduces measures of bulk biomass but highlights several molecules of interest that diverge from the overall trend.

4.6 Methods

4.6.1 Sample collection

Environmental samples were collected from the North Pacific Subtropical Gyre near Station ALOHA during two research cruises that targeted strong mesoscale eddy features during June/July 2017 and March/April 2018, traversing an area between 28° N, 156° W and 23° N, 161° W. An eddy dipole off the coast of Hawaii was detected using sea-level anomaly (SLA) satellite data and targeted for both a transect across the cyclonic and anticyclonic poles of the eddy dipole. The cyclonic pole of the eddy had a maximum negative SLA of -15 cm in 2017 and -20 cm in 2018, while the anticyclonic center reached +24 cm in 2017 and +21 cm in 2018. The 2017 cruise samples were taken along a transect across the eddy dipole while the 2018 cruise targeted only the center of each eddy.

Environmental samples were obtained using the onboard CTD rosette to collect water from 15 meters, the deep chlorophyll maximum (DCM), and 175 meters during the 2017 MESOSCOPE cruise and from 25 meters and the DCM during the 2018 Falkor cruise. The DCM was determined visually from fluorometer data during the CTD downcast and Niskin bottles were tripped during the return trip to the surface. Seawater from each depth was sampled in triplicate by firing one Niskin bottle for each sample. Samples were brought to the surface and decanted into prewashed (3x with DI, 3x with sampled seawater) polycarbonate bottles for filtration. Samples were filtered by peristaltic pump onto 142mm 0.2 μ m Durapore filters held by polycarbonate filter holders on a Masterflex tubing line. Pressures were kept as low as possible while still producing a reasonable rate of flow through the filter, approximately 250-500 mL per minute. Samples were then removed from the filter holder using solvent-washed tweezers and

placed into pre-combusted aluminum foil packets that were then flash-frozen in liquid nitrogen before being stored at -80 °C until extraction. A methodological blank was also collected by running filtrate through a new filter and then treated identically to the samples.

Culture samples used as the validation sets for this paper have been previously described by Durham et al. (2022) and on Metabolomics Workbench (Project ID PR001317).

4.6.2 Sample processing

Extraction of the environmental samples followed a modified Bligh & Dyer approach as detailed in Boysen et al. (2018). Briefly, filters were added to PTFE centrifuge tubes with a 1:1 mix of 100 μm and 400 μm silica beads, approximately 2mL -20 °C Optima-grade DCM, and approximately 3mL -20 °C 1:1 methanol/water solution (both also Optima-grade). Extraction standards were added during this step. The samples were then bead-beaten three times, followed by triplicate washes with fresh methanol/water mixture. Samples were then dried down under clean nitrogen gas and warmed using a Fisher-Scientific Reacti-Therm module. Dried aqueous fractions were re-dissolved in 380 μL of Optima-grade water and amended with 20 μL isotope-labeled injection standards. Additional internal standards were added at this point to measure the variability introduced by chromatography and ionization, and the reconstituted fraction was syringe-filtered to remove any potential clogging material. This aqueous fraction was then aliquoted into an HPLC vial for injection on the HILIC column and diluted 1:1 with Optima-grade water. A pooled sample was created by combining 20 μL of each sample into the same HPLC vial, and a 1:1 dilution with water half-strength sample was aliquot from that to assess matrix effects and obscuring variation (Boysen et al. 2018). Also run alongside the environmental samples were two mixes of authentic standards in water and in an aliquot of the pooled sample at a variety of concentrations for quality control, annotation, and absolute

concentration calculations. HPLC vials containing the samples were frozen at -80 °C until thawing shortly before injection.

The CultureData samples were re-run from the frozen aliquots for this paper. The Pptime sample processing is documented on Metabolomics Workbench where it has been assigned Project ID PR001317.

4.6.3 LC conditions

For the MESOSCOPE, Falkor, and CultureData samples a SeQuant ZIC-pHILIC column (5 µm particle size, 2.1 mm x 150 mm, from Millipore) was used with 10 mM ammonium carbonate in 85:15 acetonitrile to water (Solvent A) and 10 mM ammonium carbonate in 85:15 water to acetonitrile (Solvent B) at a flow rate of 0.15 mL/min. The column was held at 100% A for 2 minutes, ramped to 64% B over 18 minutes, ramped to 100% B over 1 minute, held at 100% B for 5 minutes, and equilibrated at 100% A for 25 minutes (50 minutes total). The column was maintained at 30 °C. The injection volume was 2 µL for samples and standard mixes. When starting a batch, the column was equilibrated at the starting conditions for at least 30 minutes. To improve the performance of the HILIC column, we maintained the same injection volume, kept the instrument running water blanks between samples as necessary, and injected standards in a representative matrix (the pooled sample) in addition to standards in water. After each batch, the column was flushed with 10 mM ammonium carbonate in 85:15 water to acetonitrile for 20 to 30 minutes. LC conditions for the Pptime samples are documented on Metabolomics Workbench where it has been assigned Project ID PR001317.

4.6.4 MS conditions

Environmental metabolomic data was collected on a Thermo Q Exactive HF hybrid Orbitrap (QE) mass spectrometer. The capillary and auxiliary gas heater temperatures were maintained at

320°C and 100°C, respectively. The S-lens RF level was kept at 65, the H-ESI voltage was set to 3.3 kV and sheath gas, auxiliary gas, and sweep gas flow rates were set at 16, 3, and 1, respectively. Polarity switching was used with a scan range of 60 to 900 m/z and a resolution of 60,000. Calibration was performed every 3-4 days at a target mass of 200 m/z. DDA data was collected from the pooled samples for high-confidence annotation of knowns and unknowns. All files were then converted to an open-source mzML format and centroided via Proteowizard's msConvert tool. For the Ptime samples, files were pulled directly from Metabolomics Workbench via Project ID PR001317 and used in their existing mzXML format.

4.6.5 Peakpicking, alignment, and grouping with XCMS

The R package XCMS was used to perform peakpicking, retention time correction, and peak correspondence (Colin A. Smith et al. 2006; Tautenhahn et al. 2008). Files were loaded and run separately for each dataset using the "OnDiskMSnExp" infrastructure. Default parameters for the CentWave peakpicking algorithm were used except for: ppm, which was set to 5; peakwidth, which was widened to 20-80 seconds; prefilter, for which the intensity threshold was raised to 10^6 ; and integrate, which was set to 2 instead of 1. snthresh was set to zero because there are known issues with background estimation in this algorithm (Myers et al. 2017a), and both verboseColumns and the extendLengthMSW parameter were set to TRUE. For retention time correction, the Obiwrap method was used except for the CultureData dataset, which was visually inspected and determined not to require correction (Benton et al. 2010). For the Obiwrap algorithm, the binsize was reduced to 0.1 but all other parameters were left at their defaults or equivalents.

Peak grouping was performed on the two environmental datasets and the Ptime data with a bandwidth of 12, a minFraction of 0.1, binSize of 0.001, and minSamples of 2 but otherwise

default arguments. CultureData's minFraction was raised to 0.4 but was otherwise identical.

Sample groups were constructed to consist of the biological replicates for all datasets. After peak grouping, peak filling was performed using the fillChromPeaks function with the ppm parameter set to 2.5. Finally, mass features with a retention time less than 30 seconds or larger than 20 minutes were removed to avoid interference from the initial and final solvent washes.

4.6.6 Manual inspection and classification

After the full XCMS workflow was completed, the mass features were visually inspected by a single qualified MS expert. For the Falkor and MESOSCOPE datasets, every mass feature was inspected, while only those features with a predicted probability of 0.9 or higher according to the two-parameter model produced above were inspected for the CultureData and Pptime datasets. Inspection consisted of plotting the raw intensity values against the corrected retention-time values for all data points within the m/z by RT bounding box determined by the most extreme values for the given feature. For this step, we decided to plot the entire feature across all files simultaneously rather than viewing each sample individually to both accelerate labeling and to more accurately represent what MS experts typically do when assessing the quality of a given mass feature (Figure 4.7). We also decided to ignore missing values and linearly interpolate between known data points rather than filling with zeroes. These EICs were then shown to an MS expert for classification into one of 4 categories: Good, Bad, Ambiguous, or Stans only if the feature appeared to only show up in the standards. The inclusion of the Ambiguous category allowed us to reduce the likelihood of disagreements between MS experts, as while we expect some interpersonal overlap between Good and Ambiguous and between Ambiguous and Bad, our heuristic exploration with several qualified individuals showed minimal overlap between Good and Bad between experts. Features classified as Ambiguous or Stans only were dropped

from the logistic regression fitting downstream. A few randomly-chosen features from the manually-assigned Good and Bad classifications are shown in Figure 4.7.

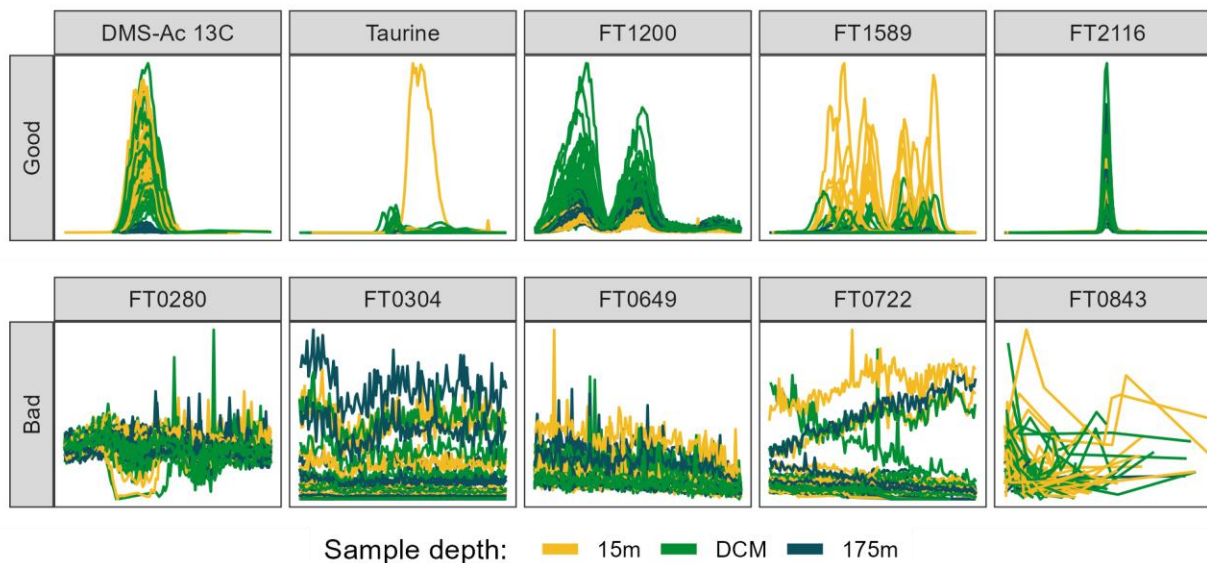


Figure 4.7: Randomly selected ion chromatograms from both “Good” (top row) and “Bad” (bottom row) manual classifications. Plots show retention time along the x-axis in a 1-minute window around the center of the feature and show measured intensity on the y. Features are from the MESOSCOPE dataset and colored by the depth from which the biological sample was taken. DCM = deep chlorophyll maximum, approximately 110 meters. Mass feature identifications are provided as the title of each panel, starting with “FT” and followed by 4 digits except for the two features annotated using authentic standards run alongside: the ^{13}C isotope of dimethylsulfonioacetate (DMS-Ac) and taurine.

4.6.7 Peak feature extraction and metric calculation

Our process of feature engineering involved querying several MS experts in our lab about their intuition for what they thought best distinguished poor-quality MFs and noise from good ones.

The simplest metrics to calculate were summary statistics of those parameters reported directly by XCMS. These features consisted of the mean retention time (RT) of each MF and the standard deviation (SD) within the feature and the mean peak width (calculated by subtracting the max RT from the minimum) and its SD. We also calculated the mean m/z ratio and the SD in parts-per-million (PPM) by dividing each peak’s reported m/z ratio by the m/z ratio of the feature as a whole, then multiplying by one million. Mean peak area was calculated by taking the $\log_{10}$

of the individual areas then taking the mean, and the same process ($\log_{10}$ then mean) was repeated for the SD of the peak areas. XCMS's default signal-to-noise parameter, `sn`, was also summarized in this way, but we only used `sn` values that were greater than or equal to zero and replaced any zeros with ones to avoid negative infinities after taking the $\log_{10}$. We also used the mean of other parameters reported by XCMS (`f`, `scale`, and `lmin`) as features. We additionally calculated several design-of-experiments metrics, using the number of peaks in each feature divided by the total number of files as well as the fraction of files in which a peak initially found by the peakpicker. This last metric was further subset into the fraction of samples in which a peak was initially found and the fraction of standards in which a peak was found (for those datasets in which standards were run). Finally, the coefficient of variance was estimated for the pooled sample peak areas by dividing the SD of the pooled sample peak areas by the mean of the same and additionally done in a robust way by using the median absolute deviation and median, respectively. For all of the above features, missing values were dropped silently from the summary calculations. We were unable to use any of the columns produced by enabling the `verboseColumns = TRUE` option in `findChromPeaks` because all of the values returned were NAs.

We also calculated several novel metrics from the raw m/z /RT/intensity values by extracting the data points falling within each individual peak's m/z and RT bounding box (values between the XCMS-reported min and max) separately for each file. The data points were then linearly scaled to fall within the 0-1 range by subtracting the minimum RT and dividing by the maximum RT, then each scaled RT was fit to a beta distribution with α values of 2.5, 3, 4, and 5, and a fixed β value of 5. This approach allowed us to approximate a bell curve with increasing degrees of right-skewness and the beta distribution was chosen because it is constrained between

0 and 1 and simple and speedy to generate in R. For each α value, Pearson's correlation coefficient (r) was calculated between the beta distribution and the raw data, with the highest value returned as a metric for how peak-shaped the data were (Figure 4.8). The beta distribution with the highest r was also then used to estimate the noise level within the peak by scaling both the beta distribution probability densities and the raw data intensity values as described above, then subtracting the scaled beta distribution from the scaled intensity values, producing the residuals of the fit (Figure 4.8). The signal-to-noise ratio (SNR) was calculated by dividing the maximum original peak height by the standard deviation of the residuals multiplied by the maximum height of the original peak. This method of SNR calculation allowed us to rapidly estimate the noise within the peak itself rather than relying on background estimation using data points outside the peak, which may not exist or may be influenced by additional mass signals (Myers et al. 2017b). If there were fewer than 5 data points, a missing value was returned and dropped in subsequent summary calculations. Accessing the raw data values also allowed us to calculate the proportion of "missed" scans in a peak for which an RT exists at other masses in the same sample but for which no data was produced at the selected m/z ratio, divided by the total number of scans between the min and max RTs.

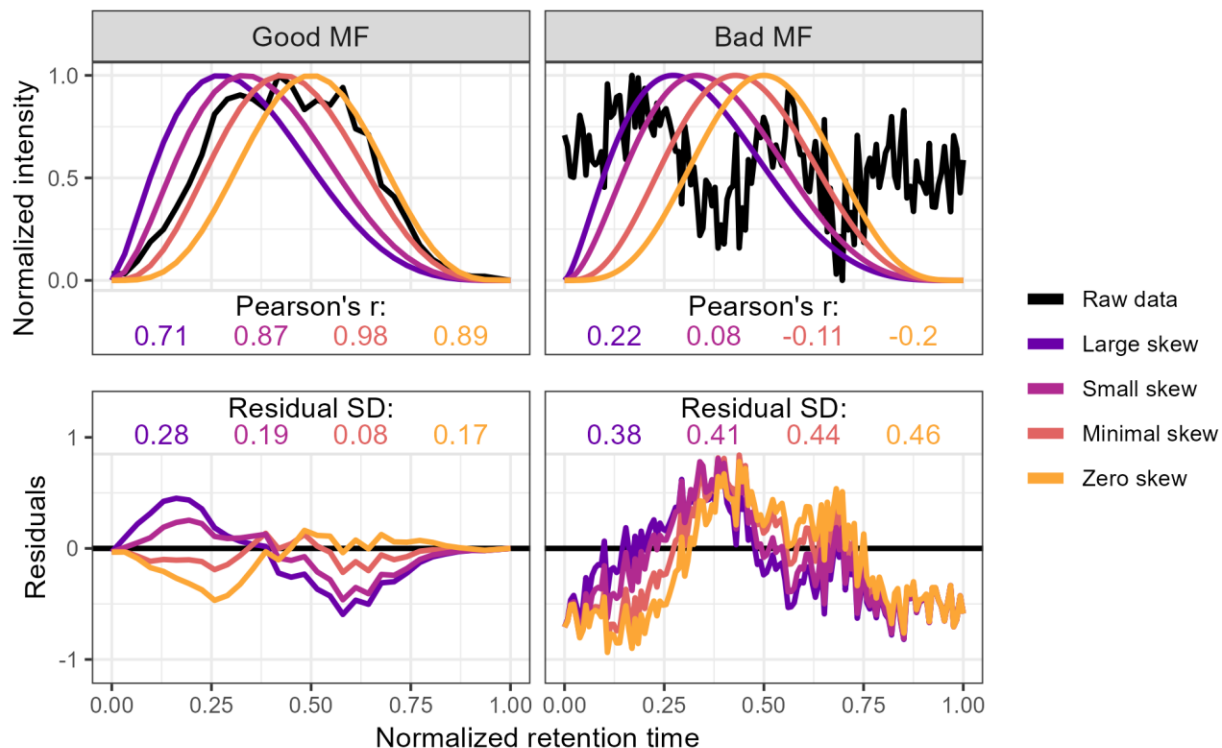


Figure 4.8: Method used to calculate the metrics for the two-parameter model from the raw data via comparison to an idealized pseudo-Gaussian peak for both manually identified “Good” and “Bad” peaks. Normalization was performed by linearly scaling the raw values into the 0-1 range by subtracting the minimum value and dividing by the maximum. Peak shape similarity was measured with Pearson’s correlation coefficient and the noise level is estimated as the standard deviation of the residuals after the raw data is subtracted from the idealized peak.

We additionally estimated the presence or absence of a ^{13}C isotope using a similar method to extract the raw m/z /RT/intensity values within the peak bounding box, then searched the same RT values at an m/z delta of $+1.003355 \pm 4$ PPM. In places where more than 5 data points existed at both the original mass and the ^{13}C mass, we again used Pearson’s correlation coefficient to estimate the similarity between the two mass traces and used a trapezoidal Riemann sum to estimate the area of the original and isotope peaks. The overall feature isotope shape similarity was calculated by taking the median of the correlation coefficients. We also calculated the correlation coefficient of the ratio of the $\frac{^{13}\text{C}}{^{12}\text{C}}$ peak areas across multiple files,

expecting that a true isotope would have a fixed $\frac{^{13}\text{C}}{^{12}\text{C}}$ ratio. Both the isotope shape similarity and the isotope area correlation were used as metrics in the downstream analysis. Peaks for which no isotope signal was detected or had too few scans to calculate the above metrics were imputed with NA values that were again dropped in the calculation of summary statistics for the mass feature as a whole. Because these isotope metrics typically had highly skewed distributions with most values very close to one, we normalized them by taking the $\log_{10}$ of one minus the value.

Distributions were visually inspected using a pairs plot and highly correlated (above a Pearson's $r \sim 0.9$) metrics had one of the redundant metrics removed.

4.6.8 Regressions and model development

We used three different multiple logistic regression models to predict the likelihood of each MF being categorized as “Good”. The first model included all metrics calculated as described above in Methods, the second contained only those parameters immediately available from the XCMS output without revisiting the raw data (the four core peak metrics m/z , RT, peak width, area and their standard deviations plus the mysterious lmin, f, and scale values as well as the fraction of peaks, samples, and standards found), and the final model was a simple two-parameter model using only the peak shape and novel SNR metrics.

In each case, we categorized each mass feature as a true positive (TP) if it was predicted to be Good and was manually classified as Good, a true negative if both predicted and classified as Bad, a false positive if predicted to be Good but manually classified as Bad, and a false negative if predicted to be Bad but was in fact manually classified as Good. This allowed us to additionally define two useful measures of success, the traditionally-defined false discovery rate (FDR, defined as 1-precision or the number of false positives divided by the total number of predicted positives) and the percentage of good features found (GFF, also known as the recall or

sensitivity and defined as the number of true positives divided by the total number of actual positives).

To further explore questions of model stability and the potential for overfitting, we compared the predictions from a Falkor-trained model to a MESOSCOPE-trained model. This comparison was done in both the raw probability space as well as a rank-ordered space to test whether the most extreme likelihood (i.e., very best and very worst) MFs were consistently found to be most extreme independently of the actual likelihood predicted. For the raw probability space we compared the predictions using Pearson's correlation coefficient, while Spearman's rank-ordered coefficient was used for the ranked space. We additionally looked at the estimates produced by these two models and compared them with the combined model trained on both datasets combined to assess the model stability directly.

We also measured the robustness of the model under a smaller training set, emulating a situation in which only a fraction of the data was available or only a portion of the mass features had been labeled. This allowed us to test the required sample size for the different models, with a larger sample size presumably required for the models with more parameters. Because no parameter was present in all 3 models, we looked at the top 2 most significant parameters from each model: average m/z and peak shape for the full model, average m/z and the standard deviation in retention time for the XCMS model, and peak shape and SNR for the two-parameter model.

Finally, we tested whether the performance could be improved with regularized regression or random forest models. These models handle correlated variables better than ordinary least squares regression, so we also included several additional implementations of the peak shape and novel SNR parameters when summarizing across multiple files, using a max and

a median of the top-three best values rather than just the overall median as well as a log-transformed version of the median peak shape calculated as $median(\log_{10}(1 - r))$ where r is Pearson's correlation coefficient, as described above (Figure 4.2). Cross-validation was used to select the optimal tuning parameter λ with `glmnet` package's `cv.glmnet` for an elastic net penalty (α) of 0, 0.5, and 1. Random forests were implemented using the `randomForest` package with default settings and a factor-type response vector to ensure classification was applied rather than regression.

4.6.9 Application of the model to novel datasets

After exploring the different models described above and determining that the two-parameter model would likely perform most consistently on novel datasets, we applied this trained model on two additional datasets that differed significantly from the training data. The CultureData dataset was produced in the Ingalls lab like MESOSCOPE and Falkor, but represent data from a variety of phytoplankton and bacterial cultures in fresh and salt water rather than environmental samples.

The Pptime dataset was discovered on Metabolomics Workbench where it has been assigned Project ID PR001317. The data can be accessed directly via its Project DOI: 10.21228/M8GH6P. This project dataset consists of *Phaeodactylum tricornutum* cultures collected at a variety of timepoints from both pelleted cells and the released exudate. This dataset was chosen because of the similar LC-MS setup used as a benchmark for the performance that other labs with similar setups may expect to achieve using the trained model directly.

Each of these datasets was only fractionally labeled, with those MFs above the 0.9 likelihood threshold according to the two-parameter model reviewed manually and categorized. This stricter threshold was chosen because we felt less comfortable interpreting results based on

mass features that were only 50% likely to be real, but did not feel the need to be so strict with this exploratory analysis that we wanted to limit it to 99+% likelihood MFs.

4.6.10 Using variable thresholds to determine effects on biological conclusions

We explored the implications of applying this model to the MESOSCOPE dataset at a variety of thresholds. In univariate space, we used nonparametric Kruskal-Wallis analyses of variance to measure the difference between the surface (15m), DCM (~110m), and 175m samples because the metabolite peak areas could not be assumed to be normally distributed. These univariate tests were then controlled for multiple hypothesis testing using R's `p.adjust` function with method `fdr` (Benjamini and Hochberg 1995). We also performed post-hoc Dunn tests provided by the `rstatix` package to categorize the response to depth for those mass features for which the KW test was significant, with responses falling into one of the 14 classes possible when permuting the sign and significance of the Dunn test outputs (Dunn 1964). p-values obtained from the Dunn tests were not FDR controlled because it was used as a categorization tool rather than a null hypothesis test. In multivariate space, we used a permutational MANOVA (PERMANOVA) (Anderson 2017) provided by the `vegan` package's `adonis2` function to test for multivariate differences in structure of the metabolome with depth (Oksanen et al. 2022). We ran multiple PERMANOVAs with a different subset of mass features included each time, corresponding to using the output from XCMS directly, likelihood thresholds of 0.01, 0.1, 0.5, 0.9, and finally only those MFs manually annotated as good.

All analyses were run in R (R Core Team 2022), version 4.3.1, and code is available on GitHub at https://github.com/wkumler/MS_metrics.

4.7 Abbreviations

- DCM: Deep Chlorophyll Maximum

- EIC: Extracted Ion Chromatogram
- FDR: False Discovery Rate
- GFF: Good Feature Found
- HILIC: Hydrophilic Interaction Liquid Chromatography
- LC: Liquid Chromatography
- MF: Mass Feature
- MS: Mass Spectrometry
- PPM: parts-per-million
- RT: Retention time
- SNR: Signal to Noise Ratio

4.8 Declarations

4.8.1 Ethics approval and consent to participate

Not applicable

4.8.2 Consent for publication

Not applicable

4.8.3 Availability of data and materials

The raw mzML files are all available on Metabolomics Workbench. The Falkor and

MESOSCOPE datasets can be found under project ID PR001738 via

<http://dx.doi.org/10.21228/M82719>. The CultureData samples were appended to the previously existing culturing collection, accessible at project ID PR001021 via

<http://dx.doi.org/10.21228/M8QM5H>. Pptime is located under Project ID PR001317 and can also be accessed directly using its Project DOI: <http://dx.doi.org/10.21228/M8GH6P>. Code and other raw data are available on the GitHub repository at https://github.com/wkumler/MS_metrics. The manuscript has been rendered as a single R Markdown document with analyses contained within for reproducibility.

4.8.4 Competing interests

The authors declare that they have no competing interests

4.8.5 Funding

This work was supported by grants from the Simons Foundation (SCOPE Award ID 329108 to AEI, SF Award ID 385428 to AEI).

4.8.6 Authors' contributions

WK extracted the Falkor samples, processed the data, performed the analyses, and wrote the manuscript. BJH helped design the metrics and implement the regressions as well as providing support and context for the analysis. AEI provided funding and data and helped to interpret the conclusions and edit the manuscript.

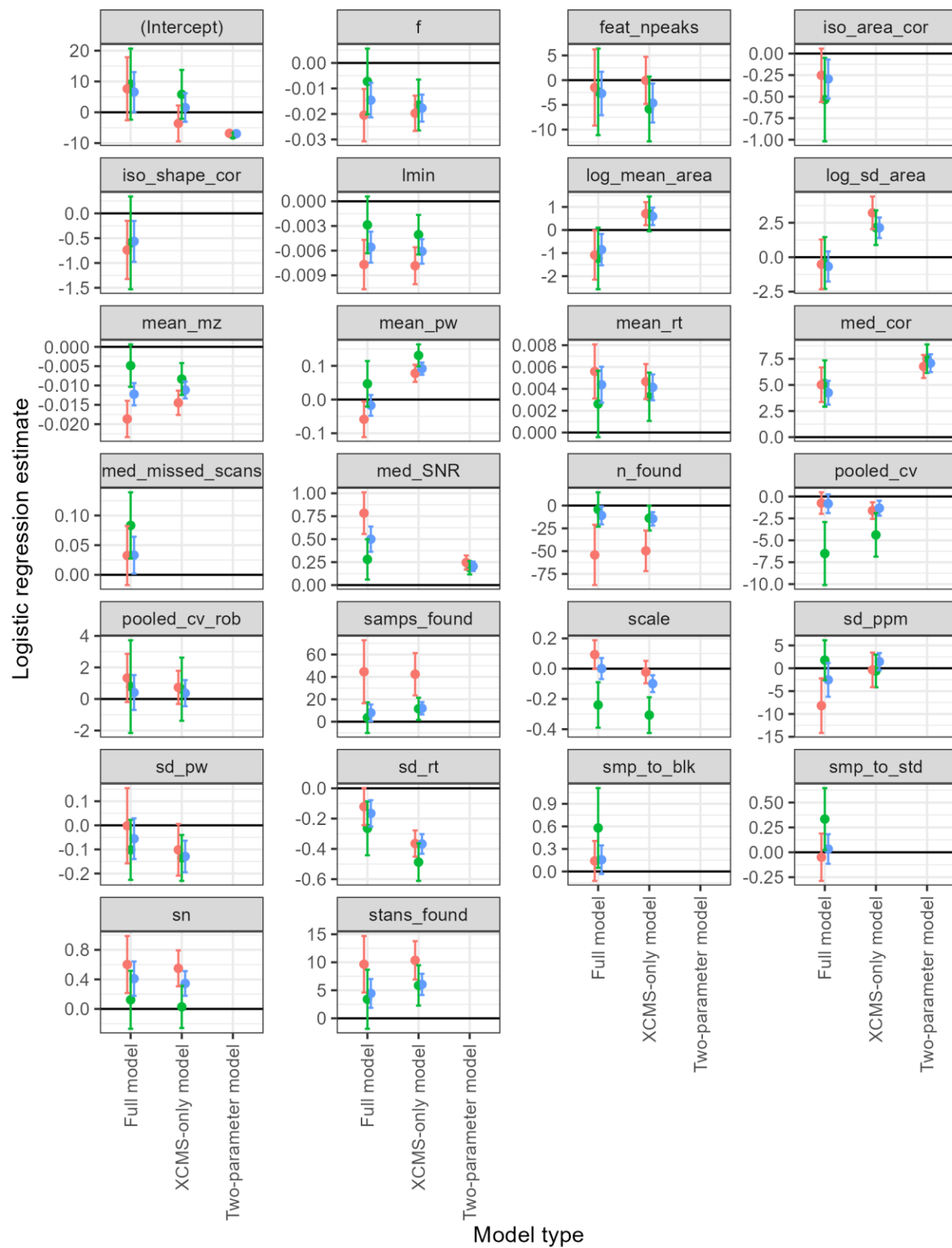
4.8.7 Acknowledgements

This work was supported by the University of Washington eScience Institute through their Data Science Incubator program. The authors would also like to acknowledge Laura Carlson for her expertise in obtaining the CultureData, MESOSCOPE, and Falkor datasets; Katherine Heal, Angie Boysen, and Bryn Durham for their assistance with culturing and collecting the cultured samples; the captain and crew of the R/Vs *Kilo Moana* and *Falkor*, Wei Qin, Rachel Lundeen, and the SCOPE ops team for collecting and processing the environmental samples; Joshua Sacks, Dave Beck, and the entire eScience incubator team for their feedback on the project development and scope; and Brisson Vanessa and LLNL for making their Pptime dataset available for reuse on Metabolomics Workbench. Additionally, the R packages plotly, dbscan, ClusterR, RaMS, and the entire tidyverse were crucial for preliminary data exploration.

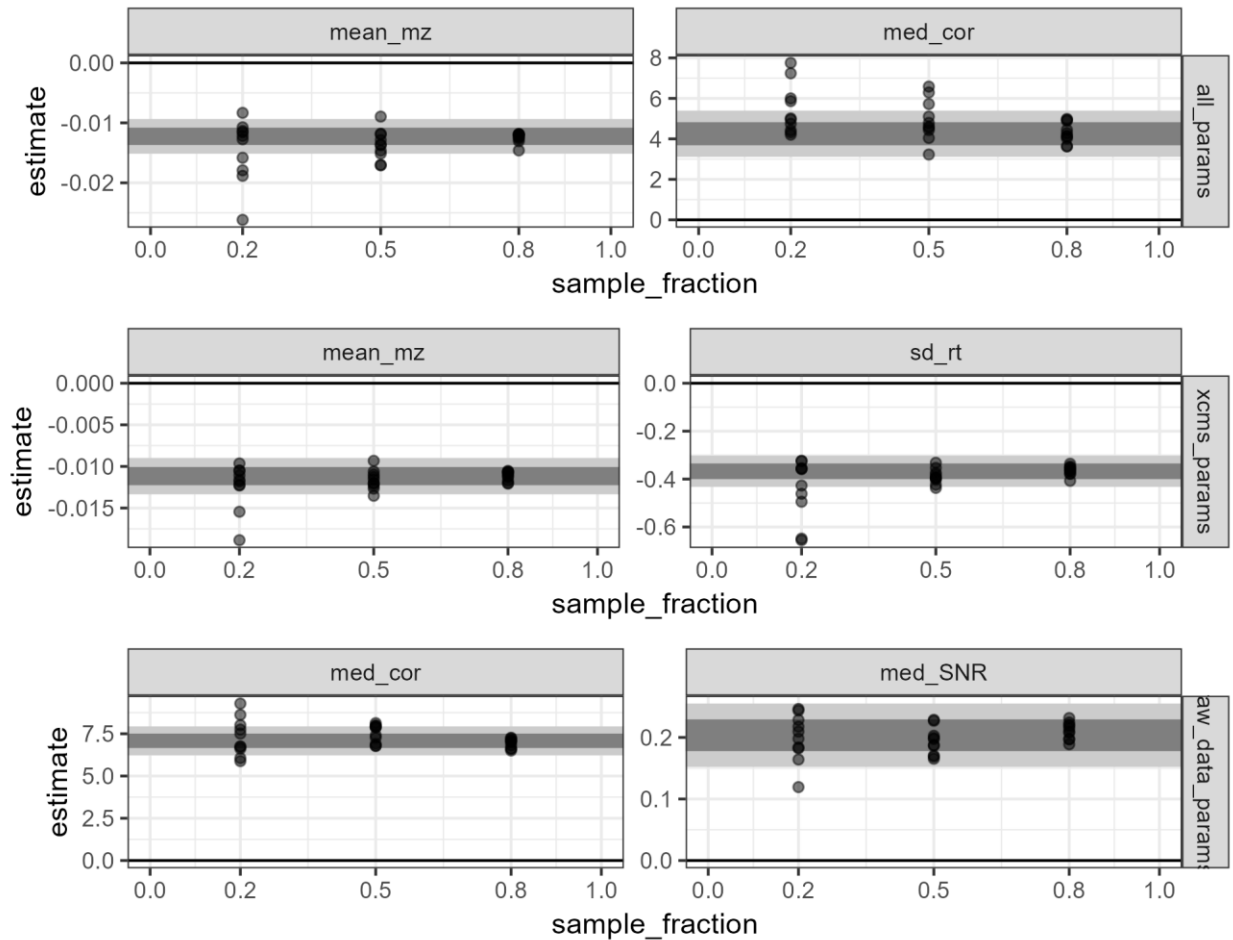
4.9 Supplemental Figures

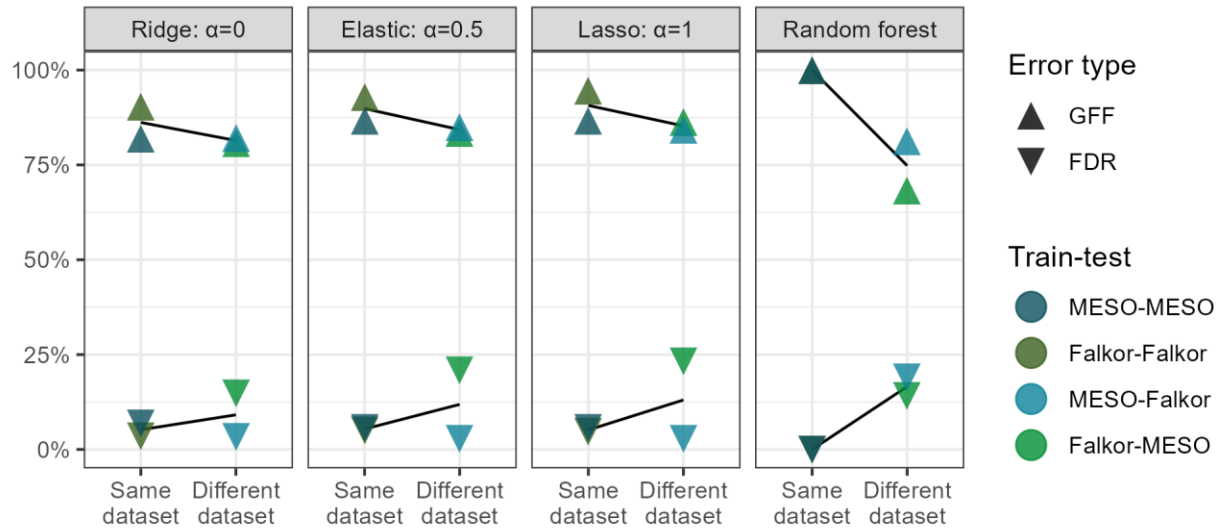
Supplemental Figure 4.9 (Available as a separate extended data file): Distribution and single-parameter logistic curves for each metric extracted for model training, shown separately for the MESOSCOPE and Falkor datasets. Histograms show the distribution of good and bad mass features by color across the span of the data on the x-axis with the number of MFs in each bin shown on the y-axis. Scatterplots show the same x-axis but show the results of a logistic regression on the single parameter, with the line of best fit in black and a ± 1 standard error ribbon around it in grey. Vertical jittering has been applied when plotting to reduce the number of overlapping points.

Training set — MESOSCOPE — Falkor — Combined



Supplemental Figure 4.10: Model parameter estimates for each of the metrics in the full model, additionally broken down by their inclusion in the two-parameter (*raw_data*) and XCMS-exclusively models. Colors correspond to the dataset used to train the logistic regression model, with “both” indicating a combined model using all manually-labeled features across both datasets.





Supplemental Figure 4.12: Performance of regularized regression and random forest models on internally (same train-test) and externally (different train-test) validated datasets.

4.10 Supplemental Tables

Available as an Extended Data file. Captions reproduced here for clarity.

Supplemental Table 4.1: Model parameters and significance according to both logistic and random forest models: Model outputs from the full (all parameters) model trained on the combined Falkor and MESOSCOPE datasets. The parameter name is reported alongside its more detailed name, with the logistic model estimate, standard error, test statistic, and p-value for each predictive term. We also included the estimate obtained from the elastic net model ($\alpha = 0.5$) and the accuracy and Gini index decreases per term. NA values are reported for several terms that were removed from the non-regularized regression because they were too highly correlated with other parameters.

Supplemental Table 4.2: Confusion matrices for logistic, regularized, and random forest regressions: Confusion matrices reported in long format for the different models reported in this manuscript. Models were all both internally tested using the same dataset for testing and training as well as cross-validated against the other fully labeled dataset. For the two-parameter model,

we also report the values obtained when training on the two datasets combined (train = Both). For each model and test-train set we report the absolute number of false positives (manually labeled as “Bad” but model labeled as “Good”), false negatives (manually labeled as “Good” but model labeled as “Bad”), true positives (both manual and model labelled as “Good”), and true negatives (both manual and model labeled as “Bad”) as well as the calculated false discovery rate (FDR, calculated as $FP/(FP + TP)$) and the fraction of good features found (GFF, calculated as $TP/(TP + FN)$). Unless otherwise specified, all values reported here used a likelihood threshold of 0.5 for categorization.

Supplemental Table 4.3: Internal standards used for normalization: Isotope-labeled internal standards used for the best-matched internal standard normalization procedure as denoted in Boysen et al. (2018)

Chapter 5: Metabolites Reflect Variability Introduced by Mesoscale Eddies in the North Pacific Subtropical Gyre

5.1 Abstract³

Mesoscale eddies significantly alter open ocean environments such as those found in the subtropical gyres that cover a large fraction of the global ocean. Previous studies have explored eddy effects on biogeochemistry and microbial community composition but not on the molecular composition of particulate organic matter. This study reports the absolute concentration of 67 metabolites and relative abundances for 640 molecular features, measured using liquid chromatography-mass spectrometry (LC-MS) following both targeted and untargeted approaches. This approach allowed us to better understand how mesoscale eddies impact the metabolome of the North Pacific Subtropical Gyre during two cruises in 2017 and 2018. We find that many metabolites track biomass trends, but metabolites like isethionic acid, homarine, and trigonelline linked to eukaryotic phytoplankton were enriched at the deep chlorophyll maximum of the cyclonic features, while degradation products such as arsenobetaine were enriched in anticyclones. In every analysis, the metabolites with the strongest responses were identified using LC-MS through untargeted metabolomics approaches, highlighting that the molecules most sensitive to environmental perturbation were not among the previously characterized metabolome. By analyzing depth variability (accounting for 20-40% of metabolomic variability across ~150 meters) and the vertical displacement of isopycnal surfaces (explaining 10-20% of variability across a sea level anomaly range of 40 centimeters and a spatial distance of 300

³ This chapter was published as Kumler, W., Qin, W., Lundeen, R.A., Barone, B., Carlson, L.T., and Ingalls, A.E. 2024. “Metabolites reflect variability introduced by mesoscale eddies in the North Pacific Subtropical Gyre.” *Frontiers in Marine Science* 11-2024. <https://doi.org/10.3389/fmars.2024.1481409>

kilometers), this analysis constrains the importance of mesoscale eddies in shaping the chemical composition of particulate matter in the largest biomes on the planet.

