



OPEN ACCESS

EDITED BY

Edgardo Donati,
National University of La Plata, Argentina

REVIEWED BY

Zhenyan Zhang,
Zhejiang University of Technology, China
Theodor Sperlea,
Leibniz Institute for Baltic Sea Research
(LG), Germany

*CORRESPONDENCE

Emelia Chamberlain
✉ emelia.chamberlain@whoi.edu
Heather Kim
✉ hkim@whoi.edu

RECEIVED 12 January 2026

REVISED 29 March 2026

ACCEPTED 03 April 2026

PUBLISHED 18 May 2026

CITATION

Chamberlain EJ, Boulton W, Connors E,
Calianos T, Bowman JS, Creamean JM,
Mock T and Kim HH (2026) From
microbial diversity to functional
potential using dimensionality
reduction.
Front. Microbiol. 17:1786397.
doi: 10.3389/fmicb.2026.1786397

COPYRIGHT

© 2026 Chamberlain, Boulton, Connors,
Calianos, Bowman, Creamean, Mock and
Kim. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original author(s)
and the copyright owner(s) are credited
and that the original publication in this
journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

From microbial diversity to functional potential using dimensionality reduction

Emelia J. Chamberlain^{1*}, William Boulton²,
Elizabeth Connors¹, Theodore Calianos¹, Jeff S. Bowman³,
Jessie M. Creamean⁴, Thomas Mock⁵ and Heather H. Kim^{1*}

¹Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, United States, ²School of Computing Sciences, University of East Anglia, East Anglia, United Kingdom, ³Scripps Institution of Oceanography, University of California, San Diego, San Diego, CA, United States, ⁴Department of Atmospheric Science, Colorado State University, Fort Collins, CO, United States, ⁵School of Environmental Sciences, University of East Anglia, East Anglia, United Kingdom

The high dimensionality of microbial diversity data from 'omics observations can be reduced using Machine Learning, with many recent studies showcasing ML utility for exploratory ecological feature finding and process prediction. Here, we compare the Self Organizing Map (SOM) dimensionality reduction method to the well-documented sample-based Principal Coordinate Analysis (PCoA) and taxa-based Weighted Gene Correlation Network Analysis (WGCNA) using near daily 16S rRNA gene amplicon sequencing data from the 2019 to 2020 MOSAiC International Arctic Drift Expedition. We then map k-means clustering outputs from each method to available metagenomes, extracting functionally distinct seasonal microbial ecotypes in the surface Arctic Ocean. Our results indicate the SOM method better represented expected seasonal transitions and identified a greater number of metabolically distinct functional groups than the more traditional PCoA ordination. Ultimately, we identified four community ecotypes with distinct taxonomic and functional cut-offs driven by seasonality, water mass, and substrate turnover, highlighting the importance of succession in functional diversity for the central Arctic Ocean. These results reinforce ML dimensionality reduction as a meaningful translator in the mining of historical amplicon datasets to address modern mechanistic questions and potentially provide 'omics informed ecotype diversity to leverage in mechanistic biogeochemical models.

KEYWORDS

Arctic Ocean, bacteria, ecosystem function, machine learning, microbial diversity

1 Introduction

Microorganisms represent the most abundant and diverse form of life on Earth (Spang and Ettema, 2016). The tremendous taxonomic and functional diversity of microbial communities serves as a fundamental driver of ecosystem functions, marine carbon and biogeochemical cycles, and trophic energy transfer, with different species and functional groups contributing uniquely to these processes (Henson et al., 2011). Among these microbes, marine heterotrophic bacteria (hereafter bacteria) and archaea are particularly important, processing vast quantities of carbon through their complex metabolic networks. Bacteria alone respire up to 85–90% of the 50 petagrams of carbon fixed annually by phytoplankton (Field et al., 1998; Giering et al., 2014), an amount equivalent to half of global photosynthesis, making them essential players in regulating whether this carbon returns to the upper ocean-atmosphere boundary as CO₂ or sinks to the deep ocean for centuries to millennia (Devries et al., 2012; Honjo et al., 2014; Boyd et al., 2019). However, the role diversity plays in regulating this process is not well constrained and many bacterial growth traits can vary widely among microbial groups (Li et al., 2025), with the most abundant taxa contributing little to overall rate processes (Munson-McGee et al., 2022).

Over recent decades, advances in sequencing technologies have enhanced our ability to profile bacterial communities, yielding vast, ‘omics datasets (Mock et al., 2022). As the importance of microbial contributions to climate processes becomes increasingly recognized, so too does the need to translate these complex community patterns into functional units that can inform predictive biogeochemical models (Hellweger, 2020). Machine learning (ML) has emerged as a powerful approach for reducing dimensionality across complex microbial datasets and identifying ecologically and environmentally relevant features. While all ‘omics data are rich in biological information, high-dimensional amplicon sequence variant (ASV) data are particularly suitable for ML techniques (McElhinney et al., 2022). Recent studies have highlighted this, using both supervised and unsupervised ML methodologies to distill the complexity of microbial community structure into trait-based groupings, i.e., “ecotypes” linked to specific ecosystem functions (Bowman et al., 2017) and extract the best taxonomic predictors for the key biogeochemical rate processes involved in bacterial growth efficiency, such as bacterial production (Connors et al., 2024) and oxygen utilization (Chamberlain et al., 2025).

Highlighting the field’s excitement around these methods, a flurry of recent reviews have explored why and how to apply this growing toolkit for microbial ecologists (Asnicar et al., 2023). For example, Qu et al. (2019) demonstrated how supervised classifiers can predict host phenotypes from microbiome profiles, highlighting the potential of ML to capture complex ecological patterns. Topçuoğlu et al. (2020) introduced an interpretable and reproducible ML pipeline tailored to microbiome data, addressing barriers around accessibility and transparency in ecological applications. Bowman (2021) highlighted the predictive capacity of microbial ML applications, while Hernández Medina et al. (2022) emphasized integrating environmental metadata and outlined modeling-specific considerations for microbiome data. While ML

shows clear utility for microbial trait extraction, mechanistic connections between microbial community composition and ecosystem-level processes remain underexplored, despite evidence suggesting their value for quantifying ecosystem dynamics through biogeochemical modeling (Li et al., 2025). This is largely because the translation of microbial community-level patterns into functional representations suitable for, e.g., ocean biogeochemical models, presents significant methodological and conceptual challenges. As computational approaches increasingly interface with these predictive ocean models, understanding when and which ML frameworks can be effectively used to translate microbial community composition into the functional units needed for such modeling frameworks becomes increasingly relevant.

In this study, we compare and evaluate three dimensionality reduction approaches compatible with k-means clustering as a strategy to navigate the high-dimensionality challenge of bacterial data: (1) Principal Coordinate Analysis (PCoA) (Anderson and Willis, 2003), (2) Self-Organizing Maps (SOMs) (Kohonen, 2013), and (3) Weighted Gene Correlation Network Analysis (WGCNA) (Zhang and Horvath, 2005). We first employ such dimensionality reduction as k-means clustering directly on community data has the potential to impose artificial boundaries on continuous ecological gradients. These specific methodological representatives were chosen because they are well established in the literature and highly interpretable (not black-box prediction), while still providing a gradient of analytical complexity, capturing different levels of ecological structure. While all three approaches account for the underlying structure embedded in high-dimensional microbial datasets, they differ significantly in complexity, assumptions, outputs, and interpretability. Each method offers distinct advantages: PCoA provides a traditional statistical ordination baseline that summarizes dominant variation among samples by projecting it into a lower-dimensional space (Legendre and Legendre, 2012). SOMs provide a ML and non-linear alternative, training an unsupervised neural network via competitive learning algorithm to identify emergent groupings (Kohonen, 2013; Oja et al., 2003). WGCNA resolves taxon-level modules by constructing statistical weighted correlation networks to identify discrete modules of co-varying taxa, offering a more taxa-driven framework suitable for hypothesis-driven ecological analysis (Zhang and Horvath, 2005).

We test these approaches using a multi-seasonal bacterial dataset from the surface waters of the central Arctic Ocean, in a region where climate warming is occurring at almost four times the global average rate (Rantanen et al., 2022). This ocean represents an ideal testbed for comparing dimensionality reduction methods due to its extreme seasonal bacterial shifts (Priest et al., 2025) and the critical roles these communities play in carbon cycling processes that could significantly impact Arctic amplification (Edwards et al., 2020). The rapid environmental transitions in the Arctic have the potential to release previously sequestered carbon (Schuur et al., 2015), while fundamentally altering bacterial community structure and function (Smith et al., 2023). Identifying which translational approach best captures these complex ecological patterns in high-dimensional Arctic bacterial datasets is therefore essential for improving future model predictions of regional and global climate responses mediated by microbial feedbacks.

