Nitrogen fixation rates increase with diazotroph richness in the global ocean

Dominic Eriksson[1,2]*, Nicolas Gruber[1], Fabio Benedetti[1], Damiano Righetti[1,3], Lucas Paoli[2], Guillem Salazar[2], Shinichi Sunagawa[2]*, Meike Vogt[1]*

[1]Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, 8092 Zürich, Switzerland.

[2]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, 8092 Zürich, Switzerland

[3]Centre for Ocean Life, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

* Corresponding author: Dominic Eriksson, Shinichi Sunagawa, Meike Vogt

Email: deriksson@ethz.ch, nicolas.gruber@env.ethz.ch, fabio.benedetti@usys.ethz.ch, drig@aqua.dtu.dk, lucas.paoli@biol.ethz.ch, guillem.salazarguiral@biol.ethz.ch, ssunagawa@ethz.ch, meike.vogt@env.ethz.ch

Author Contribution: Dominic Eriksson: Conceptualization, Data Curation, Methodology, Software, Formal analysis, Writing – Original Draft, Writing – Review & Editing, Visualization **Nicolas Gruber:** Conceptualization, Writing – Review & Editing, Supervision, Funding acquisition **Fabio Benedetti:** Conceptualization, Methodology, Formal analysis, Writing – Review & Editing **Damiano Righetti:** Conceptualization, Data Curation, Methodology **Lucas Paoli:** Data Curation, Writing – Review & Editing **Guillem Salazar:** Conceptualization **Shinichi Sunagawa:** Conceptualization, Writing – Review & Editing, Supervision, Funding acquisition **Meike Vogt:** Conceptualization, Writing – Review & Editing, Supervision, Funding acquisition

Keywords: Nitrogen fixation, Diazotrophs, Biodiversity, Species Distribution Models, Ecosystem function

Abstract

Marine diazotrophs, a highly specialized group of marine prokaryotes, convert atmospheric nitrogen gas into bioavailable forms of nitrogen and are thus critical to maintain the fertility of the ocean. However, little is known about the link between global-scale diazotroph diversity and marine N2 fixation rates. Here, we address this question by integrating more than 22'000 DNA sequencing and microscopy-based observations for 14 diazotroph species into species distribution models. We identify distinct biogeographic patterns for the major known taxa of diazotrophs, including colony-forming, unicellular, symbiotic, and non-cyanobacterial diazotrophs. Non-cyanobacterial diazotrophs show a higher annual mean number of presences in upwelling regions compared to their cyanobacterial counterparts. In addition, the identified biogeographic patterns reveal a strong latitudinal gradient in diazotroph species richness, which is highest in tropical and subtropical regions and declines towards the poles. Temperature and nutrient-related parameters rank as the most important predictors of the biogeography explaining up to 36% of the variance in the data for specific taxa. We find diazotroph richness to be positively correlated with independently estimated nitrogen fixation rates, suggesting efficient resource partitioning rather than competitive exclusion as the dominant driver of the observed biodiversity patterns. Our work reveals that important biodiversity-ecosystem functioning relationships associated with global biogeochemical cycling exist in the marine plankton, and suggests that global nitrogen fixation rates and diazotroph diversity are likely to increase in a warming ocean.

Significance Statement

The diversity patterns of marine nitrogen fixers (diazotrophs) are poorly known, despite the central role that these organisms play in providing bioavailable nitrogen to fuel the ocean's productivity.

We enhance our current knowledge by combining observations with species distribution models for an unprecedented number of marine diazotrophic species, permitting us to derive the first observation-based map of diazotroph species richness. By showing that an increase in richness of diazotrophic species is positively correlated with biological nitrogen fixation, we imply the importance of biodiversity for ecosystem function and the existence of biodiversity-ecosystem function relationships within the marine realm. We also demonstrate the importance of temperature for structuring diazotrophic richness, implying that global ocean warming might positively impact marine pelagic nitrogen fixation.

Introduction

Nitrogen-fixing microorganisms, collectively termed diazotrophs, convert atmospheric nitrogen gas into fixed forms of bioavailable nitrogen. This way, they supply a very substantial fraction of the nitrogen needed to support primary production and export in many oligotrophic regions of the tropics and subtropics (1, 2). Biological nitrogen fixation (BNF) is also key for maintaining the fertility of the ocean, as it resupplies most of the fixed nitrogen that is lost from the ocean as a consequence of denitrification processes (3). Biological nitrogen fixation is a highly specialized process that only a handful of organismal groups can perform (4). The nifH gene encoding the central enzyme that enables the splitting of the N_2 molecule, i.e., the nitrogenase, is highly conserved and found exclusively within the domain of Bacteria and Archaea (4). Thus, the overall richness of diazotroph species is relatively low, especially when compared to the richness of species supplying more widespread ecosystem functions such as photosynthesis. Still, research in the last few years has uncovered several new groups of marine species capable of BNF, substantially expanding the diversity of known diazotrophs (5, 6). This highlights how little is known about the key marine diazotrophs and their biogeography, and consequently, how their diversity might be related to the magnitude of BNF. This is a major shortcoming, hampering not only our ability to model the global

distribution of this important group of species in space and time but also limiting our ability to assess how this group responds to future climate change and other changes in oceanic stressors (7).

Historically, it was believed that BNF in the global oceans was primarily driven by one group of cyanobacteria, i.e., the colony-forming species in the genus *Trichodesmium* (8, 9). *Trichodesmium* spp. can easily be recognized using microscopy-based sampling strategies and has been studied for decades. The second studied group refers to diazotrophs of the genera *Richelia* and *Calothrix* (10), who live in symbiosis with diatoms of the genera *Chaetoceros*, *Hemiaulus* and *Rhizosolenia* (10). Their contribution to global BNF was believed to be much smaller than that of *Trichodesmium*, mainly owing to the strong silicic acid limitation of the host diatoms that prevent them from growing in many low-nutrient regions (11).

The advent of culture-independent methodologies, such as amplification of the nifH gene by PCR or shotgun metagenomics, revolutionized traditional methods of identifying diazotrophs and led in the past two decades to the discovery of several additional groups of diazotrophs (6, 12). The first newly discovered group of marine diazotrophs are unicellular cyanobacteria collectively referred to as UCYN, with three known subgroups *UCYN-A*, *UCYN-B*, and *UCYN-C*. Besides free-living cells, this group also contains species that are found in symbiosis with photosynthetic eukaryotes or as aggregates (13, 14). Within the Candidatus species *Atelocyanobacterium thalassa* (*UCYN-A*), several strains have been recognized (A1 to A6) (15). The strain *UCYN-A*2 is known as a symbiont of the prymnesiophyte algae *Braarudosphaera bigelowii*, and the smaller-sized *UCYN-A1* is associated with a yet unidentified relative of *B. bigelowii* (13). Representatives from the genus *Crocosphaera* (*UCYN-B*) are either free-living or aggregate-forming depending on the specific strain (14). Cyanothece-like *UCYN-C* diazotrophs are presumably free-living small diazotrophs that can form aggregates of up to 500 μm that contribute to the rapid sinking of particulate organic carbon (16). The second group of newly recognized diazotrophs are potentially heterotrophic

bacteria and archaea (often referred to as non-cyanobacterial diazotrophs), which might be capable of BNF based on the presence of the nifHDK operon in their genomes (12). The most widely distributed non-cyanobacterial diazotroph is the phylotype *Gamma-A*, a gamma-proteobacteria that has been found in oligotrophic oxygenated waters of subtropical and tropical latitudes (17). Thus, these molecular methods led to the recognition that diazotrophic niches are much wider than previously recognized.

As to their niche characteristics and drivers, at first, diazotrophs were thought to thrive only in the oligotrophic regions of the ocean, where they benefit from the low concentration of nitrate, either because these low concentrations hamper the growth of their competitors or because the nitrogen-fixing capacity allows them to inhabit a nitrogen-poor milieu (18). Diazotrophs were also believed to be limited to high temperatures since elevated temperatures provide the higher energies needed for the diazotrophs to better cope with the energetically very expensive process of splitting the N_2 molecule. However, more recent work indicates that diazotrophs can be found in nearly all environments, ranging from euphotic oligotrophic waters (6) to cold polar nutrient-rich waters (19) down to aphotic environments (20) and oxygen minimum zones (21). Whereas cyanobacterial diazotrophs are observed predominantly in euphotic surface waters (22), active BNF in aphotic waters has been attributed to non-cyanobacterial diazotrophs which are believed to be ubiquitous in marine waters and can reach higher relative abundances than their cyanobacterial counterparts (6, 12).

As to their impact on global nitrogen cycling, our knowledge about the ecosystem function performed by diazotrophs in terms of BNF advanced significantly in recent decades (4, 23–25). While initial estimates were based nearly exclusively on the incubation of Trichodesmium (9), whole community assays and geochemical approaches (26–29) are now paying better attention to the potential diversity of diazotrophs and their contribution to BNF, both in modeling and observational studies (24, 30). Yet, in situ BNF measurements remain sparse and extrapolations tentative, with

current best estimates of global BNF hovering around 150 Tg N yr^{-1} (23, 24). While work has been conducted to quantify overall diversity and gross BNF rates, what has not been assessed so far is the relationship between global diazotroph species diversity and BNF. The identification of such a relationship would provide evidence for an important biodiversity-ecosystem functioning relationship (31–33) in marine plankton ecosystems associated with global biogeochemical cycling. Together with the identification of its underlying drivers, the existence of such a biodiversity-ecosystem functioning relationship may have far-reaching implications for the response of marine plankton and global biogeochemical cycles to climate change.

As BNF is exclusively driven by diazotrophs, their community and diversity structure could directly control global BNF rates in alternative ways. On one hand, high degrees of niche similarity and overlap could lead to competitive exclusion for resources (34), where species in well-mixed environments with little spatial environmental heterogeneity compete for the same resources, which would result in a negative biodiversity-ecosystem functioning relationship. On the other hand, higher rates of ecosystem function with increasing diversity could emerge through niche partitioning, where species adapt and specialize in different ecological niches and resources within a community (35). The difference in resource use ultimately leads to an augmentation in the overall exploitation of the available resources increasing resource use efficiency. While pelagic diazotrophs occupy similar environments, different kinds of diazotrophs may co-exist by relying on diverse ecophysiological strategies to achieve BNF. For example, *Trichodesmium* spp. fix nitrogen during the day with the highest nitrogen fixation rates around noon (36), while other species from the genera *Crocosphaera* spp. (*UCYN-B*) and *Cyanothece* spp. (*UCYN-C*) fix nitrogen at night (37, 38). Others, such as *Richelia* and *Calothrix* form heterocysts that protect the oxygen-sensitive nifH gene from being irreversibly damaged by photosynthetically produced oxygen (39). Therefore, we

hypothesize that due to this variety of complementary strategies, higher species richness of diazotrophs increases the resource use efficiency, resulting in higher BNF rates.

Historically, data limitations hampered the testing of such biodiversity-ecosystem functioning relationships in the marine environment on a global scale. Although there have been efforts to compute data compilations for observations of cyanobacterial diazotrophs mostly (24, 25, 40), we fill the gap by compiling a new database about the distribution of marine diazotrophic species and by using species distribution models (SDMs) to derive their biogeographic pattern. This provides us a means to address three key scientific questions: (i) What is the biogeographic distribution of the different diazotrophic species in the global ocean? (ii) What species richness pattern emerges from the assemblage of these individual distributions? And (iii) how does species richness relate to BNF?

Materials and Methods

In this study, we update the existing databases of in situ diazotroph observations (24, 40) with records retrieved from recent surveys, thereby combining diazotroph records detected via traditional (microscopy) or sequence-based (qPCR and metagenomic) techniques. This leads to a database of marine diazotrophs spanning more than 22'000 observations and information on 29 species (SI Appendix, Fig. S1). Each taxon has been modeled individually using a specific set of varying environmental predictor combinations at a resolution of 1° longitude by 1° latitude in space and monthly climatological basis in time. SDMs are empirical models that estimate the realized environmental niche of a taxon by fitting a response curve between the distributions of occurrence data and variables (Table S1) that depict the environmental conditions associated with these occurrences (41). Based on such response curves, SDMs predict habitat suitability indices (HSI) that we converted to presence-absence distribution maps. We then determined the global biogeographical patterns of successfully modeled species (n = 14) via ensembles of SDMs that were optimized to filter out spurious patterns that may emerge from sparse and biased

sample distributions (42, 43). Our use of an ensemble approach in terms of environmental predictors, model types, and background selection strategies permits us to determine global plankton species richness patterns in a robust manner (42, 43). By stacking up taxa-specific predictions of presence-absence maps, we estimate global diazotroph richness in space and time and analyze its emergent correlation with marine BNF retrieved from model-based (44) and observational (25) studies. A detailed description of the methodology applied can be found in the supplementary information provided.

Results

Diazotroph richness and beta diversity

Our ensemble modeling framework predicts a strong latitudinal gradient in diazotroph richness (Fig. 1A and B). Hotspots of diazotroph richness are found in the North Pacific Gyre and the central Indian Ocean, where nearly 70% of the species occur, whereas polar waters display the lowest ensemble richness estimates with less than 1% of the diazotrophs being present. The highest richness estimates of the Atlantic Ocean reach up to 59% of total diazotrophs modeled with maxima located in the South Atlantic Subtropical Gyre and along the western central Atlantic. The emergent global richness pattern is retained when accounting for differences in sampling methodology as shown in Fig. S2 (SI Appendix), where the input data to the modeling pipeline has been re-run using only microscopy-based and sequence-based observations (Pearson $r = 0.98$, $p < 0.001$). When analyzed per 1° latitudinal bins, the ensemble mean richness reaches its maxima at around ~10° north, and ~13° south of the equator with an averaged normalized richness across bins of around 50%. At the equator, our projections show a drop in richness harboring up to 35% of the diazotrophic species. This drop results from the low richness estimates predicted for the upwelling region of the equatorial Pacific Ocean. The global latitudinal

trend of high richness in tropical regions and a poleward decrease in diazotroph richness is consistent across all ensemble members (Fig. 1B).

