Environmental Microbiology

ENVIRONMENTAL MICROBIOLOGY

RESEARCH ARTICLE

Marine Bacterioplankton Composition Predicts Oxygen Consumption During Dissolved Organic Matter Degradation Experiments

Cecilia Alonso[1] | Juan Zanetti[1] | Luciana Griffero[1] | Emiliano Pereira-Flores[1] | Belén González[2] | Carolina Lescano[3] | Andrés Pérez-Parada[4] | Carolina Crisci[5] | Rudolf Amann[6]

[1]Microbial Ecology of Aquatic Ecosystems. Interdisciplinary Department of Coastal and Marine Systems. Centro Universitario Regional del Este, Universidad de la República, Rocha, Uruguay | [2]PEDECIBA Post-Graduate Program. Centro Universitario Regional del Este, Universidad de la República, Rocha, Uruguay | [3]Interdisciplinary Department of Coastal and Marine Systems. Centro Universitario Regional del Este, Universidad de la República, Rocha, Uruguay | [4]Technological Development Department, Centro Universitario Regional del Este, Universidad de la República, Rocha, Uruguay | [5]Department of Statistical Data Modelling and Artificial Intelligence, Centro Universitario Regional del Este, Universidad de la República, Rocha, Uruguay | [6]Molecular Ecology Department, Max Planck Institute for Marine Microbiology, Bremen, Germany

Correspondence: Cecilia Alonso (calonso@cure.edu.uy)

Received: 25 April 2025 | **Revised:** 1 September 2025 | **Accepted:** 15 October 2025

Funding: This work was supported by the Agencia Nacional de Investigación e Innovación (grant ANII_MPI_ID_2017_1_1007663) and the Comisión Sectorial de Investigación Científica (grant Grupos I+D 2022).

Keywords: coastal ecosystems | dissolved organic matter | microbial indicators | microbial-driven processes | statistical modelling

ABSTRACT

Microbial communities play pivotal roles in ocean biogeochemistry, yet linking their composition to ecosystem functions remains a significant challenge. In this study, we demonstrate the predictive power of bacterioplankton taxonomic composition in explaining oxygen consumption during dissolved organic matter (DOM) degradation. Using 4 years of experimental data, we integrated 'omics with statistical modeling, applying feature selection and dimensionality reduction to develop high-performance linear regression models with strong predictive accuracy. Our framework also identifies key microbial groups driving oxygen consumption, including taxa known for their differential capabilities in DOM processing and recently shown to exhibit distinct respiration rates. Flavobacteriales emerge as central contributors to oxygen consumption, underscoring their ecological importance in nutrient-rich, highly productive coastal systems often referred to as 'green seas'. Their consistent dominance across varying oxygen consumption categories highlights their pivotal role in sustaining ecosystem functions in these environments. Beyond oxygen consumption, this framework provides a versatile tool for investigating microbially driven biogeochemical processes. By linking community composition with ecosystem functions, our study advances predictive microbial ecology. These findings deepen our understanding of microbial contributions to the ocean's carbon and oxygen cycles, improving our ability to anticipate their responses to environmental change.

Luciana Griffero and Emiliano Pereira-Flores contributed equally to this work.

1 | Introduction

Massive sequencing technologies and 'omics' techniques have unveiled the vast taxonomic and functional diversity of marine microbial communities, laying the groundwork for developing predictive models of microbially driven processes. However, integrating this complexity into robust models remains a pressing challenge amid rapid global environmental change (Li et al. 2025).

The microbial degradation of dissolved organic matter (DOM) is a cornerstone of marine biogeochemistry, fueling carbon turnover and driving oxygen consumption across marine ecosystems (Kujawinski 2011). Heterotrophic bacterioplankton transform DOM through biomass incorporation, remineralisation to CO_2 and the production of refractory compounds (Jiao et al. 2010), regulating carbon and oxygen dynamics in the ocean (Robinson 2019). Coastal environments, in particular, act as biogeochemical hotspots, where microbial-mediated DOM processing plays a crucial role in shaping ecosystem function (Hedges et al. 1997; Breitburg et al. 2018).

Understanding the microbial mechanisms underpinning these processes requires integrating information on community composition with direct measurements of biogeochemical rates. Historically, microbial community profiling and physiological rate measurements have been applied independently, yet their integration is now recognised as essential for improving predictive models of ecosystem function (Strzepek et al. 2022). Recent regression-based approaches have successfully identified microbial functional groups in relation to steady-state variables (e.g., nitrate concentration in the TARA dataset) (Shan et al. 2023) or theoretical simulations of resource utilisation (Zhao et al. 2024), highlighting the robustness of relatively straightforward statistical frameworks. However, their direct application to predicting microbially driven processes from empirical data remains largely unexplored, particularly in marine biogeochemistry.

Here, we present a quantitative framework to predict oxygen consumption during DOM degradation from microbial community composition and to identify the bacterioplankton taxa driving this process, integrating microbial taxonomic and functional profiling, environmental characterisation and statistical modelling This approach draws on 50 experiments conducted over 4 years at the South Atlantic Microbial Observatory (SAMO), located in a region recognised as a global warming hotspot.

2 | Methods

2.1 | Sampling

Sampling was conducted at the South Atlantic Microbial Observatory (SAMO, 34° 42′ 43.36″ S, 54° 14′ 08.64″ W), a coastal site within a national protected area and part of international initiatives such as Ocean Sampling Day and the AMOLat network (Kopf et al. 2015; Fermani et al. 2024). Subsurface water samples were collected using acid-washed carboys during 50 sampling campaigns from March 2018 to October 2021.

2.2 | Environmental Characterisation

Physicochemical parameters (conductivity, turbidity, temperature, dissolved oxygen, salinity, total dissolved solids, density and pH) were measured in situ with a Horiba multiparameter sensor. Chlorophyll and phycocyanin fluorescence were assessed with a Turner fluorometer, and light penetration with a Secchi disc. Water samples were frozen at −20°C for nutrient analysis following standard colorimetric protocols (APHA 1995) to quantify silica (Müllin and Riley 1955), ammonium (Koroleff 1970), nitrate (Mackereth et al., Mackereth et al. 1978), nitrite (Bendschneider and Robinson 1952), phosphate (Murphy and Riley 1962), total nitrogen (Valderrama 1981 and then Mackereth et al. 1978) and total phosphorus (Valderrama 1981, and then Murphy and Riley 1962) with a UV–Vis Genesys 150 spectrophotometer (Thermo Fisher Scientific). Chlorophyll-a was determined spectrophotometrically after acetone extraction (Lorenzen 1967; Parsons et al. 1984).

2.3 | Dissolved Organic Matter Quality

Samples for DOM characterisation were filtered through precombusted GF/F filters (0.7 μm, 450°C, 4 h) and stored frozen. Absorbance spectra (240–800 nm) were measured using a Perkin Elmer Lambda 35 spectrophotometer, with a 1 cm quartz cuvette and Milli-Q water blank; values were blank-corrected and converted to Napierian absorption coefficients (Helms et al. 2008).

Fluorescence properties were analysed using a Fluoromax 4 (Horiba) spectrofluorometer, acquiring excitation-emission matrices (EEMs) for emission wavelengths from 280 to 600 nm (2 nm steps) and excitation wavelengths from 240 to 450 nm (5 nm steps). The *StaRdom* R package (Pucher et al. 2019) was employed for parallel factor analysis (PARAFAC) (Murphy et al. 2013), with spectra corrected for inner-filter effects and with instrument-specific correction factors. Components were standardised to Raman units (nm^{-1}) and validated via split-half analysis, random initialization and examination of the model's residuals (Murphy et al. 2013). Identified components were compared against the OpenFluor database (Murphy et al. 2014).

DOM composition was further analysed via high-performance liquid chromatography coupled with fluorescence detection (HPLC-FLD) (Li et al. 2013) and UV detection, following acidification (pH 2, HCl), solid-phase extraction (Oasis HLB cartridges) and methanol elution (Dittmar et al. 2008).

Chromatographic separations were conducted on a Thermo Scientific Hypersil Gold C18 column using reversed-phase chromatography (RPC) in a Thermo Scientific Ultimate 3000 System equipped with a sequential Diode Array Detector (DAD) 3000DAD and a 3400RS Fluorescence Detector (FLD). The mobile phase consisted of (A) 0.1% formic acid in water and (B) 0.1% formic acid in LC-grade acetonitrile. The gradient increased from 10% B to 25% B at 10 min, to 95% B at 18 min, held for 2 min; the flow rate was 1 mL/min, injection volume 10 μL. Hydrophilic and hydrophobic fractions were classified based on retention time (Li et al. 2013).

FLD operated in emission scan mode with fast scan speed (λ_exc: 240 nm, λ_em: 290–600 nm) and DAD scanned 220–800 nm. Peaks identified by FLD within the emission maxima of PARAFAC components were quantified. The DAD targeted UV absorption peaks at 254, 273, 280 and 290 nm—indicative of aromatic compounds, humic substances, lignin derivatives (254 nm), polyaromatic compounds, polyphenols and amino acids or protein-bound aromatic residues (273 and 280 nm) and oxidised aromatic compounds and extended conjugation systems (290 nm) (Weishaar et al. 2003; Thomas and Burgess 2007). Instrument control and data processing were performed using Chromeleon v.7.2.9 (Thermo Scientific). Full DOM characterisation results are presented in Data S1.

2.4 | Bacterioplankton Taxonomic and Functional Composition

Water samples were sequentially filtered through 25 μm and 0.2 μm cellulose filters; the latter, containing the microbial biomass were stored at −80°C until DNA extraction, using a manual protocol (Alonso et al. 2010). Bacterioplankton taxonomic composition was determined by amplifying a fragment of the V4-V5 hypervariable regions of the 16S rRNA gene with primers 515F-Y (5′-GTG YCA GCM GCC GCG GTA A-3′) and 926R (5′-CCG YCA ATT YM TTT RAG TTT-3′) (Parada et al. 2016), sequenced on an Illumina MiSeq at the Integrative Genomics Core (City of Hope, US).

Amplicon sequences were pre-processed using bbduk (http://seqanswers.com/forums/showthread.php?t=42776) and cutadapt (https://cutadapt.readthedocs.io/en/stable). The DADA2 R package (Callahan et al. 2016) was then used for quality filtering and ASV inference, following these steps: (1) Trimming R1 and R2 reads to 220 and 175 bp, respectively and removing reads with >2 expected errors; (2) Modeling error rates; (3) Dereplicating reads; (4) inferring ASVs using the dada function with the pool option; (5) Merging paired-end reads with a minimum overlap of 12 nucleotides; and (6) Removing chimera sequences. ASVs were taxonomically annotated using the Naive Bayes Classifier in DADA2, with SILVAv138.1 NR 99 as the reference database (https://www.arb-silva.de) (Quast et al. 2013). Pre-processing tasks were automated through a custom pipeline available at https://github.com/pereiramemo/amplicon_pipelines.

For functional characterisation, metagenomic DNA was sequenced on the Illumina NovaSeq 6000 SP FC platform at LGC Genomics GmbH (Berlin, Germany) and the Genomics and Cell Characterisation Core Facility (GC3F) at the University of Oregon (Eugene, US). Raw metagenomic sequences were pre-processed by merging paired-end reads using Pear (https://cme.h-its.org/exelixis/web/software/pear), with a minimum overlap of 12, trimming low-quality regions (Q < 20) and discarding reads shorter than 50 bp with bbduk, The complete custom pipeline is available at https://github.com/pereiramemo/metagenomic_pipelines.

Functional annotation was performed by predicting open reading frames (ORFs) with FragGeneScan (https://omics.informatics.indiana.edu/FragGeneScan) and comparing translated sequences to the dbCAN database of HMM profiles for carbohydrate-active enzymes (CAZymes) (Drula et al. 2022) using hmmsearch (Eddy 2011).

To evaluate the commensurability of amplicon- and metagenome-derived taxonomic profiles, we compared community composition across multiple taxonomic ranks after applying harmonised filtering criteria and sequencing depth normalisation. A detailed description of the methodology, together with results on microbial taxonomic and functional composition—including amplicon versus metagenome comparisons—is provided in Data S2.

2.5 | DOM Degradation Experiments

Fifty DOM biodegradation experiments were conducted between March 2018 and October 2021. Water samples were prefiltered through 1.6 μm GF filters to reduce eukaryotic biomass, then incubated in 1 L airtight glass bottles, sealed with plastic film and caps. Incubations were performed in triplicate, along with a killed control, for ~1 week in a dark environmental chamber at in situ temperature recorded at the time of sampling (ranging from 11.1°C to 27.8°C).