5.2 Introduction

High frequency observations at Station ALOHA in the North Pacific Subtropical Gyre (NPSG) over the past 25 years have revealed temporal and spatial variability in what had previously been considered a relatively homogenous environment (Karl 1999; Karl and Church 2017; Karl et al. 2021). A major source of variability comes in the form of mesoscale eddies in which water is entrained into circular surface currents tens to hundreds of kilometers in diameter (Karl and Church 2017; McGillicuddy 2016).

Eddies can be observed via satellite altimetry, which measures anomalies in sea surface height. Those that have a positive sea level anomaly (SLA) typically indicate regions where deep water layers and isopycnal surfaces are depressed and are associated with clockwise rotation in the Northern Hemisphere. In contrast, a negative SLA corresponds to counterclockwise rotation in the Northern Hemisphere and the uplift of deep water layers into the sunlit region of the ocean (McGillicuddy 2016). Mode-water eddies are an exception to the conventions established above (Sweeney et al. 2003; McGillicuddy et al. 2007) but here we only focus on the cyclonic (negative SLA) and anticyclonic (positive SLA) mesoscale features that are commonly observed near station ALOHA (Barone et al. 2019).

The uplift of deep, nutrient-rich seawater into the euphotic zone alters microbial communities (Rii et al. 2008, 2022). Measurements of chlorophyll in cyclones reveal a shallower and more intense maximum (Barone et al. 2022; Cornec et al. 2021) as a result of eukaryotic phytoplankton thriving in the higher nutrient concentrations while cyanobacterial biomass is reduced (Hawco et al. 2021). This growth of large eukaryotes corresponds to a net increase in

biomass and primary productivity (Benitez-Nelson et al. 2007), especially near the deep chlorophyll maximum (Barone et al. 2022). In contrast, anticyclones produce conditions favorable to nitrogen fixation by accumulating diazotrophs such as *Crocospaera* and *Trichodesmium* (Fong et al. 2008; Olson et al. 2015; Dugenne et al. 2020, 2023).

This shift in community structure and productivity should result in a corresponding shift in the composition of the particulate organic matter produced. Eukaryotic organisms have distinct chemical fingerprints from those of cyanobacteria and heterotrophic bacteria (Heal et al. 2021; Durham et al. 2022; Kuhlisch et al. 2023) and the relative contribution of these different taxa to total biomass should reflect their chemical composition. However, recent work from Harke et al. (2021) showed that overall community function may be robust across eddy types at the surface while Gleich et al. (2024) detected significant differences in protistan community composition and metabolic potential down to 250 meters, with heterotrophy-associated protistan transcripts enriched in the cyclone. The chemical composition of organic matter is one metric of community function since community metabolism and interactions are in part mediated through organic molecules, making their measurement useful for determining shifts in community dynamics.

In this work, we directly tested for differences in the chemical composition of particulate matter due to changes in eddy state in the NPSG. We sampled from cyclonic and anticyclonic eddies spatially near each other and used liquid chromatography-mass spectrometry (LC-MS) for both targeted and untargeted metabolomics to explore changes in the composition of small, polar molecules. We expected to find that overall metabolite abundance would reflect shifts in biomass over multiple depths. We also expected to see enrichment in the deep chlorophyll maximum of cyclonic features for those metabolites especially abundant in eukaryotic organisms. Finally, we

predicted that these patterns would be robust across years and sampling regimes for a reliable way to link these ocean features with the chemical composition of organic matter in the open ocean.

5.3 Materials and Methods

5.3.1 Cruise information

Samples were collected from two cruises in the North Pacific Subtropical Gyre near Station ALOHA that targeted strong mesoscale eddy features as described in Dugenne et al. (2023) and Gleich et al. (2024) (Figure 5.1). Briefly, the 2017 MESO-SCOPE cruise (Microbial Ecology of the Surface Ocean-Simons Collaboration on Ocean Processes and Ecology, KM1709 on the R/V *Kilo Moana*) consisted of a transect across adjacent cyclonic and anticyclonic eddies as well as two long-term Lagrangian stations at the center of each eddy. The cyclonic eddy had a maximum negative sea level anomaly (SLA) of -20 centimeters and the anticyclonic eddy reached +24 centimeters. The transect samples were taken at various times of day as the ship transected the adjacent eddies while the eddy center samples were all collected between 5 and 8 pm. In 2018, the Hawaiian Eddy Experiment (HEE, FK180310 on the R/V *Falkor*) targeted new cyclonic and anticyclonic eddies in approximately the same location with samples taken at the center of each eddy (maximum negative anomaly in the cyclone = -15 cm, maximum positive anomaly in the anticyclone = +26 cm) (Figure 5.1).

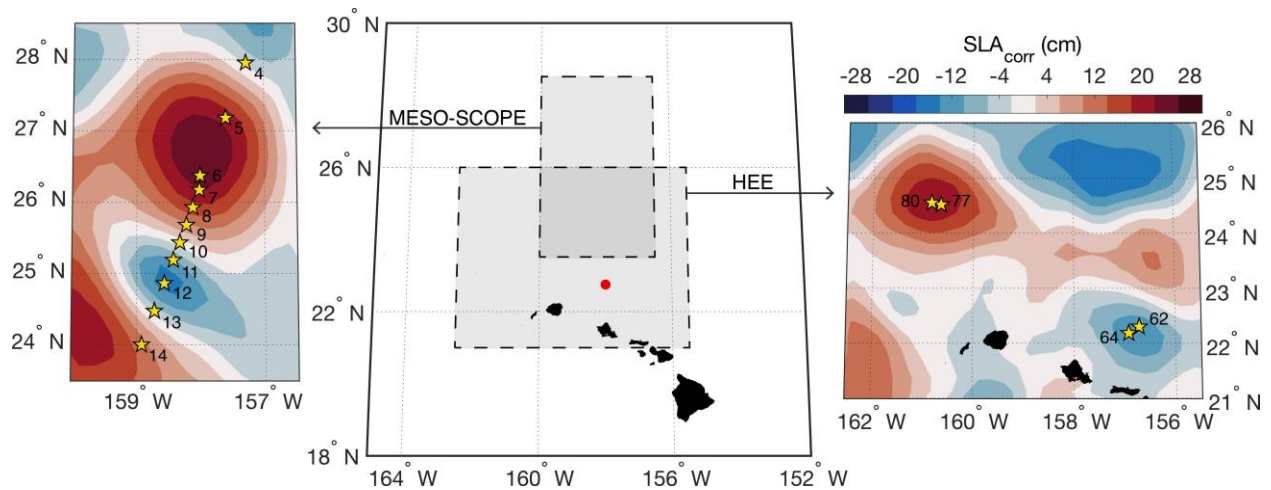


Figure 5.1: Sampling during MESO-SCOPE and the Hawaiian Eddy Experiment (HEE). Cruise bounds are shown in the large central map with Station ALOHA colored as a point in red near the Hawaiian Islands. Yellow stars denote sampling locations and station numbers relative to the sea level anomaly contours in the background during 28 June 2017 for the MESO-SCOPE cruise (left) and the 6 April 2018 for the HEE cruise (right). Lagrangian sampling near the eddy centers of the MESO-SCOPE cruise was performed by following drifters deployed in proximity of Stations 6 and 12.

5.3.2 Biogeochemical data

Biogeochemical measurements were collected as described in Barone et al. (2022) and Dugenne et al. (2023) mostly following protocols used by the HOT program (<http://hahana.soest.hawaii.edu/hot/methods/results.html>). Briefly, all environmental data was collected via CTD rosette except SLA which was measured via satellite. Particulate carbon and particulate nitrogen were measured using an elemental analyzer. Nitrate + nitrite (N+N) was measured on an autoanalyzer except where concentrations were below 100 nM in which case they were measured using chemiluminescence. Soluble reactive phosphorus (phosphate, PO_4^{3-}) was measured on an autoanalyzer or following the magnesium induced coprecipitation method. Values were interpolated linearly to the depths at which metabolite samples were taken using all data collected at the given station. Two particulate carbon values were determined to be spurious, with concentrations 2-3 times higher than typical Station ALOHA values, and were instead

estimated using a best-fit linear regression against beam attenuation. CTD data is available here in Supplemental Tables 5.1 and 5.2.

5.3.3 Sample collection for particulate metabolites

Samples were obtained using the onboard CTD rosette to collect water from 15 meters, the deep chlorophyll maximum (DCM), and 175 meters during the MESO-SCOPE eddy transect and from the DCM ± 10 and ± 20 meters during the eddy center sampling in the MESO-SCOPE cruise and from 25 meters and the DCM during the HEE cruise. The DCM was determined visually from fluorometer data during the CTD downcast and Niskin bottles were tripped during the return trip to the surface. Seawater from each depth was sampled in triplicate by firing one Niskin bottle for each sample. Samples were brought to the surface and decanted into prewashed (3x with DI, 3x with sampled seawater) polycarbonate bottles for filtration. 10L samples were filtered by peristaltic pump onto 142mm 0.2 μm Durapore filters held by polycarbonate filter holders on a Masterflex tubing line. Pressures were kept as low as possible while still producing a reasonable rate of flow through the filter, approximately 250-500 mL per minute. Samples were then removed from the filter holder using solvent-washed tweezers and placed into pre-combusted aluminum foil packets that were then flash-frozen in liquid nitrogen before being stored at $-80\text{ }^{\circ}\text{C}$ until extraction. A methodological blank was also collected by running filtrate through a new filter and then treated identically to the samples.

5.3.4 Metabolite sample extraction

Extraction followed a modified Bligh & Dyer approach as detailed in Boysen et al. (2018).

Briefly, filters were randomized and added to PTFE centrifuge tubes with a 1:1 mix of 100 μm and 400 μm silica beads, approximately 2 mL $-20\text{ }^{\circ}\text{C}$ Optima-grade dichloromethane, and approximately 3 mL $-20\text{ }^{\circ}\text{C}$ 1:1 methanol/water solution (both also Optima-grade). Isotope

labeled extraction standards were also added during this step (Supplemental Table 5.4). The samples were then bead-beaten three times, followed by triplicate extraction of the aqueous layer into a separate vial after spontaneous separation and replacement with fresh methanol/water mixture using a glass Pasteur pipette with additional bead-beatings in between. The aqueous fraction was then dried down under ultrapure nitrogen gas and warmed using a Fisher-Scientific Reacti-Therm module. Dry samples were reconstituted in 380 μ L Optima-grade water and amended with 20 μ L isotope-labeled injection standards (Supplemental Table 5.4) to measure the variability introduced by chromatography and ionization. The reconstituted fraction was syringe-filtered into precombusted glass inserts in liquid chromatography (LC) vials to remove any potential clogging material. Samples were then additionally diluted 1:1 with Optima-grade water to prevent overloading on the column and to reduce salt effects.

A pooled sample was created by combining 20 μ L of each sample into the same LC vial, and a 1:1 dilution with water created a half-strength pooled sample from that to assess matrix effects and obscuring variation (Boysen et al. 2018). Also run alongside the environmental samples were samples of authentic standards split into two mixes (Supplemental Table 5.5). Standards were dissolved in Optima-grade water to help with identification and into an aliquot of the pooled sample to quantify matrix response factors. LC vials containing the samples were frozen at -80 °C until thawing shortly before injection.

5.3.5 HPLC-MS methods

Separate LC-MS runs were used for the MESO-SCOPE eddy transect, the MESO-SCOPE eddy center, and the HEE cruise data. Eddy center samples were run in February 2018, eddy transect samples were run later during August of that year, and the HEE samples were run in July 2019. Each run was treated as a single batch and injected in sequence while maintaining the

randomization that had occurred during extraction to minimize chromatographic shifts from solvent or column switching.

For each run, a Waters Acquity I-Class UPLC with a SeQuant ZIC-pHILIC column (5 μ m particle size, 2.1 mm x 150 mm, from Millipore) was used with 10 mM ammonium carbonate in 85:15 acetonitrile to water (Solvent A) and 10 mM ammonium carbonate in 85:15 water to acetonitrile (Solvent B) at a flow rate of 0.15 mL/min. The column was held at 100% A for 2 minutes, ramped to 64% B over 18 minutes, ramped to 100% B over 1 minute, held at 100% B for 5 minutes, and equilibrated at 100% A for 25 minutes (50 minutes total). The column was maintained at 30 °C. The injection volume was 2 μ L for samples and standard mixes. When starting a batch, the column was equilibrated at the starting conditions for at least 30 minutes. To improve the performance of the HILIC column, we maintained the same injection volume, kept the instrument running water blanks between samples as necessary, and injected standards in a representative matrix (the pooled sample) in addition to standards in water. After each batch, the column was flushed with 10 mM ammonium carbonate in 85:15 water to acetonitrile for 20 to 30 minutes.

The Waters Acquity UPLC was coupled to a Thermo Q Exactive HF hybrid Orbitrap high resolution mass spectrometer equipped with a heated electrospray ionization source (H-ESI). The H-ESI voltage was set to 3.3 kV and sheath gas, auxiliary gas, and sweep gas flow rates were set at 16, 3, and 1, respectively. The capillary and auxiliary gas heater temperatures were maintained at 320°C and 100°C, respectively. Full scan analyses were performed with polarity switching and a scan range of 60 to 900 m/z at a resolution of 60,000. The instrument was mass calibrated at 200 m/z before each run began and once again during the eddy transect sample set to ensure calibrations were always performed within 3-4 days. All data files were then

converted to an open-source mzML format and vendor centroided via Proteowizard's msConvert tool (version 3.0.19297, Chambers et al. (2012)). All mzML files have been uploaded to Metabolomics Workbench under Project ID PR001738.

5.3.6 Metabolomic data analysis

5.3.6.1 Full scan feature extraction for known molecules

Compounds for which we had an authentic standard (Supplemental Table 5.5) were manually integrated using Skyline (versions 4.1 and 21.2.0.470, Adams et al. (2020)) and MSDIAL (version 4.36, Tsugawa et al. (2015)) with the standard mixes used to ensure the correct peak was integrated by matching retention time. Compounds were removed from the analysis if they were visually assessed to have poor peak quality in the samples or peak areas in the methodological blanks similar to those in the samples. Skyline was used for both the eddy center samples and the HEE cruise data while MSDIAL was used for the eddy transect data due to the larger number of files in that dataset. Raw peak areas were then normalized to their best-matched internal standard (Supplemental Table 5.4) per Boysen et al. (2018) except that all compounds were normalized to their best match no matter how minimal the improvement. While significant analytical drift was detected during the longer eddy transect sample run, the use of internal standards was able to negate the instrument's general increase in sensitivity. Normalized peaks were then calibrated via a single-point standard curve as determined uniquely for each compound from the authentic standard mixes and used to convert normalized peak area into moles per μL injected and scaled to the amount of seawater filtered to provide a final estimate of environmental concentration for each compound in moles per liter of seawater filtered. Scripts used for this analysis are available at <https://github.com/wkumler/MesoscopeMetabolomicsManuscript/tree/master/targeted>.

5.3.6.2 Automated processing for detection of unknown features

Samples were also processed via the untargeted workflow detailed in Kumler et al. (2023). In short, XCMS (version 3.10.0, Colin A. Smith et al. (2006), Tautenhahn et al. (2008)) was used for feature detection with CentWave peakpicking, Obiwrap alignment, and peak density correspondence/grouping. This was followed by manual inspection of the extracted ion chromatogram for each mass feature by an expert for quality with only those peaks classified as “Good” used in the analysis here. Features were then matched with and normalized to an internal standard (Supplemental Table 5.4) as described above. Scripts used for this analysis are available at <https://github.com/wkumler/MesoscopeMetabolomicsManuscript/tree/master/untargeted>.

5.3.7 Statistics

All statistics were run in R version 4.4.0. We used multivariate statistics provided by the vegan package (version 2.6-6.1, Oksanen et al. (2022)) to assess the overall impact of various metrics across the metabolome followed by univariate statistics to investigate individually significant responses. Non-parametric tests were used when data were unlikely to obey parametric assumptions and permutational statistics were preferred wherever possible. Multiple comparison problems were controlled where necessary using a false-discovery rate correction (Benjamini and Hochberg 1995). For the analyses involving the eddy transect samples, sea level anomaly was treated as a continuous variable with the exception of the PERMANOVA, for which SLA was categorized as cyclonic if less than -5 cm and anticyclonic if greater than 5 cm, with a third category for stations in between (Anderson 2001). Both the MESO-SCOPE eddy center and HEE analyses treated SLA as a categorical variable. Where time of day was included as an explanatory factor in the PERMANOVA, it was binned into 6-hour increments starting from 3AM (e.g. Midday contains casts collected between 9AM and 3PM). Marginal effects are reported for the PERMANOVA analyses to avoid influencing the results with the order of the

terms in the model formula. Total sample size (n) for the eddy transect was 33 per depth (3 biological triplicates at each of 11 stations), while the eddy centers had an n of 15 per eddy (3 biological replicates at 5 depths) and the HEE data had an n of 24 (3 biological replicates at 4 stations (2 in the cyclone and 2 in the anticyclone) at 2 depths each).

5.4 Results

We explored the impacts of isopycnal uplift and depression on the composition of particulate organic matter across pairs of mesoscale eddies of opposite polarity in the North Pacific Subtropical Gyre (NPSG). The first eddy pair was the focus of the July 2017 cruise, Microbial Ecology of the Surface Ocean-Simons Collaboration on Ocean Processes and Ecology (MESO-SCOPE, KM1709 aboard the R/V *Kilo Moana*). During the first phase of our two-phase expedition, a transect was taken from the north to the south consisting of 11 stations across a pair of eddies (the same sampling sites as in Barone et al. (2022) and Dugenne et al. (2023), Figure 5.1). Samples for particulate metabolomic analysis were collected from 15 meters, the deep chlorophyll maximum (DCM), and 175 meters at each station onto 0.2 μm filters. The next phase of the MESO-SCOPE cruise focused on Lagrangian sampling in the two eddy centers, starting with the cyclone (SLA = -14 cm) before progressing to the anticyclone (SLA = 24 cm). There, metabolomics samples were taken from the DCM as well as 10 and 20 meters above and below it. The second eddy pair was targeted during the March 2018 Hawaiian Eddy Experiment (HEE, FK180310 aboard the R/V *Falkor*). During this followup cruise, a strong anticyclone (SLA = 21 cm) and a nearby cyclone (SLA = -13 cm) were targeted for a similar set of biogeochemical measurements described in Dugenne et al. (2023) and Gleich et al. (2024) (Figure 5.1). Metabolomics samples were taken from the center of each eddy as described above at both 25

meters and the DCM. We analyzed these three different datasets separately and discuss the findings from each in sequence below.

5.4.1 Metabolome variability across adjacent mesoscale eddies of opposing polarity during MESO-SCOPE explained by sea level anomaly variations

The metabolome clearly differed between cyclonic and anticyclonic samples, though the magnitude of this difference varied by depth. The largest differences in particulate matter composition were detected in the DCM and 175 meter samples, as shown by the NMDS in Figure 5.2.

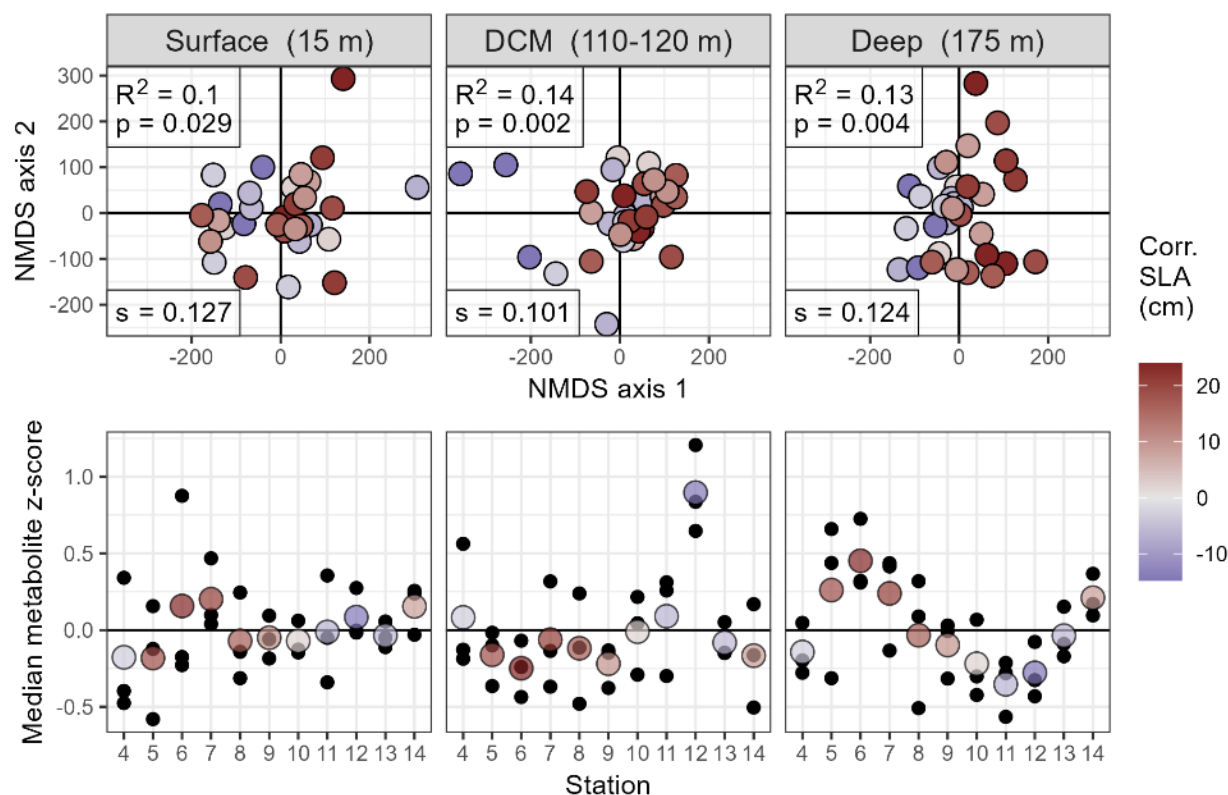


Figure 5.2: Distribution of metabolites in multivariate space across adjacent eddies of opposite polarity during the MESO-SCOPE transect, broken down by depth. Top panels depict non-metric multidimensional scaling (NMDS) plots with individual samples colored based on their corrected sea level anomaly. NMDS stress values (s) have been reported in the bottom left corner, while PERMANOVA R^2 and p -values are reported in the top left. SLA trends are visible in the DCM and 175 meter samples, with dark blue circles consistently discriminating from the dark red circles along the first multidimensional axis. Bottom panels depict the direction and

magnitude of this effect by plotting the mean value of all z-scored metabolites across three biological triplicates in color with the raw values in black behind.

At the DCM, approximately 14% of the variation in the metabolome could be explained by sea level anomaly (SLA) alone (PERMANOVA, $p = 0.002$). Additionally, SLA was strongly correlated with the first principal component (PC) of the metabolome with a Pearson's r of 0.615 and 27% of the variance explained by PC1 (Supplemental Figure 5.7A), indicating that this was one of the largest sources of variation in the dataset. Notably, the samples taken from the exact center of the cyclonic eddy at Station 12 were highly distinct and likely drove much of the explained variance. In the 175 meter samples, 13% of the variance was explained by SLA (PERMANOVA, $p = 0.004$) and SLA was highly correlated with the second PC of the dataset ($r = 0.697$, % variance explained by PC2 = 15.4, Supplemental Figure 5.7A), though the first PC did not seem to have any visible pattern with metadata and appeared to largely capture variation between biological triplicates (Supplemental Figure 5.7B). At 15 meters, SLA trends are slightly less evident with a larger p -value (PERMANOVA, $p=0.029$) and a lower R^2 (0.10) (Figure 5.2). Time of day was also a significant factor in the PERMANOVA at the surface and DCM (p -values of 0.018 and 0.033, respectively) while the diel effects in the 175 meter samples were unsurprisingly much lower ($p = 0.140$). The significance of SLA as an explanatory factor was similar even when excluding time of day from the model expression (R^2_{SLA} and p -values within permutation error).

To characterize the difference between eddies in a lower-dimensional space, we plotted an average metabolite peak area per station at each depth. We performed z-score normalization on each mass feature to give them all equal weight while preserving the variance between samples then calculated the median z-score for each sample, resulting in large positive values when metabolites are more abundant in a given sample than in the median sample (Figure 5.2).

The average metabolite collected at the exact center of the cyclonic eddy (Station 12) had higher values than other samples at this same depth, indicating that most metabolites were more abundant at the cyclone's DCM than elsewhere in the transect. The 175 meter samples, on the other hand, showed the largest metabolite abundances in the anticyclone while the cyclone had consistently lower median z-scores. This method highlighted the negative correlation between the average metabolite and SLA at the DCM ($r_{DCM} = -0.716$) while the 15 meter and 175 meter samples showed the opposite trend ($r_{15m} = 0.211$, $r_{175m} = 0.830$).

This shift in metabolite abundance was largely driven by a corresponding shift in biomass. Particulate carbon and summed metabolite concentration were tightly correlated across all samples and depths (Type I linear regression $\beta = 24.9 \pm 1.82$ (SE), Pearson's $r = 0.813$, p-value < 0.001 , Supplemental Figure 5.8). This overall trend was largely driven by higher values at 15 meters and the DCM than at depth, but when analyzed at each depth individually the general correspondence held (Supplemental Figure 5.8). The DCM and 175 meter samples had the stronger trends ($\beta_{DCM} = 22.6 \pm 5.53$, $r = 0.593$, p-value < 0.001 ; $\beta_{175m} = 23.9 \pm 7.36$, $r = 0.503$, p-value = 0.003) while 15 meter samples showed no significant relationship ($\beta_{15m} = 3.4 \pm 12.9$, $r = 0.049$, p-value = 0.792). This meant that metabolites contributed a relatively fixed fraction of carbon to the particulate pool with our 53 quantified metabolites representing between 1.5% and 5% of the carbon in the system (Figure 5.3). Particulate nitrogen was tightly correlated with particulate carbon ($r = 0.933$) but the fraction of particulate nitrogen represented by the quantified metabolites was slightly higher across the transect with contributions typically between 2% and 6%.

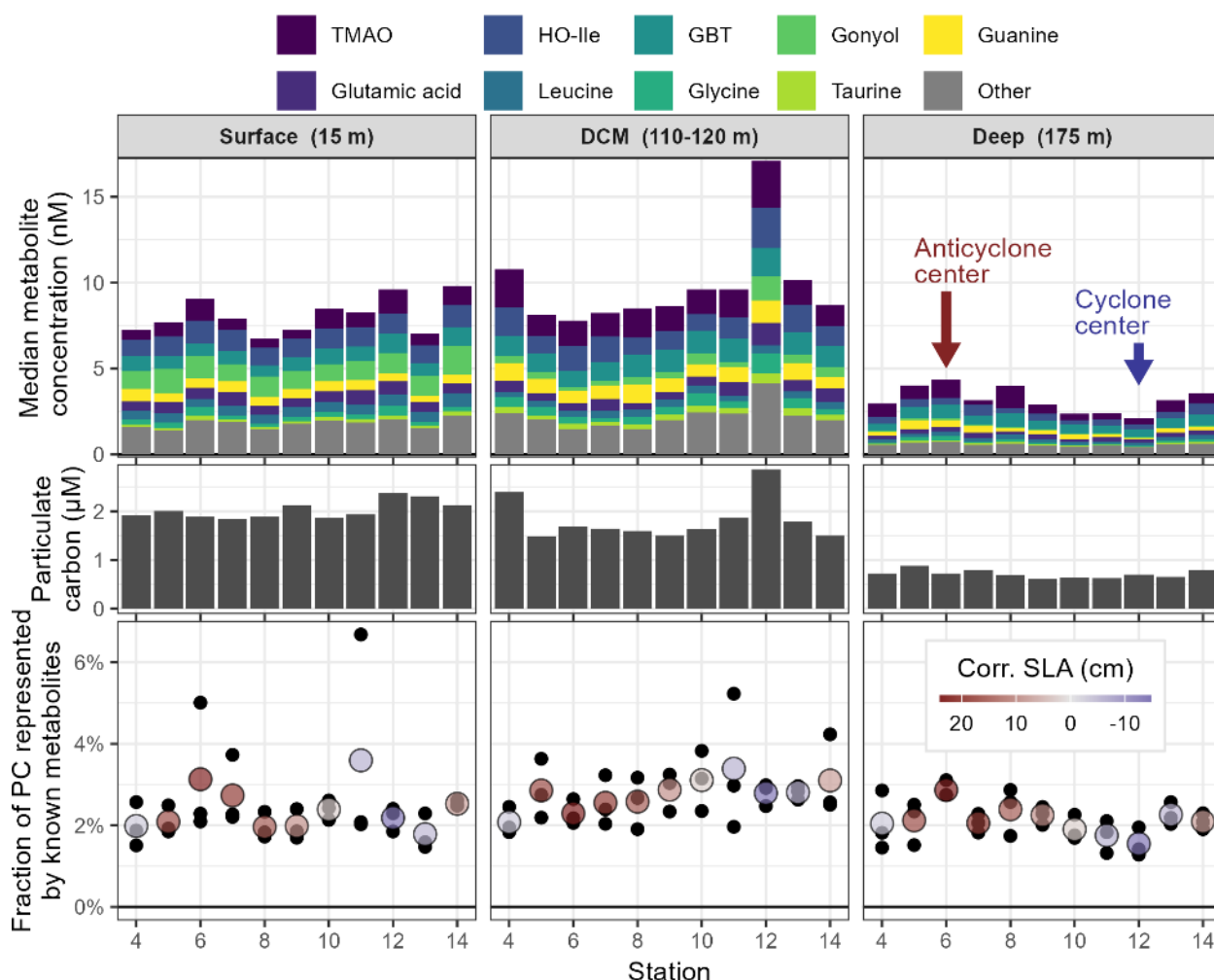


Figure 5.3: Differences in known metabolite concentration across the pair of adjacent eddies in the MESO-SCOPE transect separated by depth. Top panels depict concentrations of known compounds measured, where bar height corresponds to median triplicate concentration for each metabolite with the top 9 shown and the 44 other identified metabolites summed in grey. TMAO = trimethylamine N-oxide, HO-Ile = hydroxyisoleucine, GBT = glycine betaine, DCM = deep chlorophyll maximum. Center panels show the corresponding measurements of particulate carbon (PC), while the lower panels depict the fraction of total particulate carbon in the known metabolites. The mean value of three biological triplicates is shown in color and the raw values in black behind. Colors correspond to corrected sea level anomaly (Corr. SLA), with dark red indicating anticyclonic (positive SLA) and dark blue indicating cyclonic (negative SLA) eddy state.

Although the untargeted pool contains more molecular features, we expect that a majority of the signal has been captured by our 53 targeted compounds. Of the total peak area that passed quality control, approximately 67% was captured by these 53 molecules. An additional 15% of

the total peak area was putatively annotated as various inorganic ions for which no authentic standard was available but could be matched by mass and isotope pattern. Additionally, our targeted list accounted for 13 out of the top 20 largest molecular features by total peak area with an additional 4 features annotated as the inorganic ions for 17/20 of the largest features known. However, these calculations should be taken as estimates because peak area does not correspond directly to environmental concentration and it is entirely possible for abundant environmental compounds to have small peak areas if they ionize poorly on the mass spectrometer or are removed during the sampling and extraction process.

Samples taken from the DCM of the cyclone center (Station 12, Figure 5.2) showed especially high particulate carbon (67% increase, rising from 1.71 μM across all other DCM samples to 2.85 μM at Station 12) and the highest measured total metabolite concentrations (73% higher, increasing from 10.0 nM to 17.2 nM). This explains the large separation observed between these samples and the rest of the data in the NMDS plot of Figure 5.2.

The most abundant intracellular metabolite quantified in the eddy transect samples was trimethylamine N-oxide, with an average concentration of 1.38 nM and a distinct DCM maximum (Figure 5.3), though a few samples had concentrations an order of magnitude higher for unknown reasons (Supplemental Figure 5.9). A majority of the known molecules (37/53) had a similar pattern with a subsurface maximum at the DCM, including glycine betaine, glutamate, and guanine. Thirteen molecules obeyed a different pattern and decreased monotonically with depth, including the molecules hydroxyisoleucine, gonyol, and dimethylsulfoniopropionate (DMSP). Two known molecules (arsenobetaine and O-acetylcarnitine) increased with depth, potentially representing particle degradation during export.

We also measured several phosphorus-containing compounds and their non-phosphorus equivalents (Supplemental Figure 5.10A). Phosphocholine is the polar headgroup for phosphatidylcholine, a class of intact polar diacylglycerols that has been previously described at this study site as showing phosphorus stress relief among eukaryotic organisms in response to the upwelling of phosphate-rich deep sea water (Bent et al. 2024). We measured this metabolite as well as choline (the non-phosphate containing equivalent) and glycerophosphocholine (GPC, which contains an additional glycerol group) with the expectation that phosphorus stress would result in higher choline:phosphocholine and choline:GPC ratios, as shown in Seelen et al. (in review). However, we saw no differences in this ratio across the eddy transect after accounting for depth effects (Supplemental Figure 5.10C). Additionally, choline:GPC ratios were highest at 175m where PO_4^{3-} concentrations are highest, indicating that this metric may not accurately measure P-stress. Other phosphorus-containing lipid headgroups glycerophosphoglycerol and glycerophosphoethanolamine also showed no trend across the eddy at a given depth and instead showed trends functionally identical to the average metabolite in Figure 5.2 (Supplemental Figure 5.10B).

5.4.2 Particulate organic matter compositional shifts across eddy transect

In addition to absolute shifts in metabolite concentration due to biomass differences, we also explored compositional shifts in the metabolomes by assessing the fraction of the total metabolite pool contributed by each compound across the eddy transect. This allowed us to isolate the signal due to shifts in community composition and organismal response from those introduced by changes in biomass of the community as a whole.

In the 15 meter surface samples, only seven compounds had significant changes in relative peak area across the eddy transect, two of which were known: 5-oxoproline, which

decreased from 0.2-0.3% of the total peak area at the center of the anticyclone to 0.05-0.1% in the cyclone; and a combined peak of sarcosine and beta-alanine which increased from 0.01% to 0.02-0.03%. The largest shift significant at an $\alpha = 0.05$ level was an unknown mass feature with $m/z = 189.12338$ and a retention time around 8.6 minutes, the relative contribution of which increased from approximately 0.2-0.3% in the cyclone to ~1.5% in the anticyclone.

At the DCM, twenty-two compounds changed significantly in relative peak area across the transect. Seven of these were known metabolites, of which trigonelline and homarine were most enriched in the cyclone while arsenobetaine was the only metabolite with a larger fraction of the total peak area in the anticyclone. Homarine showed a very large and highly significant shift, representing about 2.5% of the total peak area in the anticyclone but 7-8% in the cyclone center. Trigonelline (N-methyl niacin), an isomer structurally very similar to homarine but biologically distinct, had a surprisingly strong correlation with homarine (Pearson's $r = 0.855$). Its peak areas were consistently around one third of homarine's but still large enough to have the second-largest shift in relative peak area of the significantly different compounds. Similarly, arsenobetaine represented about 0.1% of the peak area in the cyclone and 0.5-0.6% of the total peak area in the anticyclone. The most significantly different compounds at the DCM in each direction, however, were both unknowns. The lowest p-value (6.4×10^{-6} after FDR correction) in the DCM data was a mass feature with an m/z of 173.09211, a retention time also around 8.6 minutes, and enrichment in the anticyclone with a putative chemical formula of $[M+H] = C_7H_{13}N_2O_3$, possibly glycylproline or prolylglycine. The next-lowest ($p = 2.5 \times 10^{-4}$ after FDR correction) was enriched in the cyclone and had an m/z of 275.0712 and a retention time of 11.4 minutes with a putative formula of $C_{18}H_{11}O_3$.

Finally, in the 175 meter samples we detected 43 mass features with a significant association with SLA. Notably, all detected nucleobases (guanine, adenine, and cytosine) were positively associated with SLA (higher values in the anticyclone than the cyclone), though the strongest associations were found in O-acetylcarnitine, betonicine, and tyrosine. Both O-acetylcarnitine and betonicine effectively quintupled their contribution to total peak area, shifting from approximately 0.1% to ~0.5% over the transect. Acetylcholine was the only known compound more relatively abundant in the cyclone than in the anticyclone along with one unknown of m/z 131.0340. The unknown at m/z 173.0921 with the strongest trend at the DCM again had the strongest trend in the 175 meter samples, here with an FDR-corrected p-value of 2.0×10^{-7} .

5.4.3 High vertical resolution sampling near eddy center DCM reveals distinct metabolome between cyclone and anticyclone

We further investigated the metabolomic response to SLA at the DCM by collecting samples taken at high-resolution depth intervals around the DCM at the center of both eddy poles. Here, we found that this ~40 meter depth range and the transition between the eddies explained similar amounts of variation in the data ($R_{depth}^2 = 0.253$, $R_{SLA}^2 = 0.238$) and were both highly significant factors (permutational p-values $\ll 0.001$, Figure 5.4). Additionally, both sample depth and SLA were correlated with the first principal component of the metabolite matrix ($r_{depth} = 0.766$, $r_{SLA} = 0.684$, fraction of variance explained by PC1 = 30.8%). The cyclonic samples in particular show a much larger spread than the anticyclonic ones, indicating larger sample variability in the cyclone DCM relative to the anticyclone. A depth gradient is visible along NMDS 2, with the deepest samples generally at the top right of the plot and the shallowest ones closer to the bottom (Figure 5.4).

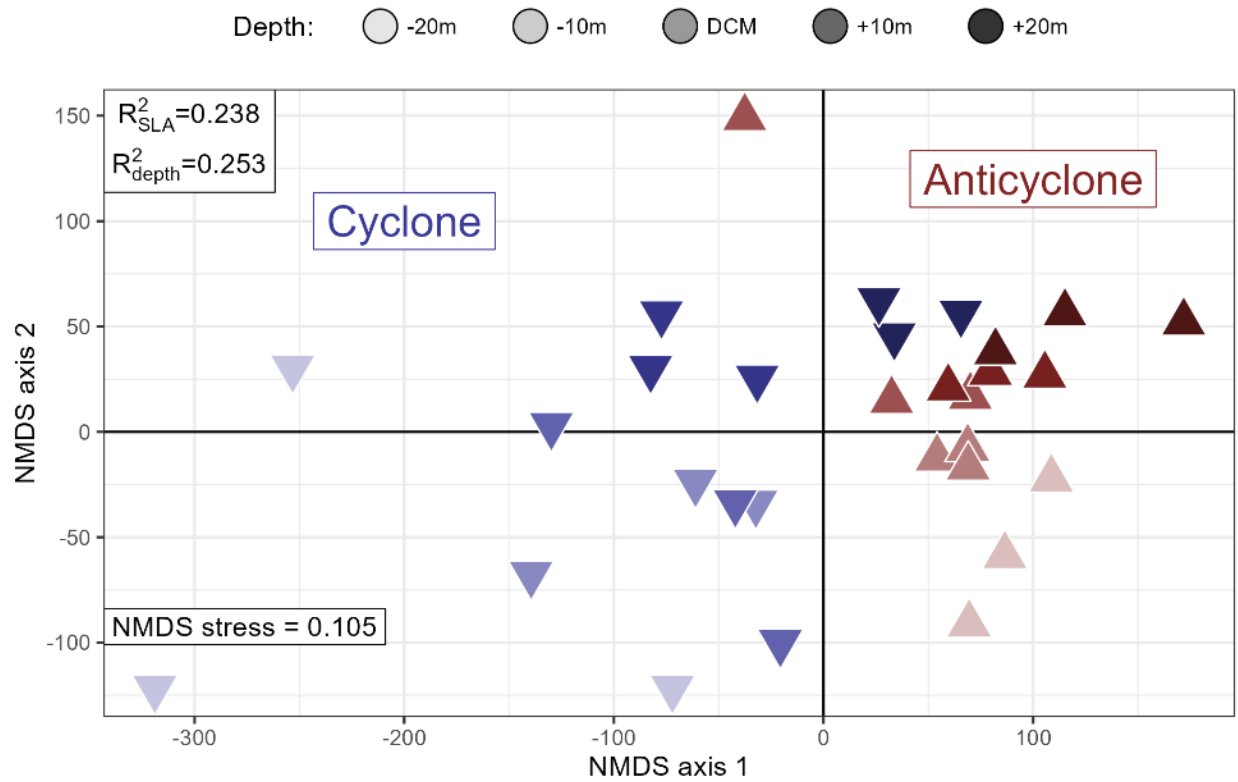


Figure 5.4: NMDS plot of high-resolution depth sampling around the deep chlorophyll maximum (DCM, ~115 meters) at the two eddy centers during MESO-SCOPE. Red upward-pointed triangles are from the anticyclone and blue downward-pointed ones are from the cyclone. Shading intensity reflects the depth above or below the DCM. PERMANOVA estimates of the variance explained by depth and sea level anomaly (SLA) are noted in the upper left corner and the NMDS stress value is reported in the bottom left.