2 Materials and methods

2.1 Sample selection and collection

The Multidisciplinary drifting Observatory for the Study of Arctic Climate Expedition (MOSAiC; 2019–2020) was an interdisciplinary year-long drift experiment in the central Arctic Ocean that created a wealth of ecological (Fong et al., 2024) and genomic (Mock et al., 2022) data which we have leveraged for this study. From the publicly available MOSAiC 16S rRNA gene amplicon sequencing dataset (Chamberlain and Bowman, 2022), a subset of 215 suitable surface seawater samples were selected and assessed. These were collected as close to daily as possible (average sampling time 15:00 UTC) between 10/29/2019 and 09/18/2020 over the course of the MOSAiC drift, covering a 10° latitudinal gradient 89.1°–79.1° across the Amundsen and Nansen Basins (Fong et al., 2024). Data gaps exist where project personnel were not on board (12/13/2019–02/20/2020) and during periods of ship relocation (07/31/2020–08/22/2020). Samples were collected from the underway seawater system (RV Polarstern; 11 m inlet), with highly comparable community structure results to samples from similar depths collected using a CTD Rosette (Chamberlain et al., 2025). Additionally, 18 of the underway sampling dates matched with whole metagenome data collected at either 10 or 11 m depth from the CTD rosette (Mock et al., 2022). Ancillary environmental data (surface temperature and salinity) were collected continuously from the underway system and corrected as described in Schulz et al. (2024). Chlorophyll-a concentrations were collected and filtered at the same time as collections for microbial community structure and are described by Hoppe et al. (2023).

2.2 DNA sequencing and metabolic function

Sample and data processing for water column bacterial and archaeal community structure (16S rRNA gene amplicon sequencing) is described in detail by Chamberlain et al. (2025). Briefly, 1 L of underway seawater was filtered through a 25 mm 0.2 µm Pall Corporation Supor membrane filter and frozen (–80 °C). DNA was extracted using a ThermoFisher Scientific KingfisherTM Flex MagMax Microbiome Ultra Nucleic Acid extraction kit and sequencing for the amplified V4 region of 16S rRNA gene took place at Argonne National Laboratory on the Illumina MiSeq platform (universal primers 515F and 806R; Walters et al., 2016). Illumina reads were cleaned using dada2 (Callahan et al., 2016) and classified using PATHWAY PRediction by phylogenetic placement (paprica v0.7.0; Bowman and Ducklow, 2015) with a database derived reference tree (Genbank RefSeq; Haft et al., 2018). Point of placement was additionally used to perform metabolic inference estimating genetic characteristics, including genome size, 16S rRNA gene copy number, and GC content (Erazo et al., 2021) and doubling time (gRodon; Weissman et al., 2021). Note, doubling times from codon usage reflect theoretical growth rates, and neglect environmental conditions, serving only as a theoretical maximum to compare between samples. Additionally, taxonomic associations from paprica represent an ASV's closest relative among the published Genbank RefSeq completed genomes and don't

always indicate the presence of an exact strain. Estimated 16S rRNA gene copy number was used to normalize read counts prior to calculations of relative abundance. Singleton ASVs were not included, and data was normalized using the Hellinger Transformation (performed in “vegan”; Oksanen et al., 2015), making it more suitable for downstream clustering analyses with assumptions of Euclidean distances (Legendre and Gallagher, 2001).

The extraction of DNA from the 18 metagenomes is the same as described in Boulton et al. (2025) (Boulton et al., 2025). Seawater was filtered through 0.22 µm Sterivex filters, and DNA was extracted using the Qiagen PowerWater DNA kit (Qiagen N.V., Hilden, Germany), following the Qiagen DNeasy Power Water SOP v1. Samples were sequenced on an Illumina NovaSeq S4 device, with 151 base pair paired-end reads, using the Illumina regular concentration library protocol. Metagenomes were assembled, annotated, and reads mapped by the JGI Metagenome Annotation Pipeline (Clum et al., 2021). To assess potential metabolic function within the 18 metagenomes, we calculated the relative abundance of COG categories (Tatusov et al., 2000), based on transcripts per million (tpm). The tpm abundance of each COG category was then calculated based on the normalized combined coverage of each CDS containing a COG annotation.

2.3 Machine learning and statistical analysis

All analyses were conducted in R (R Core Team, 2023). The PCoA was constructed using the “vegan” package (Oksanen et al., 2015) to provide a low-dimensional projection of samples based on Euclidean distances which was separated into distinct clusters using the unsupervised ML k-means algorithm. The final selection of k was determined using both the elbow method, or inflection in within-clusters sum of squares (WSS), and the Silhouette method. While the elbow method identifies the point of diminishing returns in terms of cluster compactness, silhouette scores also account for between-cluster separation. Experimentally testing k values around the perceived optimums (experimental process described in Supplementary Text 1), we compared how well clusters separated in the first two ordination axes and used a post-hoc Permutational Multivariate Analysis of Variance (PERMANOVA) test to confirm that clusters were statistically significant.

The unsupervised SOM model was constructed using the “kohonen” package (Wehrens and Kruisselbrink, 2018). Grid parameters were experimentally selected, while aiming for an optimal distribution of samples assigned to each map unit (Supplementary Figure 1). The final map was constructed using a 4 × 4 toroidal grid with hexagonal map units and a decay scaling alpha from 0.05 to 0.01. Unit assignment was based on Euclidean distances and clustering also took place via k-means. This is a similar method to what was used on a full water column MOSAiC dataset in Chamberlain et al. (2025) but with a unique SOM construction and sensitivity due to the differing resolution of input data (215 surface underway vs. 693 samples underway and CTD samples, 0–4,000 m). Resulting differences are described in detail in Supplementary Figure 2. Ultimately, through the combination of dimension reduction (PCoA or SOM) and k-means clustering, each sample was assigned to unique community compositional clusters

which we were then statistically compared to metabolic inference and environmental parameters using Kruskal Wallis and *post hoc* Wilcoxon signed-rank tests (rstatix package; Kassambara, 2023).

The WGCNA analysis was constructed using the “WGCNA” package (Langfelder and Horvath, 2008) and a hybrid network construction through the application of k-means clustering instead of the standard hierarchical clustering and dynamic tree cut, inspired by the framework described in Botía et al. (2017). A soft-thresholding power of 8 was chosen based on the scale-free topology fit index ($R^2 > 0.8$) and mean connectivity profiles. An adjacency matrix was computed using a signed network derived from Pearson correlations, followed by transformation into a topological overlap matrix to capture shared patterns of co-abundance among ASVs. To detect modules of co-varying ASVs, we applied k-means clustering to the TOM-based dissimilarity matrix. Following module detection, each module was summarized by its eigengene—the first principal component of the ASV profiles within the module. These module eigengenes were then correlated with functional metadata (estimated 16S rRNA gene copy number, doubling time, genome size, GC content, and total chlorophyll-a concentration) and environmental parameters (temperature, salinity, and location) to identify ecologically relevant modules. Module-trait relationships were visualized in a heatmap with both correlation coefficients and *p*-values. Similar modules were merged using a correlation-based threshold of 0.7 to reduce redundancy. To assess the importance of individual ASVs within each module, we calculated module membership scores (kME), defined as the Pearson correlation between each ASV’s abundance profile and the module eigengene. These scores were used to rank ASVs by their centrality within a module, allowing identification of the taxa with the strongest association to overall module structure and likely ecological influence.

To provide further comparison across methods using alternative methods from those used in clustering, a non-metric multidimensional scaling (NMDS) ordination was constructed from the Bray–Curtis dissimilarity matrix of Hellinger transformed ASV relative abundances (vegan package; Oksanen et al., 2015) and correlated with basic environmental parameters representative of season and location.

3 Results

The underway surface water amplicon dataset contained a total of 5,135 unique ASVs across 215 samples. The PCoA and SOM models, which are robust to zeros and outliers, used all samples and sequences for model construction. The WGCNA correlation network approach, which is sensitive to outliers and rare abundances, has a standard data QC and selection procedure applied during analysis (Langfelder and Horvath, 2008). This removed 3 of the original samples and retained only the most abundant 1,322 sequences for clustering. Ultimately, K-means clustering yielded 3 PCoA (Supplementary Figure 3), 5 SOMs (Supplementary Figures 1, 3), and 4 WGCNA (Supplementary Figure 4) unique microbial community clusters respectively.

