scientific reports

Check for updates

OPEN

# Microbial signatures define the ecosystem functions of the pelagic microbiome in a basin-scale, Southwest Atlantic Ocean

Natascha Menezes Bergo[1,6]✉, Francielli Vilela Peres[1,6], Danilo Candido Vieira[1], Flúvio Modolon[1], Julio Cezar Fornazier Moreira[1], Rebeca Graciela Matheus Lizárraga[1], Renato Gamba Romano[1], Amanda Goncalves Bendia[1], Leandro Nascimento Lemos[2,5], Alice de Moura Emilio[1], Augusto Miliorini Amendola[1], Diana Carolina Duque Castano[1], Mateus Gustavo Chuqui[1], Fabiana S. Paula[1], William Soares Gattaz Brandão[1], Gustavo Fonseca[3], Ana Tereza R. Vasconcelos[5], Célio Roberto Jonck[4], Daniel Leite Moreira[4], Frederico Pereira Brandini[1] & Vivian Helena Pellizari[1]

**The pelagic environment represents a mosaic of biogeographical domains shaped by regional oceanographic processes. Here, a coastal-to-open ocean microbiome investigation was conducted from 64 water samples of the Santos Basin (SB), located in the subtropical South Atlantic Ocean. We combined shotgun metagenomics with a hybrid machine learning workflow to investigate the taxonomic diversity, community structure, and ecosystem functions of pelagic microbiomes. The workflow integrated self-organizing maps (unsupervised) for pattern discovery and Random Forest (supervised) for predictive modeling. Unsupervised machine learning revealed a clear spatial and vertical (light-driven) distribution, with indicator taxa reflecting biogeochemical patterns consistent with global surveys. Supervised learning identified phosphate, salinity, and nitrate, influenced by local upwelling and La Plata River plume, as the primary environmental drivers of microbial community structure. In terms of functionality, the SB microbiome displayed depth- and region-specific patterns: photoautotrophs and nitrogen fixers dominated photic waters (with differences between coastal and oceanic stations), whereas chemolithoautotrophs and mixotrophs prevailed in the aphotic zone. Notably, nitrification signatures were more frequent in northern mesopelagic communities, while sulfur-oxidation pathways were enriched toward the south. Genes for CO bio-oxidation and dimethylsulfoniopropionate (DMSP) degradation were present across all depths. Furthermore, potential non-cyanobacterial diazotrophs were detected in the deep waters, underscoring previous underappreciated to nitrogen cycling. Our findings indicated that the Santos Basin hosts a functionally diverse microbiome including putative novel lineages. The taxonomic and functional patterns observed in the SB might provide insights into potential ecological responses to shifts in nutrient dynamics and physical processes. This investigation provides an ecogenomic baseline for understanding the microbial ecosystem services in subtropical oceans and reveals the potential of machine learning to uncover ecological patterns in underexplored marine regions.**

The oceans are the largest biome on Earth[1]. They are heterogeneous across scales of depth, nutrient availability, microbial distribution, dispersion, and interaction. These habitat variations in the pelagic environment directly influence ecosystem functioning and services of the planet[2]. For example, in the photic layer, photoautotrophic microorganisms are responsible for 46% of the global primary productivity, contributing substantially

[1]Oceanographic Institute, Universidade de São Paulo, São Paulo, Brazil. [2]Ilum-School of Science, Brazilian Center for Research in Energy and Materials, Campinas, Brazil. [3]Instituto Do Mar, Universidade Federal de São Paulo, São Paulo, Brazil. [4]PETROBRAS Research Center, Centro de Pesquisas Leopoldo Américo Miguez de Mello (PETROBRAS/CENPES), Rio de Janeiro, Brazil. [5]National Laboratory for Scientific Computing, Petrópolis, Brazil. [6]Natascha Menezes Bergo and Francielli Vilela Peres contributed equally to this work. ✉email: nataschabergo@gmail.com

to atmospheric carbon fixation by photosynthesis[3–5]. In both photic and aphotic zones, heterotrophic microorganisms drive the remineralization of nutrients and organic matter, facilitating carbon export to higher trophic levels and the deep ocean[6]. In addition to light-driven primary production, the ocean also supports a significant chemosynthetic productivity through dark carbon fixation fueled by the oxidation of inorganic compounds (e.g., ammonium and sulfur)[7]

The metabolic diversity of prokaryotes in the water column plays an important role in marine biogeochemical cycles[8,9]. Over the past decade, global-scale expeditions and advances in omics technologies have greatly improved our understanding of microbial metabolic functions and their roles in oceanic element cycling[10–14]. Nevertheless, despite the increasing number of studies on the global ocean, knowledge about planktonic prokaryotic metabolisms in the South Atlantic Ocean (SAO) remains limited[15,16]. Most SAO microbiome studies have focused on prokaryotic abundance, biomass[17–19], and taxa diversity[20–22], with genomic surveys indicating that bacterial and archaeal diversity is poorly characterized in this region[15,16]. These findings suggest that microbial communities in the SAO exhibit small- to large-scale shifts in distribution and functionality, yet their contribution to ecosystem services is still insufficiently resolved.

The Santos Basin (SB), located in the SAO, is ecologically and economically significant. It sustains ecosystem services relevant to regional fisheries and is an area of intense offshore oil exploration[23]. As a typical deep-water petroliferous region, the SB faces several environmental stressors, including coastal pollution[24–26], overfishing[27], and anthropogenic climate change impacts[28]. Characterizing its marine microbiome under these ongoing pressures is therefore timely and relevant[30].

Recent advances in metagenomic sequencing combined with machine learning (ML) approaches provide a powerful framework for capturing microbial diversity and ecological function. Previous studies have demonstrated that ML can identify ecological patterns, integrate high-dimensional data, and enhance predictive understanding of microbial communities[30,31]. These tools offer critical opportunities for monitoring, preserving, and managing marine ecosystems that are increasingly impacted by human activity[29,32].

In this study, we investigated the diversity, function, and spatial distribution of planktonic prokaryotic communities in the Santos Basin (South Atlantic Ocean) using integrated metagenomic and ML approaches. Our objectives were to (1) characterize the taxonomic and functional diversity of pelagic Bacteria and Archaea across pelagic environment; (2) recover metagenome-assembled genomes (MAGs) and assess their potential roles in regional biogeochemical cycles; and (3) identify the oceanographic drivers shaping microbial community structure.

## Results

### Characterization of santos basin metagenomic reads using unsupervised learning

Machine learning analysis identified ecologically relevant prokaryotes by assessing statistically supported associations among reads' taxonomy in each sample collected in the Santos Basin (Fig. 1 and 2). The SOM analysis stabilized with a learning rate of approximately 0.0006 for genus-level and 0.06 for geographical data. The resulting SOM network explained 88.70% of total data variance, with a mean topographic error of 0.36. Hierarchical clustering, using the elbow method and split moving window analysis to determine the dendrogram cut, identified four taxonomic associations (Assoc) (Fig. 1B-C), each exhibiting clear vertical and spatial distribution patterns across the SB (Fig. 2A-E): (i) In the photic environment (1 m and DCM depths), Assoc 1 (black) predominated in coastal epipelagic samples and Assoc 2 (cyan) was dominant in oceanic epipelagic samples; (ii) Assoc 3 (orange) and 4 (red) were both prevalent in the aphotic environment (250, 900 and 2300 m depths), occurring in northern and southern regions of the basin, respectively.

*Bacteria* accounted for most classified reads across all taxonomic associations (93.2%), while *Archaea* represented 6.7% (Table S1). We focused on the 15 most abundant bacterial and archaeal genera and five 5 indicator taxa per association and observed vertical and spatial distribution patterns (Fig. S1 and Fig. 2).

The photic layer (1 m and DCM) generally presented a similar taxonomic composition was dominated by Cyanobacteria *Synechococcus* and *Prochlorococcus*, as well as Indicator taxa analysis supported these patterns (Assoc 1 and 2, Fig. S1). As expected, distinct patterns were observed within the order *Synechococcales*: *Synechococcus* was dominant in the coastal epipelagic samples (Assoc 1), whereas *Prochlorococcus* predominated in the oceanic waters (Assoc 2, Fig. S1). In terms of indicator taxa (Fig. 2F), *Synechococcus*, *Cyanobium, Candidatus Puniceispirillum, Formosa* and *Planktomarina (Roseobacter clade)* were indicator taxa for the coastal waters and the representatives of *Prochlorococcus*, *Trichodesmium*, *Candidatus Atelocyanobacterium*, *Candidatus Nesciobacter* and *Paraphotobacterium* for oceanic waters. Metabolic profiles in the photic layer (Fig. S2 and Table S3) reflected phototrophic lifestyles, such as the enrichment of photosystem and phycobilisome pathways (present in the *Synechococcales*) and photoheterotrophy with rhodopsin pump (generally present in *Candidatus Pelagibacter).*

The taxonomic composition changed below the photic zone, with distinct communities at mesopelagic (250 and 900 m) and bathypelagic (2300 m) zones, corresponding with Assoc 3 and 4 (Fig. S1). Spatial heterogeneity was evident within the aphotic microbiome. For example, bacterial representatives *Alternomonas, Pseudoalteromonas* and *Sulfitobacter* were abundant in southern samples (Assoc 4), while the *Candidatos Nitrosomarinus*, *Nitrosopumilus, Gordonia* and *Candidatos Thioblobus* were more abundant in northern samples (Assoc 3). Among Archaea, *Nitrosopumilus* and *Candidatus Nitrosopelagicus* were prevalent across all aphotic samples (Assoc 3 and 4, Fig. S1). Indicator taxa analysis supported these patterns (Fig. 2F): *Nitrosopumilus, Nitrososphaera* and *Candidatus Nitrosotalea* were identified as characteristic of northern aphotic microbiome, whereas *Sulfitobacter, Pseudoalteromonas* and *Alteromonas* were associated with southern region.