The ensemble spread is highest at 7.5°- 10.5° northern latitude, 11.5°- 14.5° southern latitude and smallest at 35.5°- 43.5° northern latitude, 33.5°- 40.5° southern latitude (Fig. 1B). Sixteen out of 18 models retain the characteristic dip in richness at the equator. Two non-conforming models show an evenly high richness estimate between 30° north and south of the equator. The global median normalized richness across ensemble members ranges between 0.3 and 0.4 with a slight difference in the range of richness estimates between ensemble members (SI Appendix, Fig. S3A).

To further visualize how differences in community composition are related to the global diazotroph richness gradient, we computed beta diversity using the Jaccard dissimilarity index as the sum of species turnover and species nestedness (45). Both species turnover and nestedness are indices ranging between zero and one, with higher values for species turnover indicating a more complete turnover of species composition and higher values in nestedness indicating a higher fraction of shared species between two locations. We computed species turnover and nestedness for each grid cell as the average across all grid cells. While nestedness estimates are generally lower when compared to species turnover on a global scale, the highest estimates of species turnover are found in tropical and subtropical regions with ocean basin-dependent differences. Species turnover has a higher range with minimum values of 0.19 +/- 0.14 and maxima of 0.66 +/-0.17, while nestedness ranges between 0.12 +/-0.06 and 0.39 +/- 0.10 (Fig. 1C and D). The highest species turnover can be found in the North Atlantic (0.66 +/- 0.17), while being lowest in the eastern part of the North Equatorial Pacific (0.19 +/-0.14; Fig. 1C) For nestedness, the highest estimates are found in the eastern tropical Pacific (0.39 +/-0.10) and Indian Ocean (0.35 +/- 0.13) (Fig. 1D), while for polar regions, we see an even contribution of nestedness and species turnover to Jaccard dissimilarity. To disentangle the relative contribution of species turnover and nestedness to total dissimilarity and thus identify which ecological

process structures the global diazotroph richness gradient, we computed the beta ratio as the ratio between nestedness and the Jaccard dissimilarity index. Overall, species turnover contributes more to the richness gradient and nestedness dominates only in richness hotspots where the highest richness estimates are found, indicating that within the modeled diazotroph community (n = 14), turnover is the dominant process underlying the global diazotroph richness pattern (SI Appendix, Fig. S4). However, when looking at the beta diversity patterns for cyanobacterial diazotrophs alone, we find the highest nestedness in regions with the highest diazotroph richness estimates (SI Appendix, Fig. S12). Differences regarding the global Jaccard dissimilarity estimates between each ensemble member are comparably small, but the background selection strategy can have a strong effect on species turnover (SI Appendix, Fig. S3B).

Global biogeography of diazotrophs

To further understand the biogeography of each diazotrophic taxa individually, we computed the annual mean number of presences across all 12 months for each taxa modeled. The genus *Trichodesmium* displays the highest annual mean number of presences in tropical and subtropical regions of the North Atlantic (annual mean number of presences > 0.98). It is projected as absent in polar regions, though penetrating further north in the North Atlantic when compared to the North Pacific (SI Appendix, Fig. S5A). All unicellular cyanobacteria (*UCYN-A*, *UCYN-B*, and *UCYN-C*; SI Appendix, Fig. S5C, E and G) show strong overlapping biogeographies that are limited to the subtropical and tropical regions and that show a strong decline in the annual mean number of presences towards the poles. For unicellular cyanobacterial diazotrophs, the annual mean number of presences reaches their maxima around ~35° north and south of the equator, while the annual number of mean presences drops in upwelling-influenced regions. This latitudinal pattern holds true across all ensemble members

within the unicellular cyanobacterial species. Similar to *Trichodesmium*, *UCYN-A*, *UCYN-B*, and *UCYN-C* are projected to have a low annual mean number of presences in upwelling regions but increases in oligotrophic gyres, such as in the subtropical north and southern Pacific Gyre, the northern subtropical Atlantic Gyre, and the Indian Ocean with exceptions of the northwestern regions of the Indian Ocean that are more strongly influenced by upwelling. For the genus *Richelia* (SI Appendix, Fig. S5I), the highest annual mean number of presences is projected around ~35° north and south of the equator, with a drop in the annual mean number of presences towards more tropical regions and a sharp decline above the ~35° threshold polewards. This latitudinal pattern is well-conserved across all ensemble members. *Richelia* are projected to be more closely affiliated to marine regions whose environmental conditions are more affected by continental inputs such as the western Atlantic close to the Amazon River delta.

We find a general spatial difference in the global distribution of cyanobacterial and non-cyanobacterial diazotrophs (Fig. 2A and B). While photoautotrophic diazotrophs show low diversity in the productive waters influenced by upwelling processes, non-cyanobacterial diazotrophs do appear to thrive in such nutrient-rich waters. For cyanobacterial organisms, the Indian Ocean shows the highest richness estimates, indicating a potential hotspot of cyanobacterial diazotroph diversity (Fig. 2A). Our model predicts a high richness of non-cyanobacterial diazotrophs in the Pacific, especially in the region affected by the Peruvian Coastal Upwelling System (Fig. 2B).

Finally, to identify the drivers of diazotroph biogeography, we used an ensemble of single factor analysis on each diazotrophic species where we computed a mean rank that was later used to select environmental predictors (SI Appendix, Fig. S6). Sea surface temperature emerged as the most important predictor across all diazotroph taxa on a global scale, explaining up to 36% of the variability in the data for some taxa and with an interquartile range across all taxa between 10-20%. After sea surface temperature, nitrate and phosphate concentrations rank as the

second and third most important predictors with median R-squared values between 0.10 and 0.15.

Diazotroph richness and N2-fixation rates

To test whether a biodiversity-ecosystem function relationship exists between diazotroph diversity and BNF, we analyze the correlation between global marine BNF rates and diazotroph richness using a global model-based estimate of BNF from (44) and in situ measurements derived from a recently compiled database (25). We matched each ensemble richness estimate with the estimated gridded BNF rate and found a significant positive correlation (Spearman's $\rho = 0.8$, $p < 0.001$) (Fig. 3A). BNF rates are higher in the low latitudes where diazotroph richness tends to be high and decrease towards the poles. Furthermore, regions of intense upwelling are recognized as regions of lower BNF and lower species richness. We further analyzed the correlation between our projected ensemble richness and in situ nitrogen fixation measurements. We plotted the in situ BNF rates compiled by Shao et al. 2023 (25) as a function of our projected ensemble richness (Fig. 2B). It is important to keep in mind that the scales covered by the two types of BNF estimates differ significantly. Indeed, while the in situ measured BNF rates correspond to local discrete measurements integrated over 24 hours influenced by submeso- and mesoscale processes, the estimate from Fig. 2A comes from a mechanistic model that is representative of mean annual scales. Due to the high range of observed BNF rates spanning several orders of magnitudes within the in situ measurements, the positive correlation between in situ measurements and richness holds true only on a log-scale (Spearman's $\rho = 0.19$, $p < 0.001$) (Fig. 2B).

Discussion

This work provides the first estimates for global diazotroph diversity and its relationship with in situ measured nitrogen fixation rates. The integration of classical microscopy-based and sequence-based (qPCR and metagenomic) data sources allowed us to model the biogeography of 14 diazotroph species representing major life history strategies of oceanic diazotrophs and included marine non-cyanobacterial diazotrophs that have not been included previously in any global scale biogeographic study. In accordance with previous studies based on SDMs of phyto- and zooplankton species (42, 43), as well as studies based on metagenomic surveys (22, 25, 46), we found diazotroph diversity to be highest in tropical and subtropical regions and to decrease towards the poles (Fig. 1B).

The inclusion of non-cyanobacterial diazotrophs into previously cyanobacteria-centric models of diazotroph diversity unveiled complementary patterns of their spatial distribution and overlap with cyanobacterial diazotrophs in richness hotspots (SI Appendix, Fig. S5 A-L). The importance of non-cyanobacterial diazotrophs has been recognized only within the last decade after extensive marine surveys such as Tara Ocean collected metagenomes from the marine environment. Recent evidence showed that non-cyanobacterial diazotrophs considerably expand the known diversity of abundant marine nitrogen fixers (47). A further analysis of samples taken during the Tara Ocean expedition identified non-cyanobacterial diazotrophs within the particle-attached fraction (22). These findings suggest that non-cyanobacterial diazotrophs are present in nutrient-rich upwelling regions where high primary production leads to higher amounts of particulate organic matter, a carbon-rich resource that may create oxygen-low micro-niches that can be occupied by non-cyanobacterial diazotrophs. Our results show the highest richness of non-cyanobacterial diazotrophs is found in upwelling-influenced marine regions such as the Peruvian Coastal Upwelling and are therefore in line with such studies. Thus, the distribution of non-cyanobacterial diazotrophs needs to be taken into account in future modeling and

experimental studies aiming at the quantification of ecosystem functions associated with the global biogeochemical cycling of nitrogen.

Based on the diazotroph richness diagnosed by our models and using either model-based BNF rates from Wang et al. (44) or in-situ measurements (25), we found a positive correlation between BNF and diazotroph richness (Spearman's $\rho = 0.8$, $p < 0.001$ in Fig. 3A and Spearman's $\rho = 0.19$, $p < 0.001$ in Fig. 3B), supporting the view that diazotroph diversity may be a major control on global BNF rates. Since our hypothesis is based on cyanobacterial nitrogen fixation strategies that vary depending on optimal environmental conditions present in the marine environment, we analyzed the correlation between diazotroph richness either based on cyanobacterial and non-cyanobacterial species (SI Appendix, Fig. S10A and B). We found a stronger correlation between cyanobacterial-based diazotroph richness and BNF (Spearman's $\rho = 0.26$, $p < 0.001$) which further supports the hypothesis according to which cyanobacterial niche complementarity leads to increased BNF. We found the correlation between in situ measured BNF and non-cyanobacterial richness to be much weaker (Spearman's $\rho = 0.09$, $p < 0.05$). While our correlative methods preclude the identification of causal links between BNF and diazotroph richness, the current diazotroph literature provides us with ample evidence to formulate plausible hypotheses that may be tested in future mechanistic or experimental work.

One mechanism that may drive the positive relationship between diazotroph species diversity and N2 fixation rate is the ecological complementary effect (48). According to this mechanism, differences in the ecological strategies of taxa lead to greater partitioning of the available resources, enhanced coexistence, enhanced resource use efficiency and thus enhanced community-level productivity, in particular under stable environmental conditions such as those prevailing in the tropics (48), where highest diazotroph richness estimates have been projected by our SDMs (Fig. 1A). Those slight differences involved in metabolic processes within marine diazotrophs link to differences in the amount of nitrogen fixed depending on ideal environmental

conditions being present, and the accumulated effect of strategies, hereby reflected as increased species richness, may lead to higher BNF with increasing diazotroph richness. These elements suggest that the correlative relationships emerging from our data may have a physiological and ecological foundation. As shown in Fig. S11 (SI Appendix), cyanobacterial nestedness estimates are highest in regions where we find the highest diazotroph richness and BNF rates. This supports the hypothesis that niche complementarity can promote species diversity and coincide with a higher resource use efficiency ultimately leading to higher BNF rates.

We partially overcame the major hurdle of data limitations to model microbial communities on a global scale by merging observations that originate from varying sampling methodologies to maximize the number of observations and analyzed if the richness projections differ when using microscopy or sequence-based data only. Here we show that observations either retrieved from microscopy or sequence-based methodologies lead to the same global richness patterns of marine diazotrophs if considered in isolation, and can thus be analyzed in combination (SI Appendix, Fig. S2) indicating that the origin of the data source has no effect that substantially skews our SDM results. Hence, our results highlight the potential compatibility of observations originating from different sampling strategies to alleviate the effect of data scarcity in modeling work and show that reliable projections can be extracted from both data types, microscopic and sequence-based. This increases our confidence in merging observations from different sampling methodologies for microbial plankton taxa to increase the pool of observations for future studies that aim to constrain uncertainties related to small sampling sizes.

Taken together, our work links diazotroph diversity to BNF at the global scale and supports the existence of an important biodiversity-ecosystem functioning relationship in marine plankton. As such, this highlights the potential of the diversity of individual plankton functional groups to govern the strength and global distribution of essential processes relevant to global biogeochemical cycles. To our knowledge, this is the first study that shows an emergent positive

relationship between diazotroph species diversity and the performance of BNF and suggests, that the diversity of functional groups must (a) be taken into account in future earth system models, and (b) needs to be taken into account in efforts that aim at the conservation of essential ecosystem functions associated with biogeochemical cycling.

Data availability: All codes and data used for this analysis are publicly available at the ETH Zurich Research collection with the doi: 10.3929/ethz-b-000635803.

Acknowledgments

This project has received financial support from the European Union's Horizon 2020 research and innovation program under grant agreement No. 862923 (AtlantECO), the NOTION project funded by the Fondation BNP Paribas grant 1-006245-000 and the Swiss National Science Foundation (SNSF) through project grants 205321_184955. The authors thank Mar Benavides for leading the NOTION project. This output reflects only the author's view and the European Union cannot be held responsible for any use that may be made of the information contained therein. A special thanks to Mridul Thomas for the fruitful discussion and ideas.

References

1. T. Shiozaki, *et al.*, Linkage Between Dinitrogen Fixation and Primary Production in the Oligotrophic South Pacific Ocean. *Global Biogeochem. Cycles* **32**, 1028–1044 (2018).
2. D. Karl, *et al.*, The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature*, 533–538 (1997).
3. N. Gruber, J. N. Galloway, An Earth-system perspective of the global nitrogen cycle. *Nature* **451**, 293–296 (2008).