Oxygen concentrations were recorded every 15 min using non-invasive optical sensors attached to the inner bottle walls, connected to a 4-channel controller/transmitter (OXY-mini, PreSens Precision Sensing GmbH, Regensburg, Germany). Oxygen consumption curves were analyzed to determine a standardised comparison time point across experiments, selecting 120 h as the reference when consumption generally stabilised.

Oxygen consumption in mg served as the response variable in regression models and the percentage of initial oxygen consumed was used in WARD hierarchical clustering (Euclidean distance) to classify experiments by consumption patterns. These clusters were subsequently used to identify bacterial indicators of different consumption levels.

2.6 | Variable Selection for Modelling

To identify the most informative components of the microbial community, we applied a sequential strategy involving feature selection followed by dimensionality reduction through ordination analysis (Figure 1).

First, ASVs with a minimum abundance of 0.005% were retained ($ASV_{0.005}$ dataset). Sparse partial least squares regression (sPLSR) analysis (Lê Cao et al. 2008), implemented in the mixOmics R package (Rohart et al. 2017) was then used to identify ASVs most associated with oxygen consumption (ASV_{sel} dataset), which were aggregated at the genus level (GEN_{sel} dataset). A similar sPLSR approach was applied to CAZyme annotations to define the CAZ_{sel} dataset.

Dimensionality reduction was performed using PCA, PCoA and NMDS with the vegan R package. (Oksanen et al. 2012). ASV and CAZ datasets were Hellinger-transformed prior to analysis (Legendre and Gallagher 2001); Bray-Curtis distances were used for PCoA and NMDS. Environmental variables (ENV dataset,

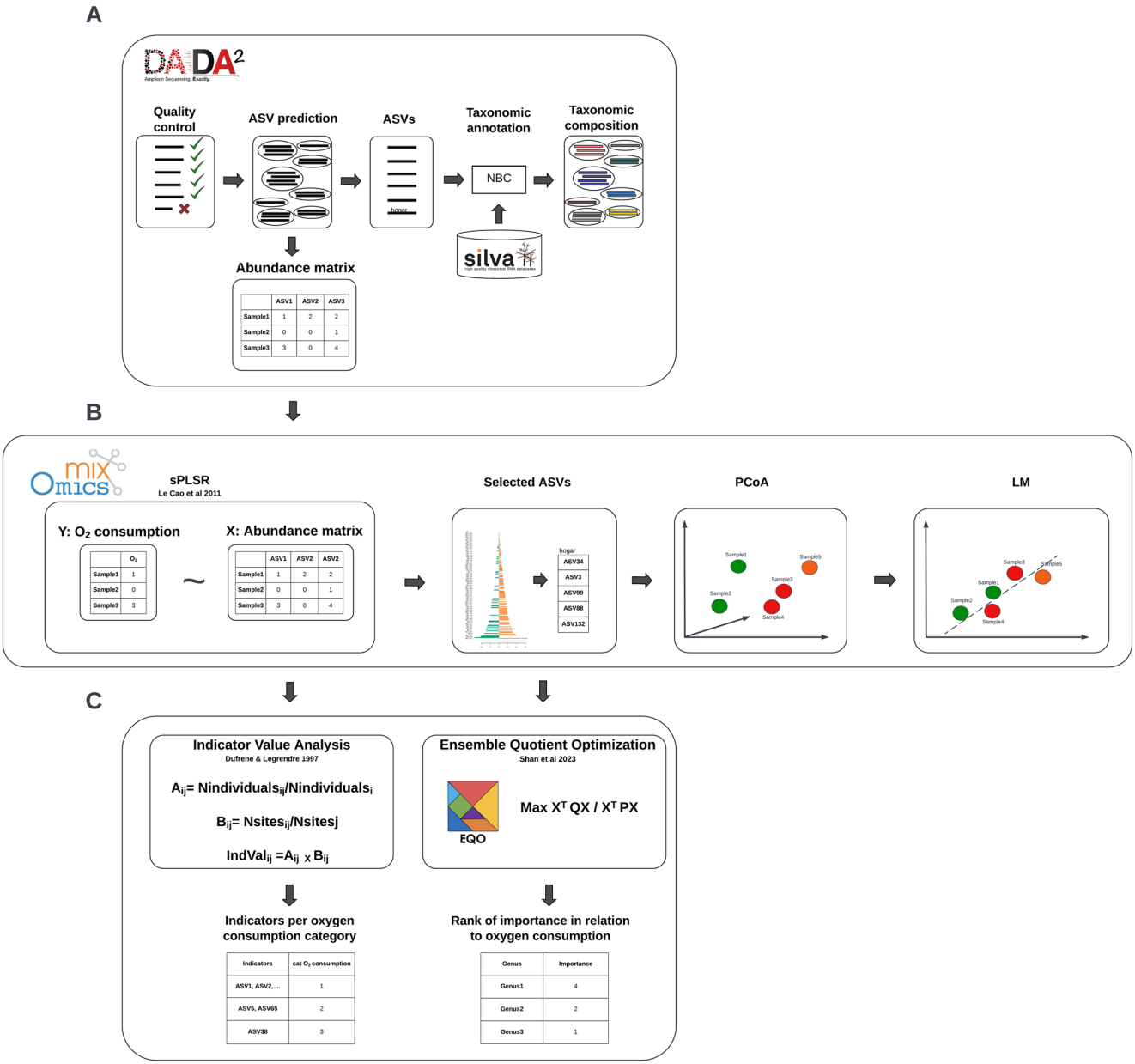

FIGURE 1 | Workflow for variable selection, reduction and identification of key microbial groups related to oxygen consumption. (A) Generation of an ASV table from amplicon sequences using DADA2 with the Silva database as a reference for taxonomic annotation. The table is filtered to retain the most abundant ASVs. (B) Selection of the most relevant ASVs associated with oxygen consumption through sparse Partial Least Squares Regression (sPLSR) implemented in MixOmics. Principal Coordinate Analysis (PCoA) is then performed, and the extracted PCoA axes are used as predictive variables in linear models. (C) Identification of indicator taxa from the most abundant ASVs matrix using Indicator Value Analysis (IndVal) to differentiate categories of oxygen consumption. From the sPLSR-selected ASVs matrix, Ensemble Quotient Optimization (EQO) is applied to identify key taxa associated with continuous oxygen consumption values.

$n = 52$) were reduced to 30 after filtering for multicollinearity (correlation analysis, VIF) and PCA was run on the standardized matrix. Axes for modeling were selected to explain ≥80% of variance (max. 15, axes) for PCA and PCoA, whereas for NMDS, the number of axes was determined by scree plots, using a stress threshold of 0.05 (Tables S1, S4).

2.7 | Data Exploration and Statistical Modelling

Data exploration and statistical modelling followed established guidelines (Zuur et al. 2010; Zuur and Ieno 2016) including identifying outliers, examining the distribution of the response variable, assessing collinearity and interactions and testing for temporal dependence.

Oxygen consumption was log-transformed to approximate normality. Linear Regression Models (LMs) were built separately for ASV_{sel}, GEN_{sel}, CAZ_{sel} and ENV datasets, using retained ordination axes as predictors.

Backward selection was applied using ANOVA tests, retaining only statistically significant variables. Model selection was refined based on the Akaike Information Criterion (AIC)

(Akaike 1973) and Bayesian Information Criterion (BIC) (Schwarz 1978). Diagnostic plots were visually inspected to identify potential influential points and to assess homogeneity and normality of residuals—supplemented by Shapiro-Wilk and Breusch–Pagan tests. Model evaluation was carried out using the R packages performance (Lüdecke, Ben-Shachar, et al. 2021), see (Lüdecke, Patil, et al. 2021), patchwork (Pedersen 2024a) and car (Fox and Weisberg 2019).

Since oxygen consumption was well explained by LMs within each dataset, Lasso regression was applied to integrate them and enhance predictive accuracy. While ordination axes are orthogonal within datasets, combining them may introduce multicollinearity by capturing overlapping gradients. Lasso mitigates this through coefficient regularisation and sparsity constraints, reducing overfitting in high-dimensional data (Tibshirani 1996).

Predictive accuracy of LM and Lasso models was assessed using Leave-One-Out Cross-Validation (LOOCV), comparing root mean square error (RMSE) and Pearson correlations between observed and predicted values.

2.8 | Identification of Key Taxa

To identify microbial taxa most associated with oxygen consumption, we applied the Ensemble Quotient Optimization (EQO) approach for functional group discovery (Shan et al. 2023) using the mEQO R package. EQO was performed at the genus level, as originally implemented (Shan et al. 2023), to optimise computational time. From the Gen_{sel} dataset, ensembles of genera maximising correlation with oxygen consumption were identified using the EQ_optim function with a genetic algorithm, testing from 2 to 24 genera (in steps of 2) and selecting the optimal model based on R^2 improvement. Model robustness was assessed by cross-validation (100 iterations of 80:20 train-test splits) and cumulative R^2 of individual and paired genera was used to determine their relative importance. Analyses were conducted using mEQO (Shan et al. 2023), GA (Scrucca 2013) and AICcmodavg (Mazerolle 2023). Network visualisation was performed with ggraph (Pedersen 2024b) and tidygraph (Pedersen 2024c).

Indicator Species Analysis (IndVal) (Dufrene and Legendre 1997) was conducted to identify taxa indicative of each oxygen-consumption category following Alonso et al. (2022) with the indicspecies R package (De Cáceres and Legendre 2009). From the $ASV_{0.005}$ dataset, candidate ASVs were identified for each oxygen-consumption category by selecting those with a frequency threshold value (Bt) of at least 0.5 (i.e., present in 50% of samples within the target category). The indicators function was used to evaluate the predictive value of combinations of up to three ASVs. From which optimal indicators were selected by maximising their positive predictive value (component A).

The same analysis was repeated using genus-level abundances derived from the $ASV_{0.005}$ dataset ($Gen_{0.005}$ dataset). The predictive accuracy of all selected indicators was validated using the predict function, with LOOCV, as recently implemented in the indicspecies package (Alonso et al. 2022).

All visualisations were produced using the ggplot2 R package (Wickham 2016), for graphs with a wide colour range, the Viridis palette was chosen for accessibility (Garnier et al. 2024). All analyses were conducted in R version 4.2.2 Patched and RStudio version 2023.09.1 + 494 (R Development Core Team 2011).

3 | Results

3.1 | Oxygen Consumption

Oxygen consumption during the experiments from 0.82 to 11.01 mg (median: 2.14 mg). WARD hierarchical clustering identified four distinct experiment clusters (Figure 2A) that significantly differed in oxygen consumption levels (Welch's ANOVA test, $F = 58.548$, $p = 0.0006$; Figure 2B).

3.2 | Feature Selection and Dimensionality Reduction

The sPLSR analysis on the $ASV_{0.005}$ dataset identified the optimal number of components and selected ASVs that maximised the co-variance between bacterioplankton composition and oxygen consumption (Figure 3A). From these results, 298 ASVs from the first two components were retained as the ASV_{sel} dataset. Similarly, for CAZymes, 450 out of 543 CAZymes were selected by the first sPLSR component, defining the CAZ_{sel} dataset (Figure 3B).

Bray-Curtis-based Principal Coordinates Analysis (PCoA) ordinations of ASV and CAZ datasets (Figure 4) show that the ASV_{sel} dataset achieves a slightly better sample separation along PC1 compared to $ASV_{0.005}$ (Figure 4A,B). A similar improvement is observed between CAZ_{all} and CAZ_{sel} (Figure 4C,D). These enhancements align with the higher variance explained by the selected datasets, although a clear separation of oxygen consumption categories is not evident in either case.

A summary of a comparison of the linear regression models obtained for the $ASV_{0.005}$ and CAZ_{all} datasets with different ordination techniques is shown in Table S2. Model performance was assessed using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and adjusted R^2. PCoA-based models generally outperformed PCA-based ones, while NMDS-based models performed poorly for ASVs.

Given these results, further linear regression models were constructed using PCoA ordination axes for taxonomic (ASV and GEN) and functional (CAZ) datasets. Environmental variable models were based on PCA axes from the standardised ENV matrix, as PCA is well-suited for continuous positive variables, and standardisation ensures all variables contribute equally, regardless of their original units or magnitude of variation.

