LIMNOLOGY and OCEANOGRAPHY

ASLO

Limnol. Oceanogr. 9999, 2025, 1–17

doi: 10.1002/lno.70125

RESEARCH ARTICLE

Predictive links between microbial communities and biological oxygen utilization in the Arctic Ocean

Emelia J. Chamberlain [1,2*], **Sebastian Rokitta** [3], **Björn Rost** [3,4], **Alessandra D'Angelo** [5], **Jessie M. Creamean** [6], **Brice Loose** [5], **Adam Ulfsbo** [7], **Allison A. Fong** [3], **Clara J. M. Hoppe** [3], **Elise S. Droste** [8], **Daiki Nomura** [9], **Kirstin Schulz** [10], **Jeff Bowman** [1]

[1]Scripps Institution of Oceanography, University of San Diego California, La Jolla, California, USA; [2]Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA; [3]Alfred-Wegener-Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany; [4]University of Bremen, Bremen, Germany; [5]Graduate School of Oceanography, University of Rhode Island, Kingston, Rhode Island, USA; [6]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA; [7]Department of Marine Sciences, University of Gothenburg, Gothenburg, Sweden; [8]School of Environmental Sciences, University of East Anglia, Norwich, UK; [9]Hokkaido University, Hakodate, Japan; [10]Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, Texas, USA

Abstract

Microbial metabolism influences rates of net community production (NCP), exerting a direct biological control on marine oxygen and carbon fluxes. In the Arctic, it is increasingly important to understand and quantify this process, as ecological and oceanographic conditions shift due to changing climate. Here, we describe potential ecological links between pelagic microbial diversity and an NCP precursor, biological oxygen utilization, using machine learning and paired observations of community structure and metabolic activity from a seasonally and spatially variable transect of the Arctic Ocean (2019–2020 MOSAiC Expedition). Community structure was determined using 16S (prokaryotic) and 18S (eukaryotic) rRNA gene amplicon sequencing, and metabolic activity was derived from $\Delta O_2/Ar$. Using self-organizing maps, we identified clear successional patterns in observed microbial community structure that were seasonally driven in the upper ocean and vertically stratified with depth. Metabolic activity was also stratified, with a primarily net heterotrophic water column (median −1.5% biological oxygen saturation), excepting periodic oxygen supersaturation (maximum: 13.6%) within the mixed layer. Using DNA sequences as predictor variables, we then constructed a random forest regression model that reliably reconstructed biological oxygen concentrations (root mean squared error = 4.14 μmol kg^{-1}). Top predictors from this model were from heterotrophic (bacteria) or potentially mixotrophic (dinoflagellate) taxa. These analyses highlight biologically driven diagnostic tools that can be used to expand biogeochemical datasets and improve the microbial perspectives and metabolisms represented in ecological models of net productivity and carbon flux in a changing Arctic Ocean.

*Correspondence: emelia.chamberlain@whoi.edu

Associate editor: Bingzhang Chen

Data Availability Statement: All sequence data used in this manuscript are findable through the Arctic Data Center (https://doi.org/10.18739/A2CC0TV5X) and archived at the National Center for Biotechnology Information, NCBI Accession PRJNA895866 (Chamberlain and Bowman 2022). Continuous raw (https://doi.org/10.1594/PANGAEA.969630) and processed (https://doi.org/10.1594/PANGAEA.971603) O_2 and Ar gas data used in this manuscript are available through the Pangaea repository (Rokitta et al. 2024a, 2024b). Scripts for the processing of discrete data from this time series, as used in this manuscript, will be made available upon request. Ancillary oceanographic data associated with Schulz et al. (2024) are available through the Arctic Data Center (https://doi.org/10.18739/A21J9790B; Schulz et al. 2023). Chlorophyll *a* data available through Pangaea repository: Chl *a* from the underway system is available via https://doi.org/10.1594/PANGAEA.963277 (Hoppe et al. 2023a), Chl *a* data from the CTD rosette casts is available via https://doi.org/10.1594/PANGAEA.963277 (Hoppe et al. 2023b).

Net community production (NCP), or the metabolic balance between photosynthetic carbon fixation and heterotrophic aerobic respiration, imposes a biological control on the ocean carbon sink (Volk and Hoffert 1985). The Arctic Ocean is considered a net carbon sink: abiotically, due to the cooling and downwelling of inflow waters (Yasunaka et al. 2023), and biotically, due to high levels of net primary production along coastal shelves (MacGilchrist et al. 2014). However, published estimates of surface Arctic Ocean NCP are relatively low (average ~ 4, maximum ~ 20% biological oxygen supersaturation) with both temporal and spatial variability (Eveleth et al. 2014; Ulfsbo et al. 2014). In a recent time-series study from the western Arctic Ocean, surface NCP displayed heterogeneous and non-linear responses to environmental change, likely due to compounding ecosystem adjustments (Zhou et al. 2024). Understanding how the complex physical and biological controls on NCP interact and shift under changing Arctic conditions, such as ocean warming and sea ice decline (Meredith et al. 2019), is critical for predicting the future of the Arctic Ocean carbon sink.

Microbial community structure exerts a strong control on NCP (Guidi et al. 2016; Wang et al. 2018) and thus a biological control on carbon sequestration. Recent advancements in sequencing technology (Edwards et al. 2020; Mock et al. 2022) and comprehensive field efforts aimed at monitoring the microbial community of the Arctic Ocean (Royo-Llonch et al. 2021) and associated ice habitats (Campbell et al. 2022b) have provided a nice baseline for measurements of Arctic microbial community structure. Previous studies have found that such measurements can be utilized to estimate metabolic rates (Erazo et al. 2021; Connors et al. 2024), predict key transitions in ecosystem state (Bowman et al. 2017), and better understand the biological mechanisms underlying variability in observed biogeochemical processes (Dutta et al. 2022). In polar regions specifically, clear relationships between community structure (taxonomic composition) and NCP have been observed (Campbell et al. 2022a), with some eukaryotic community members acting as viable predictors of measured rates (Lin et al. 2017). However, prokaryotic predictors of NCP have been less well studied despite their critical role in global respiration (Worden et al. 2015) and observed predictive capacity for other ecological processes (Bowman et al. 2017; Dutta et al. 2022).

Machine learning allows us to extract the embedded patterns and predictive relationships between biogeochemical processes and diversity data (Bowman 2021). In recent years, an increasing number of studies have leveraged machine learning algorithms when analyzing community structure data for such purposes, with great progress in predicting environmental information from community structure alone (Dutta et al. 2022; Connors et al. 2024). To train such models, large paired datasets of both the taxonomic makeup of the microbial community and the biogeochemical parameter of interest are required. The 2019–2020 Multidisciplinary Drifting Observatory for the Study of Arctic Climate Expedition (Nicolaus et al. 2022; Rabe et al. 2022; Fong et al. 2024) provided a novel opportunity to collect paired measurements of both community structure and net productivity estimates across seasonal cycles in the under-sampled region of the central Arctic Ocean. MOSAiC was designed to carry out a year-long cross-disciplinary and process-level study of the interconnected central Arctic ocean–ice–atmosphere climate system across a seasonally and spatially variable drifting transect of the central Arctic Ocean.

In this study, we investigate the biogeochemical potential and predictive capacity of microbial (bacterial, archaeal, and eukaryotic) community structure on the relative ratio of biologically transformed oxygen (O_2) and biologically inert argon (Ar), $\Delta O_2/Ar$, which can be used as a signal of biological activity reflecting the net metabolic balance between photosynthesis and respiration, that is, the net trophic state controlling NCP (Eveleth et al. 2014). First, we analyzed the general trends and oceanographic context of water column microbial community composition and succession during the MOSAiC Drift using Self Organizing Maps (SOMs; Wehrens and Kruisselbrink 2018). Next, we identified specific microbial taxa with clear predictive linkages to observed biological oxygen utilization through the construction of a random forest (RF; Breiman 2001) regression model. Taken together, our findings help advance ecological understanding of the current Arctic Ocean and raise new questions regarding the crucial role which microbial organisms, particularly heterotrophs, play in regulating pelagic net community production and carbon sequestration.

Materials and methods

Oceanographic data collection—MOSAiC Expedition

In October 2019, the German icebreaker RV *Polarstern* was tethered to a multi-year ice floe in the Amundsen Basin and an ice camp was established to conduct research while passively drifting (4 October 2019–2031 July 2020; Fig. 1). A break in this drift occurred when *Polarstern* was required to temporarily leave, then re-establish, the ice camp during resupply (27 May–15 June 2020; considered "active transit" in Fig. 1). On July 31st, *Polarstern* reached the marginal ice zone and the original ice floe broke apart. *Polarstern* then relocated further north (considered "active transit" in Fig. 1), and a second ice camp was established to continue observations for the remainder of the MOSAiC time series (21 August 2020–2020 September 2020). There are two significant gaps in the data presented here, a period where no project personnel were onboard (December 2019–March 2020) and another during the May–June 2020 resupply mission.

Ecological and biogeochemical data used in this study were collected from either *Polarstern*'s underway seawater system (inlet at 11 m depth) or from Niskin bottles attached to a Conductivity, Temperature and Depth (CTD) rosette package (Fig. 1). Details on CTD operations and ecological sampling from the rosette can be found in Rabe et al. (2022) and Fong

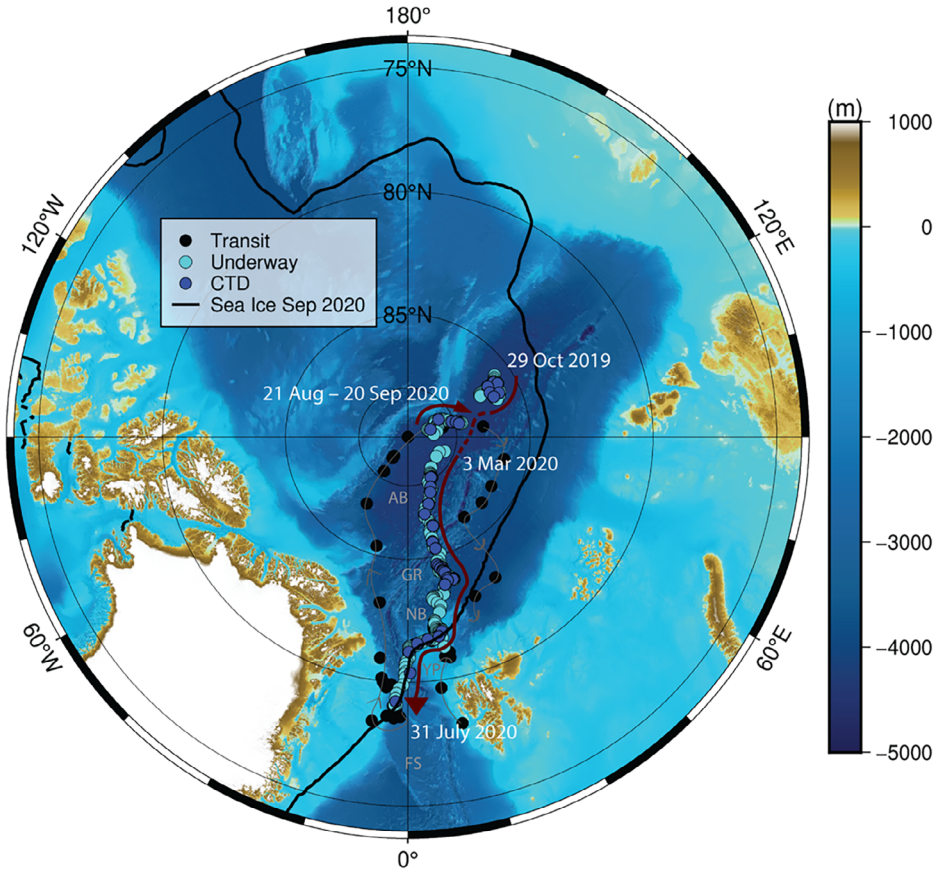

Fig. 1. Sampling locations. Geographic location and bathymetry (IBCAO v5; Jakobsson et al. 2024) of daily underway (UW) sample collections (cyan) and CTD stations (navy) during the MOSAiC Drift. Passive drift direction is indicated by the maroon arrows, with the dashed portion representing our data gap. Samples collected during active transit of the *Polarstern* are marked in black (i.e., not collected at the MOSAiC floe location during drift, thin gray arrows indicate sailing direction). Average sea ice extent from the minimum month, September 2020, is marked by the thick black line (Hersbach et al. 2023) and geographic features are labeled in gray, where AB = Amundsen Basin, GR = Gakkel Ridge, NB = Nansen Basin, YP = Yermak Plateau, and FS = Fram Strait.

et al. (2024). CTDs were equipped with sensors for measuring pressure, temperature, conductivity, and total dissolved oxygen (Tippenhauer et al. 2023a, 2023b). Values for dissolved oxygen were then corrected using periodic Winkler titrations (Ulfsbo et al. 2023). Temperature and salinity of the underway seawater system were collected by the remote temperature sensor SBE38 integrated with the SBE21 Thermosalinograph (Seabird) system installed on *Polarstern* (Haas et al. 2021; Kanzow et al. 2021; Rex et al. 2021a, 2021b, 2021c). Calculated oceanographic parameters used in this study, such as mixed layer depth and water mass influence, are from Schulz et al. (2024). A total of 334 of the water column (CTD rosette or underway system) observations for microbial community structure were sampled at the same time as water column chlorophyll *a* (Chl *a*) concentration (Hoppe et al. 2023b, 2023a).