Functionally, the aphotic zone of the SB exhibited high metabolic diversity. The most abundant pathway was related to nitrogen cycling, including nitrate/nitrite ammonification, denitrification (yielding nitrogen gas) and dissimilatory nitrate reduction to ammonium. Other metabolic of pathways coexisting in dark waters included

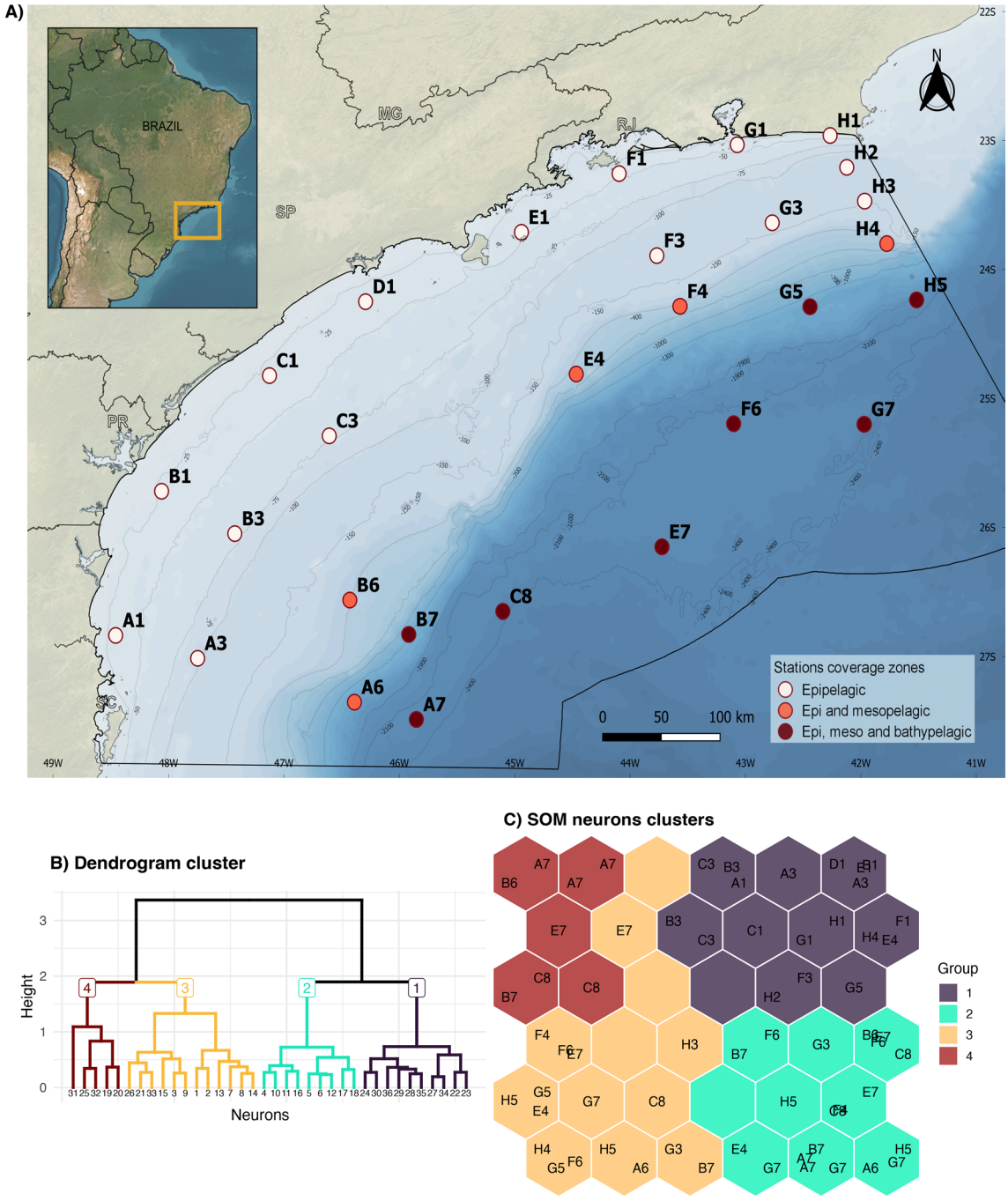


**Fig. 1**. Sampling map of the Santos Basin in the South Atlantic Ocean and results from the unsupervised phase of the hybrid machine learning. (**A**): The map shows the basin's location relative to the Brazil coastline and the oceanographic stations sampled during the SANAGU expeditions (August – October 2019). Color dots represent sampling coverage: black indicates stations with epipelagic sampling, orange for epipelagic and mesopelagic and white for epipelagic, mesopelagic, and bathypelagic sampling. Abbreviations represent the depth zones: Epi (epipelagic) and Meso (mesopelagic). (**B**): Dendrogram obtained by the hierarchical clustering of the neurons from the Self Organizing Map (SOM); (**C**): SOM with neurons grouped by the respective clusters. Colors represent the associations: association 1 (black), association 2 (cyan), association 3 (orange) and association 4 (red).

methanogenesis from methylated compounds, inorganic sulfur assimilation, lysine fermentation and anaerobic respiratory reductases (Fig. S2 and Table S3).

In agreement with spatial taxa distribution and potential metabolic function, Shannon diversity index differed significantly across the four associations (Kruskal–Wallis, $\chi^2 = 33.18$, $p < 0.001$) (Fig. 2G). Post-hoc Dunn tests for the Shannon index revealed that Assoc 3 and 4 had significantly higher median values than Assoc 1 and 2

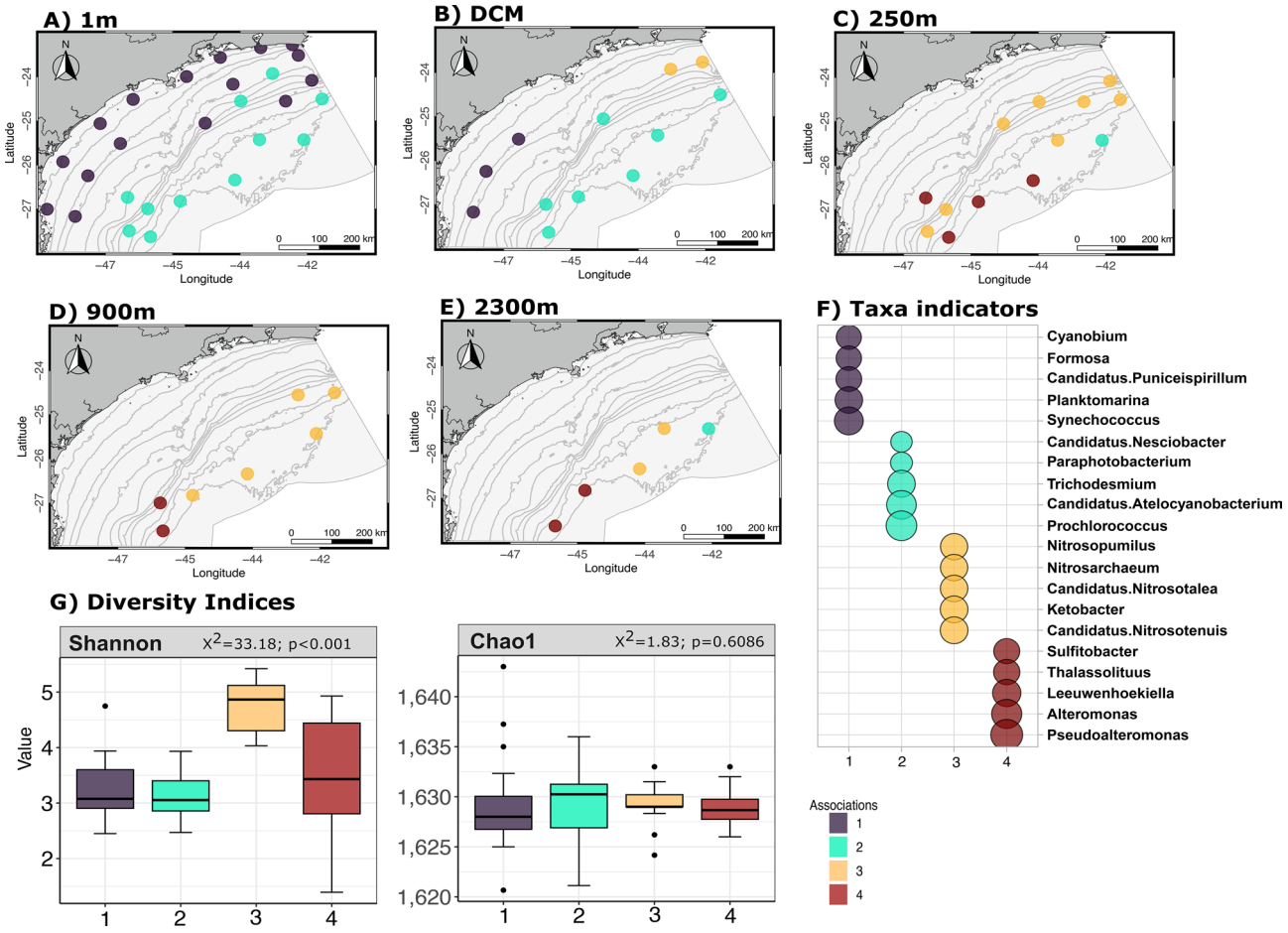


**Fig. 2**. Results from the unsupervised phase of the hybrid machine learning. (**A-E**): Spatial distribution of taxonomic associations across sampled depths in the Santos Basin. (**F**): Top five significant prokaryotic ($p < 0.05$) associated with each cluster, where point size represents the IndVal statistic, reflecting the association strength. (**G**): Shannon and Chao1 indices calculated for each association, with corresponding chi-square ($\chi^2$) values and p-values shown for each statistical comparison. Colors represent the associations: association 1 (black), association 2 (cyan), association 3 (orange) and association 4 (red).

(Dunn test, $p < 0.01$; Table S4). In contrast, no significant differences were found for the Chao1 richness index (Kruskal–Wallis, $\chi^2 = 1.88$, $p = 0.6$) (Fig. 2G), suggesting that genus richness is relatively constant across depths. These results indicate that taxonomic diversity is elevated in the aphotic microbiome, represented by Assoc 3 and 4.

## Predictive modeling of Santos Basin metagenomic reads using supervised learning

To assess the relationship between taxonomic structure and environmental conditions, we applied a RF classification model using the associations defined earlier and environmental metadata (Table S2). The RF model achieved an accuracy of 82% (training) and 90% (test) dataset (Fig. 3A). Within the training data, Assoc 1 and 2 had the lowest classification error (15%), while Assoc 4 had the highest error (23%). In the test dataset, Assoc 3 misclassified 50% as Assoc 1. Among the 30 environmental variables evaluated, 15 were identified as significant predictors ($p < 0.05$). Phosphate was the most important feature with a mean minimal depth of 2.35, followed by salinity, nitrate and colored dissolved organic matter (CDOM) (Fig. 3B). Environmental conditions differed significantly across associations (Fig. 3C): Assoc 1 was characterized by higher CDOM and phycoerythrin, lower salinity and moderated temperature; Assoc 2 showed elevated salinity and temperature and lower CDOM and phosphate; Assoc 3 and 4 were defined by higher phosphate and nitrate, lower temperature, and moderated salinity; CDOM also was a distinguishing feature of Assoc 4.

## Genome-recovery metagenomics reveals taxonomic novelty

The co-assembly of pelagic metagenomes, combined with automated binning and manual refinement, allowed for the reconstruction of 307 medium- or high-quality Santos Basin Metagenome-Assembled Genomes (SB-MAGs). Of these MAGs, 59 were classified as high-quality drafts, and 248 were classified as medium quality drafts, with an average genome completeness of 76.9%, according to the genome quality standards suggested by Bowers et al.[33]. Summary statistics of SB-MAGs are provided in Table S5.

