Title

Microbial community composition predicts bacterial production across ocean ecosystems.

Journal

The ISME Journal: Multidisciplinary Journal of Microbial Ecology, 18(1)

Authors

Connors, Elizabeth

Dutta, Avishek

Trinh, Rebecca

et al.

Publication Date

2024-01-08

DOI

10.1093/ismejo/wrae158

Peer reviewed

The ISME Journal, 2024, **18(1)**, wrae158
https://doi.org/10.1093/ismejo/wrae158
Advance access publication: 6 August 2024
Original Article

Microbial community composition predicts bacterial production across ocean ecosystems

Elizabeth Connors[1,2,*], Avishek Dutta[3,4], Rebecca Trinh[5], Natalia Erazo[1], Srishti Dasarathy[1], Hugh Ducklow[5], J.L. Weissman[6,7], Yi-Chun Yeh[6], Oscar Schofield[8], Deborah Steinberg[9], Jed Fuhrman[6], Jeff S. Bowman[1,2]

[1]Scripps Institution of Oceanography, UC San Diego, La Jolla, CA 92037, United States
[2]Scripps Polar Center, UC San Diego, La Jolla, CA 92037, United States
[3]Department of Geology, University of Georgia, Athens, GA 30602, United States
[4]Savannah River Ecology Laboratory, University of Georgia, Aiken, SC 29802, United States
[5]Lamont-Doherty Earth Observatory, Columbia University, New York, NY 10964, United States
[6]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, United States
[7]Department of Biology, The City College of New York, New York, NY 10003, United States
[8]Coastal Ocean Observation Laboratory, Institute of Marine and Coastal Sciences, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, NJ 08901-8520, United States
[9]Virginia Institute of Marine Science, College of William & Mary, Gloucester Point, VA 23062, United States

*Corresponding author: Elizabeth Connors, Scripps Institution of Oceanography, UC San Diego, La Jolla, CA 92037, United States. Email: econnors@ucsd.edu

Abstract

Microbial ecological functions are an emergent property of community composition. For some ecological functions, this link is strong enough that community composition can be used to estimate the quantity of an ecological function. Here, we apply random forest regression models to compare the predictive performance of community composition and environmental data for bacterial production (BP). Using data from two independent long-term ecological research sites—Palmer LTER in Antarctica and Station SPOT in California—we found that community composition was a strong predictor of BP. The top performing model achieved an R^2 of 0.84 and RMSE of 20.2 pmol L^{-1} hr^{-1} on independent validation data, outperforming a model based solely on environmental data ($R^2 = 0.32$, RMSE = 51.4 pmol L^{-1} hr^{-1}). We then operationalized our top performing model, estimating BP for 346 Antarctic samples from 2015 to 2020 for which only community composition data were available. Our predictions resolved spatial trends in BP with significance in the Antarctic (P value $= 1 \times 10^{-4}$) and highlighted important taxa for BP across ocean basins. Our results demonstrate a strong link between microbial community composition and microbial ecosystem function and begin to leverage long-term datasets to construct models of BP based on microbial community composition.

Keywords: microbial ecological function, community structure, bacterial production, random forest regression

Introduction

Microbial ecosystem functions, defined as microbial activity at the community scale, are an essential component of Earth's biogeochemical cycles, including carbon and nitrogen [1, 2]. Typically measured via a stable isotope tracer or via enzymatic activity, microbial functions are often—but not always—strongly linked to microbial community composition [1]. Previous work has identified strong links between community composition and various components of the carbon and nitrogen cycles and demonstrated that community composition data can be used to make quantitative predictions of some functions [3–9]. For instance, because microbial community composition strongly influences decomposition and respiration rates in soil [10], bacterial community composition can be used to predict dissolved organic carbon concentrations in leaf litter [4]. Identifying the connections between microbial composition and function is of utmost importance in the context of global change, where microbial diversity loss is predicted to increase with unknown consequences to microbial function and subsequently carbon and nitrogen cycling [11].

In one example of a microbial ecosystem function, marine heterotrophic bacteria (here meaning heterotrophic members of the bacteria and archaea) incorporate phytoplankton-derived dissolved organic matter (DOM) into new bacterial biomass through bacterial production (BP). This repackaging of DOM into microbial biomass is a key step in the microbial loop, where organic matter is recycled to the higher trophic levels via bacterivory by protists [12]. As the abundance and productivity of bacteria thus rely on the availability of phytoplankton-derived DOM, BP is strongly related to primary production (PP), with an average global BP:PP ratio of ~10% [13, 14]. However, this ratio of BP:PP is highly variable (~0.5%–25%) depending greatly on the time and space scales analyzed [14, 15].

Here, we leveraged long-term time series of BP and other microbial and environmental data across two coastal regions to

Received: 5 March 2024. **Revised:** 28 June 2024. **Accepted:** 5 August 2024

construct predictive models and better understand the links between microbial community composition and BP. The selected long-term study sites were the Palmer Long-Term Ecological Research site along the western Antarctic Peninsula (wAP) and the long-running San Pedro Ocean Time Series (SPOT) located in the San Pedro Channel in Southern California. Unfortunately, many long running time series do not regularly measure BP, which can be prohibitive or difficult to measure in some settings as it requires a radioactive tracer. The Palmer Long-Term Ecological Research (LTER) site off the western Antarctic Peninsula (wAP) is a good example; there have been hundreds of measurements for bacterial community composition since 2015 with a concurrent BP measurement only 24% of the time. However, previous work has shown that BP is strongly related to bacterial community composition and bacterial abundance along the wAP [3, 16]. These works showed that community composition and bacterial abundance were the two most important variables in a linear model that best described BP over five spring–summer seasons along the wAP [3]. Other work showed that increasing BP coincided with a change in community composition during the 2014 summer season [17]. A more robust dataset of BP is necessary as (along with PP) BP has the most direct effect on the production and consumption of dissolved organic carbon and nitrogen in the upper ocean of the western Antarctic Peninsula [18].

In this study, we utilized the random forest algorithm to construct and compare several models that predict BP from bacterial community composition using data from the Palmer LTER along the wAP and from Station SPOT in Southern California, where more observations of BP are available. This approach has previously been shown to be effective for predicting biogeochemical standing stocks that are strongly influenced by microbial processes [4, 19]. We expand on previous studies by comparing our amplicon models of BP to one constructed from only environmental and fluorescence data, demonstrating that community composition may be a better predictor of BP than environmental data and highlighting a fundamental microbial composition–function relationship. Overall, our findings demonstrate a strong link between the composition of microbial communities and their ecosystem functions in two disparate coastal research sites.

Materials and methods

Palmer Station Long-Term Ecological Research data

Palmer Long-Term Ecological Research Project (PAL) amplicon data used in this study were downloaded from the NCBI SRA database under BioProject PRJNA901488. A minority of the samples used in this study (113 samples, 25%) were collected weekly from 10 m depth at PAL Station B (lat: −64.774167, long: −64.0544) over austral summer seasons from 2015 to 2020 (PAL samples, November–March, Fig. 1C). The majority (340 samples, 75%) of samples were collected during two cruises, aboard the ASRV *Laurence M. Gould* (LMG) in January of 2019 and 2020, respectively (LMG1901 and LMG2001 samples), and from the small vessel *Hadar* on 7 March 2020 (PD2001). Samples collected during the two LMG cruises were collected along a sampling grid of stations 10 km apart arranged in 10 onshore to offshore lines spaced 100 km apart along the Peninsula and opportunistically along the ship track throughout the nine LTER subregions (along-shore regions of offshore, shelf, coastal regions; cross-shore regions of north, south, and farther south, Fig. 1B) extensively outlined in previous work [20, 21].