Microbial community structure and segmentation

A total of 693 seawater samples (approximately 0.5–1 L) were collected in acid-cleaned and thrice sample-rinsed 1 L HDPE plastic bottles once daily from the ship's underway system (11 m) and on an approximately weekly basis from Niskin bottles attached to the CTD rosette (full water column; 0–4000 m). Collected samples were filtered immediately or kept dark at 0°C until filtered (within 4 h) through a sterile 0.2 μm Supor membrane disc filter (Pall Corporation) using a vacuum filtration manifold. The manifold was flushed with MilliQ and ethanol sterilized before, in between, and after samples. Filters were stored at −80°C onboard *Polarstern* until the end of the expedition when they were shipped to Scripps Institution of Oceanography on dry ice and stored at −80°C.

A KingFisher™ Flex Purification system and MagMax Microbiome Ultra Nucleic Acid Extraction kit (ThermoFisher Scientific) were used for extractions. Extracted material was then sent to the Environmental Sample Preparation and Sequencing Facility (ESPSF) at Argonne National Laboratory where amplicon library preparation and sequencing were conducted using a 151 bp × 151 bp paired-end run on the Illumina Miseq platform. rRNA gene amplification was

conducted using universal primers 515F and 806R (16S, V4; Walters et al. 2016) and 1380F and 1510R (18S, V9; Amaral-Zettler et al. 2009). The R package dada2 (Callahan et al. 2016) was used to denoise and merge Illumina reads prior to taxonomic classification using PAthway PRediction by phylogenetIC placement (paprica v0.7.0; Bowman and Ducklow 2015; Erazo et al. 2021). Further details on amplification, data QC, and taxonomic assignment can be found in Supporting Information Text S1.

Briefly, the paprica pipeline was used to assign individual amplicon sequence variants (ASVs) to its closest relative among completed genomes in either the Genbank RefSeq (16S; Haft et al. 2018) or PR2 (18S; v4.13.0 Guillou et al. 2013) reference databases (which does not necessarily indicate the presence of that exact strain, but a similar organism). For bacteria and archaea, paprica also uses the point of placement in the reference tree to estimate genomic characteristics such as genome size and 16S rRNA gene copy number (Erazo et al. 2021), as well as theoretical growth rates using the gRodon package (Weissman et al. 2021; Connors et al. 2024). Estimated doubling times from codon usage do not reflect in situ temperature conditions and therefore represent theoretical maximums only used to compare between samples and not as in situ rates. Final read counts for 16S ASVs were normalized to estimated 16S rRNA gene copy number before calculations of relative abundance.

Self Organizing Maps were trained using the "kohonen" package (Wehrens and Kruisselbrink 2018) to assign each sample to a group of similar samples (Bowman et al. 2017), effectively segmenting the community into statistically coherent groups. Self Organizing Map training was conducted using the Hellinger transformed relative abundance matrices of either 16S or 18S ASVs (independent models) on a 6×6 toroidal grid with hexagonal map units. Grid size was selected qualitatively based on the iterative distribution of samples assigned to each map unit in grids up to 10×10. Segmentation was achieved through k-means clustering of the map units, with a final k (number of clusters) selected qualitatively based on the within-clusters sum of squares scree plot and iterative experimental variation around the perceived optimum, or "elbow," aiming for a reasonable distribution of clusters and samples within the map grid. This results in each sample assigned to a specific cluster representing a unique prokaryotic/eukaryotic community ecotype, here termed "mode."

Kruskal–Wallis and Wilcoxon signed-rank tests (rstatix package; Kassambara 2023a) were used to compare mean environmental and genetic parameters between modes. Modes were additionally compared to the results of a non-metric multidimensional scaling (NMDS) ordination. This was derived from a Bray–Curtis dissimilarity matrix of Hellinger transformed 16S or 18S ASV relative abundances and correlated with basic environmental parameters (vegan package; Oksanen et al. 2015). A permutational ANOVA (PERMANOVA, 999 permutations) was used to test whether there was a significant ($\alpha = 0.05$) difference between modes (vegan package; Oksanen et al. 2015). All statistical analyses were conducted in R, version 4.3.1 (R Core Team 2023) with plots generated using packages pheatmap (Kolde 2019), ggplot2 (Wickham 2016), and associated packages (e.g., ggpubr; Kassambara 2023b).

Total oxygen (O_2) and biological oxygen anomalies (O_2/Ar)

Biological oxygen saturation, estimated from the ratio of O_2 and Ar concentrations, was used as a qualitative indicator of recent metabolic activity and a functional estimate of trends in NCP (Eq. 1). This value was obtained following the methods and assumptions of Craig and Hayward (1987) where the ratio of physical O_2 concentration to O_2 at air-saturation equals the ratio of Ar concentration to Ar at air-saturation.

$$\Delta O_2/Ar = \left[\left(\frac{[O_2]_{meas}/[Ar]_{meas}}{[O_2]_{sat}/[Ar]_{sat}}\right)\right] - 1 \quad (1)$$

Dissolved O_2 and Ar concentrations were measured using a Pfeiffer QMG 220 quadrupole membrane inlet mass spectrometer (MIMS) with a flow-through silicon capillary membrane inlet and pulse tube cooler for water vapor removal. Underway seawater was measured continuously throughout the entirety of the MOSAiC Expedition as described in Rokitta et al. (n.d.), except during daily instrument calibrations (2-point; 0% and saturated O_2/Ar gas) and bottle sampling events. Discrete seawater was collected from the CTD rosette in Milli-Q rinsed air-tight 300 mL Biological Oxygen Demand (BOD) bottles and pumped directly through the MIMS to analyze dissolved O_2 and Ar. These were always paired with sample collection for microbial community structure (although not all microbial community structure samples have a paired gas measurement). To align continuous underway data with the discrete underway collections for microbial community structure, we calculated 6-h averages centered around the sampling time. Missing values represent periods of missing or poorly calibrated data. Details regarding the MIMS system and bottle sampling can be found in the Supporting Information Text S2.

Theoretical concentrations of O_2 and Ar at air saturation ($[O_2]_{sat}$ and $[Ar]_{sat}$) were calculated with the solubility equations from Garcia and Gordon (1992; 1993 – combined fit) and Hamme and Emerson (2004), respectively, using temperature recorded at the MIMS inlet and salinity recorded by the thermosalinograph. Comparing $[O_2]_{sat}$ and $[Ar]_{sat}$ to 21 : 1% synthetic air standard calibrations (Supporting Information Text S2), we translated the seawater spectrometer readings into total concentrations of $[O_2]_{meas}$ and $[Ar]_{meas}$ in μmol kg^{-1}. Following the approach used in Eveleth et al. (2014), these values were then used to extract only the biological

component of total oxygen concentration, $[O_2]_{bio}$ in μmol kg^{-1} (Eq. 2).

$$[O_2]_{bio} = \frac{[Ar]_{meas}}{[Ar]_{sat}}[O_2]_{sat}(\Delta O_2/Ar) \quad (2)$$

Seawater $[O_2]_{sat}$ and $[Ar]_{sat}$ were calculated individually for each sample using in situ practical salinity and potential temperature. Positive (or negative) values of $[O_2]_{bio}$ represent a metabolic tracer for net oxygen production (or consumption), or net productivity (or heterotrophy, respectively).

Random forest regressions

A RF regression model was used to extract the key microbial predictors of $[O_2]_{bio}$ in the upper water column ("randomForest" package; Liaw and Wiener 2002). The model was constrained to CTD samples collected from 122 m and above, encompassing both the halocline and mixed layer, but without potential impacts from the long residence times of O_2 in deep water ($n = 232$; 112 with $[O_2]_{bio}$ measurements, although two datapoints were not included in model construction due to indicators of deep-water upwelling). In certain models, feature selection using the Boruta algorithm (Kursa and Rudnicki 2010) was applied prior to model construction. All predictor variables were further evaluated following model construction by comparing the increase of model mean squared error (%IncMSE) which occurs when the variable is randomly permuted (Liaw and Wiener 2002). To optimize our analysis, eight RF models were tested using 70–30 stratified sub-sampling techniques (Frick et al. 2025) and varying combinations of microbial and environmental predictor variables (Supporting Information Table S1). To extract the best taxonomic predictors of $[O_2]_{bio}$, the final model was constructed using all 110 upper water column CTD samples, with only feature-selected 16S and 18S ASV relative abundances as predictor variables. Internal model accuracy for the final RF model was assessed through a bootstrapped cross-validation analysis, randomly withholding and re-predicting 10% of the data for 999 iterations ("rfUtilities" package; Evans and Murphy 2018). We then used this model to predict $[O_2]_{bio}$ values in the surface ocean for periods of the drift where non-paired CTD measurements were collected ($n = 120$), as well as the daily underway sample collections for microbial community structure ($n = 236$). "External" model accuracy for the final RF model was assessed by comparing the predicted discrete underway data to the integrated average of continuous underway $[O_2]_{bio}$ measurements ($n = 212$).

Results

Microbial community composition and Self Organizing Map segmentation

Denoising and initial QC yielded a total of 4613 16S and 4638 18S ASVs across all samples after removing 655 suspected organelle (16S) and 18 intracellular parasite (18S) sequences. The prokaryotic community was dominated by bacteria, with only 5.8% of the remaining 16S ASVs assigned to archaea. An overview of total community composition (including chloroplast and parasite sequences) is presented in Supporting Information Fig. S1.

The SOM analysis identified seven coherent 16S (Fig. 2a) and six coherent 18S (Fig. 3a) microbial modes occurring across the time series. Our NMDS analysis reaffirmed that both the prokaryotic (Fig. 2b; PM) and eukaryotic (Fig. 3b; EM) modes were composed of a distinctive community structure, with little overlap in the 2D ordination space. In both ordinations, distribution across the first two axes highlighted significant separation by time and depth within the water column (16S: $r^2 = 0.190$ and 0.619 respectively, $p = 0.001$; 18S: $r^2 = 0.333$ and 0.434 respectively, $p = 0.001$). Temperature (16S: $r^2 = 0.502$, $p = 0.001$; 18S: $r^2 = 0.526$, $p = 0.001$) and salinity (16S: $r^2 = 0.645$, $p = 0.001$; 18S: $r^2 = 0.524$, $p = 0.001$) were also significantly correlated. Latitude was only significant for the eukaryotic community (16S: $r^2 = 0.0067$, $p = 0.095$; 18S: $r^2 = 0.0841$, $p = 0.001$). The PERMANOVA on the resulting Bray–Curtis dissimilarity matrices confirmed that dispersion among groups was significantly different across modes for both the 16S ($p = 0.001$; $F[6, 671] = 163.39$) and 18S communities ($p = 0.001$; $F[5, 640] = 87.62$).