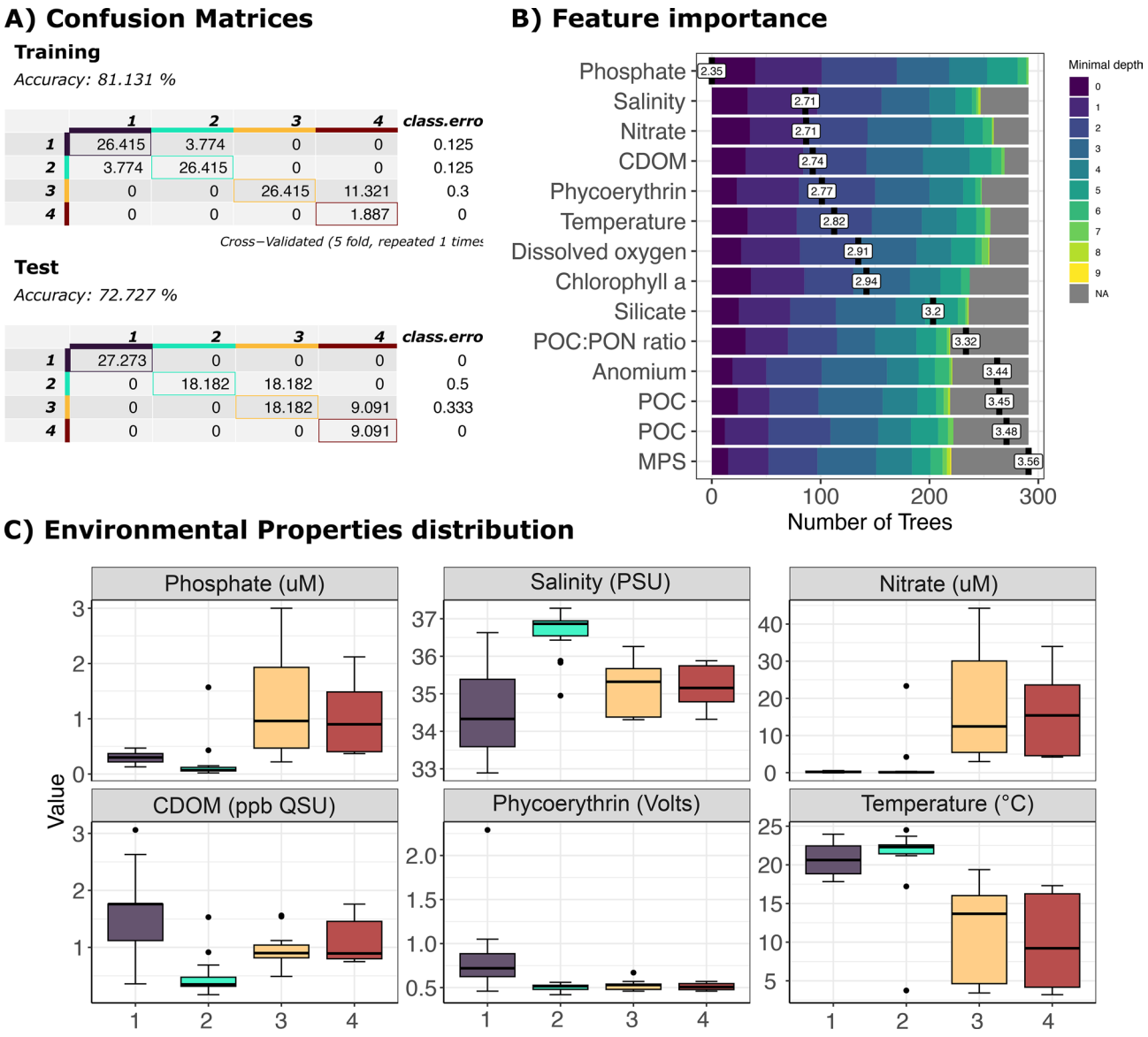


**Fig. 3**. Model performance, predictor importance, and environmental variant across taxonomic associations. (**A**): Classification accuracy and errors (misclassification rates) for training and test datasets based on random forest models, displaying actual versus predicted classes. (**B**): Relative importance of environmental predictors ranked by mean minimal tree depth across decision trees. (**C**): Box plots showing the distribution of key environmental parameters across the four associations. Colors codes: association 1 (black), association 2 (cyan), association 3 (orange) and association 4 (red).

SB-MAGs were recovered from all SB depths (Fig. S3). The number of recovered MAGs was higher at 250 m depth, followed by 1 m, 2300 m, DCM and 900 m. A total of 256 bacterial SB-MAGs were taxonomically assigned to 17 phyla, while 51 archaeal SB-MAGs were assigned to 2 phyla based on the GTDB taxonomy classification (Table S5). Proteobacteria (class *Gammaproteobacteria* and *Alphaproteobacteria*, n = 81) followed by *Planctomycetota* (n = 48) were the most abundant bacterial phyla in the SB-MAGs, while archaeal MAGs were assigned to the phyla *Thermoplasmatota* (class *Nitrososphaeria*, n = 31) and *Thermoproteota* (class *Poseidonia*, n = 18).

The SB-MAGs included putative taxonomic novelty, with 45% of bacterial and 42% of archaeal MAGs classified as novel species according to GTDB-Tk (Fig. 4). In addition, two MAGs recovered from 900 m depth were not classified at any taxonomic level by GTDB-Tk.

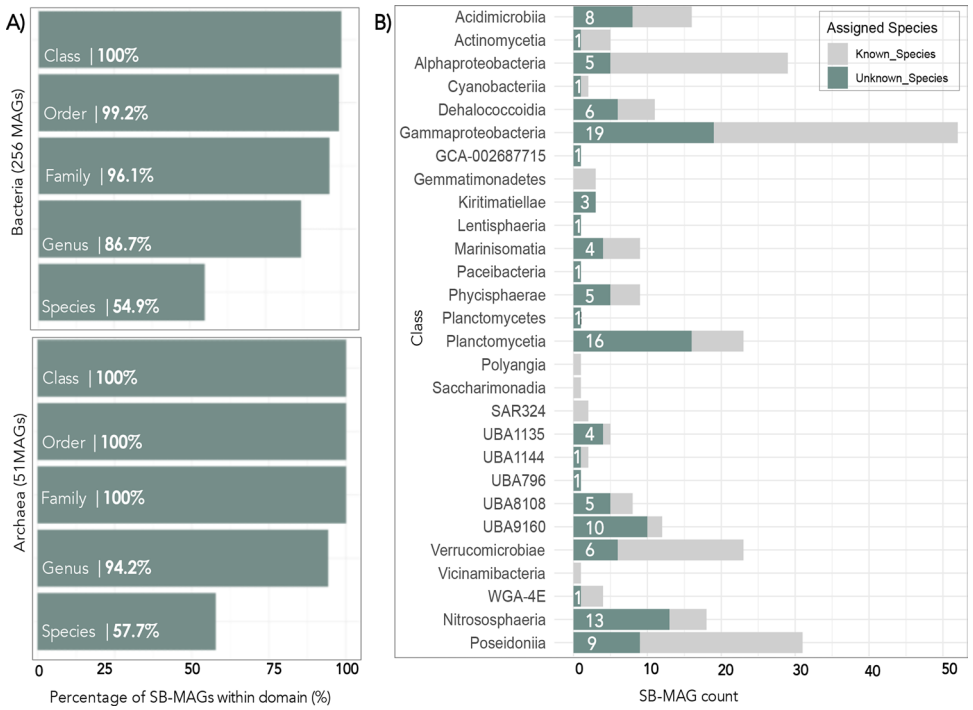


**Fig. 4**. Stacked bar plot for illustrating the taxonomic novelty among the Santos Basin Metagenome-Assembled Genomes. (**A**): Proportion of MAGs (X axis) with classification according to their taxonomic ranks (Y axis) for archaea and bacteria. (**B**): Number of MAGs (X axis) exhibiting novelty at class level (Y axis). Taxa classified by GTDB-Tk are show in gray; unclassified taxa are shown in green.

## The relative abundance of Santos Basin Genomes aligns with machine learning-based associations

Despite use of a co-assembly strategy, the differences in the relative abundances of the genomes explored in the next section followed spatial distribution patterns consistent with those inferred from the SOM clustering (Fig. S5 and Fig. S6).

In the photic layer, particularly in the MAGs recovered from 1 m depth, high relative abundances were observed for: *UBA868* clade, cyanobacterium *NIES-981 genus* and *MED-G52 sp001627375* in the coastal waters (Assoc 1); *Pirellulaceae* MAGs and cyanobacterium *Atelocyanobacterium thalassa_A spp* in the oceanic waters (Assoc 2). Some genomes showed a higher relative abundance in both coastal and oceanic photic samples (Assoc 1 and 2) such as *Gordonia sp002700145* and *MED-G52 sp002457055*. Within the genomes from the DCM layer, four *Pirellulaceae* MAGs exhibited enrichment in the oceanic waters (Assoc 2). Some genomes showed a higher relative abundance into two associations: *Rhodobacteraceae* MAG in both coastal photic samples (Assoc 1) and north aphotic samples (Assoc 3), and the *Sphingomonadaceae* MAG (PMC_MAG_59) in both oceanic photic (Assoc 2) and north aphotic (Assoc 3) samples (Fig. S4).

In the aphotic zone, several MAGs showed distinct spatial enrichment patterns. At 250 m, higher relative abundances in northern samples (Assoc 3) were observed for *Ketobacteraceae*, *Sphingomonadaceae*, *Pseudomonadales*, *Nitrosopumilaceae* MAGs and *Rhizobiaceae*. In contrast, *Alteromonadaceae*, *UBA826* and *Rhodobacteraceae* MAGs were enriched in in the southern waters (Assoc 4). Two MAGs, *Pseudohongiellaceae* and *Acidimicrobiales*, showed broad distribution in both Assoc 3 and 4.

At 900 m, the *Mycobacteriaceae* MAG (900M_MAG_05) was higher relative abundance in the north samples (Assoc 3) while *Nitrosopumilaceae* MAG (900M_MAG_29) was enriched in the south samples (Assoc 4). Three of the genomes, *Sphingomonadaceae Pseudohongiellaceae* and *UBA868*, were enriched in both northern and southern samples (Assoc 3 and 4). Finally, at 2300 m depth (Assoc 2 and 4), a higher abundance of *Nitrosopumilaceae* MAG was observed in the oceanic waters (Assoc 2) compared to other associations (Fig. S5). These depth- and association-specific abundance patterns suggest ecological partitioning and functional adaptation to environmental gradients in the SB. Abundance for all SB-MAGs are provided in Table S6.