For the PAL amplicon samples, 1 L of seawater was filtered through a sterile 0.2 μm Supor membrane disk filter (Pall Corporation, Port Washington, NY, USA) and stored at −80°C until extraction. Filters were extracted using the KingFisher Flex Purification System and MagMax Microbiome Ultra Nucleic Acid Extraction kit (ThermoFisher Scientific, Waltham, MA, USA). Extracted DNA was sent to Argonne National Laboratory for amplicon library preparation and sequencing using the MiSeq platform (Illumina) with the universal primers 515F and 806R [22], and a 2 × 151 bp library architecture. Illumina reads were then filtered, denoised, and merged with dada2 [23].

BP samples were collected alongside amplicon data for 108 of the 453 samples (24%) used in this study. PAL BP data were downloaded from the ERRDAP database [24]. All samples were collected and processed according to PAL LTER standard protocols using radioactively labeled 3H-leucine [15].

Bacterial abundance data via flow cytometry were also downloaded from the PAL LTER ERRDAP database [24] and collected alongside amplicon data for all 453 samples (100%). Flow cytometry samples were prefiltered (with a Coring Falcon 40 μm Cell Strainer) before running on an AccuriC6 flow cytometer (BD Biosciences, Franklin Lakes, NJ, USA) equipped with a blue (488 nm) laser. All samples were stained and incubated in the dark for 15 min with the nucleotide stain SYBR Green 1 (Molecular Probes, Inc., Eugene, OR, USA) and the manufacturer's recommended concentration. Quality control for absolute cell counts were confirmed by spiking 10 μl of 1:2500 diluted 1 μm Fluoresbrite Yellow Microspheres (Polyscience Inc., Fishers, IN, USA) to each sample. All samples were run on "slow" with a flow rate of 14 μl min^{-1} for 1 min and measured for forward scatter, side scatter, and green emission (488/533 nm excitation/emission).

Bacterial populations were identified using a self-organizing map (SOM) from forward scatter, side scatter, and green emission following previous methods [3, 25]. In brief, a training set was constructed with data from five randomized sample days, with one from each year. These data were trained using a toroidal map with a grid size of 41 × 41 using the "kohonen" package in R [26]. Populations were identified using k-means clustering and $k = 6$ was chosen through a priori knowledge of populations and the visual evaluation of a within-cluster sum of squares scree plot. This k-means cluster model was then used to classify events in all flow cytometry samples. HNA and LNA bacterial populations were identified from the flow cytometry clusters and were converted into cells mL^{-1}. These two bacterial populations were then combined to form a total cell count (bacterial abundance in cells mL^{-1}) for each sample. Total cell count outliers (two observations) were removed when their values were outside the range Q1–1.5 × (Q3 − Q1), Q3 + 1.5 × (Q3 − Q1), where Q1 and Q3 are the first and third quartiles, respectively.

San Pedro Ocean Time Series data

Amplicon data from the upper 200 m were downloaded from the EMBL database under accession PRJEB48162 and processed following developed protocols [27]. This dataset includes monthly measurements at the San Pedro Time Series (SPOT, Fig. 1A) off the coast of California from 2005 to 2018 for community composition, via amplicon sequencing of two distinct filter size classes (0.2–1 and 1–80 μm). Once retrieved from the database, amplicon sequences were trimmed with cutadapt [28], split into 16S rRNA gene reads using bbtools [29] and denoised and merged with dada2 [23]. For each sampling day, the distinct filter size classes were combined (via simple addition of the absolute read counts

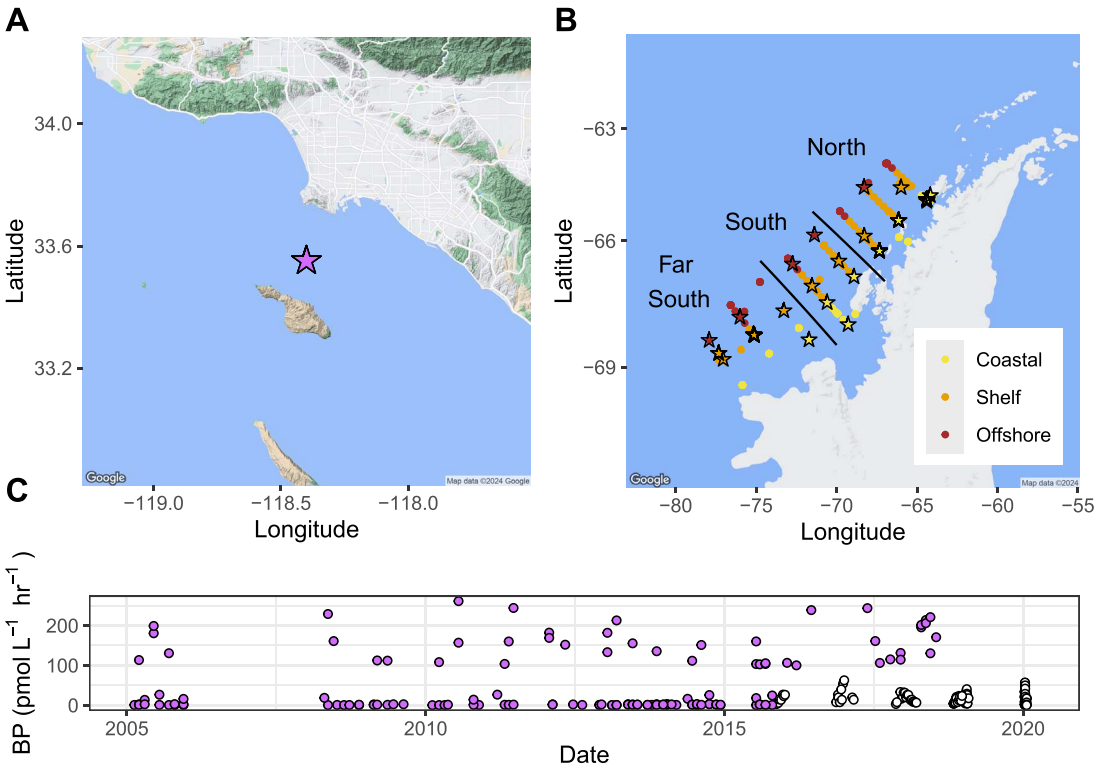

Figure 1. (A) Sampling location of the San Pedro Ocean Time Series off the southern California coast (SPOT, star). SPOT has a total of 133 amplicon samples from monthly samples from 2005 to 2020 with an associated bacterial production (BP) measurement (purple values in 1C). (B) Sampling locations along the Palmer LTER (PAL) grid along the western Antarctic Peninsula. The PAL dataset contains 453 amplicon samples, 108 of which have an associated bacterial production (BP, pmol leucine per L^{-1} h^{-1}) measurement (24%, black stars). (C) Timing and amplitude of the total 241 observations of BP used in this study (purple from SPOT, white from PAL). Underlying map courtesy of Google Maps.

of each ASV in both size fractions) for our final analysis to better match sequences with no filter size classes from the PAL dataset.

BP samples were collected alongside amplicon data for all 133 monthly samples (100%) used in this study [30]. All samples were collected and processed according to SPOT standard protocols using radioactively labeled 3H-leucine [31]. Finally, environmental data were also collected for all SPOT samples, including NO_3, PO_4, and CTD measurements for temperature, oxygen, salinity, and fluorescence following standard protocols [27, 30]. These environmental variables were used to create the SPOT-ENV model.