The 10 most abundant ASVs from each PM combined to only 36 unique 16S ASVs, indicating some ASVs were in the top 10 most abundant of more than one mode. Here, we compare the average relative abundance of these top sequences across all samples within each mode (Fig. 2c). PM-2 and PM-3 clustered together (Bray–Curtis dissimilarity) as deep water modes and contained high abundances of ASVs identified as Archaea, Proteobacteria, and Acidobacteria. PM-5 (early summer) and PM-7 (cosmopolitan) clustered together and contained similarly high abundances of Archaea and Proteobacteria taxa in addition to *Glaciecola amylolytica*. The other upper water column cluster contained late summer and fall modes PM-1, PM-6, and PM-4 and had high abundances of ASVs identified to reference genomes of complex carbon degraders from Flavobacteria, Proteobacteria, and Actinomycetia (Fig. 2c). Higher order taxonomy for each PM is presented in Supporting Information Fig. S2.

The 10 most abundant ASVs from each EM combined to 35 unique 18S ASVs. EM-1 and EM-6 both contained deep water samples with high abundances of heterotrophic eukaryotes and dinoflagellate genomes, although their community composition was largely distinct at the ASV level. EM-6 was occasionally sampled near the surface as well and additionally contained relatively high abundances of diatom ASVs, which overlapped with top species in other modes. Top ASVs in surface fall and winter modes EM-3 and EM-4 were similarly non-overlapping, except for a few specific dinoflagellate ASVs. EM-4 had significantly higher abundances of diatom sequences, particularly *Attheya longicornis*. The final cluster contained EM-2 and EM-5, which were sampled above 100 m

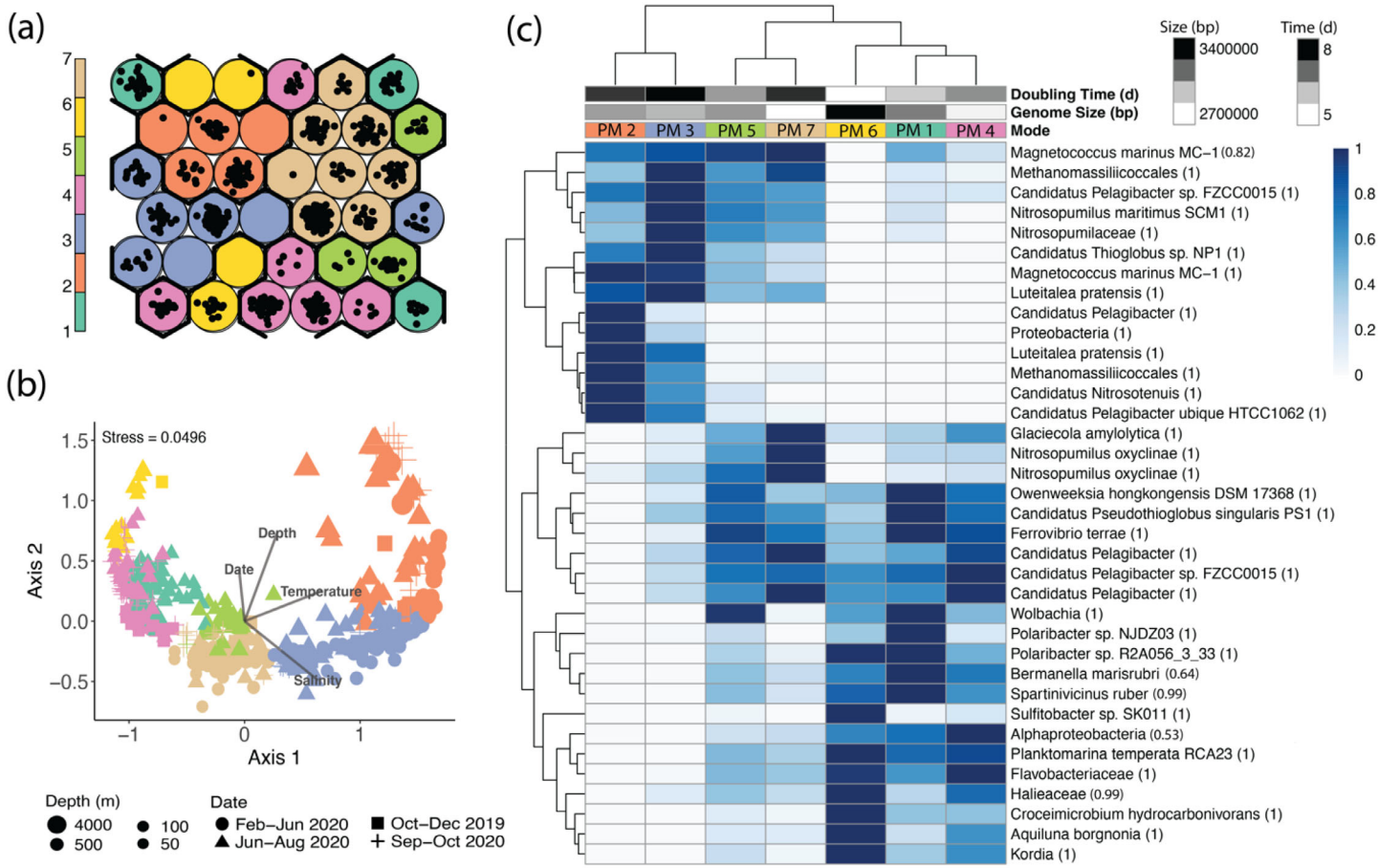

Fig. 2. Segmentation of the prokaryotic community. (**a**) The Self Organizing Map (SOM) grid with individual samples (black dots) sorted into their respective map units. Arranged as a toroid, units on opposite sides of the 2-D grid connect. K-means clustering of map units is indicated by color, with thick lines surrounding each prokaryotic mode (PM). (**b**) Individual samples presented in the first two dimensions of a non-metric multi-dimensional scaling ordination created from the Hellinger transformed relative abundances of 4613 16S ASVs (total community, not including suspected chloroplasts). Samples are colored by their assigned PM, sized by collection depth in the water column, and shaped by collection timeframe during the MOSAiC Drift. (**c**) PM community structure as represented by the 10 most abundant ASVs from each mode, when averaged across all their samples. Duplicate ASVs (within the top 10 most abundant of more than 1 PM) are shown only once. Hellinger transformed 16S ASV relative abundances were scaled using a z-score between 0 and 1. Bray–Curtis dissimilarity distances were used to cluster both ASVs (rows) and PMs (columns). Taxonomic IDs were assigned using the paprica pipeline, representing the closest relative among Refseq genomes and placement proportion. Columns are additionally colored by average predicted minimum doubling time (in hours) and average genome size (in base pairs) across each PM community.

from summer into fall and contained overlapping top ASVs assigned to a variety of dinoflagellate, diatom, and prasinophyte reference genomes (Fig. 3c). Higher order taxonomy for each EM is presented in Supporting Information Fig. S3.

The Kruskal–Wallis tests confirmed statistically significant variability in the average environmental conditions between modes (Supporting Information Fig. S4). Deep water modes EM-1, EM-6, PM-2, and PM-3 persisted below the mixed layer (Fig. 4) over the entirety of the drift and contained, on average, higher temperatures and salinities than all other modes (Table 1). PM-3 and EM-1 corresponded to significantly higher mean temperatures (Supporting Information Fig. S4; Supporting Information Table S2) and their vertical stratification in the upper water column closely followed that of Atlantic-influenced water (Fig. 4), indicating that the variation in deep water community structure was likely related to water mass influence. Estimated theoretical doubling times indicated significantly faster growth rates for PM-6 and PM-1 than for PM-2 and PM-3. PM-6 also had the largest genome size, while PM-7 had the smallest (Fig. 2c; Supporting Information Table S2). Average Chl *a* concentration was highest in EMs 2 and 5 and lowest in EMs 1 and 3 (Fig. 3c).

Total and biological oxygen concentration

Water column total dissolved oxygen concentrations $[O_2]_{tot}$ as calculated from the discrete BOD bottle samples ranged between 184.69 and 550.21 μmol kg^{-1} and generally