4. J. P. Zehr, D. G. Capone, Changing perspectives in marine nitrogen fixation. *Science* **368** (2020).

5. C. Wu, *et al.*, Heterotrophic Bacteria Dominate the Diazotrophic Community in the Eastern Indian Ocean (EIO) during Pre-Southwest Monsoon. *Microb. Ecol.* **78**, 804–819 (2019).

6. H. Farnelid, *et al.*, Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLoS One* **6**, e19223 (2011).

7. L. Wrightson, A. Tagliabue, Quantifying the Impact of Climate Change on Marine Diazotrophy: Insights From Earth System Models. *Frontiers in Marine Science* **7** (2020).

8. B. Bergman, G. Sandh, S. Lin, J. Larsson, E. J. Carpenter, Trichodesmium--a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol. Rev.* **37**, 286–302 (2013).

9. D. G. Capone, *et al.*, Nitrogen fixation byTrichodesmiumspp.: An important source of new nitrogen to the tropical and subtropical North Atlantic Ocean. *Global Biogeochem. Cycles* **19** (2005).

10. R. A. Foster, J. P. Zehr, Characterization of diatom-cyanobacteria symbioses on the basis of nifH, hetR and 16S rRNA sequences. *Environ. Microbiol.* **8**, 1913–1925 (2006).

11. F. Hashihama, J. Kanda, Y. Maeda, H. Ogawa, K. Furuya, Selective depressions of surface silicic acid within cyclonic mesoscale eddies in the oligotrophic western North Pacific. *Deep Sea Res. Part I* **90**, 115–124 (2014).

12. T. O. Delmont, *et al.*, Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* **3**, 804–813 (2018).

13. A. Thompson, *et al.*, Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ. Microbiol.* **16**, 3238–3249 (2014).

14. S. R. Bench, I. Frank, J. Robidart, J. P. Zehr, Two subpopulations of Crocosphaera watsonii have distinct distributions in the North and South Pacific. *Environ. Microbiol.* **18**, 514–524 (2016).

15. K. A. Turk-Kubo, H. M. Farnelid, I. N. Shilova, B. Henke, J. P. Zehr, Distinct ecological niches of marine symbiotic N -fixing cyanobacterium Candidatus Atelocyanobacterium thalassa sublineages. *J. Phycol.* **53**, 451–461 (2017).

16. A. N. Knapp, *et al.*, Nitrogen isotopic evidence for a shift from nitrate- to diazotroph-fueled export production in the VAHINE mesocosm experiments. *Biogeosciences* **13**, 4645–4657 (2016).

17. F. M. Cornejo-Castillo, J. P. Zehr, Intriguing size distribution of the uncultured and globally widespread marine non-cyanobacterial diazotroph Gamma-A. *ISME J.* **15**, 124–128 (2021).

18. B. A. Ward, S. Dutkiewicz, C. M. Moore, M. J. Follows, Iron, phosphorus, and nitrogen supply ratios define the biogeography of nitrogen fixation. *Limnol. Oceanogr.* **58**, 2059–2075 (2013).

19. T. Shiozaki, *et al.*, Biological nitrogen fixation detected under Antarctic sea ice. *Nat. Geosci.* **13**, 729–732 (2020).

20. S. G. Acinas, *et al.*, Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun Biol* **4**, 604 (2021).

21. C. F. Reeder, C. R. Löscher, Nitrogenases in Oxygen Minimum Zone Waters. *Frontiers in Marine Science* **9** (2022).

22. J. J. Pierella Karlusich, *et al.*, Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. *Nat. Commun.* **12**, 4160 (2021).

23. Y.-W. Luo, I. D. Lima, D. M. Karl, C. A. Deutsch, S. C. Doney, Data-based assessment of environmental controls on global marine nitrogen fixation. *Biogeosciences* **11**, 691–708 (2014).

24. W. Tang, N. Cassar, Data-driven modeling of the distribution of diazotrophs in the global ocean. *Geophys. Res. Lett.* **46**, 12258–12269 (2019).

25. Z. Shao, *et al.*, Version 2 of the global oceanic diazotroph database. *Earth Syst. Sci. Data Discuss.* (2023) https:/doi.org/10.5194/essd-2023-13.

26. M. R. Gradoville, *et al.*, Latitudinal constraints on the abundance and activity of the cyanobacterium UCYN-A and other marine diazotrophs in the North Pacific. *Limnol. Oceanogr.* **65**, 1858–1875 (2020).

27. M. Chen, *et al.*, Biogeographic drivers of diazotrophs in the western Pacific Ocean. *Limnol. Oceanogr.* **64**, 1403–1421 (2019).

28. A. M. S. Detoni, A. Subramaniam, S. T. Haley, S. T. Dyhrman, P. H. R. Calil, Cyanobacterial diazotroph distributions in the western South Atlantic. *Front. Mar. Sci.* **9** (2022).

29. C. Martínez-Pérez, *et al.*, The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat Microbiol* **1**, 16163 (2016).

30. D. Louchard, M. Münnich, N. Gruber, On the role of the Amazon River for N_2 fixation in the western tropical Atlantic. *Global Biogeochem. Cycles* **37** (2023).

31. D. Tilman, F. Isbell, J. M. Cowles, Biodiversity and Ecosystem Functioning. *Annu. Rev. Ecol. Evol. Syst.* **45**, 471–493 (2014).

32. L. Gamfeldt, *et al.*, Marine biodiversity and ecosystem functioning: what's known and what's next? *Oikos* **124**, 252–265 (2015).

33. S. Lehtinen, T. Tamminen, R. Ptacnik, T. Andersen, Phytoplankton species richness, evenness, and production in relation to nutrient availability and imbalance. *Limnol. Oceanogr.* **62**, 1393–1408 (2017).

34. I. T. Carroll, B. J. Cardinale, R. M. Nisbet, Niche and fitness differences relate the maintenance of diversity to ecosystem function. *Ecology* **92**, 1157–1165 (2011).

35. D. Muratore, *et al.*, Complex marine microbial communities partition metabolism of scarce resources over the diel cycle. *Nat Ecol Evol* **6**, 218–229 (2022).

36. I. Berman-Frank, *et al.*, Segregation of nitrogen fixation and oxygenic photosynthesis in the marine cyanobacterium Trichodesmium. *Science* **294**, 1534–1537 (2001).

37. J. Toepel, E. Welsh, T. C. Summerfield, H. B. Pakrasi, L. A. Sherman, Differential transcriptional analysis of the cyanobacterium Cyanothece sp. strain ATCC 51142 during light-dark and continuous-light growth. *J. Bacteriol.* **190**, 3904–3913 (2008).

38. T. Shi, I. Ilikchyan, S. Rabouille, J. P. Zehr, Genome-wide analysis of diel gene expression in the unicellular N(2)-fixing cyanobacterium Crocosphaera watsonii WH 8501. *ISME J.* **4**, 621–632 (2010).

39. P. Fay, Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiol. Rev.* **56**, 340–373 (1992).

40. Y.-W. Luo, *et al.*, Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates. *Earth Syst. Sci. Data* **4**, 47–73 (2012).

41. J. Elith, J. R. Leathwick, Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annu. Rev. Ecol. Evol. Syst.* **40**, 677–697 (2009).

42. F. Benedetti, *et al.*, Major restructuring of marine plankton assemblages under global warming. *Nat. Commun.* **12**, 5226 (2021).

43. D. Righetti, M. Vogt, N. Gruber, A. Psomas, N. E. Zimmermann, Global pattern of phytoplankton diversity driven by temperature and environmental variability. *Sci Adv* **5**, eaau6253 (2019).

44. W.-L. Wang, J. K. Moore, A. C. Martiny, F. W. Primeau, Convergent estimates of marine nitrogen fixation. *Nature* **566**, 205–211 (2019).

45. A. Baselga, C. D. L. Orme, betapart : an R package for the study of beta diversity. *Methods Ecol. Evol.* **3**, 808–812 (2012).

46. S. Sunagawa, *et al.*, Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).

47. T. O. Delmont, Discovery of nondiazotrophic species abundant and widespread in the open ocean. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021).

48. M. Striebel, S. Behl, S. Diehl, H. Stibor, Spectral niche complementarity and carbon dynamics in pelagic ecosystems. *Am. Nat.* **174**, 141–147 (2009).

Figures

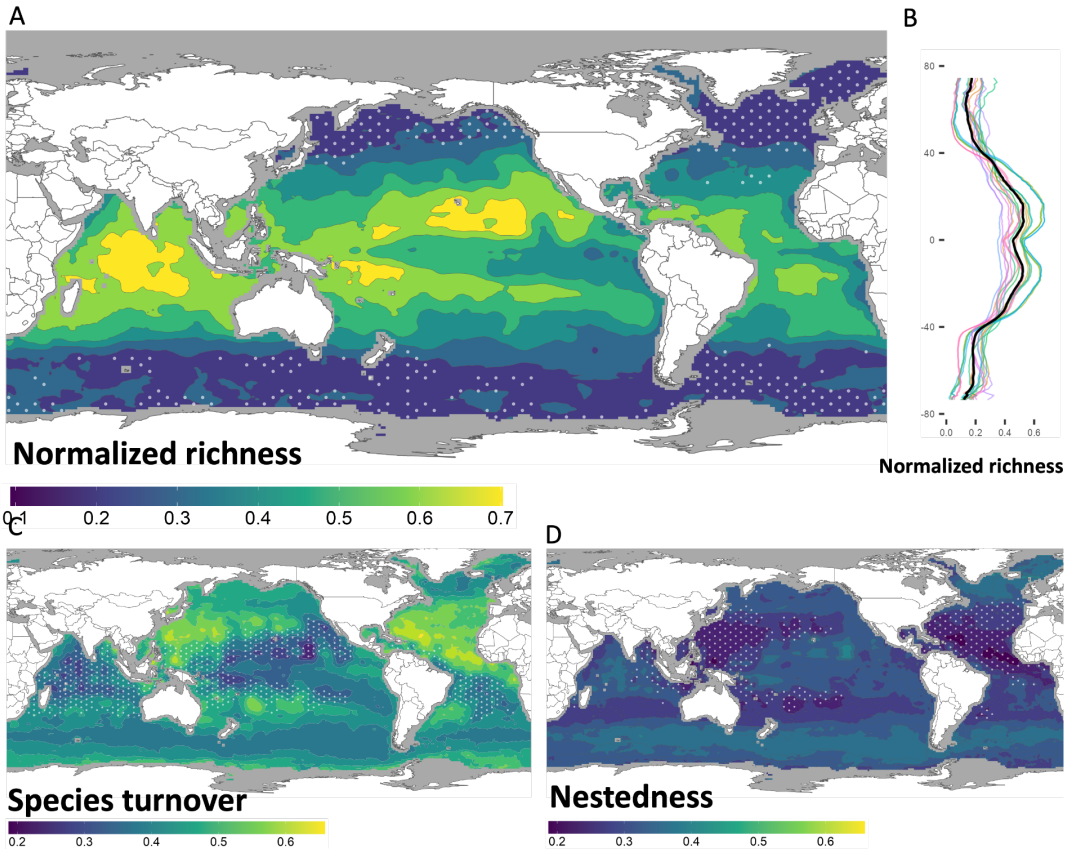

Figure 1. Global diazotroph ensemble diversity. A) Global annual ensemble mean of diazotroph species richness. B) Global 1° binned latitudinal richness gradients for each one of the 18 models (colored lines) the ensemble has been generated from. The black line is the mean across all 18 models. White stipples indicate areas where the coefficient of variation was above the 70th percentile marking greater differences between model projections. C) Global annual ensemble species turnover. D) Global ensemble nestedness. Species turnover measures the degree of species replacement and nestedness observes species assemblages that are smaller subsets of larger sets, based on Jaccard's dissimilarity index.