## Genome-resolved metabolic profiles reveal distinct roles in biogeochemical cycles

MAGs classified as *Alphaproteobacteria* (73% of average of genome completeness—AGC), *Gammaproteobacteria* (84% of AGC), *Acidimicrobiia* (81% of AGC) and *Nitrososphaeria* (79% of AGC), encoded the highest numbers of genomes with high-level completeness and biochemical cycles metabolic marker genes (Fig. S6A). Among the total recovered MAGs, we selected 44 genomes (> 60% pathway complete) with potential for autotrophy, chemolithoautotrophy, mixotrophy, and nitrogen fixation across depths (Fig. 5). Complete information on MAGs with > 60% or more of the complete metabolic pathway is shown in Table S7.

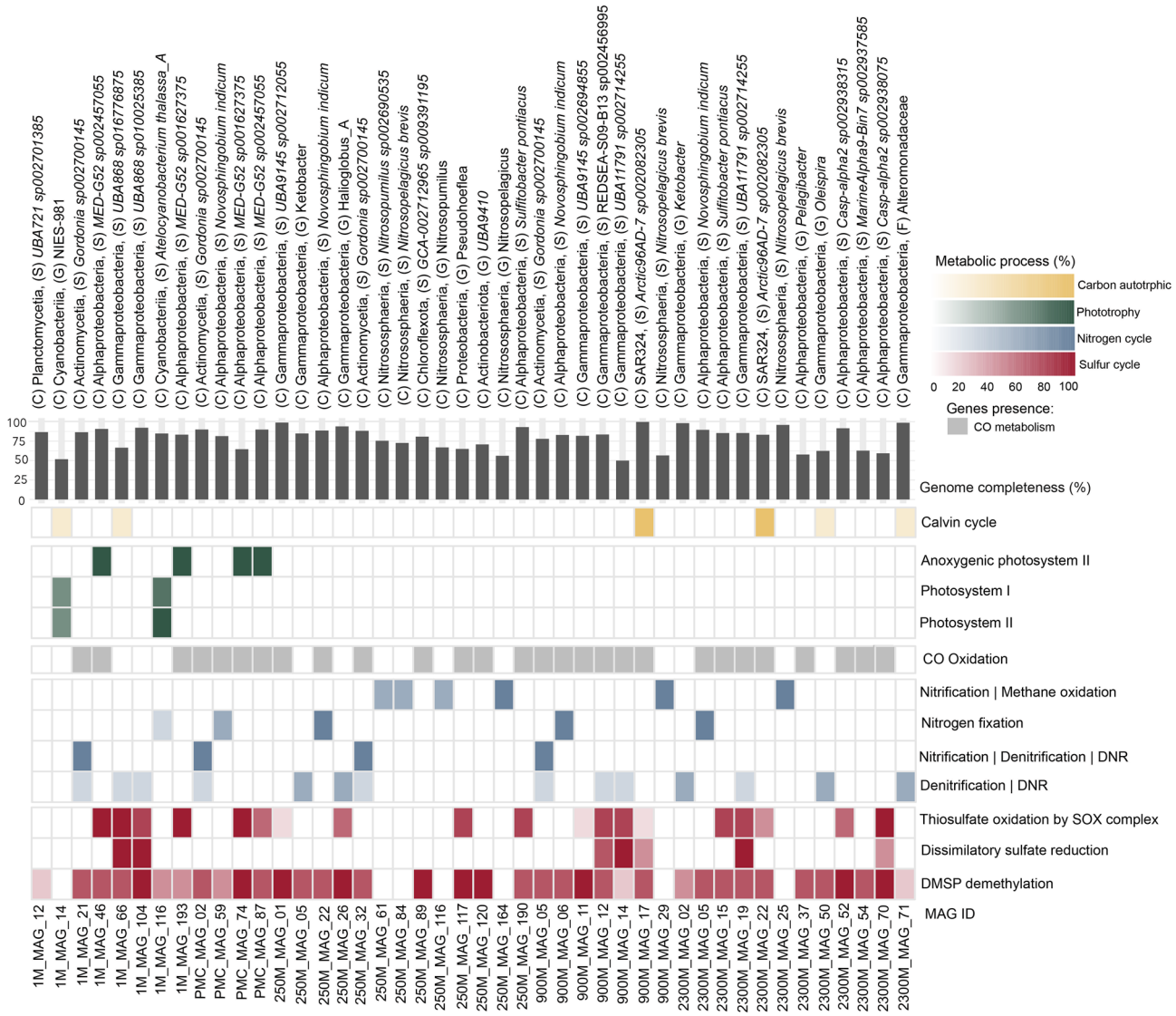


**Fig. 5**. Metabolic potential of Santos Basin in selected metagenome-assembled genomes (SB-MAGs). Heatmap of metabolic pathway completion percentages across 44 SB-MAGs. The percentage module completeness for each pathway is coded by color intensity. Only module completeness equal to or greater than 60% were plotted. The phylogenomic taxonomic classification of the MAGs is indicated at the top of the figure: Class (C), Genus (G), and Species (S).

## MAGs from the photic layer

As expected, key phototrophic and photoheterotrophic microorganisms were exclusive to the photic zone, including *Cyanobacteria* and *Rhodobacteraceae* (genus *MED-G52*), which were exclusive to this layer (Fig. S6B and Fig. 5). Genes encoding subunits of Photosystem I (*psaABCDEF*) and II (*psbABCDEF*) were affiliated with the cyanobacterium *NIES-981 genus* (1M_MAG_14) and *Atelocyanobacterium thalassa_A spp* (1M_MAG_116). These genomes display 67% and 100% completeness of the photosynthetic metabolic pathway, respectively, and both contain the *nifK* gene for nitrogen fixation (Fig. 5).

Photoheterotrophic lifestyle, using all genes that encode subunits in the anoxygenic photosynthesis type II reaction center (*pufML*) was detected in four *Rhodobacteraceae* MAGs: *MED-G52 sp002457055* (1M_MAG_46 and PMC_MAG_87) and *MED-G52 sp001627375* (1M_MAG_193 and PMC_MAG_74) (Fig. 5). These genomes also encoded > 65% completeness of the thiosulphate oxidation by SOX complex (*soxABCDXYZ*), genes for CO-Oxidation (*coxL, cutL, coxM, cutM and coxS*) and the 3-hydroxypropionate bi-cycle (*mct* and *meh*), indicating potential for chemoautotrophy. Notably, in the *MED-G52 sp001627375* (PMC_MAG_74), no *mct* and *meh* genes were detected. Genes encoding type I reaction center for photoheterotrophic anoxygenic photosynthesis were absent (Fig. S6A).

Unexpectedly, two MAGs from the *UBA868* family exhibited genes for sulfur metabolism (Fig. 5): *UBA868 sp016776875* (1M_MAG_66) and *UBA868 sp010025385* (1M_MAG_104). Both encoded partial (83%) thiosulfate oxidation by SOX complex and complete dissimilatory reduction of sulfate (*sat* and *aprAB* genes,

respectively). In addition, *UBA868 sp010025385* has a complete DMSP demethylation pathway (*dmdA, dmdB, dmdC* and *dmdD* genes).

Evidence for chemoautotrophy in the photic SB was observed in two MAGs high-quality genomes (>92% completeness) (Fig. 5). The genome of *Novosphingobium indicum* (PMC_MAG_59) encoded genes for CO oxidation and nitrogen fixation (*nifH, nifD* and *nifK*), and partial (>70%) DMSP demethylation, indicating a potential non-cyanobacterial diazotroph. Conversely, MAGs of *Gordonia sp002700145* (1M_MAG_21 and PMC _MAG_02) suggested chemoautotrophy supported by reduced nitrogen compounds, encoding all genes for CO oxidation and nitrate reduction (*narG, narZ, nxrA, narH, narY* and *nxrB*) and partial DMSP demethylation (>60%).

## MAGs from the twilight and aphotic zones

A comparative analysis revealed that several functional traits were shared between photic and aphotic waters, particularly involving sulfur, nitrogen and carbon monoxide metabolisms (Fig. S6A and Fig. 5). These functions, however, are more prevalent in the twilight and aphotic zones. For example, DMSP demethylation pathway in the sulfur metabolism were more abundant in the exclusive aphotic than photic genomes (98 vs. 43 of 155 MAGs, respectively) (Fig. 5).

Sulfur cycling was a dominant process. Six twilight MAGs contained all markers genes for both DMSP demethylation and CO oxidation markers genes (Fig. 5): (i) two *Pseudohongiellaceae* family; (ii) *UBA9145 sp002712055* (250M_MAG_01), *UBA9145 sp002694855* (900M_MAG_11), *GCA-002712965 sp009391195* (250M_MAG_89); (iii) new *UBA9410* (250M_MAG_120) species. *SAR324* genomes represented by *Arctic96AD-7 sp002082305* (900M_MAG_17 and 2300M_MAG_22) encoded genes for partial DMSP demethylation (75%), and all genes for Calvin cycle (P*RK, prkB, rbcL* and *rbcS* genes) and for CO oxidation. Thiosulfate oxidation by SOX complex was another abundant pathway (Fig. S4 and Fig. 5) with complete or partial (60–80%) systems identified in 17 bacterial genomes and were more abundant in deep-sea MAGs (21 vs. 10 of 155, respectively) (Fig. S6B and Table S7). Partial Sox system (60–80%) and DMSP demethylation (<75%) were detected in *Sulfitobacter pontiacus* (250M_MAG_190 and 2300M_MAG_15) and possible new *Halioglobus_A* (250M_MAG_26) specie. The 2300M_MAG_15 genome also had all CO oxidation genes (Fig. 5).

Genes for dissimilatory sulfate reduction (*sat* and *aprAB*) that encode for enzymes for dissimilatory sulfate reduction (DRS) to sulfite were detected at teen MAGs (Fig. S6B and Table S7). However, a complete or near-complete (75%) pathway were limited to three *UBA868* family genomes: *REDSEA-S09-B13 sp002456995* (900M_ MAG_12), *UBA11791 sp002714255* (900 M _MAG_14 and 2300M_MAG_19) (Fig. 5). These genomes also have genes for partial (83%) thiosulfate oxidation and all genes for CO oxidation, indicating a mixotrophic lifestyle.

Specialized sulfur-based chemoautotrophy was identified at four exclusive deep-sea MAGs (Fig. 5): (i) the *Casp-alpha2 sp002938315* and *Casp-alpha2 sp002938075* (2300M_MAG_52 and 2300M_MAG_70, respectively), showed complete DMSP demethylation and all genes for CO oxidation. Also, the *Casp-alpha2 sp002938075* with complete Sox system, and the *Casp-alpha2 sp002938315* with partial Sox system (>60%); (ii) one possible new *Pseudohoeflea* (250M_MAG_117) species with the complete DMSP demethylation, partial Sox system (>60%) and all genes for CO oxidation; (iii) one *Pelagibacter* (2300M_MAG_37) with partial (>60%) DMSP demethylation, genes for CO oxidation, and urea metabolism (*UreB* and *UreC* genes).