3.3 | Linear Regression Model Performance

A summary of the linear regression models comparison for the ASV, GEN_{sel}, CAZ and ENV datasets is shown in Table 1. Feature selection notably improved model performance, with

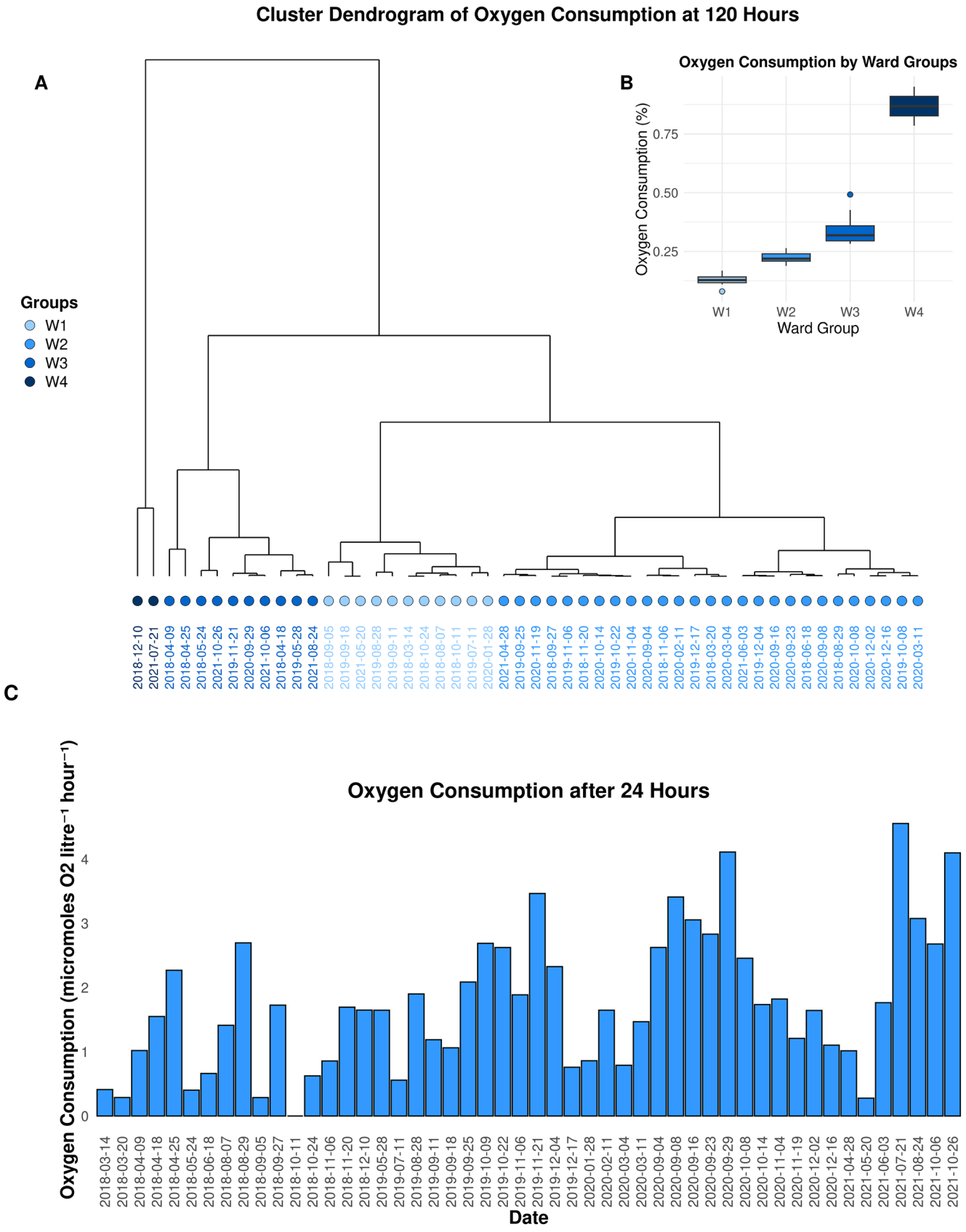

FIGURE 2 | Oxygen consumption during DOM degradation experiments. (A) Ward clustering of the percentage of initial oxygen consumed at 120 h of incubation, using Euclidean distance. (B) Oxygen consumption grouped by Ward clusters. The boxplots represent the distribution of oxygen consumption for each predefined Ward group, identified in the clustering analysis. Each box shows the interquartile range (IQR), with the horizontal line representing the median value. Whiskers extend to 1.5 times the IQR, and points outside this range indicate outliers. (C) Oxygen consumption (millimoles per litre per hour) measured at the 24-h mark for each experiment.

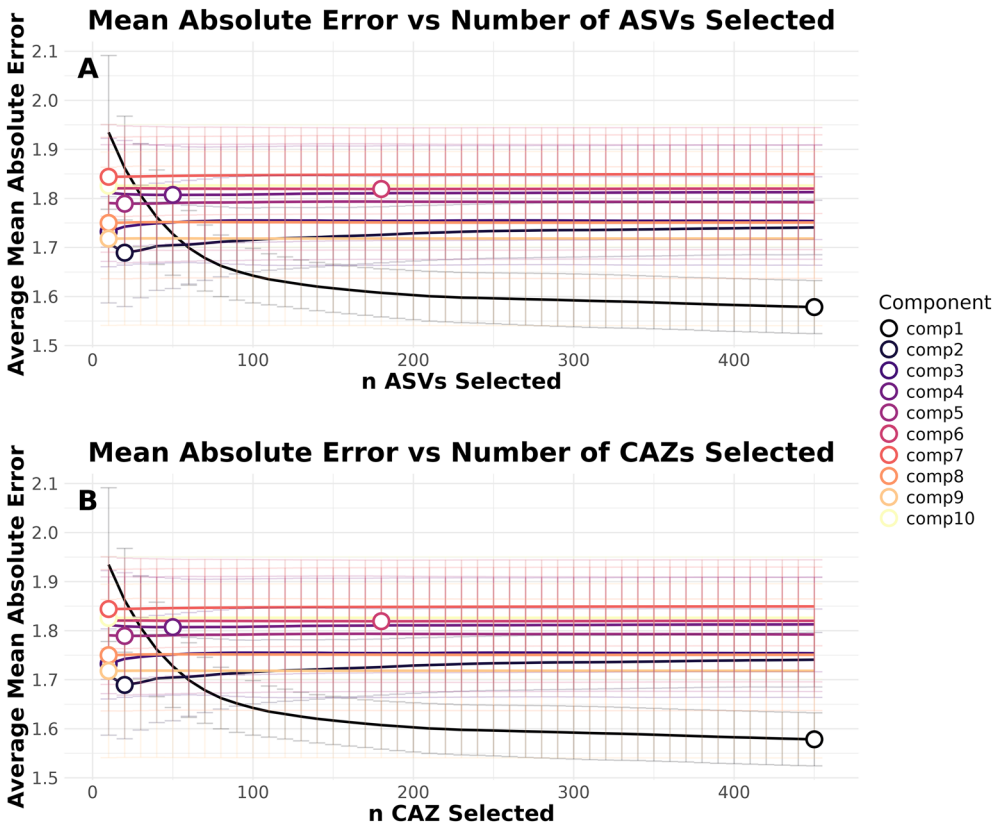

FIGURE 3 | Feature selection using sparse Partial Least Squares Regression (sPLSR). The number of variables (features) used to construct each latent component was determined using the tune.spls function from the mixOmics R package. Components were tuned iteratively, selecting the optimal number of variables (keepX) based on the lowest Mean Square Error (MSE). The error rate was averaged across cross-validated folds and repeats, with standard deviations represented by error bars. The bubbles indicate the optimal keepX values that achieved the lowest classification error rates for each component, determined using a one-sided *t*-test. Panel A show the Results for the ASV dataset, and Panel B for the CAZ dataset.

ASV_{sel}- and CAZ_{sel}-based models showing higher adjusted R^2 values and lower AIC/BIC scores than $ASV_{0.005}$ and CAZ_{all} models, respectively. Within taxonomy-based datasets, ASV_{sel} outperformed GEN_{sel} in explanatory power.

When modelled independently, bacterioplankton taxonomy explained 68% of the variance in oxygen consumption during DOM degradation experiments, CAZymes 36% and environmental variables 11%, reflecting the relative explanatory power of each dataset. Detailed model results are provided in Data S3.

Since oxygen consumption was effectively modeled using linear regressions within each dataset, Lasso regression was applied to integrate them and potentially enhance predictive accuracy.

In the first analysis (*lassoaxes*), we used ordination axes explaining ≥80% of variance in ASV_{sel}, CAZ_{sel} and ENV datasets (38 axes total), retaining 21 in the final model (Table S3A). In the second analysis (*lassoraw*), Lasso was applied directly to raw data (298 ASVs, 450 CAZymes, 30 environmental variables), retaining 43 variables (Table S3B).

Predictive performance was evaluated using leave-one-out cross-validation (LOOCV), comparing root mean square error (RMSE) and Pearson correlations between observed and predicted values for both Lasso models and the linear regression models based on ASV_{sel} and GEN_{sel} datasets.

The *lassoaxes* model outperformed *lassoraw*, confirming that dimensionality reduction improves predictive accuracy while maintaining model stability. Notably, ASV-based and GEN-based linear models performed comparably to *lassoaxes*, underscoring the high predictive power of bacterioplankton taxonomic composition alone (Figure 5).

3.4 | Identification of Key Groups

Ensemble Quotient Optimization (EQO) on the GEN_{sel} dataset identified combinations of up to 18 genera as optimal for maximising correlation with oxygen consumption, beyond which R^2 plateaued (Figure S1). The importance of individual genera and their pairwise associations is illustrated in an aggregation network, with node size and edge width representing their relative significance (Figure 6). Flavobacteriales was the most highly represented order, followed by Rhodobacterales. Together, these groups accounted for ca. 30% of the genera identified by EQO, while the contribution of each of the remaining orders did not exceed 5%. Indicator Species Analysis (IndVal) applied to the $ASV_{0.005}$ dataset identified taxa indicative of each of the oxygen consumption categories—W1 (low), W2 (moderate), W3 (high) and W4 (very high)—as previously defined by Ward clustering. Many ASVs selected as indicators were also highlighted by EQO (bold in Table 2A). A genus-level IndVal analysis (Table 2B) similarly retrieved several genera

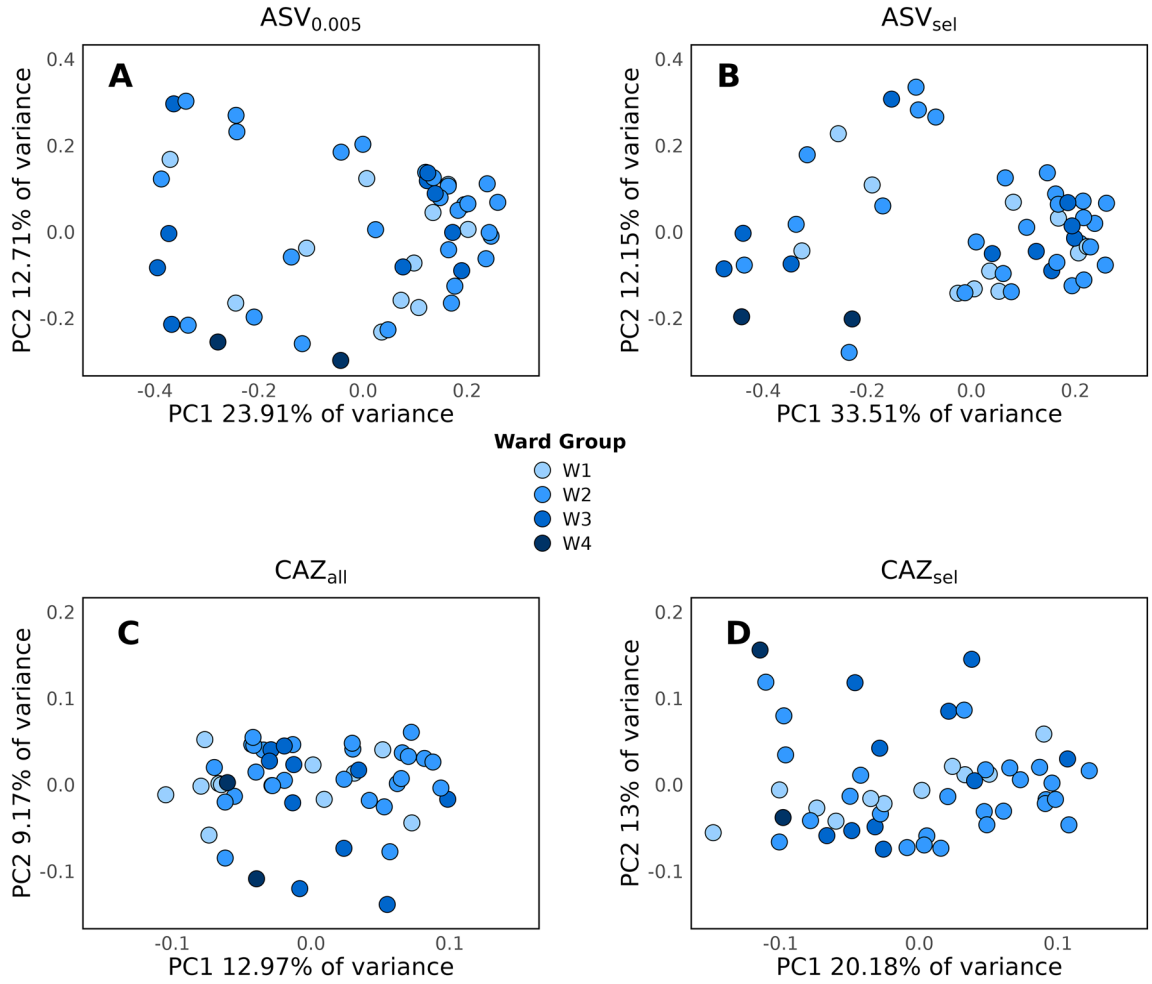

FIGURE 4 | Principal Coordinate Analysis (PCoA) ordination plots.