To characterize the observed differences in multivariate space we used k-means clustering as an unsupervised way to identify dominant trends within the data (Figure 5.5). We found that a majority of the metabolites (Clusters 1 and 2, 63% of the total) fell into clusters with larger peak areas in the cyclone while also detecting 32 metabolites that clustered such that the mean metabolite was enriched in the anticyclone (Cluster 4, Figure 5.5). Cluster 2 had a distinct decrease in relative peak area with depth, while the other clusters had much less clear depth trends. Cluster 1 also showed a DCM maximum for the samples from the anticyclone, correlating

well with flow cytometry counts of picoeukaryotes (Pearson's $r = 0.851$) that was not present in the cyclone ($r = 0.013$).

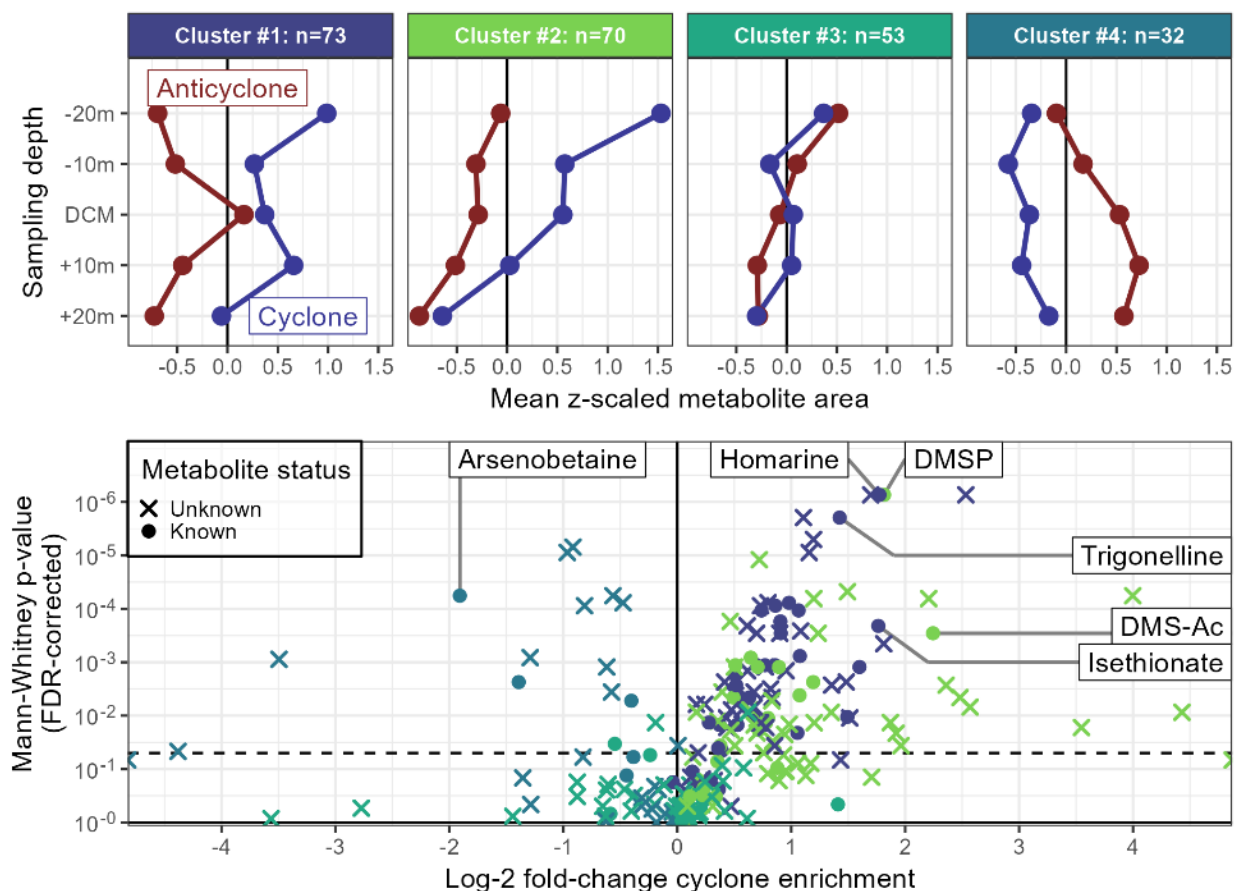


Figure 5.5: Distribution of metabolites in the high-resolution depth samples from the centers of each MESO-SCOPE eddy. The upper row of plots shows k -means clusters where points denote the average z -scored peak area for both known and unknown metabolites across the samples and are colored by the eddy from which they were taken. Clusters have been ordered by number of metabolites in each group and the total is denoted in the panel titles. Both depth trends (mostly a net decrease in metabolites with depth) and eddy effects (cyclonic enrichment in clusters 1 and 2, anticyclone enrichment in cluster 4) are observable. The lower plot shows the individual known and unknown metabolites where points correspond to the FDR-corrected p -value estimated by the nonparametric Mann-Whitney U test and the $\log_2$ fold-change calculated with the average peak area in the cyclone divided by the average peak area in the anticyclone. Colors have been assigned using the k -means clusters and shapes have been assigned based on the status of the mass feature as either a known metabolite that was matched to an authentic standard or an unidentified metabolite. The dashed line across the figure represents the 0.05 level of significance as a visual cue for metabolites above which the differences between the eddies are unlikely to be due to chance. DMSP = dimethylsulfoniopropionate, DMS-Ac = dimethylsulfonioacetate.

The majority of the individual metabolites in these high-resolution depth profiles differed significantly between the two eddies (121/228, $\alpha = 0.05$). Of those, 15 were enriched in the anticyclone and 106 were more abundant in the cyclone. As expected, all compounds enriched in the anticyclone were part of Cluster 4 and all those enriched in the cyclone belonged to either Cluster 1 or 2 (Figure 5.5).

Many of the known metabolites enriched in the cyclone function as osmolytes in the cell and might be enriched due to the increased eukaryotic phytoplankton biomass. However, some metabolites such as isethionate, the reduced sulfur osmolytes dimethylsulfoniopropionate (DMSP) and dimethylsulfonioacetate (DMS-Ac), and the isomers homarine and trigonelline, were enriched in excess of biomass (Figure 5.5). A few metabolites more abundant in the anticyclone were also given putative identifications based on RT and m/z matching with internal standards run at a later time on the mass spectrometer. Of these, the putative arsenobetaine was the most significantly different among these with a peak area at the anticyclone DCM nearly quadruple that of the cyclone (Figure 5.5). Of note, the 173.0921 m/z mass feature noted above as strongly enriched in the anticyclone was among the most significantly different between the two eddies, along with two isomers at an m/z of 170.1176 and retention times around 8-9 minutes that increased by a factor of 1.5 in the anticyclone (putative formula $C_7H_{14}N_4O$).

5.4.4 Hawaiian Eddy Experiment data reveals a different community and response to mesoscale eddies

Given the strong signals detected in the samples from the MESO-SCOPE cruise both across the eddy transect as well as in the high-resolution DCM sampling, we expected to find similar results in an analogous dataset. Samples were collected during a 2018 cruise on the R/V *Falkor* that again targeted both a cyclonic and an anticyclonic eddy in the North Pacific Subtropical Gyre near Station ALOHA as part of the Hawaiian Eddy Experiment (HEE).

The HEE data was characterized by high inter-replicate variability relative to the samples from the MESO-SCOPE cruise, with 25 meter samples in particular highly variable in multivariate space (Figure 5.6). Despite this, we still saw the importance of sea level anomaly as a significant explanatory factor in the dataset (PERMANOVA $R^2_{SLA} = 0.090$, p-value = 0.017) as well as the larger depth differences between the surface (25 meters) and DCM (110 - 120 meters) (PERMANOVA $R^2_{depth} = 0.251$, p-value $\ll 0.001$). We saw no difference between the 6 AM and 6 PM samples (PERMANOVA $R^2_{time} = 0.031$, p-value = 0.367) despite a few known compounds (trehalose and sucrose) showing large differences, likely due to a majority of the mass features demonstrating no diel effect (Supplemental Figure 5.11).

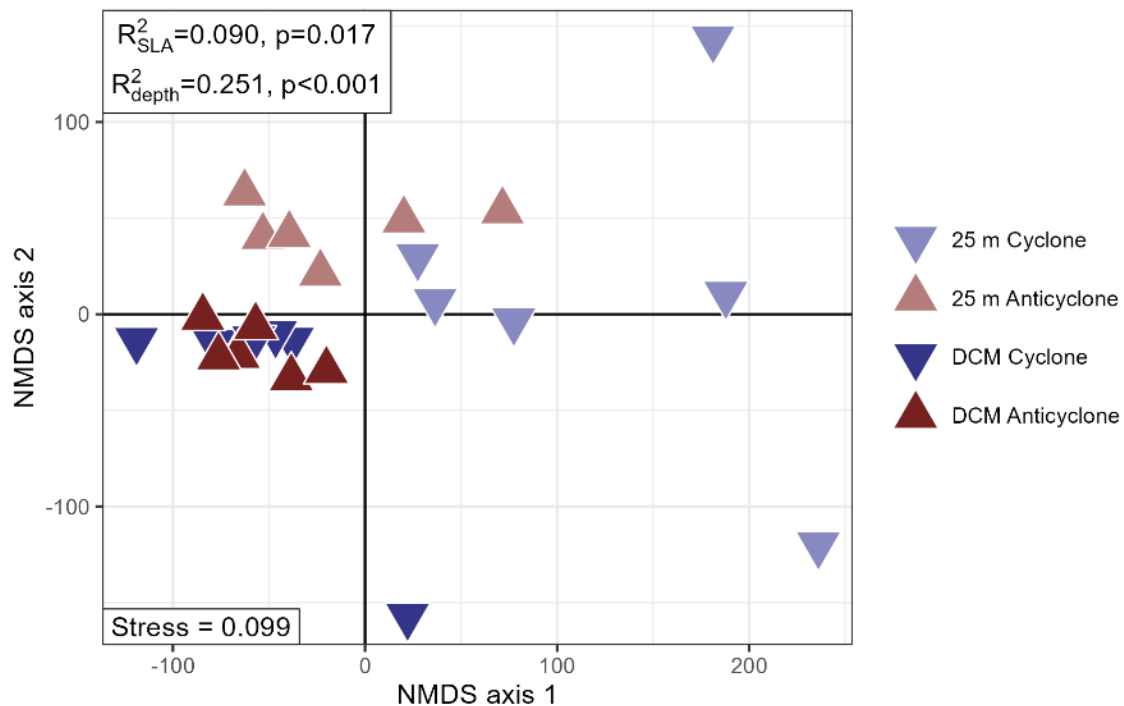


Figure 5.6: Non-metric multidimensional scaling (NMDS) plot from the Hawaiian Eddy Experiment cruise data, in which points correspond to individual samples and have been colored and shaped by their source eddy status and shaded by the depth from which they were collected. The NMDS stress value has been reported in the bottom left corner, while PERMANOVA R^2 and p-values are reported in the top left. Samples from 25 meters deep are visibly distinct from the deep chlorophyll maximum (DCM) samples and an SLA signal is visible in the 25 meter samples only.

When analyzed separately as distinct depths, the 25 meter samples had higher variances explained by eddy and a lower p-value (PERMANOVA, $R^2=0.26$, p-value=0.013), while the DCM samples were no longer likely to be distinct between the two eddies (PERMANOVA p-value = 0.171) (Figure 5.6, Supplemental Figure 5.12).

We were also able to use the HEE dataset to test whether the compounds detected as significantly different in the MESO-SCOPE dataset were also different in this eddy pair. We expected to find abundant TMAO and more hydroxyisoleucine at the surface than the DCM in addition to most compounds enriched in the cyclone DCM with biomass, especially the six compounds mentioned above in the results of Figure 5.5 (trigonelline, homarine, DMS-Ac, DMSP, taurine, and isethionic acid).

As expected, TMAO was again one of the most abundant metabolites detected in the particulate matter with concentrations around 1.2 nM at 25 meters and 0.4 nM at the DCM, second only to the high levels of the amino acid glycine (2 nM at 25 meters, 0.9 nM DCM), though the differences between 25 meters and DCM were not significant for either compound. We also detected a significant difference between depths for hydroxyisoleucine (25 meter mean = 0.32 nM, DCM mean = 0.21 nM, t-test p-value=0.022 with n=12 samples at each depth). These results imply that the overall depth structure of the metabolome of the gyre is relatively fixed for the most abundant compounds.

The eddy effects, on the other hand, were much less strong during this cruise. Of the six known metabolites with the strong enrichment in the cyclone at the DCM in the MESO-SCOPE cruise, only isethionic acid was significantly different in the HEE dataset (isethionic acid t-test p-value_{FDR} = 0.043, other five p-values > 0.25). This difference in isethionic acid concentration was only found after normalizing each sample to the sum of all metabolites in the sample to

control for biomass. Arsenobetaine was again found to be strongly enriched in the anticyclone at the DCM both when normalizing to biomass and when not, just as in MESO-SCOPE.

Surprisingly, the 173.0921 m/z mass feature was also slightly enriched but this time in the *cyclone* DCM with peak areas approximately 1.4 times the values in the anticyclone (t-test p -value = 0.067, Mann-Whitney p -value = 0.045). The 170.1176 m/z mass feature was not detected in the HEE samples.

5.5 Discussion

5.5.1 Multivariate approaches reveal metabolome-wide shifts across pairs of eddies of opposite polarity

We measured the metabolome of samples collected across two sets of adjacent eddies of opposite polarity to explore the effect of sea level anomaly (SLA) on metabolite composition and found that the altered biogeochemistry and microbial community composition between cyclonic and anticyclonic eddies explain a significant portion of the observed variations in their particulate metabolomes.

In all of the datasets analyzed here, we detected a significant difference in the composition of the metabolome between the adjacent eddies. This effect was strongest in the samples taken during the Lagrangian stations at the center of each eddy in the 2017 MESO-SCOPE cruise, with nearly a quarter of the total variance explained by the eddy from which the samples were taken. In the samples taken along the eddy transect, the largest effect was detected in the deepest samples taken from 175 meters and from the DCM, with much less of a response at 25 meters. This lack of SLA effect at the surface is unsurprising given the similarity in the biogeochemistry of the well-lit upper ocean previously reported (Barone et al. 2022). In the HEE samples, however, the opposite was true with a surprisingly larger SLA response at 25 meters than that at the DCM. In each case, the differences introduced by SLA were smaller than those of

depth when compared directly, with even the high-resolution sampling around the DCM finding slightly more variance explained by the 40 meter difference in sampling depth than the 40 centimeter difference in SLA. This result agrees well with prior research showing the large effect of sampling depth on the metabolome (Heal et al. 2021; Kumler et al. 2023; Bent et al. 2024). Time of day was included as a significant factor in the MESO-SCOPE transect, with trehalose and sucrose in particular showing large differences between 6 AM and 6 PM as has been previously noted (Muratore et al. 2022), though this effect was diminished at 175 meters.

Although possible that the observed differences in metabolomes across adjacent eddies differing in polarity are due to latitudinal shifts or simply background variation in the gyre environment, the transect data in particular implies that this is unlikely to be the case. Samples taken when the absolute value of the SLA was less than 5 cm were more similar to each other than to those samples taken within the eddies despite larger differences in latitude. The transect sampled both outside of each eddy and between the two across a large spatial gradient ($\sim 4^\circ$ latitude and 2° longitude), yet the non-eddy (absolute value of $SLA < 5$ cm) samples from all three locations grouped together, showed similar median metabolite concentrations, and were visually similar in the most abundant metabolite composition. These results imply that strong cyclonic and anticyclonic eddies represent endpoints in NPSG composition, in agreement with results from Barone et al. (2019) which found the most extreme values in the Hawaii Ocean Time series data typically detected when eddies passed over Station ALOHA. We also detected the largest differences in the DCM metabolome at the exact center of the eddy, with Station 12 highly distinct from even the nearest stations, while no such stark difference was found in the anticyclone (Figures 5.1 and 5.2). This indicates that the exact center of the cyclone has a unique

metabolome at the DCM relative to the rest of the transect, rather than existing as a smooth gradient with SLA.

Shifts in total biomass were a large driver of metabolomic differences, with earlier work in the same location by Barone et al. (2022) showing that particulate carbon, chlorophyll, and beam attenuation were all 20-80% higher in the DCM of the cyclone relative to the anticyclone. We see this reflected particularly well in the targeted metabolites measured here, with clear depth trends and a strong correlation between total metabolite concentration and particulate carbon values. However, even after controlling for biomass effects by normalizing each sample to the sum of signal measured within it we still found a significant SLA effect, likely due to shifts in the community composition and in particular the increased eukaryote presence in the cyclone DCM.

The NMDS plots of the high-resolution depth samples were also illustrative of sample dissimilarity with clear SLA and depth trends. The anticyclonic samples grouped tightly and showed little variance with depth or triplicate when compared to the cyclonic samples with the exception of DCM triplicate B, which was highly distinct for unknown reasons. In contrast, the cyclonic samples were much more variable as is perhaps expected when the biomass is in the form of larger phytoplankton that are less homogenous in the environment. One notable aspect of their clustering was the way in which the deepest samples (DCM plus 20m) grouped most closely to the anticyclonic samples, perhaps indicating that below the DCM the metabolome rapidly approaches a uniform deep-water signal. This result was noted previously in the 175 meter samples of the eddy transect in Kumler et al. (2023), where samples from the deep euphotic zone showed greater intra-depth similarity than samples from the DCM or 25 meters.

Given the confluence of both depth and SLA signals and the way that the median metabolite plots of Figure 5.2 would confound the SLA signal with the depth variance, we used k-means clustering to group similar compounds and provide a reduced dimensionality space for visualization. This revealed four major patterns of metabolite response, with a majority of compounds responding to eddy state (Clusters 1, 2, and 4 in Figure 5.5). It is interesting to note that the sole cluster in which abundances in the anticyclone are greater than those of the cyclone is also the only cluster to generally increase in abundance with depth (Cluster 4), likely because compounds more concentrated in the anticyclone are the same kind of degradation products and recalcitrant carbon typically found at depth.

The Hawaiian Eddy Experiment (HEE) data upset many of the expectations we developed during the MESO-SCOPE cruise in the previous year. Most surprising was the large difference between the eddies detected at the *surface*, while the DCM samples were functionally indistinguishable. This contrasts directly with the results from MESO-SCOPE and the results in Gleich et al. (2024), who found eddy-driven shifts in protistan community composition to be larger at depth than at the surface. However, Dugenne et al. (2023) found differences in nitrogen fixation rate and nitrogen fixer composition varying widely throughout the upper water column. The large surface differences were especially surprising given the large inter-replicate differences between the HEE surface samples in this study, while the DCM samples tended to be much more consistent.

5.5.2 Univariate approaches highlight individual metabolites responding to lifted and depressed isopycnals

Several molecule- or pathway-specific narratives emerged from the metabolomic data. The untargeted approach used here allowed us to describe and characterize signals from small molecules whose identity was unknown, a particularly promising approach in open ocean gyres

where the largest fraction of unknowns is found (Heal et al. 2021). We found that our list of authentic standards covered fewer metabolites in anticyclonic eddies at all depths, with the molecular diversity of the cyclone much better characterized. Many of the trends detected persisted even after normalizing within each sample, indicating that the shifts discussed have implications beyond simple scaling with biomass.

Of those molecules whose identity was known, the clearest response to SLA was that of the enigmatic osmolyte homarine at the DCM. This abundant compound increased approximately threefold in concentration from < 50 pM in the anticyclone to around 150 pM in the center of the cyclone in both the eddy transect and eddy center datasets. The pattern was weaker in the noisier HEE data but the largest concentrations were still detected in the cyclone and the median cyclone measurement was greater than the maximum value for the anticyclone. This molecule has a well-established role as an osmolyte in eukaryotic phytoplankton and cyanobacteria (Gebser and Pohnert 2013; Dawson et al. 2020; Heal et al. 2021; Durham et al. 2022) as well as a documented decrease in abundance with depth (Heal et al. 2021), indicating that its response to SLA is potentially due to differences in physical and chemical attributes that in turn control biomass, community composition, and recycling rates. Curiously, the distribution of homarine was also tightly correlated with that of trigonelline. Although they have structural similarity, they are not known to have a relationship beyond their shared function as osmolytes.

Isethionate is a known osmolyte thus far found exclusively in eukaryotes (Durham et al. 2022) and strongly associated with a few diatoms in particular (Heal et al. 2021). This compound was enriched in the cyclone of the DCM in both the MESO-SCOPE and HEE cruises along with its precursor taurine, validating earlier results from Barone et al. (2022) that showed strong enrichment of eukaryotic phytoplankton at the DCM and *Pseudo-nitzschia* in particular.

Additionally, the abundance of other sulfur-containing compounds (e.g. dimethylsulfonioacetate, DMS-Ac and dimethylsulfoniopropionate, DMSP) in the cyclone may indicate links between the formation of eukaryote-rich cyclonic features and the production of dimethylsulfide, a climate-active volatile gas (Moran and Durham 2019).

The strong SLA signal detected among the 175 meter samples was partially surprising to us given our expectation about the strongest effect at the DCM where eddy effects lift large concentrations of nutrients above the 1% light level. Instead, we found more compounds overall to be significantly different across the eddy transect in the 175 meter samples than we did at the DCM ($\alpha=0.05$, 43 compounds at 175 meters vs 22 at the DCM). This was likely driven by the enhancement of organic matter and heterotrophic picoplankton below the DCM as seen in Barone et al. (2019), with the enrichment of nucleobases in the 175 meter samples additionally hinting at increases in bacterial biomass given that nucleobases tend to be highly abundant in bacterial cultures (Heal et al. 2021). The abundance of a mass feature putatively identified as acetylcholine in the 175 meter cyclone samples followed the same general trend as arsenobetaine, both enriched at depth and in the anticyclone, and may be another compound that results from heterotrophic degradation of phytoplankton metabolites (Durham et al. 2022). These results raise important questions about the depth at which the community is isolated from SLA effects, with additional data from Barone et al. (2022) and Gleich et al. (2024) indicating that even at 250 meters eddy effects are still discernable.

5.5.3 Metabolites link changes in biogeochemistry to community biology

The composition of particulate matter controls the turnover of organic carbon in the ocean, but it is still unclear how this composition is altered in the environment and the plankton community. Measuring metabolites can quantify the variability of the central molecules responsible for the

flow of energy and matter through the marine ecosystem at many different scales. In these results, we focused on mesoscale eddies, which perturb the ocean ecosystem on relatively small horizontal scales on the order of tens to hundreds of kilometers.

Here, we reported how changes in the distribution of water layers of different density driven by mesoscale physical dynamics cascading effects in the biological and chemical characteristics of the eddies. The enrichment of eukaryotic organisms at the DCM of the cyclone mirrors other ocean processes where upwelling of deep, nutrient-rich water leads to increases in sulfur-rich metabolites representative of a shift towards organisms such as diatoms and dinoflagellates that undergo bloom dynamics (Durham et al. 2022; Kuhlisch et al. 2023). Barone et al. (2022) proposed that a “chemical wake” is formed after cyclone intensification when a transient phytoplankton community with net photosynthetic production induces export and leaves deficits and excesses of inorganic nutrients and oxygen, respectively. This conceptual model that can be extended to include many of the metabolites named here, as some metabolites are more bioavailable than others and the “flavor” of the chemical wake will change throughout the eddy life cycle.

The prevalence of unknown metabolites relative to known metabolites in the anticyclone may also in part explain why the increased biomass in the cyclone did not result in large increases in export (Barone et al. 2022; Bent et al. 2024). Our list of known compounds covers a large fraction of the small molecules common to most organisms, indicating that unknown molecules are more likely to be rare and therefore more difficult to degrade or a result of prokaryote dominance. If the increased biomass in the cyclone is dominantly in the form of highly bioavailable compounds, it is not surprising that much of it can be easily degraded. This hypothesis is supported by Gleich et al. (2024), who found that heterotrophy-associated

transcripts were more abundant in the 2018 cruise cyclone than the anticyclone, indicating that there was fresh material for the heterotrophic bacteria to readily consume.

Almost all large phytoplankton classes in the 2017 eddy pair were enriched in the cyclone DCM relative to the anticyclone (particularly prymnesiophytes, dinoflagellates, diatoms, and green algae), while *Prochlorococcus* was significantly less abundant and *Synechococcus* was approximately equal (Barone et al. 2022). Thus, the metabolite data presented here does not help distinguish between large phytoplankton classes but instead provides additional cues about the shift from a metabolome dominated by cyanobacteria producers into one dominated by eukaryotic organisms. This shift occurs naturally over the life cycle of an eddy, with distinctions typically drawn between the eddy intensification, its mature phase, and eventual decay (Rii et al. 2008; McGillicuddy 2016; Kuhlisch et al. 2023; Barone et al. 2022). Increased diapycnal mixing of nutrients in mature cyclones can alter the DCM community without large effects on integrated primary productivity or export even months after eddy intensification (Barone et al. 2022). We here demonstrated how this dynamic is linked with large changes in the chemical characteristics of the particulate material.

5.6 Conclusion

This study reports how changes in the biogeochemistry and community composition due to mesoscale eddies affect the metabolome and thereby the composition of particulate organic matter. A transect across adjacent mesoscale eddies of opposite polarity showed that many metabolites track closely with metrics of overall biomass, with eddy effects stronger at 175 meters than at 15 meters or the deep chlorophyll maximum (DCM). High-resolution depth sampling of the DCM at the center of each eddy elucidated several known and unknown biomarkers likely corresponding to the previously documented increase in eukaryotic

phytoplankton in the cyclone. Metabolite clusters aligned well with expected trends in abundance with depth and between the two eddies, with convergence below the DCM towards a general deep-sea metabolite signal for many metabolites. By contrasting these effects with a follow-up analysis in the centers of a separate pair of mesoscale eddies of opposite polarity the following year, we learned that the impact of eddies may be influenced by their origin, age, and period of the year. Finally, the metabolites with the strongest responses to eddy effects were unknown, and a relatively smaller proportion of known metabolites in the anticyclone demonstrates the utility of untargeted metabolomics in exploring environmental variation and the large amount of information that is missed by only investigating known compounds. In this work, we have shown the sensitivity of the metabolome to various environmental and community factors and constrained the importance of eddy effects in models of elemental cycling.

5.7 Acknowledgements

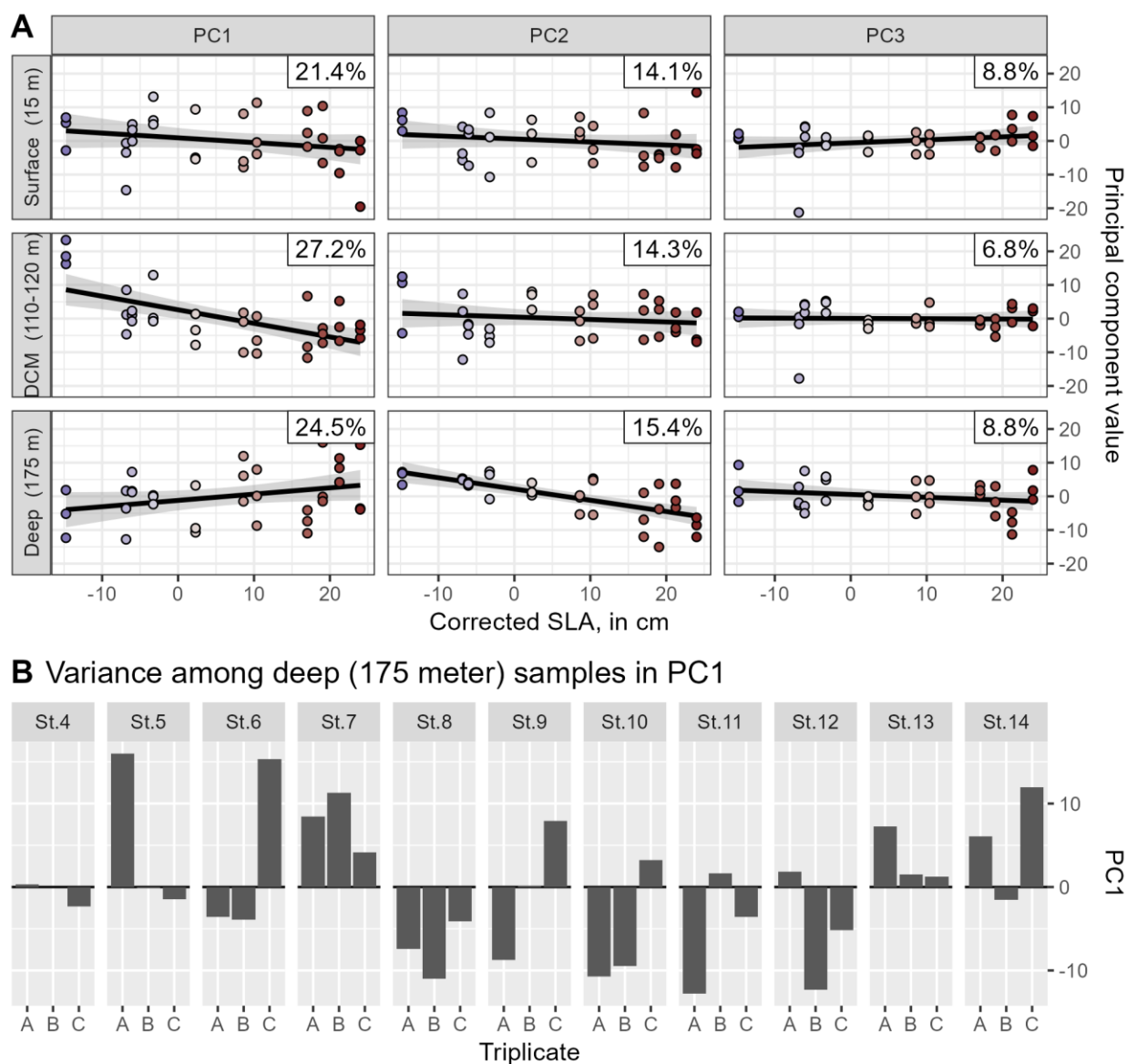
The authors would like to thank the other members of the Ingalls Lab who provided assistance in sample processing, integration, and analysis. We are also grateful to the SCOPE science team (Tara Clemente and Tim Burrell) for sample collection during the Hawaiian Eddy Experiment. We would like to acknowledge the captain and crew of the R/V *Kilo Moana* and R/V *Falkor* during the KM1709 and FK180310 cruises for making this science possible. We would also like to acknowledge the analysts involved in the biogeochemical measurements for their willingness to provide data and answer questions (Eric Grabowski, Karin Bjorkman, and Rhea Foreman). This work was supported by grants from the Simons Foundation (SCOPE Award ID 329108 to AI, SF Award ID 385428 to AI, Postdoctoral Fellowship in Marine Microbial Ecology ID 548565 to WQ, and SCOPE Award ID 721264 to David Karl which supported BB).

The authors have no conflicts of interest to declare.

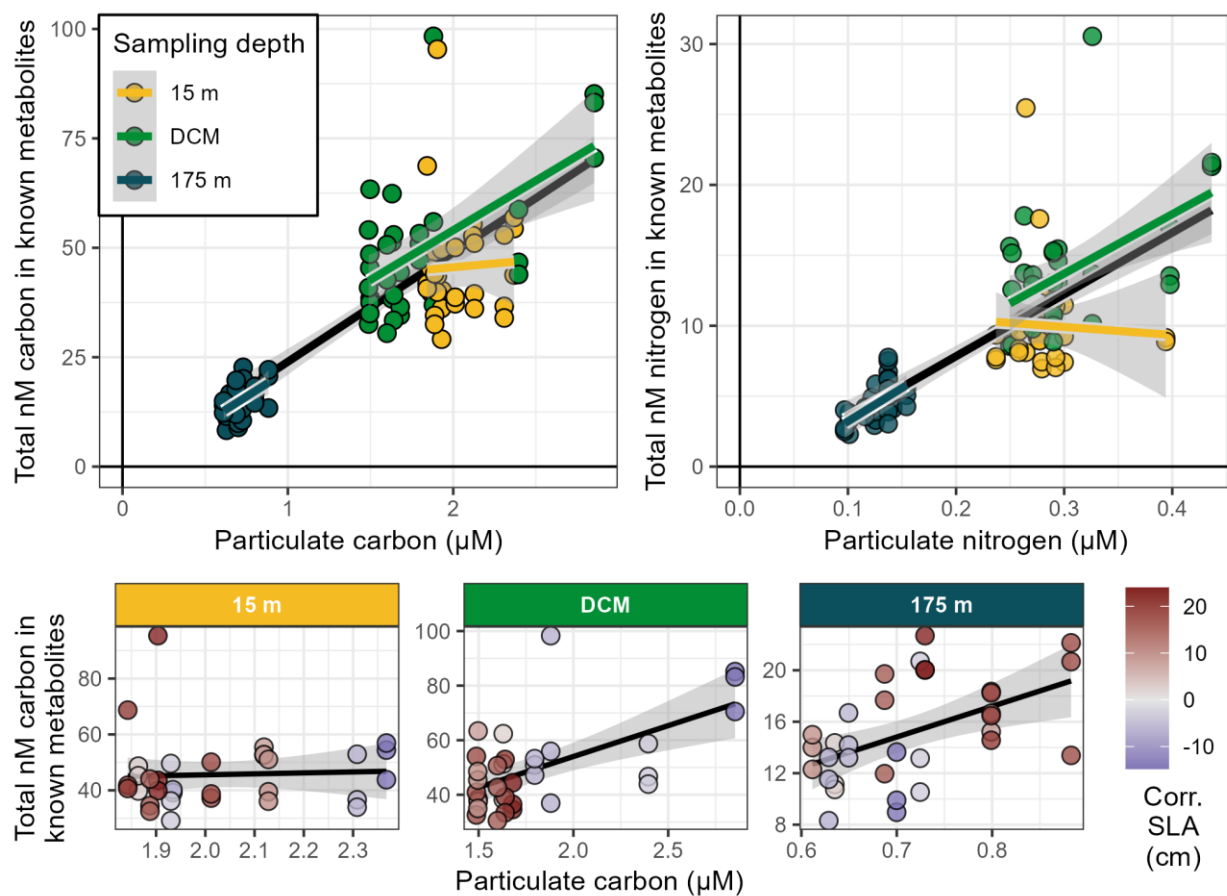
5.8 Data availability

Data and code used for this manuscript are all available online. Biogeochemical data can be sourced from the Simons SCOPE website at <https://scope.soest.hawaii.edu/data/> and has been provided in Supplemental Tables 5.1-5.3. Metabolomics data has been uploaded to Metabolomics Workbench under Project ID PR001738, where the data can be accessed directly via its DOI: 10.21228/M82719. Scripts used to process metabolomics data are available on GitHub at <https://github.com/wkumler/MesoscopeMetabolomicsManuscript>.

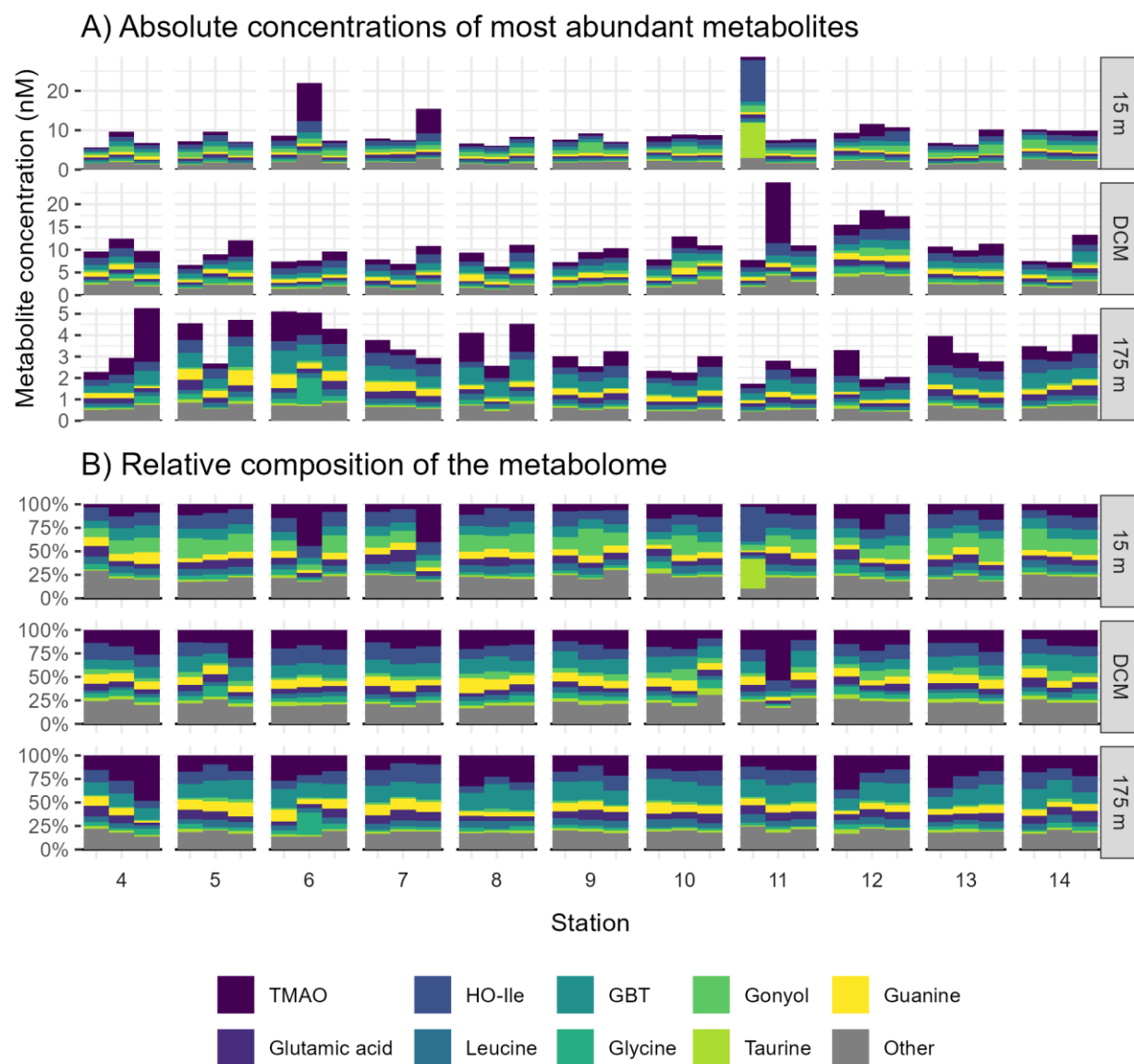
5.9 Supplemental Figures



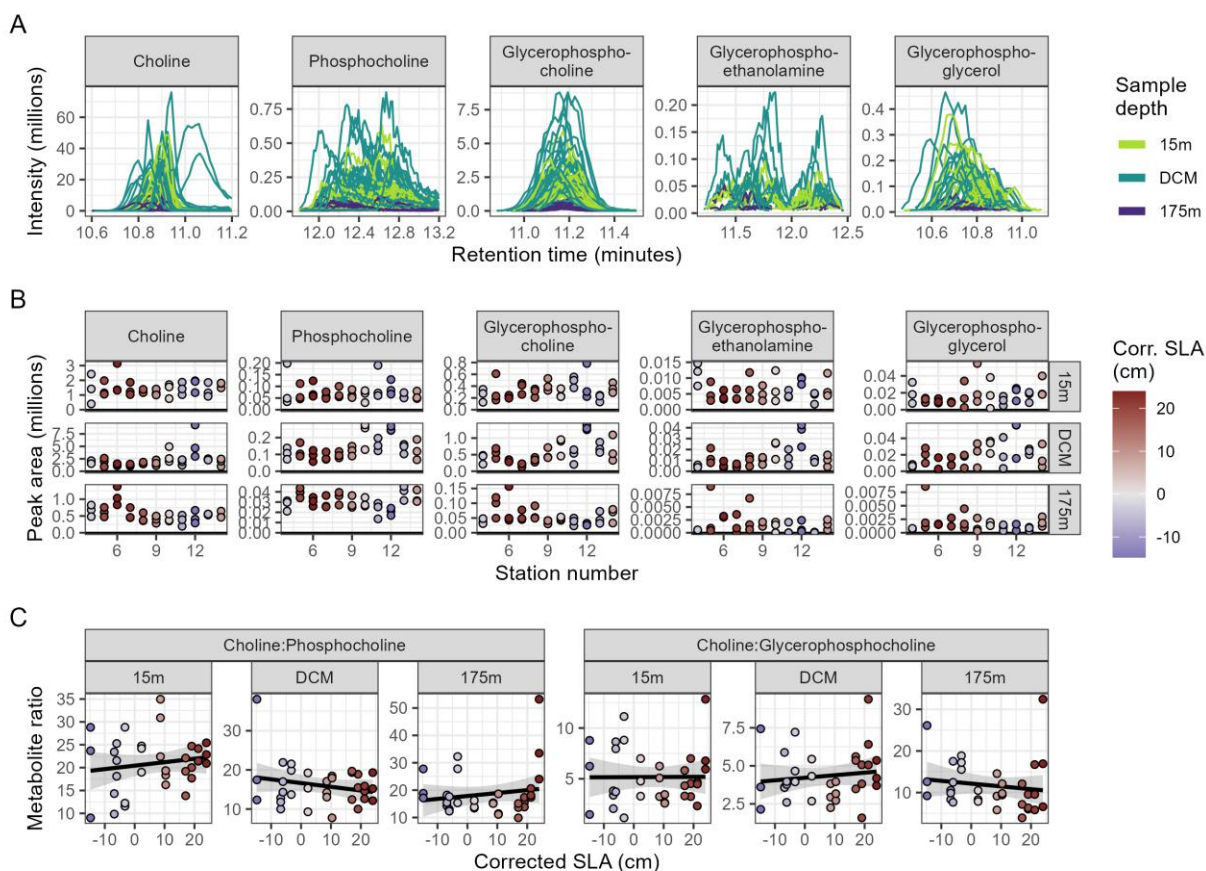
Supplemental Figure 5.7: Results of a principal component analysis performed separately on each depth of the eddy transect metabolome. (A) Regressions of SLA against various principal components, broken down by depth. % variance explained by each PC is noted in the upper right corner of the plot. Colors have been added corresponding to the SLA value for clarity but are redundant with the x-axis values. Lines of best fit and the associated standard error have been added behind the points in black and translucent grey, respectively. Strong associations exist between PC1 and SLA in the DCM samples and PC2 and SLA in the 175 meter samples. (B) Principal component 1 values for the 175 meter samples at each station showing large variance between triplicates rather than any external factor.