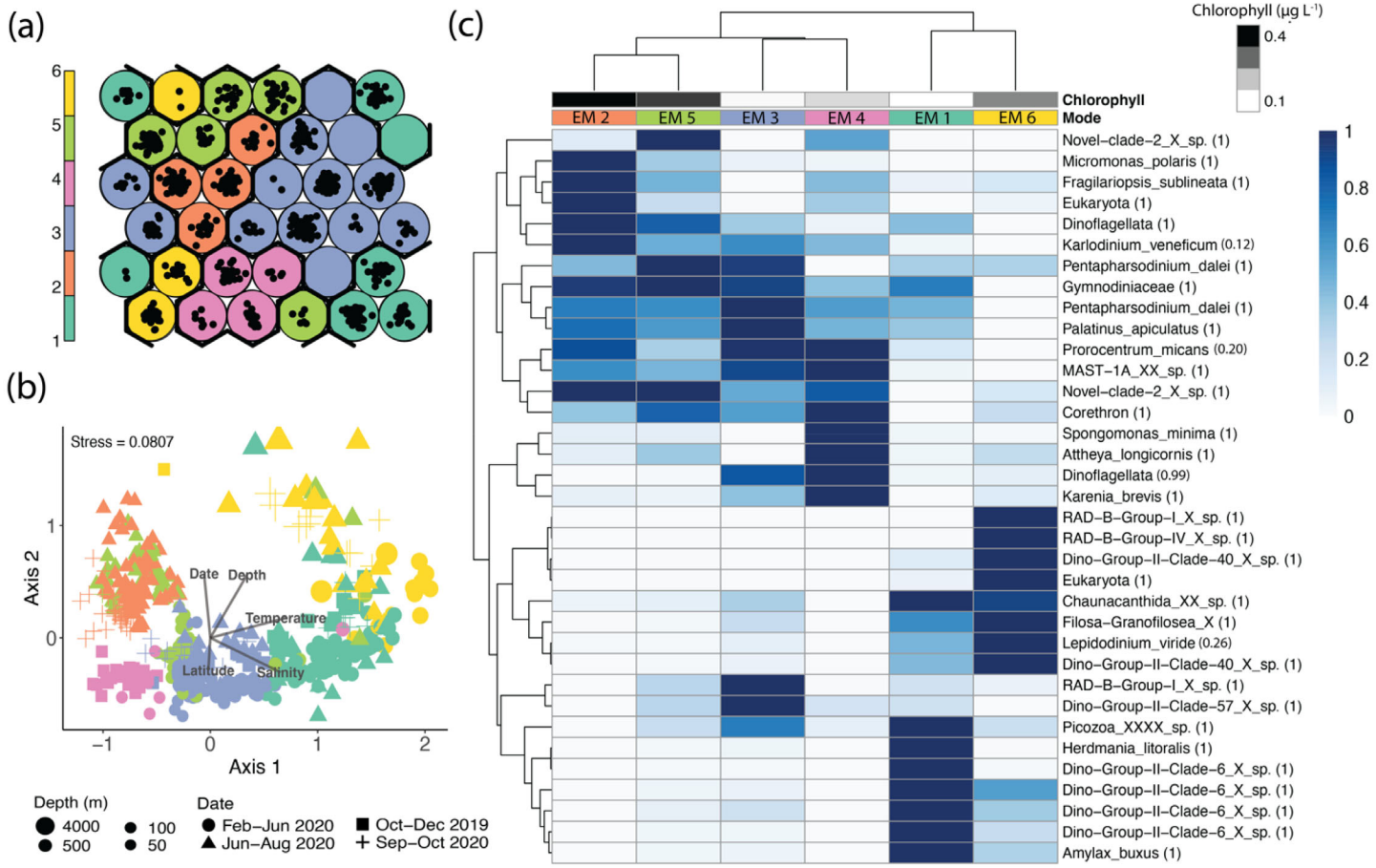

Fig. 3. Segmentation of the eukaryotic community. (**a**) The Self Organizing Map (SOM) grid (6 × 6) with samples (black dots) sorted into their respective map units. Arranged as a toroid, units on opposite sides of the 2-D grid connect. K-means clustering of map units is shown by color and thick lines surrounding each eukaryotic mode (EM). (**b**) Individual samples are presented in the first two dimensions of a non-metric multi-dimensional scaling (NMDS) ordination created from the Hellinger transformed relative abundances of 6870 18S ASVs (total community, not including suspected intracellular parasites). Samples are colored by their assigned EM, sized by collection depth in the water column, and shaped by collection timeframe during the MOSAiC Drift. (**c**) EM community structure is represented by the 10 most abundant ASVs averaged across all samples within each mode. Duplicate ASVs (within the top 10 most abundant of more than one EM) shown only once. Hellinger transformed 18S ASV relative abundances were scaled using a z-score between 0 and 1. Bray–Curtis dissimilarity distances were used to cluster both ASVs (rows) and EMs (columns). Taxonomic IDs were assigned using the paprica pipeline, representing the closest relative among PR2 genomes and placement proportion. Columns are additionally colored by average chlorophyll *a* (Chl *a*) concentration for each EM.

followed the pattern of $[O_2]_{tot}$ measured by the CTD sensors with an average deviation of -38.84 ± 53.46 μmol kg^{-1} (Supporting Information Fig. S5). Water column $[O_2]_{bio}$ (including transit stations) ranged between -72.73 and 29.51 μmol kg^{-1}. $[O_2]_{bio}$ was generally near 0 or negative, indicating oxygen consumption (net respiration) except in early spring and late summer where oxygen accumulation (net photosynthesis) was observed (Fig. 4). Below 11 m, negative $[O_2]_{bio}$ values indicated net heterotrophic conditions for most of the upper water column, with stronger undersaturation beneath the mixed layer (approx. 25 μmol kg^{-1} additional consumption; Fig. 4). Net heterotrophic conditions persisted across the entirety of the deeper water column (below 200 m) with the lowest $[O_2]_{bio}$ values occurring in the deepest water of the Amundsen Basin (Supporting Information Fig. S6) or with shoaling near bathymetric features (Fig. 4). Surface values (11 m and above) from the CTD rosette and averaged values from the underway system followed similar trends in time and space (Fig. 5c). Replicate BOD bottles indicated a measurement standard deviation of, on average, $\pm$ 1.85 μmol kg^{-1}.

Random forest regressions

Test model results indicated superior model performance when using only the feature selected ASVs compared to all other predictor variable combinations (Supporting Information Table S1). After combining all samples to train the final comprehensive ASV-based RF model (Fig. 5a), Boruta feature selection identified 128 ASVs (79 16S, 49 18S; Supporting information Table S3) as being statistically significant in making model predictions and were included as predictor variables

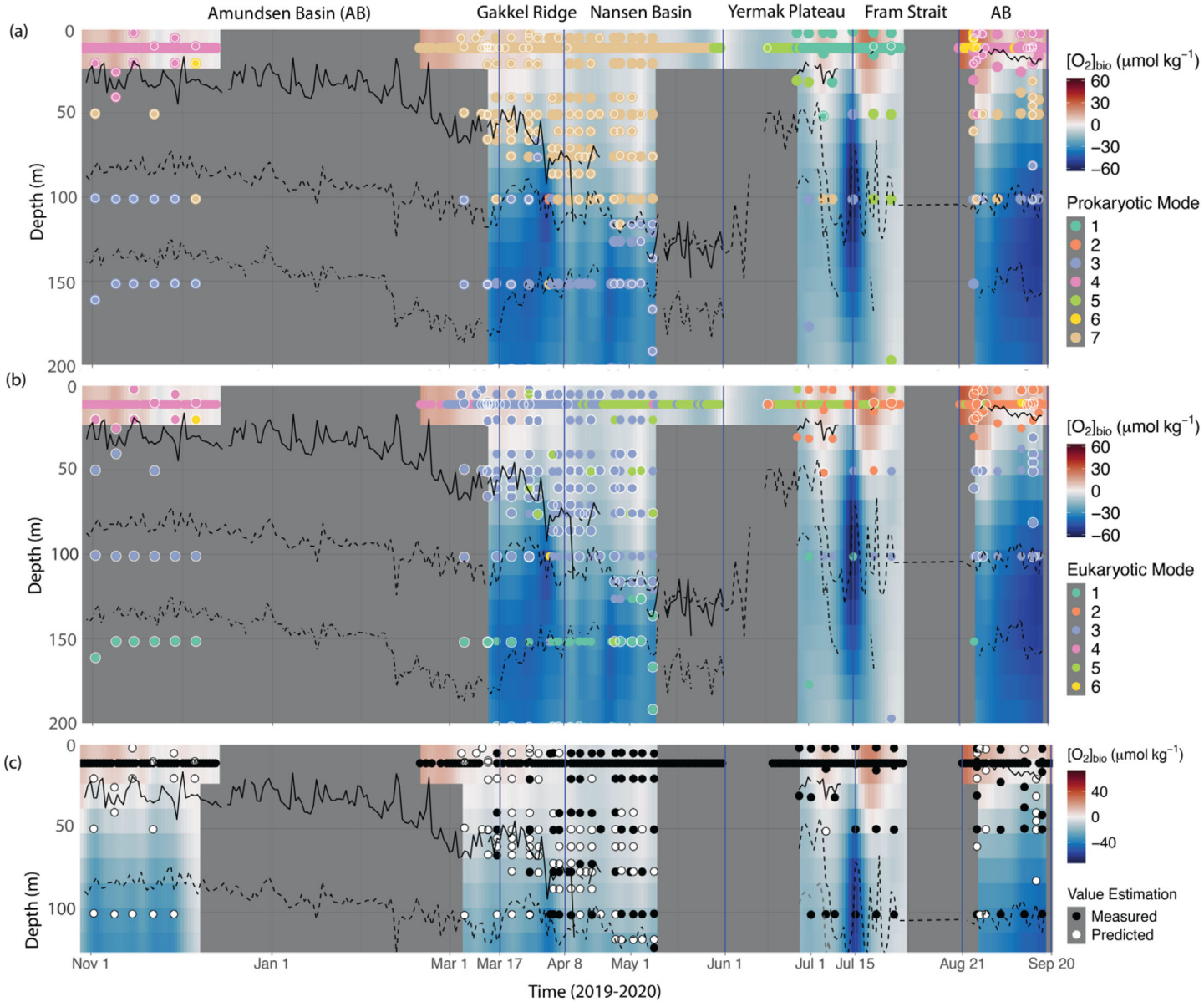

Fig. 4. Timeseries of upper water column biological oxygen ($[O_2]_{bio}$ in μmol kg^{-1}) and taxonomic mode during the MOSAiC Drift. Interpolated measurements for $[O_2]_{bio}$ are shown with (**a**) Prokaryotic Mode or (**b**) Eukaryotic Mode along the upper 200 m of the MOSAiC Drift. Panel (**c**) focuses on the upper 130 m and includes predicted values for $[O_2]_{bio}$ from the final random forest model. In all panels, points represent water collections for microbial community structure. In panels (a) and (b) they are colored by taxonomic mode. Full colored points (a, b) or black points (c) indicate paired DNA and $[O_2]_{bio}$ measurements. Points encircled by white (a, b) or colored white (c), indicate no matching $[O_2]_{bio}$ measurement, or predicted values (c). All interpolations were made using Multilevel B-spline approximation (Finley et al. 2022). Transit data during periods of ship relocation are not shown. Vertical blue lines represent geographic transitions across the Arctic Ocean basins and are indicated by name at the top of the chart. The black solid line represents mixed layer depth, the dashed black line indicates the top end of Atlantic influence, determined by where temperature begins to increase with depth, and the dashed gray line indicates the upper end of the Artic Atlantic Water mass (zero-degree isotherm).

in the final model. Top predictors (%IncMSE of greater than 4, $n = 25$) included 10 bacterial sequences, 4 archaeal sequences, and 11 total eukaryotic sequences, with 9 assigned to dinoflagellate species (Fig. 6). The ASV with the highest % IncMSE (7.42) was assigned to Syndiniales (Dino Group 1, clade 1), followed by the next four top predictors of an unidentified species of the Rhizaria order Chaunacanthida (% IncMSE = 7.17), *Actinomarinicola tropica* (%IncMSE = 6.41), an unidentified Bacteria genome with an NCBI BLAST 100% ID to Nitrospinaceae (%IncMSE = 6.12), and an unidentified Eukaryote genome with a 90% NCBI BLAST identity to an uncultured dinoflagellate (%IncMSE = 5.49). Overall, archaea

Table 1. Summary characteristics for prokaryotic (PM) and eukaryotic (EM) modes.

Mode	Seasonal/ spatial dominance	Estimated mean (16S)		Mean Chl *a* (μg L^{-1})	Mean temp (°C)	Mean salinity (g kg^{-1})	Primary dates observed	Primary depths observed
		Genome size (10^6 bp)	Doubling time[†] (h)					
PM 1	Summer bloom	3.13 ± 0.28	5.68 ± 0.53	0.61 ± 0.27	−1.60 ± 0.16	32.62 ± 1.08	Jul	Surface (< 50 m)
PM 2	Deep water	3.05 ± 0.22	7.44 ± 0.79	0.001 ± 0.00	−0.09 ± 0.81	34.91 ± 0.07	Year-round	Deep (> 500 m)
PM 3	Atlantic water	2.99 ± 0.12	8.02 ± 0.22	0.01 ± 0.008	0.43 ± 1.15	34.67 ± 0.25	Year-round	Mid (100–500 m)
PM 4	Winter surface	2.78 ± 0.32	6.43 ± 0.66	0.26 ± 0.19	−1.51 ± 0.61	31.39 ± 1.56	Sep–Dec	Surface (< 50 m)
PM 5	Summer surface	3.07 ± 0.27	6.29 ± 0.68	0.55 ± 0.28	−0.82 ± 1.59	33.94 ± 0.37	Jun–Jul	Surface–mid (< 100 m)
PM 6	Transpolar Drift*	3.43 ± 0.33	4.66 ± 0.71	0.40 ± 0.21	−1.65 ± 0.08	30.77 ± 0.90	Aug, Dec	Surface (< 50 m)
PM 7	Winter–spring	2.70 ± 0.18	7.51 ± 0.39	0.03 ± 0.03	−1.78 ± 0.22	33.91 ± 0.53	Feb–may	Surface–mid (< 100 m)
EM 1	Atlantic water	3.0 ± 0.17	7.99 ± 0.50	0.007 ± 0.008	0.90 ± 0.75	34.80 ± 0.15	Year-round	Mid–deep (> 150 m)
EM 2	Late summer bloom	2.9 ± 0.33	6.04 ± 0.64	0.45 ± 0.22	−1.51 ± 0.55	31.32 ± 1.70	Jul–Sep	Surface (< 50 m)
EM 3	Winter–spring	2.7 ± 0.17	7.59 ± 0.31	0.03 ± 0.02	−1.63 ± 0.55	33.93 ± 0.52	Mar–Apr	Surface–mid (< 100)
EM 4	Fall bloom	2.7 ± 0.32	6.71 ± 0.75	0.11 ± 0.15	−1.43 ± 1.06	32.54 ± 0.48	Nov–Dec	Surface (< 20 m)
EM 5	Summer bloom	3.0 ± 0.31	6.44 ± 1.09	0.37 ± 0.33	−1.30 ± 1.02	33.56 ± 1.27	May–Jun	Surface–mid (< 100 m)
EM 6	Deep water	3.1 ± 0.24	7.20 ± 0.97	0.26 ± NA	−0.58 ± 0.46	34.73 ± 0.92	Year-round	All (10–4000 m)

*Transpolar Drift here refers to the advective pathway of low salinity, river-rich surface water of Siberian origin.
[†]Doubling time here refers to theoretical doubling times as estimated from codon usage (Weissman et al. 2021).

were over-represented in the top predictors, making up 11% of the feature selected ASVs compared to only 6% of the overall prokaryotic community composition. Seventy-one percent of the eukaryotic taxa within the feature selected predictors were dinoflagellates, with another 16% assigned to protists with known heterotrophic (e.g., phagotrophy) behavior.