16S rRNA gene amplicon data

PAL and SPOT data were QC'd and denoised to amplicon sequence variants (ASVs) with dada2 [23] following previously published procedures [19]. Because different primers were used for these datasets the ASVs are consistent only with each dataset. PAL and SPOT ASVs data were then analyzed with paprica v0.7.1 [32]. Paprica utilizes phylogenetic placement with Gappa [33] EPA-ng [34] and Infernal [35] to place query reads on a reference tree constructed from the full-length 16S rRNA genes from all completed genomes in GenBank [36]. All unique reads are assigned to both internal branches (closest estimated genomes, see paprica documentation for further description) and if possible, terminal branches (closest completed genomes) on the reference tree. Once assigned, unique reads that were assigned as mitochondria or chloroplasts were omitted, as well as any reads that only appeared once (25% of all ASVs). The phylogenetic placement approach that is inherent to paprica results in ASV aggregation by phylogenetic edges. This edge-level data unified the data across primers and were used for relative mean maximal growth rate calculations and joint model construction.

The latest version of paprica includes a prediction for relative mean minimal doubling time from codon usage patterns adapted from the R package gRodon [37]. To make doubling time predictions, paprica applies gRodon by calculating the relative mean minimal doubling time on each completed genome in the paprica database. Because the goal is to estimate the theoretical maximum growth rate based on genetic signature, predictions are presented only in relative terms without correction for temperature. Mean minimal doubling times are then assigned to reads according to their point of placement on the paprica reference tree. Placements to terminal branches (i.e. closest completed genomes) are assigned the rate corresponding with that genome, placements to internal branches (i.e. closest estimated genomes) are assigned the average of all rates of terminal nodes belonging to that clade. A predicted relative mean minimal doubling time for the community is calculated by taking the average of the rates assigned to all edges.

Random forest regression modeling

All random forest regression models were created using the "randomForest" package in R [38]. For each model, samples were restricted to those that had an observed BP (BP_{obs}) measurement. Those samples with a BP_{obs} were further randomly separated into training (80%) and validation (20%) datasets. A random forest regression was run on the training dataset and then used to predict BP (BP_{pred}) for the validation dataset. Model performance was assessed using residual mean square error (RMSE) and R^2 values for the validation dataset. The optimal number of decision

trees (ntree in randomForest, the point where more trees did not improve model performance) was set to 500 and number of variables to randomly sample as candidates at each split (mtry in randomForest, where too few variables can lead to overly biased results and too many is inefficient) was set to 10 after random forest hypertuning over the range of 100–1200 ntree and 1–20 mtry; where ntree = 500 and mtry = 10 produced the highest R^2 in a linear regression of BP_{obs} and BP_{pred} for the validation data.

Palmer Long-Term Ecological Research Project absolute abundance calculation and model input variables

For the PAL data only, we were able to leverage the available flow cytometry data to produce an absolute abundance for each of the closest completed genome matches from the amplicon reads. We multiplied relative abundance of quality controlled unique 16S rRNA gene reads—corrected for 16S rRNA gene copy number by paprica—by the total bacterial abundance (as stated above, the HNA and LNA combined cell count from flow cytometry). For PAL random forest models only, we compared a model with absolute abundance as input (PAL-CCG) to a model with relative abundance as input (PAL-CEG). As flow cytometry data were not available for SPOT, only relative abundance data were used in the SPOT random forest model (SPOT-CEG) and for the joint model (JOINT-CEG) with data from both PAL and SPOT, with an additional categorical variable for region. The fifth and final model was created with only the environmental data from Station SPOT (SPOT-ENV) as input. A PAL-ENV model was not created as there were not enough environmental data available.

Feature selection on input variables with Boruta

All models included a feature selection step before the random forest regression model. The ASV relative or absolute abundances were reduced by the Boruta feature selection algorithm [39]. This algorithm finds those variables that contribute most to model performance by iteratively removing those variables for which randomization does not diminish model performance. For SPOT-ENV, feature selection did not reduce the number of variables in the model (as all were deemed relevant).

Cross validation of random forest and predictions of bacterial production

For each of the five random forest models we tested (PAL-CEG, PAL-CCG, SPOT-CEG, SPOT-ENV, and JOINT-CEG), a final model was then trained with all samples with BP_{obs} and we validated our results via random forest cross validation performed with the package "rfUtilities" [40]. This is an independent assessment of model performance that provides the average mean standard error and average variance explained with 10% of the data omitted randomly over 99 iterations of the model. Out best performing model, the cross-validated JOINT-CEG model, was then used to predict BP for those samples where BP data were missing (346 PAL samples). For statistical comparisons of BP_{pred} and BP_{obs} across multiple spatial scales, *P* values were adjusted for multiple comparisons via the Holm–Bonferroni method. In addition to predicting BP, the increase in mean standard errors of BP predictions because of each ASV being randomly shuffled was calculated (% IncMSE). A high % IncMSE indicates a taxon that was important for model performance.

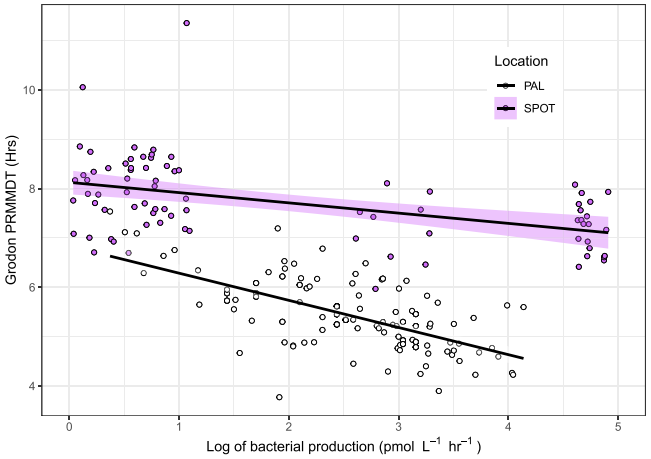

Figure 2. Relationship between log of observed BP (pmol leucine per L^{-1} hr^{-1}) and predicted relative mean minimal doubling time (PRMMDT) from gRodon for both the San Pedro Ocean Time Series (SPOT, $R^2 = 0.30$) and palmer LTER (PAL, $R^2 = 0.49$).

Results

Trends in observed bacterial production, abundance, and community composition

Rates of observed BP (BP_{obs}) ranged from 0.42 to 62.70 pmol L^{-1} hr^{-1} across all sampling locations and years at PAL (Fig. 1C). In the cruise samples (LMG1901, LMG2001, and PD2001) BP_{obs} from the surface (0 m) and above the 50 m mixed layer samples were significantly higher than samples below 50 m (Kruska–Wallis adj *P* value = 6.29×10^{-5}). BP_{obs} showed no significant trends across-shore (Coast/Shelf/Offshore) or alongshore (North/South/Far South, see Fig. 1B) for the LTER grid (Kruskal–Wallis adj *P* values = 0.13 and 0.45, respectively). At SPOT, rates of observed BP (BP_{obs}) ranged from 0.07 to 261.36 pmol L^{-1} hr^{-1} across all years (Fig. 1C). Observed BP was negatively correlated to minimum doubling time estimated by gRodon for both SPOT and PAL samples (Fig. 2, $R^2 = 0.30$ and $R^2 = 0.49$, respectively).

PAL cell abundance ranged from 7.43×10^3 to 6.91×10^5 cells ml^{-1} across all sampling sites and years. In the cruise samples (LMG1901, LMG2001, and PD2001) total cell abundances from the surface (0 m) were significantly higher than samples from the mixed layer and below 50 m (Kruskal–Wallis adj *P* value = 6.60×10^{-16}). Cell abundances showed no significant trends across-shore (Coast/Shelf/Offshore) or alongshore (North/South/Far South) the LTER grid (Kruskal–Wallis adj *P* values = .19 and .78, respectively), and cell abundance was significantly correlated with observed BP (linear regression $R^2 = 0.18$, *P* value = .013).