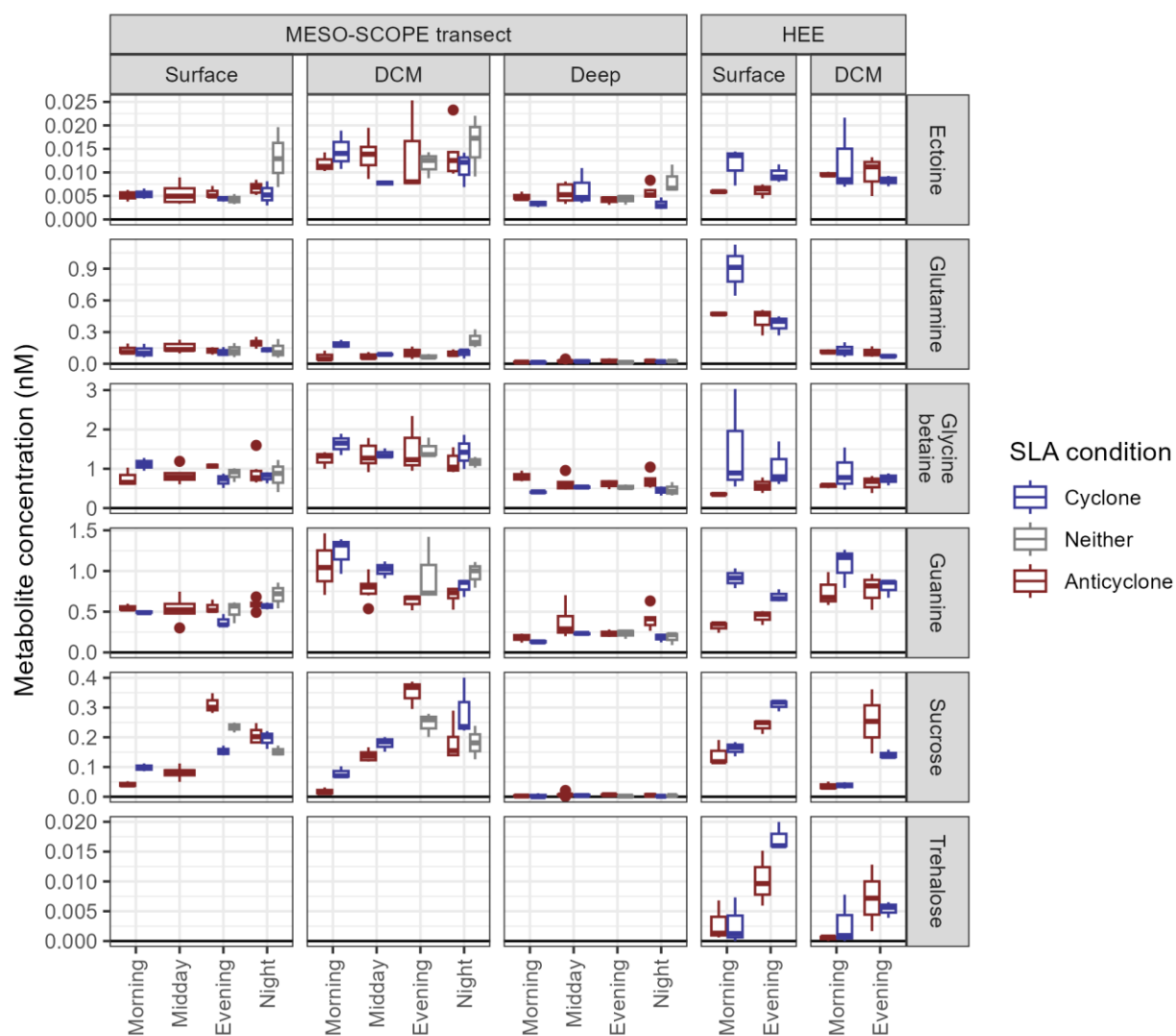
Supplemental Figure 5.8: Type I linear regressions of particulate carbon and nitrogen against total nM carbon and nitrogen in known metabolites both as a whole and divided by depth. Points have been colored according to depth (top row of plots) or SLA (bottom row of plots) and have been placed on independent axes. Two outlier particulate carbon points have been interpolated from beam attenuation for 15 meter values at stations 10 and 11.



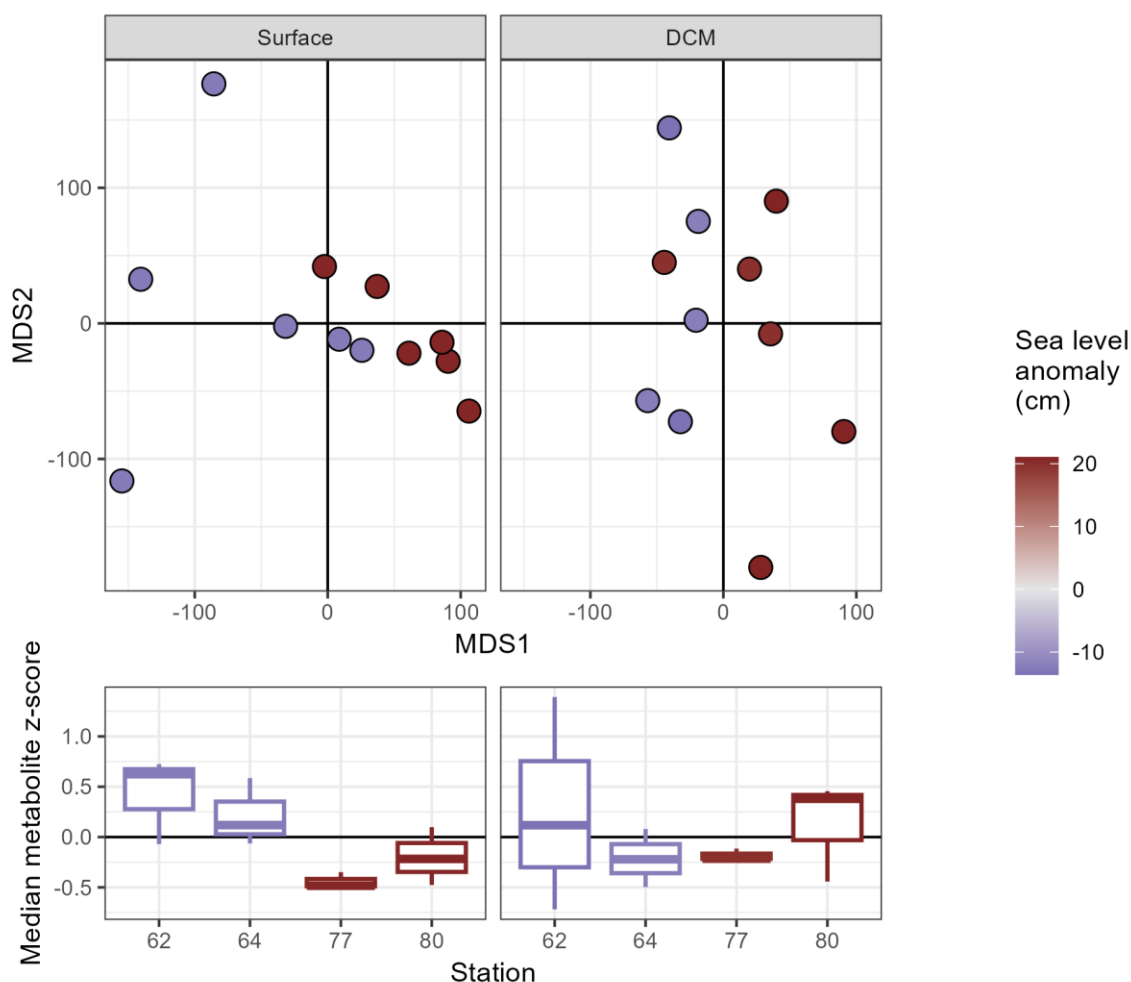
Supplemental Figure 5.9: Stacked barplot of top metabolites by triplicate. Stacked barplots of known metabolite concentration as shown in main text (Figure 3) but with individual triplicates shown. The same data is shown in A) as B) but B) is rendered in relative contribution space instead of absolute concentration. TMAO = trimethylamine N-oxide, HO-Ile = hydroxyisoleucine, GBT = glycine betaine, DCM = deep chlorophyll maximum.



Supplemental Figure 5.10: Evidence against phosphorus limitation across the MESO-SCOPE eddy transect. Top row of plots (A) correspond to extracted ion chromatograms for 5 metabolites useful for estimating the phosphorus stress of an environment, colored by the depth from which the sample was taken. Middle row of plots shows the integrated peak area for each metabolite across the 11 stations of the transect with sub-panels for each depth independently, with each point colored by the corrected sea level anomaly (SLA). Bottom row of plots shows the ratio of choline to phosphocholine (left) and choline to glycerophosphocholine (right) as a function of corrected SLA with no statistically significant trend visible at any of the three depths.



Supplemental Figure 5.11: Boxplots showing the concentrations of six selected compounds between cruises and depths as well as time and SLA classes. The abscissa is separated by the time category into which sampling was binned, referring to 6-hour increments starting from 3AM (e.g. Midday contains casts collected between 9AM and 3PM). Colors denote the sea level anomaly bins with a $\pm 5\text{cm}$ cutoff. For the eddy transect, six stations are included in the anticyclone, three in the cyclone, and two were neither. There were two morning stations, 4 midday stations, 2 evening stations, and 3 night stations. Diel effects are clearly visible for sucrose and trehalose while other compounds show very little diel differences. Trehalose was not quantified during the MESO-SCOPE transect for lack of an authentic standard to quantify the response factor.



Supplemental Figure 5.12: Plots of the metabolome during the Hawaiian Eddy Experiment, broken down by depth. The top row of plots are non-metric multidimensional scaling (NMDS) plots, in which points correspond to individual samples and have been colored by their corrected sea level height anomaly. SLA trends are visible in the 25 meter samples, with dark blue circles consistently discriminating from the dark red circles. The bottom row of plots show the direction and magnitude of this effect by plotting the grand mean of the normalized metabolite peak areas in the stations taken between the adjacent eddies of opposite polarity.

5.10 Supplemental Tables

Available as an Extended Data file. Captions reproduced here for clarity.

Supplemental Table 5.1: CTD cast list

Supplemental Table 5.2: CTD cast data by bottle

Supplemental Table 5.3: Interpolated CTD data

Supplemental Table 5.4: Internal standards

Supplemental Table 5.5: Authentic standards

Chapter 6: Nitrogen's Form Determines its Fate in the North Pacific Subtropical Gyre

6.1 Abstract

Nitrogen (N) availability limits the growth of many surface ocean microbial communities, particularly in the ocean gyres. Yet, even before growth is limited, some metabolic pathways or specific taxa may experience lower than optimal N flux due to cellular allocation and competition for various forms of N. Here, we present a stable isotope probing metabolomics study that illustrates the metabolic pathways taken by four different forms of nitrogen: ammonia, nitrate, arginine, and guanosine monophosphate. In the morning and evening, we followed isotope labeled substrates through metabolic pathways for either 26 hours (inorganic N additions) or 73 hours (organic N additions) in both surface (25m) and deep (175m) microbial communities in the North Pacific Subtropical Gyre. We found that nearly the entire surface glutamine pool (1-2 nM) is labeled by ammonia in less than an hour and goes on to label the metabolome more broadly after 24 hours, though abundant osmolytes like betaines (e.g. glycine betaine, 0.4-0.5 nM) had slow turnover rates (<5% in 24 hours). Nitrate amendments resulted in the labeling of only a small selection of molecules such as ultraviolet light absorbing compounds (mycosporine-like amino acids, or MAAs), whose incorporation patterns suggest that this form of nitrogen was available to a subset of the community, likely eukaryotes. The two organic nitrogen amendments labeled the metabolome more slowly than the inorganic forms of nitrogen and showed similar usage at 25 meters and 175 meters, hinting at the major role that heterotrophic bacteria play in cycling organic matter. The organic nitrogen added was mostly converted into glutamate after 24 hours rather than used as intact substrates, though labeled arginine was incorporated into protein directly and we were able to trace the carbon backbone

from both arginine and GMP into several additional metabolites. We also show that nitrogen assimilation and use was largely the same between morning and evening with the exception of MAAs. These measurements reveal the forms and fluxes of nitrogen in the natural environment and provide constraints on biogeochemical models that require characterizing the transformations between organic and inorganic nitrogen.

6.2 Introduction

In most of the surface ocean primary productivity is limited by the availability of nitrogen, an element with a complex and incompletely characterized biogeochemical cycle (Deborah A. Bronk and Steinberg 2008; Moore et al. 2013; Hutchins and Capone 2022). Organic nitrogen in particular is often treated as a black box despite substantial variability in its bioavailability and chemical nature. Turnover times for labile organic compounds like urea and amino acids can be orders of magnitude shorter than semi-labile or refractory DOM (Benner and Amon 2015; Deborah A. Bronk et al. 2024). Additionally, the availability of the precious nitrogen atom is determined not only by the form of the organic material it composes but also the environment in which it's found.

While extensive work has been done to characterize the forms and fluxes of nitrogen within the inorganic pool, the equivalent for the organic pool is woefully underdeveloped (Zehr and Kudela 2011; Deborah A. Bronk and Steinberg 2008). This is largely due to the difficulty of comprehensively measuring the many organic molecules that contain nitrogen in the marine environment (Boysen et al. 2018; Moran, Kujawinski, et al. 2022). Recently, however, we have improved our understanding of community composition via genomics and transcriptomics and made progress in our ability to characterize organic matter via proteomics and metabolomics. These advancements, paired with the increased availability of pure and isotope-labeled

substrates, mean that we can now apply metabolically resolved isotope tracing to resolve how organic nitrogen is transformed in the marketplace of metabolites that powers the surface ocean (Heal et al. 2021; McParland et al. 2021; Moran, Kujawinski, et al. 2022).

Nitrogen assimilation is a well-documented process in the ocean after decades of research and demonstrates how our understanding of non-marine organisms can be applied to the ocean (Mulholland and Lomas 2008). Central to the process is ammonia, which can be directly taken up from the environment by all organisms and combined with glutamate. This forms glutamine, which is then combined with alpha-ketoglutarate to form two glutamate molecules that can then be used as a nitrogen source throughout the cell (Walker and Van Der Donk 2016). Nitrate, on the other hand, must first be reduced to ammonia in a metabolically expensive process that some organisms have stripped from their genomes (Partensky and Garcezarek 2010; Zehr and Kudela 2011).

Unlike the relatively well-resolved pathways for inorganic nitrogen, the use of organic nitrogen remains underexplored in the marine environment. Organic nitrogen is generally thought to be accessible to most organisms but the high levels of DON make it clear that this is highly substrate-specific (Antia et al. 1991; D. A. Bronk et al. 2007; Aluwihare and Meador 2008; Zehr and Kudela 2011; Deborah A. Bronk et al. 2024). Additionally, many of the transformations possible are only active in specific communities or times of day (Bender et al. 2012). Linking the transformations that nitrogen undergoes to known molecular pathways from other environments or axenic cultures is a particularly promising way to understand what we observe on the global scale.

The lack of a comprehensive understanding of nitrogen's fate after it is assimilated makes molecular modeling difficult and undermines our ability to connect the community to the organic

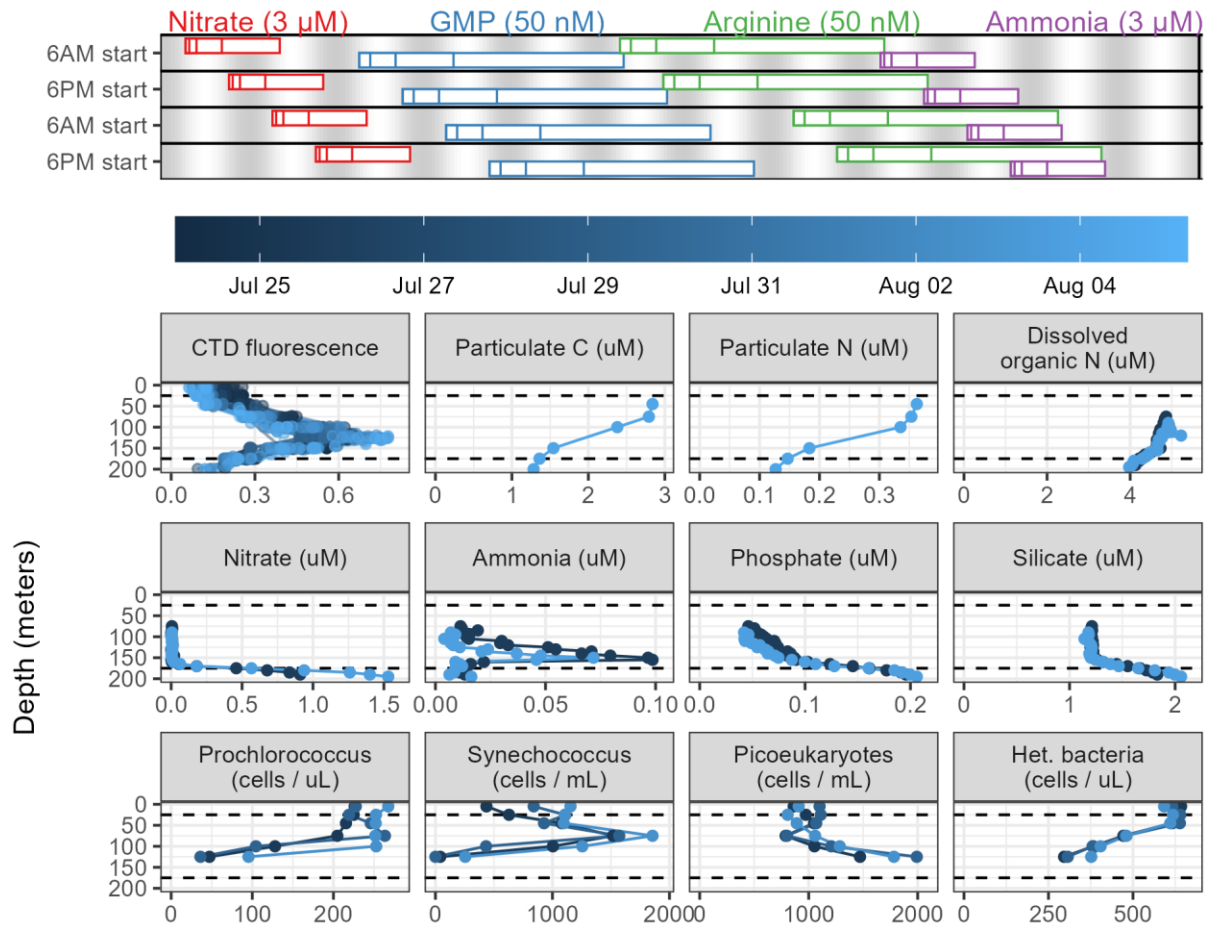
nitrogen it produces (Jones et al. 2024; Gonsior et al. 2022). Here, we trace the uptake and use of labeled ammonia, nitrate, arginine, and guanosine monophosphate in a natural community from the North Pacific Subtropical Gyre at multiple depths and diel conditions. Our application of metabolomics to these samples allows us to map the pathways and restructuring that organic nitrogen experiences in the largest biomes on the planet, quantifying the pool sizes and turnover rates for various low molecular weight compounds that serve as building blocks and intermediates of cell biology.

6.3 Results

Samples were collected from 25 meters and 175 meters during the 2021 SCOPE PARAGON cruise (KM2112, July 23rd - August 5th) near Station ALOHA that targeted a bloom within an anticyclonic eddy. The bloom was initially dominated by the diatom genus *Hemiaulus* with a large surface chlorophyll signal that decayed into a more typical North Pacific Subtropical Gyre community composed mostly of *Prochlorococcus* and SAR11. Community composition at our sampling depth below the deep chlorophyll maximum (DCM) was fairly typical of ALOHA, with a prokaryotic community dominated by SAR11, SAR282, and SAR 406, low levels of picoeukaryotes, and a microeukaryotic community equally distributed between diatoms, haptophytes, and dinoflagellates (Figure 1).

Samples collected from 25 meters were within the mixed layer and had PAR levels of 20-35% of the surface light while the 175 meter samples had a PAR level between 0.05 and 0.1%. Nitrate, nitrite, and ammonia data indicated that inorganic N was low in the surface ($<0.01 \mu\text{M}$), while the deeper samples at 175 meters were collected in the middle of the nutricline, with a clear DCM at 125 meters, an ammonia maximum at 150 meters, and a nitrite max at 160-165 meters. Environmental nitrate concentrations were about $0.5 \mu\text{M}$ at 175 meters (Figure 6.1).

We performed the incubations in succession, with the nitrate incubations first, followed by arginine, then guanosine monophosphate and finally ammonia. 6AM and 6PM data was always collected the same day and we alternated days between the surface and deep sampling depths to allow for reasonable sample processing times (Figure 6.1).



*Figure 6.1: Biogeochemical and community composition data collected during the PARAGON cruise alongside the incubation sampling plan. The top plot shows the start and end times of each incubation as rectangles aligned with the dates of the cruise in the middle color bar, with each rectangle representing a biological triplicate sacrificed at the given time. Grey shading in the background denotes light levels with grey corresponding to nighttime. The four amendments are shown as different colors. The lower panels show various biogeochemical and flow cytometry measurements colored by the date they were collected during the cruise. Note that *Prochlorococcus* and heterotrophic bacteria (Het. bacteria) are reported in cells per microliter while *Synechococcus* and picoeukaryotes are reported per milliliter. The 25 meter and 175 meter sampling depths have been added as dashed lines. C = carbon, N = nitrogen, GMP = guanosine monophosphate.*

6.3.1 Differences in nitrogen labeling at the surface between amendments

We observed significant differences in ^{15}N labeling patterns resulting from incubation with different forms of nitrogen (Figure 6.2). Ammonia incubations resulted in very rapid labeling, with glutamine more than 75% labeled even at the T0 timepoint, corresponding to the amount of time required to spike and immediately filter the 2 liters of seawater (~15 minutes). The dominant labeling pattern at T0 was a single ^{15}N , though all later timepoints showed more doubly- ^{15}N labeled glutamine than singly-labeled. After 24 hours, glutamine was more than 95% singly and doubly labeled and had approximately doubled in pool size (initially ~0.3 nM becoming 0.8 nM).

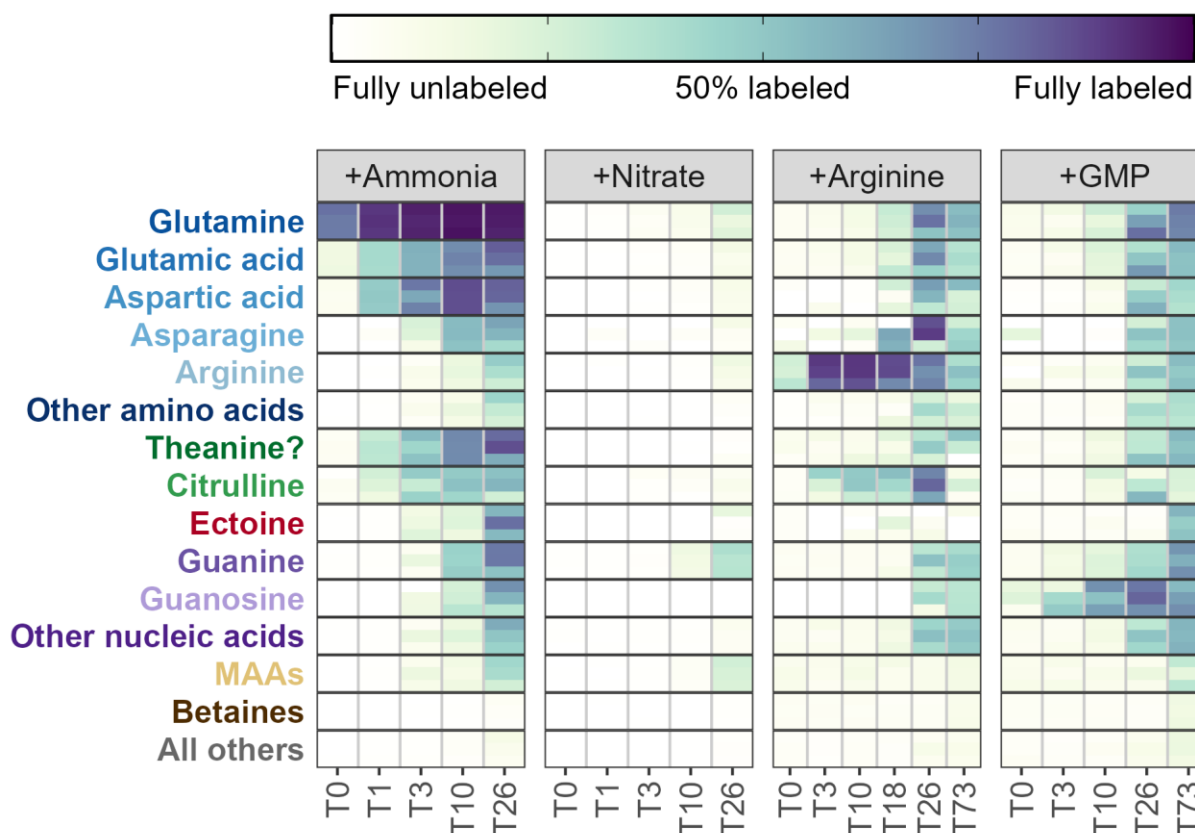


Figure 6.2: Fraction of each metabolite or metabolite group containing at least one ^{15}N label during the surface seawater amendments with various substrates after the number of hours shown on the x-axis. Darker colors indicate a larger percentage of the compound signal was isotopically labeled. Three biological replicates are stacked vertically at each timepoint and the

categories with multiple entries have been averaged. MAA = mycosporine-like amino acids, GMP = guanosine monophosphate.

The ^{15}N was then quickly propagated to glutamate and aspartate, creating a metabolite pool 40-50% labeled for each of these after a single hour. Aspartate was labeled slightly faster than glutamate, likely as a result of its smaller pool size (0.2 nM Asp versus 0.5 nM Glu). Also rapidly labeled was citrulline, reaching 20% labeled after a single hour, though arginine was not labeled until much later despite its location immediately downstream of this metabolite. Untargeted detection of other molecular features revealed an additional metabolite that was 30% labeled at the one hour timepoint with an m/z value of 175.1077 and a retention time of 8.2 minutes which we have putatively identified as theanine (N5-ethyl glutamine) based on m/z match and labeling pattern (Figure 6.2).

Other amino acids and nucleobases/nucleosides had trace amounts of labeling after 3 hours and were extensively labeled (>25%) after 24 hours (Figure 6.2). Cytidine (42.2% labeled) and adenosine (59.9% labeled) were more heavily ^{15}N labeled than their nucleobase derivatives at T26 (22.9% for cytidine and 42.1% for adenosine), corresponding to the expected synthesis pathway. Guanine, however, had the inverse trend with guanosine being 44.9% labeled while guanine was only 30% labeled. The deoxyribose forms were in all cases less completely labeled than their ribose counterparts (deoxycytidine = 18.9%, deoxyadenosine = 20.3%, and deoxyguanosine = 17.2%). In each case, the fully ^{15}N labeled nucleobase was the dominant constituent of the isotope envelope ($^{15}\text{N}_3$ for cytosine and $^{15}\text{N}_5$ for adenine and guanine). Other compounds, including betaines (here restricted to quaternary amines with a carboxylic acid group) and choline derivatives showed very little labeling at all (<10%).

Amendments with $^{15}\text{NO}_3$ showed a significantly different labeling pattern with very few metabolites incorporating the ^{15}N label at the surface even after 24 hours. Only trace amounts of

glutamine were labeled by the added nitrate even after 24 hours. The most heavily labeled compound was guanine, which saw trace amounts of label at the T10 timepoint and was ~40% labeled at T26. Adenine and adenosine were both ~15% labeled, while deoxyadenosine, deoxyguanosine, deoxycytidine, and guanosine were all 5% labeled or less. The only other group of compounds that did see compelling degrees of label incorporation from nitrate were the mycosporine-like amino acids (MAAs), a group of compounds whose primary function is thought to be protection from UV light (see below). These compounds, like guanine, were labeled in excess of the bulk glutamate/amino acid pool from which their nitrogen is sourced.

The organic nitrogen amendments also showed labeling patterns distinct from the ammonia incubations, though a large fraction of the nitrogen was clearly remineralized into ammonia/glutamate and used to label a similar suite of compounds (Figure 6.2). At least 50 picomoles of fully C and N labeled ($^{13}\text{C}_6$, $^{15}\text{N}_4$) arginine was taken up at the T0 timepoint (within 20 minutes, creating an arginine pool 30% labeled), and 80-90% of the pool was fully labeled after 12 hours with a maximum arginine concentration of 0.73 nM. Surprisingly, the fraction of arginine labeled decreased from this point, with the T26 and T73 timepoints dropping to 74% and 48% labeled, respectively, as the total amount of both labeled and unlabeled arginine dropped back to initial concentrations (~0.08 nM). This labeling rapidly propagated into citrulline, likely via the urea cycle, indicating that labeled urea ($^{13}\text{C}_1$, $^{15}\text{N}_2$) was probably created and available in the system, though we were unable to measure this directly. Glutamate and glutamine were labeled much more slowly, reaching 10% labeled after 10 hours and a maximum of 66% labeled after 24 hours. Much of this nitrogen was then cycled very quickly among other compounds which also saw an abrupt increase in fraction labeled at 24 hours. This was particularly true for other amino acids and the nucleobases/nucleosides and unlike the inorganic

N incubations guanine was not labeled more in excess of the other nucleobases. Mycosporine-like amino acids and betaines were unlabeled above background even after 3 days of incubation.

The guanosine monophosphate (GMP) incubations showed a pattern very similar to the arginine incubations in the identity, timing, and magnitude in which metabolites were labeled (Figure 6.2). While we were unable to measure GMP directly we saw clear labeling of fully C and N labeled guanosine ($^{13}\text{C}_{10}$, $^{15}\text{N}_5$) after 3 hours that increased to a maximum of 75% after 24 hours before decreasing slightly to 64% labeled at the end of the three-day incubation. A small part of this signal was reflected in guanine, with $^{13}\text{C}_5$, $^{15}\text{N}_5$ guanine representing about 10% of the signal at T3 and reaching a maximum of 40% at T26. At T26 we also saw extensive labeling in adenine, adenosine, and hypoxanthine though their dominant labeling patterns ($^{13}\text{C}_0$ $^{15}\text{N}_5$, $^{13}\text{C}_5$ $^{15}\text{N}_5$, and $^{13}\text{C}_0$ $^{15}\text{N}_4$ respectively, Supplemental Figure 6.11) suggested that the adenine and hypoxanthine were synthesized from glutamate and that the ribose pool had been swamped by guanosine cleavage. Here, unlike the arginine incubations, we also saw some MAA labeling at the end of the 3 day incubation.

We also explored the degree of labeling incorporation into proteins by measuring the total hydrolyzable amino acids (THAAs) at both the start of each incubation and the end (Figure 6.3). THAAs come from protein (which can be in living biomass or detritus) and include minor contributions from free amino acids as well as metabolites and peptides with hydrolyzable amino acid moieties (e.g. MAAs and glutathione). THAAs were 26% ($\pm 6.5\%$, SD, $n=24$) of the total surface particulate carbon pool ($\text{PC} = 2.84 \mu\text{M}$ in mixed layer) at T0 and 55% ($\pm 14.3\%$, SD, $n=24$) of the particulate nitrogen ($\text{PN} = 0.36 \mu\text{M}$ in mixed layer) at the start of the incubations. Free amino acids composed between 0.2% and 14% of the total hydrolyzable fractions with the highest contributions from free glutamate (Table 6.1).

Table 6.1: Concentrations of 13 amino acids detected in both the total hydrolyzable (THAA) and dissolved free (DFAA) pools for unlabeled data in the surface averaged across all treatments. Values are reported in nanomoles per liter (nM).

Amino acid	THAA (T0)	Free (T0)	% free (T0)	THAA (final)	Free (final)	% free (final)
Glutamate + glutamine	19.57	1.66	8.47%	29.41	4.04	13.72%
Histidine	2.53	0.04	1.59%	2.76	0.07	2.58%
Arginine	6.87	0.08	1.13%	8.15	0.20	2.51%
Aspartate + asparagine	14.54	0.33	2.28%	20.93	0.48	2.28%
Alanine	15.24	0.17	1.09%	20.79	0.46	2.22%
Proline	7.22	0.04	0.56%	8.47	0.16	1.86%
Serine	20.26	0.21	1.04%	21.25	0.27	1.27%
Glycine	29.57	0.17	0.59%	29.28	0.34	1.15%
Leucine + isoleucine	22.49	0.18	0.78%	26.95	0.30	1.12%
Threonine	12.06	0.03	0.22%	14.71	0.06	0.38%

There were also differences in THAA labeling patterns between the amendments (Figure 6.3). Despite their shorter incubation time (26 hours vs 73 hours), the inorganic nitrogen amendments had approximately equal fractions of labeled protein. The nitrate amendments in particular were surprisingly heavily labeled given the minimal signal seen in free metabolites.

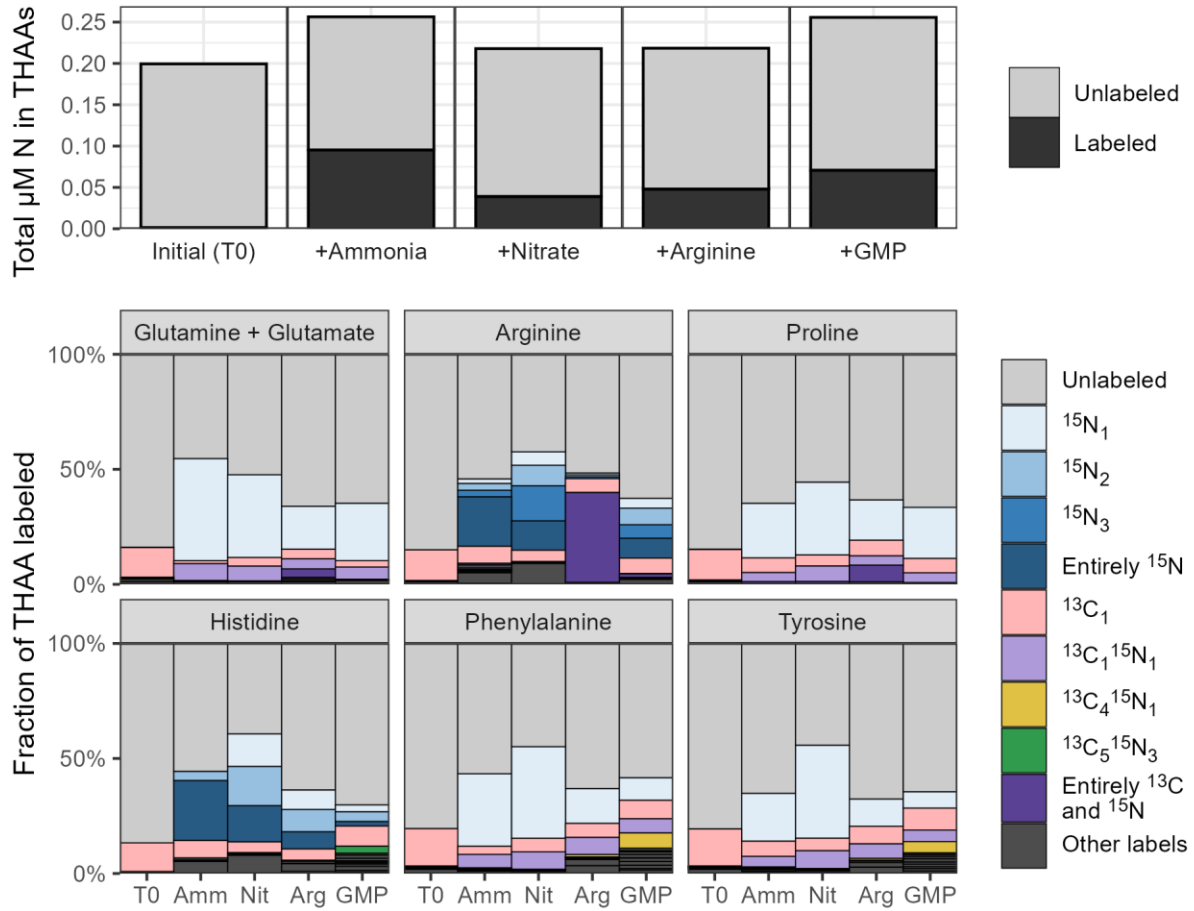


Figure 6.3: Stacked bar plots showing the fraction of various total hydrolyzable amino acids (THAAs) labeled at the end of the incubations (26 hours for nitrate and ammonia, 73 hours for guanosine monophosphate (GMP) and arginine) across the four amendments. Total μM nitrogen (N) in the protein pool is shown in the top plot while the fraction of individual amino acids labeled is shown in the rows below. Colors denote specific labeling patterns detected in the THAAs.

In the THAAs we were able to see the carbon backbone from the organics propagate into other compounds, resulting in ^{13}C labeling in excess of natural abundance. Proline and glutamate both had detectable $^{13}\text{C}_5^{15}\text{N}_1$ labels in the arginine treatments. We also saw histidine $^{13}\text{C}_5^{15}\text{N}_3$ synthesized during the guanosine monophosphate treatment, with the $^{15}\text{N}_3$ pattern indicating that glutamate, glutamine, and adenine were also heavily ^{15}N labeled. Finally, both phenylalanine and

tyrosine had significant $^{13}\text{C}_4$ $^{15}\text{N}_1$ labeling in the GMP amendments (Figure 6.3, Supplemental Figure 6.12).

However, this increase in THAA abundance was not reflected in the osmolytes (Figure 6.4). Instead, osmolytes showed small amounts of labeling as a whole and did not increase in concentration over the course of the experiment with the exception of proline and glycine, which were ~50% labeled after 24 hours in the ammonia, arginine, and GMP additions. Ectoine also saw extensive labeling at T26 in all amendments.