The cross-validation analysis of this model indicated a median root mean squared error of 4.13 μmol kg^{-1} (Supporting Information Fig. S9). When this model was applied to an independent validation dataset (underway $[O_2]_{bio}$), the model had poorer performance, with a median absolute error of ± 5.77 μmol kg^{-1} and a root mean squared error of 10.93 μmol kg^{-1} with a trend toward undersaturation of the measured values (Fig. 5c). However, overall trends were retained within the water column predictions, and the application of the model to community structure data lacking MIMS observations allowed us to expand the dataset from 112 upper water column CTD samples to 238 and cover the critical fall–winter transitional period (Fig. 4c).

Discussion

In this study, we identify the biological controls on NCP (proxy $[O_2]_{bio}$) using a unique, year-long time series of paired microbial and biogeochemical data over a spatially and temporally variable drift track across the central Arctic Ocean. Our results showed that the magnitude and trophic signal of water column biological oxygen utilization and microbial community structure were tightly coupled. We demonstrated that microbial community structure outperforms physical predictors in a random forest model to predict $[O_2]_{bio}$ (Supporting Information Table S1), although both measurements showed indicators of environmental shaping (Fig. 4; Table 1). Top predictors suggest strong bottom-up controls of NCP from the heterotrophic community (Fig. 6), but more experimental research is needed to determine the difference between correlative and causative effects.

Depth resolved $[O_2]_{bio}$ over the MOSAiC drift

$[O_2]_{bio}$ in the water column remained close to 0 or net heterotrophic at depth for almost the entirety of the MOSAiC Drift (Fig. 4). Even during summer, $\Delta O_2/Ar$ in the surface ocean (< 11 m) was, on average, only oversaturated by 0.8%. Previous studies have found spatially variable, but on average higher $\Delta O_2/Ar$ supersaturation (3.1 ± 2.9%; $n = 9157$) in surface waters of the Eurasian Arctic (Eveleth et al. 2014; Ulfsbo et al. 2014). However, most previous studies only report surface water data from July–October, which is the period in our data that showed the greatest $[O_2]_{bio}$ accumulation from in situ production and ice melt (Fig. 4). As expected, such net phototrophic conditions would not persist during the spring/winter periods, but our data highlight that limited seasonal data on high Arctic net production cannot provide accurate estimates on an annual time scale. It is likely that external carbon sources (Oziel et al. 2025), such as terrestrial input from rivers (as indicated by PM-6) or ice melt (as indicated by PM-5 and EM-2), would be required to sustain the levels of remineralization observed beneath the mixed layer over the course of the drift (Fig. 4).

Extremely negative $[O_2]_{bio}$ signals in the upper water column (e.g., < −60 μmol kg^{-1} at 100 m on July 15th; Fig. 4) can be explained by tidal influences (Rabe et al. 2022) or enhanced upwelling caused by turbulence from bottom topography as

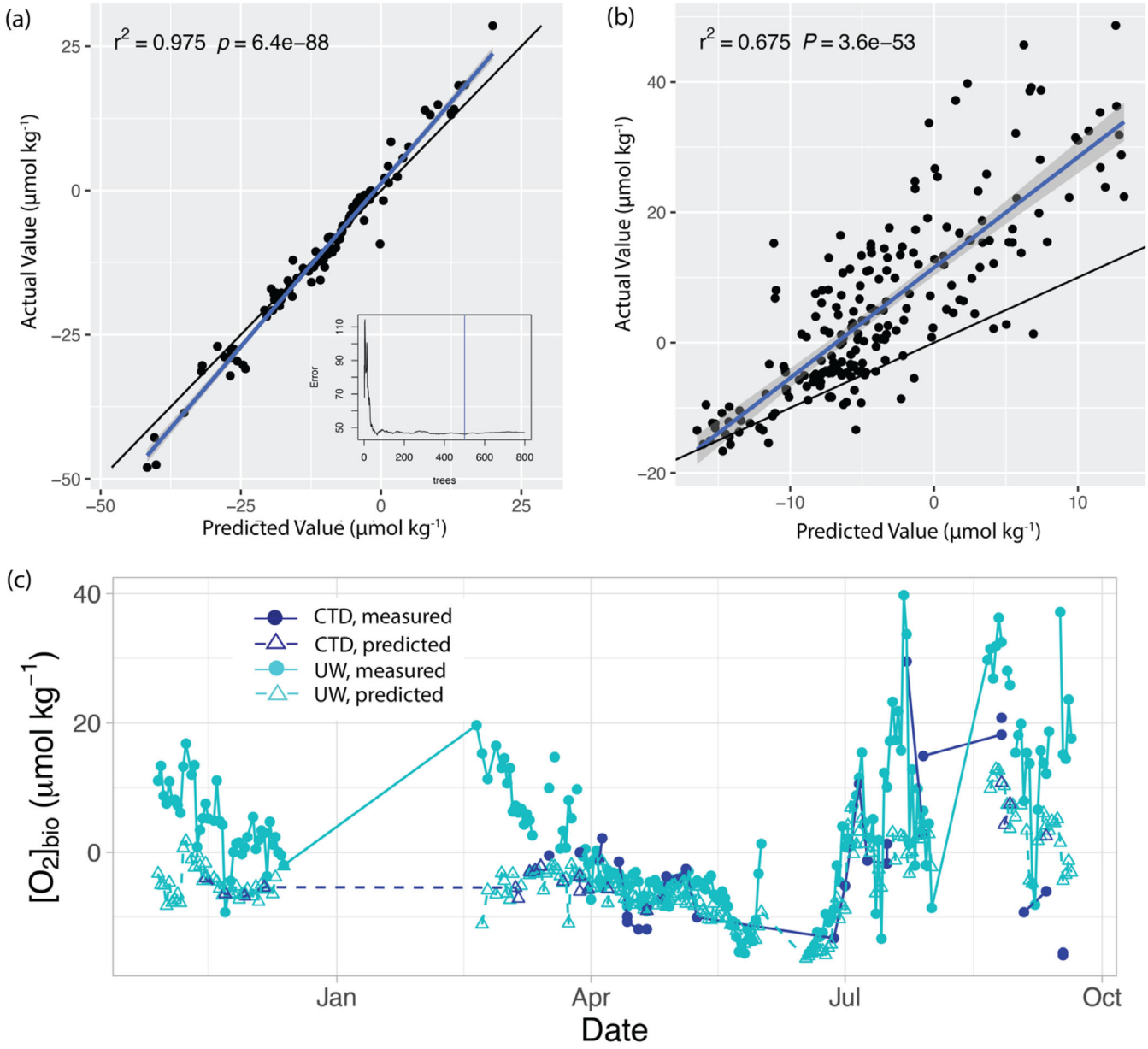

Fig. 5. Upper water column random forest (RF) model construction and external dataset validation. (**a**) 1 : 1 comparison of predicted $[O_2]_{bio}$ and measured $[O_2]_{bio}$ values with a linear regression from the RF training dataset comprised of all suitable paired surface ocean (< 122 m) DNA and $[O_2]_{bio}$ measurements collected from the CTD rosette. (**b**) 1 : 1 comparison of predicted and measured $[O_2]_{bio}$ values with a linear regression from the validation dataset comprised of all integrated daily underway $[O_2]_{bio}$ values generated from 6-h averages around the DNA sampling time. (**c**) Time-series view of the predicted (triangles, dashed) and measured (circles, solid) values for both the underway system (cyan) and surface (≤ 11 m) CTD measurements (dark blue).

the ship passed over the western slope of the Yermak Plateau layer (Schulz et al. 2024). The strongly positive $[O_2]_{bio}$ signal in the Amundsen basin during early spring is also likely linked to physical rather than biogeochemical processes given the decoupling from observed changes in community structure (Fig. 4), low light levels, and predicted undersaturation of these samples from our ASV-generated RF model (Fig. 5). Potential explanations include gas entrapment beneath high winter sea ice extent (Nicolaus et al. 2022) or frontal dynamics due to crossing bathymetric features, that is, the Gakkel ridge (Eveleth et al. 2014; Ulfsbo et al. 2014).

Seasonal and spatial context from water column microbial community structure

Prokaryotic mode and EM distribution corresponded with observed oceanographic conditions during MOSAiC, especially