Nitrogen cycling was primarily driven by ammonia-oxidizing archaea (AOA) of the class *Nitrososphaeria* (6 of 155 MAGs) including *Nitrosopelagicus brevis* (250M_MAG_84, 900M_MAG_29 and 2300M_MAG_25) *Nitrosopumilus sp002690535* (250M_MAG_61) and two possible novel species of *Nitrosopelagicus* (250M_ MAG_164) and *Nitrosopumilus* (250M_MAG_116) which possessed >60% of genes required for nitrification (*amoABC*) (Fig. S6B and Table S7). Most of these AOA MAGs also encoded key enzyme for the archaeal 3-hydroxypropionate–4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway (K14534: abfD; 4-hydroxybutyryl- CoA dehydratase/vinylacetyl-CoA-delta-isomerase (Table S7). The potential ammonia-oxidizing bacteria (**AOB)** *MarineAlpha9-Bin7 sp002937585* (2300M_MAG_54) have key gene for nitrification (*hao*) and all the genes required for CO oxidation.

Surprisingly, the *Novosphingobium indicum* (250M_MAG_22, 900M_MAG_06 and 2300M_MAG_05) genomes had all genes for nitrogen fixation (*nifH, nifD* and *nifK*) and CO oxidation, and partial (>70%) DMSP demethylation. These results suggest a potential non-cyanobacterial diazotrophs. *Gordonia sp002700145* MAGs (250M_MAG_32 and 900M_MAG_05) with almost complete genomes (>94% completeness) encoded genes for nitrate reduction *(NarG, narZ, nxrA, narH, narY* and *nxrB)*, CO oxidation and partial DMSP demethylation (>60%), indicating a reduced nitrogen-based chemoautotrophy (Fig. 5). Marker genes for denitrification (*nar, nir, nor* and *nos*) with partial (>60%) pathway were identified only in novel species of *Ketobacter* genus (250M_ MAG_05 and 2300M_MAG_02, 78%—ANI), *Halioglobus* genus (250M_MAG_26, 77%—ANI), *Oleispira* genus (2300M_MAG_50), and *Alteromonadaceae* (2300M_MAG_71) (Fig. 5).

## Summary: metabolic stratification and niche partitioning

The genome-resolved data reveals a clear stratification of microbial processes in the SB pelagic environment with distinct specializations and similarities (Fig. 6). The fundamental difference is the primary energy source, in the photic zone is dominated by phototrophy (oxygenic photosynthesis by Cyanobacteria) and photoheterotrophy (anoxygenic photosynthesis by *Rhodobacteraceae*). In contrast, the aphotic zone is driven by chemotrophy, fueled through the oxidation of inorganic compounds. The chemoautotrophic potentials exist in the photic zone (e.g., in *Gordonia*), however, the sulfur and nitrogen cycling processes is more abundant and diverse in the dark zone. This includes a higher abundance of SOX complex, DMSP demethylation, dissimilatory sulfate reduction and nitrogen transformers (AOA, AOB and denitrifiers) genes. The aphotic zone hosts unique obligate chemoautotrophs, such as the *Casp-alpha2* group, which used sulfur compounds and CO for energy and carbon fixation.

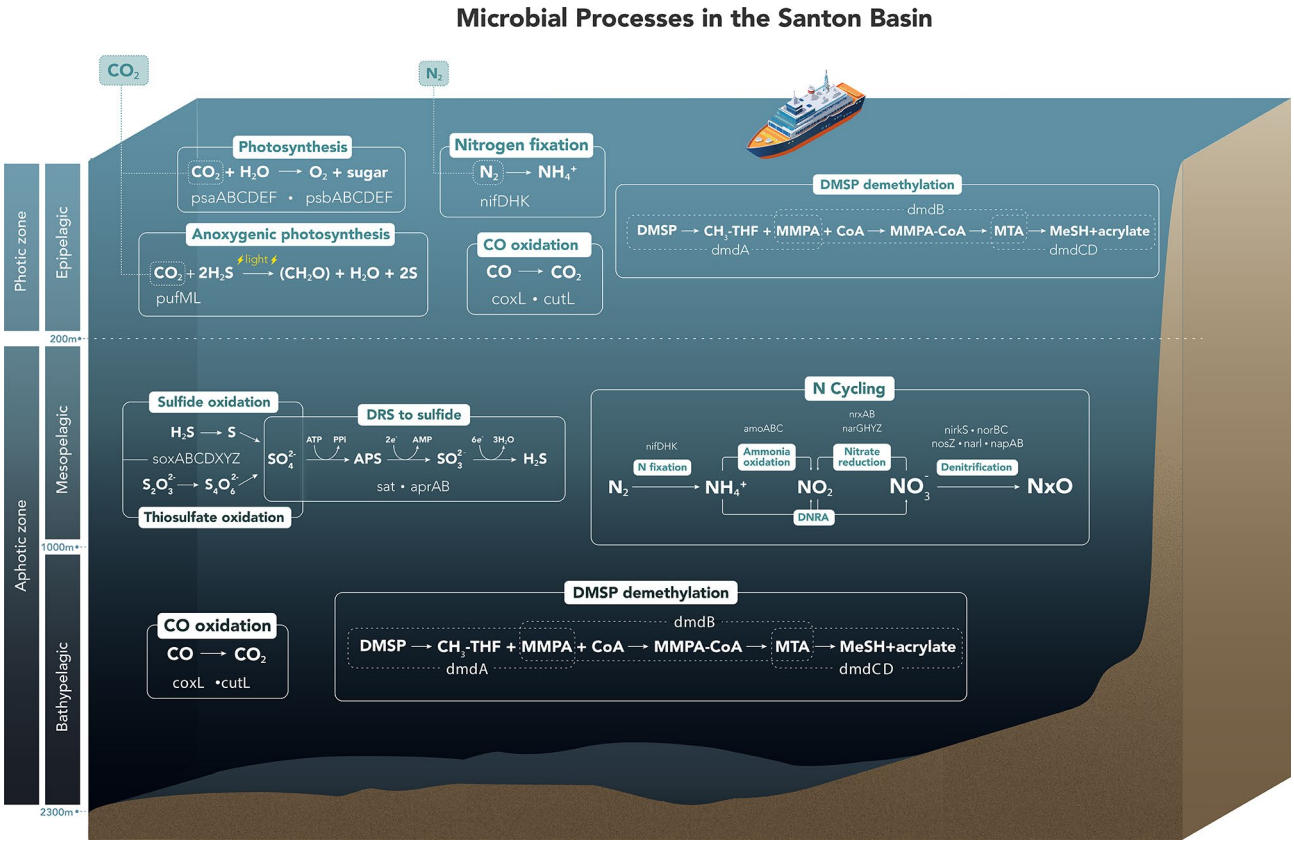


**Fig. 6**. Conceptual diagram summarizing the key microbial biogeochemical process identified in the Santos Basin,South Atlantic Ocean. Metabolic microbial pathways involved in the carbon, nitrogen, and sulfur cycles are represented. Metabolic pathways in the aphotic environment (< 200 m depth) were grouped due to the substantial taxonomic and functional redundancies observed in the SB-MAGs. Acronyms: $CO_2$ (carbon dioxide), $N_2$ (nitrogen gas), $NH_4^+$ (ammonium), $NO_2^-$ (nitrite), $NO_3^-$ (nitrate), N (nitrogen), $NH_3$ (ammonia), $H_2S$ (hydrogen sulfide), $SO_4^{2-}$ (sulfate), $S_2O_3^{2-}$ (thiosulfate), $S_4O_6^{2-}$ (tetrathionate), DMSP (dimethylsulfoniopropionate), $CH_3$-THF (methyltetrahydrofolate), MMPA (3-methiolpropionate), CoA (coenzyme A), MMPA-CoA (3-methiolpropionyl-CoA), MeSH or MTA (methanethiol), ATP (adenosine triphosphate), PPI (pyrophosphate), AMP (adenosine monophosphate), APS (adenosine 5'-phosphosulfate).

Some similarities and connecting processes are observed throw all water column (Fig. 6). For instance, the potential mixotrophy. In the photic zone, *Rhodobacteraceae* combine photoheterotrophy with chemotrophic potential (SOX and CO oxidation). In the dark zone, numerous taxa (e.g., *UBA868*, *SAR324* and *Casp-alpha2*) combine inorganic energy sources (sulfur compounds, CO and nitrogen compounds) with organic carbon metabolism. Furthermore, the vertical distribution of metabolically versatile taxa, such as *Gordonia* (nitrate reduction, CO oxidation) and *Novosphingobium indicum* (nitrogen fixation, CO oxidation, DMSP demethylation), supported metabolic connectivity between the photic and aphotic zones. The oxidation of carbon monoxide is an important process for energy generation in all pelagic SB environment.

## Discussion

This study revealed that the Santos Basin (SB) harbors a diverse and depth-stratified pelagic microbiome with approximately 45% of bacterial and 42% of archaeal MAGs classified as novel species, proportions higher than typically reported in global surveys such as Tara Oceans. This reveals the SB as a regional hotspot of microbial novelty within the South Atlantic. While large-scale expeditions such as Tara Oceans[9,34] and Malaspina[35,36] provided knowledge into the global ocean microbiome, their spatial resolution in the South Atlantic was minimal, particularly in the SB. Our results expand those global efforts by revealing that a subtropical basin harbors putatively novel lineages linked to key metabolic processes. These results align with the evidence that subtropical marine systems remain underexplored in global microbiome surveys[31] and highlight the SB as a functional and taxonomic hotspot for novel microbial lineages.

Santos Basin microbiome was powerfully predicted using a hybrid machine learning, achieving a classification accuracy of 90%. The results reflected depth-stratified ecological niches: photoautotrophic and photoheterotrophic lifestyles adapted to light and nutrient variations dominated photic waters, while heterotrophic and chemosynthetic prokaryotes prevailed in aphotic zone, consistent with adaptations to depth-related physicochemical conditions. Although vertical and spatial stratification patterns of marine microbiomes

are well-documented on global scale[9,34,35], the application of ML represents an innovative approach to microbiome monitoring in dynamic coastal ecosystems and data-scarce regions.

In contrast to global studies that identify temperature and salinity as the primary drivers of marine microbiome distribution[9,12], our model identified phosphate concentration as the primary predictor of microbial distribution in the SB, followed by salinity and nitrate. This result illustrates the influence of nutrient availability on microbial community structure in the region, suggesting that future changes in nutrient regimes could alter local ecosystem function[37]. The distribution of phosphate and nitrate in the SB is influenced by regional oceanographic processes, such as South Atlantic Central Water upwelling in the north and the La Plata River plume in the south[23,38]. These processes may enhance local biomass productivity[17,38], leading to an increase in prokaryotes and associated ecosystem dynamics.