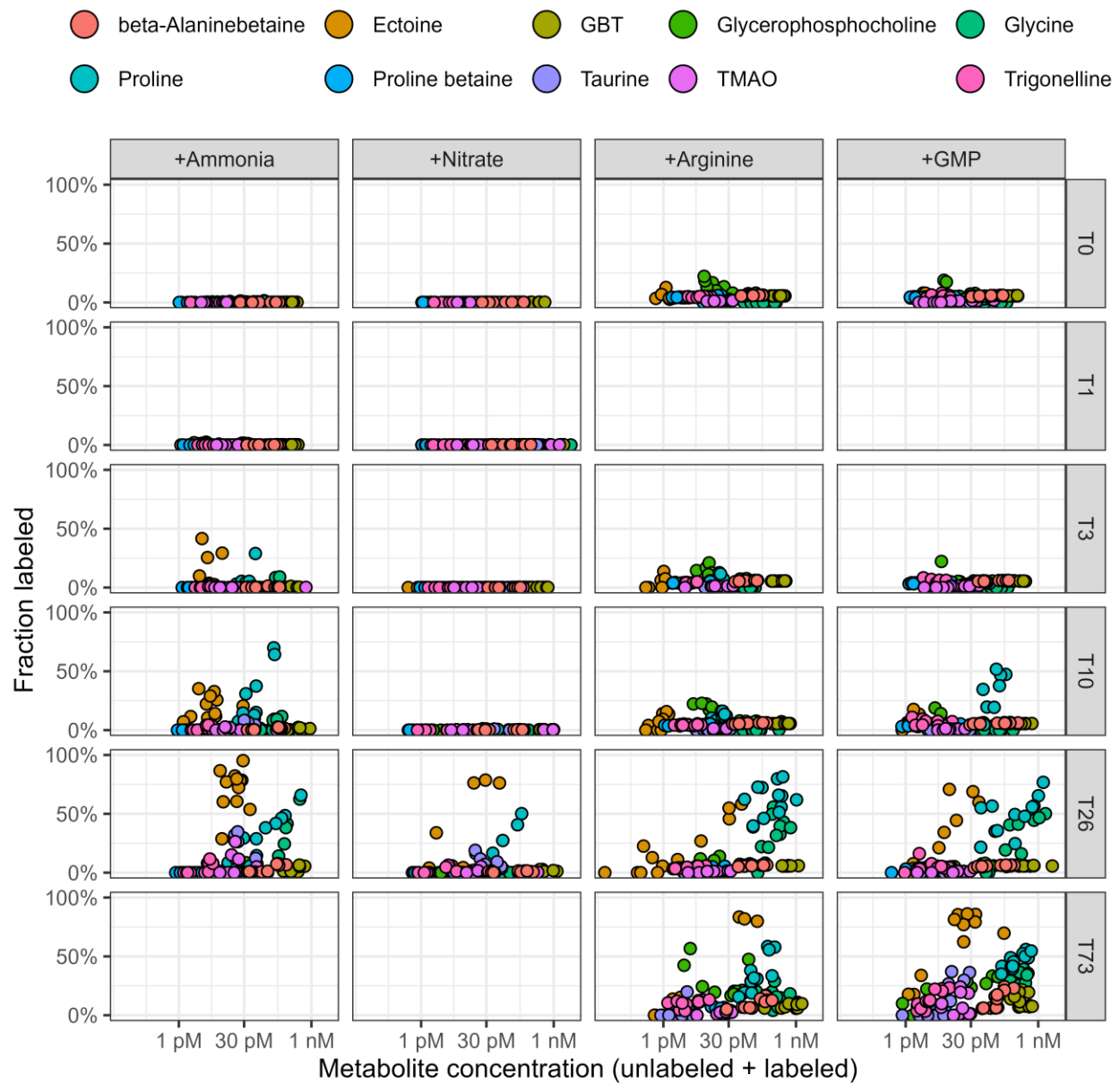


Figure 6.4: Concentration of common osmolytes plotted against the fraction labeled. Time increases in panels from top to bottom while the four different amendments are shown as columns. Colors denote individual metabolites and both the 25 meter, 175 meters, 6AM, and 6PM samples are shown together in each facet. GBT = glycine betaine, TMAO = trimethylamine N-oxide.

6.3.2 Minimal diel signal observed except in the UV-absorbing mycosporine-like amino acids

We performed experiments with each of the 4 forms of nitrogen at both 6AM and at 6PM to test for any diel effects in labeling patterns. Very few compounds showed a detectable difference

between the 6AM start time and the 6PM start time with the exception being the mycosporine-

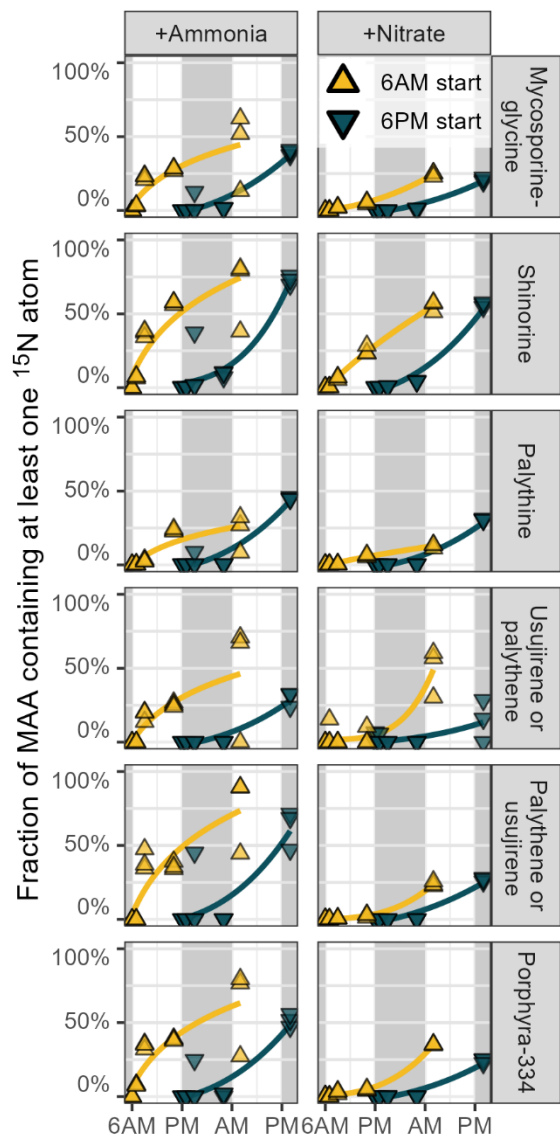


Figure 6.5: Fraction of UV-absorbing mycosporine-like amino acids (MAAs) containing at least one ¹⁵N label over one day during the treatments amended with ¹⁵N labeled ammonia or nitrate. Incubations started at 6AM are shown in yellow while incubations started at 6PM are shown in blue, with grey background regions highlighting nighttime. Three replicates are included at each of the 5 timepoints and best-fit power law curves have been fit behind the data.

Nitrate data showed similar degrees of labeling which was again surprising given the reduced amount of bulk metabolite labeling in those amendments, though the synthesis of MAAs from nitrate appeared less clearly diel. Shinorine and porphyr-334 were particularly well-

like amino acids (MAAs, Figure 6.5). Shinorine and palythine were confirmed via MS² fragments while the other annotations remain putative, though their presence exclusively in the surface is compelling. During the ammonia incubations, six MAAs showed clear ¹⁵N incorporation while exposed to sunlight, with the labeled fraction increasing sharply over the first 10 hours in the incubations started at 6AM. Incubations started at 6PM showed very little labeling from ammonia until a full 24 hours had passed. Shinorine, mycosporine-glycine, palythine, and porphyr-334 were all clearly labeled (>20% ¹⁵N) as well as two of three molecular features detected at an *m/z* of 285.1445 likely corresponding to palythene and usujirene.

labeled under nitrate, approaching the values achieved in the ammonia incubations. Very little MAA labeling was observed during the arginine and GMP amendments until the 73 hour timepoint, though at that point the fraction labeled was approximately equivalent to the inorganic amendments and slightly higher for GMP than arginine (Supplemental Figure 6.13).

6.3.3 Ammonia more bioavailable at surface, organics equally bioavailable at 25m and 175m

We also performed the four nitrogen amendments at both the surface (25 meters) and below the euphotic zone (175 meters, <0.1% PAR) to investigate the effect of depth on organic and inorganic nitrogen use. We found that ammonia was used much more readily at the surface as measured by the fraction of metabolite labeled, with ^{15}N incorporation for amino acids 5-25% greater in the surface than at depth (Figure 6.6). Glutamate, theanine, aspartate, and asparagine were among the most different with glutamate and theanine surface values about 10 times those at depth while ^{15}N aspartate was undetectable until hour 3 of the incubation and ^{15}N asparagine was never detected. In contrast, glutamine labeling was similar between the two depths, with the fraction labeled 10-20 percentage points higher at the surface across all timepoints; a trend which was clearly not propagated to glutamate or other amino acids and instead must have represented accumulation or protein incorporation.

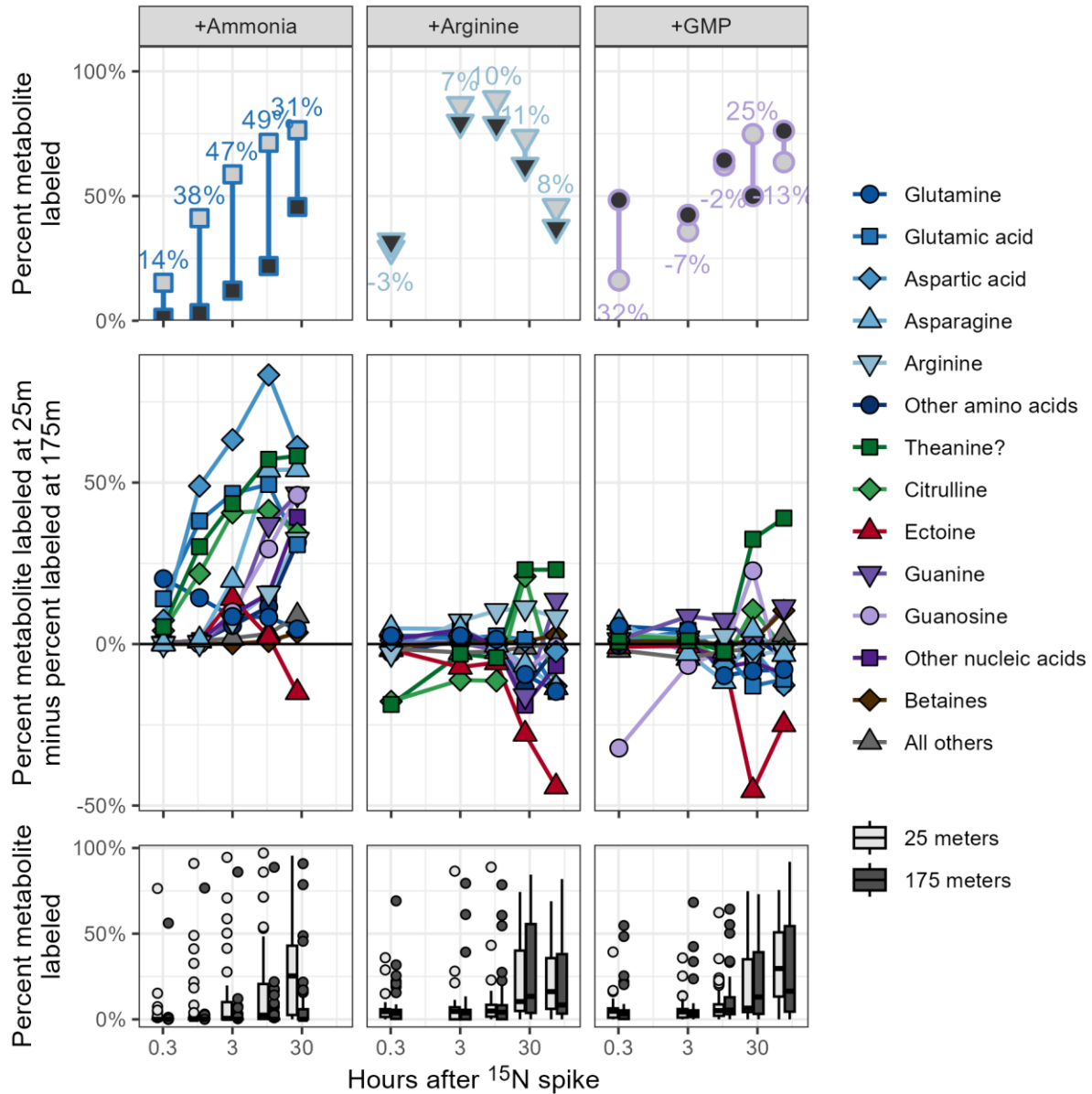


Figure 6.6: Plots of the metabolite fraction labeled at the surface (25 meters) and the fraction labeled at 175 meters. Incubations with added ammonia, arginine, and guanosine monophosphate (GMP) are shown. Top row of plots show the calculation of % labeled at 25 meters minus % labeled at 175 meters for three example compounds. All compounds are shown in the middle row of plots, grouped by chemical similarity and labeling pattern where positive values indicate increased synthesis in the surface samples relative to those at depth. Finally, the bottom row of plots shows the relative difference between the surface and deep as boxplots for all compounds without aggregating. Points in the first two rows are the average of six biological replicates (incubations started in the morning and the evening are combined here for additional power) while points in the bottom row denote metabolites with labeled fractions falling outside 1.5 times the interquartile range.

Other metabolites varied in their degree of labeling (Figure 6.6). Nucleobases and nucleosides were much more heavily labeled from ammonia at the surface, with maximum labeling 3-10 times more extensive around T3 and T10 if the isotope was measurable at both depths and infinitely more so if not. Betaines and other nitrogen-containing metabolites, however, were almost entirely unlabeled at 175 meters until the 24 hour mark, at which time the fraction labeled in both was 5-20 times higher at the surface. An important exception to this rule is the osmolyte ectoine, an aspartate derivative that nonetheless was synthesized with nitrogen from ammonia at depth with the fraction labeled equal to or exceeding that of the surface in an equal amount of time.

We were unfortunately unable to present the data collected from the 175 meter labeled nitrate incubations here due to our inability to confidently connect data to the correct timepoint for those samples but the raw numbers are provided in the supplement (Supplemental Figure 6.11, Supplemental Figure 6.15).

^{15}N use from the organic nitrogen substrates was similar between the 25 meter and 175 meter samples aside from the direct products added to incubations. In both cases, uptake and use of the labeled substrate was equal or slightly higher at depth for a majority of the metabolites measured. Exceptions to this were theanine, which was 20-30 percentage points more heavily labeled at the surface by the end of the incubations, and ectoine, which was synthesized more at depth even to a larger degree than it was during the ammonia treatments, reaching nearly 50 percentage points more labeled at 175 meters than it was at 25 meters.

At 175 meters, the difference between organic and inorganic amendments also becomes apparent (Figure 6.7). Organic compounds result in increases in absolute metabolite concentration and fraction labeled while the inorganic substrates are essentially unused. This is

in direct contrast to the surface data, where ammonia also resulted in an increased metabolite load likely corresponding to overall growth. Much of the labeling at 175 meters appears to be in the less-abundant compounds. Glutamine (~1 nM) was almost entirely labeled after 24 hours and half the glutamate pool (3-5 nM) had been synthesized from the ^{15}N in arginine and GMP, while the other 7 most abundant compounds were not nearly as heavily labeled. The less-abundant compounds, however, were also on average 50% labeled, representing an additional 1-2 nanomoles of ^{15}N label incorporation per liter (Supplemental Figure 6.14).

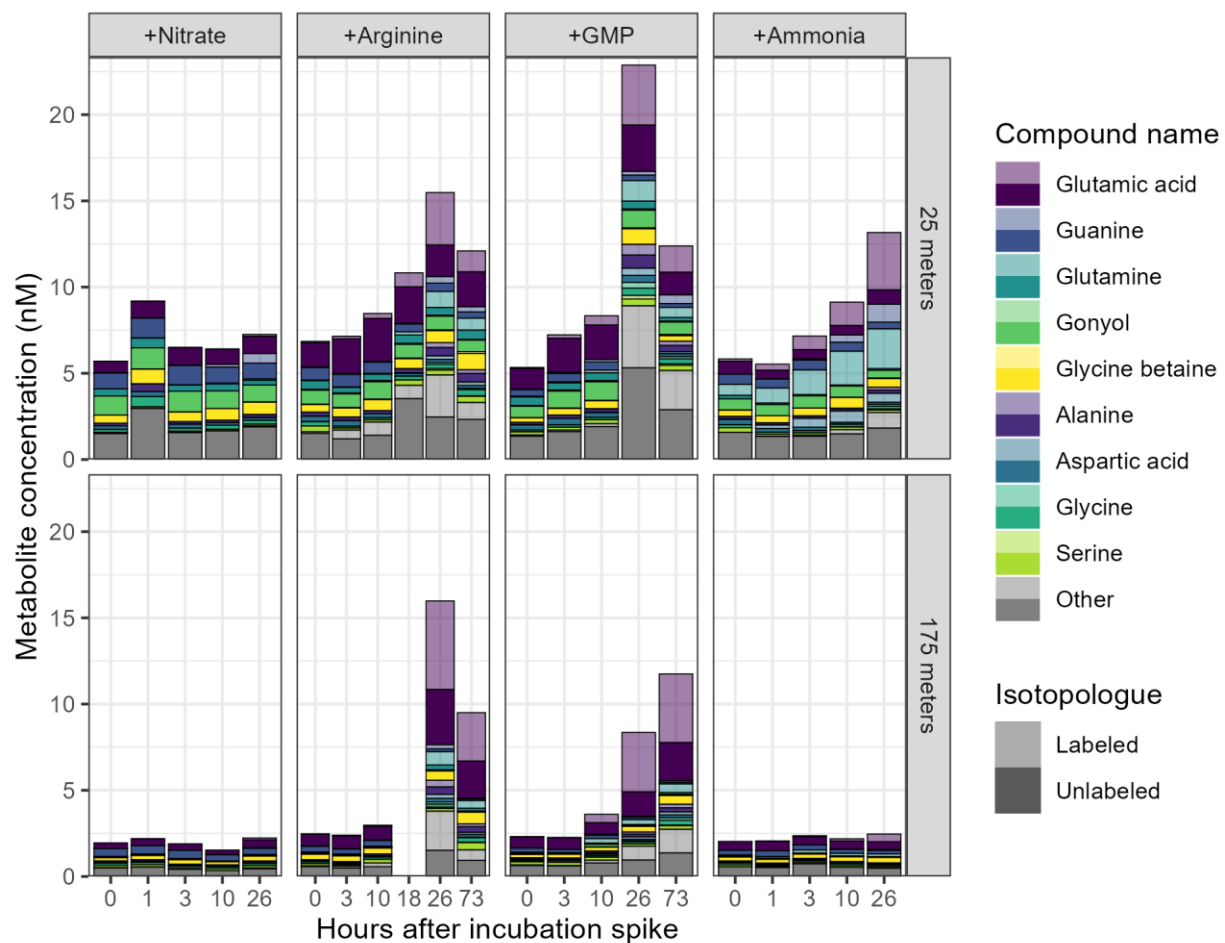


Figure 6.7: Quantified metabolite concentrations in nanomolar throughout the incubations for the 9 most abundant metabolites as distinguished by color. Concentrations have been broken down by whether they are fully unlabeled or contain any amount of ^{15}N (or ^{13}C in the case of the arginine and GMP amendments). Compounds other than the top 9 have been averaged and are shown in grey. Totals for all four amendments are shown here in the order they were done during the cruise, with separate panels for the 25 meter samples and the 175 meter samples.

Biological triplicates from both 6AM and 6PM have been averaged here to produce the single value with an n of 6 except for the 25 meter arginine incubations where the morning T3 timepoint was filtered after 18 hours instead for an n of 3 at T3 and T18.

We explored this further by investigating the total amount of nitrogen contained within the metabolites and THAAs in the experiments. We saw that the absolute amount of ^{14}N remained relatively constant around 10 nM at the surface and 5 nM at 175 meters, while the absolute amount of ^{15}N increased smoothly. The same pattern was observed in the THAAs, where ^{14}N concentrations hovered around 150-200 nM at the surface and ~100 nM at 175 meters while the ^{15}N concentrations jumped up between the initial and final timepoints for arginine and GMP after the three day incubation while the increase was only observed for ammonia at the surface.

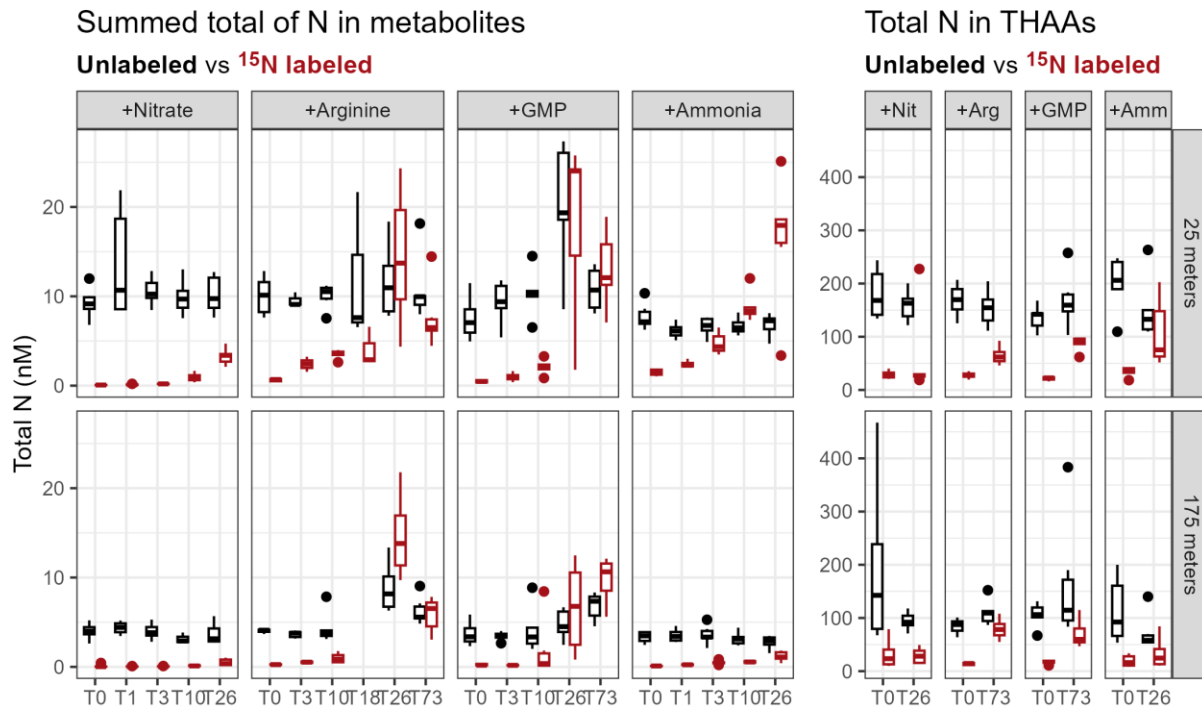


Figure 6.8: Quantified metabolite concentrations in nanomolar throughout the incubations for both metabolites and THAAs. Colors denote whether they are fully unlabeled or contain any amount of ^{15}N (or ^{13}C in the case of the arginine and GMP amendments). Totals for all four amendments are shown here in the order they were done during the cruise.

Finally, to investigate what fraction of this labeling was due to net synthesis or metabolite turnover for the metabolome as a whole, we plotted the relative concentration change between the initial timepoint and T26 against the fraction of the metabolite labeled (Figure 6.9). While most metabolites increased their labeling in sync with their total increase in abundance, two metabolites stood out in particular. The amino acid methionine experienced large amounts of synthesis and jumped from below the limit of detection at T0 (a few pM) to hundreds of picomolar. This did not occur in the 175 meter ammonia or nitrate incubations, and the 25 meter nitrate incubation's methionine concentration actually decreased between T0 and T26. Unlike methionine and most other compounds, adenosine also stood out for its high levels of turnover without much increase in abundance (Figure 6.9). Adenosine concentrations consistently decreased after 24 hours despite accumulating large amounts of label (40-80%), indicating high turnover and use of this metabolite even in the GMP incubation where a similar nucleoside was abundant in the environment. This pattern was especially unique because the nucleobase adenine did not follow this pattern and instead mapped onto most other metabolites with simultaneous synthesis and labeling, as did the other nucleosides guanosine and cytidine.

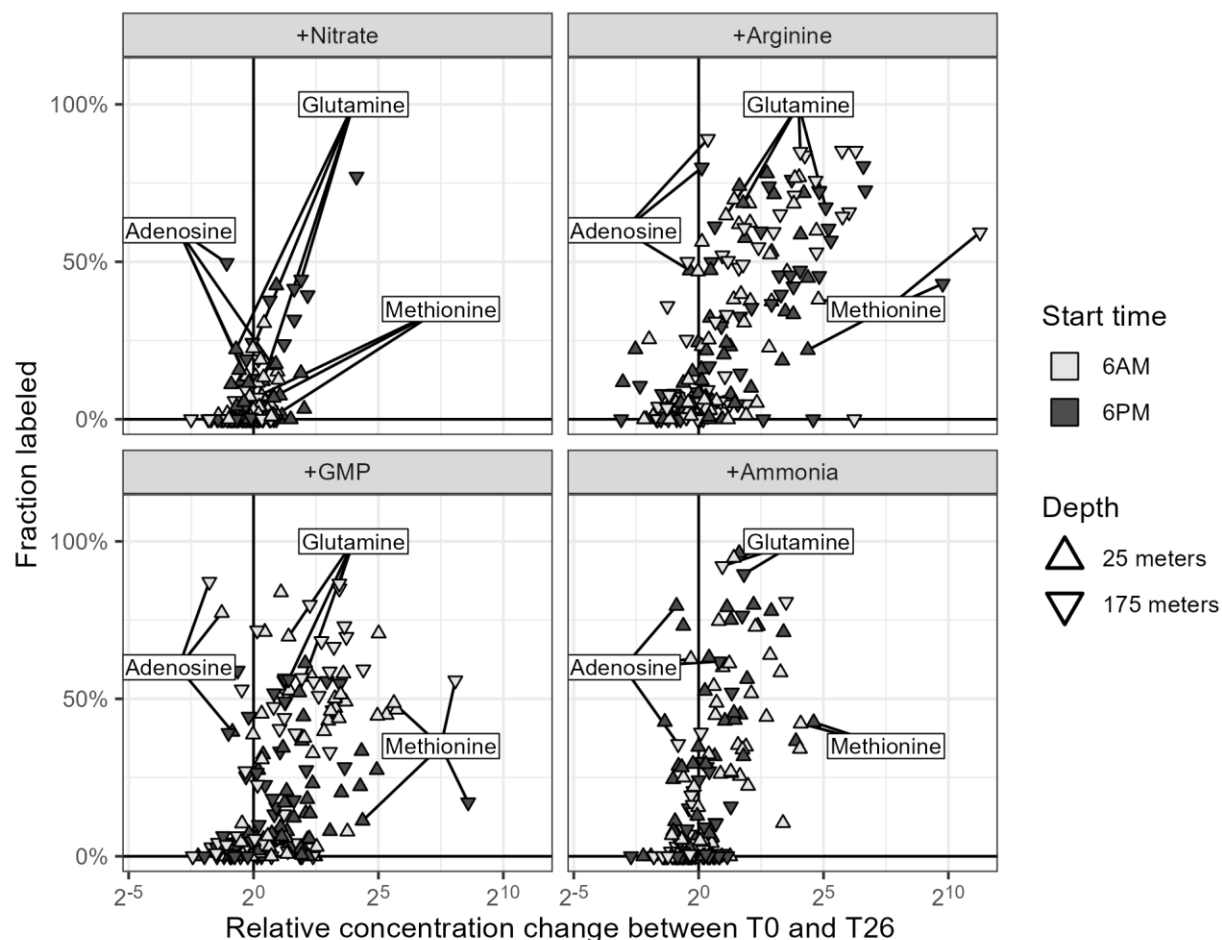


Figure 6.9: Turnover vs new synthesis of metabolites after ~24 hours. Points correspond to individual metabolites with the x-coordinate indicating the log₂-fold change difference between the compound's initial concentration and its final amount and the y-coordinate indicating the fraction of the metabolite labeled at T26. Points near zero on the x-axis but with large y-axis values underwent significant turnover without much synthesis, while values sloping up to the right are the result of net growth. Grey fills denote the start time (either 6AM or 6PM) and the shape denotes the depth from which the incubation community was taken. The three consistently outermost metabolites have been manually labeled, though zero methionine signal was detected at T0 for the 175 meter samples in the 6AM nitrate and ammonia incubations as well as the 6PM ammonia incubation, so no points can be annotated.

6.3.4 Untargeted methods reveal widespread labeling throughout the metabolome

After a heuristic compound discovery approach revealed that the putatively identified

compounds theanine, deoxyguanosine, deoxycytidine, and MAAs had interesting labeling

patterns, we decided to perform a fully untargeted analysis to detect additional compounds that

were labeled over the course of the incubations. Fairly lax peakpicking settings for xcms

returned nearly 10,000 chromatographic peaks in each of the ~280 samples across the four LC runs used, one for each amendment sample set. At this point we noted that the nitrate amendment samples had an unusually large number of peaks detected due to the inclusion of a contaminant suite, likely plastic, which fortunately eluted before many of our compounds of interest.

Correspondence requiring the presence of a peak in all three biological triplicates for a given treatment/timepoint combination reduced that number to approximately 5,000 unique features. Random inspection revealed that the majority of these features were noise so we calculated additional peak quality metrics as well as reintegrated each peak and its predicted isotope (see Methods). Of the initial 5,602 features, over half (2,956) failed quality control and an additional 1,690 either eluted within the first two minutes or were likely adducts of bisulfate (eluting between 14 and 14.5 minutes) and were removed from downstream analysis as well.

Of the 219 features that remained, the majority (150) had a ^{15}N isotope with the correct m/z value and a chromatographic similarity score between the unlabeled and labeled data above 0.8 (see Methods). 80 of these 150 also showed linear increases in peak area across the timepoints, denoting likely ^{15}N -labeled metabolites. We manually inspected each of these 80 metabolites and were able to identify the majority as either compounds in our list of authentic standards (19/80), an isotope of a known compound (39/80), or an in-source fragment of a known compound (12/80, consisting of fragments from proline (70.0658), glutamic acid/glutamate (84.0449, 102.0550, and 130.0499), and the loss of ribose from nucleosides). Five compounds were putatively identified using isotope labeling patterns and literature searches, four additional compounds could not be putatively identified but had plausible molecular formulae, and one feature we were unable to annotate at all (Supplemental Table 3).

To derive biological meaning from the untargeted analysis, we performed a k-means clustering of the features across all the treatments and timepoints and included our targeted data in the clustering. Clear patterns of isotope incorporation emerged in 6 of the 10 clusters we used (Figure 6.10). Cluster 1, the largest with 64 molecular features, contained many of the unlabeled compounds including those without any nitrogen and all of the betaines. Cluster 2 contained most of the amino acids and beta-alanine. Cluster 3 contained the nucleobases other than cytosine and the ribose nucleosides as well as asparagine and proline, while cluster 5 contained the deoxyribose nucleosides and cytosine which did not get labeled as rapidly in the ammonia treatment. Cluster 4 contained 27 rapidly-labeled compounds including glutamine, glutamic acid, and aspartic acid. The sole known compound in cluster 6 was guanosine, with heavy labeling in the GMP incubations and mild labeling elsewhere. Cluster 7 consisted exclusively of the MAAs and their isotopes. Finally, cluster 8 contained arginine, citrulline, ectoine, theanine, and N-acetyllysine, all of which had significant labeling in the arginine amendment. None of the 7 compounds in clusters 9 or 10 had known metabolites and these appear to be largely noise signals with 50% or more labeled at T0 (Figure 6.10, Supplemental Figure 6.15).

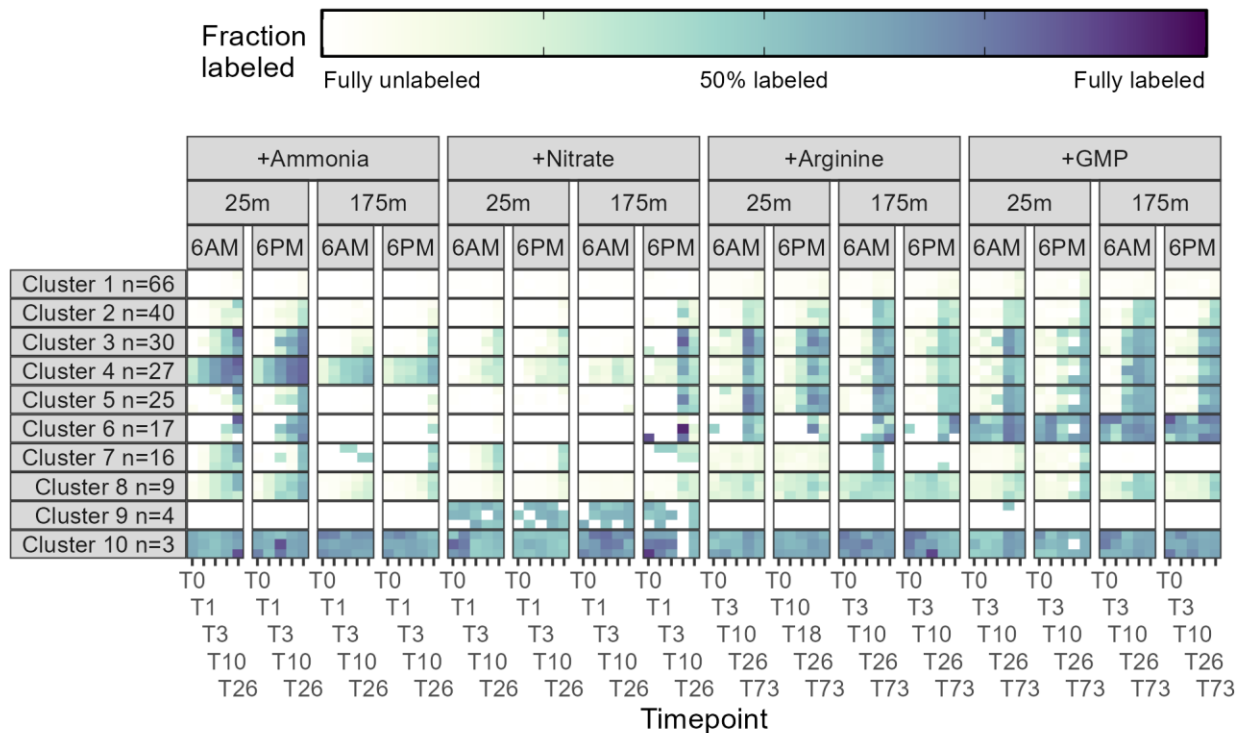


Figure 6.10: Fraction of each cluster containing at least one ^{15}N label during the surface seawater amendments with various substrates after the number of hours shown on the x-axis. Darker colors indicate a larger percentage of the feature signal was isotopically labeled. Three biological replicates are stacked vertically at each timepoint and the labeling within each clusters has been averaged across all the composing features. The number of features in each cluster is specified by the n in the facet name. GMP = guanosine monophosphate.

6.4 Discussion

6.4.1 Contrasting the metabolic pathways taken by four forms of nitrogen

Ammonia was taken up immediately and resulted in extensive labeling of glutamine even within the first 20 minutes of the incubation at both 25 and 175 meters. This indicates that ammonia transporters and glutamine synthetase were present, despite low in-situ concentrations. This ^{15}N signal rapidly propagated to the expected amino acids via well-established pathways, though much more slowly at depth. One surprise was the extensive labeling of citrulline, a central urea cycle metabolite, without the associated propagation to arginine and ornithine. This indicates that the ammonia was likely being processed into carbamoyl phosphate via the CPS (carbamoyl

phosphate synthetase) enzyme instead of ornithine via the Arg* synthesis pathway, potentially as a detoxification mechanism to prevent the accumulation of ammonia to dangerous levels (Armbrust et al. 2004; Armbrust 2009). Also surprising to us was the rate at which nucleobases were labeled, most commonly with all nitrogens as ^{15}N . This indicates that either nucleobase synthesis was already highly active or was triggered by the addition of ammonia. Notably, the deoxyribose forms of the nucleosides were labeled more slowly than the ribose forms, hinting at a slower turnover of DNA components than RNA.

In contrast, nitrate was available to only a subset of the community and did not result in widespread labeling either at the surface or at depth. Not all organisms are able to use nitrate, most notably absent in a majority of clades in the abundant cyanobacteria *Prochlorococcus* (Berube et al. 2015; Partensky and Garczarek 2010; Martiny et al. 2009). Typically, nitrate amendments result in the growth of diatoms, with smaller contributions from haptophytes or pelagophytes (Alexander et al. 2015; Rii et al. 2018; Seelen et al. 2024). Diatoms are capable of nitrate reduction and we suspect that they are the primary source of the ^{15}N labeling seen in the nitrate incubations. We see clear patterns of a community subset in the labeling patterns of guanine, MAAs, and the THAAs, all of which had a fraction labeled greater than the bulk glutamine/glutamate pool and is thus not possible with a homogeneously mixed pool. The extensive labeling of the THAAs in particular relative to the metabolite pool indicates that a majority of the protein synthesis was done by a subset of the community able to assimilate nitrate. Additionally, the MAAs are synthesized by adding amino acids to the 4-deoxygadusol core (Wada et al. 2015), requiring that the glycine, serine, and threonine pools also be at least as heavily labeled as the MAAs they compose.

Importantly, the amount of nitrate added is massively in excess of the quantity normally available at the surface which likely resulted in a lag as nitrate-sensitive genes were turned on so it is possible that a longer incubation time would have resulted in additional labeling if translation time was a major limiting factor. However, we see similar results in Zimmerman et al. (2023) who also observed significantly more labeling from ammonia than nitrate. The clear upward vertical propagation of increasingly reduced forms of nitrogen (ammonia max < nitrite max < nitrate max) also hints at the role the more oxidized forms may play as electron sinks in the surface ocean where light is abundant.

The two forms of organic nitrogen were used almost identically at both the surface and below the DCM as well as at 6AM and 6PM. In all eight circumstances, the first 10 hours showed essentially no labeling aside from the amended molecule and its direct products (e.g. citrulline & guanosine), then the fraction labeled spiked up at T26. This similarity leads us to believe that the heterotrophic community was responsible for most of the labeling that occurred in these incubations, in agreement with the general philosophy that heterotrophic bacteria are the primary sinks of DON (Berthelot et al. 2021; Bradley et al. 2010) despite clear evidence that autotrophs are capable of taking up and using organic substrates (Muñoz-Marín et al. 2020; Antia et al. 1991; Deborah A. Bronk et al. 2024). The especially heavy labeling (>75% across the board) of the nucleobases and nucleosides from organic substrates indicate that their synthesis was a priority for the community. This, combined with the large amount of THAA synthesis indicates net synthesis of biomass, as seen in the majority of the metabolite concentrations also increasing over the experiment duration.

Organic nitrogen additions may have relieved N limitation at the surface and C limitation at depth, allowing similar degrees of growth in both locations. This is supported by the reuse of

the carbon backbone in the THAAs, with proline and glutamate both showing $^{13}\text{C}_5$ $^{15}\text{N}_1$ labels in the arginine treatments and histidine showing $^{13}\text{C}_5$ $^{15}\text{N}_3$ synthesized during the guanosine monophosphate treatment, confirming that the ribose pool must have been swamped with fully ^{13}C labeled ribose cleaved from guanosine in order to enter the phosphoribosyl pyrophosphate pathway. The $^{15}\text{N}_3$ pattern also indicates that for these communities, glutamate, glutamine, and adenine were also >50% labeled. Finally, both phenylalanine and tyrosine had significant $^{13}\text{C}_4$ $^{15}\text{N}_1$ labeling, likely from erythrose 4-phosphate via the pentose phosphate pathway indicating new synthesis through shikimate with nitrogen addition from a heavily labeled glutamate pool again.

In contrast, the lack of corresponding osmolyte synthesis would indicate that total biomass did not increase enough to demand the synthesis of additional salinity-balancing metabolites. The concentration of major osmolytes in the cell is thought to be a function of cell volume and moderated by salinity (Yancey 2005), but we did not see significant labeling or net increases in most concentrations. The lack of correlation between the metabolite pool size and the fraction of it labeled indicates that cells are carefully allocating intracellular nitrogen. The exceptions to this were proline and glycine, which have many roles in the cell as amino acids, and ectoine, which is a metabolite thought to be produced almost exclusively by heterotrophic bacteria (McParland et al. 2021). Ectoine also showed unique labeling patterns with nearly equal amounts of labeling observed at the surface and 175 meters, reinforcing the idea that it's largely synthesized by heterotrophic bacteria to whom organic nitrogen is readily available.

The general patterns shown in the known metabolites were generally reflected in the unknown metabolome as well. Our list of standards covered a majority of the signals that were high quality, had a ^{15}N isotope, and increased in peak area over the course of the incubation

according to fairly strict cutoffs. Of the novel molecules with putative identifications, the deoxyribonucleic acids were interesting for their smaller fraction labeled relative to the ribonucleic acids, implying that significantly more ^{15}N was allocated to RNA synthesis than DNA in good agreement with our general understanding of the relative DNA vs RNA turnover times. The deoxyribose forms did still show a large amount (~20%) labeling after 24 hours, however, so DNA synthesis was occurring and likely therefore DNA replication and cell division. Another interesting metabolite discovered by untargeted methods was theanine (N5-ethyl glutamine), best known for its role in tea flavoring (Vuong et al. 2011) that has since been found in fungi and can be produced from bacterial enzymes (Mu et al. 2015). This metabolite was almost exclusively singly-labeled, indicating that the ethylamine pool was not heavily labeled as it is predominately synthesized from glutamate and ethylamine. Its labeling pattern also tracked closely with that of glutamate, however, hinting at active synthesis of this compound in the marine environment for unknown reasons. Finally, the labeling of nicotinamide could be a result of vitamin B3 synthesis or the broader synthesis of nicotinamide adenine dinucleotide (NAD^+), the latter of which would also explain why adenosine turns over rapidly but does not increase in concentration throughout the incubations. However, we must note that the identification of these compounds remain putative and are based exclusively on MS^1 data via m/z match and isotope incorporation patterns.