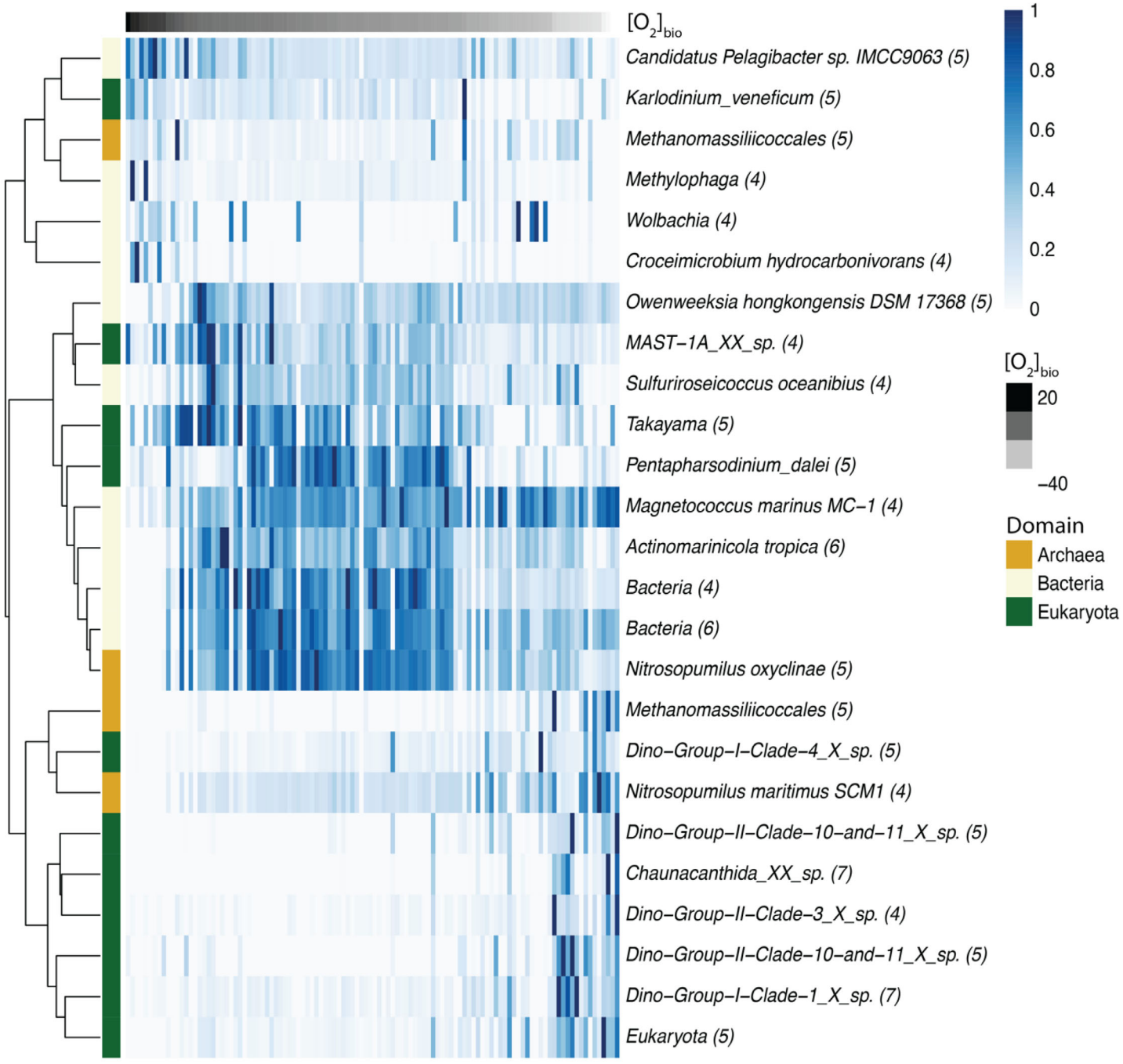

Fig. 6. Heatmap of the key microbial predictors for upper water column biological oxygen utilization. Scaled relative abundance (z-score between 0 and 1) of predictor variables (ASVs) with a percent Increase in Mean Squared Error (%IncMSE) ≥ 4% across all model training samples ($n = 110$). Samples (columns) are colored by biological oxygen ($[O_2]_{bio}$) concentration in μmol kg^{-1}. Bray–Curtis dissimilarity distances were used to cluster ASVs (rows), which are labeled both by Domain (color) and the highest resolved paprica taxonomic placement (closest relative among RefSeq or PR2 genomes) with its rounded %IncMSE included in parentheses.

vertical stratification within the water column (Fig. 4). This supports previous studies (Carter-Gates et al. 2020; Priest et al. 2023), which suggest that water mass influence plays an important role in the environmental structuring of taxonomic distribution in the pelagic Arctic Ocean. For example, PM-3 appeared almost exclusively below the upper boundary of Arctic Atlantic Water influence and reflected the average environmental characteristics of this water mass, including high

temperature and salinity (Table 1). Atlantic-influenced EM-1 was not present above the true demarcation of the Arctic Atlantic water mass between 100 and 150 m (Fig. 4). The strong density gradient of the halocline likely acted as a physical barrier for vertical exchange and dispersal of microbial communities, indicating that populations above and below this barrier developed independently of each other.

Disentangling seasonal and spatial (water mass driven) successional influences in the upper ocean was more complicated. PM-6, dominated by the freshwater indicator taxa Actinobacter (Supporting Information Fig. S2), was observed briefly in December 2019 and then again in late August of 2020. At those times, the MOSAiC floe was in waters influenced by the Transpolar Drift as indicated by increased river water fraction (Laukert et al. 2025) and reduced $\delta^{18}O$ isotopes (Schulz et al. 2024). The presence of Actinobacter, putative complex carbon degraders (Fig. 2c), and a low PM average salinity (Table 1) suggest a community spatial signal driven by fresh riverine input.

Seasonal mode composition was largely defined by metabolic strategy. In Fall 2019, the community composition (Fig. 3) and elevated Chl *a* concentrations of EM-4 (Table 1) indicate the remnants of an *Attheya*-rich under-ice bloom with a copiotrophic heterotrophic prokaryotic community (PM-4) and higher capacity for NCP (mixed layer $[O_2]_{bio}$ supersaturation; Fig. 4). Enhanced mixing in early March led to the entrainment of deeper water masses into the mixed layer. This transported the previously subsurface heterotroph/dinoflagellate dominated eukaryotic EM-3 and primarily chemotroph dominated PM-7 to the surface, which persisted in the mixed layer until crossing over the Gakkel Ridge (EM-3) or Yermak Plateau (PM-7). While there could be some spatial component, the gradual transition between EM-3 and EM-5 observed in April also corresponded to the return of sunlight and a surprisingly early exponential increase in Chl *a* that already started during EM-3 prevalence (Hoppe et al. 2024). The top ASVs between EM-3 and EM-5 showed a great deal of community overlap, with the main difference being a shift from dinoflagellate dominance toward increased phototrophic species, for example, a diatom sp. of *Corethron* (Fig. 3). This, in addition to the continuity of PM-7 and persistence of EM-5 well into the Fram strait, supports homogeneous selection pressures across basins with a seasonally driven vernal evolution within the eukaryotic community.

Our analyses suggest that summer succession in the prokaryotic community was driven by both season and substrate availability. The accumulated production signal (positive $[O_2]_{bio}$) from summer bloom EM-5 peaked in July. The prokaryotic summer community shifted in response to these bloom conditions (Fig. 4). Pre-bloom summer mode PM-5 was enriched in Alphaproteobacteria with top ASVs assigned to terrestrial taxa (Fig. 2), potentially from sediments picked up during ice formation melting out of the ice (Nicolaus et al. 2022). Bloom summer mode PM-1 contained high abundances of particle-associated taxa (Fig. 2), highlighting organic matter availability, here derived from either ice melt (Fig. 6) or in situ phytoplankton growth (Fig. 4) as a determining factor for bacterial taxonomic composition (Underwood et al. 2019; Piontek et al. 2021) over the MOSAiC year. In mid-July, EM-5 was gradually succeeded by EM-2, which had a greater abundance of phototrophic pico-plankton (i.e., *Micromonas polaris*) and ice-associated *Fragilariopsis sublineata*. These taxa are not surprising in low salinity waters (Table 1) and likely indicative of cryo-pelagic coupling related to ice melt and breakup during this time (Nicolaus et al. 2022).

Microbial community structure as a predictor for net community production

The results from our community-based RF model highlight the key role microbial community structure and metabolic balance play in determining water column NCP. Including physical parameters within the test RF models did not significantly improve model accuracy (Supporting Information Table S1), underscoring the strong predictive relationship between community composition and $[O_2]_{bio}$. The regions where our model performed poorly (significant underprediction of measured values from early spring in the validation dataset; Fig. 5) were also those that were most likely to be impacted by residence time driven decoupling in tracer $[O_2]_{bio}$ and active biogeochemical cycling. Another explanation for the general trend in underestimation of the underway $[O_2]_{bio}$ data, and limitation of the CTD RF model, is that the reduced temporal and vertical resolution of CTD casts led to an underrepresentation of surface (< 15 m) water samples in the model training data (Supporting Information Table S1). To better validate and constrain our RF model, we suggest future collections of similarly analyzed depth discrete paired DNA and O_2/Ar data in similar geographic locations to compare to and/or build a better predictive model constrained by multiple years of data. The more training data, the better suited such an RF model will be to predicting $[O_2]_{bio}$ in the highly heterogeneous Arctic Ocean.

Prokaryotic processes in particular play an important role in the mechanisms driving net carbon export in the pelagic ocean (Worden et al. 2015). Our results highlight this, with over half of the top predictor taxa in our final RF model belonging to prokaryotes (Fig. 6). Out of the eukaryotic ASVs which were feature selected as important $[O_2]_{bio}$ predictor taxa, the majority were dinoflagellates (potentially mixotrophic) or heterotrophic protists (e.g., Telonema). A similar phenomenon was observed in Guidi et al. (2016), which examined planktonic networks driving carbon export in the subtropics and found that the driving eukaryotic taxa in subnetworks correlated with carbon export were primarily dinoflagellates. Other studies from more nutrient-rich environments, such as Lin et al. (2017) from the Southern Ocean and Wang et al. (2018) from the mid-Atlantic Bight, found a mix of both autotrophic and heterotrophic taxa, which were good predictors and/or strongly correlated with NCP estimates. Many of these, however, were associated with bloom

conditions and when bloom samples were excluded from analysis in the Wang et al. study, only cryptophytes and bacteria remained associated with NCP rates.

When looking at the relative abundance across samples of top predictor sequences in our RF model, three distinctive and ecologically coherent groups of taxa emerge (Fig. 6). The top predictors associated with samples with high net production $[O_2]_{bio}$ sample group were primarily made up of sequences assigned to taxa directly associated with CO_2 fixation, for example, the dinoflagellate *Karlodinium veneficum*, which was dominant in late summer mode EM-2, or taxa associated with hydrocarbon degradation, for example, *Croceimicrobium hydrocarbonivorans* and *Methylophaga* (Gutierrez and Aitken 2014). Top predictors associated with the net neutral $[O_2]_{bio}$ sample group were primarily chemotrophs highly represented in winter surface PM-7 or motile heterotrophic taxa. Top predictors in samples with strong oxygen consumption ($[O_2]_{bio}$) were parasitic dinoflagellates and ASVs assigned to the domain Archaea.

Chemosynthetic archaea were generally overrepresented in top predictor taxa with a representative in each of the three "predictor groups" (Fig. 6). In previous studies, Archaea taxa were not clearly associated with, or assessed in relation to, NCP rates. However, most assessed only surface water within the mixed layer. Our depth-discrete water column data therefore provide a new insight into the predictive capacity of a less abundant microbial group. While predictive capacity does not necessarily equal mechanistic importance, several other recent studies have emphasized the role of chemoautotrophic Archaea (in addition to Bacteria) in altering carbon pools (Bayer et al. 2023) and regulating organic carbon export (Dithugoe et al. 2023) in the water column. Therefore, it is highly likely that the relative abundance of these key prokaryotic taxa may closely track changes in oxygen production and consumption rates, suggesting process-level biological mechanisms that directly influence the overall net metabolic balance and carbon sequestration potential in the Arctic Ocean.

Conclusions

Altogether, our findings highlight the influence of microbial community structure and biogeochemical potential on the net trophic state in the central Arctic Ocean. Microbial community structure (SOMs) displayed a strong seasonal and/or substrate-driven signal in the upper ocean but was oceanographically constrained at depth. These microbial signals were reflected in the observed biogeochemistry, and our upper ocean community structure-driven RF model predicted $[O_2]_{bio}$ with high fidelity. By specifically developing our RF model using only unique amplicon sequence predictor variables, we were able to directly highlight the crucial role microbes play in regulating this critical NCP precursor and assess potential discrepancies/shifts in the dominance of its biological and environmental drivers. Given the expensive nature of measuring such chemical rates directly, there is additional utility in these methods for constructing more complete biogeochemical datasets to accompany microbial studies.

Most of the key predictor ASVs from our RF model were assigned to potentially mixotrophic or heterotrophic taxa, indicating O_2 consumption as a dominant component in constraining total NCP within the water column. This is supported by recent modeling studies, which have predicted unexpected decreases in Arctic biological carbon pump efficiency due to accelerated remineralization rates under climate change scenarios (Oziel et al. 2025). While there are other O_2 loss mechanisms that can impact the integrated $[O_2]_{bio}$ signal and were not measured directly in conjunction with our data (such as zooplankton respiration and grazing), our results still show that microbial community structure holds at least predictive power for this signal.

$[O_2]_{bio}$ is closely related to atmospheric CO_2 uptake, which is strongly mediated by biological CO_2 consumption in the open water season (Juranek et al. 2019; Mathis et al. 2024). Large uncertainties in future CO_2 uptake by the Arctic Ocean are partly due to uncertainties in how trophic productivity will evolve with changes in the sea ice and seawater conditions that shape microbial community structures (Campbell et al. 2022b). This and other future work toward understanding the relationship between microbial community structure and biological oxygen utilization will therefore be important to accurately assessing the future of the Arctic Ocean carbon sink.