All clustering methods displayed evidence of seasonal community succession, but there was substantial variance between the temporal cut-offs for each methodology (Figure 1). Projecting

the underlying data structure of both the samples (Figure 1D) and individual ASVs (Figure 1E) using a comparative projection method (NMDS ordination), separation in the ordination space was most closely aligned with the SOM segmentation (samples) and WGCNA modules (ASVs). Numerical month ($r^2 = 0.56$), temperature ($r^2 = 0.66$), salinity ($r^2 = 0.70$), latitude ($r^2 = 0.65$), and longitude ($r^2 = 0.79$) were all significant explanatory variables ($p < 0.001$) in the NMDS ordination (Figures 1D,E). The distribution of values for these associated environmental variables and estimated genetic traits are presented in the supplemental material for each PCoA (Supplementary Figure 5) and SOM (Supplementary Figure 6) cluster. Pearson correlation coefficients indicate a significant relationship with all environmental variables and genetic traits ($p < 0.05$) except for the brown module which was not significantly correlated with genome size ($p = 0.08$) turquoise module which was not significantly correlated with Temperature ($p = 0.6$), Numeric month ($p = 0.5$), and % GC content ($p = 0.8$) (Supplementary Figure 7).

Seasonal cutoffs between most clusters were reinforced by clear taxonomic and genetic shifts. SOM5, for example, is present in both late Fall 2019 and early Fall 2020 and was enriched in ASVs assigned to taxa in the Brown module, including some of the top module members (e.g., *Planktomarina temperata*, and an unidentified *Flavobacterium* ASV; Figure 2B). Most of these top taxa represent copiotrophs associated with the rapid colonization and degradation of organic material (Dithugoe et al., 2023). SOM5 samples were additionally enriched in carbohydrate, amino acid, and lipid transport genes (Figure 3). The PCoA analysis however, distinguished two clusters during the fall periods (PCoA2 in Fall 2019 and PCoA3 in late Summer/early Fall 2020), with significant environmental drivers of water mass salinity and temperature (Figure 1D). The primary taxonomic difference between these clusters were the additional enrichment in ASVs assigned to *Bermanella marisrubri*, *Owenweeksia hongkongensis*, and *Polaribacter* sp. (Figure 2A) and cell motility/translational genes (Figure 3) in PCoA3. However, these features were also prominent in the additional SOM cluster of SOM4 which is only present in July 2020 (Figure 1) and also marked by significantly lower mean doubling times (Figure 3B), higher mean 16S copy numbers, and higher mean genome sizes (Supplementary Figure 6). While the July 1 shift between PCoA2 and 3 coincides with the onset of Brown Module dominance, the taxa with both high kME and significant negative correlation to mean doubling time (Figure 4) align exactly with the topmost abundant taxa in SOM4. The Brown Module then reaches its abundance peak with the transition to SOM 5.

During early summer (June 2020), the SOM (SOM1) and WGCNA (Turquoise Module) clustering methods capture additional community complexity which was not captured in the PCoA (Figure 1). Despite being the module with the most ubiquitous background taxa (Figure 1E), the Turquoise module was significantly ($p < 0.001$) negatively correlated with mean doubling time, with its top kME and correlated taxa strongly related to phytoplankton groups (e.g., *Micromonas* mitochondrial DNA) or associated heterotrophs known for exploiting sudden pulses of organic matter (e.g., *Polaribacter*, *Lutibacter*). The Turquoise module was also the most significantly correlated with chlorophyll-a concentration ($p < 0.001$; Supplementary Figure 7) and had several high kME ASVs assigned to either

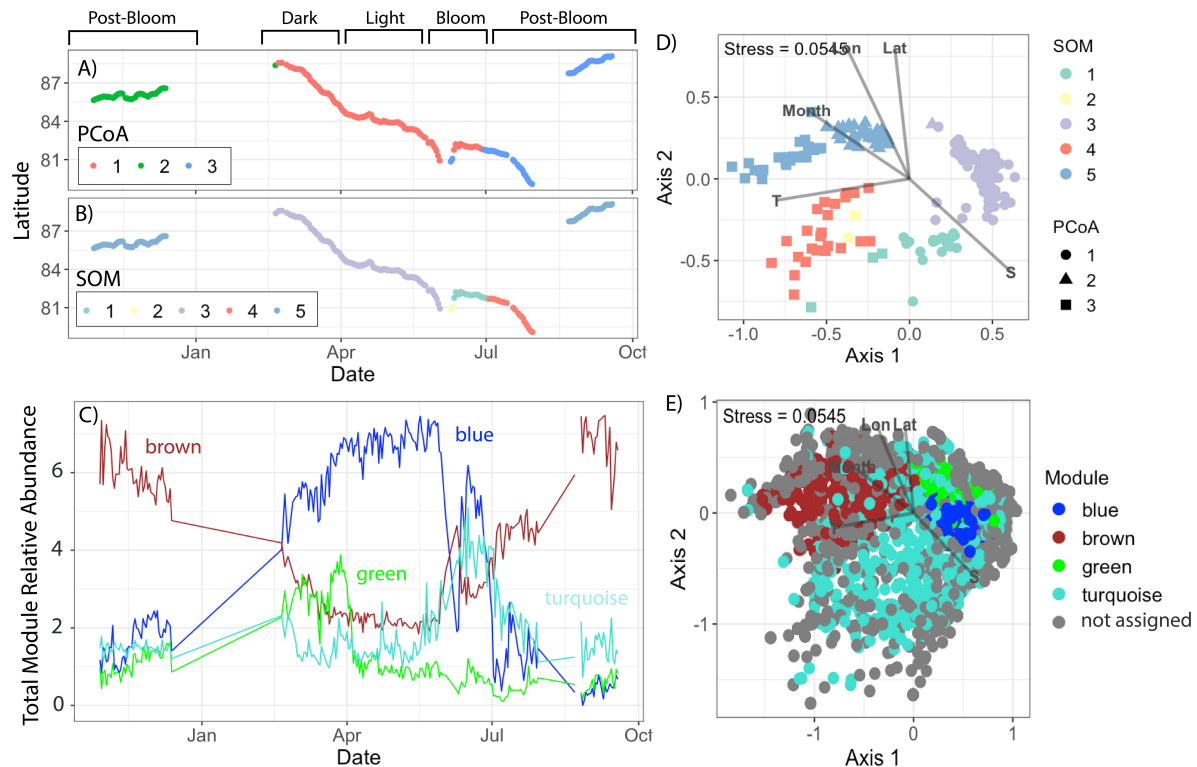


FIGURE 1

Clustering output and spatio-temporal distributions during MOSAiC. PCoA (A) and SOM (B) clusters identified distinct seasonal shifts in bacterial community structure throughout the MOSAiC Drift, here visualized as latitude. In (C), the relative abundances of all ASVs identified to each WGCNA module were summed for each sampling date. Environmental separation among clusters for latitude (Lat), longitude (Lon), numerical month, temperature (T), and salinity (S) when samples (D) and ASVs (E) were visualized in the first two axes of an NMDS ordination created from the Bray dissimilarities of Hellinger transformed 16S ASV relative abundances. The NMDS stress value, indicating ordination fit, is listed in the top left-hand corner. In (D) samples are colored by SOM and shaped by PCoA cluster assignment. In (E) ASVs are colored by module assignment. Gray points represent ASVs with total relative abundances too low to be included in the WGCNA analysis.

phycrophilic biofilm producers or algal associates/pathogens. These functional changes in genetic information were captured with in the segmentation of SOM1, which had high chlorophyll-*a* concentrations, the second lowest mean doubling times, second highest mean genome size, and a low 16S copy number (Supplementary Figure 6). However, this SOM cluster was also composed of highly abundant ASVs assigned across the Blue, Brown, and Turquoise modules, which were all relatively abundant during this seasonal transition as well (Figure 1). While there was only one matching metagenome sample for this period, it was enriched in genes related to the mobilome, cell division, and transcription (Figure 3). This resulted in a unique bloom-associated community which quickly grew in abundance during this time.