PAL community composition (the relative abundance of unique ASVs) varied significantly over the sampling locations in a nonmetric multidimensional scaling (NMDS) of the square root of Bray–Curtis distances of Hellinger-transformed relative abundance. An ANOSIM analysis of station samples indicated significant differences in community composition across years (ANOSIM R statistic = 0.07, *P* value = .002) and months across years (i.e. all January data binned together, ANOSIM R statistic 0.17, *P* value = .001). For 2019 and 2020 cruise samples, both alongshore and across-shore stations sites had statistically different community compositions (alongshore N/S/Far S: ANOSIM R statistic 0.02, *P* value .001; across-shore C/S/O: ANOSIM R statistic = 0.06, *P* value = .001).

Table 1. Random forest model performance statistics in our initial model validation (20%) and median cross validation (10% of the data, $n = 99$) for the five models compared in this study.

Model name	Training data (80%) adj R^2	Validation data adj R^2	Training data (100%) adj R^2	Median cross-validation permuted % Var Exp	Median cross-validation RMSE
JOINT-CEG	0.65	0.84	0.98	89.3	20.2
SPOT-CEG	0.87	0.93	0.98	88.2	26.8
SPOT-ENV	0.64	0.32	0.32	47.2	51.4
PAL-CEG	0.76	0.57	0.94	47.4	9.81
PAL-CCG	0.96	0.82	0.96	65.9	7.87

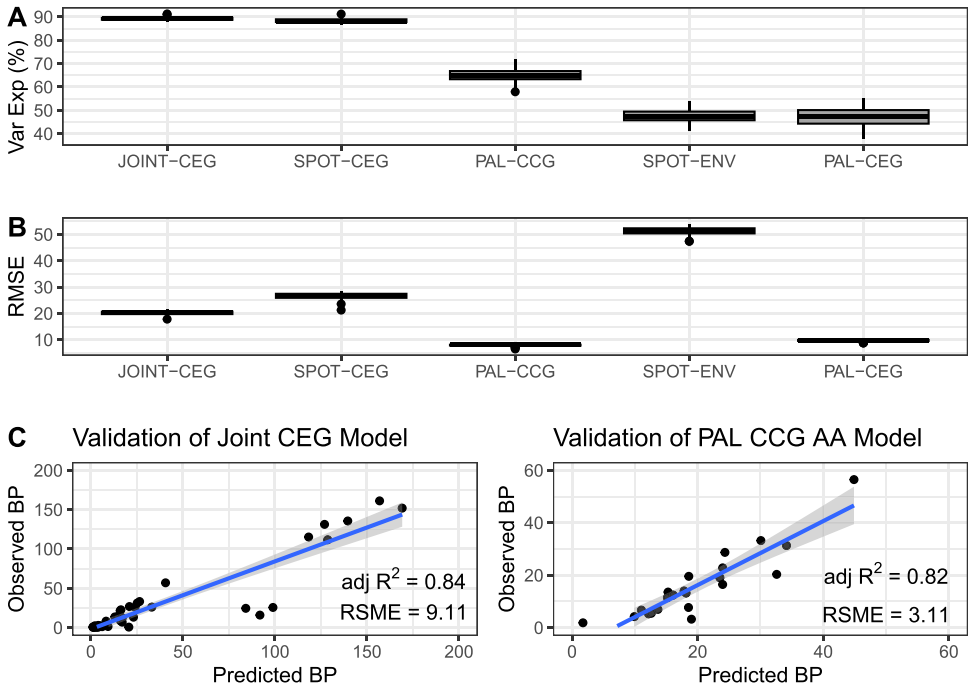

Figure 3. Comparison of cross validation of random forest models where (A) is the percent variance explained (% Var Exp) of each of the models and (B) is square root of the mean square error (RMSE) of each of the models over their cross validation ($n = 99$, with 10% of data withheld). The Joint-CEG model is built from the relative abundance of closest estimated genomes (CEG) of amplicon data from both the San Pedro Ocean Time Series (SPOT) and Palmer LTER (PAL) for all samples with an observation of BP ($n = 241$) and location ("PAL" or "SPOT"). This model is compared to one each of CEG from each site (SPOT-CEG model and PAL-CEG model), from environmental variables at SPOT (SPOT-ENV), and a model built from absolute abundance (relative abundance multiplied by bacterial cell count) of closest completed genomes from PAL (PAL-CCG). Individual models from PAL performed more poorly than the joint model in cross validation, and the environmental model from SPOT performed the worst (with the lowest percent variance explained and highest RMSE). (C) are validation data from the two best performing models, the Joint-CEG and PAL-CCG model.

Comparing random forest model performance

The Joint-CEG model performed best both in our initial model testing (Table 1) and in cross validation when compared to models from the two regions on their own (Fig. 3). BP_{pred} from the random forest regression models matched BP_{obs} from observed samples with high fidelity when testing each of the five models (testing 20% of samples, adj R^2 are listed in Table 1). BP_{pred} and BP_{obs} matched while initially training the model for comparison (80% of samples, adj R^2 are listed in Table 1) and while training the final model (100% of samples). Cross validation of the final random forest regression model indicated a high average variance explained and a low average square root of the mean square error for each of the models (Fig. 3). The Joint-CEG model and PAL-CCG model performed the best in cross validation, while the environmental model from SPOT performed the worst (with the lowest percent variance explained and highest RMSE in Fig. 3). For a direct comparison of the environmental model to our best performing model, we created a Joint-CEG model with the same number of variables (The six taxa with the highest %IncMSE). This truncated Joint-CEG model performed very similarly to the Joint-CEG (In cross validation, % Var Exp = 82.2 and RSME = 19.1).

Palmer Long-Term Ecological Research Project BP_{pred} from the Joint-CEG model

BP_{pred} ranged from 1.09 to 40.17 pmol L^{-1} hr^{-1} incorporated leucine over all sampling locations and years (Fig. 4). In the cruise samples (LMG1901, LMG2001, and PD2001) BP_{pred} from the surface (0 m) and above the 50 m mixed layer samples were significantly higher than samples below 50 m (Kruskal–Wallis adj P value = 6.60×10^{-16}). BP_{pred} showed no significant trends alongshore the LTER grid (Kruskal–Wallis adj P value = .30, North/-South/Far South, see Fig. 1B). However, offshore measurements were significantly lower than coastal and shelf measurements in the across-shore comparison (Wilcox rank sum test adj P value of Offshore vs. Shelf = 8.29×10^{-3}). Finally, we used a Mantel test to examine geospatial correlation for BP_{obs} and BP_{pred}. A geospatial matrix of sample latitude, longitude, and depth was not significant when compared to BP_{obs} (P value = .093) but was