6.4.2 Environmental context of the incubations

We collected water for the incubations from an anticyclonic eddy near Station ALOHA that was experiencing the decay of a surface diatom bloom, specifically *Hemiaulus*. *Hemiaulus* is known to be associated with the diazotroph *Richelia* and has been observed in blooms worldwide (Carpenter et al. 1999; Villareal 1994; White et al. 2018; Castillo Cieza et al. 2024). Over the

course of the incubations, we saw a return to typical Station ALOHA conditions with a well-defined deep chlorophyll maximum (DCM) and an ingrowth of *Synechococcus* and *Prochlorococcus* while the amount of heterotrophic bacteria decreased. Data from 175 meters was significantly sparser, with only a few CTD casts that nonetheless indicated relatively stable conditions below the DCM.

The incubations were amended with nitrate and ammonia in excess of environmental concentrations to ensure extensive labeling relative to the total pool and to ensure a detectable signal. The final 3 μM concentration is comparable to other recent incubations with nitrate and ammonia (Rii et al. 2018; Zimmerman et al. 2023; Shilova et al. 2017) and for a comparable length of time. Fewer incubations have been done with organic nitrogen, particularly with a whole seawater community, but we estimate that the arginine and guanosine monophosphate (GMP) that we added (50 nM) was at least a hundred times the normal environmental concentration of these compounds and likely saturating transporters (Boysen et al. 2022). However, the 4-5 μM concentration of dissolved organic nitrogen (DON) was another hundred times the concentrations we used, so while we almost certainly swamped the arginine and GMP pools (and likely the labile organic pool as a whole) we did not exceed the total organic nitrogen available in the system. Additionally, our use of GMP instead of guanosine means that organisms may have taken up GMP for the phosphate, especially at the surface where environmental concentrations were $<0.05 \mu\text{M}$, though this part of the molecule was unlabeled and we were unable to trace it.

We also did not measure the final concentrations of the amended molecules remaining in the dissolved phase at the end of the experiments. However, Rii et al. (2018) reports nitrate and ammonia uptake rates in the hundreds of nanomolar per day indicating the excess DIN was likely

available after the 26 hour incubations. We did see a sizable fraction of the organics used, however, with up to 25 nM ^{15}N labeling in metabolites and nearly 100 nM concentrations of labeled THAAs at both depths. The fact that the summed quantity of label in the THAAs exceeds the initial concentration is a result of the amended molecule providing more than 1 mole of ^{15}N per mole of substrate, with arginine containing 4 nitrogens and GMP containing 5 nitrogens for a final ^{15}N concentration of 200 nM N and 250 nM N, respectively.

An important question when assessing the applicability of this study to in situ processes is the likelihood of a community shift during the incubation. While these incubations were relatively short (1 day for the inorganics and 3 days for the organics), prior studies have shown that a significant amount of diatoms, haptophytes, and pelagophytes are capable of growing in within a few days (Rii et al. 2018; Alexander et al. 2015; McAndrew et al. 2007). The heterotrophic bacteria appear to respond even more quickly, with Shilova et al. (2017) observing a growth of Alteromonadaceae and picoeukaryotes and Goldberg et al. (2017) detecting an extensive cascade of heterotrophic organisms during a one-week incubation. In contrast, Zimmerman et al. (2023) found very few changes in community structure during their 11-hour incubations. Thus, the first few timepoints of our study likely reflected Station ALOHA conditions quite well, while the 1 day and 3 day incubations probably resulted in a community shift towards larger eukaryotes.

6.5 Conclusion

Together, these results highlight the diverse strategies employed by marine microbial communities to obtain nitrogen under oligotrophic conditions. Ammonia was rapidly and broadly utilized, underscoring its high bioavailability and central role in nitrogen metabolism across depths. In contrast, nitrate uptake was limited to a smaller subset of the community, likely

eukaryotic phytoplankton, and occurred more slowly. Organic nitrogen substrates, though used more slowly initially, were ultimately incorporated into a wide range of metabolites with striking similarity across depths and diel cycles, pointing to the central role of heterotrophic bacteria in dissolved organic nitrogen assimilation. These findings reinforce the view of the ocean's nitrogen pool as a dynamic and structured marketplace, shaped not only by elemental availability but the form it is in and by the biochemical capabilities and ecological roles of distinct microbial players.

6.6 Methods

6.6.1 Cruise information

The PARAGON cruise (KM2112) departed Honolulu, HI on July 22nd, 2021 and returned on August 5th. Satellite chlorophyll and altimetry were used to target an anticyclonic eddy with a large bloom signal. Samples were collected during the CTD upcast between 21.4 °N, 155.1 °W and 25.5 °N, 156.6 °W.

6.6.2 Biogeochemical data

Biogeochemical measurements were collected following protocols used by the HOT program (<http://hahana.soest.hawaii.edu/hot/methods/results.html>). Particulate carbon and particulate nitrogen were measured using an elemental analyzer. Nitrate + nitrite (N+N) was measured on an autoanalyzer. Soluble reactive phosphorus (phosphate, PO_4^{3-}) was measured on an autoanalyzer or following the magnesium induced coprecipitation method.

6.6.3 Community composition data

Water column community composition was estimated for the initial community via flow cytometry (FCM), imaging flowcytobot (IFCB), and amplicon sequence variants (ASVs). FCM was done three times during the cruise on July 25th, July 29th, and August 2nd at six depths

from 5 to 125 meters. Depth profiles from the IFCB were collected twice. The first IFCB profile took place on July 27th at nine depths from 5 to 300 meters and the second on August 5th at four depths from 5 to 125 meters. Annotation of the selected particles was done via a convolutional neural net. One ASV depth profile was collected during the PARAGON cruise on July 30th, also at nine depths from 5 meters down to 500 meters and processed according to the methods detailed in Li et al. (2023).

6.6.4 ^{15}N incubations

Morning CTD casts were returned to the surface between 6:40 and 7:15AM, with amendments happening ~20 minutes after the CTD surfaced. Evening casts were sometimes delayed, with most incubations beginning between 7:15 and 7:45PM but the nitrate 175m cast amended at 9:30PM, the arginine 175m cast amended at 8:30PM, and the guanosine monophosphate 25m cast amended at 8:25PM. ~30 liters of water from multiple CTD bottles were distributed into 15 prewashed (3x DI, 3x with sampled seawater) polycarbonate 2L bottles (3x biological triplicates at each of 5 timepoints) for metabolites and 6 250 mL bottles for THAAs (3x triplicates of initial and final conditions). Clean 53 μm mesh was used when transferring from CTD to incubations bottles to exclude zooplankton and large aggregates from the incubations.

2L samples were given a 1 mL spike with 6 mM $\text{Na}^{15}\text{NO}_3$, 6 mM $^{15}\text{NH}_4\text{Cl}$, 0.1 mM $^{13}\text{C}_5$ $^{15}\text{N}_4$ arginine, or $^{13}\text{C}_{10}$ $^{15}\text{N}_5$ guanosine monophosphate to obtain a 3 μM final concentration of the inorganics and a 50 nM final concentration of the organics. 250 mL samples were given a 125 μL spike to the same final concentrations. T0 samples were then immediately filtered using a peristaltic pump and Masterflex tubing onto 47 mm 0.2 μm Omnipore (PTFE) filters held by polycarbonate holders. We tried to filter the entire sample in less than 20 minutes (flow rate ~6 L/hour) but some samples took up to an hour. All 250 mL THAA samples finished within 5

minutes and were filtered onto Durapore (PVDF) instead. Filters were then removed using ethanol-washed tweezers and placed into pre-combusted foil packets before flash-freezing in liquid nitrogen and then stored at -80 °C until processing.

Later timepoints followed the same process but were stored in flow-through incubators with blue shading to match mixed layer light and temperature conditions. Inorganic amendments were sampled after 1, 3, 10, and 26 hours to avoid sampling overlap with the next set. The organic amendments were sampled at 3, 10, 26, and 73 hours except for the arginine 25 meter 6AM samples, where the T3 timepoint was missed and sampled after 18 hours instead. 175 meter samples were stored in black trash bags to minimize light exposure during the incubations as well.

6.6.5 Incubation sample processing

6.6.5.1 Metabolite extractions

A modified Bligh & Dyer dichloromethane:water/methanol extraction was used as detailed in Boysen et al. (2018). Sample filters were removed from the freezer and randomly assigned to PTFE centrifuge tubes with a 1:1 mix of 100 µm and 400 µm silica beads, approximately 2 mL -20 °C Optima-grade dichloromethane, and approximately 3 mL -20 °C 1:1 methanol/water solution (both also Optima-grade). Samples were then run through a bead-beater and had the aqueous layer removed and replaced three times. The aqueous fraction was then warmed and dried under clean nitrogen gas before reconstitution into 300 µL of Optima-grade water.

At this point the samples were split into two 30 µL aliquot for analysis both with and without a full suite of isotope-labeled internal standards. Samples that received the internal standards were used for quantification of the unlabeled metabolite while those that did not were used to estimate the isotope distributions though they still received an internal standard mix of

$^2\text{H}_8$ phenylalanine and $^2\text{H}_4$ taurine to ensure all samples were successfully injected. The full list of internal standards can be found in the Supplemental Table 2.

A pooled sample was created from 5-10 μL of each sample for a total volume of 300-600 μL that was used to assess matrix suppression and obscuring variation. We also ran samples of pure, unlabeled compounds split into two mixes in the pooled sample used to estimate the response factor of the instrument for quantification (Boysen et al. 2018). Samples were then placed directly into the autosampler for injection.

6.6.5.2 Total hydrolyzable amino acid extractions

Amino acids were hydrolyzed as in Fountoulakis and Lahm (1998) with minor modifications.

Samples were heated at 120 $^{\circ}\text{C}$ for 20 hours instead of 110 $^{\circ}\text{C}$ for 20-24 hours. Initial recovery tests with bovine serum albumin (BSA) resulted in better recovery of the amino acids at 120 $^{\circ}\text{C}$ as compared to 110 $^{\circ}\text{C}$. A shorter hydrolysis with BSA at 150 $^{\circ}\text{C}$ was also tried but recovery of amino acids was highest at 120 $^{\circ}\text{C}$.

Filters were transferred into combusted 4 mL glass vials with enough 6N hydrochloric acid to cover the filter. Each sample was purged under nitrogen gas for 30 seconds before immediately sealing with a solvent rinsed cap. The samples were then heated at 120 $^{\circ}\text{C}$ for 20 hours. The acid was then transferred to a clean, combusted glass vial. The original glass vial and filter were rinsed with approximately 500 μL of Optima grade water and transferred to the new glass vial. A rinsing step was repeated with an equal volume of Optima grade methanol. The acid mixture was concentrated to dryness under nitrogen gas and medium heat. Once dried, approximately 500 μL of water was used to rinse each vial and samples were returned to the nitrogen gas. Dried samples were re-dissolved in 400 μL of Optima grade water and syringe filtered into LCMS vials.

Amino acids were derivatized as in Gray et al. (2017) using the AccQ Tag Ultra derivatization kit from Waters. 10 μ L of each sample was transferred into a new glass LCMS vial and 70 μ L of borate buffer was added to each vial. After vortexing, 20 μ L of AccQ Tag Ultra derivatizing agent was added to each sample and the vials were heated at 55 °C for 10 minutes before being cooled and transferred to the autosampler for injection.

6.6.5.3 HPLC-MS methods

Separate LC-MS runs were used for each amendment experiment. Nitrate samples were run in August 2021, ammonia samples were run in September, and the organics were run in July 2022. Each run was treated as a single batch and injected in sequence while maintaining the randomization that had occurred during extraction to minimize chromatographic shifts from solvent or column switching.

For each run, a Waters Acquity I-Class UPLC with a SeQuant ZIC-pHILIC column (5 μ m particle size, 2.1 mm x 150 mm, from Millipore) was used with 10 mM ammonium carbonate in 85:15 acetonitrile to water (Solvent A) and 10 mM ammonium carbonate in 85:15 water to acetonitrile (Solvent B) at a flow rate of 0.15 mL/min. The column was held at 100% A for 2 minutes, ramped to 64% B over 18 minutes, ramped to 100% B over 1 minute, held at 100% B for 5 minutes, and equilibrated at 100% A for 25 minutes (50 minutes total). The column was maintained at 30 °C. The injection volume was 2 μ L for samples and standard mixes. When starting a batch, the column was equilibrated at the starting conditions for at least 30 minutes. To improve the performance of the HILIC column, we maintained the same injection volume, kept the instrument running water blanks between samples as necessary, and injected standards in a representative matrix (the pooled sample) in addition to standards in water. After each batch, the column was flushed with 10 mM ammonium carbonate in 85:15 water to acetonitrile for 20 to 30 minutes.

The Waters Acquity UPLC was coupled to a Thermo Q Exactive HF hybrid Orbitrap high resolution mass spectrometer equipped with a heated electrospray ionization source (H-ESI). Each metabolite sample was run on a full scan method three times: once from the vial with internal standards with polarity switching at a resolution of 60,000, once from the vial without internal standards at a resolution of 240,000 in positive mode, and once again from the vial without internal standards at a resolution of 240,000 in negative mode. During all injections we set a scan range of 60 to 900 m/z and with an AGC target of 1 million and a max injection time of 200 ms. The capillary temperature was set to 320 °C, the H-ESI spray voltage to 3.5 kV, and the auxiliary gas heater temperature to 90 °C. The S-lens RF level was 65. Sheath gas, auxiliary gas, and sweep gas flow rates were maintained at 16, 3, and 1, respectively. For the quality control pooled samples, both high-resolution MS scans and data dependent MS/MS scans were collected and separate injections were done for positive and negative ion modes (30,000 resolution and a dynamic exclusion time of 20 seconds for MS/MS data, top five most abundant ions in each cycle were fragmented, collision energies of 20, 35, and 50 eV). The MS/MS AGC target was set to 50,000 and the maximum injection time was 100 ms. The mass spectrometer was calibrated weekly in positive and negative mode using solutions provided by the manufacturer.

For analysis of derivatized amino acids, a Waters Acquity UPLC HSS T3 column (1.8 μ m particle size, 2.1 mm x 150 mm) was used with 0.1% formic acid in water (Solvent A) and 0.1% formic acid in acetonitrile (Solvent B) at a flow rate of 0.6 mL/min. The column was held at 4% B for 0.5 minutes, ramped to 10% B over 2 minutes, ramped 28% B over 2.5 minutes, ramped to 95% B over 0.5 minutes, held at 95% B for 1 minute, and equilibrated back to 4% B for 2.5 minutes (total run time is 9 minutes). The column temperature was maintained at 45°C. A

full scan method in positive ion mode was used with a scan range of 100 to 600 m/z and a resolution of 60,000 on a Thermo Scientific Q Exactive HF Orbitrap. The capillary temperature was 320 °C, the H-ESI spray voltage was 3.8 kV, and the auxiliary gas heater temperature was 90 °C. The S-lens RF level was 65. Sheath gas, auxiliary gas, and sweep gas flow rates were maintained at 50, 15, and 1, respectively.

6.6.6 Data analysis

Raw files were converted to mzML format via Proteowizard's MSConvert command line tool (version 3.0.25009, Chambers et al. (2012)) with vendor centroiding and separated into positive and negative mode files.

6.6.6.1 Targeted analysis

We used the RaMS R package (version 1.4.3, Kumler and Ingalls (2022)) to extract ion chromatograms from the samples without internal standards for each of the ~200 compounds for which we had a pure standard available. Each chromatogram was then visualized and the start and end retention times manually entered for each compound and MS run that had a high-quality mass feature that matched the pure standard in retention time. We then calculated all the possible isotopes for each known compound and extracted those EICs as well, using the same retention time bounds as for the unlabeled feature. We integrated each of those peak areas using a trapezoidal Riemann sum (`trapz` function in the RaMS package) if the unlabeled and labeled chromatograms shared more than 5 scans and had a correlation between the ^{14}N and ^{15}N peak intensities with a p-value below 0.01 to remove interfering ions. Samples which accidentally received isotope-labeled internal standards had the peak areas for that specific isotopologue set to zero. After the untargeted analysis we repeated this process for any compounds for which we were confident of an annotation (e.g. theanine and the MAAs).

The same process was used for the samples with internal standards except that the isotopes were not considered. These peak areas were then normalized to their best-matched internal standard (Boysen et al. 2018) and quantified using the response factors estimated from the standards in the pooled sample. Samples that accidentally did not receive isotope-labeled internal standards had their peak areas interpolated as the average value of the samples immediately preceding and following.

6.6.6.2 Untargeted analysis

We used the R package *xcms* (version 4.6.0, Colin A. Smith et al. (2006)) to perform an untargeted analysis on the positive-mode pooled and environmental samples without internal standards across all four MS runs simultaneously (282 files). We used the CentWave algorithm (Tautenhahn et al. 2008) with default settings except for *ppm* which was set to 5, *peakwidth* which was set to between 20 and 80, *prefilter* which was set to 5 scans above 10,000, *snthresh* which we set to 5, and *integrate* which we set to 2. We also enabled *verboseColumns* and *extendLengthMSW*. We did not perform retention time correction since manual inspection of the TIC and a few selected compounds did not reveal significant retention time drift. We then performed correspondence via peak density with each set of replicates counting as a sample group, the *bandwidth* set to 12, *minFraction* set to 1, *binSize* set to 0.001, and all other settings kept at their defaults. Finally, we filled peaks using the original *xcms* algorithm and default settings.

Once features were identified, we then re-integrated the individual peaks and used the same retention time bounds to integrate any data at a mass difference of -0.997035, in essence assuming that the peak detected by *xcms* was a ^{15}N peak for which a ^{14}N peak should exist to get high-quality isotope ratios. High quality features were those with a retention time above 2 minutes and a median *beta_snr* value larger than the median *beta_snr* value times -50 plus 50,

thus drawing a diagonal line through the med_snr:med_cor space above which high quality features are more likely to be found (Kumler et al. 2023). Additionally, high-quality features were only considered to be isotopes if the correlation between the intensity values of the ^{14}N and ^{15}N peaks was above 0.8. We also removed peaks that eluted between 14 and 14.5 minutes because many of them appeared to be adducts or in-source transformations of bisulfate.

6.6.7 Statistics

Statistics were run in R version 4.5.0 (R Core Team 2025). We fit logistic regression curves to the fraction of the metabolite peak area labeled using R's `glm` function to model labeling fractions and error distributions except where only the final timepoint had signal, in which case we used a one-sample t-test to determine whether the labeling was significantly different from zero using only the final timepoint's data. When testing whether untargeted compounds were increased over time we used Type-I linear models (`lm` in R) with a false-discovery rate correction (`p.adjust`, Benjamini and Hochberg (1995)) and log-scaled time and peak area. Analyses were performed on triplicate data points across the 5 timepoints for a sample size of 15 with the exception of the 175 meter 6PM nitrate incubations for which T10 was removed due to anomalous labeling patterns and unclear sampling timing (resulting in $n=12$) and replicate B in the 25 meter 6AM guanosine monophosphate amendments at T26 which was lost during sample processing ($n=14$). k-means clustering was performed via `kmeans` from the `stats` package with 10 clusters after no clear jump in variance explained was observed with other cluster numbers, indicating the labeling patterns likely exist as a gradient instead of discrete groups.

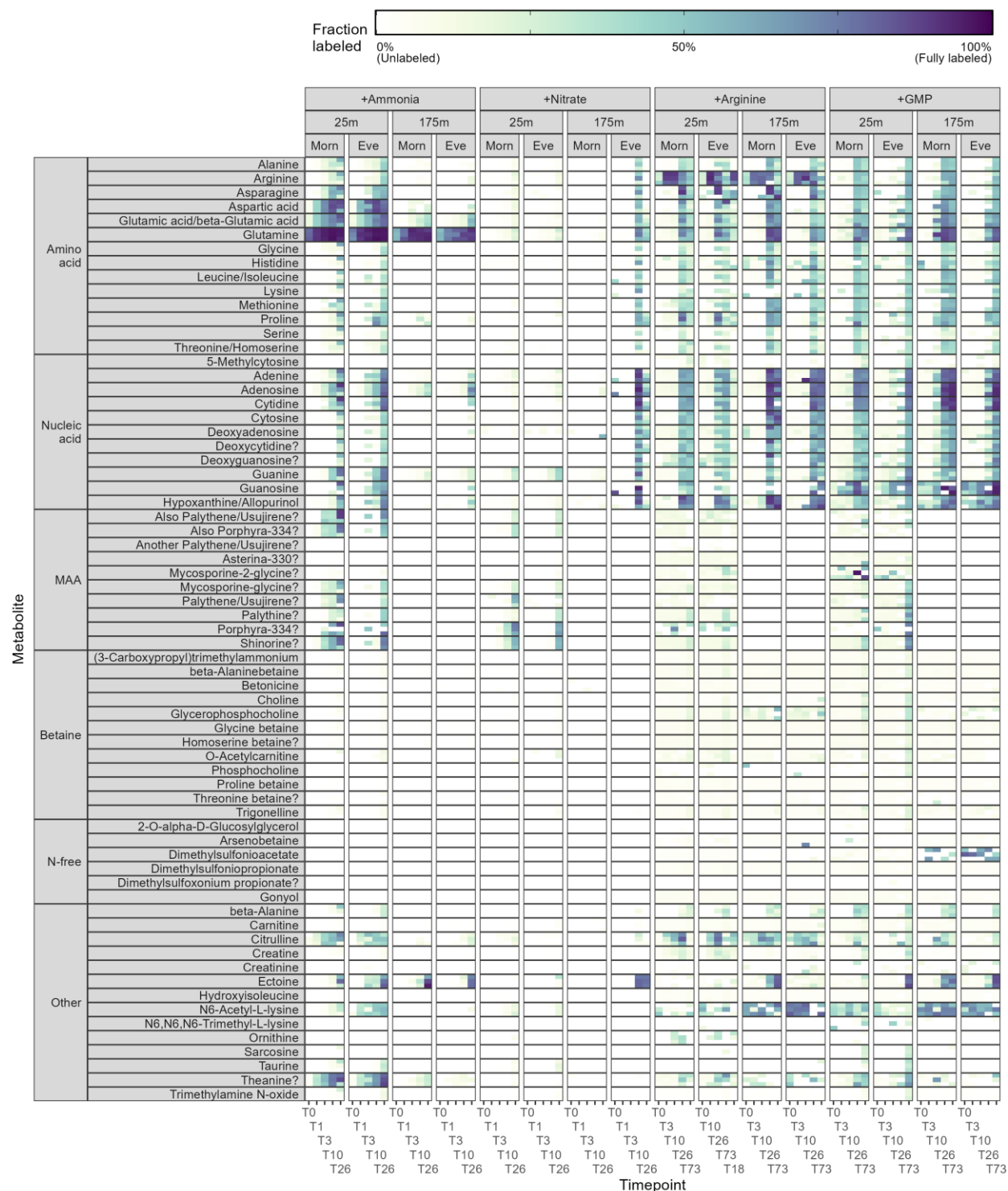
6.7 Acknowledgements

We would like to acknowledge the captain and crew of the R/V *Kilo Moana* for making this science possible. We would also like to thank the other members of the Ingalls, White, and

BEACH labs who provided assistance in sample processing, integration, and analysis. This work was supported by grants from the Simons Foundation (SCOPE Award ID 329108 to AI, SF Award ID 385428 to AI, and SCOPE Award ID 721264 to David Karl which supported BB).

The authors have no conflicts of interest to declare.

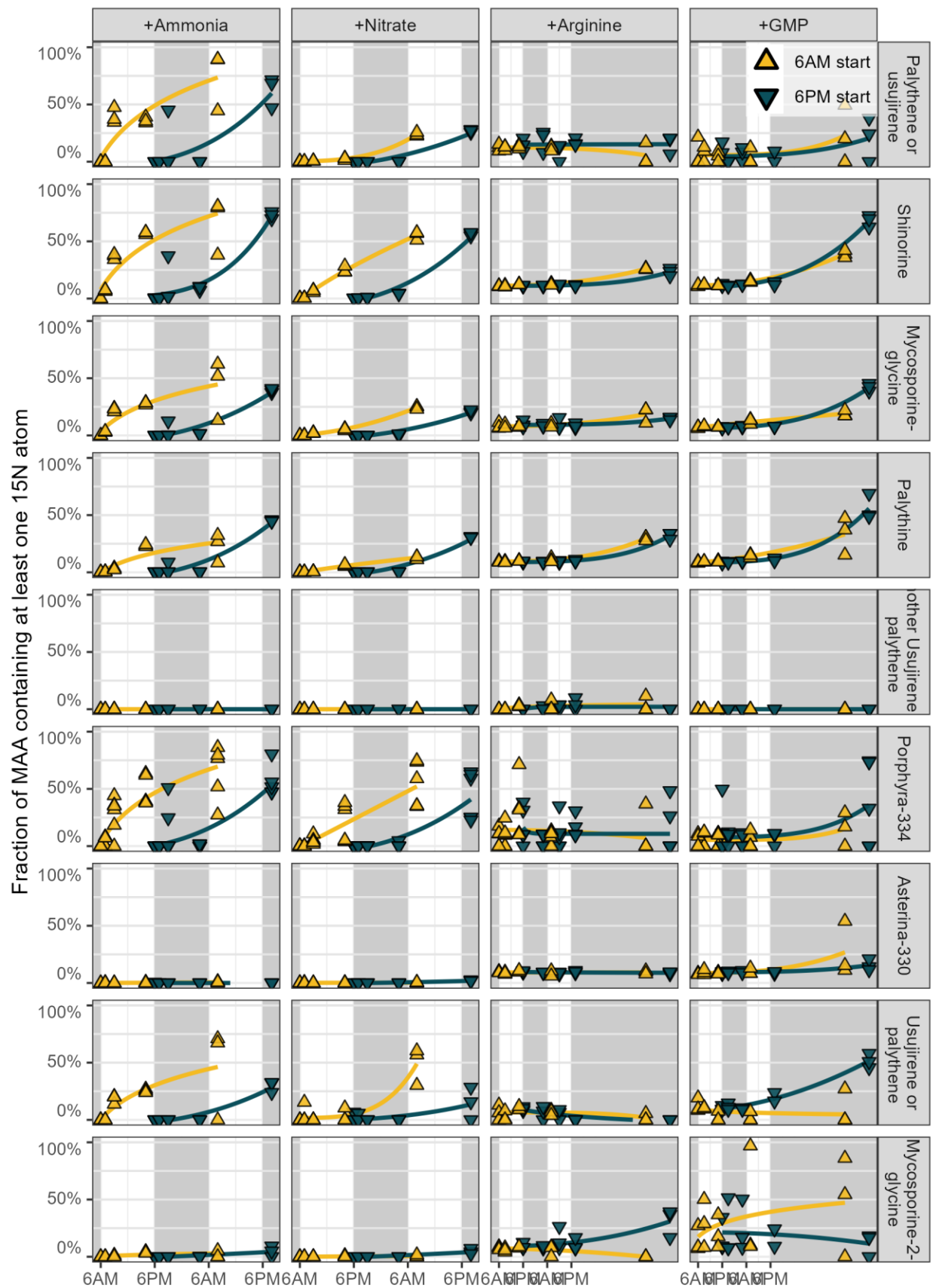
6.8 Supplemental Figures



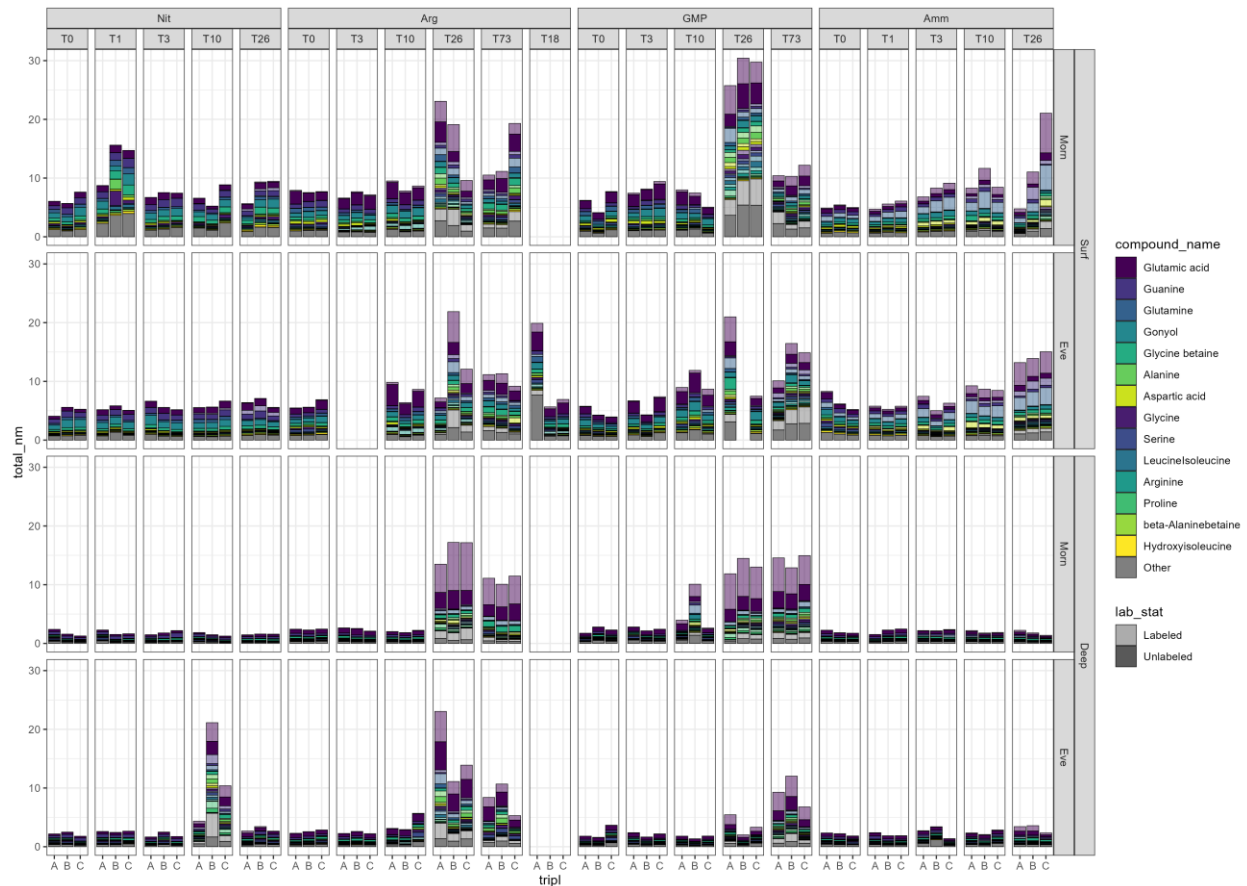
Supplemental Figure 6.11: Fraction of each metabolite containing at least one ^{15}N label during the surface seawater amendments with various substrates after the number of hours shown on the x-axis. Darker colors indicate a larger percentage of the compound signal was isotopically

labeled. Three biological replicates are stacked vertically at each timepoint and the categories with multiple entries have been averaged. GMP = guanosine monophosphate.

Supplemental Figure 6.12 (Available as a separate extended data file): Stacked bar plots showing the fraction of various total hydrolyzable amino acids (THAAs) labeled at the end of the incubations (26 hours for nitrate and ammonia, 73 hours for guanosine monophosphate (GMP) and arginine) across the four amendments. Total μM nitrogen (N) in the protein pool is shown in the top plot while the fraction of individual amino acids labeled is shown in the rows below. Colors denote specific labeling patterns detected in the THAAs.



Supplemental Figure 6.13: Fraction of UV-absorbing mycosporine-like amino acids (MAAs) containing at least one ^{15}N label over one day during the treatments amended with ^{15}N labeled ammonia or nitrate. Incubations started at 6AM are shown in yellow while incubations started at 6PM are shown in blue, with grey background regions highlighting nighttime. Three replicates are included at each of the 5 timepoints and best-fit power law curves have been fit behind the data. GMP = guanosine monophosphate.



Supplemental Figure 6.14: Quantified metabolite concentrations in nanomolar throughout the incubations for the 14 most abundant metabolites as distinguished by color. Concentrations have been broken down by whether they are fully unlabeled or contain any amount of ^{15}N (or ^{13}C in the case of the arginine and GMP amendments). Compounds other than the top 14 have been averaged and are shown in grey. Totals for all four amendments are shown here in the order they were done during the cruise, with separate panels for the 25 meter samples and the 175 meter samples as well as those started at 6AM vs 6PM. Lighter colors denote the isotopically labeled version of the compound.

Supplemental Figure 6.15 (Available as a separate extended data file): Individual untargeted metabolites containing at least one ^{15}N label during the surface seawater amendments with various substrates after the number of hours shown on the x-axis. Darker colors indicate a larger percentage of the feature signal was isotopically labeled. Three biological replicates are stacked vertically at each timepoint and the labeling within each clusters has been averaged

across all the composing features. The number of features in each cluster is specified by the n in the facet name. GMP = guanosine monophosphate.

6.9 Supplemental Tables

Available as an Extended Data file. Captions reproduced here for clarity.

Supplemental Table 6.1: Sample metadata

Supplemental Table 6.2: Authentic and internal standards

Supplemental Table 6.3: Untargeted feature identifications

Chapter 7: Conclusions

This thesis makes novel contributions to the field of marine microbial metabolomics in two ways. First, I demonstrated the need for new tools that streamline metabolomic analysis and built several of those tools for widespread use. Then, I showed how the application of those tools allowed me to produce high-quality oceanographic datasets that identified several metabolites for the first time in the open ocean. Finally, I discussed the way interactions between microbes and the marine environment depend on these small molecules.

Untargeted metabolomics is rapidly increasing in popularity, largely due to its automated methods that scale well with larger datasets unlike targeted manual approaches. In theory, it's possible for an untargeted analysis to fully encompass a targeted one and do so more quickly and reproducibly than traditional approaches. However, this is an unrealized dream at the moment due to two major bottlenecks: first, the peakpicking algorithms that return many low-quality or noise peaks and second, the low confidence in automated annotations.

This thesis focused primarily on solving the first of these problems. Most importantly, I made it enormously easier to access the raw datasets that compose metabolomics data in Chapter 2. This had multiple advantages: visualization of chromatographic peaks, development of novel metrics for assessing peakpicking, and simplifying the extraction of fragmentation data. Chromatographic peaks visualization is the primary way in which we can assess our confidence in peakpicking algorithms and assess their performance across an entire dataset at once. This means that it is a critical step in untargeted analyses but was extraordinarily painful, typically requiring minutes or hours to view a single compound. Another benefit to straightforward raw data access was the ability to develop new methods for determining the quality of

chromatographic peaks, which also improved our ability to isolate signal features from noise as shown in Chapter 3.

The second problem, that of low confidence automated annotation, also stands to benefit from the work I've done in this thesis. Fragmentation data is the primary way in which compounds are annotated but linking MS/MS information to the associated chromatographic feature is surprisingly non-trivial due to the complex data structures used in the behemoth black-box pipelines at play. My initial focus was largely on MS¹ data due to the complexity of optimizing fragment collection, the difficulty in interfacing with online databases, and ongoing questions around diagnostic fragments and consensus spectra. Thus, it came as a surprise that the RaMS package has been mostly cited by other labs mostly for use in MS/MS access. Clearly, there is much room to continue improving in both automation and untargeted methodologies in metabolomics.

Of course, the best tools in the world are useless if there are no useful questions to be asked with them. Fortunately the marine environment is full of microbial mysteries that can be answered by their molecular signals. This thesis contains one of the largest published collections (by # of small, polar molecules measured) of relative and absolute microbial metabolite abundances for the North Pacific Subtropical Gyre and possibly the ocean as a whole. Marine metabolomics is still in its infancy given that methods for quantifying metabolites in seawater were developed only recently and assembling our fundamental understanding of relative and absolute metabolite compositions is crucial basic research.

We still do not understand the scales in time and space across which metabolites vary. This thesis in particular places a special importance on the effect of depth in chapters 4 and 6, which is a major factor in community composition and therefore metabolite distribution but

previous studies have largely measured metabolites at the surface. Only a few depth profiles have been taken and reported in the literature (Heal et al. 2021; Johnson et al. 2023) and this gap is one my thesis begins to bridge. Here, I presented a total of 57 environmental samples taken from 175 meters (33 from Chapter 5 and 24 from the initial timepoints of Chapter 6) and 45 samples from the deep chlorophyll maximum in Chapter 5 for a low-resolution assay of metabolite abundance throughout the euphotic zone across multiple years.

Another major goal of metabolomics is harmonization with other 'omics fields. Genomic and transcriptomic studies are rapidly becoming more widespread, but aligning data between these possible or predicted behaviors with the molecular reality is difficult due to a lack of shared language. Pathway maps such as KEGG (Kanehisa 2000) or BioCyc (Karp et al. 2019) attempt to provide common ground but often the data itself is ambiguous, with some metabolites increased in a pathway that the transcripts would predict to be downregulated. This makes building cellular-level models of marine microbes difficult. This gap is again one that I attempted to bridge with this thesis by tracing the flow of nitrogen through a natural NPSG community in Chapter 6 and identifying additional compounds for inclusion in such maps in Chapters 4 and 5.

This thesis consists of a collection of work that advances our ability to perform metabolomics in the marine environment and leverages those advances to make novel contributions to the field. Here, I have produced both data science perspectives on the role of mass spectrometry in diving deeper into environmental data as well as novel results about the composition and variability of the marine metabolome.

Bibliography

Adams, Kendra J., Brian Pratt, Neelanjan Bose, Laura G. Dubois, Lisa St. John-Williams, Kevin M. Perrott, Karina Ky, et al. 2020. "Skyline for Small Molecules: A Unifying Software Package for Quantitative Metabolomics." *Journal of Proteome Research* 19 (4): 1447–58. doi:[10.1021/acs.jproteome.9b00640](https://doi.org/10.1021/acs.jproteome.9b00640).

Alexander, Harriet, Mónica Rouco, Sheean T. Haley, Samuel T. Wilson, David M. Karl, and Sonya T. Dyhrman. 2015. "Functional Group-Specific Traits Drive Phytoplankton Dynamics in the Oligotrophic Ocean." *Proceedings of the National Academy of Sciences* 112 (44). doi:[10.1073/pnas.1518165112](https://doi.org/10.1073/pnas.1518165112).

Aluwihare, Lihini I., and Travis Meador. 2008. "Chemical Composition of Marine Dissolved Organic Nitrogen." In *Nitrogen in the Marine Environment*, 2nd ed., 95–140. Academic Press Burlington, MA.

Anderson, Marti J. 2001. "A New Method for Non-Parametric Multivariate Analysis of Variance." *Austral Ecology* 26 (1): 32–46. doi:[10.1111/j.1442-9993.2001.01070.pp.x](https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x).

———. 2017. "Permutational Multivariate Analysis of Variance (PERMANOVA)." In *Wiley StatsRef: Statistics Reference Online*, 1–15. Wiley. doi:[10.1002/9781118445112.stat07841](https://doi.org/10.1002/9781118445112.stat07841).

Antia, N. J., P. J. Harrison, and L. Oliveira. 1991. "The Role of Dissolved Organic Nitrogen in Phytoplankton Nutrition, Cell Biology and Ecology." *Phycologia* 30 (1): 1–89. doi:[10.2216/i0031-8884-30-1-1.1](https://doi.org/10.2216/i0031-8884-30-1-1.1).

Armbrust, E. Virginia. 2009. "The Life of Diatoms in the World's Oceans." *Nature* 459 (7244): 185–92. doi:[10.1038/nature08057](https://doi.org/10.1038/nature08057).

Armbrust, E. Virginia, John A. Berges, Chris Bowler, Beverley R. Green, Diego Martinez, Nicholas H. Putnam, Shiguo Zhou, et al. 2004. "The Genome of the Diatom *Thalassiosira Pseudonana* : Ecology, Evolution, and Metabolism." *Science* 306 (5693): 79–86. doi:[10.1126/science.1101156](https://doi.org/10.1126/science.1101156).

Aron, Allegra T., Emily C. Gentry, Kerry L. McPhail, Louis-Félix Nothias, Mélissa Nothias-Esposito, Amina Bouslimani, Daniel Petras, et al. 2020. "Reproducible Molecular Networking of Untargeted Mass Spectrometry Data Using GNPS." *Nature Protocols* 15 (6): 1954–91. doi:[10.1038/s41596-020-0317-5](https://doi.org/10.1038/s41596-020-0317-5).

Askenazi, Manor, Hisham Ben Hamidane, and Johannes Graumann. 2017. "The Arc of Mass Spectrometry Exchange Formats Is Long, but It Bends Toward HDF5." *Mass Spectrometry Reviews* 36 (5): 668–73. doi:[10.1002/mas.21522](https://doi.org/10.1002/mas.21522).

Bai, Caihong, Suyun Xu, Jingyi Tang, Yuxi Zhang, Jiahui Yang, and Kaifeng Hu. 2022. "A 'Shape-Orientated' Algorithm Employing an Adapted Marr Wavelet and Shape Matching Index Improves the Performance of Continuous Wavelet Transform for Chromatographic Peak Detection and Quantification." *Journal of Chromatography A* 1673: 463086. doi:[10.1016/j.chroma.2022.463086](https://doi.org/10.1016/j.chroma.2022.463086).