Author Contributions

Brice Loose and Jeff Bowman designed the project with input from Emelia J. Chamberlain, Sebastian Rokitta, Björn Rost, and Adam Ulfsbo. Allison A. Fong, Sebastian Rokitta, Jessie M. Creamean, Adam Ulfsbo, Jeff Bowman, Clara J. M. Hoppe, Emelia J. Chamberlain, Alessandra D'Angelo, Kirstin Schulz, Elise S. Droste, and Daiki Nomura contributed to data collection in the field. Sebastian Rokitta and Björn Rost developed the membrane inlet mass spectrometer, and Sebastian Rokitta conducted data QC and analysis for underway oxygen and argon concentrations. Emelia J. Chamberlain conducted data QC and analysis for discrete CTD oxygen and argon concentrations. Jeff Bowman and Emelia J. Chamberlain conducted the bioinformatics analysis. Clara J. M. Hoppe provided chlorophyll data, and Kirstin Schulz provided oceanographic parameters. Emelia J. Chamberlain merged data products, ran the machine learning models, and wrote the first manuscript draft with assistance from Jeff Bowman. Figure 1 was produced by Adam Ulfsbo; all others were produced by Emelia J. Chamberlain. All authors contributed to manuscript development/revisions and approved the submitted version for publication.

Acknowledgments

The data used in this research were produced as part of the Multidisciplinary Drifting Observatory for the Study of Arctic Climate (MOSAiC) project, which is published under the data tag MOSAiC20192020. We would like to thank all persons involved in MOSAiC 2019–2020 as listed in the MOSAiC extended acknowledgement (Nixdorf et al. 2021). We would specifically like to acknowledge the members of Team Ocean for their efforts in processing and making the CTD data and other core hydrographic parameters available. Additionally, we would like to thank Jacqueline Stefels, Annette Rinke, and all additional co-authors included in Hoppe et al. (2023a, 2023b) for help with sampling and processing Chl *a* and DNA samples. We would also like to thank current and former Bowman Lab members, particularly Jesse Wilson, Avishek Dutta, Elizabeth Connors, and Riley Hale, for assistance with data processing and the random forest analysis. We would additionally like to acknowledge the funding which made this research possible. Primary scientific analyses and Emelia J. Chamberlain, Jeff Bowman, Alessandra D'Angelo, and Brice Loose were funded through the US National Science Foundation (NSF) award OPP 1821911 to PI JB. This award also supported travel for Jessie M. Creamean. Emelia J. Chamberlain was additionally funded through an NSF Graduate Research Fellowship. Jessie M. Creamean was supported by the U.S. Department of Energy's Atmospheric System Research (DOE ASR) program (DE-SC0019745). Daiki Nomura was supported by the Japan Society for the Promotion of Science (grant numbers: JP18H03745; JP18KK0292; JP17KK0083; JP17H04715; JP20H04345) and by a grant from the Joint Research Program of the Japan Arctic Research Network Center. Adam Ulfsbo was supported by the Swedish Research Council Formas (grant 2018-01398). Allison A. Fong, Clara J. M. Hoppe, Sebastian Rokitta, and Björn Rost were supported by MOSAiC_ECO funds provided by the Alfred-Wegener-Institut Helmholtz Zentrun für Polar-und Meeresforschung. Elise S. Droste was supported by UK Natural Environment Research Council (NERC) through the EnvEast Doctoral Training Partnership (NE/L002582/1), as the Department for Business, Energy & Industrial Strategy (BEIS) through the UK Arctic Office. This work was additionally supported by the Swedish Polar Research Secretariat as part of the MOSAiC 2019–2020 expedition.

Conflicts of Interest

None declared.

References

Amaral-Zettler, L. A., E. A. McCliment, H. W. Ducklow, and S. M. Huse. 2009. "A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes G." *PLoS One* 4: e6372. https://doi.org/10.1371/journal.pone.0006372.

Bayer, B., K. McBeain, C. A. Carlson, and A. E. Santoro. 2023. "Carbon Content, Carbon Fixation Yield and Dissolved Organic Carbon Release From Diverse Marine Nitrifiers." *Limnology and Oceanography* 68: 84–96. https://doi.org/10.1002/LNO.12252.

Bowman, J. S. 2021. "Making Sense of a Scent-Sensing Metaphor for Microbes and Environmental Predictions." *mSystems* 6: 00993-21. https://doi.org/10.1128/MSYSTEMS.00993-21.

Bowman, J. S., L. A. Amaral-Zettler, J. J. Rich, C. M. Luria, and H. W. Ducklow. 2017. "Bacterial Community Segmentation Facilitates the Prediction of Ecosystem Function along the Coast of the Western Antarctic Peninsula." *ISME Journal* 11: 1460–1471. https://doi.org/10.1038/ismej.2016.204.

Bowman, J. S., and H. W. Ducklow. 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. https://doi.org/10.1371/journal.pone.0135868.

Breiman, L. 2001. Random Forests. Machine Learning. Kluwer Academic Publishers. The Netherlands. Vol. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. 2016. "DADA2: High-Resolution Sample Inference From Illumina Amplicon Data." *Nature Methods* 13: 581–583. https://doi.org/10.1038/nmeth.3869.

Campbell, K., B. A. Lange, J. C. Landy, et al. 2022*a*. "Net Heterotrophy in High Arctic First-Year and Multi-Year Spring Sea Ice." *Elementa: Science of the Anthropocene* 10: 00040. https://doi.org/10.1525/ELEMENTA.2021.00040.

Campbell, K., I. Matero, C. Bellas, et al. 2022*b*. "Monitoring a Changing Arctic: Recent Advancements in the Study of Sea Ice Microbial Communities." *Ambio* 51: 318–332. https://doi.org/10.1007/s13280-021-01658-z.

Carter-Gates, M., C. Balestreri, S. E. Thorpe, et al. 2020. "Implications of Increasing Atlantic Influence for Arctic Microbial Community Structure." *Scientific Reports* 10: 1–13. https://doi.org/10.1038/s41598-020-76293-x.

Chamberlain, E., and J. Bowman. 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. Santa Barbara, CA. https://doi.org/10.18739/A2CC0TV5X.

Connors, E., A. Dutta, R. Trinh, et al. 2024. "Microbial Community Composition Predicts Bacterial Production across Ocean Ecosystems." *ISME Journal* 18: 158. https://doi.org/10.1093/ISMEJO/WRAE158.

Craig, H., and T. Hayward. 1987. "Oxygen Supersaturation in the Ocean: Biological Versus Physical Contributions." *Science (1979)* 235, no. 4785: 202–205. https://doi.org/10.1126/science.235.4785.199.

Dithugoe, C. D., O. K. I. Bezuidt, E. L. Cavan, W. P. Froneman, S. J. Thomalla, and T. P. Makhalanyane. 2023. "Bacteria and Archaea Regulate Particulate Organic Matter Export in Suspended and Sinking Marine Particle Fractions." *M Sphere* 8: e00420-22. https://doi.org/10.1128/msphere.00420-22.

Dutta, A., T. Goldman, J. Keating, et al. 2022. "Machine Learning Predicts Biogeochemistry From Microbial Community Structure in a Complex Model System." *Microbiology Spectrum* 10: e0190921. https://doi.org/10.1128/SPECTRUM.01909-21.

Edwards, A., K. A. Cameron, J. M. Cook, et al. 2020. "Microbial Genomics Amidst the Arctic Crisis." *Microbial Genomics* 6: 1–20. https://doi.org/10.1099/mgen.0.000375.

Erazo, N. G., A. Dutta, and J. S. Bowman. 2021. "From Microbial Community Structure to Metabolic Inference Using Paprica." *STAR Protocols* 2: 101005. https://doi.org/10.1016/J.XPRO.2021.101005.

Evans, J. S., and M. A. Murphy. 2018. "*rfUtilities*." R package Version 2.1-3. https://cran.r-project.org/package=rfUtilities.

Eveleth, R., M.-L. Timmermans, and N. Cassar. 2014. "Physical and Biological Controls on Oxygen Saturation Variability in the Upper Arctic Ocean." *Journal of Geophysical Research, Oceans* 119: 7420–7432. https://doi.org/10.1002/2014JC009816.

Finley, A., S. Banerjee, and Ø. Hjelle. 2022. "*MBA: Multilevel B-Spline Approximation*." R Package Version 0.1-0. https://CRAN.R-project.org/package=MBA.

Fong, A. A., C. J. M. Hoppe, N. Aberle, et al. 2024. "Overview of the MOSAiC Expedition: Ecosystem." *Elementa: Science of the Anthropocene* 12: 00135. https://doi.org/10.1525/ELEMENTA.2023.00135.

Frick, H., F. Chow, M. Kuhn, M. Mahoney, J. Silge, and H. Wickham. 2025. rsample: General Resampling Infrastructure. R package version 1.3.0.9000, Posit. Boston Massachusetts. http://github.com/tidymodels/rsample

Garcia, H. E., and L. I. Gordon. 1992. "Oxygen Solubility in Seawater: Better Fitting Equations." *Limnology and Oceanography* 37: 1307–1312. https://doi.org/10.4319/lo.1992.37.6.1307.

Guidi, L., S. Chaffron, L. Bittner, et al. 2016. "Plankton Networks Driving Carbon Export in the Oligotrophic Ocean." *Nature* 532: 465–470. https://doi.org/10.1038/nature16942.

Guillou, L., D. Bachar, S. Audic, et al. 2013. "The Protist Ribosomal Reference Database (PR2): A Catalog of Unicellular Eukaryote Small Sub-Unit rRNA Sequences With Curated Taxonomy." *Nucleic Acids Research* 41: D597–D604. https://doi.org/10.1093/nar/gks1160.

Gutierrez, T., and M. D. Aitken. 2014. "Role of Methylotrophs in the Degradation of Hydrocarbons during the Deepwater Horizon Oil Spill." *ISME Journal* 8: 2543–2545. https://doi.org/10.1038/ISMEJ.2014.88.

Haas, C., M. Hoppmann, S. Tippenhauer, and G. Rohardt. 2021. Continuous Thermosalinograph Oceanography Along RV POLARSTERN Cruise Track PS122/2. PANGAEA. Germany. http://doi.org/10.1594/PANGAEA.930024

Haft, D. H., M. DiCuccio, A. Badretdin, et al. 2018. "RefSeq: An Update on Prokaryotic Genome Annotation and Curation." *Nucleic Acids Research* 46, no. D1: D851–D860. https://doi.org/10.1093/nar/gkx1068.

Hamme, R. C., and S. R. Emerson. 2004. "The Solubility of Neon, Nitrogen and Argon in Distilled Water and Seawater." *Deep Sea Research, Part I: Oceanographic Research Papers* 51: 1517–1528. https://doi.org/10.1016/j.dsr.2004.06.009.

Hersbach, H., B. Bell, P. Berrisford, et al. 2023. ERA5 Monthly Averaged Data on Single Levels From 1940 to Present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Brussels. https://doi.org/10.24381/cds.f17050d7.

Hoppe, C. J. M., J. Creamean, J. S. Bowman, et al. 2023*b*. Year-Round Discrete Underway Water Column Chlorophyll a Concentrations From the Central Arctic. PANGAEA. Germany. http://doi.org/10.1594/PANGAEA.962597.

Hoppe, C. J. M., L. Heitmann, T. Brenneis, et al. 2023*a*. Water Column Chlorophyll a concentrations During the MOSAiC Expedition (PS122) in the Central Arctic Ocean 2019–2020. PANGAEA. Germany https://doi.org/10.1594/PANGAEA.963277.

Hoppe, C. J. M., N. Fuchs, D. Notz, et al. 2024. "Photosynthetic Light Requirement Near the Theoretical Minimum Detected in Arctic Microalgae." *Nature Communications* 15: 1–10. https://doi.org/10.1038/s41467-024-51636-8.