Moreover, functional genes related to carbon, nitrogen, and sulfur cycling were broadly distributed across the SB and offer new insights into its pelagic ecosystem function. For example, the SB dataset reveals a significant potential for the degradation of dissolved organic sulfur compounds, with a notable focus on algal-derived dimethylsulfoniopropionate (DMSP). This compound plays a dual role in marine systems: serving as a critical sulfur and carbon source for microbial communities and acting as a defense mechanism for phytoplankton[38]. The *Sulfitobacter pontiacus* was identified as an ecologically relevant heterotrophic bacterium involved in the sulfur cycle in the deep SB. Genes for the DMSP demethylation pathway were found in these SB genomes, consistent with the role of *Sulfitobacter* in DMSP degradation in marine environments[40–42].

In addition, genes for carbon monoxide (CO) bio-oxidation were observed in all SB dataset. The potential CO bio-oxidation is increasingly recognized as a significant contributor to microbial energy acquisition[12,43,44], particularly in oligotrophic and deep-sea waters where alternative energy sources are limited. In the Santos Basin, its widespread detection across depths and suggests that CO oxidation supports metabolic flexibility, complementing phototrophy in surface waters and chemoautotrophy in the aphotic zone, thereby influencing carbon turnover in the region. In SB photic waters, we found evidence that potential aerobic anoxygenic phototroph (AAP) could use light and CO oxidation as complementary energy sources, as previously reported[45,46]. Also, these CO oxidation genes co-occurred with phototrophic aerobic anoxygenic photosynthesis and thiosulphate oxidation genes in Rhodobacteraceae SB-MAGs, indicating metabolic flexibility in the form of light- and CO-driven photoheterotroph. All AAP taxa recovered in the SB also carried genes for thiosulphate oxidation, providing evidence that these taxa use sulfur as an electron donor for photosynthesis, as previously observed[31,47], supporting energy-efficient carbon retention.

Similar to previous deep-sea surveys[48,49], the SB exhibits increased biodiversity in the mesopelagic layer. That could explain the higher recovery of MAGs at these depths, especially among mixotrophic and chemolithoautotrophic. Although the ecological implications of mixotrophy on marine ecosystem dynamics are unclear[50], it has been proposed to enhance dissolved inorganic carbon (DIC) fixation and increase the efficiency of the biological carbon pump[51,52]. In the SB dataset, *SAR*324 and *UBA868* emerged as potential mixotrophic groups, as previously described[53,54]. Notably, only SAR324 possesses genes for DIC fixation by RuBisCO genes, while both *SAR*324 and *UBA868* harbor genes for degrading dissolved organic sulfur compounds (DMSP), underscoring their roles in carbon and sulfur cycling, supporting dual roles in carbon and sulfur transformation.

The ecological importance of chemolithoautotrophs in the dark ocean is well established[55–57], although our knowledge of the key chemolithoautotrophy microorganisms in the deep waters remains unclear[12]. In the oxygenated, and nutrient-rich deep waters of the SB, evidence was found that microbes use DMSP, thiosulphate/sulfide and ammonia (reduced inorganic compounds), as energy sources for dark carbon fixation. Chemolithoautotrophs in the deep SB were proposed by the presence of ammonia-oxidizers *MarineAlpha9-Bin7 sp002937585, Nitrosopelagicus brevis* and *Nitrosopumilus sp002690535*. Another, potential SB chemolithotroph via CO and thiosulfate oxidation is *Casp-alpha2*, as previously observed in other regions[58]. SB's *Casp-alpha2* genomes also carried genes for degradation DMSP, pointing to a potential relation with the degradation of dissolved organic sulfur.

Surprisingly, potential non-cyanobacterial diazotrophs were observed in the deep SB, the *Novosphingobium indicum*. The ecological relevance of these deep sea diazotrophs is unclear as the availability of carbon substrates to support their energetic demand of $N_2$ fixation[12,59]. *Novosphingobium* members are metabolically versatile and usually associated with the biodegradation of aromatic compounds[60]. Interestingly, in addition to the *nif* genes, these *Novosphingobium* genomes have genes for DMSP degradation and CO oxidation, pointing to potential chemolithoautotrophic diazotroph. Their occurrence in oxygenated, nutrient-rich waters raises questions about energy strategies and carbon source availability in SB diazotrophic activity.

Traditionally, the ecosystem function of the pelagic microbiome has been associated with the autotroph–heterotroph dichotomy. Consequently, the organic matter sources supporting the deep ocean come from the photic layer or are generated by chemoautotrophic[7]. However, our results provide evidence that mixotrophic taxa may play a central role in organic sulfur transformations and nutrient cycling in the South Atlantic. The co-occurrence of pathways for CO oxidation, DMSP demethylation, nitrogen fixation, and sulfur respiration across MAGs reveals a metabolic plasticity, that is essential for resilience under environmental fluctuations.

Together, these findings provide a genomic baseline for future assessments of microbial ecosystem services in subtropical oceanic regions. The identification of depth-structured metabolic pathways and taxonomically novel lineages illustrates the potential of microbiome-based indicators to detect environmental changes. Given the increasing anthropogenic pressures in the SB (including offshore drilling, nutrient loading, and climate-driven changes in water mass dynamics), our integrative genome-resolved and machine learning approach offers a framework for microbial ecosystem monitoring. These results emphasize the need to incorporate microbial dimensions into ocean observation systems and environmental management strategies in the South Atlantic.

## Materials and methods

### Sampling strategy

A total of 67 water samples were collected during the SANAGU 2019 expedition, corresponding to 28 different oceanographic stations distributed across the SB (Fig. 1A). At each station, up to five discrete sampling depths (1 m, deep chlorophyll maximum layer [DCM], 250, 900 and 2300 m) were selected based on in situ profiles of temperature, salinity, and fluorescence, and corresponding to distinct water masses, including Tropical Water, Coastal Water, South Atlantic Central Water, Antarctic intermediate Water, and Upper Circumpolar Deep Water. These depths were defined using the temperature, salinity and fluorimeter profiles. All samples for biological and chemical analysis, in addition to physical data (temperature [°C], salinity [psu], dissolved oxygen [mg/L], colored dissolved organic matter [CDOM, $mg.L^{-1}$] and fluorescence [Relative Fluorescence Units, RFU]) were obtained using a combined Sea-Bird CTD/Carrousel 9 Plus system with Niskin bottles and equipped with calibrated sensors. For molecular biology, seawater (15L) was filtered through a peristaltic pump with 0.22 µm membrane pores (Sterivex, Millipore, MA). The filters were stored at − 80 °C for 30 days until DNA extraction. Samples for inorganic nutrients were also filtered through 0.22 µm filters onboard and frozen (at − 20 °C) until analysis (up to 30 days). Additional sampling details are provided in the SB scientific cruise report[23].

### Oceanographic parameters

Inorganic nutrients (nitrate, nitrite, phosphate, and silicate) concentrations were determined using a flow injection auto-analyzer (Auto-Analyzer 3, Seal Inc.), after filtration through 0.22 µm membrane filters. Ammonium concentrations were quantified using a Hitachi U1100 spectrophotometer. Dissolved organic carbon (DOC) concentrations were determined using an Elementar® Vario TOC CUBE analyzer (Elementar®). Further details related to environmental variables are provided in SB scientific cruise report[23].

### DNA extraction and sequencing

DNA was extracted using the DNeasy PowerWater kit (Qiagen, EUA) according to the manufacturer's instructions. Negative (no sample) extraction controls were used to ensure the extraction quality. DNA integrity was determined by electrophoresis on a 1% (v/v) agarose gel prepared with 1X TAE buffer (0.04 M Tris, 1 M glacial acetic acid, 50 mM EDTA, pH 8), and stained with SYBR Green (Thermo Fisher Scientific, São Paulo, Brazil). DNA concentration and integrity were determined using the Qubit dsDNA HS assay kit (Thermo Fisher Scientific, São Paulo, Brazil), according to the manufacturer's instructions.

DNA library preparation and sequencing were conducted at the National Laboratory of Bioinformatics (LNCC, Rio de Janeiro, Brazil, https://www.gov.br/lncc/pt-br, Rio de Janeiro, Brazil) using Illumina technology. Samples were arranged into sequencing plates based on their DNA concentrations to normalize PCR amplification cycles, according to the Illumina protocol. Libraries were constructed using the Nextera DNA Flex Library Preparation Kit (Illumina, USA), following the manufacturer's instructions. Sequencing was performed on a NextSeq 500 System (Illumina, USA) using the NextSeq 500/550 High Output Kit v2.5 (300 cycles), generating 2 × 150 bp paired-end reads.

### Taxonomic and functional annotation of metagenomic reads

Metagenomic reads were quality-filtered and trimmed using BBduk software (http://jgi.doe.gov/data-and-tools/bb-tools/). Illumina adapters, PhiX sequences and low-quality reads (Phred score < 20) were removed using the following parameters: minlength = 50, mink = 8, qout = auto, hdist = 1, k = 31, trimq = 10, qtrim = rl, ktrim = l, minavgquality = 20 and statscolumns = 5. Prokaryotic taxonomic annotation of short reads was assigned through best-hit classification using Kraken 2 v2.1.2[61,62]. with parameters default and the k2_pluspfp database (https://benlangmead.github.io/aws-indexes/k2). This expanded database includes archaeal and bacterial genomes from RefSeq. Functional profiling was conducted using the SEED subsystems and KEGG databases via MG-RAST platform[64,65], with the following parameters: e-value 0.00001, minimum 50 bp alignment and > 60% sequence identity. The top hits were analyzed to identify specific genes and pathways based on the SEED database. Hits counts were normalized to the total number of hits per counts for each sample and a row Z-score was computed to assess statistical differences among samples for each metabolic pathway. All functional analyses were performed in R software (R Core Team, 2021) with the "PhyloSeq"[66], "ggplot2"[67], and "vegan"[68] packages.

### Machine learning

Machine learning analyses were conducted using taxonomic annotation of metagenomics reads and associated environmental data (Table S1). A hybrid approach combining unsupervised and supervised learning methods was implemented through the iMESC platform—An Interactive Machine Learning App for Environmental Science[69,70]. For unsupervised learning, we applied a self-organizing map (SOM) algorithm, using a neural network method[71] to group similar samples into neurons -Best-Matching Units (BMUs). Due to the high diversity of genera (2,265 genera) identified by the read's annotation (Table S2), we transformed the genus-level data using the Hellinger method to explore the microbiome distribution in the SB (to account for compositional constraints). In this study, we applied a multi-layered SOM, in which genus-level taxonomic profiles were trained together with the geographic coordinates of the sampling stations. One layer represented microbial taxonomic composition and another represented spatial information, allowing the SOM to preserve both taxonomic similarity and spatial structure in the clustering process. The first layer incorporated microbial genus-level profiles (weight = 0.95) based on the Euclidean similarity index; the second layer integrated sample coordinates and depth (scaled and centered; weight = 0.05), also using Euclidean distance. The SOM grid dimensions were determined following Vesanto & Alhoniem (2000)[72]: The initial number of map nodes was determined using the heuristic recommendation $5^{*}\sqrt{x}$, where x is the number of samples. Next, we determined the eigenvectors and

eigenvalues from the autocorrelation matrix. We then set the ratio between the two sides of the grid to match the ratio between the two largest eigenvalues. Finally, we scaled the side lengths so that their product (xdim * ydim) was as close as possible to the number of map units determined above. This process resulted in the use of 42 neurons (map units).