- Bajad, Sunil U., Wenyun Lu, Elizabeth H. Kimball, Jie Yuan, Celeste Peterson, and Joshua D. Rabinowitz. 2006. "Separation and Quantitation of Water Soluble Cellular Metabolites by Hydrophilic Interaction Chromatography-Tandem Mass Spectrometry." *Journal of Chromatography A* 1125 (1): 76–88. doi:[10.1016/j.chroma.2006.05.019](https://doi.org/10.1016/j.chroma.2006.05.019).
- Barone, Benedetto, Matthew J. Church, Mathilde Dugenne, Nicholas J. Hawco, Oliver Jahn, Angelicque E. White, Seth G. John, Michael J. Follows, Edward F. DeLong, and David M. Karl. 2022. "Biogeochemical Dynamics in Adjacent Mesoscale Eddies of Opposite Polarity." *Global Biogeochemical Cycles* 36 (2): e2021GB007115. doi:[10.1029/2021GB007115](https://doi.org/10.1029/2021GB007115).
- Barone, Benedetto, Ashley R. Coenen, Stephen J. Beckett, Dennis J. McGillicuddy, Joshua S. Weitz, and David M. Karl. 2019. "The Ecological and Biogeochemical State of the North Pacific Subtropical Gyre Is Linked to Sea Surface Height." *Journal of Marine Research* 77 (2): 215–45. doi:[10.1357/002224019828474241](https://doi.org/10.1357/002224019828474241).
- Barrett, Tyson, Matt Dowle, Arun Srinivasan, Jan Gorecki, Michael Chirico, Toby Hocking, Benjamin Schwendinger, and Ivan Krylov. 2006. "Data.table: Extension of 'Data.frame'." Comprehensive R Archive Network. doi:[10.32614/CRAN.package.data.table](https://doi.org/10.32614/CRAN.package.data.table).
- Barupal, Dinesh Kumar, Sadjad Fakouri Baygi, Robert O. Wright, and Manish Arora. 2021. "Data Processing Thresholds for Abundance and Sparsity and Missed Biological Insights in an Untargeted Chemical Analysis of Blood Specimens for Exposomics." *Frontiers in Public Health* 9 (June). doi:[10.3389/fpubh.2021.653599](https://doi.org/10.3389/fpubh.2021.653599).
- Bauermeister, Anelize, Helena Mannocho-Russo, Leticia V. Costa-Lotufo, Alan K. Jarmusch, and Pieter C. Dorrestein. 2022. "Mass Spectrometry-Based Metabolomics in Microbiome Investigations." *Nature Reviews Microbiology* 20 (3): 143–60. doi:[10.1038/s41579-021-00621-9](https://doi.org/10.1038/s41579-021-00621-9).
- Beagley, Nathaniel, Chad Scherrer, Yan Shi, Brian H. Clowers, William F. Danielson, and Anuj R. Shah. 2009. "Increasing the Efficiency of Data Storage and Analysis Using Indexed Compression." In *2009 Fifth IEEE International Conference on e-Science*, 66–71. Oxford, United Kingdom: IEEE. doi:[10.1109/e-Science.2009.18](https://doi.org/10.1109/e-Science.2009.18).
- Becker, Kevin W., James R. Collins, Bryndan P. Durham, Ryan D. Groussman, Angelicque E. White, Helen F. Fredricks, Justin E. Ossolinski, et al. 2018. "Daily Changes in Phytoplankton Lipidomes Reveal Mechanisms of Energy Storage in the Open Ocean." *Nature Communications* 9 (1): 5179. doi:[10.1038/s41467-018-07346-z](https://doi.org/10.1038/s41467-018-07346-z).
- Bender, Sara J., Micaela S. Parker, and E. Virginia Armbrust. 2012. "Coupled Effects of Light and Nitrogen Source on the Urea Cycle and Nitrogen Metabolism over a Diel Cycle in the Marine Diatom *Thalassiosira Pseudonana*." *Protist* 163 (2): 232–51. doi:[10.1016/j.protis.2011.07.008](https://doi.org/10.1016/j.protis.2011.07.008).
- Bengtsson, Henrik. 2021. "A Unifying Framework for Parallel and Distributed Processing in R Using Futures." *The R Journal* 13 (2): 208. doi:[10.32614/RJ-2021-048](https://doi.org/10.32614/RJ-2021-048).
- Benitez-Nelson, Claudia R., Robert R. Bidigare, Tommy D. Dickey, Michael R. Landry, Carrie L. Leonard, Susan L. Brown, Francesco Nencioli, et al. 2007. "Mesoscale Eddies Drive

Increased Silica Export in the Subtropical Pacific Ocean.” *Science* 316 (5827): 1017–21. doi:[10.1126/science.1136221](https://doi.org/10.1126/science.1136221).

Benjamini, Yoav, and Yoel Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300. doi:[10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x).

Benner, Ronald, and Rainer M. W. Amon. 2015. “The Size-Reactivity Continuum of Major Bioelements in the Ocean.” *Annual Review of Marine Science* 7 (1): 185–205. doi:[10.1146/annurev-marine-010213-135126](https://doi.org/10.1146/annurev-marine-010213-135126).

Bent, Shavonna M., Daniel Muratore, Kevin W. Becker, Benedetto Barone, Tara Clemente, Helen F. Fredricks, Henry C. Holm, David M. Karl, and Benjamin A. S. Van Mooy. 2024. “Lipid Biochemical Diversity and Dynamics Reveal Phytoplankton Nutrient-Stress Responses and Carbon Export Mechanisms in Mesoscale Eddies in the North Pacific Subtropical Gyre.” *Frontiers in Marine Science* 11 (September): 1427524. doi:[10.3389/fmars.2024.1427524](https://doi.org/10.3389/fmars.2024.1427524).

Benton, H. Paul, Elizabeth J. Want, and Timothy M. D. Ebbels. 2010. “Correction of Mass Calibration Gaps in Liquid Chromatography–Mass Spectrometry Metabolomics Data.” *Bioinformatics (Oxford, England)* 26 (19): 2488–89. doi:[10.1093/bioinformatics/btq441](https://doi.org/10.1093/bioinformatics/btq441).

Berthelot, Hugo, Solange Duhamel, Stéphane L’Helguen, Jean-François Maguer, and Nicolas Cassar. 2021. “Inorganic and Organic Carbon and Nitrogen Uptake Strategies of Picoplankton Groups in the Northwestern Atlantic Ocean.” *Limnology and Oceanography* 66 (10): 3682–96. doi:[10.1002/lno.11909](https://doi.org/10.1002/lno.11909).

Bertrand, Erin M., John P. McCrow, Ahmed Moustafa, Hong Zheng, Jeffrey B. McQuaid, Tom O. Delmont, Anton F. Post, et al. 2015. “Phytoplankton–Bacterial Interactions Mediate Micronutrient Colimitation at the Coastal Antarctic Sea Ice Edge.” *Proceedings of the National Academy of Sciences* 112 (32): 9938–43. doi:[10.1073/pnas.1501615112](https://doi.org/10.1073/pnas.1501615112).

Berube, Paul M., Steven J. Biller, Alyssa G. Kent, Jessie W. Berta-Thompson, Sara E. Roggensack, Kathryn H. Roache-Johnson, Marcia Ackerman, et al. 2015. “Physiology and Evolution of Nitrate Acquisition in *Prochlorococcus*.” *The ISME Journal* 9 (5): 1195–1207. doi:[10.1038/ismej.2014.211](https://doi.org/10.1038/ismej.2014.211).

Bhamber, Ranjeet S., Andris Jankevics, Eric W. Deutsch, Andrew R. Jones, and Andrew W. Dowsey. 2021. “mzMLb: A Future-Proof Raw Mass Spectrometry Data Format Based on Standards-Compliant mzML and Optimized for Speed and Storage Requirements.” *Journal of Proteome Research* 20 (1): 172–83. doi:[10.1021/acs.jproteome.0c00192](https://doi.org/10.1021/acs.jproteome.0c00192).

Bilbao, Aivett, Dylan H. Ross, Joon-Yong Lee, Micah T. Donor, Sarah M. Williams, Ying Zhu, Yehia M. Ibrahim, Richard D. Smith, and Xueyun Zheng. 2023. “MZA: A Data Conversion Tool to Facilitate Software Development and Artificial Intelligence Research in Multidimensional Mass Spectrometry.” *Journal of Proteome Research* 22 (2): 508–13. doi:[10.1021/acs.jproteome.2c00313](https://doi.org/10.1021/acs.jproteome.2c00313).

Bouyssié, David, Marc Dubois, Sara Nasso, Anne Gonzalez De Peredo, Odile Burlet-Schiltz, Ruedi Aebersold, and Bernard Monsarrat. 2015. “mzDB: A File Format Using Multiple Indexing

Strategies for the Efficient Analysis of Large LC-MS/MS and SWATH-MS Data Sets *.” *Molecular & Cellular Proteomics* 14 (3): 771–81. doi:[10.1074/mcp.O114.039115](https://doi.org/10.1074/mcp.O114.039115).

Boyd, C., and D. Grädmann. 2002. “Impact of Osmolytes on Buoyancy of Marine Phytoplankton.” *Marine Biology* 141 (4): 605–18. doi:[10.1007/s00227-002-0872-z](https://doi.org/10.1007/s00227-002-0872-z).

Boysen, Angela K., Laura T. Carlson, Bryndan P. Durham, Ryan D. Groussman, Frank O. Aylward, François Ribalet, Katherine R. Heal, et al. 2021. “Particulate Metabolites and Transcripts Reflect Diel Oscillations of Microbial Activity in the Surface Ocean.” Edited by Jeff Bowman. *mSystems* 6 (3): e00896–20. doi:[10.1128/mSystems.00896-20](https://doi.org/10.1128/mSystems.00896-20).

Boysen, Angela K., Bryndan P. Durham, William Kumler, Rebecca S. Key, Katherine R. Heal, Laura T. Carlson, Ryan D. Groussman, E. Virginia Armbrust, and Anitra E. Ingalls. 2022. “Glycine Betaine Uptake and Metabolism in Marine Microbial Communities.” *Environmental Microbiology* 24 (5): 2380–2403. doi:[10.1111/1462-2920.16020](https://doi.org/10.1111/1462-2920.16020).

Boysen, Angela K., Katherine R. Heal, Laura T. Carlson, and Anitra E. Ingalls. 2018. “Best-Matched Internal Standard Normalization in Liquid Chromatography–Mass Spectrometry Metabolomics Applied to Environmental Samples.” *Analytical Chemistry* 90 (2): 1363–69. doi:[10.1021/acs.analchem.7b04400](https://doi.org/10.1021/acs.analchem.7b04400).

Bradley, Paul B., Marta P. Sanderson, Marc E. Frischer, Jennifer Brofft, Melissa G. Booth, Lee J. Kerkhof, and Deborah A. Bronk. 2010. “Inorganic and Organic Nitrogen Uptake by Phytoplankton and Heterotrophic Bacteria in the Stratified Mid-Atlantic Bight.” *Estuarine, Coastal and Shelf Science* 88 (4): 429–41. doi:[10.1016/j.ecss.2010.02.001](https://doi.org/10.1016/j.ecss.2010.02.001).

Bronk, D. A., J. H. See, P. Bradley, and L. Killberg. 2007. “DON as a Source of Bioavailable Nitrogen for Phytoplankton.” *Biogeosciences* 4 (3): 283–96. doi:[10.5194/bg-4-283-2007](https://doi.org/10.5194/bg-4-283-2007).

Bronk, Deborah A., Rachel E. Sipler, Robert T. Letscher, and Matthew D. McCarthy. 2024. “Dissolved Organic Nitrogen.” In *Biogeochemistry of Marine Dissolved Organic Matter*, 343–404. Elsevier. doi:[10.1016/B978-0-443-13858-4.00016-2](https://doi.org/10.1016/B978-0-443-13858-4.00016-2).

Bronk, Deborah A., and Deborah K. Steinberg. 2008. “Nitrogen Regeneration.” In *Nitrogen in the Marine Environment*, 2nd ed., 385–450. Academic Press Burlington, MA.

Buszewski, Bogusław, and Sylwia Noga. 2012. “Hydrophilic Interaction Liquid Chromatography (HILIC)—a Powerful Separation Technique.” *Analytical and Bioanalytical Chemistry* 402 (1): 231–47. doi:[10.1007/s00216-011-5308-5](https://doi.org/10.1007/s00216-011-5308-5).

Cajka, Tomas, and Oliver Fiehn. 2016. “Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics.” *Analytical Chemistry* 88 (1): 524–45. doi:[10.1021/acs.analchem.5b04491](https://doi.org/10.1021/acs.analchem.5b04491).

Carpenter, Ej, Jp Montoya, J Burns, Mr Mulholland, A Subramaniam, and Dg Capone. 1999. “Extensive Bloom of a N₂-fixing Diatom/Cyanobacterial Association in the Tropical Atlantic Ocean.” *Marine Ecology Progress Series* 185: 273–83. doi:[10.3354/meps185273](https://doi.org/10.3354/meps185273).

Castillo Cieza, S. Alejandra, Rachel H. R. Stanley, Pierre Marrec, Diana N. Fontaine, E. Taylor Crockford, Dennis J. McGillicuddy Jr., Arshia Mehta, et al. 2024. “Unusual *Hemiaulus* Bloom

Influences Ocean Productivity in Northeastern US Shelf Waters.” *Biogeosciences* 21 (5): 1235–57. doi:[10.5194/bg-21-1235-2024](https://doi.org/10.5194/bg-21-1235-2024).

Chambers, Matthew C, Brendan Maclean, Robert Burke, Dario Amodei, Daniel L Ruderman, Steffen Neumann, Laurent Gatto, et al. 2012. “A Cross-Platform Toolkit for Mass Spectrometry and Proteomics.” *Nature Biotechnology* 30 (10): 918–20. doi:[10.1038/nbt.2377](https://doi.org/10.1038/nbt.2377).

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihue Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barb Borges. 2024. “Shiny: Web Application Framework for R.”

Coble, Jamie B., and Carlos G. Fraga. 2014. “Comparative Evaluation of Preprocessing Freeware on Chromatography/Mass Spectrometry Data for Signature Discovery.” *Journal of Chromatography A* 1358 (September): 155–64. doi:[10.1016/j.chroma.2014.06.100](https://doi.org/10.1016/j.chroma.2014.06.100).

Codd, E. F. 1970. “A Relational Model of Data for Large Shared Data Banks.” *Communications of the ACM* 13 (6): 377–87. doi:[10.1145/362384.362685](https://doi.org/10.1145/362384.362685).

Collette, Andrew, James Tocknell, Thomas A Caswell, Darren Dale, Ulrik Kofoed Pedersen, Aleksandar Jelenak, Andrea Bedini, et al. 2017. “H5py/H5py: REL: 2.7.1.” Zenodo. doi:[10.5281/ZENODO.877338](https://doi.org/10.5281/ZENODO.877338).

Cornec, Marin, Rémi Laxenaire, Sabrina Speich, and Hervé Claustre. 2021. “Impact of Mesoscale Eddies on Deep Chlorophyll Maxima.” *Geophysical Research Letters* 48 (15): e2021GL093470. doi:[10.1029/2021GL093470](https://doi.org/10.1029/2021GL093470).

Dawson, H M, E Connors, N G Erazo, J S Sacks, V Mierzejewski, S M Rundell, L T Carlson, et al. 2023. “Microbial Metabolomic Responses to Changes in Temperature and Salinity Along the Western Antarctic Peninsula.” *The ISME Journal* 17 (11): 2035–46. doi:[10.1038/s41396-023-01475-0](https://doi.org/10.1038/s41396-023-01475-0).

Dawson, H M, K R Heal, A Torstensson, L T Carlson, A E Ingalls, and J N Young. 2020. “Large Diversity in Nitrogen- and Sulfur-Containing Compatible Solute Profiles in Polar and Temperate Diatoms.” *Integrative and Comparative Biology* 60 (6): 1401–13. doi:[10.1093/icb/icaa133](https://doi.org/10.1093/icb/icaa133).

DeLong, Edward F. 2005. “Microbial Community Genomics in the Ocean.” *Nature Reviews Microbiology* 3 (6): 459–69. doi:[10.1038/nrmicro1158](https://doi.org/10.1038/nrmicro1158).

Deutsch, Eric W. 2010. “Mass Spectrometer Output File Format mzML.” In *Proteome Bioinformatics*, edited by Simon J. Hubbard and Andrew R. Jones, 604:319–31. Totowa, NJ: Humana Press. doi:[10.1007/978-1-60761-444-9_22](https://doi.org/10.1007/978-1-60761-444-9_22).

Dugenne, Mathilde, Mary R. Gradoville, Matthew J. Church, Samuel T. Wilson, Uri Sheyn, Matthew J. Harke, Karin M. Björkman, et al. 2023. “Nitrogen Fixation in Mesoscale Eddies of the North Pacific Subtropical Gyre: Patterns and Mechanisms.” *Global Biogeochemical Cycles* 37 (4): e2022GB007386. doi:[10.1029/2022GB007386](https://doi.org/10.1029/2022GB007386).

Dugenne, Mathilde, Fernanda Henderikx Freitas, Samuel T. Wilson, David M. Karl, and Angelicque E. White. 2020. “Life and Death of *Crocospaera* Sp. In the Pacific Ocean: Fine

Scale Predator–Prey Dynamics.” *Limnology and Oceanography* 65 (11): 2603–17.
doi:[10.1002/lno.11473](https://doi.org/10.1002/lno.11473).

Dunn, Olive Jean. 1964. “Multiple Comparisons Using Rank Sums.” *Technometrics : A Journal of Statistics for the Physical, Chemical, and Engineering Sciences* 6 (3): 241–52.
doi:[10.1080/00401706.1964.10490181](https://doi.org/10.1080/00401706.1964.10490181).

Durham, Bryndan P., Angela K. Boysen, Katherine R. Heal, Laura T. Carlson, Rachel Boccamazzo, Chloe R. Deodato, Wei Qin, Rose Ann Cattolico, E. Virginia Armbrust, and Anitra E. Ingalls. 2022. “Chemotaxonomic Patterns in Intracellular Metabolites of Marine Microbial Plankton.” *Frontiers in Marine Science* 9 (September): 864796. doi:[10.3389/fmars.2022.864796](https://doi.org/10.3389/fmars.2022.864796).

Edwards, Bethanie R. 2023. “Lipid Biogeochemistry and Modern Lipidomic Techniques.” *Annual Review of Marine Science* 15 (1): 485–508. doi:[10.1146/annurev-marine-040422-094104](https://doi.org/10.1146/annurev-marine-040422-094104).

El Abiead, Yasin, Maximilian Milford, Reza M Salek, and Gunda Koellensperger. 2021. “mzRAPP: A Tool for Reliability Assessment of Data Pre-Processing in Non-Targeted Metabolomics.” Edited by Pier Luigi Martelli. *Bioinformatics (Oxford, England)* 37 (20): 3678–80. doi:[10.1093/bioinformatics/btab231](https://doi.org/10.1093/bioinformatics/btab231).

Falkowski, Paul G. 1994. “The Role of Phytoplankton Photosynthesis in Global Biogeochemical Cycles.” *Photosynthesis Research* 39 (3): 235–58. doi:[10.1007/BF00014586](https://doi.org/10.1007/BF00014586).

Falkowski, Paul G., Tom Fenchel, and Edward F. Delong. 2008. “The Microbial Engines That Drive Earth’s Biogeochemical Cycles.” *Science* 320 (5879): 1034–39.
doi:[10.1126/science.1153213](https://doi.org/10.1126/science.1153213).

Ferrer-González, Frank Xavier, Brittany Widner, Nicole R. Holderman, John Glushka, Arthur S. Edison, Elizabeth B. Kujawinski, and Mary Ann Moran. 2021. “Resource Partitioning of Phytoplankton Metabolites That Support Bacterial Heterotrophy.” *The ISME Journal* 15 (3): 762–73. doi:[10.1038/s41396-020-00811-y](https://doi.org/10.1038/s41396-020-00811-y).

Fong, Allison A, David M Karl, Roger Lukas, Ricardo M Letelier, Jonathan P Zehr, and Matthew J Church. 2008. “Nitrogen Fixation in an Anticyclonic Eddy in the Oligotrophic North Pacific Ocean.” *The ISME Journal* 2 (6): 663–76. doi:[10.1038/ismej.2008.22](https://doi.org/10.1038/ismej.2008.22).

Fountoulakis, Michael, and Hans-Werner Lahm. 1998. “Hydrolysis and Amino Acid Composition Analysis of Proteins.” *Journal of Chromatography A* 826 (2): 109–34.
doi:[10.1016/S0021-9673\(98\)00721-3](https://doi.org/10.1016/S0021-9673(98)00721-3).

Garnier, Simon, Noam Ross, BoB Rudis, Antoine Filipovic-Pierucci, Tal Galili, Timelyportfolio, Alan O’Callaghan, et al. 2023. “Sjmgarnier/Viridis: CRAN Release V0.6.3.” Zenodo.
doi:[10.5281/ZENODO.4679423](https://doi.org/10.5281/ZENODO.4679423).

Gatto, Laurent, Sebastian Gibb, and Johannes Rainer. 2021. “MSnbase, Efficient and Elegant R-Based Processing and Visualization of Raw Mass Spectrometry Data.” *Journal of Proteome Research* 20 (1): 1063–69. doi:[10.1021/acs.jproteome.0c00313](https://doi.org/10.1021/acs.jproteome.0c00313).

Gebser, Björn, and Georg Pohnert. 2013. “Synchronized Regulation of Different Zwitterionic Metabolites in the Osmoadaption of Phytoplankton.” *Marine Drugs* 11 (6): 2168–82. doi:[10.3390/md11062168](https://doi.org/10.3390/md11062168).

Giera, Martin, Oscar Yanes, and Gary Siuzdak. 2022. “Metabolite Discovery: Biochemistry’s Scientific Driver.” *Cell Metabolism* 34 (1): 21–34. doi:[10.1016/j.cmet.2021.11.005](https://doi.org/10.1016/j.cmet.2021.11.005).

Gika, Helen, Christina Virgiliou, Georgios Theodoridis, Robert S. Plumb, and Ian D. Wilson. 2019. “Untargeted LC/MS-based Metabolic Phenotyping (Metabonomics/Metabolomics): The State of the Art.” *Journal of Chromatography B* 1117 (June): 136–47. doi:[10.1016/j.jchromb.2019.04.009](https://doi.org/10.1016/j.jchromb.2019.04.009).

Gleich, Samantha J., Sarah K. Hu, Arianna I. Krinos, and David A. Caron. 2024. “Protistan Community Composition and Metabolism in the North Pacific Subtropical Gyre: Influences of Mesoscale Eddies and Depth.” *Environmental Microbiology* 26 (1): e16556. doi:[10.1111/1462-2920.16556](https://doi.org/10.1111/1462-2920.16556).

Gloaguen, Yoann, Jennifer A. Kirwan, and Dieter Beule. 2022. “Deep Learning-Assisted Peak Curation for Large-Scale LC-MS Metabolomics.” *Analytical Chemistry* 94 (12): 4930–37. doi:[10.1021/acs.analchem.1c02220](https://doi.org/10.1021/acs.analchem.1c02220).

Goldberg, S. J., C. E. Nelson, D. A. Viviani, C. N. Shulse, and M. J. Church. 2017. “Cascading Influence of Inorganic Nitrogen Sources on DOM Production, Composition, Lability and Microbial Community Structure in the Open Ocean.” *Environmental Microbiology* 19 (9): 3450–64. doi:[10.1111/1462-2920.13825](https://doi.org/10.1111/1462-2920.13825).

Gonsior, Michael, Leanne Powers, Madeline Lahm, and Shannon Leigh McCallister. 2022. “New Perspectives on the Marine Carbon Cycle—The Marine Dissolved Organic Matter Reactivity Continuum.” *Environmental Science & Technology* 56 (9): 5371–80. doi:[10.1021/acs.est.1c08871](https://doi.org/10.1021/acs.est.1c08871).

Gray, Nicola, Rabiya Zia, Adam King, Vishal C. Patel, Julia Wendon, Mark J. W. McPhail, Muireann Coen, Robert S. Plumb, Ian D. Wilson, and Jeremy K. Nicholson. 2017. “High-Speed Quantitative UPLC-MS Analysis of Multiple Amines in Human Plasma and Serum via Precolumn Derivatization with 6-Aminoquinolyl- *N* -Hydroxysuccinimidyl Carbamate: Application to Acetaminophen-Induced Liver Failure.” *Analytical Chemistry* 89 (4): 2478–87. doi:[10.1021/acs.analchem.6b04623](https://doi.org/10.1021/acs.analchem.6b04623).

Guo, Jian, Sam Shen, Shipei Xing, Ying Chen, Frank Chen, Elizabeth M. Porter, Huaxu Yu, and Tao Huan. 2021. “EVA: Evaluation of Metabolic Feature Fidelity Using a Deep Learning Model Trained with over 25000 Extracted Ion Chromatograms.” *Analytical Chemistry* 93 (36): 12181–86. doi:[10.1021/acs.analchem.1c01309](https://doi.org/10.1021/acs.analchem.1c01309).

Handy, Kyle, Jebediah Rosen, André Gillan, and Rob Smith. 2017. “Fast, Axis-Agnostic, Dynamically Summarized Storage and Retrieval for Mass Spectrometry Data.” Edited by Jon M. Jacobs. *PLOS ONE* 12 (11): e0188059. doi:[10.1371/journal.pone.0188059](https://doi.org/10.1371/journal.pone.0188059).

Harke, Matthew J., Kyle R. Frischkorn, Gwenn M. M. Hennon, Sheean T. Haley, Benedetto Barone, David M. Karl, and Sonya T. Dyhrman. 2021. “Microbial Community Transcriptional

Patterns Vary in Response to Mesoscale Forcing in the North Pacific Subtropical Gyre.” *Environmental Microbiology* 23 (8): 4807–22. doi:[10.1111/1462-2920.15677](https://doi.org/10.1111/1462-2920.15677).

Haug, Kenneth, Keeva Cochrane, Venkata Chandrasekhar Nainala, Mark Williams, Jiakang Chang, Kalai Vanii Jayaseelan, and Claire O’Donovan. 2019. “MetaboLights: A Resource Evolving in Response to the Needs of Its Scientific Community.” *Nucleic Acids Research*, November, gkz1019. doi:[10.1093/nar/gkz1019](https://doi.org/10.1093/nar/gkz1019).

Hawco, Nicholas J., Benedetto Barone, Matthew J. Church, Lydia Babcock-Adams, Daniel J. Repeta, Emma K. Wear, Rhea K. Foreman, et al. 2021. “Iron Depletion in the Deep Chlorophyll Maximum: Mesoscale Eddies as Natural Iron Fertilization Experiments.” *Global Biogeochemical Cycles* 35 (12): e2021GB007112. doi:[10.1029/2021GB007112](https://doi.org/10.1029/2021GB007112).

Heal, Katherine R., Bryndan P. Durham, Angela K. Boysen, Laura T. Carlson, Wei Qin, François Ribalet, Angelique E. White, Randelle M. Bundy, E. Virginia Armbrust, and Anitra E. Ingalls. 2021. “Marine Community Metabolomes Carry Fingerprints of Phytoplankton Community Composition.” Edited by Manuel Liebeke. *mSystems* 6 (3): e01334–20. doi:[10.1128/mSystems.01334-20](https://doi.org/10.1128/mSystems.01334-20).

Heal, Katherine R., Wei Qin, Francois Ribalet, Anthony D. Bertagnolli, Willow Coyote-Maestas, Laura R. Hmelo, James W. Moffett, et al. 2017. “Two Distinct Pools of B₁₂ Analogs Reveal Community Interdependencies in the Ocean.” *Proceedings of the National Academy of Sciences* 114 (2): 364–69. doi:[10.1073/pnas.1608462114](https://doi.org/10.1073/pnas.1608462114).

Heuckeroth, Steffen, Tito Damiani, Aleksandr Smirnov, Olena Mokshyna, Corinna Brungs, Ansgar Korf, Joshua David Smith, et al. 2024. “Reproducible Mass Spectrometry Data Processing and Compound Annotation in MZmine 3.” *Nature Protocols* 19 (9): 2597–2641. doi:[10.1038/s41596-024-00996-y](https://doi.org/10.1038/s41596-024-00996-y).

Hohrenk, Lotta L., Fabian Itzel, Nicolai Baetz, Jochen Tuerk, Maryam Vosough, and Torsten C. Schmidt. 2020. “Comparison of Software Tools for Liquid Chromatography–High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples.” *Analytical Chemistry* 92 (2): 1898–1907. doi:[10.1021/acs.analchem.9b04095](https://doi.org/10.1021/acs.analchem.9b04095).

Houriet, Joelle, Warren S. Vidar, Preston K. Manwill, Daniel A. Todd, and Nadja B. Cech. 2022. “How Low Can You Go? Selecting Intensity Thresholds for Untargeted Metabolomics Data Preprocessing.” *Analytical Chemistry* 94 (51): 17964–71. doi:[10.1021/acs.analchem.2c04088](https://doi.org/10.1021/acs.analchem.2c04088).

Hutchins, David A., and Douglas G. Capone. 2022. “The Marine Nitrogen Cycle: New Developments and Global Change.” *Nature Reviews Microbiology* 20 (7): 401–14. doi:[10.1038/s41579-022-00687-z](https://doi.org/10.1038/s41579-022-00687-z).

Ipsen, Andreas, Elizabeth J. Want, John C. Lindon, and Timothy M. D. Ebbels. 2010. “A Statistically Rigorous Test for the Identification of Parent-Fragment Pairs in LC-MS Datasets.” *Analytical Chemistry* 82 (5): 1766–78. doi:[10.1021/ac902361f](https://doi.org/10.1021/ac902361f).

Iversen, Morten H. 2023. “Carbon Export in the Ocean: A Biologist’s Perspective.” *Annual Review of Marine Science* 15 (1): 357–81. doi:[10.1146/annurev-marine-032122-035153](https://doi.org/10.1146/annurev-marine-032122-035153).

Jarmusch, Alan K., Allegra T. Aron, Daniel Petras, Vanessa V. Phelan, Wout Bittremieux, Deepa D. Acharya, Mohammed M. A. Ahmed, et al. 2022. "A Universal Language for Finding Mass Spectrometry Data Patterns." doi:[10.1101/2022.08.06.503000](https://doi.org/10.1101/2022.08.06.503000).

Johnson, Winifred M., Melissa C. Kido Soule, and Elizabeth B. Kujawinski. 2017. "Extraction Efficiency and Quantification of Dissolved Metabolites in Targeted Marine Metabolomics." *Limnology and Oceanography: Methods* 15 (4): 417–28. doi:[10.1002/lom3.10181](https://doi.org/10.1002/lom3.10181).

Johnson, Winifred M., Melissa C. Kido Soule, Krista Longnecker, Maya P. Bhatia, Steven J. Hallam, Michael W. Lomas, and Elizabeth B. Kujawinski. 2023. "Particulate and Dissolved Metabolite Distributions Along a Latitudinal Transect of the Western Atlantic Ocean." *Limnology and Oceanography* 68 (2): 377–93. doi:[10.1002/lno.12275](https://doi.org/10.1002/lno.12275).

Johnson, Winifred M., Krista Longnecker, Melissa C. Kido Soule, William A. Arnold, Maya P. Bhatia, Steven J. Hallam, Benjamin A. S. Van Mooy, and Elizabeth B. Kujawinski. 2020. "Metabolite Composition of Sinking Particles Differs from Surface Suspended Particles Across a Latitudinal Transect in the South Atlantic." *Limnology and Oceanography* 65 (1): 111–27. doi:[10.1002/lno.11255](https://doi.org/10.1002/lno.11255).

Jones, Catriona L. C., Judith Camps-Castella, Mike Smykala, Morgan S. Sobol, and Keisuke Inomura. 2024. "Seeing Through the Gray Box: An Integrated Approach to Physiological Modeling of Phytoplankton Stoichiometry." *Frontiers in Ecology and Evolution* 12 (December): 1505025. doi:[10.3389/fevo.2024.1505025](https://doi.org/10.3389/fevo.2024.1505025).

Kanehisa, M. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28 (1): 27–30. doi:[10.1093/nar/28.1.27](https://doi.org/10.1093/nar/28.1.27).

Kantz, Edward D., Saumya Tiwari, Jeramie D. Watrous, Susan Cheng, and Mohit Jain. 2019. "Deep Neural Networks for Classification of LC-MS Spectral Peaks." *Analytical Chemistry* 91 (19): 12407–13. doi:[10.1021/acs.analchem.9b02983](https://doi.org/10.1021/acs.analchem.9b02983).

Karl, David M. 1999. "Minireviews: A Sea of Change: Biogeochemical Variability in the North Pacific Subtropical Gyre." *Ecosystems* 2 (3): 181–214. doi:[10.1007/s100219900068](https://doi.org/10.1007/s100219900068).

Karl, David M., and Matthew J. Church. 2017. "Ecosystem Structure and Dynamics in the North Pacific Subtropical Gyre: New Views of an Old Ocean." *Ecosystems* 20 (3): 433–57. doi:[10.1007/s10021-017-0117-0](https://doi.org/10.1007/s10021-017-0117-0).

Karl, David M., Ricardo M. Letelier, Robert R. Bidigare, Karin M. Björkman, Matthew J. Church, John E. Dore, and Angelique E. White. 2021. "Seasonal-to-Decadal Scale Variability in Primary Production and Particulate Matter Export at Station ALOHA." *Progress in Oceanography* 195 (July): 102563. doi:[10.1016/j.pocean.2021.102563](https://doi.org/10.1016/j.pocean.2021.102563).

Karp, Peter D, Richard Billington, Ron Caspi, Carol A Fulcher, Mario Latendresse, Anamika Kothari, Ingrid M Keseler, et al. 2019. "The BioCyc Collection of Microbial Genomes and Metabolic Pathways." *Briefings in Bioinformatics* 20 (4): 1085–93. doi:[10.1093/bib/bbx085](https://doi.org/10.1093/bib/bbx085).

Kido Soule, Melissa C., Krista Longnecker, Winifred M. Johnson, and Elizabeth B. Kujawinski. 2015. “Environmental Metabolomics: Analytical Strategies.” *Marine Chemistry* 177 (December): 374–87. doi:[10.1016/j.marchem.2015.06.029](https://doi.org/10.1016/j.marchem.2015.06.029).

Kind, Tobias, and Oliver Fiehn. 2006. “Metabolomic Database Annotations via Query of Elemental Compositions: Mass Accuracy Is Insufficient Even at Less Than 1 Ppm.” *BMC Bioinformatics* 7 (1): 234. doi:[10.1186/1471-2105-7-234](https://doi.org/10.1186/1471-2105-7-234).

Kuhlish, Constanze, Adva Shemi, Noa Barak-Gavish, Daniella Schatz, and Assaf Vardi. 2023. “Algal Blooms in the Ocean: Hot Spots for Chemically Mediated Microbial Interactions.” *Nature Reviews Microbiology*, October. doi:[10.1038/s41579-023-00975-2](https://doi.org/10.1038/s41579-023-00975-2).

Kujawinski, Elizabeth B., Rogier Braakman, Krista Longnecker, Jamie W. Becker, Sallie W. Chisholm, Keven Dooley, Melissa C. Kido Soule, Gretchen J. Swarr, and Kathryn Halloran. 2023. “Metabolite Diversity Among Representatives of Divergent *Prochlorococcus* Ecotypes.” Edited by Jack A. Gilbert. *mSystems* 8 (5): e01261–22. doi:[10.1128/msystems.01261-22](https://doi.org/10.1128/msystems.01261-22).

Kumler, William, Bryna J. Hazelton, and Anitra E. Ingalls. 2023. “Picky with Peakpicking: Assessing Chromatographic Peak Quality with Simple Metrics in Metabolomics.” *BMC Bioinformatics* 24 (1): 404. doi:[10.1186/s12859-023-05533-4](https://doi.org/10.1186/s12859-023-05533-4).

Kumler, William, and Anitra Ingalls. 2022. “Tidy Data Neatly Resolves Mass-Spectrometry’s Ragged Arrays.” *The R Journal* 14 (3): 193–202. doi:[10.32614/RJ-2022-050](https://doi.org/10.32614/RJ-2022-050).

Li, Fuyan, Andrew Burger, John M. Eppley, Kirsten E. Poff, David M. Karl, and Edward F. DeLong. 2023. “Planktonic Microbial Signatures of Sinking Particle Export in the Open Ocean’s Interior.” *Nature Communications* 14 (1): 7177. doi:[10.1038/s41467-023-42909-9](https://doi.org/10.1038/s41467-023-42909-9).

Libiseller, Gunnar, Michaela Dvorzak, Ulrike Kleb, Edgar Gander, Tobias Eisenberg, Frank Madeo, Steffen Neumann, et al. 2015. “IPO: A Tool for Automated Optimization of XCMS Parameters.” *BMC Bioinformatics* 16 (1): 118. doi:[10.1186/s12859-015-0562-8](https://doi.org/10.1186/s12859-015-0562-8).

Ligrone, Roberto. 2019. “The Great Oxygenation Event.” In *Biological Innovations That Built the World*, 129–54. Cham: Springer International Publishing. doi:[10.1007/978-3-030-16057-9_5](https://doi.org/10.1007/978-3-030-16057-9_5).

Longnecker, Krista, Melissa C. Kido Soule, and Elizabeth B. Kujawinski. 2024. “A Witches’ Brew: Dissolved Metabolites in Seawater.” In *Biogeochemistry of Marine Dissolved Organic Matter*, 69–99. Elsevier. doi:[10.1016/B978-0-443-13858-4.00007-1](https://doi.org/10.1016/B978-0-443-13858-4.00007-1).

Lu, Miaoshan, Shaowei An, Ruimin Wang, Jinyin Wang, and Changbin Yu. 2022. “Aird: A Computation-Oriented Mass Spectrometry Data Format Enables a Higher Compression Ratio and Less Decoding Time.” *BMC Bioinformatics* 23 (1): 35. doi:[10.1186/s12859-021-04490-0](https://doi.org/10.1186/s12859-021-04490-0).

Martens, Lennart, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H. Tang, et al. 2011. “mzML—a Community Standard for Mass Spectrometry Data.” *Molecular & Cellular Proteomics* 10 (1): R110.000133. doi:[10.1074/mcp.R110.000133](https://doi.org/10.1074/mcp.R110.000133).

Martiny, Adam C., Satish Kathuria, and Paul M. Berube. 2009. “Widespread Metabolic Potential for Nitrite and Nitrate Assimilation Among *Prochlorococcus* Ecotypes.” *Proceedings of the National Academy of Sciences* 106 (26): 10787–92. doi:[10.1073/pnas.0902532106](https://doi.org/10.1073/pnas.0902532106).

McAndrew, Pm, Km Björkman, Mj Church, Pj Morris, N Jachowski, leB Williams Pj, and Dm Karl. 2007. “Metabolic Response of Oligotrophic Plankton Communities to Deep Water Nutrient Enrichment.” *Marine Ecology Progress Series* 332 (March): 63–75. doi:[10.3354/meps332063](https://doi.org/10.3354/meps332063).

McGillicuddy, Dennis J. 2016. “Mechanisms of Physical-Biological-Biogeochemical Interaction at the Oceanic Mesoscale.” *Annual Review of Marine Science* 8 (1): 125–59. doi:[10.1146/annurev-marine-010814-015606](https://doi.org/10.1146/annurev-marine-010814-015606).

McGillicuddy, Dennis J., Laurence A. Anderson, Nicholas R. Bates, Thomas Bibby, Ken O. Buesseler, Craig A. Carlson, Cabell S. Davis, et al. 2007. “Eddy/Wind Interactions Stimulate Extraordinary Mid-Ocean Plankton Blooms.” *Science* 316 (5827): 1021–26. doi:[10.1126/science.1136256](https://doi.org/10.1126/science.1136256).

McParland, Erin L., Harriet Alexander, and Winifred M. Johnson. 2021. “The Osmolyte Ties That Bind: Genomic Insights Into Synthesis and Breakdown of Organic Osmolytes in Marine Microbes.” *Frontiers in Marine Science* 8 (July): 689306. doi:[10.3389/fmars.2021.689306](https://doi.org/10.3389/fmars.2021.689306).

Melnikov, Arsenty D., Yuri P. Tsentalovich, and Vadim V. Yanshole. 2020. “Deep Learning for the Precise Peak Detection in High-Resolution LC–MS Data.” *Analytical Chemistry* 92 (1): 588–92. doi:[10.1021/acs.analchem.9b04811](https://doi.org/10.1021/acs.analchem.9b04811).