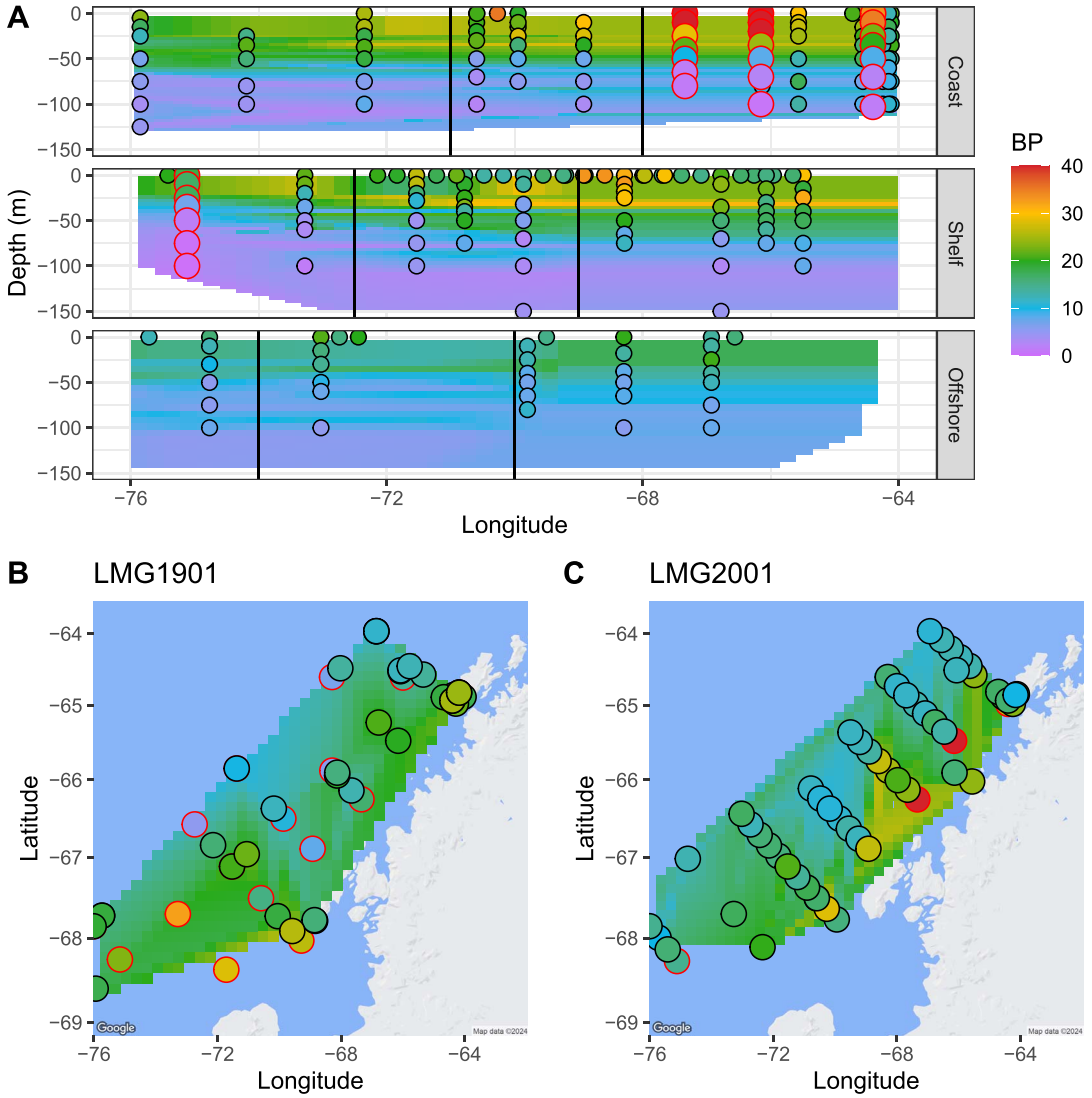

Figure 4. Joint model bacterial production (BP, pmol leucine per L^{-1} hr^{-1}) predictions expand our understanding of Palmer LTER (PAL) BP across the surface PAL LTER 2019 (LMG1901) and 2020 cruises (LMG2001) and depth profiles from LMG2001 and Palmer Deep (PD2001) cruises. Joint model predictions of BP where it is missing in PAL dataset (with linear interpolation behind) closely match observations of BP.

significant for all BP (BP_{obs} & BP_{pred} P value = 1.0×10^{-4}). Finally, BP_{pred} from the Joint-CEG model was correlated to BP_{pred} from the PAL-CCG model (linear regression $R^2 = 0.76$, P value = 2.20×10^{-16}).

Important taxa for predicting bacterial production in both Joint-CEG and PAL-CCG

Random forest regression models report the taxa (ASVs or edges from paprica) that are most important for model performance as percent increase in mean standard error (% IncMSE) when the variable is randomly shuffled. Here, we report the top 20 most important taxa in our two best performing models of BP, the Joint-CEG model (highest cross validation % Variance Explained) and the PAL-CCG model (lowest cross-validation RSME). The top 20 most important taxa from the Joint-CEG random forest model (Fig. 5, with every % IncMSE listed in Supp Table 1) had values that ranged from 555.8% (*Formosa*) to 72% (FCB Group). For this model, only 8 of the top 20 most important taxa had a relative abundance in PAL samples (for instance, *Formosa* is not present in any PAL samples). In SPOT samples with higher BP_{obs} and BP_{pred} than any of the PAL data (<100), the relative abundance of the two most important taxa (which mapped to *Formosa* and *Pelagibacter ubique*) were also high, along with *Puniceispirillum marinum* and *Planktomarina temperata*. In the second best performing model, PAL-CCG model (Fig. 6), the top 20 most important taxa ranged from 70% IncMSE (*Sulfitobacter* spp.) to 2.1% (*Tenacibaculum todarodis*).

Discussion

The application of machine learning–based models in the field of microbial ecology is steadily increasing, including models that can predict diseases caused by microorganisms [41], species interactions [42], and even biogeochemical processes [19]. In our study, we leveraged large datasets of observations from the Palmer LTER and from the SPOT to expand our understanding of how microbial communities lead to specific biogeochemical outcomes. Because community composition data were comparatively common and easy to collect, models that quantitatively link community composition with ecophysiology parameters can greatly improve our understanding of the distribution of rates and standing stocks and can suggest microbial mechanisms (i.e. shifts

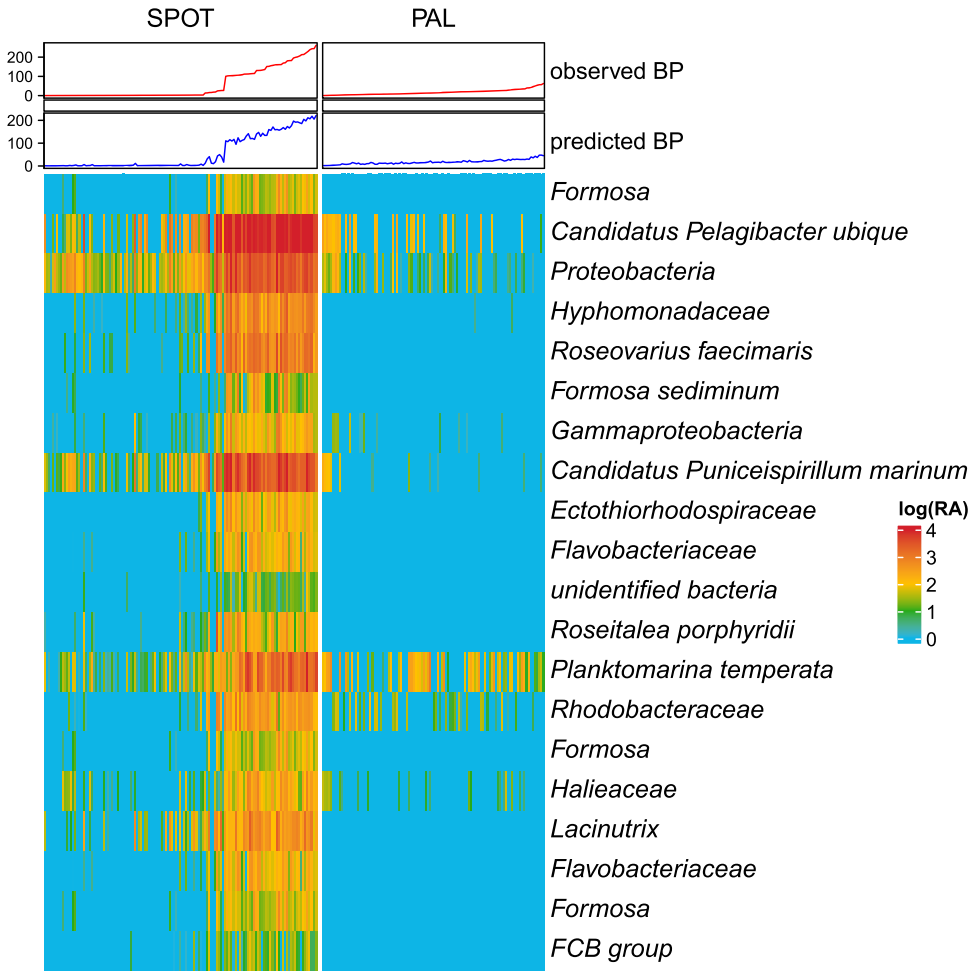

Figure 5. List of top twenty important taxa for the joint model (JOINT-CEG) performance, by percent increase in mean square error (%IncMSE). Heatmap is relative abundance (RA) of each of these taxa over all samples with a matching measurement of BP from the SPOT and PAL datasets. The top heatmap annotation is observed bacterial production (BP, in pmol leucine per L^{-1} hr^{-1}) and predicted BP (pmol leucine per L^{-1} hr^{-1}) from the joint model, ordered from smallest to largest.