Each neuron in the final SOM map contains a weighted list of microbial genus-level, termed a codebook. The codebook generated by the SOM analysis was then subjected to hierarchical clustering (using the Ward method with squared differences) to group similar BMUs and their corresponding samples. The number of clusters formed by the hierarchical clustering (HC) was determined using a split moving window analysis, which identifies discontinuities in the relationship between cluster number and within-cluster sum of squares. These neurons serve as descriptors of the microbial communities, and they will be referred to as taxonomic associations (Assoc). To characterize community structure within each Assoc, we calculated Shannon and Chao1 indices and compared them using Kruskal–Wallis ANOVA followed by Dunn's post hoc test. We further applied IndVal (Indicator Species Analysis) to determine genus-level indicators specific to each Assoc, assigning features exclusively to the cluster where their association was strongest. Statistical significance was evaluated with 999 permutations.

In the supervised phase, each Assoc was used as a response variable in a supervised training phase. This phase aimed to identify the optimal oceanographic variables (Table S1) for modeling and predicting microbial community Assoc using machine-learning classification algorithms. Initially, samples were split for validation, ensuring a balanced partition across Assoc. The training dataset (80%) was used for model fitting, while the test dataset (20%) served for subsequent evaluation. Random Forest (RF) classification algorithms were employed with 500 trees, a cross-validation method using 5 folds and 10 repetitions, evaluating up to 10 variables randomly sampled ('mtry') at each split. The performance of each RF model was assessed using standard metrics. This included evaluating accuracy, analyzing confusion matrices to understand predictive performance across different associations, and determining feature importance to identify which oceanographic variables had the greatest influence on SB microbiome associations.

## Metagenome assembly, binning, and genome quality control and distribution across Santos Basin

We choose a co-assembly strategy to increase the throughput and maximize the number of recovered metagenome-assembled genomes (MAGs) per depth (1 m, DCM, 250 m, 900 m and 2300 m; Table S1 to check dataset details). This approach has been successfully used in previous water column genomes reconstruction studies[73,74]. The 64 metagenomic reads were co-assembled using MEGAHIT v1.0.6[75], and contigs smaller than 1000 bp were removed. Individual metagenomic reads were mapped back to their respective depth-specific co-assembly using Bowtie2 v2.3.4.3 with default parameters[76]. To process FASTA files, each of the 5 resulting assemblies was processed using the anvi'o v8 workflow[77], which generated a contigs database, identified open reading frames (ORFs) using Prodigal v2.60[78], identified single-copy Bacteria and Archaea genes using HMMER v3.2.1[79], and annotated genes with functions using the NCBI's Clusters of Orthologous Groups (COG)[80].

Individual BAM files (read-mapping outputs) were profiled with a minimum contig length of 2500 bp. Genome binning was performed using the CONCOCT[81] program with default parameters. The resulting bins were manually refined using the anvi'o interactive interface[82]. Bin quality was assessed with CheckM v.1.0.7[83] and metagenome-assembled genomes (MAG) were categorized according to the standards proposed by Bowers et al,[33]: high-quality MAGs (> 90% completeness, < 5% contamination), medium-quality MAGs (> 50% completeness, < 10% contamination), or low-quality MAGs (< 50% completeness, < 10% contamination).

MAGs were taxonomically classified based on genome phylogeny using the Genome Taxonomy Database Toolkit (GTDB-Tk v2.4.0)[84] and functionally annotated using the Distilled and Refined Annotation of Metabolism (DRAM v1.0)[85]. In addition to the DRAM annotations, we used the Sulfur Metabolism Gene Integrative Database[86] to identify genes for inorganic and organic sulfur cycling. A curated list of 109 genes based on Kyoto Encyclopedia of Genes and Genomes (KEGG) representative of main biogeochemical cycling from the pelagic realm that were searched at the SB reconstructed genomes, is provided in Table S8. All genomic analyses were performed in R software (R Core Team, 2021) with the "ggplot2"[66] and "vegan"[68] packages.

To investigate the distribution of MAGs across depths in SB, relative abundance was calculated as the number of reads recruited to each contig divided by the total reads recruited across all samples, using the anvi'o v.8[77]. First, MAGs were profiled using the "*anvi-summarize*" function, which provided detailed information about each MAG in the bin_by_bin and bins_across_samples subdirectories. The bin_by_bin subdirectory contained per-bin statistics, including mean coverage and relative abundance across samples, while the bins_across_samples subdirectory aggregated these data for all MAGs in tab-delimited matrices. The relative abundance of each MAG in different samples was retrieved from the mean_coverage.txt and rel_abundance.txt files in bins_across_samples, ensuring consistency and accuracy in quantifying MAG distributions. For visualization, centered log ratio (CLR) transformation was performed by dividing the relative abundance of each MAG by the geometric mean of all MAG in a sample, followed by the application of the natural logarithm. To ensure comparability, all zero values were replaced with a pseudocount ($\varepsilon = 1e\text{-}^{6}$) prior to transformation. CLR-transformed values were normalized using Min–Max scaling to a of 0—1 range for visualization. Analyses were conducted independently for each depth layer to emphasize depth-specific MAG patterns. Final distributions were visualized using heatmaps.

## Data availability

The sequencing reads dataset generated and analyzed during the current study are available in the National Centre for Biotechnology Information (NCBI) repositor under the BioProject PRJNA1191028. The genomic dataset

generated during and analyzed during the current study are available in the Figshare repositor under https://doi.org/10.6084/m9.figshare.30043690. The iMESc and iMESc Savepoints used for the machine learning analyses are available at Github via https://github.com/DaniloCVieira/imesc_savepoints/tree/2c5a4441ad447c71cc8f5d3e11811fb5a76a3427/Bergo_2025/Microbial-signatures-define-the-ecosystem-functions-of-the-pelagic-microbiome-in-a-basin-scale-Southwest-Atlantic-Ocean

Received: 4 June 2025; Accepted: 21 January 2026
Published online: 22 April 2026

## References

1. Stal, L. J. A Sea of Microbes: What's So Special about Marine Microbiology. In *The Marine Microbiome* (eds Stal, L. J. & Cretoiu, M. S.) 1–44 (Springer, 2022).
2. Jiao, N. et al. The microbial carbon pump and climate change. *Nat. Rev. Microbiol.* **22**, 408–419 (2024).
3. Li, W. K. W. Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: measurements from flow cytometric sorting. *Limnol. Oceanogr.* **39**, 169–175 (1994).
4. Jardillier, L. et al. Significant $CO_2$ fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME J.* **4**, 1180–1192 (2010).
5. Worden, A. Z. et al. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594 (2015).
6. Buchan, A., LeCleir, G. R., Gulvik, C. A. & González, J. M. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat. Rev. Microbiol.* **12**, 686–698 (2014).
7. Baltar, F. & Gerhard, J. H. Ideas and perspectives: is dark carbon fixation relevant for oceanic primary production estimates?. *Biogeosciences* **16**, 3793–3799 (2019).
8. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
9. Sunagawa, S. et al. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
10. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5**, 170203 (2018).
11. Yoshitake, K. et al. Development of a time-series shotgun metagenomics database for monitoring microbial communities at the Pacific coast of Japan. *Sci. Rep.* **11**, 12222 (2021).
12. Acinas, S. G. et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Nat. Commun.* **12**, 604 (2021).
13. Junger, P. C. et al. Global biogeography of the smallest plankton across ocean depths. *Sci. Adv.* **9**, eadg9763 (2023).
14. Sánchez, P. et al. Marine picoplankton metagenomes and MAGs from eleven vertical profiles obtained by the Malaspina Expedition. *Sci. Data* **11**, 154 (2024).
15. Alves, J. N. et al. Microbial community diversity and physical–chemical features of the Southwestern Atlantic Ocean. *Arch. Microbiol.* **197**, 165–179 (2015).
16. Coutinho, F. H. et al. Ecogenomics and metabolic potential of the South Atlantic Ocean microbiome. *Sci. Total Environ.* **765**, 142758 (2021).
17. Bergo, N. M., Signori, C. N., Amado, A. M., Brandini, F. P. & Pellizari, V. H. The partitioning of carbon biomass among the pico- and nano-plankton community in the South Brazilian Bight during a strong summer intrusion of South Atlantic central water. *Front. Mar. Sci.* **4**, 238 (2017).
18. Ribeiro, C. G. et al. Pico and nanoplankton abundance and carbon stocks along the Brazilian Bight. *PeerJ* **4**, e2587 (2016).
19. Ribeiro, C. G. et al. Estimating microbial populations by flow cytometry: Comparison between instruments. *Limnol. Oceanogr. Methods* **14**, 750–758 (2016).
20. Agostini, L. et al. Deep-sea plastisphere: long-term colonization by plastic-associated bacterial and archaeal communities in the Southwest Atlantic Ocean. *Sci. Total Environ.* **793**, 148335 (2021).
21. Ferreira, J. C. N. et al. Abundance and microbial diversity from surface to deep water layers over the Rio Grande Rise. *South Atlantic. Prog. Oceanogr.* **201**, 102736 (2022).
22. Moreira, J. C. F. et al. A Machine learning approach elucidates spatial patterns of environmental properties driving microbial composition over Santos Basin. *South Atlantic bioRxiv* https://doi.org/10.1101/2025.02.26.640418 (2025).
23. Moreira, D. L. et al. Multidisciplinary scientific cruises for environmental characterization in the Santos Basin – methods and sampling design. *Ocean Coast. Res.* **71**, e23022 (2023).
24. Corrêa, A. M. S. et al. Hydrocarbon composition of oils spilt on the Brazilian coast in 2019 and 2022. *Mar. Pollut. Bull.* **207**, 116821 (2024).
25. Sartori, É. et al. Trace metal concentration along the Brazilian coast: An assessment of the influence of the Doce River plume. *Mar. Pollut. Bull.* **188**, 114640 (2023).
26. Senez-Mello, T. M. et al. Assessment of anthropogenic pollution in Guanabara Bay (SE Brazil) through biogeochemical data and stable isotope mixing models. *Environ. Sci. Pollut. Res.* **31**, 32972–32997 (2024).
27. Ramírez, F. et al. Overfishing species on the move may burden seafood provision in the low-latitude Atlantic Ocean. *Sci. Total Environ.* **836**, 155480 (2022).
28. Franco, B. C. et al. Climate change impacts on the atmospheric circulation, ocean, and fisheries in the southwest South Atlantic Ocean: a review. *Clim. Change* **162**, 2359–2377 (2020).
29. Tara Ocean Foundation et al. Priorities for ocean microbiome research. *Nat. Microbiol.* **7**, 937–947 (2022).
30. Asnicar, F., Thomas, A. M., Passerini, A., Waldron, L. & Segata, N. Machine learning for microbiologists. *Nat. Rev. Microbiol.* **22**, 191–205 (2024).
31. Prabhu, A. et al. Machine learning and metagenomics identifies uncharacterized taxa inferred to drive biogeochemical cycles in a subtropical hypereutrophic estuary. *ISME Commun.* **4**, 067 (2024).
32. Pinnaka, A. K. & Tanuku, N. R. S. Marine microbial diversity for sustainable development. In *Microbial Diversity in Ecosystem Sustainability and Biotechnological Applications* (eds Gupta, J. et al.) 117–158
33. Bowers, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol.* **35**, 725–731 (2017).
34. Sunagawa, S. et al. Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
35. Arístegui, J., Gasol, J. M., Duarte, C. M. & Herndl, G. J. Microbial oceanography of the dark ocean's pelagic realm. *Limnol. Oceanogr.* **54**, 1501–1529 (2009).
36. Duarte, 2015 – Malaspina
37. Zhang, Z. et al. Global biogeography of microbes driving ocean ecological status under climate change. *Nat. Commun.* **15**, 4657 (2024).
38. Brandini, F. P., Turra, P. M. & Santos, P. P. G. M. Ecosystem responses to biogeochemical fronts in the South Brazil Bight. *Prog. Oceanogr.* **164**, 52–62 (2018).

