Insights from Automated and Untargeted Marine Microbial Metabolomics

William Kumler

2025-05-26

# Acknowledgements

* Lab
  + Anitra
  + Laura
  + Josh
  + Susan
  + Angie/Katherine/Frank
  + Regina
  + Natalie/Iris/Claudia/Raisha/Bryn
  + Leland/Lindsey/Everetta/Raafay/Natalie/Anna/Amy
* Broader Ocean/UW
  + Committee
  + Armbrust lab
  + Cohort
  + Bowling team
  + OCEAN front office
  + OSB cleaning staff
  + eScience (Dave and Bryna esp.)
  + Fisheries friends (SEAS+Helena/Zoe)
* Research things
  + Funding (ARCS/SCOPE)
  + SCOPE folks
    - Dave K/Sonya/Ed/Angel/Matt/Ben/Daniel
    - SCOPE Ops
    - Captain and crew of the KM
    - PERISCOPE team
  + R packages xcms/data.table/tidyverse/knitr/plotly/shiny/vegan/broom/arrow/duckdb/xml2/XML/Rcolorbrewer/viridis/ggh4x/ggtext/patchwork/pbapply
    - RStudio/Posit, R Markdown
  + Sisweb/Chemcalc/Metlin/HMDB/GNPS/Metabolights/Metabolomics Workbench
* Friends and family
  + Not all my time was spent on boats, in the lab, or the office
    - Stone Gardens/Edgeworks/Seattle Public Libraries
    - Land acknowledgement
  + Past mentors
    - Bethanie, Jim, Mimi, Pete
  + Kaylyn/Hannah/Shawnee/Wave/Kate/Sarra/Rachel Liu/Rachel Lai/Tasha/Sylvia/Colby/Allison/Este/Cassie?/Carly?/Tori?
  + Mom/Dad/Ben/Maggie/Gmas
* And you, dear reader?

# Chapter 1: Introduction

## 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 dioxide into biomass and has performed this continuously for the last two billion years [@ref]. Most of this is performed by single-celled organisms too small to see with the naked eye known as microbes [@ref]. The process by which they transform air 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 millenia in marine sediments [@ref].

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. 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. Determining the molecular composition of marine carbon and its fluxes through the environment is therefore paramount in improving our ability to accurately model the marine ecosystem [@Jones2024].

## 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 (<600 Dalton) organic molecules that act as currencies within the cell while excludes 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 storage[@ref], antioxidation[@ref], osmotic balance[@Yancey], buoyancy[@Yancey], and cell signaling (both beneficial and antagonistic interactions, @Thukral2023). There are likely hundreds of thousands of individual molecules composing the metabolome in the environment, making their comprehensive analysis challenging [@Schrimpe-Rutledge2016].

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, @Edwards2023). 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 [@Bauermeister2022; @KEGG; @Biocyc].

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 [@KidoSoule2015; @Cajka2016]. 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-Rutledge2016]. Third, the diversity and novelty of many compounds makes pure standards often unavailable, let alone isotopically-labeled equivalents necessary for the construction of a multi-point calibration curve [@Cajka2016].

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 [@Heal2021; @Sacks2022; @Longnecker2024]. 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 105 to 10^10 [@Longnecker2024]. In contrast to other metabolomics specialties where the organism of interest is well studied and genetically documented, environmental metabolomics struggles with a lack of clear genetic representation where less than 5% of the genetic diversity in the ocean has been captured by reference genomes [@DeLong2005; @Salazar2017]. 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 [@Heal2021; @Durham2022; @Kujawinski2023]. 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 [@Karl2017].

## Automated and untargeted liquid-chromatography mass spectrometry

Mass spectrometry (MS) is the dominant analytical platform in metabolomics [@Cajka2016]. 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 method 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, @Buszewski2012) and compounds near the limit of detection, this becomes a challenge. The conventional solution is manual integration, in which a mass-spectrometrist 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 integration.

Automatic peakpicking and annotation software has been developed in parallel for the better part of two decades by both open-source and commercial endeavours [@xcms; @mzmine; @MSDIAL; @CompoundDiscoverer; @comparisonrefs]. 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 problem with a list of known compounds in advance. 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. While untargeted analysis is traditionally associated with hypothesis generation because of its compound discovery capability [@Giera2023; @Thukral2023], 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 (@ref) or even Microsoft Excel as long as the data itself is used to drive discovery instead of a priori knowledge about the expected compounds. For example, one could imagine an ADAP-style algorithm [@Myers2017a] that recursively extracts the largest intensities in a file 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 and documented sufficiently.

## 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 @Wickham2014. 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.

Chapter 4 demonstrates the utility of allowing for rapid raw data access by showing now 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. Chapter 4.5 extends this further by showing how the raw data itself can be treated as a multidimensional array with the largest “signals” being that 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. Chapter 5 documents an exploratory metabolomics dataset collected 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 with several compounds detected for the first time in the open ocean and required the use of complex ecological statistics to unravel the impacts of sea level anomaly on the ocean’s metabolome. 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 each metabolite’s role in ocean biogeochemistry.

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

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

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

# Chapter 4.5: Speedy Quality Assurance via Lasso Labeling for Untargeted Mass-Spectrometry Data

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

# Chapter 6: The Form of Nitrogen Determines its Fate in the North Pacific Subtropical Gyre

# Conclusions?

# Bibliography