in taxon abundance) underlying changes across space and time. We also demonstrated that it is possible to create a single model that predicts rates with good accuracy across regions.

Though we included only two regions here, our results suggest that it is possible to build a single global model that can predict BP from community composition for any region included in the training data. Finally, in our study, all four models built from community composition data outperformed a model built from environmental data. On a regional scale, our best performing model allowed us to predict values where measurements were missing and to determine abundance trends in taxa that are most important for predicting BP. The increased data density from our predictions allowed us to make conclusions about the distribution of BP across the western Antarctic Peninsula study region that were not possible from BP_{obs} alone.

Our input variables for the PAL random forest regression model—measurements for bacterial abundance via flow cytometry and bacterial community composition via 16S rRNA gene sequencing—follow similar patterns as previously reported for Palmer LTER data [3]. Bacterial abundance from 2003 to 2014 ranged from 1×10^3 to 4×10^6 cells ml^{-1}, with higher abundances in coastal waters as in this study [43]. Community composition data from previous studies show similar significant differences across sampling months—with notable community changes in January—and with depth in the water column [16, 44]. BP_{obs} are also within the range of most data from previous years (majority 0–60 pmol L^{-1} hr^{-1} from 2003 to 2014) where higher production was measured in inshore regions [43].

We saw significant geospatial trends in BP when we included BP_{pred} in our analysis, highlighting the potential for gap-filling biogeochemical data with observations of microbial community composition. Our model would be improved by additional measurements of community composition, bacterial abundance, and BP_{obs}, when BP_{obs} is anomalously high (> 100 pmol L^{-1} hr^{-1} in previous years), which are currently unrepresented in our training data from PAL (highest value of 60 pmol L^{-1} hr^{-1}). Even with that limitation, our random forest regression model greatly outperformed linear models on similar Palmer LTER data in previous years [3] at predicting BP. This comes with the caveat that it might not be possible to determine predictive taxa at all sites from a joint model, as some of the predictive data from the joint model seems to have no relationship with BP in PAL (see *Planktomarina temperate* in Fig. 5).

Our comparison of BP to predicted relative mean minimal doubling time (Fig. 2) demonstrated a significant negative correlation between average minimum doubling time and BP. The differences in predicted relative mean minimal doubling time across the two sites may demonstrate ecological differences, where the

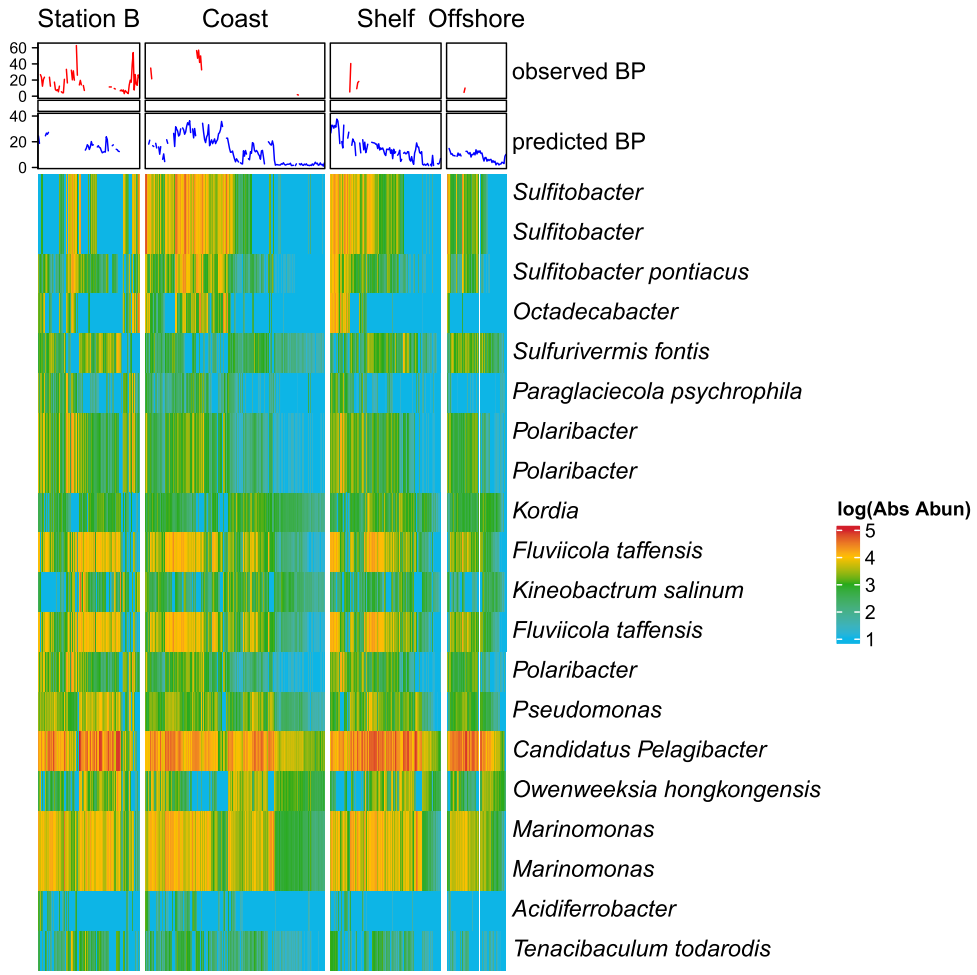

Figure 6. List of top twenty important taxa for the Palmer Station closest completed genome model (PAL-CCG) model, by percent increase in mean square error (%IncMSE). Heatmap is absolute abundance (Abs Abun) of each of these taxa over the PAL dataset, partitioned into LTER station locations. The top heatmap annotation is observed bacterial production (BP, pmol leucine per L^{-1} hr^{-1}) and predicted BP (pmol leucine per L^{-1} hr^{-1}) from the PAL-CCG model.

SPOT bacterial community is always primed for fast growth, while at Palmer, the bacterial community is replaced if it is not actively growing. The minimum doubling time estimated by gRodon in paprica should be treated as an incomplete measure of doubling time, given that the prediction depends on bacteria present in all available completely assembled genomes in RefSeq, where Antarctic seawater bacteria are poorly represented [45]. Improvements in the representation of Antarctic seawater bacteria in sequencing databases, the inclusion of assembled Antarctic genomes in our analysis, as well as more lab-based culture work on Antarctic bacterial growth rate would improve our estimates and our comparison. gRodon output and productivity may have a slightly noisy relationship in that gRodon always estimates a maximum and does not consider interactions between community members. Even with these caveats, it is encouraging to see a significant negative correlation between predicted relative mean doubling time and BP, as it demonstrates the inherent connection between bacterial community genetics (codon usage) and cellular processes (growth rate and BP).