The PCoA plots represent the dissimilarity between samples, calculated using Bray-Curtis distances on Hellinger-transformed data. Panels correspond to different datasets: (A) $ASV_{0.005}$, (B) ASV_{sel}, (C) CAZ_{all} and (D) CAZ_{sel}. Each point represents a sample, and the distances between points reflect their compositional differences. The axes (PCoA1 and PCoA2) show the percentage of variance explained by the respective principal coordinates. Points are coloured according to oxygen consumption categories, defined via WARD clustering of oxygen consumption at 120 h of incubation (W1 = low, W2 = moderate, W3 = high, W4 = very high).

TABLE 1 | Comparison of model performance metrics across independent datasets. The table summarises the number of features, components, and model performance metrics—Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and explained variance (adjusted R^2)—for the full and optimised models across six datasets: $ASV_{0.005}$, ASV_{sel}, GEN_{sel}, CAZ_{all}, CAZ_{sel}, (PCoA axes) and ENV (PCA axes) in explaining oxygen consumption during DOM degradation experiments.

	$ASV_{0.005}$	ASV_{sel}	GEN_{sel}	CAZ_{all}	CAZ_{sel}	ENV
n features	1160	298	106	543	450	30
n components full model	14	11	10	15	14	13
n components optimised model	4	8	5	3	3	2
AIC full model	53	20	35	69	59	77
AIC optimised model	43	18	27	53	47	63
BIC full model	84	45	58	101	89	106
BIC optimised model	55	37	41	62	57	70
R^2_{adj} full model	0.39	0.68	0.56	0.18	0.32	0.004
R^2_{adj} optimised model	0.42	0.68	0.59	0.29	0.36	0.11

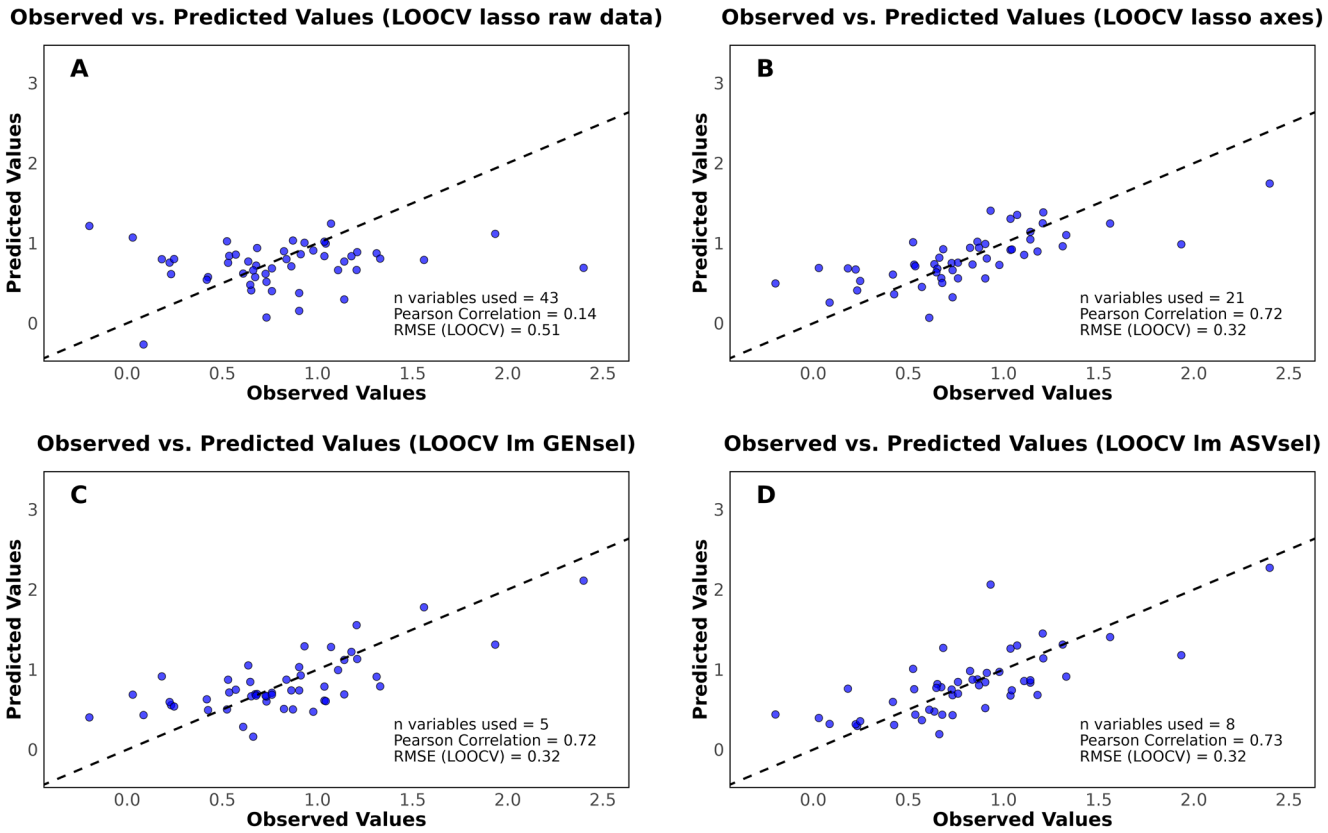

FIGURE 5 | Predictive performance of Lasso and linear models using LOOCV. Scatterplots compare observed versus predicted values for different regression models evaluated using Leave-One-Out Cross-Validation (LOOCV):(A) Lasso model applied to raw data from the ASV_{sel}, CAZ_{sel} and ENV datasets, (B) Lasso model applied to ordination axes of the ASV_{sel}, CAZ_{sel} and ENV datasets, (C) Linear model applied to ordination axes of the GEN_{sel} dataset, (D) Linear model applied to ordination axes of the ASV_{sel} dataset. The dashed line represents the 1:1 relationship between observed and predicted values. The Pearson correlation coefficient and Root Mean Square Error (RMSE) are reported for each model, along with the number of variables included.

identified by EQO (bold), reinforcing their role in oxygen consumption patterns. Flavobacteriales accounted for approximately 20% of the indicator taxa, followed by Rhodobacterales with about 10%. Other relevant orders included Cytophagales, Alteromonadales and Burkholderiales, each contributing up to 8% of the indicator taxa.

ASV-based indicators achieved perfect classification of oxygen consumption categories using leave-one-out cross-validation (LOOCV, Figure 7A), whereas genus-based indicators correctly classified 84% of samples, with misclassifications mainly between W2 and W3 (Figure 7B).

Correlation analysis between the ASV_{sel} and CAZ_{sel} datasets, based on strong positive correlations (>0.5) revealed a clear pattern of associations (Figure S2). Flavobacteriales clearly dominated these associations, accounting for more than 25% of the correlated taxa. They were followed, at a considerable distance, by Rhodobacterales and Puniceispirillales, each contributing around 6%–7% of the correlations. Other noteworthy groups, each representing approximately 4% of the associations, included SAR11 as well as the Gammaproteobacterial orders Oceanospirillales, Cellvibrionales and Burkholderiales. Notably, most of these ASVs belonged to the key groups identified by EQO and/or IndVal (Table S4).

ASV_{sel} and CAZ_{sel} correlations revealed a distinct prevalence pattern: Glycoside Hydrolases (GH) were dominant, followed by Glycosyltransferases (GT) and Polysaccharide Lyases (PL). Auxiliary Activities (AA), Carbohydrate Esterases (CE) and Carbohydrate-Binding Modules (CBM) were less abundant (Table S4). Among GH families, GH13 was most prevalent, followed by GH5, GH43, GH65, GH57 and GH42, mirroring overall dataset trends (Figure 3 in Data S2). GH13 and GH65 CAZymes were primarily associated with Bacteroidia (mainly Flavobacteriales), while GH42 correlated with Gammaproteobacteria. Alphaproteobacteria were overrepresented in correlations with GH5 and GH57 (Table S4).

3.5 | Environmental Influence on Microbial Oxygen Consumption

Although environmental variables explained only a small fraction of oxygen consumption variability, analyzing their contributions provides insights into the underlying drivers of this microbial process. Since PCA axes are derived from linear relationships, each variable's contribution can be directly quantified. The two retained ENV model axes—PC2 and PC11—highlight key factors grouped into four categories: DOM quality, water column structure, nutrients and temperature (Data S3).

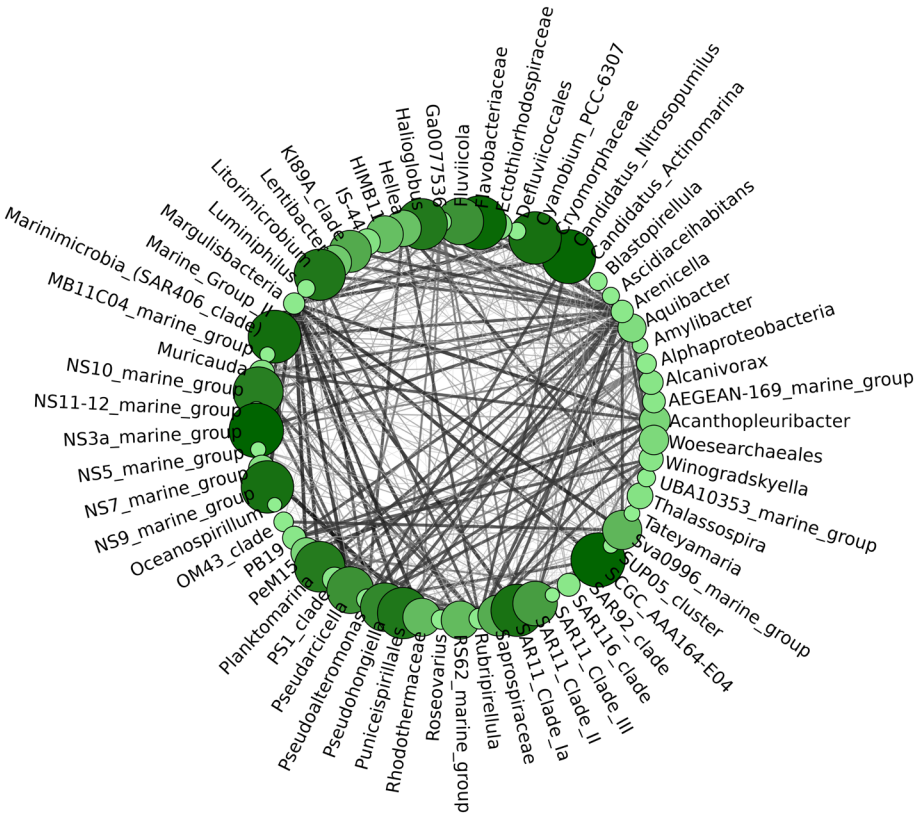

FIGURE 6 | Aggregation network of genera and their pairwise associations in relation to oxygen consumption as revealed by EQO. The network illustrates the relative importance of individual genera (nodes) and their pairwise associations (edges) in relation to oxygen consumption. Node size and colour intensity indicates the importance of each genus, while edge thickness represents the importance of the association between genera.