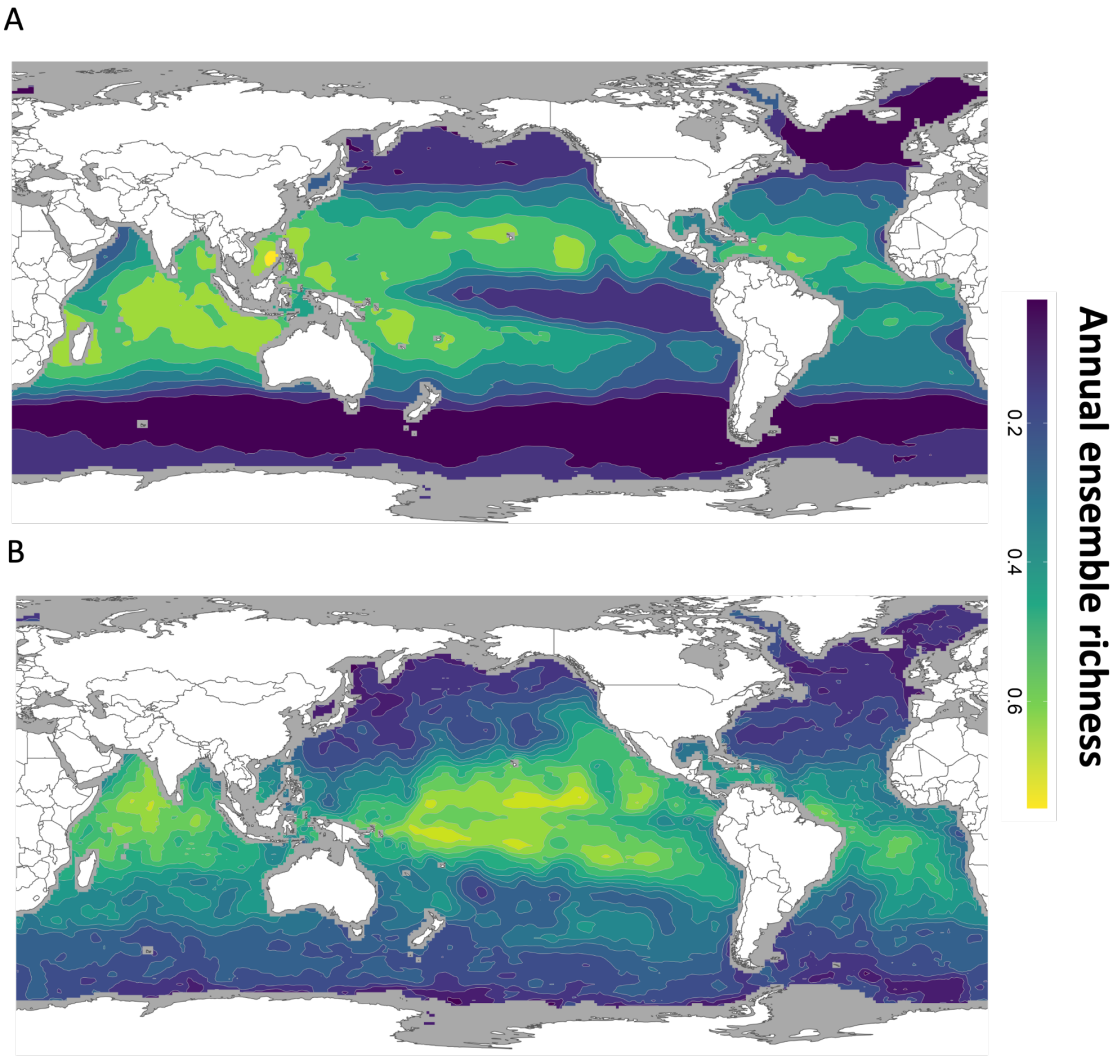

Figure 2. Annual ensemble mean species richness of A) cyanobacterial and B) non-cyanobacterial diazotrophs. The ensemble (n = 18) has been computed from 1° longitudinal and 1° latitudinal monthly Species Distribution Model outputs that account for uncertainties related to predictors, algorithms and background selection strategies chosen. Ensemble richness has been normalized by the number of species modeled with blue colors indicating low richness and the yellowish colors increasing the annual ensemble richness.

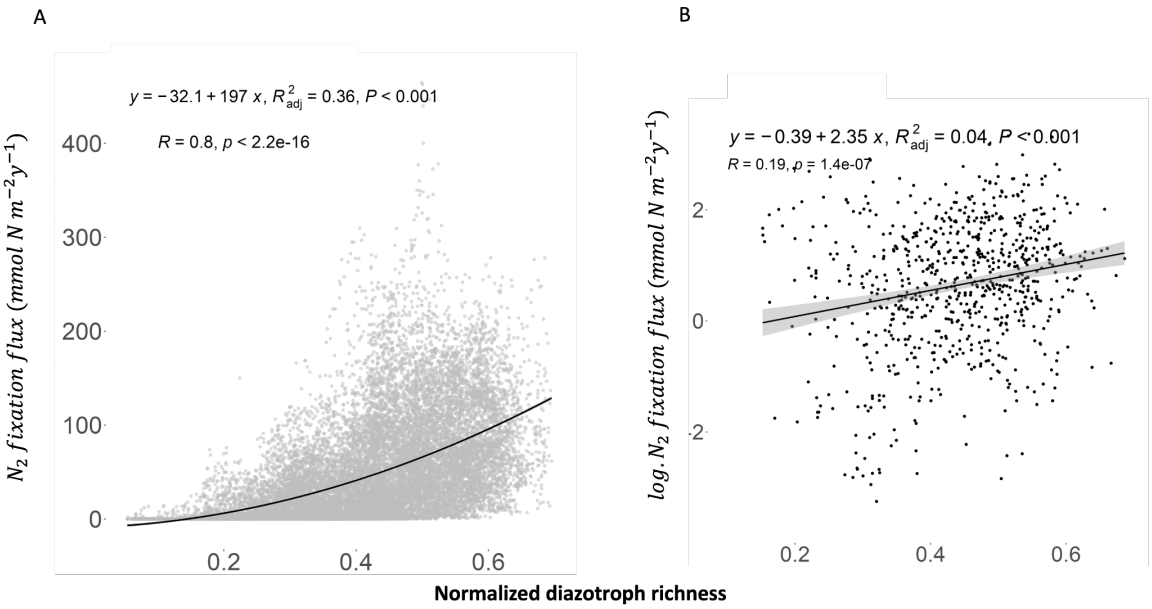

Figure 3. Relationship between diazotroph species richness and biological nitrogen fixation (BNF). Shown are correlations between nitrogen fixation rates and global annual diazotroph richness. A) Nitrogen fixation rates originate from (44). The black line indicates a 2nd order polynomial fit. Further indicated are Spearman's rank correlation coefficient and the statistical p-value. B) Correlation between grid cells that contain in situ nitrogen fixation measurements compiled by Shao et al. (2023) and projected annual diazotroph richness from an ensemble of species distribution models. Spearman correlation coefficient is given in the top left corner with the associated p-value. Grey shading indicates the 0.95 confidence interval.

Supporting Information

Extended Material and Methods

Diazotroph occurrence dataset. We compiled an exhaustive dataset of diazotroph occurrences from public sources and from recent studies that focused on either quantitative– (counts or gene reads) and qualitative (presence-absence, non-detection) field records of planktonic diazotrophs. We compiled diazotrophic data from the Global Biodiversity Information Facility (GBIF; www.gbif.org, last access: 20 October 2021), the Ocean Biogeographic Information System (OBIS; www.obis.org/, last access: 21 October 2021), Luo et al. (1), Tang and Cassar (2), Gradoville et al. (3), Detoni et al. (4), Martínez-Pérez et al. (5), Phytobase (6), Pierella Karlusich et al. (7) and Paoli et al. (8) (Figure S1).

We included further quality controls on observations that have been retrieved via public databases such as GBIF or OBIS. We used an ocean mask (9) to ensure that only marine taxa were included and any observation displaying a doubtful taxonomic assignment in the original datasets was removed. To avoid the inclusion of outdated species names from early sampling periods, each taxon name of microscopic origin was screened against the World Register of Marine Species (WoRMS, https://www.marinespecies.org) and only taxa with an accepted status were included. WoRMS was further used for taxonomic harmonization regarding microscopy retrieved observations and annotations. Scientific names whose taxonomic status was flagged as unaccepted in WoRMS were either removed or corrected by an alternative accepted name. When information about the measurement method was missing from the original datasets, we screened the associated publications to backtrack the methodology used to identify each observation. When no information on the method was found within the complete dataset, we checked the time period of the sampling event. Observations before 1980 were assumed to be

microscope-based since sequence-based taxonomy was not established in the scientific community back then. For some studies, the exact days were not recorded but a several-week period was given. In those cases, we assigned a specific day from the covering period since none of those mentioned periods were extensively long (all periods < three weeks).

We further included records of non-cyanobacterial diazotrophs (10). We used the metagenomic assembled genomes (MAGs) computed by Delmont et al. (10) and screened the Ocean Microbiomics Database (8) which compiles data from Sunagawa et al. (11), Salazar et al. (12), Biller et al. (13), Acinas et al. (14), Delmont et al. (10), Klemetsen et al. (15) and Pachiadaki et al. (16) for matching metagenomic operational taxonomic units (mOTUs) to retrieve a taxonomic annotation of those genomes using the Genome Taxonomy Database (17). Column names or data fields were adjusted and harmonized to establish compatibility in the dimensions of the different source datasets following Darwin Core standards (https://dwc.tdwg.org). To remove duplicates, an occurrence ID was created considering the columns “family”, “genus”, ”species”, “decimalLongitude”, “decimalLatitude”, “year”, “month”, “day” and “depth”.

Open ocean environmental conditions. Environmental parameters were compiled to reflect key dimensions of microbial plankton niches that shape species’ distributions via effects on physiology, growth or species competition (Table S1) (18–20). Since we focused on the diazotroph community of the global offshore ocean, we limited the confounding influences of complex and fertile coastal environments by excluding data from seas shallower than 200 m (21) and from regions characterized by climatological surface salinities below 20 (22).

Species distribution models. SDMs fit statistical associations between species’ observed occurrences and environmental variables; i.e., they estimate a species’ realized environmental

niche (19). SDMs provide a useful framework to explore large-scale distributions of microbial plankton species. They assume that: (i) species are not dispersal-limited in the open ocean (23) a trait consistent with the generally wide geographic ranges of the species in the data; (ii) species are primarily controlled by abiotic environmental factors in their global distribution (23), and rapidly respond when conditions turn suitable (24). Since the distribution patterns of diazotrophic taxa are likely to change seasonally (25), we used a monthly match-up between species' occurrences and the environmental variables to train the SDMs. Then, we projected the SDMs onto global environmental data fields at 1° and monthly resolution to obtain maps of the species' habitat suitability index (HSI; also called "presence probability" in the literature), or distribution maps of presence-absence after applying a probability threshold to the HSI maps. We follow the standard SDM ensemble framework of Righetti et al. (18) and Benedetti et al. (26) which has been shown to robustly model species distributions and the associated emergent patterns of species diversity. We further developed an ensemble of SDMs that address three key sources of uncertainty: (i) sampling bias, (ii) predictor choice, and (iii) algorithm choice.

We converted all quantitative data to presence-only data for the present study and interpreted zeros as absences. We binned the species' presences into the monthly 1° × 1° cell grid to match the resolution of the environmental predictors. Multiple observations per species and 1° cell that came from the same month although from potentially different years were counted as a single monthly presence. The final occurrence dataset recorded a total of more than 6500 gridded presences across 29 taxa available for the SDMs.

Target-group approaches to sample background data. To inform the correlative SDMs about the parts of the environmental space that are less suitable for the species to be present, we had to generate background data (also termed "pseudo-absences" in the literature). We selected

environmental background data for each species, using the target-group approach (18, 27). Here, we sampled the background data based on the target group to: (i) ensure that background sampling follows a sampling scheme similar to that of the presence data, thereby balancing presence data bias when fitting SDMs; (ii) ensure that extensive ocean areas characterized by lower sampling density were not artificially misclassified as areas of lower species' habitat suitability. We use a larger number of pseudo-absences (presence/absence ratio = 1/10) with equal weighting for regression-based techniques and a ratio of 1/1 for tree-based models as suggested by Barbet-Massin et al. (28).

We defined three different target groups: the "total target group", the "group-specific target group" and the "cruise-specific target group". The "total target group" approach included taxonomic records from Phytobase (6) which fell into the surface ocean mixed-layer, excluding records from the comparably larger-sized diatoms and dinoflagellates. The "group-specific target group" consisted of all locations from the compiled diazotroph database used in this study and the "cruise-specific target group" provided background information from cruises that used the same sampling method. This use of varying background selection strategies is a powerful tool if the particular method is applied in a sufficiently broad environmental context and across multiple taxa. In the context of mostly data-deficient diazotrophs, several methods have only been applied in certain ocean basins and without a regular grid. Therefore, to provide enough environmental variability for the modeling it is important to strive for extensive datasets, merging observations that originate from varying sampling methodologies.

Under all three configurations of the target group approach, we sampled background data in a stratified manner from the target group following the procedure of Righetti et al. (18) and summarized hereafter. We incorporated two environmental gradients (sea surface temperature and mixed layer depth) during the background selection to ensure that the breadth of the chosen key environmental factors was reflected in the background of each taxon. Background data were

sampled with overlapping and non-overlapping options. The overlapping option means that the taxon modeled remains part of the background and can provide background data itself, while the non-overlapping option refers to the case where the presence cells of the model taxon are excluded from the background. While the overlapping option generates a background that is more general (i.e., pseudoabsences reflect environmental conditions in the study domain) the latter is more specific.

Algorithm and complexity choice. Statistical algorithm choice represents the main source of uncertainty in studies relying on SDMs (29). We constructed SDMs based on either General Linear Models (GLMs; using the R package "stats"), General Additive Models (GAMs; R package "mgcv"), or Random Forests (RFs; R package "randomForest"), as three algorithms of increasing statistical response shape complexity. We used comparably few predictors (n = 4) in models and fitted simple response shapes to account for the relatively few presences of most diazotroph species. We considered species with at least 24 presences (across all possible monthly 1° cells) for modeling, following recommendations by Brun et al. (30), where one predictor per 10 presence observations would be ideal. GLM included linear and quadratic terms and a stepwise bidirectional predictor selection procedure. GAM used smoothing terms with four basis dimensions (k = 4), estimated by penalized regression splines without penalization to zero for single variables. To balance the overall weight of presences versus background data per species, background data in GAM and GLM were weighted by the ratio of species' presence to background data points. RFs included 4'000 trees, simple terms, and single-end node size. The weighting of data in individual RF trees was balanced by randomly subsampling the same amounts of background data as the species had presences as suggested by Barbet-Massin et al. (28). In cases where the sampling of absences resulted in a lower number than presences

due to the lack of potential locations valid for drawing absences, presences have been downsampled to the number of absences found, to keep the 1:1 ratio, when running the RF.

Predictor ranking and selection for member models. In addition to algorithm choice, predictor choice represents a potentially important source of uncertainty in the present SDMs, as the environmental variables controlling the spatial distributions of planktonic diazotrophs remain poorly known. We fitted single-factor GLM, GAM, and RF models to the presence versus background data of each taxon, for each candidate predictor using the same model parametrization as in the SDMs. Model explanatory skill was evaluated using the adjusted D^2 for GLM, adjusted R^2 for GAM and the Out-Of-Bag Error statistic for RF. For each species, predictors were ranked according to these statistics, and a predictor ensemble using the mean variable ranks was obtained across GLM, GAM, and RF, which served as a basis for predictor selection.