Whereas our random forest regression model trained on data from both California and Antarctica was able to predict BP with success, there are definite drawbacks to the Joint-CEG Model. Foremost, it is probable that ecological differences diminished predictive power across these widely separated regions. This is well demonstrated by *Polaribacter*, a genus that is adapted to and much more abundant in colder climates [46]. Although it is one of the top 20 most important taxa in the PAL-CCG model (by % IncMSE, Fig. 6), it is not an important taxon for the Joint-CEG Model at all (Fig. 5). Overall, SPOT taxa dominated the Joint-CEG Model, with 60% of the top 20 taxa with representatives only at SPOT. Although the RMSE of the Joint-CEG model (20.2) was lower than the SPOT-CEG model (26.8), it was much higher than the RMSE of the PAL-CCG model (7.78), which indicates regional models are important to reduce error in predictions of BP. However, the observed model fidelity and still strong predictive power of the Joint-CEG model suggests that it may be possible to construct global models for BP if training data are drawn from an adequate number of representative regions.

Although the genus *Polaribacter* demonstrated clear ecological differences, the genus *Sulfitobacter* highlighted potential ecological similarities across the two sites. The most important taxa for the PAL-CCG model to predict BP by % IncMSE, *Sulfitobacter* had elevated abundances in samples with higher BP for both the Antarctic and Station SPOT data in California (Figs 5 and 6). In previous work, the family *Rhodobacteraceae* (the family to which *Sulfitobacter* belongs) and *Polaribacter* were the two most abundant bacteria when BP was highest during a phytoplankton bloom at Station B in Antarctica [17]. Members of the *Rhodobacteracae* were

also dominant in the community when BP was the highest in a 5-year analysis of community composition along the wAP [3]. Cultured representatives of Antarctic *Sulfitobacter* have the ability to breakdown various carbohydrates, organic acids, amino acids, and peptides of phytoplankton-derived DOM [47]. In coastal California, *Sulfitobacter* can play an important role in the sulfur cycle by converting dimethylsulfoniopropinate to dimethlysulfide [48], but the extent of this process along the wAP and its relationship to BP is unknown.

Our findings suggest that machine learning methods and especially random forest regressions are important tools to understand the complex datasets inherent to microbial ecology. Random forest regression and similar techniques make it possible to predict biogeochemical rates such as BP from microbial community composition data and even compare it to predictions from environmental data. This approach provides a new technique for filling gaps in biogeochemical data sets and for making predictions where it is not practical to measure rates. By evaluating the efficacy of random forest regression models trained in one ocean biome to another, we can even determine which microbial taxa are globally vs. regionally significant for a process of interest. Our models demonstrate a strong link between microbial community composition and ecosystem functions. We anticipate that with enough training data, it will be possible to construct global BP models that captures most microbial composition–function relationships.

Acknowledgements

The authors would like to thank Emelia Chamberlain and Benjamin Klempay for their assistance with processing microbial samples for genetic sequencing. We would also like to thank all Palmer LTER collaborators, Antarctic Support Contract staff at Palmer Station and the crew of the ARSV *Laurence M. Gould*, without whom this project would not have been possible.

Author contributions

The manuscript was conceptualized by J.S.B and E.C. Funding was acquired by J.S.B, O.S., D.S. and J.F. Investigation was performed by R.T., E.C., N.E., H.D., and S.D. E.C, J.L.W., Y.C.Y., and J.F. did the data curation. J.S.B, J.L.W, A.D and E.J.C developed the software used in the study. The initial draft was written by E.C and edited by all authors.

Supplementary material

Supplementary material is available at *The ISME Journal* online.

Conflicts of interest

None declared.

Funding

Palmer Station Long-Term Ecological Research (PAL-LTER) is supported by the National Science Foundation Office of Polar Programs (NSF OPP-2026045). This work was also supported by a Simons Foundation Early Career Marine Microbial Ecology and Evolution award to J.S.B. (award year 2018) and an NSF CAREER award to JSB (NSF OPP-1846837), by the National Science Foundation (NSF-OCE-1737409 to J.A.F.), the Gordon and Betty Moore Foundation (grant #3779 to J.A.F.), and the Simons Foundation CBIOMES project (grant #549943 to J.A.F.).

Data availability

All sequences are available at NCBI SRA BioProject PRJNA901488 or EMBL under accession PRJEB48162. The code for this manuscript is located on the first author's GitHub, at https://github.com/beth-connors/wAP-predicted-bacterial-production.

References

1. Bier RL, Bernhardt ES, Boot CM *et al.* Linking microbial community structure and microbial processes: an empirical and conceptual overview. *FEMS Microbiol Ecol* 2015;**91**:fiv113. https://doi.org/10.1093/femsec/fiv113
2. Graham EB, Knelman JE, Schindlbacher A *et al.* Microbes as Engines of Ecosystem Function: when does community structure enhance predictions of ecosystem processes? *Front Microbiol* 2016;**7**:214. https://doi.org/10.3389/fmicb.2016.00214
3. Bowman JS, Amaral-Zettler LA, J JR, C ML, Ducklow HW. Bacterial community segmentation facilitates the prediction of ecosystem function along the coast of the western Antarctic peninsula. *ISME J.* 2017;**11**:1460–71. https://doi.org/10.1038/ismej.2016.204
4. Thompson J, Johansen R, Dunbar J *et al.* Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. *PLoS One* 2019;**14**:e0215502. https://doi.org/10.1371/journal.pone.0215502
5. Lin Y, Cassar N, Marchetti A *et al.* Specific eukaryotic plankton are good predictors of net community production in the western Antarctic peninsula. *Sci Rep* 2017;**7**:14845. https://doi.org/10.1038/s41598-017-14109-1
6. Zhao Y, Cordero OX, Tikhonov M. Linear-regression-based algorithms can succeed at identifying microbial functional groups despite the nonlinearity of ecological function. *bioRxiv.* 2024. Preprint at: https://www.biorxiv.org/content/10.1101/2024.01.21.576558v1.
7. Skwara A, Gowda K, Yousef M *et al.* Statistically learning the functional landscape of microbial communities. *Nat Ecol Evol* 2023;**7**:1823–33. https://doi.org/10.1038/s41559-023-02197-4
8. Shan X, Goyal A, Gregor R *et al.* Annotation-free discovery of functional groups in microbial communities. *Nat Ecol Evol* 2023;**7**:716–24. https://doi.org/10.1038/s41559-023-02021-z
9. Bertilsson S, Eiler A, Nordqvist A *et al.* Links between bacterial production, amino-acid utilization and community composition in productive lakes. *ISME J* 2007;**1**:532–44. https://doi.org/10.1038/ismej.2007.64
10. Nielsen UN, Ayres E, Wall DH *et al.* Soil biodiversity and carbon cycling: a review and synthesis of studies examining diversity–function relationships. *Eur J Soil Sci* 2010;**62**:105–16. https://doi.org/10.1111/j.1365-2389.2010.01314.x
11. Cavicchioli R, Ripple WJ, Timmis KN *et al.* Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol* 2019;**17**:569–86. https://doi.org/10.1038/s41579-019-0222-5
12. Azam F, Fenchel T, Field JG *et al.* The ecological role of water-column microbes in the sea. *Mar Ecol* 1983;**10**:257–63. https://doi.org/10.3354/meps010257
13. Cole J, Findlay S, Pace ML. Bacterial production in fresh and saltwater ecosystems: a cross-system overview. *Mar Ecol Prog Ser* 1988;**43**:1–10. https://doi.org/10.3354/meps043001
14. Kirchman DL, Moran XA, Ducklow H. Microbial growth in the polar oceans - role of temperature and potential impact