DOM composition-related variables, including FLD340 (protein-like), FLD416 and FLD452 (humic-like) and UV254 (aromatic compounds) (Data S3), were among the strongest contributors. These metrics reflect DOM characteristics such as aromaticity and chemical composition, which influence microbial degradation. Water column structure was captured by Secchi depth, turbidity and dissolved oxygen (DO), parameters that regulate light penetration, particle load and microbial interactions with DOM. Nutrients, including total nitrogen (NT), total phosphorus (PT) and nitrite (NO_2), also emerged as significant drivers. Temperature was consistently identified as a key factor in both PCA axes (Data S3).

4 | Discussion

Our framework successfully predicted oxygen consumption during DOM degradation, both quantitatively and qualitatively, demonstrating the strong link between bacterioplankton taxonomic composition and microbially driven biogeochemical processes. By integrating high-resolution community composition data with statistical modeling, we identified key microbial taxa and functional signatures driving oxygen consumption across diverse experimental conditions.

A major challenge in microbial ecology is extracting biologically meaningful patterns from high-dimensional datasets, where a high number of variables relative to a limited number of observations, combined with multicollinearity and a sparse data structure, often hinders robust statistical modelling (Lê Cao et al. 2008). We addressed this by applying sparse partial least squares regression (sPLSR), which enabled the selection of the most relevant ASVs and CAZymes associated with oxygen consumption, significantly improving model performance metrics. This feature selection step, combined with dimensionality reduction via PCoA, optimised predictive accuracy and enhanced biological interpretability.

Dimensionality reduction utilising ordination axes is a widely used and effective strategy to capture the most meaningful gradients in the data while minimising redundancy (Paliy and Shankar 2016) and has been effectively employed as a basis for modeling microbial processes (Domeignoz-Horta et al. 2020). Here, PCoA outperformed other ordination techniques in preserving variance relevant to oxygen consumption modeling, largely due to the characteristics of microbial datasets. High-dimensional and sparse structures, common in microbial ecology, pose challenges for methods that operate on raw data matrices, such as PCA. The presence of many zeros and low-abundance features can introduce instability and distort variance calculations, particularly in Euclidean-based approaches. PCoA overcomes this limitation by relying on a dissimilarity matrix, often constructed using ecological metrics like Bray-Curtis, which effectively deals with sparsity (Legendre and Legendre 1998). Additionally, PCoA (like PCA) generates orthogonal axes that reflect independent gradients of variation, improving interpretability and reducing multicollinearity—critical factors in regression-based models (James et al. 2013). In contrast, NMDS, while useful for visualizing

TABLE 2 | Bacterioplankton indicators for each oxygen consumption category. Indicators were identified at the ASV level (A) and the genus level (B). Candidate ASVs or genera met the criterion of being present in at least 50% of the samples within each category. The IndVal components of specificity (A) and sensitivity (B) were used to assess indicator performance. The final indicators represent the best combinations of candidates selected among all statistically significant possibilities ($p < 0.05$), achieving 100% coverage of their respective categories. Taxa highlighted in bold were also retrieved as key for oxygen consumption by the EQO analysis.

A. Indicators of the different consumption levels based on ASVs

		IndVal		
Group	**Candidate ASVs**	**A**	**B**	**Indicator ASVs and their taxonomic affiliation**
1	320	1	0.64	ASV14 **Fluviicola** + ASV158 **Saprospiraceae** + ASV1258 **SAR11_Clade_Ia**
		1	0.64	ASV192 **Rubripirellula** + ASV563 Ulvibacter + ASV1110 Roseivirga
		1	0.64	ASV 305 Candidatus_Puniceispirillum + ASV386 SAR324_clade (Marine_group_B) + **ASV742 Flavobacteriaceae NS5_marine_group**
2	269	1	0.52	ASV285 **Methylophilaceae OM43_clade** + ASV581 Rhodobacteraceae + ASV922 **Rhodospirillales AEGEAN-169_marine_group NA**
		1	0.44	ASV14 **Fluviicola** + ASV32 **Marinimicrobia_(SAR406_clade)** + ASV1058 Bacteriovoracaceae
		1	0.37	ASV45 **Marine_Group_II** + ASV297 **SAR11_Clade_III** + ASV509 **Arenicella**
		1	0.37	ASV190 **Flavobacteriaceae NS5_marine_group** + ASV423 Vicingus
3	197	1	0.50	ASV111 **Candidatus_Actinomarina** + ASV583 Rickettsiales S25-593 + ASV616 Aestuariicoccus
		1	0.30	ASV4 **Cyanobium_PCC-6307** + ASV505 **Microtrichaceae Sva0996_marine_group** + ASV567 **Sphingobacteriales NS11-12_marine_group**
		1	0.30	ASV119 **Flavobacteriales NS9_marine_group** + ASV398 Formosa + ASV579 **Rhodospirillales AEGEAN-169_marine_group**
4	128	1	1	ASV246 **Comamonadaceae RS62_marine_group** + ASV523 Alphaproteobacteria NRL2

B. Indicators of the different consumption levels based on genera

		IndVal		
Group	**Candidate genera**	**A**	**B**	**Indicator genera and their taxonomic affiliation**
1	117	1	0.45	Roseivirga + Ulvibacter + Allofrancisella
		1	0.36	**Acanthopleuribacter** + **PeM15** + Schlesneria
		1	0.36	Limnobacter + **Winogradskyella** + **Rhodobacteraceae HIMB11**
2	102	0.90	0.67	**Amylibacter** + **Thalassospira**
		0.89	0.59	Spongiibacteraceae BD1-7_clade + **SAR11_clade Clade_III** + Salinisphaera
		0.96	0.52	Bacteriovoracaceae + NS2b_marine_group + OCS116_clade
		0.95	0.52	**RS62_marine_group** + Algoriphagus
3	94	0.77	0.50	Alteromonas + **PS1_clade** + **Rhodobacteraceae HIMB11** + **Litorimicrobium**
		0.82	0.40	Alteromonas + Vicingus + **Puniceispirillales** + **Thalassospira**
		0.78	0.40	**Halioglobus** + Marinoscillum + **PS1_clade** + **Muricauda** + **NS10_marine_group**
4	67	1	1	Limnobacter + Algoriphagus + NRL2

community patterns, does not enforce orthogonality and is sensitive to initial configurations, potentially leading to inconsistencies across replicates. Furthermore, NMDS prioritises preserving rank order rather than absolute distances, limiting its suitability for applications requiring quantitative dimensionality reduction (Armstrong et al. 2022).

A

Sample	True_group	pW1	pW2	pW3	pW4	Pred_group
2018-03-14	W1	1	0	0	0	W1
2018-08-07	W1	1	0	0	0	W1
2018-09-05	W1	1	0	0	0	W1
2018-10-11	W1	1	0	0	0	W1
2018-10-24	W1	1	0	0	0	W1
2019-07-11	W1	1	0	0	0	W1
2019-08-28	W1	1	0	0	0	W1
2019-09-11	W1	1	0	0	0	W1
2019-09-18	W1	1	0	0	0	W1
2020-01-28	W1	1	0	0	0	W1
2021-05-20	W1	1	0	0	0	W1
2018-03-20	W2	0	1	0	0	W2
2018-06-18	W2	0	1	0	0	W2
2018-08-29	W2	0	1	0	0	W2
2018-09-27	W2	0	1	0	0	W2
2018-11-06	W2	0	1	0	0	W2
2018-11-20	W2	0	1	0	0	W2
2019-09-25	W2	0	1	0	0	W2
2019-10-08	W2	0	1	0	0	W2
2019-10-22	W2	0	1	0	0	W2
2019-11-06	W2	0	1	0	0	W2
2019-12-04	W2	0	1	0	0	W2
2019-12-17	W2	0	1	0	0	W2
2020-02-11	W2	0	1	0	0	W2
2020-03-04	W2	0	1	0	0	W2
2020-03-11	W2	0	1	0	0	W2
2020-09-04	W2	0	1	0	0	W2
2020-09-08	W2	0	1	0	0	W2
2020-09-16	W2	0	1	0	0	W2
2020-09-23	W2	0	1	0	0	W2
2020-10-08	W2	0	1	0	0	W2
2020-10-14	W2	0	1	0	0	W2
2020-11-04	W2	0	1	0	0	W2
2020-11-19	W2	0	1	0	0	W2
2020-12-02	W2	0	1	0	0	W2
2020-12-16	W2	0	1	0	0	W2
2021-04-28	W2	0	1	0	0	W2
2021-06-03	W2	0	1	0	0	W2
2018-04-09	W3	0	0	1	0	W3
2018-04-18	W3	0	0	1	0	W3
2018-04-25	W3	0	0	1	0	W3
2018-05-24	W3	0	0	1	0	W3
2019-05-28	W3	0	0	1	0	W3
2019-11-21	W3	0	0	1	0	W3
2020-09-29	W3	0	0	1	0	W3
2021-08-24	W3	0	0	1	0	W3
2021-10-06	W3	0	0	1	0	W3
2021-10-26	W3	0	0	1	0	W3
2018-12-10	W4	0	0	0	1	W4
2021-07-21	W4	0	0	0	1	W4

B

Sample	True_group	pW1	pW2	pW3	pW4	Pred_group
2018-03-14	W1	1	0.00	0.80	0	W1
2018-08-07	W1	1	0.95	0.85	0	W1
2018-09-05	W1	1	0.96	0.83	0	W1
2018-10-11	W1	1	0.90	0.00	0	W1
2018-10-24	W1	1	0.95	0.86	0	W1
2019-07-11	W1	1	0.96	0.83	0	W1
2019-08-28	W1	1	0.90	0.00	0	W1
2019-09-11	W1	1	0.90	0.79	0	W1
2019-09-18	W1	1	0.00	0.00	0	W1
2020-01-28	W1	1	0.96	0.79	0	W1
2021-05-20	W1	1	0.00	0.79	0	W1
2018-03-20	W2	0	0.96	0.81	0	W2
2018-06-18	W2	0	0.95	0.84	0	W2
2018-08-29	W2	0	0.44	0.00	0	W2
2018-09-27	W2	0	0.95	0.84	0	W2
2018-11-06	W2	0	0.88	0.80	0	W2
2018-11-20	W2	0	0.89	0.82	0	W2
2019-09-25	W2	0	0.95	0.00	0	W2
2019-10-08	W2	0	0.95	0.00	0	W2
2019-10-22	W2	0	0.96	0.00	0	W2
2019-11-06	W2	0	0.95	0.83	0	W2
2019-12-04	W2	0	0.90	0.83	0	W2
2019-12-17	W2	0	0.95	0.00	0	W2
2020-02-11	W2	0	0.89	0.78	0	W2
2020-03-04	W2	0	0.96	0.78	0	W2
2020-03-11	W2	0	0.95	0.80	0	W2
2020-09-04	W2	0	0.95	0.00	0	W2
2020-09-08	W2	0	0.96	0.00	0	W2
2020-09-16	W2	0	0.87	0.83	0	W2
2020-09-23	W2	0	0.95	0.00	0	W2
2020-10-08	W2	0	0.96	0.00	0	W2
2020-10-14	W2	0	0.96	0.00	0	W2
2020-11-04	W2	0	0.96	0.83	0	W2
2020-11-19	W2	0	0.90	0.00	0	W2
2020-12-02	W2	0	0.96	0.00	0	W2
2020-12-16	W2	0	0.96	0.00	0	W2
2021-04-28	W2	0	0.96	0.00	0	W2
2021-06-03	W2	0	0.95	0.78	0	W2
2018-04-09	W3	0	0.95	0.75	0	W2
2018-04-18	W3	0	0.00	0.77	0	W3
2018-04-25	W3	0	0.95	0.30	0	W2
2018-05-24	W3	0	0.90	0.29	0	W2
2019-05-28	W3	0	0.90	0.82	0	W2
2019-11-21	W3	0	0.95	0.77	0	W2
2020-09-29	W3	0	1.00	0.81	0	W2
2021-08-24	W3	0	0.91	0.82	0	W2
2021-10-06	W3	0	0.95	0.77	0	W2
2021-10-26	W3	0	0.00	0.77	0	W3
2018-12-10	W4	0	0.00	0.00	1	W4
2021-07-21	W4	0	0.00	0.00	1	W4

FIGURE 7 | Assignment of experiments to oxygen consumption categories according to ASV- and genus-based indicators, using leave-one-out cross-validation (LOOCV). (A) ASV-based indicators. (B) Genus-based indicators. In each panel, the first coloured column represents the true oxygen consumption category for each experiment. The subsequent four columns show the probability of assignment to each oxygen consumption category based on the respective indicators. The second coloured column highlights the group to which each sample was assigned based on the indicators. Colours correspond to the four oxygen consumption categories (W1 = low, W2 = moderate, W3 = high, W4 = very high).