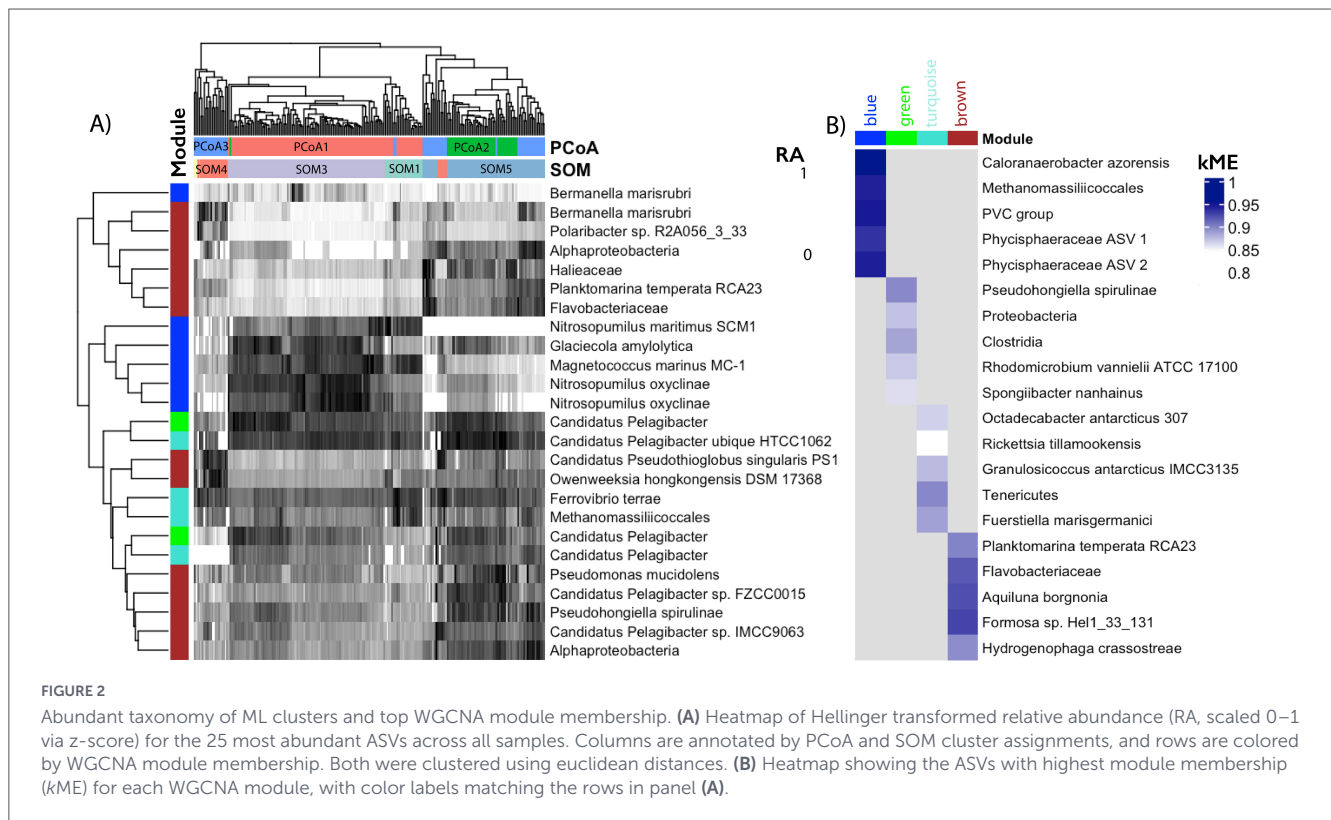
The winter-spring period of the dataset (March–June 2020) was dominated by PCoA1, SOM3, and the Blue WGCNA Module (Figure 1). Abundant taxa during this time were primarily slow growing chemoautotrophs (Figure 2), with available metagenomes showing enriched genes for such complex metabolisms (Figure 4). Both PCoA1 and SOM3 had a high mean estimated doubling time (Figure 3), but with a large sample spread. This could be due to the additional intraseasonal taxonomic community highlighted by WGCNA in the Green module prior return of sunlight at the end of March (Figure 1). Prevalent only in the dark period, the Green module was made up of a unique selection of pelagibacterales

taxa, however the module generally followed the same trait-relationships as the more abundant Blue module species (e.g., positive correlations with estimated doubling time and % GC content and negative correlations with Chlorophyll concentration, genome size, and 16S rRNA gene copy number; Supplementary Figure 7).

4 Discussion

4.1 Identifying seasonal microbial ecotypes for the surface Arctic Ocean

Taking the summation of information on cluster cut-offs, their external environmental drivers (e.g., season), internal community composition, and inferred and measured functional potential, we suggest 4 seasonal microbial ecotypes for the surface Arctic Ocean with distinct parameterizable functional traits (Table 1). To qualify as a unique functional ecotype, inferred functional estimates (e.g., doubling time, genome size) had to be statistically significant from all other clusters within the same methodology (PCoA/SOM) or significantly correlated with at least 2 physiological variables (WGCNA), contain a distinctive community composition from the dominant clusters temporally before and after, and have estimated



functional distinctions (Table 1) validated by available metagenome observations. Ecotypes were described by the timing which the representative clusters meeting the above criteria were present within the time series.

The summer Bloom-Associated Ecotype rapidly increases its abundance during the period of the pelagic phytoplankton bloom, representing a shift in carbon source utilization and increased microbial growth rates. This Ecotype is represented in our dataset during mid-late June (Fong et al., 2024) by SOM1 (lowest estimated doubling times, high enrichment of cell division genes) and the WGCNA Turquoise Module (high kME of phytoplankton associated taxa).

The summer Post-Bloom Ecotype follows directly after the prior Bloom-Associated and represents a still fast-growing microbial community with a unique taxonomy and greater capacity for organic matter processing that indicates it should be recognized as its own functional group. In our dataset, this Ecotype is represented by PCoA2/3, SOM4/5, and the Brown Module. It is possible that the high kME taxa in the Brown Module that were significantly correlated with low doubling time and tracked the most abundant taxa in the distinctive SOM4 could also be considered its own ecotype. However, most differences between these SOMs were environmentally driven (Supplementary Figure 5) and the estimated genetic trait differences were not fully reflected in the select available metagenomes. More measurements would be required to functionally separate these two groups.

The Winter/Dark Ecotype represents the appearance of the low abundance Green module in early spring and is the only defined ecotype not to be represented across multiple clustering techniques. We still include it here, however, due to the module containing multiple unique structural features. Its most abundant ASVs

(Figure 2A) were assigned to the slow growing and genomically streamlined Pelagibacterales order (Tucker et al., 2021), which were also highly abundant in the Blue module and are common in the Arctic winter and spring upper water column (Chamberlain et al., 2025). Pelagibacterales are also uniquely distinguished by their high internal diversity—even within subclades, the micro-diversity of this species is extremely high, allowing them to track subtle shifts in environmental niches across the world's oceans (Delmont et al., 2019). It is therefore not surprising that they are highly abundant across multiple community clusters as there is likely greater speciation not captured in the reference genomes used here for taxonomic assignment. The ability to capture meaningful ecological differences and functional separation even without perfect taxonomic assignment highlights this methods robustness and is one of the primary benefits to using segmentation at the ASV-level.

The primary environmental deviation for the Green Module is an opposite correlation to Latitude, potentially indicating oceanographic partitioning at this time—although the decrease in Green module relative abundance occurs slightly after the Polarstern drifted across the Gakkel Ridge (Schulz et al., 2024). Additionally, the top kME taxa for this module (Figure 2B) are primarily specialized in breaking down recalcitrant carbon and their genomic representatives all contain at least one mechanism for boosting ATP yields during periods of electron scarcity (e.g., rhodopsins and other bacterial photosystems, motility). This indicates their suitability for surviving the scarcity period of polar night and adapting early to the return of the sun in early spring, as seen in the overwintering strategies of Arctic phytoplankton and sea-ice algae (Kvernvik et al., 2018; Hoppe et al., 2024).