To capture predictor-based uncertainty, we fitted five ensemble model members per taxon, each using a different set of four predictors and built an ensemble of SDMs. We used a randomization approach to select the four predictors per member model, using the predictor pre-ranking of each taxon as a basis (Table S2). For each pair of predictors, we computed pairwise Spearman's rank correlation coefficients since collinearity between predictors can inflate the standard errors of regression model parameters and inflate their variance in regressive models leading to biased SDM projections (31). For predictor pairs with a Spearman's rank correlation higher than 0.7, only one predictor was used in the SDM.

Evaluation of member models and ensemble prediction. For each species, we evaluated the predictive skill of each ensemble model member based on fourfold cross-validation. In this cross-validation, the species' presences and background data were randomly split into four fractions

with approximately equal numbers of presences and pseudoabsences each. The ensemble model member was iteratively trained on 75% of the data and the predictions were evaluated against the remaining 25%. We used the True Skill Statistic (TSS) to quantify model skill (i.e., for each member model, per taxon). The TSS ranges from -1 to +1 with values greater than zero indicating models performing better than random. We retained members showing a TSS score of at least 0.30 to build the model ensemble. Fig. S8 shows boxplots of TSS scores of each successfully modeled diazotroph across all member models. Successful model members were then projected globally onto monthly (n = 12 months) environmental data fields, yielding species-level maps of HSI. Before estimating species richness, we converted the HSI maps to presence-absence maps based on the probability threshold that maximizes the TSS using the function "optimal.thresholds" from the PresenceAbsence package in R (32).

Diazotroph diversity. Richness is the simplest measure of diversity and corresponds to the number of species present within one location or grid cell. We stacked the monthly SDM projections from each ensemble member of the successfully modeled taxa (n = 14) within each grid cell. For each model, we then calculated the monthly richness estimate by summing up the species-level presence-absence maps. The annual ensemble mean richness was then computed as the mean richness of each grid cell across all ensemble members and all twelve months. The final map on diazotroph richness is therefore computed as an ensemble of 18 models each including 14 diazotroph species that passed the evaluation criteria (at least 24 observations and a TSS score higher than 0.3) including the three different SDMs (GLM, GAM, RF) and six different background selection strategies (total target-group, group-specific, cruise-specific with overlapping and non-overlapping options). Species richness was then normalized by the number of species to map the relative number of species present within one grid cell. The inclusion of

several SDMs has the advantage of covering one source of main uncertainty in SDMs, which results from model choice, to increase generalization.

To estimate beta diversity and investigate the ecological processes of the global gradient of diazotrophs species richness, we calculated the Jaccard index and its two components i) species turnover and ii) species nestedness for each model. We used the function beta.pair from the betapart package in R (33) according to the equation:

$$ß_{jac} = ß_{jtu} + ß_{jne} = \frac{b+c}{a+b+c} = \frac{2b}{2b+a} + \left(\frac{c-b}{a+b+c}\right)\left(\frac{a}{2b+a}\right) \quad \text{Equation 1}$$

where $ß_{jac}$ is the Jaccard dissimilarity, $ß_{jtu}$ the turnover component of Jaccard dissimilarity, and $ß_{jne}$ the nestedness component of Jaccard dissimilarity. We perform a pairwise dissimilarity between all pairs of sites and average the results for each grid cell across all other grid cells, where *a* is the number of shared species between two cells, *b* the number of species unique to the poorest site and *c* the number of species unique to the richest site. The final ensembles of the three beta diversity indices were computed the same way as the ensemble species richness: we averaged the estimates across 18 ensemble member models retrieving an ensemble estimate for each grid cell for species turnover and nestedness. Moreover, the beta ratio between nestedness was computed ($ß_{ratio} = ß_{jne}/ß_{jac}$) to understand which component has the highest contribution to total dissimilarity. A $ß_{ratio}$ that is higher than 0.5 would therefore indicate that the region is dominated by nestedness rather than species turnover and alternatively a value lower than 0.5 would indicate the opposite.

Biological nitrogen fixation rates. To analyze the correlation between global marine biological nitrogen fixation (BNF) and diazotroph richness, we used the global BNF estimates originating from two independent studies. Wang et al. (34) published global marine BNF rates from the Community Earth System Model (CESM) model simulations (https://www.cesm.ucar.edu/models), where nitrogen cycle simulations were conducted using a modified version of the ocean component. Additionally, we use in situ measured BNF compiled by Shao et al. (35). To make both BNF rates comparable to the global diazotroph richness estimate, we re-gridded both global estimates to a 1° latitude x 1° longitude resolution and correlated each richness estimate within a grid cell with the corresponding BNF rate.

Uncertainty analysis. We carefully assessed the influence of the choices made in the SDM framework on our final estimate of mean annual diazotroph species richness as such choices generate uncertainty (i.e., variability) around this estimate. To do so, we included several modeling strategies as described above (background selection, predictor choice, and algorithm complexity) to analyze how different choices influence the biogeographies of the individual diazotroph taxa. Species richness estimates were computed for each SDM output and the coefficient of variation was used to quantify differences in the global distribution of richness. To account for uncertainties related to differences in sampling methodologies, we split our total dataset into observations that are either microscopy-based or sequence-based. We then re-analyzed those datasets and further computed SDMs for each of those three datasets. We compared the individual species distributions and diversity patterns to assess if significantly different patterns arise, or if these patterns follow a similar distribution which would justify merging different types of observations (Fig. S2).

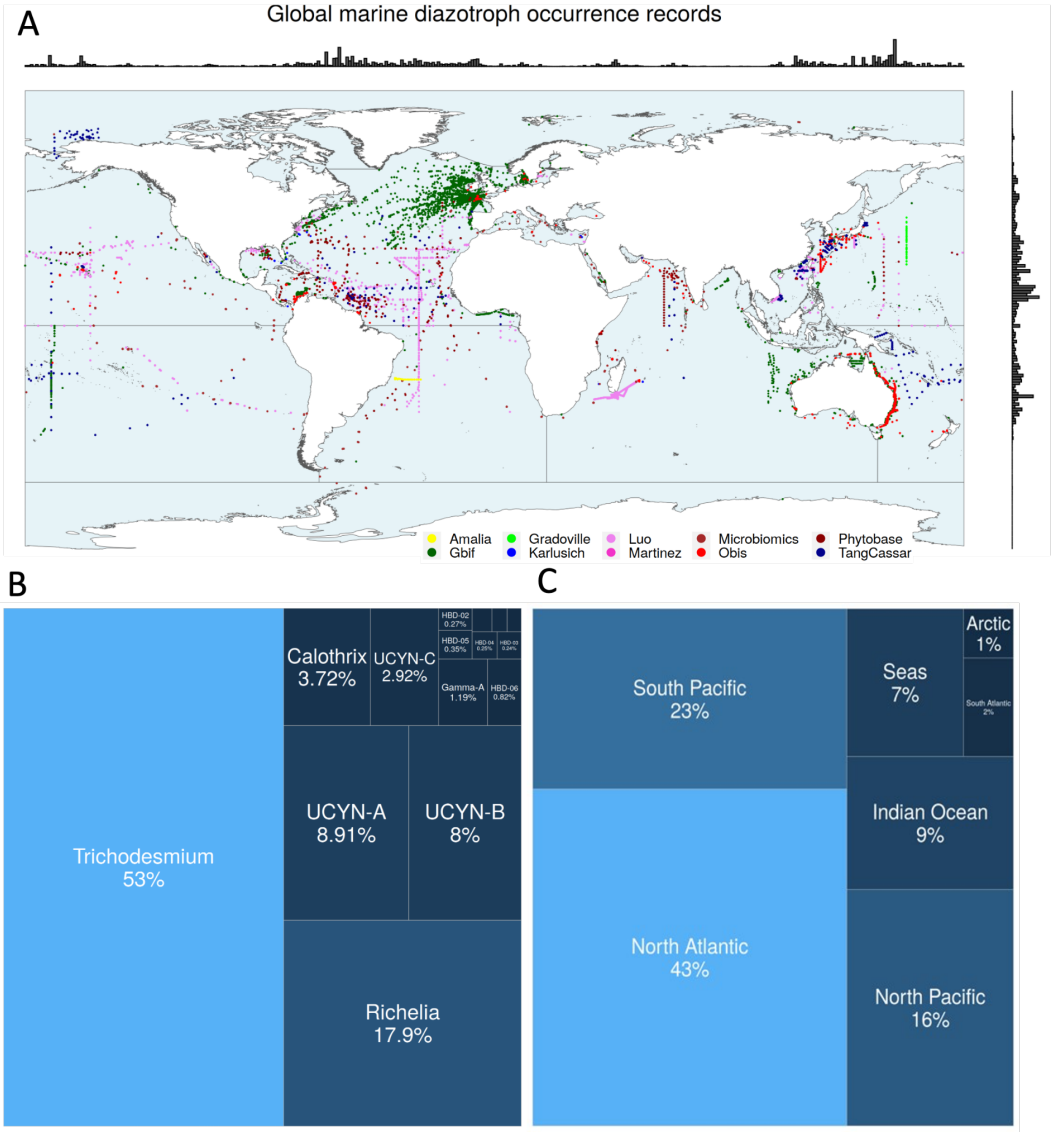

Fig. S1. Figure S1: Global map of diazotroph observations: A) Global map of diazotroph observation (n > 22.000) with longitudinal and latitudinal marginal histograms colored by sources (yellow: Detoni et al. (2022), green: Gradoville, pink: Luo et al. (2012), brown: Ocean Microbiomics Database (Paoli et al., 2022), dark red: Phytobase (Righetti 2019), dark green: GBIF, blue: Karlusich et al. (2021), dark pink: Martinez et al. (2016), red: OBIS, dark blue: Tang and Cassar (2019). B) A treemap showing the fraction of total observations/sampling effort in percentage for

each ocean basin. Percentages decrease along light to dark blue color gradient. The Southern Ocean is not shown, as the fraction of observational records falling into this region is below 1%. C) A treemap showing the percentage fraction of individual diazotroph taxa in percentage. Percentages decrease along light to dark blue color gradient.

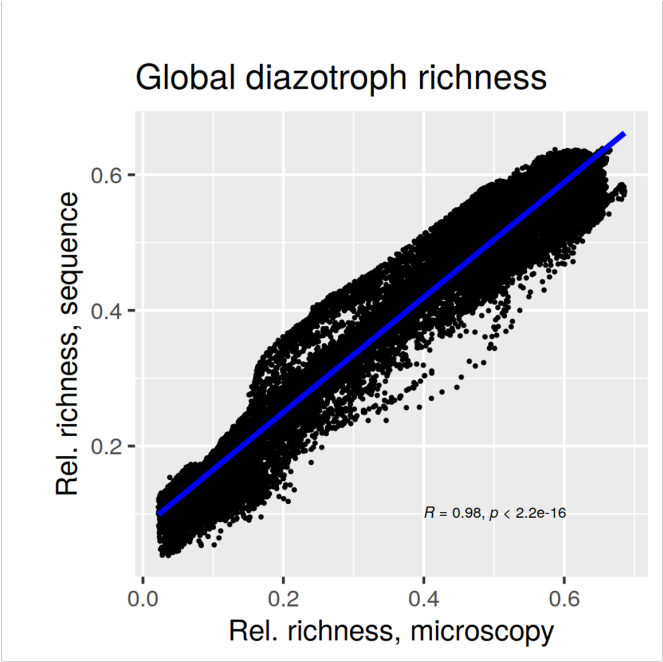

Fig. S2. Global Pearson correlation coefficient between model outputs derived from microscopy-based and sequence-based datasets from the total non-overlapping target-group approach from a GAM. Both axes show the normalized richness for each location (i.e., global ocean grid cell).

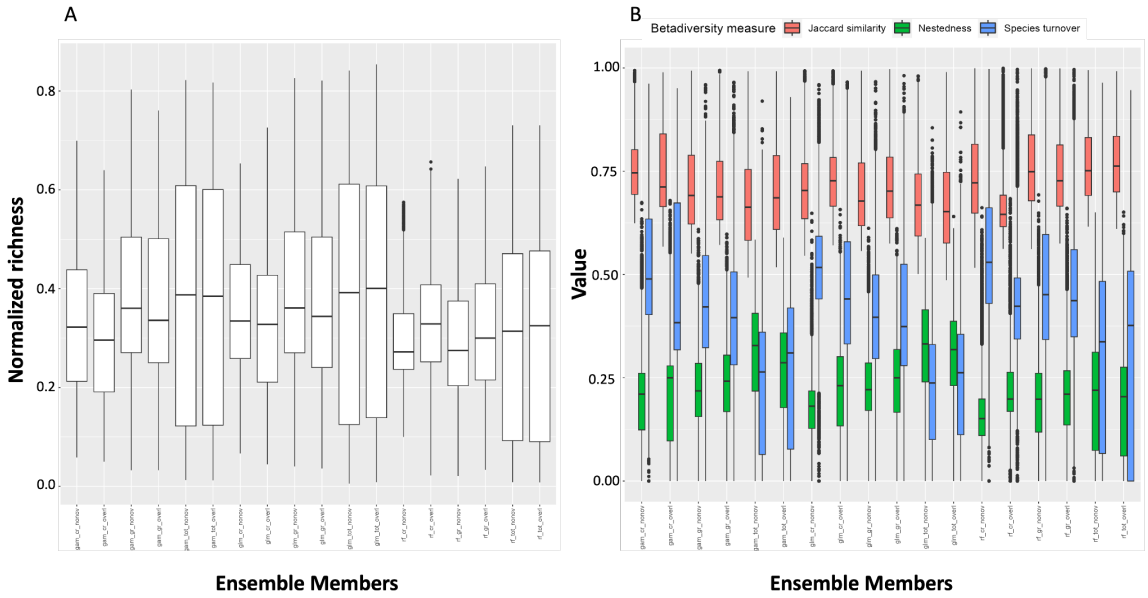

Fig. S3. Boxplots showing the range of diversity estimates across each ensemble member. The different ensemble members account for uncertainties related to algorithm choice (General Linear Model, glm; General Additive Model, gam; Random Forest, rf) and background selection strategy (total target group, tot; group-specific target group, gr; cruise-specific target group, cr) A) Shows each ensemble member on the x-axis and the normalized richness on the y-axis. B) Shows each ensemble member on the x-axis and the value of the corresponding betadiversity measure which is colored in pink (jaccard similarity), green (nestedness) or blue (species turnover).