Beyond selecting the most suitable ordination method, ensuring that retained axes contribute meaningfully to the model is essential. Defining an appropriate maximum number of axes in advance—considering variance explained and the balance between variables and observations—optimises model performance by enabling rigorous variable selection through statistical criteria such as AIC, BIC and significance tests. This approach prevents reliance on early axes based solely on explained variance and ensures that retained components provide biologically relevant information. Notably, the significant axes in our final models were not necessarily the first ones (Data S3). Eigenvector-based ordination methods, such as PCA and PCoA, capture dominant gradients in the dataset, but these do not always correspond to the ecological processes driving the response variable. Thus, relevant information for regression models may lie in later axes that capture subtler ecological patterns more directly related to the response variable.

The predictive power of taxonomic composition in explaining oxygen consumption suggests that bacterioplankton community structure inherently captures both functional and environmental influences on this process. This aligns with evidence of phylogenetic conservation in functional traits, such as the distribution of carbohydrate-active enzymes (CAZymes) across bacterial lineages, from strains to phyla (Berlemont and Martiny 2015; López-Mondéjar et al. 2022). Further traits related to the capacity to metabolise specific DOM components are often conserved at intermediate taxonomic levels, such as genus or family (Martiny et al. 2015). The strong correlation between taxonomic and functional datasets (Table S4) further supports the idea that taxonomic composition can serve as an effective

proxy for metabolic potential in the context of this community-level ecological process.

The strong link between taxonomic composition and oxygen consumption observed in this study may also result from the focus on a subset of dominant species, selected from the matrix of the most abundant taxa. Dominant taxa are theorized to have a disproportionate impact on ecosystem processes due to their high relative abundance, as proposed by the 'mass-ratio' hypothesis (Grime 1998). Additionally, the Metabolic Theory of Ecology (Brown et al. 2004) posits that the energy turnover of a population is proportional to its biomass, further supporting the central role of dominant species in driving ecosystem-level processes. While microbial communities include paradigmatic examples of specialised processes driven by rare taxa, these are typically associated with highly specific metabolic functions or unique environmental niches (Strous et al. 1999; Musat et al. 2008). In contrast, this study focuses on a broad-scale process—oxygen consumption—over a temporal span of 4 years. This extended timeframe encompasses a wide range of environmental conditions (Figure S3), ensuring that the observed relationships are robust and representative of the dynamics of a coastal system.

Most taxa identified by EQO as the most relevant for explaining oxygen consumption belonged to *Bacteroidia* (mainly Flavobacteriales), Alphaproteobacteria (mainly Rhodobacteraceae and SAR11), and several clades of Gammaproteobacteria. These groups are known for their central role in the processing of algal-derived DOM in surface waters (Buchan et al. 2014; Teeling et al. 2016). Similarly, most indicators of the different oxygen consumption categories also belonged to these groups. Interestingly, representatives of these taxa have been recently demonstrated to differentially contribute to oxygen consumption rates in marine environments (Munson-McGee et al. 2022). For instance, SAR11 members, here found as indicators of the categories of lower oxygen consumption (W1 and W2), consistently displayed low respiration rates in coastal and off-shore sites of the Atlantic and Pacific oceans. In contrast, *Planktomarina* (Rhodobacteraceae) and members of Flavobacteriales exhibited higher oxygen consumption rates in evaluations performed in the Gulf of Maine, with Flavobacteriales especially active during phytoplankton blooms (Munson-McGee et al. 2022). Here we found members of Flavobacteriales as indicators for all oxygen consumption categories (Table 2).

In our dataset, oxygen consumption rates measured in the first 24 h from 0 to 4.6 μmol of oxygen per hour (Figure 2C), similarly to the range of bulk oxygen consumption measured via Winkler in the Gulf of Maine (0.01–2.38 μmol) of oxygen per hour; data extracted from Figure 1b of Munson-McGee et al. (2022) using WebPlotDigitizer (https://automeris.io).

Nevertheless, the median value in our dataset (1.65 μmol of oxygen per hour, $n = 50$) was notably higher than for the data reported in that study (0.29 μmol of oxygen per hour, $n = 6$). Additionally, the range of chlorophyll values in this 4 year study (0.4–48.6 $\mu g\,L^{-1}$) was much higher than the range presented for a 3 year period in the Gulf of Maine (0.2–9.8 $\mu g\,L^{-1}$); data extracted from Extended Data Figure 3d of Munson-McGee et al. (2022), using WebPlotDigitizer (https://automeris.io).

Altogether, these data indicate that our system is much more productive, which could be a key factor in why Flavobacteriales comprise a high proportion of the community (28.5% of the sequences in our dataset), as they are usually found under resource-rich conditions worldwide (Simon et al. 1999; Kirchman et al. 2003; Alonso et al. 2010; Teeling et al. 2016). This may explain their identification as indicators for all oxygen consumption categories in our system, as well as the utmost importance of several Flavobacteriales for oxygen consumption revealed by EQO analysis. Their increased proportion in the ASV_{sel} dataset, compared to the $ASV_{0.005}$ or the entire dataset, further underscores their importance (Figure 2 in Data S2).

Productive conditions might also explain the importance of some Gammaproteobacteria (*Arenicella, Halioglobus, Luminiphilus, Pseudohongiella*, OM43_group, RS62_marine_group), the Verrucomicrobiota SCGC_AAA164-E04 and several Rhodobacteraceae (*Aestuariicoccus, Amylibacter, HIMB11, Lentibacter, Litorimicrobium, Planktomarina, Roseovarius, Tateyamaria*) identified as key groups for oxygen consumption by the EQO and/or IndVal analysis. Members of these groups have been repeatedly identified as being particularly abundant or active under such conditions (González et al. 2000; Cardman et al. 2014; Teeling et al. 2016; Liu et al. 2020; Francis et al. 2021).

Other taxa of high importance in the EQO analysis and/or selected as indicators of oxygen consumption categories were the Crenarchaeota *Candidatus Nitrosopumilus* and the Cyanobacteria *Cyanobium* PCC-6307. *Candidatus Nitrosopumilus* is a specialist in ammonia oxidation (Könneke et al. 2005) and has been previously identified by EQO as one of the key groups in the prediction of nitrate concentration in the TARA Oceans dataset (Shan et al. 2023). Its selection could indicate an indirect connection with oxygen consumption through nitrification, while the inclusion of the Cyanobium PCC-6307 might point to either DOM and/or oxygen production fueling heterotrophic aerobic DOM degradation.

Beyond taxonomic composition, functional traits associated with DOM degradation were strongly reflected in CAZyme correlations, particularly among glycoside hydrolase (GH) families. GH13, GH5 and GH43 emerged as dominant families, indicating the widespread utilisation of polysaccharides and glycoconjugates in the study system (Stam et al. 2006; Aspeborg et al. 2012; Mewis et al. 2016). The high prevalence of these enzymes suggests a substantial terrestrial influence on DOM composition, with plant-derived carbohydrates playing a significant role in shaping microbial metabolic pathways. The association of GH families with specific bacterial groups—GH13 and GH65 with Bacteroidia, GH42 with Gammaproteobacteria, and GH5/GH57 with Alphaproteobacteria—further supports the idea of substrate specialisation among dominant clades.

The high explanatory power of ASV-based models compared to genus-based models underscores the importance of fine-scale taxonomic resolution in linking microbial composition to oxygen consumption. This is consistent with the proposed shallow phylogenetic conservation of organic carbon substrate utilisation (Martiny et al., Martiny et al. 2013) and field observations showing that even closely related species engage in different physiological processes modulated by the environmental conditions

(Alonso et al. 2009). Nonetheless, genus-based models and indicators still demonstrate high explanatory and, notably, comparable predictive power, suggesting that the genus level may be adequate for defining functional groups within heterotrophic bacterioplankton. This aligns with recent findings that species replacement within genera plays a critical role in maintaining ecological success (Bustos-Caparros et al. 2024), reinforcing the ecological relevance of the genus as a key level of microbial diversity organisation.

The fact that taxonomic composition explained a much larger fraction of oxygen consumption variability compared to environmental variables suggests that microbial community structure inherently integrates, at least in part, the influence of environmental conditions on DOM degradation. However, identifying the specific environmental drivers remains essential for anticipating shifts in oxygen consumption rates under changing conditions. Here, DOM quality, water column structure, nutrient availability and temperature emerged as key factors—all undergoing significant transformations in coastal seas. Rising temperatures are altering DOM quality (Lønborg et al. 2020), potentially expanding bacterial functional diversity (Morán et al. 2023) while increasing ocean stratification limits vertical mixing, affecting nutrient and oxygen availability (Venegas et al. 2023). Concurrently, anthropogenic nutrient inputs drive eutrophication and the expansion of hypoxic zones (Breitburg et al. 2018) reshaping coastal ecosystem dynamics (Bindoff et al. 2019). Despite these well-documented trends, the extent to which environmental shifts will modify oxygen consumption rates and microbial community composition remains uncertain. A better understanding of these dynamics is needed to refine predictive models and assess microbial community resilience.

The South Atlantic Microbial Observatory (SAMO), located at the confluence of the Brazil and Malvinas currents and influenced by the Río de la Plata—one of the world's largest estuaries—is a hotspot of microbially mediated processes undergoing rapid environmental shifts, reflecting broader global trends (Kim et al. 2023). Investigating microbial responses in this dynamic setting offers insight into the future trajectories of coastal ecosystems.

Beyond SAMO, our approach provides a scalable framework for investigating microbially driven biogeochemical processes and refining our understanding of microbial functional groups. Applying and validating these models across diverse ecosystems will be essential to assess their generality, enhance large-scale monitoring efforts and advance the implementation of the space-for-time substitution approach in marine systems. This will ultimately improve predictions of microbial contributions to global biogeochemical cycles in the face of accelerating environmental change.

Author Contributions

Conceptualization: C.A. and R.A. designed the study framework and experimental objectives. Fieldwork: C.A., J.Z. and B.G. conducted field sampling, performed in situ measurements and coordinated the logistics for the SAMO observatory. Experimental design and execution: C.A., L.G. and B.G. implemented and carried out the biodegradation experiments. Laboratory work: C.A., J.Z., L.G. and B.G. performed DNA extractions and prepared amplicon libraries for sequencing; L.G. prepared the samples for DOM, nutrients and chlorophyll a analysis; J.Z. and B.G. conducted the spectral characterisation of DOM; A.P.-P. performed the HPLC characterisation of DOM; and C.L. carried out the nutrient and chlorophyll *a* analyses. Data analysis: E.P.-F. conducted the bioinformatic analyses and implemented the EQO routine, while C.A. performed the PARAFAC modelling. Statistical Modelling: C.A., E.P.-F. and C.C. conceptualised and applied the statistical modelling. Writing – original draft: C.A., E.P.-F. and R.A. wrote the initial draft of the manuscript. Writing – review and editing: All authors contributed to the critical revision of the manuscript, providing input and approval of the final version.

Acknowledgements

This project was funded by the ANII-MPI grant (ANII_MPI_ID_2017_1_1007663) awarded to C.A. and R.A., the full-time dedication program (DT) of the Universidad de la República, and PEDECIBA support to C.A. The National Agency for Research and Innovation is also acknowledged for supporting E.P.-F. as a postdoctoral fellow and L.G. and B.G. as PhD students during the execution and evaluation of the experiments. We further acknowledge the support of the Scientific Research Commission of the Universidad de la República Uruguay (CSIC) through the Grupos I + D 2022 program, which contributed to the completion of this work.

We are grateful to Stephan Kambach and Christian Ristok for their assistance with statistical modelling, as well as to Robert Strzepek for his valuable suggestions on an earlier version of this manuscript. Special thanks are extended to Danilo Calliari, Ana Martínez, Lorena Rodríguez, Maite Colina, Nicolás Silvera, Mariana Meerhoff and the personnel from La Paloma Naval Prefecture for their continued dedication to maintaining the SAMO observatory time series, which is part of the microbial observatories initiative within the µSudAqua network.