The Winter/Light Ecotype (SOM3, PCoA2, Blue module) on the other hand is driven by phytoplankton associated and

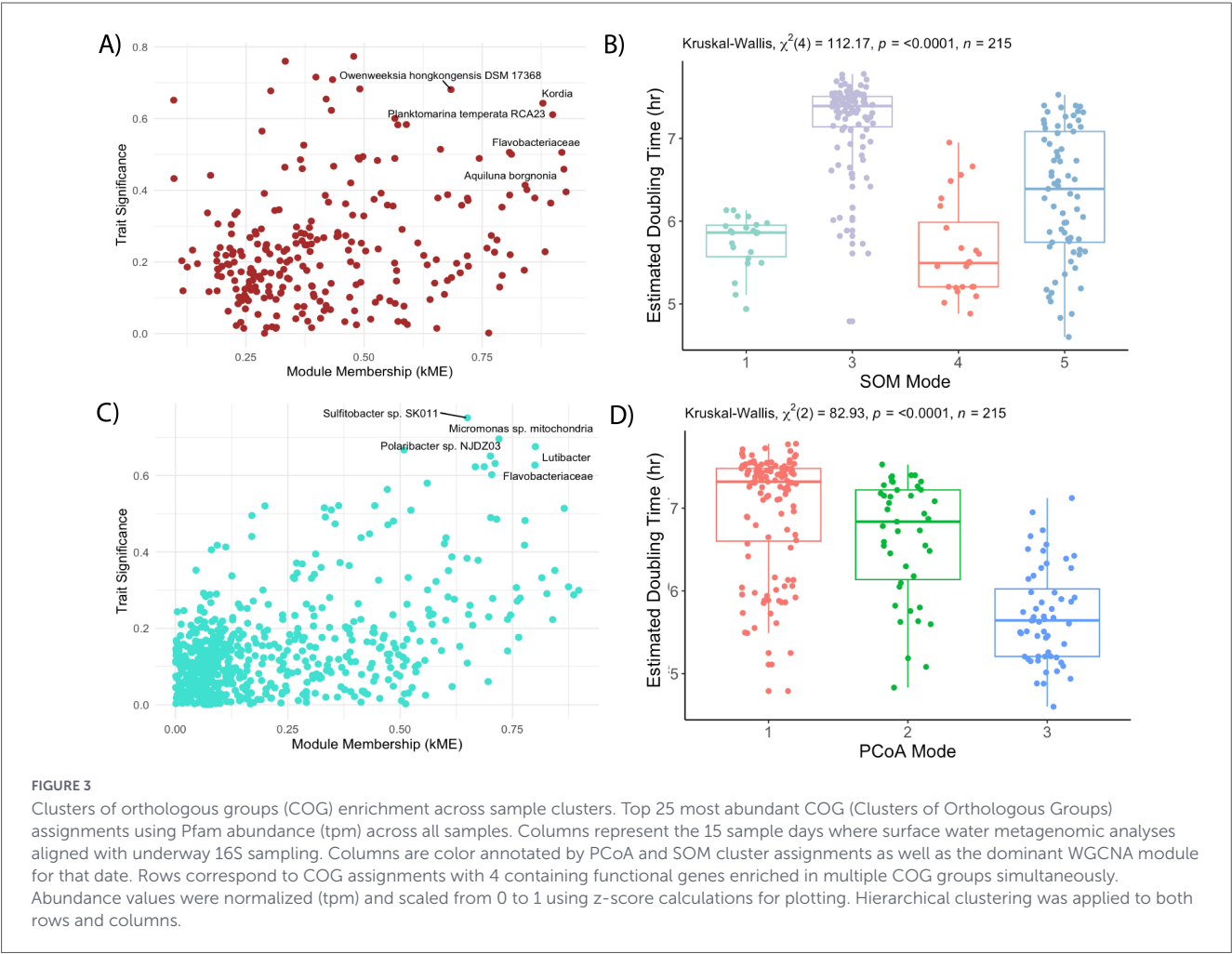


TABLE 1 Functional summary of surface ocean Arctic ecotypes.

Description	Growth rate estimate	Estimated genome size	Key taxonomic representatives	Enriched cellular functions	ML clustered output representatives
Bloom associated	Fastest	Large	Polaribacter, Suliflobacter	Motility and cell division	SOM1, Turquoise module
Post-bloom	Fast	Large	Flavobacteria, general	Carbohydrate and lipid processing	PCoA3, SOM5, Brown module
Winter/dark	Slowest	Streamlined	Various Gammaproteobacteria	Energy production and conversion	Green module
Spring/light	Slow	Streamlined	<i>Nitrospilimus</i> sp., Pelagibacter, Phycisphaeraceae sp.	Defense mechanisms, inorganic ion transport	PCoA2, SOM3, Blue module

chemosynthetic taxa (Figure 2B), which were likely brought to the surface from deeper in the water column due to intensified mixing at this time (Chamberlain et al., 2025) and representing a total functional shift that is best reflected in the most abundant ASVs in SOM3 (Figure 2A). SOM3 is also enriched in heterotrophic metabolism gene indicators, as well as defense mechanism genes—which would reflect the growing number of Spring phytoplankton and potential grazing activities by mixotrophic dinoflagellates, which were also brought to the surface and highly abundant at this time (Chamberlain et al., 2025).

Annually recurrent microbial ecotypes have been observed in other Arctic Ocean time series (Priest et al., 2025), due to strong seasonal selection pressures (Wietz et al., 2021). The layered functional organizations offered by our multi-method comparison may further inform ongoing debates over microbial functional redundancy, a concept used to explain the stability of many microbial processes despite high individual species turnover (Louca et al., 2018). The distinct temporal windows occupied by the SOM clusters and WGCNA modules (Figure 1) and clear shifts in community composition coinciding to measurable changes in dominant metabolic function (Figure 4),

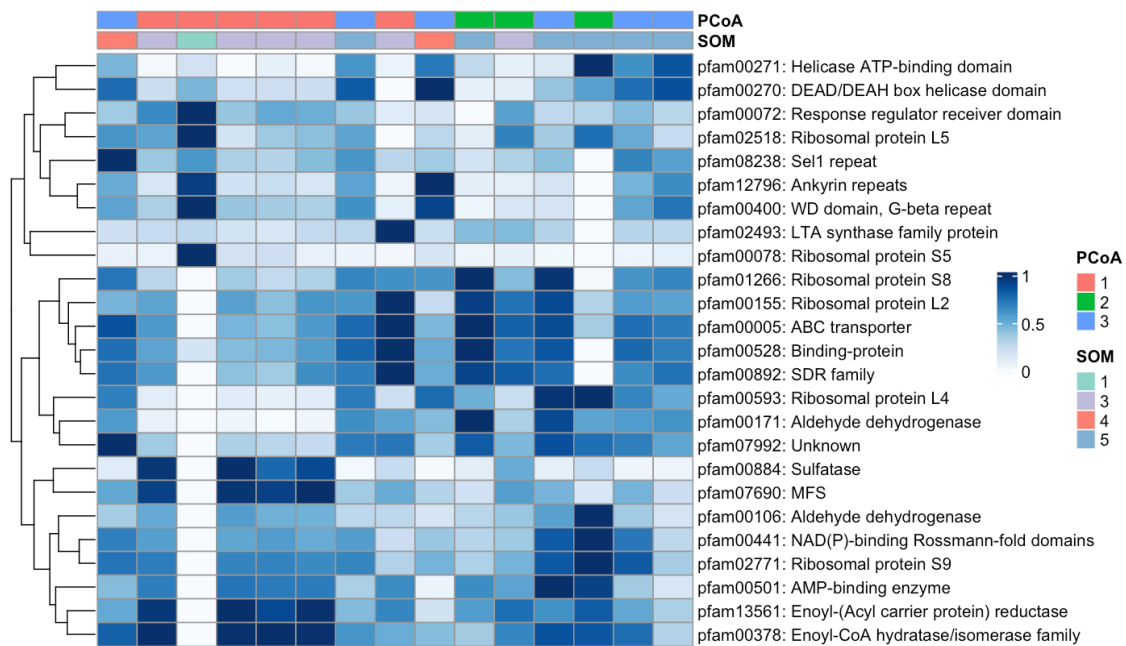


FIGURE 4

Codon-usage derived growth rates across clusters and highly correlated taxa. Growth rate is represented here by the gRodon estimated mean doubling time in hr, where a lower doubling time indicates faster growth. (A) Scatterplot of trait significance (doubling time) and module membership (kME) for all ASVs in the Brown module. (B) SOM clusters compared to doubling time. (C) Scatterplot of trait significance (doubling time) and module membership (kME) for all ASVs in the Turquoise module (D) PCoA clusters compared to doubling time. In (A,C), the top five significantly correlated and high kME ASVs are labeled by closest reference genome. In (B,D), the horizontal line of each boxplot indicates the median, the box spans the first to the third quartile, the whiskers indicate minimum and maximum values, and the dots indicate all values.

support studies emphasizing a tight coupling between taxa and function (Galand et al., 2018) and limited redundancy at ecologically relevant scales of environmental change (Larkin et al., 2025). As has been found in other study systems (Wilson et al., 2021), these methods revealed a layered picture of community succession. Seasonal shifts in community structure and overall function were captured by sample-based clustering (SOMs/PCoA) as the entire assemblage shifted coherently, but inter-seasonal responses to substrate pulses depended on specific, and not fully redundant, functional guilds that only the ASV-level co-occurrence WGCNA analysis was able to extract. By building a signed correlation network among individual ASVs that co-vary across samples, the modules emerge from shared temporal dynamics, not simply relative abundance. This explains why this method was more sensitive to dynamic environmental shifts—allowing it to better resolve fine-scale resource niches.