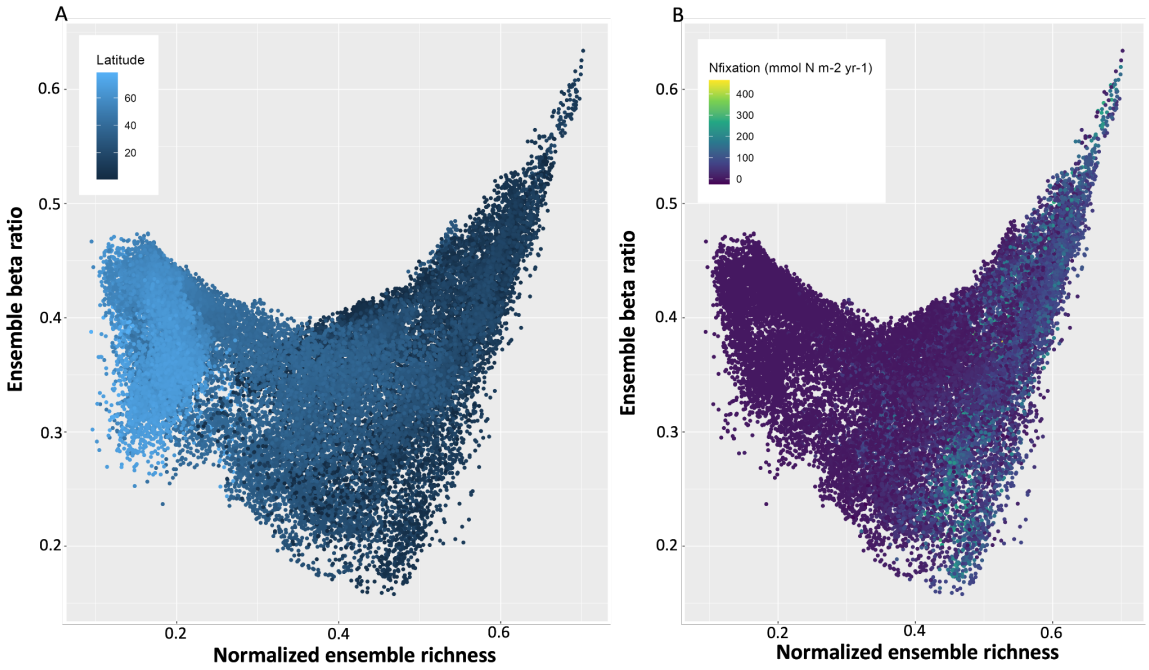

Fig. S4. Scatterplot between ensemble beta ratio and the normalized ensemble richness with points colored either by absolute latitude (A) or annual nitrogen fixation rates by Wang et al. (2019) (B). The beta ratio is calculated as the ratio between the ensemble nestedness and ensemble jacchard dissimilarity index. Each ensemble estimate is an average across eighteen Species Distribution Models that account for uncertainties related to predictors, algorithms, and background selection strategies. The spatial resolution is 1° longitude by 1° latitude on a monthly basis and coastal regions have been excluded by removing seas shallower than 200 meters and waters with surface salinity less than 20.

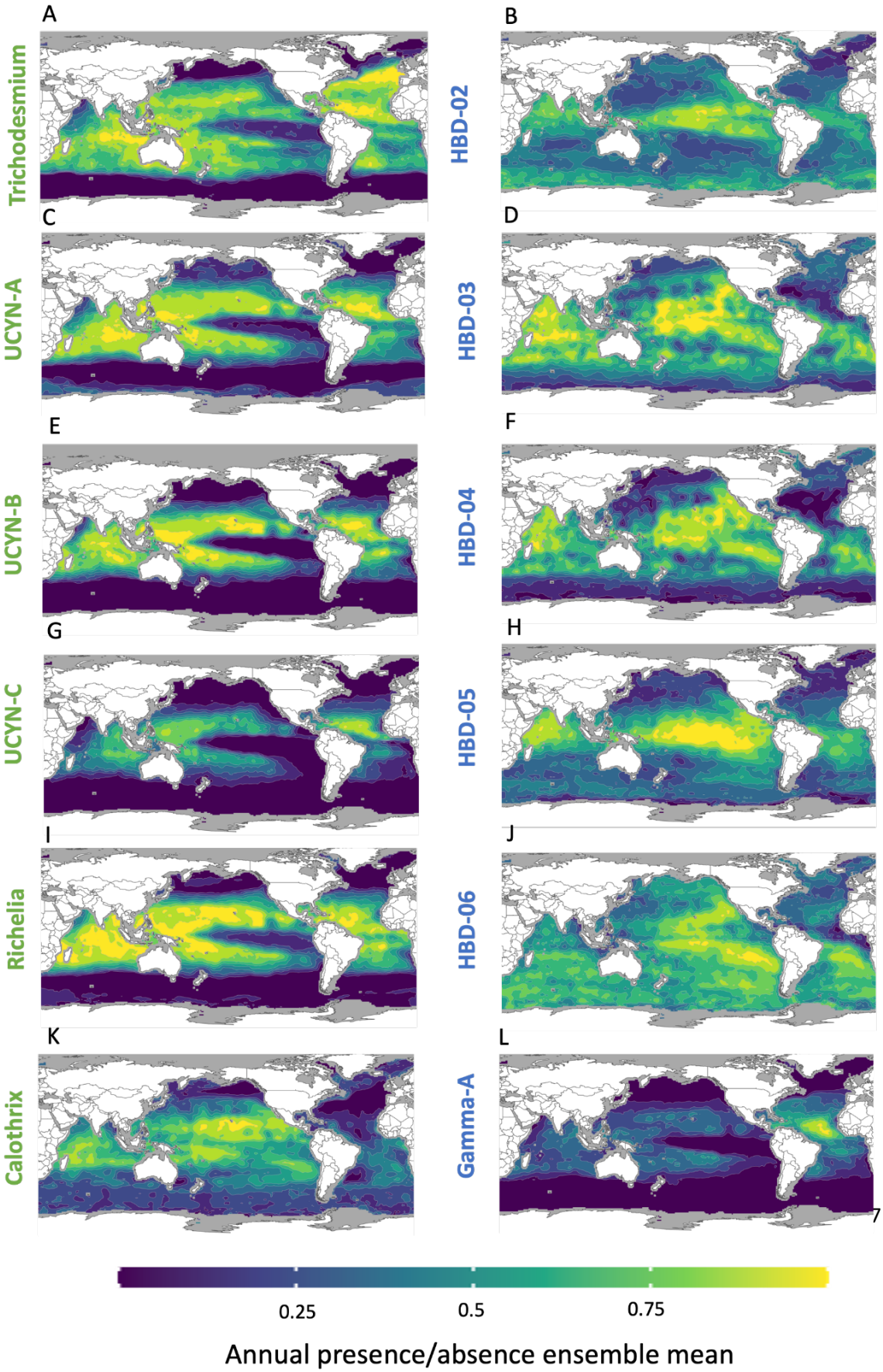
A
Trichodesmium
B
HBD-02
C
UCYN-A
D
HBD-03
E
UCYN-B
F
HBD-04
G
UCYN-C
H
HBD-05
I
Richelia
J
HBD-06
K
Calothrix
L
Gamma-A
0.25
0.5
0.75
Annual presence/absence ensemble mean

Fig. S5. Species-level spatial distribution of diazotrophs based on an ensemble across eighteen species distribution models. Each taxon was modeled separately for each month on a 1° longitude 1° latitude spatial resolution. The annual ensemble mean number of presences has been calculated by averaging across each output, with yellow indicating higher permanent annual presence while blue indicates the opposite with dark blue representing absences.

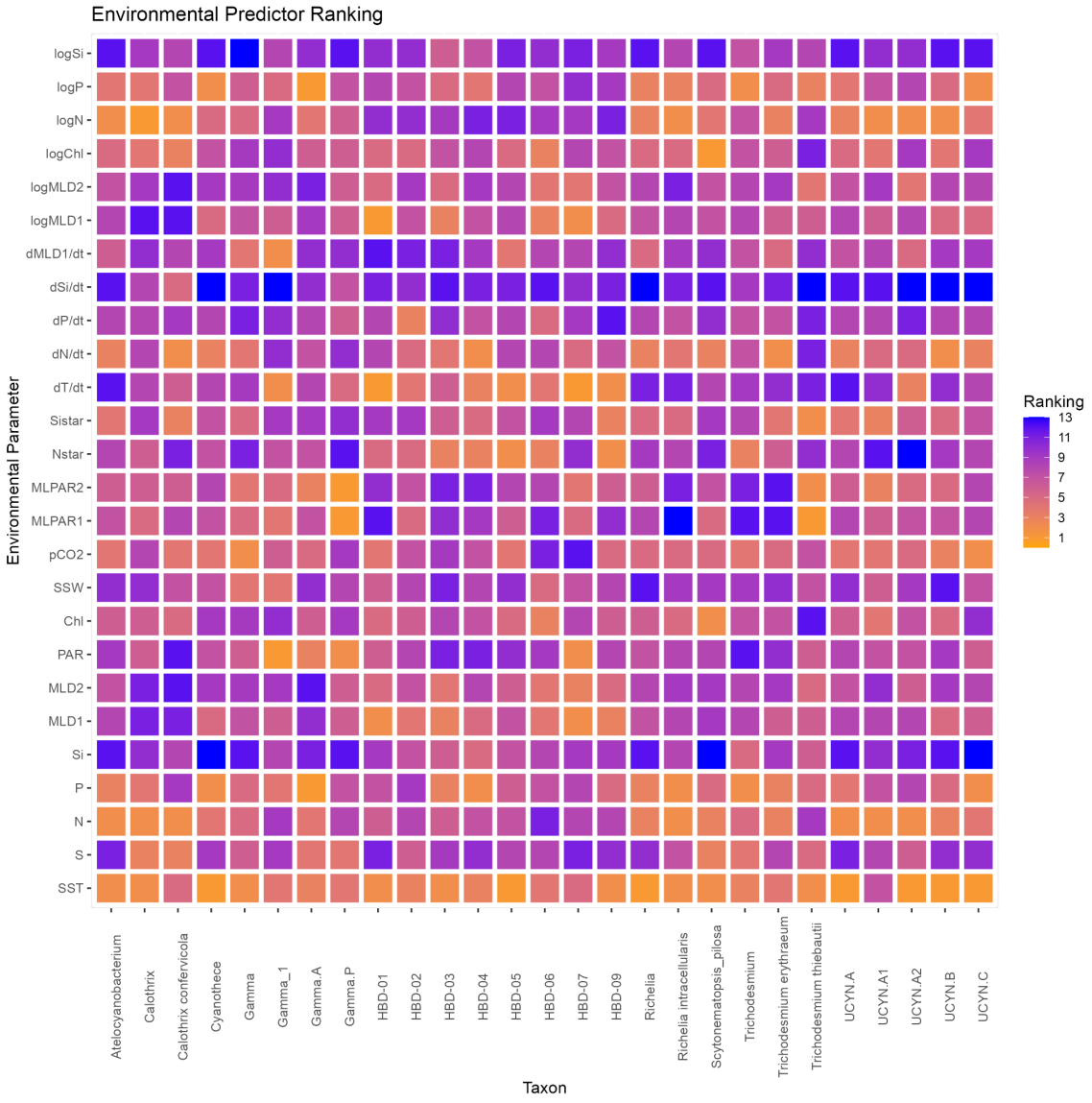

Fig. S6. Mean ranking of 26 environmental parameters based on single factor analysis on each diazotroph species applying three different algorithms namely General Linear Model (GLM), General Additive Model (GAM), and a Random Forest (RF).