of climate change. *Nat Rev Microbiol* 2009;**7**:451–9. https://doi.org/10.1038/nrmicro2115
15. Ducklow HW, Schofield O, Vernet M *et al.* Multiscale control of bacterial production by phytoplankton dynamics and sea ice along the western Antarctic peninsula: a regional and decadal investigation. *J Mar Syst* 2012;**98-99**:26–39. https://doi.org/10.1016/j.jmarsys.2012.03.003
16. Luria CM, Amaral-Zettler LA, Ducklow HW *et al.* Seasonal shifts in bacterial community responses to phytoplankton-derived dissolved organic matter in the western Antarctic peninsula. *Front Microbiol* 2017;**8**:2117. https://doi.org/10.3389/fmicb.2017.02117
17. Luria CM, Amaral-Zettler LA, Ducklow HW *et al.* Seasonal succession of free-living bacterial communities in coastal waters of the western Antarctic peninsula. *Front Microbiol* 2016;**7**:1731. https://doi.org/10.3389/fmicb.2016.01731
18. Dittrich R, Henley SF, Ducklow HW *et al.* Dissolved organic carbon and nitrogen cycling along the West Antarctic peninsula during summer. *Prog Oceanogr* 2022;**206**:102854. https://doi.org/10.1016/j.pocean.2022.102854
19. Dutta A, Goldman T, Keating J *et al.* Machine learning predicts biogeochemistry from microbial community structure in a complex model system. *Microbiology Spectrum* 2022;**10**:1–17. https://doi.org/10.1128/spectrum.01909-21
20. Martinson DG, Stammerjohn SE, Iannuzzi RA *et al.* Western Antarctic peninsula physical oceanography and spatio–temporal variability. *Deep-Sea Res II Top Stud Oceanogr* 2008;**55**:1964–87. https://doi.org/10.1016/j.dsr2.2008.04.038
21. Brown MS, Bowman JS, Lin Y *et al.* Low diversity of a key phytoplankton group along the West Antarctic peninsula. *Limnol Oceanogr* 2021;**66**:2470–80. https://doi.org/10.1002/lno.11765
22. Walters W, Hyde ER, Berg-Lyons D *et al.* Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *mSystems* 2016;**1**:1–10. https://doi.org/10.1128/mSystems.00009-15
23. Callahan BJ, McMurdie PJ, Rosen MJ *et al.* DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;**13**:581–3. https://doi.org/10.1038/nmeth.3869
24. Schofield O. Bacterial Properties in Discrete Water Column Samples. *NOAA ERDDAP*. 2019. Access online at: https://pallter-data.marine.rutgers.edu/erddap/info/index.html?page=1&itemsPerPage=1000.
25. Wilson JM, Chamberlain EJ, Erazo N *et al.* Recurrent microbial community types driven by nearshore and seasonal processes in coastal Southern California. *Environ Microbiol* 2021;**23**:3225–39. https://doi.org/10.1111/1462-2920.15548
26. Wehrens R, Kruisselbrink J. Flexible self-organizing maps in kohonen 3.0. *J Stat Softw* 2018;**87**:1–17. https://doi.org/10.18637/jss.v087.i07
27. Yeh Y-C, Fuhrman JA. Contrasting diversity patterns of prokaryotes and protists over time and depth at the San-Pedro Ocean Time series. *ISME. Communications* 2022;**2**:1–12. https://doi.org/10.1038/s43705-022-00121-8
28. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 2011;**17**:10–2. https://doi.org/10.14806/ej.17.1.200
29. Bushnell B, Rood J, Singer E. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS ONE* 2022;**12**:1–15. Access online at: https://sourceforge.net/projects/bbmap/.
30. Yeh Y, Fuhrman JA. Environmental data, nutrients, and leucine and thymidine bacterial production from samples collected by CTD during cruises in the San Pedro Channel on R/V yellowfin from 2005 to 2018. *Biological and Chemical Oceanography Data Management Office (BCO-DMO)* 2023. Access online at: https://www.bco-dmo.org/dataset/885939
31. Kirchman DL, K'Nees E, Hodson R. Leucine incorporation and its potential as a measure of protein synthesis by bacteria in natural aquatic systems. *Appl Environ Microbiol* 1985;**49**:599–607. https://doi.org/10.1128/aem.49.3.599-607.1985
32. Bowman JS, Ducklow HW. 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* 2015;**10**:e0135868. https://doi.org/10.1371/journal.pone.0135868
33. Czech L, Barbera P, Stamatakis A. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics* 2020;**36**:3263–5. https://doi.org/10.1093/bioinformatics/btaa070
34. Barbera P, Kozlov AM, Czech L *et al.* EPA-ng: massively parallel evolutionary placement of genetic sequences. *Syst Biol* 2019;**68**:365–9. https://doi.org/10.1093/sysbio/syy054
35. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;**29**:2933–5. https://doi.org/10.1093/bioinformatics/btt509
36. Haft DH, Dicuccio M, Badretdin A *et al.* RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res* 2018;**46**:D851–60. https://doi.org/10.1093/nar/gkx1068
37. Weissman JL, Hou S, Fuhrman JA. Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proc Natl Acad Sci USA* 2021;**118**:1–10. https://doi.org/10.1073/pnas.2016810118
38. Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002;**2**:5.
39. Kursa MB, Rudnicki W. Feature selection with the Boruta package. *J Stat Softw* 2010;**36**:1–13.
40. Evans JF, Murphy MA. rfUtilities: R package version 2.1–3. 2020; https://cran.r-project.org/package=rfUtilities.
41. Yuan J, Wen T, Zhang H *et al.* Predicting disease occurrence with high accuracy based on soil macroecological patterns of fusarium wilt. *ISME J* 2020;**14**:2936–50. https://doi.org/10.1038/s41396-020-0720-5
42. Lee JY, Sadler NC, Egbert RG *et al.* Deep learning predicts microbial interactions from self-organized spatiotemporal patterns. *Comput Struct Biotechnol J* 2020;**18**:1259–69. https://doi.org/10.1016/j.csbj.2020.05.023
43. Henley SF, Schofield OM, Hendry KR *et al.* Variability and change in the West Antarctic peninsula marine system: research priorities and opportunities. *Prog Oceanogr* 2019;**173**:208–37. https://doi.org/10.1016/j.pocean.2019.03.003
44. Luria CM, Ducklow HW, Amaral-Zettler LA. Marine bacterial, archaeal and eukaryotic diversity and community structure on the continental shelf of the western Antarctic peninsula. *Aquat Microb Ecol* 2014;**73**:107–21. https://doi.org/10.3354/ame01703
45. Bowman JS. Identification of microbial dark matter in Antarctic environments. *Front Microbiol* 2018;**9**:3165. https://doi.org/10.3389/fmicb.2018.03165
46. Bowman JP. Polaribacter. *Bergey's Manual of Systematics of Archaea and Bacteria* 2018;**119**:1–21.
47. Choi S-B, Kim J-G, Jung M-Y *et al.* Cultivation and biochemical characterization of heterotrophic bacteria associated with phytoplankton bloom in the Amundsen Sea polynya, Antarctica.

Deep-Sea Res II Top Stud Oceanogr 2016;**123**:126–34. https://doi.org/10.1016/j.dsr2.2015.04.027

48. Curson AR, Rogers R, Todd JD *et al.* Molecular genetic analysis of a dimethylsulfoniopropionate lyase that liberates the climate-changing gas dimethylsulfide in several marine alpha-proteobacteria and Rhodobacter sphaeroides. *Environ Microbiol* 2008;**10**:757–67. https://doi.org/10.1111/j.1462-2920.2007.01499.x