4.2 Environmental sensitivity and other dataset caveats

The strong correlations of longitude and salinity in the NMDS are notable, but not unexpected. Longitude (and latitude) represent the large-scale oceanographic gradients shaping community composition. Salinity (and temperature) are related to these gradients but also impacted by seasonal ice melt—which created critical thresholds for community structure to shift as recognized across our statistical and ML clustering (Table 1). The PCoA was the least sensitive to these finer environmental shifts, recognizing

only one community cluster both during summer melt and in fall 2020 when MOSAiC drifted into relatively fresher Transpolar Drift influenced waters (Schulz et al., 2024). Meanwhile, the period between June–September (summer melt season) had three distinct co-occurrence modules (Brown, Turquoise, and Blue; Figure 1C) and two unique SOM communities (SOM1 and SOM4; Figure 1B), indicating greater sensitivity of these methods to recognizing fine scale environmental and/or seasonal shifts in community structure. This advantage likely stems from the underlying data structure. PCoA axes are continuous and dominated by highly abundant taxa (Legendre and Gallagher, 2001). Rare ASVs get compressed near the origin within the original ordination, meaning their influence is likely to be under-represented (Shade et al., 2014). While this projection is useful in analyses specifically targeting a balanced representation of the community, our results indicate this may come at the cost of functional sensitivity and support other studies which have highlighted a mismatch between abundance and influence (Munson-McGee et al., 2022).

One result not yet discussed is the $n = 2$ segmentation of SOM2. Both samples were from a period of lower latitude transit through the marginal ice zone and likely represented a brief spatial disruption in temporal community succession. This scenario is also reflected in both the PCoA, which clustered these samples with the later dominant PCoA3 (Figure 1B), and WGCNA, which showed a brief inversion in the dominance of the Blue and Brown modules at this time (Figure 1C). These two samples are either true outliers in the dataset and the SOM map grid was the most sensitive projection, or it was a cut-off induced by insufficient grid-size or misinterpretation of map proximity (Kohonen, 2013). Given

that during the grid size sensitivity analysis, these samples clustered with what we consider SOM1 using a grid size was 3×3 , and SOM4 (nearest neighbor, Figure 1D) using the same grid size of 4×4 but choosing a k of 4, this is likely the case (Supplementary Figure 1). Because k -means clustering assumes clusters of equal variances, non-balanced clusters can result in unduly isolated, low-variance clusters (Jain, 2010). While SOM2 and SOM4 were not significantly different in most genetic parameters, there was some significant environmental variation and SOM2 did display a significantly higher GC content ($p = 0.0062$; Supplementary Figure 5). However, with only $n = 2$, we were not confident in these relationships and SOM2 was not considered functionally unique for the purposes of identifying Arctic Ocean microbial Ecotypes.

When discussing functionality, it is important to acknowledge that our analysis here is largely derived from community composition and metabolically inferred genetic profiles from amplicon data (Bowman and Ducklow, 2015), which can be biased by the available reference genomes. While this is not a measurement of functionality directly, it can still provide a valuable means of approximating metabolic potential—bridging the gap when more cost-intensive metagenomic or metatranscriptomic data are unavailable (Erazo et al., 2021). Given that the paprica output from this dataset produced functionally consistent patterns with our, albeit less temporally resolved, metagenomic observations, our study offers a cost-effective and scalable pathway to link microbial diversity with potential ecosystem function. This is especially important in data poor regions like the central Arctic Ocean where large temporal and spatial gaps exist (Mock et al., 2022). However, we do emphasize that these inferences (e.g., doubling times and genome size) represent potential and not realized function. Future work integrating multi-omics (e.g., higher resolution metagenomic or transcriptomic) and paired true biogeochemical rate measurements (e.g., Connors et al., 2024) to validate and refine amplicon-based predictions would be ideal.

4.3 Perspective on integrating data-driven ecotypes into mechanistic biogeochemical models

A central challenge in ocean modeling is representing the immense diversity of marine bacteria (Séférian et al., 2020). Our results demonstrate that segmentation approaches provide an efficient solution to the challenge of scaling between individual taxa and the bulk ecosystem compartments often used by process-based biogeochemical models by creating functionally meaningful intermediate units of complexity. We identified four distinct ecotypes in the surface Arctic Ocean, each characterized by clear environmental drivers and seasonal patterns. Two ecotypes dominated during winter and spring periods characterized by low light conditions, while two others were tightly coupled to the timing of summer phytoplankton blooms (Table 1). These ML-derived ecotypes represent interpretable ecological units ideal for model integration because they consolidate key physiological and functional attributes including growth strategies (16S copy number), metabolic potential (COGs), and salinity tolerance. Moreover, these approaches generate ready-made state

variables whose parameters, such as maximum growth rates, temperature-salinity optima, and substrate preferences, can be constrained from multi-omic datasets, eliminating dependence on outdated single-species culture estimates. Recent studies have demonstrated significant improvements achievable through incorporating such bacterial trait structures (Zakem et al., 2025).

We propose treating each ML derived ecotype as an emergent functional unit with parameter values derived from ensemble trait measurements (e.g., median maximum growth rate, half-saturation constants). Because these ecotypes maintain clear relationships with environmental covariates such as salinity, temperature, and longitude, their abundances can be dynamically adjusted using environment-dependent growth and loss terms. This approach adds adaptive realism to models without increasing computational complexity. In doing so, the selection of appropriate segmentation methods is critical. The clear difference between our ordination approaches highlights the necessity of employing non-linear, ML dimensionality-reduction techniques when deriving model compartments. While traditional axis-based ordinations (PCoA) grouped samples by water mass characteristics, they failed to capture the seasonally coupled temperature-salinity niches that more sophisticated methods like SOM and WGCNA successfully identified. Integrating comparable ML segmentation within model-data assimilation frameworks could substantially improve state estimation and parameter optimization (Kim et al., 2021, 2022), ultimately strengthening constraints on predicted bacterial feedback to ocean carbon cycling.

5 Conclusion

Statistical dimensionality reduction provides a workable bridge between microbial surveys and the coarse compartments required by biogeochemical modeling. Our results indicated that functional variation and shifts in species interaction were accurately captured by k -means segmentation—notably paired with SOMs—and we were able to identify potential “functionally coherent” groupings which crossed taxonomic boundaries (e.g., Pelagibacterales, Flavobacteria) and embodied distinct variations in the genetic traits which are most useful for predictive model parameterization. The utility of compartmentalizing the community in this way is that it goes beyond “size-class” based trait extraction under the argument of functional redundancy and still allows community-based groupings when taxonomic succession does matter. Furthermore, testing our workflow on amplicon sequencing derived ASVs demonstrates how these algorithms could be used to combine and mine previously generated legacy data for new ecological insight. Amplicon sequencing has long been the more widespread and affordable methodology compared to the richer but scarcer metagenomic and transcriptomic records. However, this work could still be improved—particularly in the Arctic, where data is scarce—by producing more *paired* ‘omics datasets for such dimensionality reduction methods to get better predictive insights into the mechanisms behind biogeochemical parameters of interest in a rapidly changing ocean.

Data availability statement

Publicly available datasets were analyzed in this study. The 16S rRNA gene amplicon sequence data are available through the National Center for Biotechnology Information (NCBI BioProject #PRJNA895866), with metadata available through the Arctic Data Center (<https://arcticdata.io/catalog/view/doi%3A10.18739%2FA2CC0TV5X>). The raw metagenomic data are also available through the NCBI (BioProject #PRJNA1160706) (Boulton et al., 2025). Data QC and ML code pipelines are available upon request from the corresponding authors.

Author contributions

ECh: Supervision, Methodology, Software, Writing – review & editing, Investigation, Funding acquisition, Conceptualization, Writing – original draft, Formal analysis, Visualization, Validation, Project administration, Data curation, Resources. WB: Investigation, Data curation, Methodology, Writing – review & editing, Formal analysis. ECo: Conceptualization, Writing – review & editing. TC: Formal analysis, Writing – review & editing. JB: Data curation, Writing – review & editing, Funding acquisition. JC: Data curation, Writing – review & editing. TM: Writing – review & editing, Resources. HK: Writing – review & editing, Supervision, Writing – original draft.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the National Science Foundation (NSF) Center for Chemical Currencies of a Microbial Planet (C-CoMP) publication #77. Amplicon sequence data collection and processing for MOSAiC was funded by the US NSF award, OPP 1821911, to JB. This award also supported MOSAiC participation for JC, JB, and ECh. JC was additionally supported by the US Department of Energy (DOE) Atmospheric System Research program (grant DE-SC0022046) and ECh was additionally supported by an NSF Graduate Research Fellowship. WB was supported by the Natural Environment Research Council and ARIES DTP (grant NE/S007334/1). Metagenome sequencing (project 10.46936/10.25585/60001271) was carried out by the DOE Joint Genome Institute (JGI) in a User Facility supported by the DOE Office of Science (grant DE-AC02-05CH11231). Development of this manuscript and TC was supported by an NSF Postdoctoral Research Fellowship award, OPP 2317681, to ECh. HK was supported by C-CoMP (NSF Award OCE-2019589).