39. Xu, X. et al. (2024) Genome-based taxonomic classification of the genus *Sulfitobacter* along with the proposal of a new genus *Parasulfitobacter* gen. nov. and exploring the gene clusters associated with sulfur oxidation. *BMC Genomics* **25**: 389.
40. Moran, M. A., González, J. M. & Kiene, R. P. Linking a bacterial taxon to sulfur cycling in the sea: studies of the marine *Roseobacter* group. *Geomicrobiol. J.* **20**, 375–388 (2003).
41. Zeng, Y.-X. et al. Diversity of bacterial dimethylsulfoniopropionate degradation genes in surface seawater of Arctic Kongsfjorden. *Sci. Rep.* **6**, 33031 (2016).
42. Sanz-Sáez, I. et al. Top abundant deep ocean heterotrophic bacteria can be retrieved by cultivation. *ISME Commun.* **3**, 92 (2023).
43. Lappan, R. et al. Molecular hydrogen in seawater supports growth of diverse marine bacteria. *Nat. Microbiol.* **8**, 581–595 (2023).
44. Dede, B. et al. Bacterial chemolithoautotrophy in ultramafic plumes along the Mid-Atlantic Ridge. *ISME J.* **18**, wrae165 (2024).
45. Giebel, H. A. et al. Complementary energy acquisition via aerobic anoxygenic photosynthesis and carbon monoxide oxidation by *Planktomarina temperata* of the Roseobacter group. *FEMS Microbiol. Ecol.* **95**. fiz050. (2019)
46. Cai, H. et al. Aerobic anoxygenic phototrophs play important roles in nutrient cycling within cyanobacterial *Microcystis*bloom microbiomes. *Microbiome* **12**, 88 (2024).
47. Luo, H. & Moran, M. A. Evolutionary ecology of the marine *Roseobacter* clade. *Microbiol. Mol. Biol. Rev.* **78**, 573–587 (2014).
48. Danovaro, R., Snelgrove, P. V. R. & Tyler, P. Challenging the paradigms of deep-sea ecology. *Trends Ecol. Evol.* **29**, 465–475 (2014).
49. Xue, C.-X. et al. Insights into the vertical stratification of microbial ecological roles across the deepest seawater column on Earth. *Microorganisms* **8**, 1309 (2020).
50. Li, J. et al. (2024) *Arcobacteraceae* are ubiquitous mixotrophic bacteria playing important roles in carbon, nitrogen, and sulfur cycling in global oceans. *mSystems* **9**: 00513–24.
51. Stoecker, D. K. et al. Mixotrophy in the marine plankton. *Annu. Rev. Mar. Sci.* **9**, 311–335 (2017).
52. Ward, B. A. & Follows, M. J. Marine mixotrophy increases trophic transfer efficiency, mean organism size, and vertical carbon flux. *Proc. Natl Acad. Sci. USA* **113**, 2958–2963 (2016).
53. Swan, B. K. et al. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**, 1296–1300 (2011).
54. Baltar, F. et al. A ubiquitous gammaproteobacterial clade dominates expression of sulfur oxidation genes across the mesopelagic ocean. *Nat. Microbiol.* **8**, 1137–1148 (2023).
55. Teira, E. et al. Distribution and activity of Bacteria and Archaea in the deep water masses of the North Atlantic. *Limnol. Oceanogr.* **51**, 2131–2144 (2006).
56. Alonso-Sáez, L. et al. High bicarbonate assimilation in the dark by Arctic bacteria. *ISME J.* **4**, 1581–1590 (2010).
57. La Cono, V. et al. Contribution of bicarbonate assimilation to carbon pool dynamics in the deep mediterranean sea and cultivation of actively nitrifying and $CO_2$-fixing bathypelagic prokaryotic consortia. *Front. Microbiol.* **9**, 3 (2018).
58. Zhou, Z., Tran, P. Q., Kieft, K. & Anantharaman, K. Genome diversification in globally distributed novel marine *Proteobacteria* is linked to environmental adaptation. *ISME J.* **14**, 2060–2077 (2020).
59. Chakraborty, S., Andersen, K. H., Merico, A. & Riemann, L. Particle-associated N2 fixation by heterotrophic bacteria in the global ocean. *Sci. Adv.* **11**(8), 4693 (2025).
60. Yuan, J., Lai, Q., Zheng, T. & Shao, Z. *Novosphingobium indicum* sp. nov., a polycyclic aromatic hydrocarbon-degrading bacterium isolated from a deep-sea environment. *Int. J. Syst. Evol. Microbiol.* **59**, 2084–2088 (2009).
61. *Acids Res.* **44**, D590-D594 (2016).
62. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
63. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
64. Wilke, A. et al. The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res.* **44**, D590–D594 (2016).
65. Keegan, K. P., Glass, E. M. & Meyer, F. MG-RAST, a metagenomics service for analysis of microbial community structure and function. In *Microbial Environmental Genomics (MEG)* 207–233.
66. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
67. Wickham, H. ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).
68. Oksanen, J. et al. vegan: Community Ecology Package. R package version 2.9 (2013).
69. Fonseca, G. & Vieira, D. C. Overcoming the challenges of data integration in ecosystem studies with machine learning workflows: an example from the Santos project. *Ocean Coast. Res.* **71**, e23021 (2023).
70. Vieira, D. C. et al. iMESc - an interactive machine learning app for environmental sciences. *Front. Environ. Sci.* **13**, 1533292 (2025).
71. Kohonen, T. The self-organizing map. *Proc. IEEE* **78**, 1464–1480 (1990).
72. Vesanto, J. & Alhoniemi, E. Clustering of the self-organizing map. *IEEE Trans. Neural Netw.* **11**, 586–600 (2000).
73. Debeljak, P. et al. Seasonal patterns in microbial carbon and iron transporter expression in the Southern Ocean. *Microbiome* **11**, 187 (2023).
74. Wu, Z. et al. Metagenomic insights into microbial adaptation to the salinity gradient of a typical short residence-time estuary. *Microbiome* **12**, 115 (2024).
75. Li, D. et al. MEGAHIT v1. 0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
76. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
77. Eren, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
78. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
79. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
80. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–D269 (2015).
81. Alneberg, J. et al. *CONCOCT: clustering contigs on coverage and composition*. *arXiv preprint* https://abs/arXiv.org/1312.4038 (2013).
82. Eren, A. M. et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2021).
83. Parks, D. H. et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
84. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
85. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
86. Yu, X. et al. SCycDB: a curated functional gene database for metagenomic profiling of sulphur cycling pathways. *Mol. Ecol. Resour.* **21**, 924–940 (2021).


## Acknowledgements

We are grateful to PETROBRAS for the PCR-BS project planning and management. We would like to thank Daniel Moreira for coordinating the PCR-BS project and Dr. Frederico Pereira Brandini for coordinating the

"Pelagic realm project". We would like to thank members of the LECOM team for their scientific support on board. We are deeply grateful for the invaluable guidance and support provided by Rosa Carvalho Gamba, your legacy wisdom continues to inspire us even in her absence. A special thanks to the following PCR-BS project's researchers and their institutions: UFRJ, UFF, PUC-Rio, UERJ, FIRJAN/SENAI, SALT, INPE, USP, UNIFESP, UFPR, UNESP, IP-SP, FURG and Socioambiental. To the University of São Paulo Foundation (FUSP) for financial administrative management. We also thank the RV Ocean Stalwart and RV Seward Johnson crew, and OceanPact for the scientific expeditions. We thank Barbara Resende Silva for providing valuable support in designing the conceptual diagram. We thank Alexandra Lehmkuhl Gerber and Ana Paula C. Guimarães from UGCDFA/LNCC for sequencing the samples. We acknowledge the anonymous reviewers' valuable comments that helped to improve this manuscript.

## Author contributions

N.M.B., V.H.P., D.L.M. and C.R.J. conceived and designed the study. J.C.F.M., A.M.A., A.M.E., R.G.R., D.C.D.C., F.S.P. and W.S.G.B. was responsible for onboard sampling and extracting DNA from water samples. M.G.C and F.P.B. was responsible for onboard sampling and generating environmental dataset. A.T.R.V. was responsible for generating the sequencing dataset. N.M.B., L.N.L., F.V.P., R.G.M.L., D.C.V. and F.M. carried out the bioinformatics and statistical analysis. N.M.B., R.G.R., and F.V.P., performed analysis of genomic data. D.C.V. and N.M.B. performed oceanographic analyses; N.M.B., F.V.P., D.C.V. and F.M. prepared the figures and tables and wrote the first draft of the manuscript. A.G.B. and G.F. provided helpful discussions and technical help. All authors discussed the results, commented, and reviewed the manuscript.

## Funding

This study was funded by Petróleo Brasileiro S.A. (PETROBRAS), through the RD&I investments clauses of Brazilian National Agency of Petroleum, Natural Gas, and Biofuels (ANP), under grant numbers 5850.0109317.18.9 and 21167–2, respectively. The project "Caracterização química e biológica do sistema pelágico da Bacia de Santos/PCR-BS" is a national collaboration between PETROBRAS and Universidade de São Paulo. NMB and LNL was financed by a Pos Postdoctoral 's fellowship from the PETROBRAS. ATRV is supported by FAPERJ (E-26/201.046/2022) and CNPq (307145/2021–2).

## Declarations

## Competing interests

The authors declare no competing interests.

## Consent for publication

All authors have read and approved the paper for submission.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-37419-9.

**Correspondence** and requests for materials should be addressed to N.M.B.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.