Jakobsson, M., R. Mohammad, M. Karlsson, et al. 2024. "The International Bathymetric Chart of the Arctic Ocean Version 5.0." *Scientific Data* 11: 1–22. https://doi.org/10.1038/s41597-024-04278-w.

Juranek, L., T. Takahashi, J. Mathis, and R. Pickart. 2019. "Significant Biologically Mediated CO_2 Uptake in the Pacific Arctic During the Late Open Water Season." *Journal of Geophysical Research: Oceans* 124: 821–843. https://doi.org/10.1029/2018JC014568.

Kanzow, T., M. Hoppmann, S. Tippenhauer, and G. Rohardt. 2021. Continuous Thermosalinograph Oceanography Along RV POLARSTERN Cruise Track PS122/3. PANGAEA. Germany. https://doi.org/10.1594/PANGAEA.930026.

Kassambara, A. 2023*a*. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.2. https://rpkgs.datanovia.com/rstatix/.

Kassambara, A. 2023*b*. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.1. https://rpkgs.datanovia.com/ggpubr/.

Kolde, R. 2019. *pheatmap: Pretty Heatmaps*. R Package Version 1.0.12. https://github.com/raivokolde/pheatmap.

Kursa, M. B., and W. R. Rudnicki. 2010. "Feature Selection With the Boruta Package." *Journal of Statistical Software* 36, no. 11: 1–13. https://doi.org/10.18637/JSS.V036.I11.

Laukert, G., D. Bauch, B. Rabe, et al. 2025. "Dynamic Ice–Ocean Pathways Along the Transpolar Drift Amplify the

Dispersal of Siberian Matter." *Nature Communications* 16: 1–15. https://doi.org/10.1038/s41467-025-57881-9.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by randomForest." *R News* 2: 18–22. http://CRAN.R-project.org/doc/Rnews/.

Lin, Y., N. Cassar, A. Marchetti, C. Moreno, H. Ducklow, and Z. Li. 2017. "Specific Eukaryotic Plankton Are Good Predictors of Net Community Production in the Western Antarctic Peninsula." *Scientific Reports* 7: 1–11. https://doi.org/10.1038/s41598-017-14109-1.

MacGilchrist, G. A., A. C. Naveira Garabato, T. Tsubouchi, S. Bacon, S. Torres-Valdés, and K. Azetsu-Scott. 2014. "The Arctic Ocean Carbon Sink." *Deep Sea Research Part I: Oceanographic Research Papers* 86: 39–55. https://doi.org/10.1016/J.DSR.2014.01.002.

Mathis, M., F. Lacroix, S. Hagemann, D. M. Nielsen, T. Ilyina, and C. Schrum. 2024. "Enhanced CO_2 Uptake of the Coastal Ocean Is Dominated by Biological Carbon Fixation." *Nature Climate Change* 14: 373–379. https://doi.org/10.1038/s41558-024-01956-w.

Meredith, M., M. Sommerkorn, S. Cassotta, et al. 2019. "Polar Regions." In IPCC Special Report on the Ocean and Cryosphere in a Changing Climate, edited by H.-O. Pörtner et al., 203–320. Cambridge University Press. Cambridge, UK. https://doi.org/10.1017/9781009157964.005.

Mock, T., W. Boulton, J. P. Balmonte, et al. 2022. "Multiomics in the Central Arctic Ocean for Benchmarking Biodiversity Change." *PLoS Biology* 20: e3001835. https://doi.org/10.1371/JOURNAL.PBIO.3001835.

Nicolaus, M., D. K. Perovich, G. Spreen, et al. 2022. "Overview of the MOSAiC Expedition: Snow and Sea Ice." *Elementa* 10:000046. https://doi.org/10.1525/ELEMENTA.2021.000046.

Nixdorf, U., K. Dethloff, M. Rex, et al. 2021. MOSAiC Extended Acknowledgement. Zenodo. https://doi.org/10.5281/ZENODO.5179738.

Oksanen, J., G. L. Simpson, F. G. Blanchet, et al. 2015. "*vegan: Community Ecology Package*." R Package vegan, Version 2.2-1. https://vegandevs.github.io/vegan/.

Oziel, L., O. Gurses, S. Torres-Valdes, et al. 2025. "Climate Change and Terrigenous Inputs Decrease the Efficiency of the Future Arctic Ocean's Biological Carbon Pump." *Nature Climate Change* 15: 171–179. https://doi.org/10.1038/s41558-024-02233-6.

Piontek, J., L. Galgani, E. M. Nöthig, I. Peeken, and A. Engel. 2021. "Organic Matter Composition and Heterotrophic Bacterial Activity at Declining Summer Sea Ice in the Central Arctic Ocean." *Limnology and Oceanography* 66: S343–S362. https://doi.org/10.1002/LNO.11639.

Priest, T., W. von Appen, E. Oldenburg, et al. 2023. "Atlantic Water Influx and Sea-Ice Cover Drive Taxonomic and Functional Shifts in Arctic Marine Bacterial Communities." *ISME Journal* 17: 1612–1625. https://doi.org/10.1038/s41396-023-01461-6.

R Core Team. 2023. R: A Language and Environment for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Rabe, B., C. Heuzé, J. Regnery, et al. 2022. "Overview of the MOSAiC Expedition: Physical Oceanography." *Elementa: Science of the Anthropocene* 10, no. 1: 00062. https://doi.org/10.1525/ELEMENTA.2021.00062.

Rex, M., M. Hoppman, S. Tippenhauer, and G. Rohardt. 2021*a*. Continuous Thermosalinograph Oceanography Along RV POLARSTERN Cruise Track PS122/5. PANGAEA. Germany. http://doi.org/10.1594/PANGAEA.930028.

Rex, M., M. Hoppmann, S. Tippenhauer, and G. Rohardt. 2021*b*. Continuous Thermosalinograph Oceanography Along RV POLARSTERN Cruise Track PS122/4. PANGAEA. Germany. https://doi.org/10.1594/PANGAEA.930027.

Rex, M., M. Hoppmann, S. Tippenhauer, and G. Rohardt. 2021*c*. Continuous Thermosalinograph Oceanography Along RV POLARSTERN Cruise Track PS122/1. PANGAEA. Germany. http://doi.org/10.1594/PANGAEA.930023.

Rokitta, S., E. J. Chamberlain, A. D'Angelo, et al. Forthcoming "Baselines Matter: Mass Spectrometric Assessments of Biological O_2 Supersaturation ($\Delta O_2 : Ar$) Benefit From Two-Point Calibrations." *Analytical Chemistry*.

Rokitta, S. D., E. Chamberlain, J. S. Bowman, et al. 2024*a*. Oxygen and Argon Concentrations in Surface Waters During the MOSAiC Expedition in the Arctic Ocean, Core Parameters. PANGAEA. Germany. https://doi.org/10.1594/PANGAEA.971603.

Rokitta, S. D., E. Chamberlain, J. S. Bowman, et al. 2024*b*. "*Oxygen and Argon Concentrations in Surface Waters During the MOSAiC Expedition in the Arctic Ocean*." Full Raw Dataset. PANGAEA. Germany. https://doi.org/10.1594/PANGAEA.969630.

Royo-Llonch, M., P. Sanchez, C. Ruiz-Gonzalez, et al. 2021. "Compendium of 530 Metagenome-Assembled Bacterial and Archaeal Genomes From the Polar Arctic Ocean." *Nature Microbiology* 6: 1561–1574. https://doi.org/10.1038/s41564-021-00979-9.

Schulz, K., Z. Koenig, and M. Muilwijk. 2023. The Eurasian Arctic Ocean along the MOSAiC Drift (2019–2020): Core Hydrographic Parameters. Arctic Data Center. Santa Barbara, CA. https://doi.org/10.18739/A21J9790B.

Schulz, K., Z. Koenig, M. Muilwijk, et al. 2024. "The Eurasian Arctic Ocean Along the MOSAiC Drift in 2019–2020: An Interdisciplinary Perspective on Physical Properties and Processes." *Elementa: Science of the Anthropocene* 12: 00114. https://doi.org/10.1525/ELEMENTA.2023.00114.

Tippenhauer, S., M. Vredenborg, C. Heuze, et al. 2023*a*. Physical Oceanography Water Bottle Samples Based on Ship CTD During POLARSTERN Cruise PS122. PANGAEA. Germany. https://doi.org/10.1594/PANGAEA.959965.

Tippenhauer, S., M. Vredenborg, C. Heuze, et al. 2023*b*. Physical Oceanography Water Bottle Samples Based on Ocean City CTD During POLARSTERN Cruise PS122. PANGAEA. Germany. https://doi.org/10.1594/PANGAEA.959966.

Ulfsbo, A., N. Cassar, M. Korhonen, et al. 2014. "Late Summer Net Community Production in the Central Arctic Ocean Using Multiple Approaches." *Global Biogeochem Cycles* 28: 1129–1148. https://doi.org/10.1002/2014GB004833.

Ulfsbo, A., E. S. Droste, and S. Torres-Valdés. 2023. Dissolved Oxygen of Seawater Samples During RV POLARSTERN Expedition PS122 MOSAiC. PANGAEA. Germany. https://doi.org/10.1594/PANGAEA.955092.

Underwood, G. J. C., C. Michel, G. Meisterhans, et al. 2019. "Organic Matter From Arctic Sea-Ice Loss Alters Bacterial Community Structure and Function." *Nature Climate Change* 9: 170–176. https://doi.org/10.1038/s41558-018-0391-7.

Volk, T., and M. I. Hoffert. 1985. "Ocean Carbon Pumps: Analysis of Relative Strengths and Efficiencies in Ocean-Driven Atmospheric CO_2 Changes." In The Carbon Cycle and Atmospheric CO2: Natural Variations Archean to Present. American Geophysical Union. Washington D.C., 99–110. https://doi.org/10.1029/GM032P0099.

Walters, W., E. R. Hyde, D. Berg-Lyons, 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. https://doi.org/10.1128/msystems.00009-15.

Wang, S., Y. Lin, S. Gifford, R. Eveleth, and N. Cassar. 2018. "Linking Patterns of Net Community Production and Marine Microbial Community Structure in the Western North Atlantic." *ISME Journal* 12: 2582–2595. https://doi.org/10.1038/s41396-018-0163-4.

Wehrens, R., and J. Kruisselbrink. 2018. "Flexible Self-Organizing Maps in kohonen 3.0." *Journal of Statistical Software* 87: 1–18. https://doi.org/10.18637/JSS.V087.I07.

Weissman, J. L., S. Hou, and J. A. Fuhrman. 2021. "Estimating Maximal Microbial Growth Rates From Cultures, Metagenomes, and Single Cells Via Codon Usage Patterns." *Proceedings of the National Academy of Sciences of the United States of America* 118: e2016810118. https://doi.org/10.1073/pnas.2016810118.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.

Worden, A. Z., M. J. Follows, S. J. Giovannoni, S. Wilken, A. E. Zimmerman, and P. J. Keeling. 2015. "Rethinking the Marine Carbon Cycle: Factoring in the Multifarious Lifestyles of Microbes." *Science* 347: 1257594. https://doi.org/10.1126/SCIENCE.1257594.

Yasunaka, S., M. Manizza, J. Terhaar, et al. 2023. "An Assessment of CO_2 Uptake in the Arctic Ocean From 1985 to 2018." *Global Biogeochemical Cycles* 37: e2023GB007806. https://doi.org/10.1029/2023GB007806.

Zhou, T., Y. Li, Z. Ouyang, W. J. Cai, and R. Ji. 2024. "Enhanced Net Community Production With Sea Ice Loss in the Western Arctic Ocean Uncovered by Machine-Learning-Based Mapping." *Geophysical Research Letters* 51: e2024GL110931. https://doi.org/10.1029/2024GL110931.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Submitted 11 November 2024
Revised 01 May 2025
Accepted 24 June 2025