Acknowledgments

The data used in this research was collected as part of the Multidisciplinary Drifting Observatory for the Study of Arctic Climate (MOSAiC) project (data tag MOSAiC20192020) and we acknowledge and thank all persons involved in this expedition as per the MOSAiC extended acknowledgement (Nixdorf et al., 2021). We would especially like to thank all members of Team Ecosystem led by Clara Hoppe, Alison Fong, and Rolf Gradinger (Fong et al., 2024) and specifically the Eco-Omics team (Mock et al., 2022), led by Katja Metfies and Thomas Mock.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. As a writing tool while revising the first draft of the manuscript to improve grammar and clarify language.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2026.1786397/full#supplementary-material>

References

- Anderson, M. J., and Willis, T. J. (2003). Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. *Ecol. Soc. Am.* 84, 511–525. doi: 10.1890/0012-9658(2003)084[0511:CAOPCA]2.0.CO;2
- Asnicar, F., Thomas, A., Passerini, A., Waldron, L., and Segata, N. (2023). Machine learning for microbiologists. *Nat. Rev. Microbiol.* 22, 191–205. doi: 10.1038/s41579-023-00984-1
- Botía, J., Vandrovčova, J., Forabosco, P., Guelfi, S., D'Sa, K., et al. (2017). An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. *BMC Syst. Biol.* 11:47. doi: 10.1186/s12918-017-0420-6
- Boulton, W., Salamov, A., Grigoriev, I., Calhoun, S., LaButti, K., Riley, R., et al. (2025). Metagenome-assembled-genomes recovered from the Arctic drift expedition MOSAiC. *Sci. Data* 12:204. doi: 10.1038/s41597-025-04525-8
- Bowman, J. (2021). Making sense of a scent-sensing metaphor for microbes and environmental predictions. *mSystems* 6:e0099321. doi: 10.1128/mSystems.00993-21
- Bowman, J., Amaral-Zettler, L., Rich, J., Luria, C., and Ducklow, H. W. (2017). Bacterial community segmentation facilitates the prediction of ecosystem function along the coast of the western Antarctic Peninsula. *ISME J.* 11, 1460–1471. doi: 10.1038/ismej.2016.204
- Bowman, J., and Ducklow, H. (2015). Microbial communities can be described by metabolic structure: A general framework and application to a seasonally variable, depth-stratified microbial community from the coastal west antarctic Peninsula. *PLoS One* 10:e0135868. doi: 10.1371/journal.pone.0135868
- Boyd, P., Claustre, H., Levy, M., Siegel, D., and Weber, T. (2019). Multi-faceted particle pumps drive carbon sequestration in the ocean. *Nature* 568, 327–335. doi: 10.1038/s41586-019-1098-2
- Callahan, B., McMurdie, P., Rosen, M., Han, A., Johnson, A., and Holmes, S. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869
- Chamberlain, E. J., Rokitta, S., Rost, B., D'Angelo, A., Creamean, J. M., Loose, B., et al. (2025). Predictive links between microbial communities and biological oxygen utilization in the Arctic Ocean. *Limnol. Oceanogr.* 9999, 1–17. doi: 10.1002/LNO.70125
- Chamberlain, E., and Bowman, J. (2022). 16S and 18S rRNA Amplicon data for the Multidisciplinary Drifting Observatory for the Study of Arctic Climate (MOSAIC) Expedition in the Central Arctic Ocean, 2019–2020. Arctic Data Center. doi: 10.18739/A2CC0TV5X
- Clum, A., Huntemann, M., Bushnell, B., Foster, B., Foster, B., Roux, S., et al. (2021). DOE JGI metagenome workflow. *mSystems* 6:e00804-20. doi: 10.1128/mSystems.00804-20
- Connors, E., Dutta, A., Trinh, R., Erazo, N., Dasarathy, S., Ducklow, H., et al. (2024). Microbial community composition predicts bacterial production across ocean ecosystems. *ISME J.* 18:wrae158. doi: 10.1093/ismej/wrae158
- Delmont, T., Kiefl, E., Kilinc, O., Esen, O., Uysal, I., Rappé, M., et al. (2019). Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife* 8:e46497. doi: 10.7554/eLife.46497
- Devries, T., Primeau, F., and Deutsch, C. (2012). The sequestration efficiency of the biological pump. *Geophys. Res. Lett.* 39:13601. doi: 10.1029/2012GL051963
- Dithugoe, C., Bezuidt, O., Cavan, E., Froneman, W., Thomalla, S., and Makhallanyane, T. (2023). Bacteria and archaea regulate particulate organic matter export in suspended and sinking marine particle fractions. *mSphere* 8:e0042022. doi: 10.1128/msphere.00420-22
- Edwards, A., Cameron, K., Cook, J., Debonnaire, A., Furness, E., Hay, M., et al. (2020). Microbial genomics amidst the Arctic crisis. *Microb. Genom.* 6:e000375. doi: 10.1099/mgen.0.000375
- Erazo, N., Dutta, A., and Bowman, J. (2021). From microbial community structure to metabolic inference using paprica. *STAR Protoc.* 2:101005. doi: 10.1016/j.xpro.2021.101005
- Field, C., Behrenfeld, M., Randerson, J., and Falkowski, P. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* 281, 237–240. doi: 10.1126/science.281.5374.237
- Fong, A. A., Hoppe, C. J. M., Aberle, N., Ashjian, C. J., Assmy, P., Bai, Y., et al. (2024). Overview of the MOSAiC expedition: Ecosystem. *Elementa Sci. Anthropocene* 12:135. doi: 10.1525/ELEMENTA.2023.00135
- Galand, P., Pereira, O., Hochart, C., Auguet, J., and Debroas, D. (2018). A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *ISME J.* 12, 2470–2478. doi: 10.1038/s41396-018-0158-1
- Giering, S., Sanders, R., Lampitt, R., Anderson, T., Tamburini, C., Boutrif, M., et al. (2014). Reconciliation of the carbon budget in the ocean's twilight zone. *Nature* 507, 480–483. doi: 10.1038/nature13123
- Haft, D., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., et al. (2018). RefSeq: An update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* 46, D851–D860. doi: 10.1093/nar/gkx1068
- Hellweger, F. (2020). Combining molecular observations and microbial ecosystem modeling: A practical guide. *Ann. Rev. Mar. Sci.* 12, 267–289. doi: 10.1146/annurev-marine-010419-010829
- Henson, S. A., Sanders, R., Madsen, E., Morris, P. J., Le Moigne, F., and Quartly, G. D. (2011). A reduced estimate of the strength of the ocean's biological carbon pump. *Geophys. Res. Lett.* 38:2011. doi: 10.1029/2011GL046735
- Hernández Medina, R., Kutuzova, S., Nielsen, K., Johansen, J., Hansen, L., Nielsen, M., et al. (2022). Machine learning and deep learning applications in microbiome research. *ISME Commun.* 2:98. doi: 10.1038/s43705-022-00182-9
- Honjo, S., Eglinton, T. I., Taylor, C. D., Ulmer, K. M., Sievert, S. M., Bracher, A., et al. (2014). Understanding the role of the biological pump in the global carbon cycle understanding an imperative for ocean science. *Oceanography* 27, 10–16. doi: 10.5670/OCEANOGRAPHY.2014.78
- Hoppe, C. J. M., Creamean, J., Bowman, J. S., Heitmann, L., Chamberlain, E., Brenneis, T., et al. (2023). Year-round Discrete Underway Water Column Chlorophyll a Concentrations from the Central ARCTIC. PANGAEA. doi: 10.1594/PANGAEA.962597
- Hoppe, C., Fuchs, N., Notz, D., Anderson, P., Assmy, P., Berge, J., et al. (2024). Photosynthetic light requirement near the theoretical minimum detected in Arctic microalgae. *Nat. Commun.* 15:7385. doi: 10.1038/s41467-024-51636-8
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recog. Lett.* 31, 651–666. doi: 10.1016/j.patrec.2009.09.011
- Kassambara, A. (2023). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. Available online at: <https://rpkgs.datanovia.com/rstatix/> (accessed April 1, 2025).
- Kim, H. H., Bowman, J. S., Luo, Y. W., Ducklow, H. W., Schofield, O. M., Steinberg, D. K., et al. (2022). Modeling polar marine ecosystem functions guided by bacterial physiological and taxonomic traits. *Biogeosciences* 19, 117–136. doi: 10.5194/BG-19-117-2022
- Kim, H. H., Luo, Y.-W., Ducklow, H. W., Schofield, O. M., Steinberg, D. K., and Doney, S. C. (2021). GMDD - WAP-1D-VAR v1.0: Development and Evaluation of a One-Dimensional Variational Data Assimilation Model for the Marine Ecosystem Along the West Antarctic Peninsula, Geoscientific Model Development, in Review. Available online at: <https://gmd.copernicus.org/preprints/gmd-2020-375/> (accessed March 21, 2021).
- Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Netw.* 37, 52–65. doi: 10.1016/j.neunet.2012.09.018
- Kvernvik, A., Hoppe, C., Lawrenz, E., Prášil, O., Greenacre, M., Wiktor, J., et al. (2018). Fast reactivation of photosynthesis in arctic phytoplankton during the polar night. *J. Phycol.* 54, 461–470. doi: 10.1111/jpy.12750
- Langfelder, P., and Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559
- Larkin, A., Brock, M., Fagan, A., Moreno, A., Gerace, S., Lees, L., et al. (2025). Climate-driven succession in marine microbiome biodiversity and biogeochemical function. *Nat. Commun.* 16:3926. doi: 10.1038/s41467-025-59382-1
- Legendre, P., and Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia* 129, 271–280. doi: 10.1007/s004420100716
- Legendre, P., and Legendre, L. (2012). Numerical ecology. Developments in environmental modeling. *Dev. Environ. Model.* 24, 1–1003.
- Li, Z., Riley, W., Marschmann, G., Karaoz, U., Shirley, I., Wu, Q., et al. (2025). A framework for integrating genomics, microbial traits, and ecosystem biogeochemistry. *Nat. Commun.* 16:2186. doi: 10.1038/s41467-025-57386-5
- Louca, S., Polz, M., Mazel, F., Albright, M., Huber, J., O'Connor, M., et al. (2018). Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* 2, 936–943. doi: 10.1038/s41559-018-0519-1
- McElhinney, J., Catacutan, M., Mawart, A., Hasan, A., and Dias, J. (2022). Interfacing machine learning and microbial omics: A promising means to address environmental challenges. *Front. Microbiol.* 13:851450. doi: 10.3389/fmicb.2022.851450
- Mock, T., Boulton, W., Balmonte, J., Barry, K., Bertilsson, S., Bowman, J., et al. (2022). Multiomics in the central Arctic Ocean for benchmarking biodiversity change. *PLoS Biol.* 20:e3001835. doi: 10.1371/journal.pbio.3001835
- Munson-McGee, J., Lindsay, M., Sintes, E., Brown, J., D'Angelo, T., Brown, J., et al. (2022). Decoupling of respiration rates and abundance in marine prokaryoplankton. *Nature* 612, 764–770. doi: 10.1038/s41586-022-05505-3
- Nixdorf, U., Dethloff, K., Rex, M., Shupe, M., Sommerfeld, A., Perovich, D. K., et al. (2021). MOSAiC Extended Acknowledgement. Geneva: Zenodo.
- Oja, M., Kaski, S., and Kohonen, T. (2003). *Bibliography of Self-Organizing Map (SOM) Papers Bibliography of Self-Organizing Map (SOM) Papers: 1998–2001 Addendum*. Available online at: <https://www.researchgate.net/publication/2605370> (Accessed July 29, 2025)
- Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P., O'Hara, R. B., et al. (2015). *Vegan: community ecology package. R package vegan, vers. 2.2.1.*
- Priest, T., Oldenburg, E., Popa, O., Dede, B., Metfies, K., von Appen, W., et al. (2025). Seasonal recurrence and modular assembly of an Arctic pelagic marine microbiome. *Nat. Commun.* 16:1326. doi: 10.1038/s41467-025-56203-3

- Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019). Application of machine learning in microbiology. *Front. Microbiol.* 10:827. doi: 10.3389/fmicb.2019.00827
- R Core Team. (2023). *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rantanen, M., Karpechko, A. Y., Lipponen, A., Nordling, K., Hyvärinen, O., Ruosteenoja, K., et al. (2022). The Arctic has warmed nearly four times faster than the globe since 1979. *Commun. Earth Environ.* 3, 1–10. doi: 10.1038/s43247-022-00498-3
- Schulz, K., Koenig, Z., Muilwijk, M., Bauch, D., Hoppe, C. J. M., Droste, E. S., et al. (2024). The Eurasian Arctic Ocean along the MOSAiC drift in 2019–2020: An interdisciplinary perspective on physical properties and processes. *Elementa Sci. Anthropocene* 12:00114. doi: 10.1525/ELEMENTA.2023.00114
- Schuur, E., McGuire, A., Schädel, C., Grosse, G., Harden, J., Hayes, D., et al. (2015). Climate change and the permafrost carbon feedback. *Nature* 520, 171–179. doi: 10.1038/nature14338
- Séférian, R., Berthet, S., Yool, A., Palmiéri, J., Bopp, L., Tagliabue, A., et al. (2020). Tracking improvement in simulated marine biogeochemistry between CMIP5 and CMIP6. *Curr. Clim. Change Rep.* 6, 95–119. doi: 10.1007/s40641-020-00160-0
- Shade, A., Jones, S., Caporaso, J., Handelsman, J., Knight, R., Fierer, N., et al. (2014). Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio* 5:e01371-14. doi: 10.1128/mBio.01371-14
- Smith, M. M., Angot, H., Chamberlain, E. J., Droste, E. S., Karam, S., Muilwijk, M., et al. (2023). Thin and transient meltwater layers and false bottoms in the Arctic sea ice pack—Recent insights on these historically overlooked features. *Elementa Sci. Anthropocene* 11:25. doi: 10.1525/ELEMENTA.2023.00025
- Spang, A., and Ettema, T. (2016). Microbial diversity: The tree of life comes of age. *Nat. Microbiol.* 1:16056. doi: 10.1038/nmicrobiol.2016.56
- Tatusov, R., Galperin, M., Natale, D., and Koonin, E. (2000). The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36. doi: 10.1093/nar/28.1.33
- Topçuoğlu, B., Lesniak, N., Ruffin, M., Wiens, J., and Schloss, P. D. (2020). A framework for effective application of machine learning to microbiome-based classification problems. *mBio* 11:e00434-20. doi: 10.1128/mBio.00434-20
- Tucker, S., Freel, K., Monaghan, E., Sullivan, C., Ramfelt, O., Rii, Y., et al. (2021). Spatial and temporal dynamics of SAR11 marine bacteria across a nearshore to offshore transect in the tropical Pacific Ocean. *PeerJ* 9:e12274. doi: 10.7717/peerj.12274
- Walters, W., Hyde, E., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., et al. (2016). Improved bacterial 16S rRNA Gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *mSystems* 1:e00009-15. doi: 10.1128/mSystems.00009-15
- Wehrens, R., and Kruisselbrink, J. (2018). Flexible self-organizing maps in kohonen 3.0. *J. Stat. Softw.* 87, 1–18. doi: 10.18637/JSS.V087.I07
- Weissman, J., Hou, S., and Fuhrman, J. (2021). Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proc. Natl. Acad. Sci. U S A* 118:e2016810118. doi: 10.1073/pnas.2016810118
- Wietz, M., Bienhold, C., Metfies, K., Torres-Valdés, S., von Appen, W., Salter, I., et al. (2021). The polar night shift: Seasonal dynamics and drivers of Arctic Ocean microbiomes revealed by autonomous sampling. *ISME Commun.* 1:76. doi: 10.1038/s43705-021-00074-4
- Wilson, J., Chamberlain, E., Erazo, N., Carter, M., and Bowman, J. (2021). Recurrent microbial community types driven by nearshore and seasonal processes in coastal Southern California. *Environ. Microbiol.* 23, 3225–3239. doi: 10.1111/1462-2920.15548
- Zakem, E., McNichol, J., Weissman, J., Raut, Y., Xu, L., Halewood, E., et al. (2025). Functional biogeography of marine microbial heterotrophs. *Science* 388:eado5323. doi: 10.1126/science.ado5323
- Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:17. doi: 10.2202/1544-6115.1128