Mojzeš, Peter, Lu Gao, Tatiana Ismagulova, Jana Pilátová, Šárka Moudříková, Olga Gorelova, Alexei Solovchenko, Ladislav Nedbal, and Anya Salih. 2020. “Guanine, a High-Capacity and Rapid-Turnover Nitrogen Reserve in Microalgal Cells.” *Proceedings of the National Academy of Sciences* 117 (51): 32722–30. doi:[10.1073/pnas.2005460117](https://doi.org/10.1073/pnas.2005460117).

Moore, C. M., M. M. Mills, K. R. Arrigo, I. Berman-Frank, L. Bopp, P. W. Boyd, E. D. Galbraith, et al. 2013. “Processes and Patterns of Oceanic Nutrient Limitation.” *Nature Geoscience* 6 (9): 701–10. doi:[10.1038/ngeo1765](https://doi.org/10.1038/ngeo1765).

Moran, Mary Ann, and Bryndan P. Durham. 2019. “Sulfur Metabolites in the Pelagic Ocean.” *Nature Reviews Microbiology* 17 (11): 665–78. doi:[10.1038/s41579-019-0250-1](https://doi.org/10.1038/s41579-019-0250-1).

Moran, Mary Ann, Frank X. Ferrer-González, He Fu, Brent Nowinski, Malin Olofsson, McKenzie A. Powers, Jeremy E. Schreier, William F. Schroer, Christa B. Smith, and Mario Uchimiya. 2022. “The Ocean’s Labile DOC Supply Chain.” *Limnology and Oceanography* 67 (5): 1007–21. doi:[10.1002/lno.12053](https://doi.org/10.1002/lno.12053).

Moran, Mary Ann, Elizabeth B. Kujawinski, William F. Schroer, Shady A. Amin, Nicholas R. Bates, Erin M. Bertrand, Rogier Braakman, et al. 2022. “Microbial Metabolites in the Marine Carbon Cycle.” *Nature Microbiology* 7 (4): 508–23. doi:[10.1038/s41564-022-01090-3](https://doi.org/10.1038/s41564-022-01090-3).

Moran, Mary Ann, Elizabeth B. Kujawinski, Aron Stubbins, Rob Fatland, Lihini I. Aluwihare, Alison Buchan, Byron C. Crump, et al. 2016. “Deciphering Ocean Carbon in a Changing World.” *Proceedings of the National Academy of Sciences* 113 (12): 3143–51. doi:[10.1073/pnas.1514645113](https://doi.org/10.1073/pnas.1514645113).

Mu, Wanmeng, Tao Zhang, and Bo Jiang. 2015. “An Overview of Biological Production of L-theanine.” *Biotechnology Advances* 33 (3-4): 335–42. doi:[10.1016/j.biotechadv.2015.04.004](https://doi.org/10.1016/j.biotechadv.2015.04.004).

Mulholland, Margaret R., and Michael W. Lomas. 2008. "Nitrogen Uptake and Assimilation." In *Nitrogen in the Marine Environment*, 303–84. Elsevier. doi:[10.1016/B978-0-12-372522-6.00007-4](https://doi.org/10.1016/B978-0-12-372522-6.00007-4).

Müller, Erik, Carolin Huber, Liza-Marie Beckers, Werner Brack, Martin Krauss, and Tobias Schulze. 2020. "A Data Set of 255,000 Randomly Selected and Manually Classified Extracted Ion Chromatograms for Evaluation of Peak Detection Methods." *Metabolites* 10 (4): 162. doi:[10.3390/metabo10040162](https://doi.org/10.3390/metabo10040162).

Muñoz-Marín, M C, G Gómez-Baena, A López-Lozano, J A Moreno-Cabezuelo, J Díez, and J M García-Fernández. 2020. "Mixotrophy in Marine Picocyanobacteria: Use of Organic Compounds by *Prochlorococcus* and *Synechococcus*." *The ISME Journal* 14 (5): 1065–73. doi:[10.1038/s41396-020-0603-9](https://doi.org/10.1038/s41396-020-0603-9).

Muratore, Daniel, Angela K. Boysen, Matthew J. Harke, Kevin W. Becker, John R. Casey, Sacha N. Coesel, Daniel R. Mende, et al. 2022. "Complex Marine Microbial Communities Partition Metabolism of Scarce Resources over the Diel Cycle." *Nature Ecology & Evolution* 6 (2): 218–29. doi:[10.1038/s41559-021-01606-w](https://doi.org/10.1038/s41559-021-01606-w).

Myers, Owen D., Susan J. Sumner, Shuzhao Li, Stephen Barnes, and Xiuxia Du. 2017a. "Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data." *Analytical Chemistry* 89 (17): 8689–95. doi:[10.1021/acs.analchem.7b01069](https://doi.org/10.1021/acs.analchem.7b01069).

———. 2017b. "One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks." *Analytical Chemistry* 89 (17): 8696–8703. doi:[10.1021/acs.analchem.7b00947](https://doi.org/10.1021/acs.analchem.7b00947).

Narainsamy, Kinsley, Sandrine Farci, Emilie Braun, Christophe Junot, Corinne Cassier-Chauvat, and Franck Chauvat. 2016. "Oxidative-Stress Detoxification and Signalling in Cyanobacteria: The Crucial Glutathione Synthesis Pathway Supports the Production of Ergothioneine and Ophthalmate." *Molecular Microbiology* 100 (1): 15–24. doi:[10.1111/mmi.13296](https://doi.org/10.1111/mmi.13296).

Neuwirth, Erich. 2002. "RColorBrewer: ColorBrewer Palettes." Comprehensive R Archive Network. doi:[10.32614/CRAN.package.RColorBrewer](https://doi.org/10.32614/CRAN.package.RColorBrewer).

Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, et al. 2001. "Vegan: Community Ecology Package." Comprehensive R Archive Network. doi:[10.32614/CRAN.package.vegan](https://doi.org/10.32614/CRAN.package.vegan).

Oksanen, Jari, Gavin L Simpson, F Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R Minchin, R B O'Hara, et al. 2022. *Vegan: Community Ecology Package*. Manual.

Olivieri, Alejandro C. 2008. "Analytical Advantages of Multivariate Data Processing. One, Two, Three, Infinity?" *Analytical Chemistry* 80 (15): 5713–20. doi:[10.1021/ac800692c](https://doi.org/10.1021/ac800692c).

Olson, Elise M., Dennis J. McGillicuddy, Glenn R. Flierl, Cabell S. Davis, Sonya T. Dyhrman, and John B. Waterbury. 2015. “Mesoscale Eddies and *Trichodesmium* Spp. Distributions in the Southwestern North Atlantic.” *Journal of Geophysical Research: Oceans* 120 (6): 4129–50. doi:[10.1002/2015JC010728](https://doi.org/10.1002/2015JC010728).

Partensky, Frédéric, and Laurence Garczarek. 2010. “*Prochlorococcus*: Advantages and Limits of Minimalism.” *Annual Review of Marine Science* 2 (1). Annual Reviews: 305–31. doi:[10.1146/annurev-marine-120308-081034](https://doi.org/10.1146/annurev-marine-120308-081034).

Patti, Gary J., Oscar Yanes, and Gary Siuzdak. 2012. “Metabolomics: The Apogee of the Omics Trilogy.” *Nature Reviews Molecular Cell Biology* 13 (4): 263–69. doi:[10.1038/nrm3314](https://doi.org/10.1038/nrm3314).

Pedersen, Thomas Lin. 2019. “Patchwork: The Composer of Plots.” Comprehensive R Archive Network. doi:[10.32614/CRAN.package.patchwork](https://doi.org/10.32614/CRAN.package.patchwork).

Pirttilä, Kristian, David Balgoma, Johannes Rainer, Curt Pettersson, Mikael Hedeland, and Carl Brunius. 2022. “Comprehensive Peak Characterization (CPC) in Untargeted LC–MS Analysis.” *Metabolites* 12 (2): 137. doi:[10.3390/metabo12020137](https://doi.org/10.3390/metabo12020137).

Portlock, T., T. Shama, S. H. Kakon, B. Hartjen, C. Pook, B. C. Wilson, A. Bhuttor, et al. 2025. “Interconnected Pathways Link Faecal Microbiota Plasma Lipids and Brain Activity to Childhood Malnutrition Related Cognition.” *Nature Communications* 16 (1): 473. doi:[10.1038/s41467-024-55798-3](https://doi.org/10.1038/s41467-024-55798-3).

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria: R Foundation for Statistical Computing.

———. 2025. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing.

Raasveldt, Mark, and Hannes Mühleisen. 2019. “DuckDB: An Embeddable Analytical Database.” In *Proceedings of the 2019 International Conference on Management of Data*, 1981–84. Amsterdam Netherlands: ACM. doi:[10.1145/3299869.3320212](https://doi.org/10.1145/3299869.3320212).

Rafiei, Atefeh, and Lekha Sleno. 2015. “Comparison of Peak-Picking Workflows for Untargeted Liquid Chromatography/High-Resolution Mass Spectrometry Metabolomics Data Analysis.” *Rapid Communications in Mass Spectrometry* 29 (1): 119–27. doi:[10.1002/rcm.7094](https://doi.org/10.1002/rcm.7094).

Rainer, Johannes, Andrea Vicini, Liesa Salzer, Jan Stanstrup, Josep M. Badia, Steffen Neumann, Michael A. Stravs, et al. 2022. “A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R.” *Metabolites* 12 (2): 173. doi:[10.3390/metabo12020173](https://doi.org/10.3390/metabo12020173).

Richardson, Tammi L. 2019. “Mechanisms and Pathways of Small-Phytoplankton Export from the Surface Ocean.” *Annual Review of Marine Science* 11 (1): 57–74. doi:[10.1146/annurev-marine-121916-063627](https://doi.org/10.1146/annurev-marine-121916-063627).

Rii, Yoshimi M., Robert R. Bidigare, and Matthew J. Church. 2018. “Differential Responses of Eukaryotic Phytoplankton to Nitrogenous Nutrients in the North Pacific Subtropical Gyre.” *Frontiers in Marine Science* 5 (March): 92. doi:[10.3389/fmars.2018.00092](https://doi.org/10.3389/fmars.2018.00092).

Rii, Yoshimi M., Susan L. Brown, Francesco Nencioli, Victor Kuwahara, Tommy Dickey, David M. Karl, and Robert R. Bidigare. 2008. "The Transient Oasis: Nutrient-phytoplankton Dynamics and Particle Export in Hawaiian Lee Cyclones." *Deep Sea Research Part II: Topical Studies in Oceanography* 55 (10-13): 1275–90. doi:[10.1016/j.dsr2.2008.01.013](https://doi.org/10.1016/j.dsr2.2008.01.013).

Rii, Yoshimi M., Logan M. Peoples, David M. Karl, and Matthew J. Church. 2022. "Seasonality and Episodic Variation in Picoeukaryote Diversity and Structure Reveal Community Resilience to Disturbances in the North Pacific Subtropical Gyre." *Limnology and Oceanography* 67 (S1). doi:[10.1002/lno.11916](https://doi.org/10.1002/lno.11916).

Robinson, David, Alex Hayes, and Simon Couch. 2014. "Broom: Convert Statistical Objects into Tidy Tibbles." Comprehensive R Archive Network. doi:[10.32614/CRAN.package.broom](https://doi.org/10.32614/CRAN.package.broom).

Römpp, Andreas, Thorsten Schramm, Alfons Hester, Ivo Klinkert, Jean-Pierre Both, Ron M. A. Heeren, Markus Stöckli, and Bernhard Spengler. 2011. "imzML: Imaging Mass Spectrometry Markup Language: A Common Data Format for Mass Spectrometry Imaging." In *Data Mining in Proteomics*, edited by Michael Hamacher, Martin Eisenacher, and Christian Stephan, 696:205–24. Totowa, NJ: Humana Press. doi:[10.1007/978-1-60761-987-1_12](https://doi.org/10.1007/978-1-60761-987-1_12).

Röst, Hannes L., George Rosenberger, Pedro Navarro, Ludovic Gillet, Saša M Miladinović, Olga T Schubert, Witold Wolski, et al. 2014. "OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data." *Nature Biotechnology* 32 (3): 219–23. doi:[10.1038/nbt.2841](https://doi.org/10.1038/nbt.2841).

Röst, Hannes L., Uwe Schmitt, Ruedi Aebersold, and Lars Malmström. 2015. "Fast and Efficient XML Data Access for Next-Generation Mass Spectrometry." Edited by Lennart Martens. *PLOS ONE* 10 (4): e0125108. doi:[10.1371/journal.pone.0125108](https://doi.org/10.1371/journal.pone.0125108).

Sacks, Joshua S., Katherine R. Heal, Angela K. Boysen, Laura T. Carlson, and Anitra E. Ingalls. 2022. "Quantification of Dissolved Metabolites in Environmental Samples Through Cation-Exchange Solid-Phase Extraction Paired with Liquid Chromatography–Mass Spectrometry." *Limnology and Oceanography: Methods* 20 (11): 683–700. doi:[10.1002/lom3.10513](https://doi.org/10.1002/lom3.10513).

Salazar, Guillem, and Shinichi Sunagawa. 2017. "Marine Microbial Diversity." *Current Biology* 27 (11): R489–94. doi:[10.1016/j.cub.2017.01.017](https://doi.org/10.1016/j.cub.2017.01.017).

Sañudo-Wilhelmy, Sergio A., Laura Gómez-Consarnau, Christopher Suffridge, and Eric A. Webb. 2014. "The Role of B Vitamins in Marine Biogeochemistry." *Annual Review of Marine Science* 6 (1): 339–67. doi:[10.1146/annurev-marine-120710-100912](https://doi.org/10.1146/annurev-marine-120710-100912).

Schmid, Robin, Steffen Heuckeroth, Ansgar Korf, Aleksandr Smirnov, Owen Myers, Thomas S. Dyrlund, Roman Bushuiev, et al. 2023. "Integrative Analysis of Multimodal Mass Spectrometry Data in MZmine 3." *Nature Biotechnology* 41 (4): 447–49. doi:[10.1038/s41587-023-01690-2](https://doi.org/10.1038/s41587-023-01690-2).

Schrimpe-Rutledge, Alexandra C., Simona G. Codreanu, Stacy D. Sherrod, and John A. McLean. 2016. "Untargeted Metabolomics Strategies—Challenges and Emerging Directions." *Journal of the American Society for Mass Spectrometry* 27 (12): 1897–1905. doi:[10.1007/s13361-016-1469-y](https://doi.org/10.1007/s13361-016-1469-y).

Seelen, Emily, Samantha Gleich, William Kumler, Hanna Anderson, Xiaopeng Bian, Karin Björkman, David Caron, et al. 2024. “A Tale of Two Nutrients: How Nitrogen and Phosphorus Differentially Control Marine Biomass Production and Stoichiometry.” In Review. doi:[10.21203/rs.3.rs-4203527/v1](https://doi.org/10.21203/rs.3.rs-4203527/v1).

Shah, Anuj R., Jennifer Davidson, Matthew E. Monroe, Anoop M. Mayampurath, William F. Danielson, Yan Shi, Aaron C. Robinson, et al. 2010. “An Efficient Data Format for Mass Spectrometry-Based Proteomics.” *Journal of the American Society for Mass Spectrometry* 21 (10): 1784–88. doi:[10.1016/j.jasms.2010.06.014](https://doi.org/10.1016/j.jasms.2010.06.014).

Shilova, I. N., M. M. Mills, J. C. Robidart, K. A. Turk-Kubo, K. M. Björkman, Z. Kolber, I. Rapp, et al. 2017. “Differential Effects of Nitrate, Ammonium, and Urea as N Sources for Microbial Communities in the North Pacific Ocean.” *Limnology and Oceanography* 62 (6): 2550–74. doi:[10.1002/lno.10590](https://doi.org/10.1002/lno.10590).

Siegel, David A., Timothy DeVries, Ivona Cetinić, and Kelsey M. Bisson. 2023. “Quantifying the Ocean’s Biological Pump and Its Carbon Cycle Impacts on Global Scales.” *Annual Review of Marine Science* 15 (1): 329–56. doi:[10.1146/annurev-marine-040722-115226](https://doi.org/10.1146/annurev-marine-040722-115226).

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. 1st ed. Chapman and Hall/CRC. doi:[10.1201/9780429447273](https://doi.org/10.1201/9780429447273).

Smith, Colin A, Grace O’Maille, Elizabeth J Want, Chuan Qin, Sunia A Trauger, Theodore R Brandon, Darlene E Custodio, Ruben Abagyan, and Gary Siuzdak. 2005. “METLIN: A Metabolite Mass Spectral Database.” *Therapeutic Drug Monitoring* 27 (6): 747–51. doi:[10.1097/01.ftd.0000179845.53213.39](https://doi.org/10.1097/01.ftd.0000179845.53213.39).

Smith, Colin A., Elizabeth J. Want, Grace O’Maille, Ruben Abagyan, and Gary Siuzdak. 2006. “XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification.” *Analytical Chemistry* 78 (3): 779–87. doi:[10.1021/ac051437y](https://doi.org/10.1021/ac051437y).

Smith, R., D. Ventura, and J. T. Prince. 2015. “LC-MS Alignment in Theory and Practice: A Comprehensive Algorithmic Review.” *Briefings in Bioinformatics* 16 (1): 104–17. doi:[10.1093/bib/bbt080](https://doi.org/10.1093/bib/bbt080).

Solymos, Peter, and Zygmunt Zawadzki. 2010. “Pbapply: Adding Progress Bar to ‘*Apply’ Functions.” Comprehensive R Archive Network. doi:[10.32614/CRAN.package.pbapply](https://doi.org/10.32614/CRAN.package.pbapply).

Sud, Manish, Eoin Fahy, Dawn Cotter, Kenan Azam, Ilango Vadivelu, Charles Burant, Arthur Edison, et al. 2016. “Metabolomics Workbench: An International Repository for Metabolomics Data and Metadata, Metabolite Standards, Protocols, Tutorials and Training, and Analysis Tools.” *Nucleic Acids Research* 44 (D1): D463–70. doi:[10.1093/nar/gkv1042](https://doi.org/10.1093/nar/gkv1042).

Sweeney, Erin N., Dennis J. McGillicuddy, and Ken O. Buesseler. 2003. “Biogeochemical Impacts Due to Mesoscale Eddy Activity in the Sargasso Sea as Measured at the Bermuda Atlantic Time-Series Study (BATS).” *Deep Sea Research Part II: Topical Studies in Oceanography* 50 (22-26): 3017–39. doi:[10.1016/j.dsr2.2003.07.008](https://doi.org/10.1016/j.dsr2.2003.07.008).

- Tautenhahn, Ralf, Christoph Böttcher, and Steffen Neumann. 2008. "Highly Sensitive Feature Detection for High Resolution LC/MS." *BMC Bioinformatics* 9 (1): 504. doi:[10.1186/1471-2105-9-504](https://doi.org/10.1186/1471-2105-9-504).
- Temple Lang, Duncan. 2000. "XML: Tools for Parsing and Generating XML Within R and S-Plus." Comprehensive R Archive Network. doi:[10.32614/CRAN.package.XML](https://doi.org/10.32614/CRAN.package.XML).
- Thukral, Monica, Andrew E. Allen, and Daniel Petras. 2023. "Progress and Challenges in Exploring Aquatic Microbial Communities Using Non-Targeted Metabolomics." *The ISME Journal* 17 (12): 2147–59. doi:[10.1038/s41396-023-01532-8](https://doi.org/10.1038/s41396-023-01532-8).
- Ting, Ying S., Jarrett D. Egertson, Samuel H. Payne, Sangtae Kim, Brendan MacLean, Lukas Käll, Ruedi Aebersold, Richard D. Smith, William Stafford Noble, and Michael J. MacCoss. 2015. "Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data." *Molecular & Cellular Proteomics* 14 (9): 2301–7. doi:[10.1074/mcp.O114.047035](https://doi.org/10.1074/mcp.O114.047035).
- Treutler, Hendrik, and Steffen Neumann. 2016. "Prediction, Detection, and Validation of Isotope Clusters in Mass Spectrometry Data." *Metabolites* 6 (4): 37. doi:[10.3390/metabo6040037](https://doi.org/10.3390/metabo6040037).
- Tsugawa, Hiroshi, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, and Masanori Arita. 2015. "MS-DIAL: Data-Independent MS/MS Deconvolution for Comprehensive Metabolome Analysis." *Nature Methods* 12 (6): 523–26. doi:[10.1038/nmeth.3393](https://doi.org/10.1038/nmeth.3393).
- Tully, Brett. 2020. "Toffee – a Highly Efficient, Lossless File Format for DIA-MS." *Scientific Reports* 10 (1): 8939. doi:[10.1038/s41598-020-65015-y](https://doi.org/10.1038/s41598-020-65015-y).
- Van Den Brand, Teun. 2021. "Ggh4x: Hacks for 'Ggplot2'." Comprehensive R Archive Network. doi:[10.32614/CRAN.package.ggh4x](https://doi.org/10.32614/CRAN.package.ggh4x).
- Van Mooy, Benjamin A. S., Helen F. Fredricks, Byron E. Pedler, Sonya T. Dyhrman, David M. Karl, Michal Koblížek, Michael W. Lomas, et al. 2009. "Phytoplankton in the Ocean Use Non-Phosphorus Lipids in Response to Phosphorus Scarcity." *Nature* 458 (7234): 69–72. doi:[10.1038/nature07659](https://doi.org/10.1038/nature07659).
- Vardi, Assaf, Fabio Formiggini, Raffaella Casotti, Alessandra De Martino, François Ribalet, Antonio Miralto, and Chris Bowler. 2006. "A Stress Surveillance System Based on Calcium and Nitric Oxide in Marine Diatoms." Edited by Jeffrey Dangel. *PLoS Biology* 4 (3): e60. doi:[10.1371/journal.pbio.0040060](https://doi.org/10.1371/journal.pbio.0040060).
- Vaughan, Davis, and Matt Dancho. 2018. "Furrr: Apply Mapping Functions in Parallel Using Futures." Comprehensive R Archive Network. doi:[10.32614/CRAN.package.furrr](https://doi.org/10.32614/CRAN.package.furrr).
- Villareal, TA. 1994. "Widespread Occurrence Of The Hemiaulus-Cyanobacterial Symbiosis In The Southwest North-Atlantic Ocean Widespread Occurrence Of The Hemiaulus-Cyanobacterial Symbiosis In The Southwest North-Atlantic Ocean Check Uw Holdings Widespread Occurrence Of The Hemiaulus-Cyanobacterial Symbiosis In The Southwest North-Atlantic Ocean." *Bulletin of Marine Science* 54 (1): 1–7.

Vohra, Deepak. 2016. "Apache Parquet." In *Practical Hadoop Ecosystem*, 325–35. Berkeley, CA: Apress. doi:[10.1007/978-1-4842-2199-0_8](https://doi.org/10.1007/978-1-4842-2199-0_8).

Vuong, Quan V, Michael C Bowyer, and Paul D Roach. 2011. "L-Theanine: Properties, Synthesis and Isolation from Tea: L -Theanine from Tea." *Journal of the Science of Food and Agriculture* 91 (11): 1931–39. doi:[10.1002/jsfa.4373](https://doi.org/10.1002/jsfa.4373).

Wada, Naoki, Toshio Sakamoto, and Seiichi Matsugo. 2015. "Mycosporine-Like Amino Acids and Their Derivatives as Natural Antioxidants." *Antioxidants* 4 (3): 603–46. doi:[10.3390/antiox4030603](https://doi.org/10.3390/antiox4030603).

Walker, Mark C, and Wilfred A Van Der Donk. 2016. "The Many Roles of Glutamate in Metabolism." *Journal of Industrial Microbiology and Biotechnology* 43 (2-3): 419–30. doi:[10.1007/s10295-015-1665-y](https://doi.org/10.1007/s10295-015-1665-y).

Wang, Jinyin, Miaoshan Lu, Ruimin Wang, Shaowei An, Cong Xie, and Changbin Yu. 2022. "StackZDPD: A Novel Encoding Scheme for Mass Spectrometry Data Optimized for Speed and Compression Ratio." *Scientific Reports* 12 (1): 5384. doi:[10.1038/s41598-022-09432-1](https://doi.org/10.1038/s41598-022-09432-1).

White, Angelique E., Katie S. Watkins-Brandt, and Matthew J. Church. 2018. "Temporal Variability of Trichodesmium Spp. And Diatom-Diazotroph Assemblages in the North Pacific Subtropical Gyre." *Frontiers in Marine Science* 5 (February): 27. doi:[10.3389/fmars.2018.00027](https://doi.org/10.3389/fmars.2018.00027).

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10). doi:[10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10).

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. doi:[10.21105/joss.01686](https://doi.org/10.21105/joss.01686).

Wickham, Hadley, Jim Hester, and Jeroen Ooms. 2015. "Xml2: Parse XML." Comprehensive R Archive Network. doi:[10.32614/CRAN.package.xml2](https://doi.org/10.32614/CRAN.package.xml2).

Widner, Brittany, Melissa C. Kido Soule, Frank Xavier Ferrer-González, Mary Ann Moran, and Elizabeth B. Kujawinski. 2021. "Quantification of Amine- and Alcohol-Containing Metabolites in Saline Samples Using Pre-extraction Benzoyl Chloride Derivatization and Ultrahigh Performance Liquid Chromatography Tandem Mass Spectrometry (UHPLC MS/MS)." *Analytical Chemistry* 93 (11): 4809–17. doi:[10.1021/acs.analchem.0c03769](https://doi.org/10.1021/acs.analchem.0c03769).

Wilhelm, Mathias, Marc Kirchner, Judith A. J. Steen, and Hanno Steen. 2012. "Mz5: Space- and Time-efficient Storage of Mass Spectrometry Data Sets." *Molecular & Cellular Proteomics* 11 (1): O111.011379. doi:[10.1074/mcp.O111.011379](https://doi.org/10.1074/mcp.O111.011379).

Wilke, Claus O., and Brenton M. Wiernik. 2020. "Ggtext: Improved Text Rendering Support for 'Ggplot2'." Comprehensive R Archive Network. doi:[10.32614/CRAN.package.ggtext](https://doi.org/10.32614/CRAN.package.ggtext).

Wilson, Samuel T., Nicholas J. Hawco, E. Virginia Armbrust, Benedetto Barone, Karin M. Björkman, Angela K. Boysen, Macarena Burgos, et al. 2019. "Kīlauea Lava Fuels Phytoplankton Bloom in the North Pacific Ocean." *Science* 365 (6457): 1040–44. doi:[10.1126/science.aax4767](https://doi.org/10.1126/science.aax4767).

Wishart, David S, AnChi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, et al. 2022. "HMDB 5.0: The Human Metabolome Database for 2022." *Nucleic Acids Research* 50 (D1): D622–31. doi:[10.1093/nar/gkab1062](https://doi.org/10.1093/nar/gkab1062).

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. Second edition. Boca Raton: CRC Press/Taylor & Francis.

Yancey, Paul H. 2005. "Organic Osmolytes as Compatible, Metabolic and Counteracting Cytoprotectants in High Osmolarity and Other Stresses." *Journal of Experimental Biology* 208 (15): 2819–30. doi:[10.1242/jeb.01730](https://doi.org/10.1242/jeb.01730).

Yancey, Paul H., Mary E. Clark, Steven C. Hand, R. David Bowlus, and George N. Somero. 1982. "Living with Water Stress: Evolution of Osmolyte Systems." *Science* 217 (4566): 1214–22. doi:[10.1126/science.7112124](https://doi.org/10.1126/science.7112124).

Yang, Runmin, Jingjing Ma, Shu Zhang, Yu Zheng, Lusheng Wang, and Daming Zhu. 2022. "mzMD: Visualization-Oriented MS Data Storage and Retrieval." Edited by Hanchuan Peng. *Bioinformatics* 38 (8): 2333–40. doi:[10.1093/bioinformatics/btac098](https://doi.org/10.1093/bioinformatics/btac098).

Zehr, Jonathan P., and Raphael M. Kudela. 2011. "Nitrogen Cycle of the Open Ocean: From Genes to Ecosystems." *Annual Review of Marine Science* 3 (1): 197–225. doi:[10.1146/annurev-marine-120709-142819](https://doi.org/10.1146/annurev-marine-120709-142819).

Zimmerman, Amy E., Justin C. Podowski, Gwendolyn E. Gallagher, Maureen L. Coleman, and Jacob R. Waldbauer. 2023. "Tracking Nitrogen Allocation to Proteome Biosynthesis in a Marine Microbial Community." *Nature Microbiology* 8 (3): 498–509. doi:[10.1038/s41564-022-01303-9](https://doi.org/10.1038/s41564-022-01303-9).

Appendix: Speedy Quality Assurance via Lasso Labeling for Untargeted Mass-Spectrometry Data

7.1 Summary

Distinguishing between high and low quality mass features remains a major bottleneck in untargeted mass spectrometry. Here, we present the R package `squallms` that streamlines the process of chromatographic feature annotation by extracting useful metrics from the raw data, labeling groups of similar features interactively, and building a logistic classification model. The package is available via Bioconductor (<https://bioconductor.org/packages/devel/bioc/html/squallms.html>) and interfaces with other mass spectrometry packages there such as XCMS (Colin A. Smith et al. 2006). We expect this functionality to significantly reduce the effort required to explore and curate increasingly common large-scale chromatography-based mass spectrometry datasets.

7.2 Statement of need

Chromatographic mass spectrometry is a powerful way to characterize the molecular composition of chemical and biological samples. The development of data-driven algorithms for untargeted mass spectrometry has massively increased the amount of information obtainable from these datasets, but these algorithms are still plagued by false positive (noise) features (Gika et al. 2019; Pirttilä et al. 2022; Gloaguen et al. 2022). Though novel metrics are actively being developed to quantify the quality of features extracted by untargeted software (Kantz et al. 2019; Kumler et al. 2023), they are difficult to translate across datasets due to differences in instrument setup and sensitivity. Additionally, removing features via sequential thresholding fails to make use of the multivariate advantage and requires the use of thresholds set arbitrarily by the user (Olivieri 2008).

A potential alternative is the use of a multivariate logistic model for the annotation of chromatographic feature quality. However, creating labeled datasets for training such models is often dull and time consuming because it requires the manual review of many individual chromatograms by an expert. Instead, we propose that many similar features can be annotated simultaneously if organized in such a way that an expert can denote an entire group of features as good or bad with a single interaction. These annotations can then be used to train a logistic regression model so that an entire dataset can be quality annotated with confidence quickly and easily (Kantz et al. 2019).

One way to organize features in such a way that similar ones are near to each other is by interpolating a multi-file extracted ion chromatogram to a shared set of retention times and treating each combination of file and retention time as a dimension in a principal component analysis (Figure A.1).

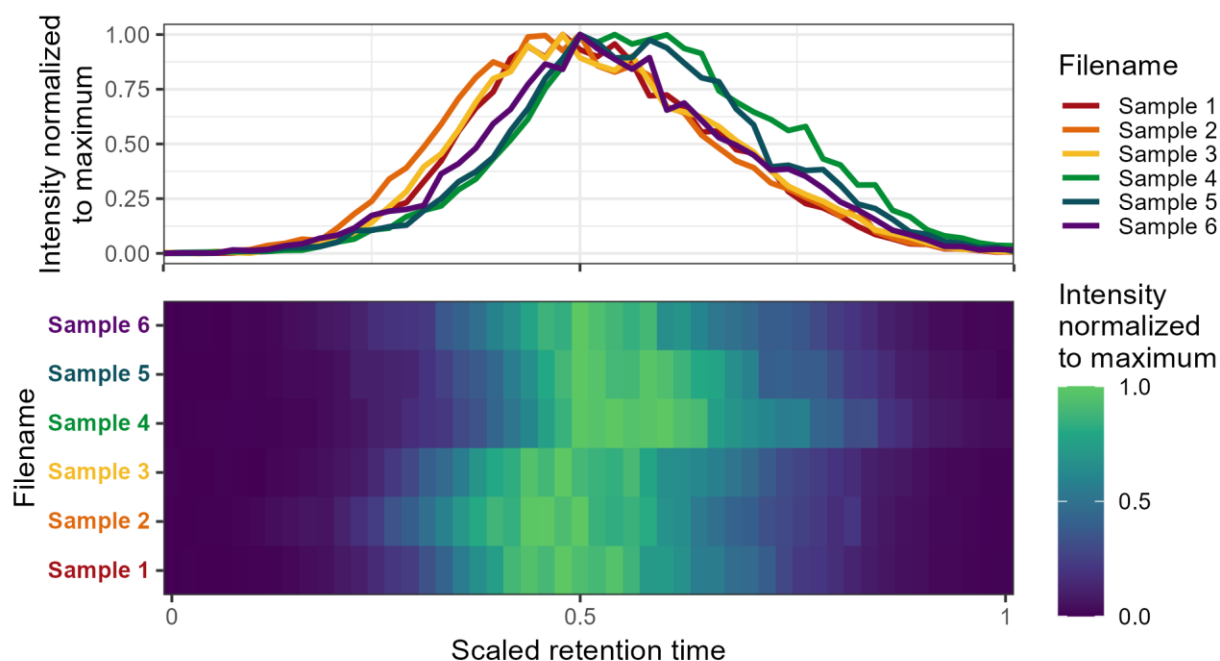


Figure A.1: Coercion of a typical multi-file extracted ion chromatogram (upper plot) to a file-by-time matrix. A single high-quality chromatographic feature is shown in all six samples after

normalization both in “side view” with intensity on the y axis (upper plot) and “top down” with intensity corresponding to fill color (lower plot).

Features that are visually similar tend to cluster tightly in the first few principal components while noise features scatter randomly near the center of this space (Figure A.2).

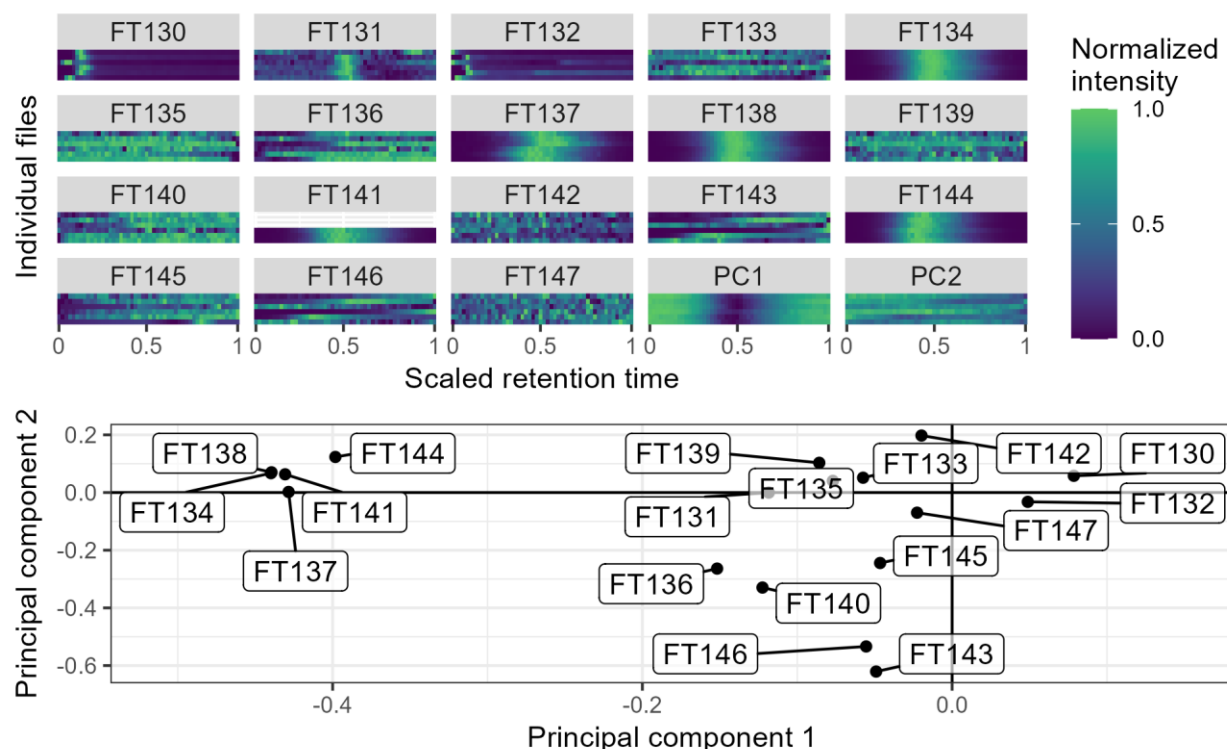


Figure A.2: Subset of chromatographic features shown both individually in top down view and in principal component space. The upper plot shows eighteen chromatographic features and their associated ID beginning with the letters FT. Features 134, 137, 138, 141, and 144 all appear to be high quality and look very similar with a bright green central stripe and dark purple edges. The two final small multiples (PC1 and PC2) show the first and second principal components for comparison. The lower plot shows how the eighteen features fall in principal component space, with the five high-quality features noted above clustering to the left side of the plot, separately from the other features.

However, the location of this high-quality cluster in principal component space is not easily predictable because feature quality is not necessarily the largest source of variance. This necessitated the development of an interactive feature selection method that we have implemented in the squallms package using R’s shiny and plotly packages (Chang et al. 2024;

Sievert 2020). The main interface is shown below in Figure A.3 for all features, not just the subset used in the demo above.

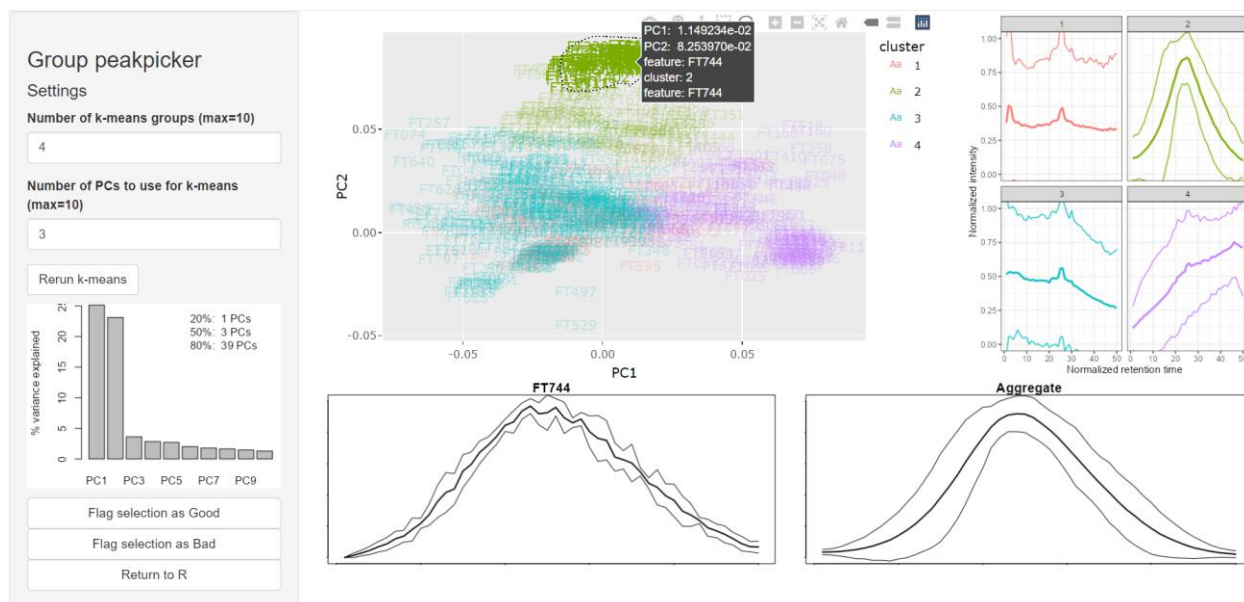


Figure A.3: Interactive group labeling dashboard of *squallms*. The central plot is an interactive version of the principal component plot shown in . Upon mouse hover, individual features are rendered in the lower left panel for rapid overview and detection of high quality regions. Clicking and dragging on this central plot will begin selecting features that can be labeled as either “Good” or “Bad” using the buttons at the bottom of the settings panel at the far left. Selected features are also shown in aggregate in the lower right panel as the median normalized intensity plus its interquartile range. Colors in this central plot correspond to clusters assigned via *k*-means algorithm and provide a way to use information from principal components beyond the first two, as clusters can be hidden by double-clicking on the associated legend entry. Multiple selections can be made for both good and bad annotations. Once the user is satisfied with the labeling they’ve done, they can either close the window or click the “Return to R” button in the settings panel to create an R object in the existing session containing the labels and their associated feature ID.

The package also contains a manual training tool for rapid single feature annotation by binding quality assessments to specific keyboard presses for ease of annotation. Finally, the package includes a basic logistic regression model framework designed to accept the feature labels and feature quality metrics which produces a best-fit estimate of each feature’s quality as not all features are typically labeled using either the aggregate or the manual methods.

7.3 Acknowledgements

This work was supported by the University of Washington eScience Institute through their Data Science Incubator program. We would like to thank Bryna Hazelton and Valentina Staneva for thoughtful discussion on algorithm design and package development. Funding was provided by the Simons Foundation (SCOPE Award ID 329108 to AEI, SF Award ID 385428 to AEI).