We are grateful to the three anonymous reviewers for their insightful and constructive comments, which significantly improved the quality and clarity of this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data supporting the findings presented here can be accessed through the Open Science Framework Repository (https://osf.io/9aupw/files/osfstorage). The sequencing data have been deposited in the European Nucleotide Archive (ENA) repository under the identifier PRJEB85301, and are also available in the µSudAqua database (Metz et al. 2022). The R code supporting the findings presented here can be accessed through the Open Science Framework Repository (https://osf.io/9aupw/files/osfstorage). Custom pipelines for bioinformatic processing are available in GitHub (https://github.com/pereiramemo/).

References

Akaike, H. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In *2nd International Symposium on Information Theory*, edited by B. N. Petrov and F. Csáki. Akadémia Kiadó.

Alonso, C., P. Gomez-Pereira, A. Ramette, L. Ortega, B. M. Fuchs, and R. Amann. 2010. "Multilevel Analysis of the Bacterial Diversity Along the Environmental Gradient Río de la Plata-South Atlantic Ocean." *Aquatic Microbial Ecology* 61, no. 1: 57–72. http://www.int-res.com/abstracts/ame/v61/n1/p57-72/.

Alonso, C., E. Pereira, F. Bertoglio, M. De Cáceres, and R. Amann. 2022. "Bacterioplankton Composition as an Indicator of Environmental Status: Proof of Principle Using Indicator Value Analysis of Estuarine

Communities." *Aquatic Microbial Ecology* 88: 1–18. https://doi.org/10.3354/ame01979.

Alonso, C., M. Zeder, C. Piccini, D. Conde, and J. Pernthaler. 2009. "Ecophysiological Differences of Betaproteobacterial Populations in Two Hydrochemically Distinct Compartments of a Subtropical Lagoon." *Environmental Microbiology* 11, no. 4: 867–876. https://doi.org/10.1111/j.1462-2920.2008.01807.x.

APHA. 1995. *Standard Methods for the Examination of Water and Wastewater*. 19th ed. American Public Health Association Inc.

Armstrong, G., G. Rahman, C. Martino, et al. 2022. "Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data." *Frontiers in Bioinformatics* 2: 821861. https://doi.org/10.3389/fbinf.2022.821861.

Aspeborg, H., P. M. Coutinho, Y. Wang, H. Brumer, and B. Henrissat. 2012. "Evolution, Substrate Specificity and Subfamily Classification of Glycoside Hydrolase Family 5 (GH5)." *BMC Evolutionary Biology* 12, no. 1: 186. https://doi.org/10.1186/1471-2148-12-186.

Bendschneider, K., and R. J. Robinson. 1952. "A New Spectrophotometric Method for Determination of Nitrite in the Sea Water." *Journal of Marine Research* 11: 87–96.

Berlemont, R., and A. C. Martiny. 2015. "Genomic Potential for Polysaccharide Deconstruction in Bacteria." *Applied and Environmental Microbiology* 81, no. 4: 1513–1519. https://doi.org/10.1128/AEM.03718-14.

Bindoff, N. L., W. Cheung, and J. Aristegui. 2019. Changing Ocean, Marine Ecosystems, and Dependent Communities (09 SROCC Ch05 FINAL-1).

Breitburg, D., L. A. Levin, A. Oschlies, et al. 2018. "Declining Oxygen in the Global Ocean and Coastal Waters." *Science* 359, no. 6371: eaam7240. https://doi.org/10.1126/science.aam7240.

Brown, J. H., J. F. Gillooly, A. P. Allen, V. M. Savage, and G. B. West. 2004. "Toward a Metabolic Theory of Ecology." *Ecology* 85, no. 7: 1771–1789. https://doi.org/10.1890/03-9000.

Buchan, A., G. R. LeCleir, C. A. Gulvik, and J. M. González. 2014. "Master Recyclers: Features and Functions of Bacteria Associated With Phytoplankton Blooms." *Nature Reviews. Microbiology* 12, no. 10: 686–698. https://doi.org/10.1038/nrmicro3326.

Bustos-Caparros, E., T. Viver, J. F. Gago, et al. 2024. "Ecological Success of Extreme Halophiles Subjected to Recurrent Osmotic Disturbances Is Primarily Driven by Congeneric Species Replacement." *ISME Journal* 18, no. 1. https://doi.org/10.1093/ismejo/wrae215.

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, no. 7: 581–583. https://doi.org/10.1038/nmeth.3869.

Cardman, Z., C. Arnosti, A. Durbin, et al. 2014. "Verrucomicrobia Are Candidates for Polysaccharide-Degrading Bacterioplankton in an Arctic Fjord of Svalbard." *Applied and Environmental Microbiology* 80, no. 12: 3749–3756. https://doi.org/10.1128/AEM.00899-14.

De Cáceres, M., and P. Legendre. 2009. "Associations Between Species and Groups of Sites:Indices and Statistical Inference." *Ecology* 90, no. 12: 3566–3574. https://doi.org/10.1890/08-1823.1.

Dittmar, T., B. Koch, N. Hertkorn, and G. Kattner. 2008. "A Simple and Efficient Method for the Solid-Phase Extraction of Dissolved Organic Matter (SPE-DOM) From Seawater." *Limnology and Oceanography: Methods* 6, no. 6: 230–235. https://doi.org/10.4319/lom.2008.6.230.

Domeignoz-Horta, L. A., G. Pold, X.-J. A. Liu, S. D. Frey, J. M. Melillo, and K. M. DeAngelis. 2020. "Microbial Diversity Drives Carbon Use Efficiency in a Model Soil." *Nature Communications* 11, no. 1: 3684. https://doi.org/10.1038/s41467-020-17502-z.

Drula, E., M.-L. Garron, S. Dogan, V. Lombard, B. Henrissat, and N. Terrapon. 2022. "The Carbohydrate-Active Enzyme Database: Functions and Literature." *Nucleic Acids Research* 50, no. D1: D571–D577. https://doi.org/10.1093/nar/gkab1045.

Dufrene, M., and P. Legendre. 1997. "Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach." *Ecological Monographs* 67, no. 3: 345–366.

Eddy, S. R. 2011. "Accelerated Profile HMM Searches." *PLoS Computational Biology* 7, no. 10: e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Fermani, P., N. Martyniuk, M. Saraceno, et al. 2024. "A Latin American Network of Microbial Observatories for Monitoring Aquatic Ecosystems." *Ecología Austral* 34, no. 3: 633–643. https://doi.org/10.25260/EA.24.34.3.0.2436.

Fox, J., and S. Weisberg. 2019. *An R Companion to Applied Regression*. Third ed. Sage.

Francis, B., T. Urich, A. Mikolasch, H. Teeling, and R. Amann. 2021. "North Sea Spring Bloom-Associated Gammaproteobacteria Fill Diverse Heterotrophic Niches." *Environmental Microbiomes* 16, no. 1: 15. https://doi.org/10.1186/s40793-021-00385-y.

Garnier, S., N. Ross, R. Rudis, P. A. Camargo, M. Sciaini, and C. Scherer. 2024. Viridis(Lite)–Colorblind-Friendly Color Maps for R.

González, J. M., R. Simó, R. Massana, et al. 2000. "Bacterial Community Structure Associated With a Dimethylsulfoniopropionate-Producing North Atlantic Algal Bloom." *Applied and Environmental Microbiology* 66, no. 10: 4237–4246. https://doi.org/10.1128/AEM.66.10.4237-4246.2000.

Grime, J. P. 1998. "Benefits of Plant Diversity to Ecosystems: Immediate, Filter and Founder Effects." *Journal of Ecology* 86, no. 6: 902–910. https://doi.org/10.1046/j.1365-2745.1998.00306.x.

Hedges, J. I., R. G. Keil, and R. Benner. 1997. "What Happens to Terrestrial Organic Matter in the Ocean?" *Organic Geochemistry* 27, no. 5–6: 195–212. https://doi.org/10.1016/S0146-6380(97)00066-1.

Helms, J. R., A. Stubbins, J. D. Ritchie, E. C. Minor, D. J. Kieber, and K. Mopper. 2008. "Absorption Spectral Slopes and Slope Ratios as Indicators of Molecular Weight, Source, and Photobleaching of Chromophoric Dissolved Organic Matter." *Limnology and Oceanography* 53, no. 3: 955–969. https://doi.org/10.4319/lo.2008.53.3.0955.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. Springer Texts in Statistics an Introduction to Statistical Learning - with Applications in R. https://doi.org/10.1007/978-1-4614-7138-7.

Jiao, N., G. J. Herndl, D. A. Hansell, et al. 2010. "Microbial Production of Recalcitrant Dissolved Organic Matter: Long-Term Carbon Storage in the Global Ocean." *Nature Reviews Microbiology* 8, no. 8: 593–599. https://doi.org/10.1038/nrmicro2386.

Kim, H. H., C. Laufkötter, T. Lovato, S. C. Doney, and H. W. Ducklow. 2023. "Projected 21st-Century Changes in Marine Heterotrophic Bacteria Under Climate Change." *Frontiers in Microbiology* 14: 1049579. https://doi.org/10.3389/fmicb.2023.1049579.

Kirchman, D. L., L. Y. Yu, and M. T. Cottrell. 2003. "Diversity and Abundance of Uncultured Cytophaga-Like Bacteria in the Delaware Estuary." *Applied and Environmental Microbiology* 69: 6587–6596. https://doi.org/10.1128/AEM.69.11.6587-6596.2003.

Könneke, M., A. E. Bernhard, J. R. de la Torre, C. B. Walker, J. B. Waterbury, and D. A. Stahl. 2005. "Isolation of an Autotrophic Ammonia-Oxidizing Marine Archaeon." *Nature* 437, no. 7058: 543–546. https://doi.org/10.1038/nature03911.

Kopf, A., M. Bicak, R. Kottmann, et al. 2015. "The Ocean Sampling Day Consortium." *GigaScience* 4, no. 1: 27. https://doi.org/10.1186/s13742-015-0066-5.

Koroleff, F. 1970. "Direct Determination of Ammonia in Natural Water as Indophenol-Blue. International Conference in the Exploration of the Sea. C.M 1969/C9. ICES. Information on Techniques and Methods for Sea Water Analysis." *Interlaboratory Reports* 3: 19–22.

Kujawinski, E. B. 2011. "The Impact of Microbial Metabolism on Marine Dissolved Organic Matter." *Annual Review of Marine Science* 3, no. 1: 567–599. https://doi.org/10.1146/annurev-marine-120308-081003.

Lê Cao, K. A., D. Rossouw, C. Robert-Granié, and P. Besse. 2008. "A Sparse PLS for Variable Selection When Integrating Omics Data." *Statistical Applications in Genetics and Molecular Biology* 7, no. 1: 35. https://doi.org/10.2202/1544-6115.1390.

Legendre, P., and E. D. Gallagher. 2001. "Ecologically Meaningful Transformations for Ordination of Species Data." *Oecologia* 129, no. 2: 271–280. https://doi.org/10.1007/s004420100716.

Legendre, P., and L. Legendre. 1998. *Numerical Ecology*. 2nd English ed. Elsevier.

Li, W.-T., Z.-X. Xu, A.-M. Li, W. Wu, Q. Zhou, and J.-N. Wang. 2013. "HPLC/HPSEC-FLD With Multi-Excitation/Emission Scan for EEM Interpretation and Dissolved Organic Matter Analysis." *Water Research* 47, no. 3: 1246–1256. https://doi.org/10.1016/j.watres.2012.11.040.

Li, Z., W. J. Riley, G. L. Marschmann, et al. 2025. "A Framework for Integrating Genomics, Microbial Traits, and Ecosystem Biogeochemistry." *Nature Communications* 16, no. 1: 2186. https://doi.org/10.1038/s41467-025-57386-5.

Liu, Y., Q. Lin, J. Feng, et al. 2020. "Differences in Metabolic Potential Between Particle-Associated and Free-Living Bacteria Along Pearl River Estuary." *Science of the Total Environment* 728: 138856. https://doi.org/10.1016/j.scitotenv.2020.138856.

Lønborg, C., C. Carreira, T. Jickells, and X. A. Álvarez-Salgado. 2020. "Impacts of Global Change on Ocean Dissolved Organic Carbon (DOC) Cycling." *Frontiers in Marine Science* 7. https://doi.org/10.3389/fmars.2020.00466.

López-Mondéjar, R., V. Tláskal, U. N. da Rocha, and P. Baldrian. 2022. "Global Distribution of Carbohydrate Utilization Potential in the Prokaryotic Tree of Life." *MSystems* 7, no. 6: e0082922. https://doi.org/10.1128/msystems.00829-22.