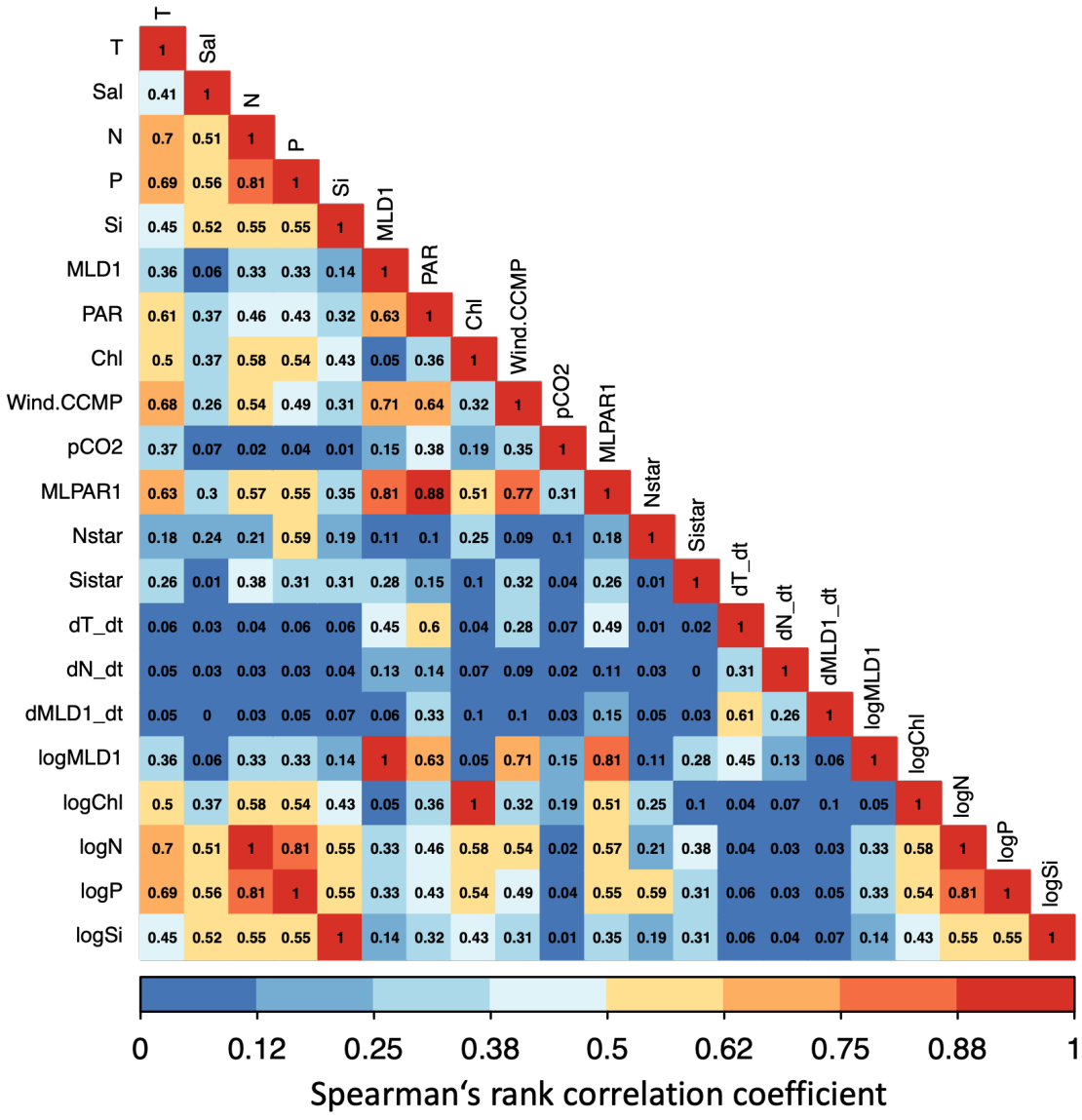

Fig. S7. Correlation heatmap showing the pairwise Spearman's rank correlation coefficients computed for the 21 environmental variables used for further analysis. When two environmental variables show a Spearman correlation coefficient > |0.7|, only one variable has been used within a set of predictors. Environmental variables: T (sea surface temperature, °C), Sal (sea surface

salinity), N (nitrate, μM), P (phosphate, μM), Si (silicic acid, μM), MLD1 (mixed layer depth, meters), PAR (photosynthetically active radiation, μmol m−2s−1), Chl (chlorophyll, μg liter−1), Wind.CCMP (sea surface wind stress, m s−), pCO2 (carbon dioxide partial pressure in the surface sea, μatm), MLPAR1 (photosynthetically available radiation over the mixed layer depth, μmol m−2s−1), Nstar (excess concentration of nitrate in relation to the redfield ratio, μM), Sistar (the ratio of nitrate to silicic acid, μM), dT_dt (temporal trends of sea surface temperature, °C), dN_dt (temporal trends of nitrate, μM), dMLD1_dt (temporal trends of mixed layer depth, meters), logMLD1 (logarithmic mixed layer depth, meters), logChl (logarithmic chlorophyll concentration, μg liter−1), logN (logarithmic nitrate concentration, μM), logP (logarithmic phosphate concentration, μM), logSi (logarithmic silicic acid concentration, μM).

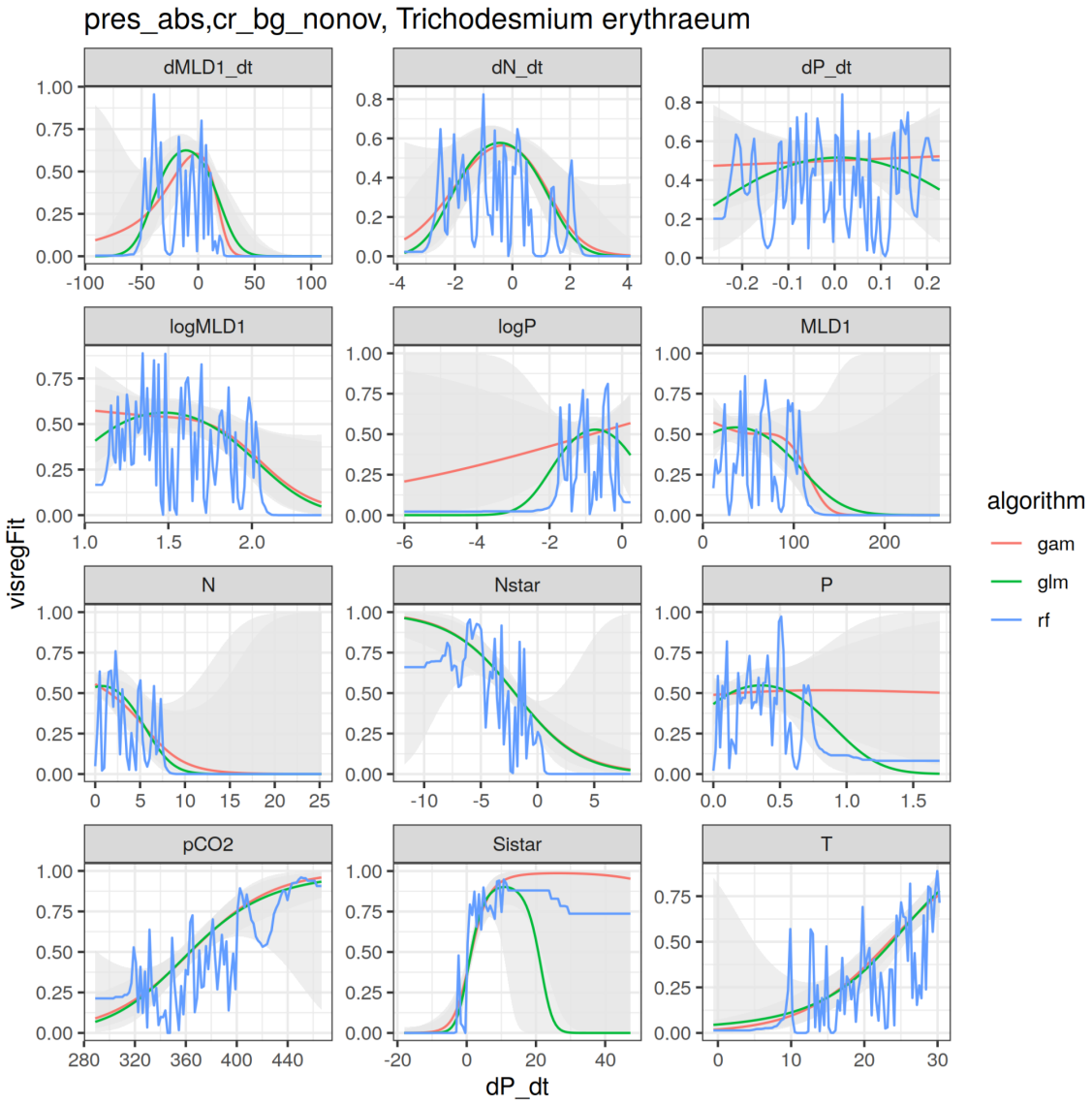

Fig. S8. Response curves *Trichodesmium*: Global Presence-Pseudoabsence maps and response curves fitted for the genus *Trichodesmium* by the three SDM algorithms (General Additive Model, GAM; General Linear Model, GLM's; Random Forest, Rf) for the environmental predictors involved in the ensemble predictor sets across the six background selection strategies. A-D) total non-overlapping background selection; B-E) total overlapping background selection; C-F) group-

specific non-overlapping; G-J) group-specific overlapping; H-K) cruise-specific non-overlapping; I-L) cruise-specific overlapping.

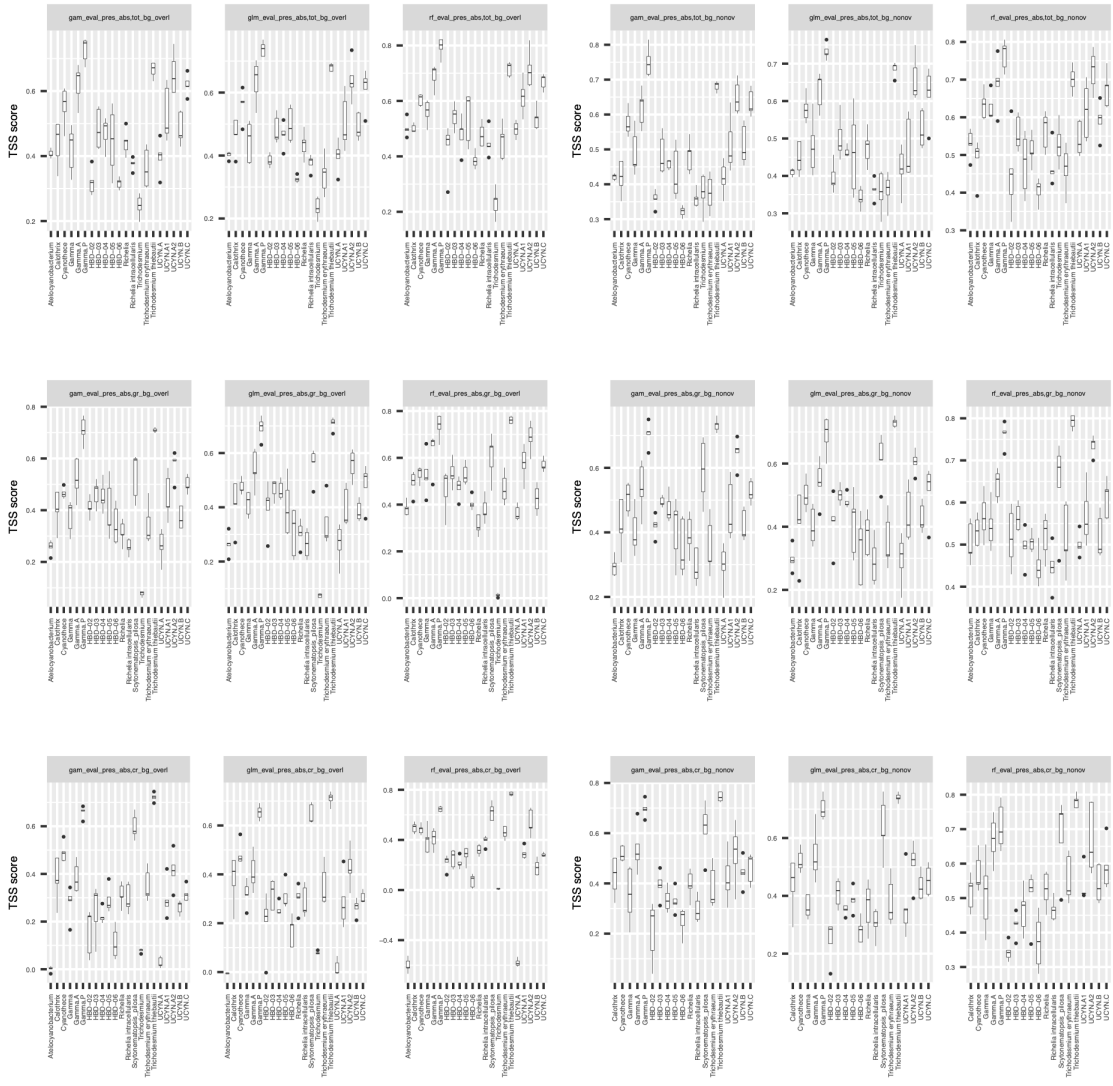

Fig. S9. Boxplots containing the TSS scores across the predictor ensemble sets. Each plot shows all diazotroph species included in the richness estimate for each background selection strategy. Richness estimates further include the phylotype *Gamma-A* and if *UCYN-A* was not successfully

modeled, we have added the individually modeled ecotypes UCYN-A1 and *UCYN-A2* to be included in the richness estimate due to their importance as a diazotroph. The boxplot titles encode for the algorithm (General Additive Model, gam; General Linear Model, glm; Random Forest, rf), background selection strategy (Total background, tot; group specific background, gr; cruise specific background, cr) and overlapping (overl) vs. non-overlapping (nonov) options.