Lorenzen, C. J. 1967. "Determination of Chlorophyll and Pheopigments: Spectrophotometric Equations." *Limnology and Oceanography* 12: 343–346.

Lüdecke, D., M. Ben-Shachar, I. Patil, P. Waggoner, and D. Makowski. 2021. "Performance: An R Package for Assessment, Comparison and Testing of Statistical Models." *Journal of Open Source Software* 6, no. 60: 3139. https://doi.org/10.21105/joss.03139.

Lüdecke, D., I. Patil, M. Ben-Shachar, B. Wiernik, P. Waggoner, and D. Makowski. 2021. "An R Package for Visualizing Statistical Models." *Journal of Open Source Software* 6, no. 64: 3393. https://doi.org/10.21105/joss.03393.

Mackereth, F. J. H., J. Heron, and J. F. Talling. 1978. "Water Analysis: Some Revised Methods for Limnologists." *Freshwater Biological Association* 36: 1–120.

Martiny, A. C., K. Treseder, and G. Pusch. 2013. "Phylogenetic Conservatism of Functional Traits in Microorganisms." *ISME Journal* 7, no. 4: 830–838. https://doi.org/10.1038/ismej.2012.160.

Martiny, J. B. H., S. E. Jones, J. T. Lennon, and A. C. Martiny. 2015. "Microbiomes in Light of Traits: A Phylogenetic Perspective." *Science* 350, no. 6261: aac9323–aac9323. https://doi.org/10.1126/science.aac9323.

Mazerolle, M. J. 2023. "AICcmodavg: Model Selection and Multimodel Inference Based on (Q)AIC(c)." In *R Package Version 2.3-3*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=AICcmodavg.

Metz, S., P. Huber, E. Mateus-Barros, et al. 2022. "A Georeferenced rRNA Amplicon Database of Aquatic Microbiomes From South America." *Scientific Data* 9, no. 1: 565. https://doi.org/10.1038/s41597-022-01665-z.

Mewis, K., N. Lenfant, V. Lombard, and B. Henrissat. 2016. "Dividing the Large Glycoside Hydrolase Family 43 Into Subfamilies: A Motivation for Detailed Enzyme Characterization." *Applied and Environmental Microbiology* 82, no. 6: 1686–1692. https://doi.org/10.1128/AEM.03453-15.

Morán, X. A. G., N. Arandia-Gorostidi, T. M. Huete-Stauffer, and L. Alonso-Sáez. 2023. "Temperature Enhances the Functional Diversity of Dissolved Organic Matter Utilization by Coastal Marine Bacteria." *Environmental Microbiology Reports* 15, no. 1: 31–37. https://doi.org/10.1111/1758-2229.13123.

Müllin, J. B., and J. P. Riley. 1955. "The Spectrophotometric Determination of Silicate-Silicon in Natural Waters With Special Reference to Sea Water." *Analytica Chimica Acta* 12: 162–170.

Munson-McGee, J. H., M. R. Lindsay, E. Sintes, et al. 2022. "Decoupling of Respiration Rates and Abundance in Marine Prokaryoplankton." *Nature* 612, no. 7941: 764–770. https://doi.org/10.1038/s41586-022-05505-3.

Murphy, J., and J. P. Riley. 1962. "A Modified Single Solution Method for the Determination of Phosphate in Natural Waters." *Analytica Chimica Acta* 27: 31–36.

Murphy, K. R., C. A. Stedmon, D. Graeber, and R. Bro. 2013. "Fluorescence Spectroscopy and Multi-Way Techniques. PARAFAC." *Analytical Methods* 5, no. 23: 6557. https://doi.org/10.1039/c3ay41160e.

Murphy, K. R., C. A. Stedmon, P. Wenig, and R. Bro. 2014. "OpenFluor–An Online Spectral Library of Auto-Fluorescence by Organic Compounds in the Environment." *Analytical Methods* 6, no. 3: 658–661. https://doi.org/10.1039/C3AY41935E.

Musat, N., H. Halm, B. Winterholler, et al. 2008. "A Single-Cell View on the Ecophysiology of Anaerobic Phototrophic Bacteria." *Proceedings of the National Academy of Sciences* 105, no. 46: 17861–17866. https://doi.org/10.1073/pnas.0809329105.

Oksanen, J., F. G. Blanchet, R. Kindt, et al. 2012. Vegan: Community Ecology Package. R Package Version 2.0–2.

Paliy, O., and V. Shankar. 2016. "Application of Multivariate Statistical Techniques in Microbial Ecology." *Molecular Ecology* 25, no. 5: 1032–1057. https://doi.org/10.1111/mec.13536.

Parada, A. E., D. M. Needham, and J. A. Fuhrman. 2016. "Every Base Matters: Assessing Small Subunit rRNA Primers for Marine Microbiomes With Mock Communities, Time Series and Global Field Samples." *Environmental Microbiology* 18, no. 5: 1403–1414. https://doi.org/10.1111/1462-2920.13023.

Parsons, T. R., Y. Maita, and C. M. Lalli. 1984. *A Manual of Chemical and Biological Methods for Seawater Analysis*. Pergamon Press.

Pedersen, T. 2024a. Patchwork: The Composer of Plots. R Package Version 1.3.0.9000.

Pedersen, T. 2024b. Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R Package Version 2.2.1.9000.

Pedersen, T. 2024c. Tidygraph: A Tidy API for Graph Manipulation.

Pucher, M., U. Wünsch, G. Weigelhofer, K. Murphy, T. Hein, and D. Graeber. 2019. "staRdom: Versatile Software for Analyzing Spectroscopic Data of Dissolved Organic Matter in R." *Water* 11, no. 11: 2366. https://doi.org/10.3390/w11112366.

Quast, C., E. Pruesse, P. Yilmaz, et al. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41, no. D1: D590–D596. https://doi.org/10.1093/nar/gks1219.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, C. 2019. "Microbial Respiration, the Engine of Ocean Deoxygenation." *Frontiers in Marine Science* 5: 533. https://doi.org/10.3389/fmars.2018.00533.

Rohart, F., B. Gautier, A. Singh, and K.-A. Lê Cao. 2017. "mixOmics: An R Package for 'Omics Feature Selection and Multiple Data Integration." *PLoS Computational Biology* 13, no. 11: e1005752. https://doi.org/10.1371/journal.pcbi.1005752.

Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6, no. 2: 461–464. https://doi.org/10.1214/aos/1176344136.

Scrucca, L. 2013. "GA: A Package for Genetic Algorithms in R." *Journal of Statistical Software* 53, no. 4: 1–37. https://doi.org/10.18637/jss.v053.i04.

Shan, X., A. Goyal, R. Gregor, and O. X. Cordero. 2023. "Annotation-Free Discovery of Functional Groups in Microbial Communities." *Nature Ecology & Evolution* 7, no. 5: 716–724. https://doi.org/10.1038/s41559-023-02021-z.

Simon, M., F. O. Glöckner, and R. Amann. 1999. "Different Community Structure and Temperature Optima of Heterotrophic Picoplankton in Various Regions of the Southern Ocean." *Aquatic Microbial Ecology* 18, no. 3: 275–284. https://doi.org/10.3354/ame018275.

Stam, M. R., E. G. J. Danchin, C. Rancurel, P. M. Coutinho, and B. Henrissat. 2006. "Dividing the Large Glycoside Hydrolase Family 13 Into Subfamilies: Towards Improved Functional Annotations of -Amylase-Related Proteins." *Protein Engineering Design and Selection* 19, no. 12: 555–562. https://doi.org/10.1093/protein/gzl044.

Strous, M., J. a. Fuerst, E. H. Kramer, et al. 1999. "Missing Lithotroph Identified as New Planctomycete." *Nature* 400, no. 6743: 446–449. https://doi.org/10.1038/22749.

Strzepek, R. F., B. L. Nunn, L. T. Bach, J. A. Berges, E. B. Young, and P. W. Boyd. 2022. "The Ongoing Need for Rates: Can Physiology and Omics Come Together to Co-Design the Measurements Needed to Understand Complex Ocean Biogeochemistry?" *Journal of Plankton Research* 44, no. 4: 485–495. https://doi.org/10.1093/plankt/fbac026.

Teeling, H., B. M. Fuchs, C. M. Bennke, et al. 2016. "Recurring Patterns in Bacterioplankton Dynamics During Coastal Spring Algae Blooms." *eLife* 5: e11888. https://doi.org/10.7554/eLife.11888.

Thomas, O., and C. Burgess. 2007. *UV-Visible Spectrophotometry of Water and Wastewater*. Elsevier.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 58, no. 1: 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Valderrama, J. C. 1981. "The Simultaneous Analysis of Total N and P in Natural Waters." *Marine Chemistry* 10: 109–122.

Venegas, R. M., J. Acevedo, and E. A. Treml. 2023. "Three Decades of Ocean Warming Impacts on Marine Ecosystems: A Review and Perspective." *Deep Sea Research Part II: Topical Studies in Oceanography* 212: 105318. https://doi.org/10.1016/j.dsr2.2023.105318.

Weishaar, J., G. Aiken, B. Bergamaschi, M. Fram, R. Fujii, and K. Mopper. 2003. "Evaluation of Specific Ultra-Violet Absorbance as an Indicator of the Chemical Content of Dissolved Organic Carbon." *Environmental Science & Technology* 37, no. 20: 4702–4708. https://doi.org/10.1021/es030360x.

Wickham, H. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://doi.org/10.1007/978-3-319-24277-4.

Zhao, Y., O. X. Cordero, and M. Tikhonov. 2024. "Linear-Regression-Based Algorithms Can Succeed at Identifying Microbial Functional Groups Despite the Nonlinearity of Ecological Function." *PLoS Computational Biology* 20, no. 11: e1012590. https://doi.org/10.1371/journal.pcbi.1012590.

Zuur, A. F., and E. N. Ieno. 2016. "A Protocol for Conducting and Presenting Results of Regression-Type Analyses." *Methods in Ecology and Evolution* 7, no. 6: 636–645. https://doi.org/10.1111/2041-210X.12577.

Zuur, A. F., E. N. Ieno, and C. S. Elphick. 2010. "A Protocol for Data Exploration to Avoid Common Statistical Problems." *Methods in Ecology and Evolution* 1, no. 1: 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Optimization of the number of genera selected for explaining oxygen consumption using the EQO approach. Applied to the GEN_{sel} dataset (ASV_{sel} dataset grouped by genera), the analysis shows that selecting up to 18 genera maximises the correlation with oxygen consumption. Beyond this threshold, no further increase in the R^2 value is observed, as indicated by the plateau in the curve. **Figure S2:** Correlation heatmap between ASVsel and CAZsel datasets. Pairwise Spearman correlations were calculated between the relative abundances of selected ASVs (*Y* axis) and selected CAZyme-encoding genes (*X* axis). Positive correlations are shown in red and negative correlations in blue, with colour intensity indicating correlation strength (ranging from −0.6 to +0.6, as indicated in the colour bar). Hierarchical clustering was applied to both axes to highlight patterns of co-association. **Figure S3:** Variability of environmental variables measured across samples. The selected variables include information on the structure of the water column (Temperature, Salinity, Dissolved Oxygen, Turbidity), the trophic status (Chlorophyll a concentration, Phycocyanin Fluorescence, Total Nitrogen and Total Phosphorous) and representatives of the DOM components quantified by HPLC (a proteinaceous hydrophilic, a proteinaceous hydrophobic and a Humic hydrophobic component, respectively). Each box represents the interquartile range (IQR), with whiskers extending to values within 1.5 times the IQR. Outliers are represented by points beyond the whiskers. **Table S1:** Summary of ordination results and axis selection for modelling microbial community data. Number of variables, ordination technique, cumulative variance explained by successive axes (when applicable) and number of axes selected for modelling are shown for each dataset, using PCoA, PCA or NMDS. **Table S2:** Comparison of the linear regression models obtained for the $ASV_{0.005}$ and CAZ_{all} datasets with different ordination techniques. The table summarises the number of features, components and model performance metrics—Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and explained variance (adjusted R^2)—for both the full and optimised models applied to the $ASV_{0.005}$ and CAZ_{all} datasets in explaining oxygen consumption during DOM degradation experiments. **Table S3A:** Variables retained in Lasso regression models using ordination axes as predictors. **Table S3B:** Variables retained in Lasso regression models using raw data as predictors. **Table S4:** ASVs with strong positive correlations (>0.5) with CAZymes. **Data S1:** DOM Characterisation. **Data S2:** Bacterioplankton taxonomic and functional composition. **Data S3:** Detailed linear models.