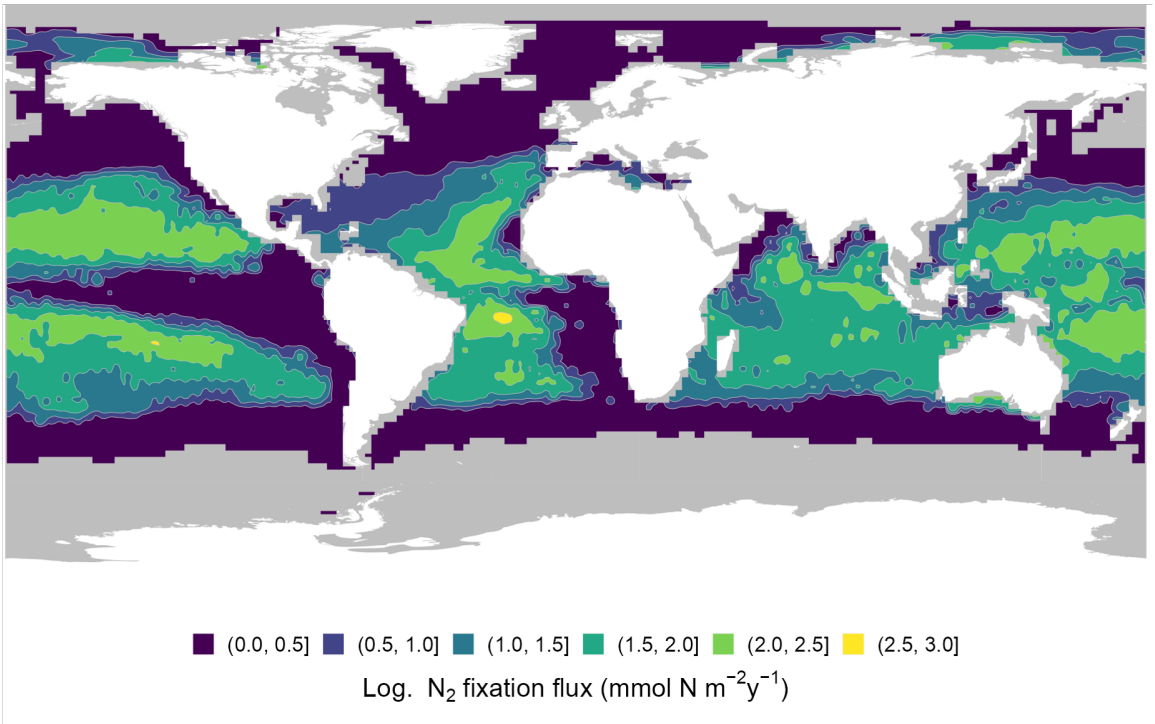

Fig. S10. Annual biological nitrogen fixation flux based on the publication by Wang et al. (2019).

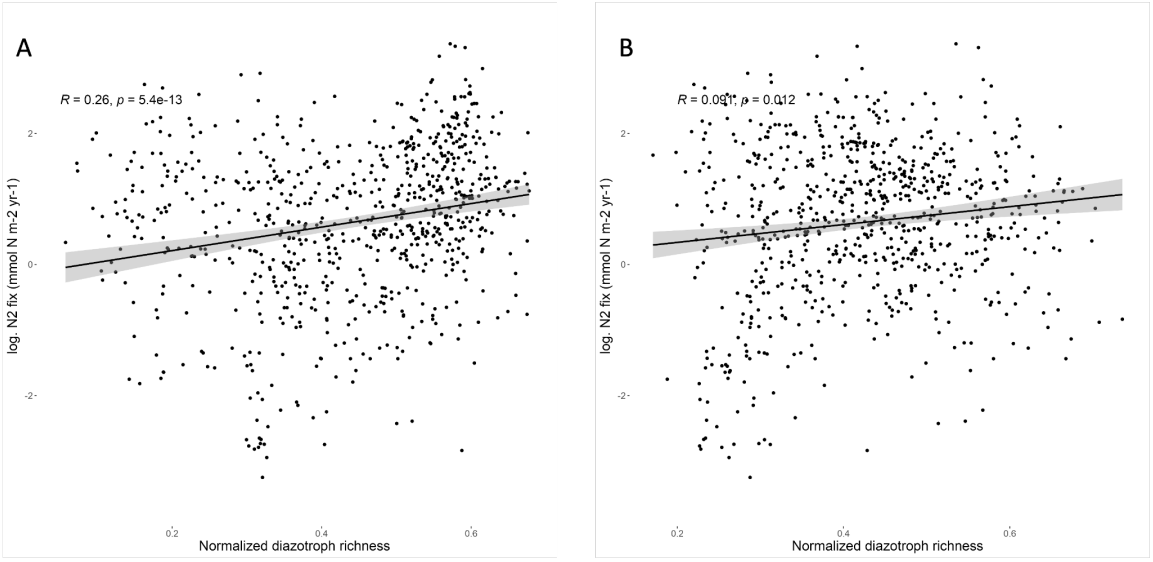

Fig. S11. Relationship between diazotroph species richness and biological nitrogen fixation (BNF). Shown are correlations between nitrogen fixation rates and global annual diazotroph richness of A) only cyanobacterial diazotrophs and B) only non-cyanobacterial diazotrophs. The black line indicates a 2nd order polynomial fit. Further indicated are Spearman's rank correlation coefficient and the statistical p-value. Grey shading indicates the 0.95 confidence interval.

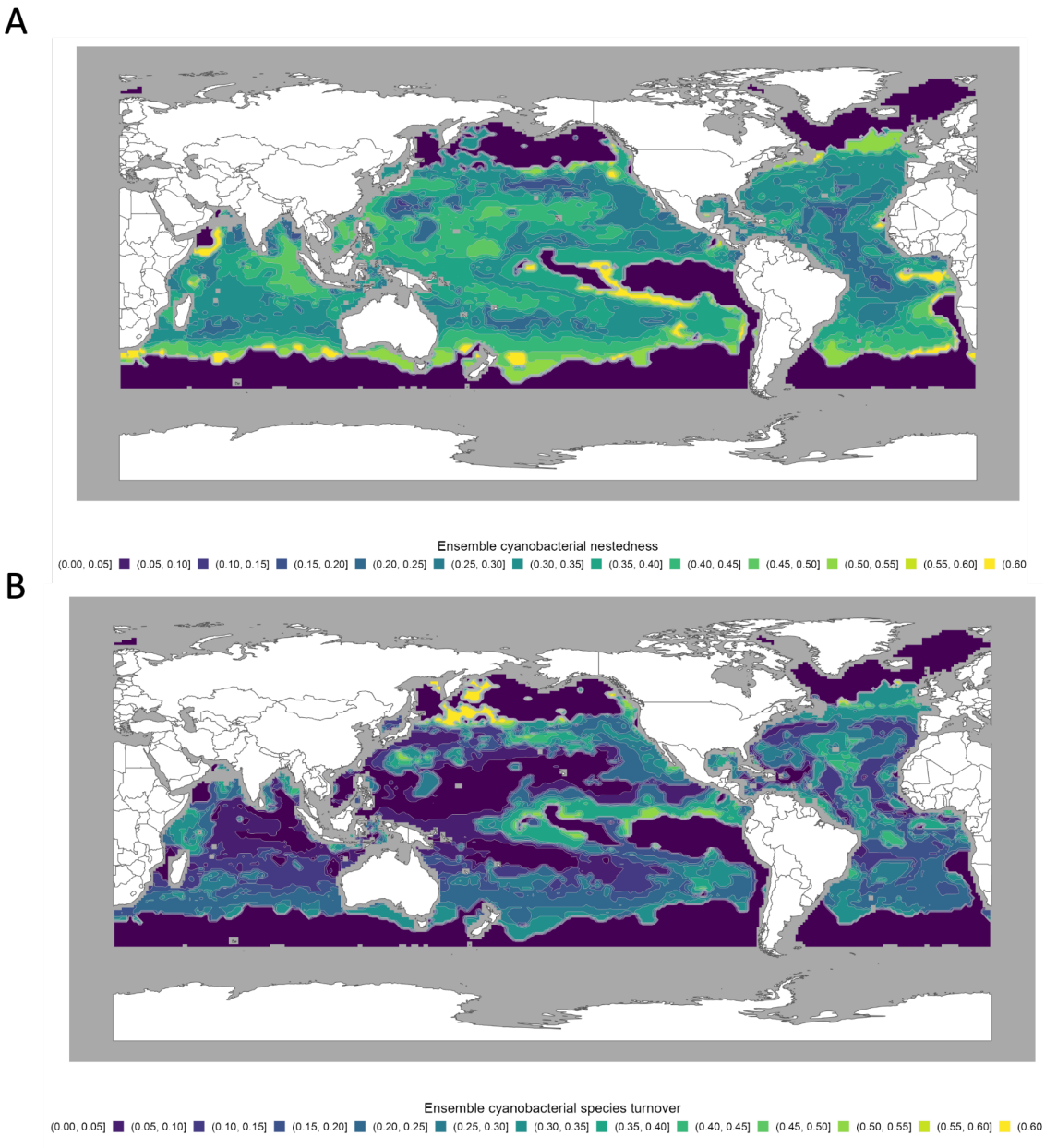

Fig. S12. Global beta diversity for cyanobacterial diazotrophs. A) Global nestedness based on a General Additive Model and B) Global species turnover. Blue color indicates low values and yellow indicates high estimates

Table S1. Environmental parameters that have been chosen in regard to reflect oceanic conditions that shape species' distributions via effects on physiology, growth, or species competition (20).

Variables were aggregated at a monthly (n = 12) climatological and globally gridded resolution (1° latitude × 1° longitude), as this was the best available resolution shared among datasets.

Candidate predictor	Variable nickname	Unit	Source
Sea surface temperatur	SST	degrees Celsius	World Ocean Atlas
Salinity	S	Practical salinity unit	World Ocean Atlas
Nitrate	N	µM	World Ocean Atlas
Phosphate	P	µM	World Ocean Atlas
Silicic acid	Si	µM	World Ocean Atlas
Mixed layer depth	MLD	meters	de Boyer Montégut 2004
Photosynthetically active radiation	PAR	µmol $m^{-2}s^{-1}$	Sea-Viewing Wide Field-of-view Sensor
Chlorophyll	Chl	µg $liter^{-1}$	Sea-Viewing Wide Field-of-view Sensor
Sea surface wind stress	SSW	$m\ s^{-1}$	Cross-Calibrated Multi-Platform
Carbon dioxide partial pressure	pCO2	μatm	Landschützer, Gruber Bakker 2015
Photosynthetically active radiation over mixed layer depth	MLPAR	µmol $m^{-2}s^{-1}$	Brun et al., 2015
Excess concentration of nitrate relative to phosphate according to Redfield Ratio	Nstar	µM	[Nitrate] - 16*[Phosphate]
Ratio of silicic acid to nitrate	Sistar	µM	[Silicic acid] / [Nitrate]
Temporal trends of sea surface temperature	dT/dt	degrees Celsius	Difference on centered mean of each month with neighboring months
Temporal trends of nitrate	dN/dt	µM	Difference on centered mean of each month with neighboring months
Temporal trends of phosphate	dP/dt	µM	Difference on centered mean of each month with neighboring months
Temporal trends of mixed layer depth	dMLD/dt	µM	Difference on centered mean of each month with neighboring months
Logarithmic mixed layer depth	logMLD	µM	
Logarithmic chlorophyll	logChl	µM	
Logarithmic nitrate	logN	µM	
Logarithmic phosphate	logP	µM	
Logarithmic silicic acid	logSi	µM	
Sea surface height anomaly	SSH	meters	Aviso

Table S2. Table showing each ensemble of predictor sets used to model each taxon to account for predictor uncertainties. Ensembles have been computed by randomly subsampling the top 10

environmental predictors that have ranked most important for each diazotroph taxa individually. Multicollinearity has been accounted for by removing parameters with Spearman's rank correlation coefficients higher than 0.7.

Taxon	Variable 1	Variable 2	Variable 3	Variable 4
Atelocyanobacterium	logP	Sistar	MLPAR2	T
Atelocyanobacterium	dN_dt	logP	Chl	T
Atelocyanobacterium	Chl	logN	dN_dt	Sistar
Atelocyanobacterium	pCO2	logN	MLPAR2	logChl
Atelocyanobacterium	N	pCO2	logChl	dMLD1_dt
Calothrix	T	MLPAR1	P	Chl
Calothrix	N	Chl	MLPAR1	T
Calothrix	Sal	logP	Nstar	logChl
Calothrix	Sal	Nstar	logP	logChl
Calothrix	P	PAR	dT_dt	dP_dt
Calothrix confervicola	dT_dt	logChl	dSi_dt	dN_dt
Calothrix confervicola	dN_dt	logChl	logN	dSi_dt
Calothrix confervicola	pCO2	Sal	Sistar	Chl
Calothrix confervicola	dT_dt	Sistar	pCO2	Sal
Calothrix confervicola	logN	Chl	T	Wind.CCMP
Cyanothece	Sistar	logN	logMLD1	pCO2
Cyanothece	pCO2	T	Sistar	N
Cyanothece	T	P	MLD1	dN_dt
Cyanothece	P	dN_dt	logMLD1	logChl
Cyanothece	MLD1	N	logChl	Nstar
Gamma	PAR	Sistar	Wind.CCMP	dMLD1_dt
Gamma	Sal	dN_dt	dMLD1_dt	MLPAR2
Gamma	T	PAR	pCO2	Sal
Gamma	pCO2	MLPAR2	Sistar	dN_dt
Gamma	T	MLPAR1	logN	dT_dt
Gamma_1	PAR	dT_dt	P	T
Gamma_1	PAR	T	Wind.CCMP	pCO2
Gamma_1	logP	dT_dt	Wind.CCMP	pCO2
Gamma_1	dMLD1_dt	P	MLPAR1	logSi
Gamma_1	MLPAR2	dMLD1_dt	logP	logSi
Gamma.A	P	logChl	PAR	Sal
Gamma.A	Sal	logN	T	pCO2
Gamma.A	T	logN	PAR	logChl
Gamma.A	MLPAR2	P	pCO2	Chl
Gamma.A	MLPAR2	N	Chl	dN_dt
Gamma.P	Sal	PAR	dT_dt	MLD2
Gamma.P	Sal	logN	MLPAR2	dT_dt
Gamma.P	MLD2	PAR	logChl	T
Gamma.P	logChl	logP	MLPAR2	T
Gamma.P	logN	MLPAR1	dP_dt	dSi_dt
HBD-01	T	Nstar	logMLD2	N
HBD-01	PAR	MLD2	dT_dt	T
HBD-01	Nstar	logMLD1	dT_dt	PAR
HBD-01	N	MLD1	pCO2	logChl
HBD-01	MLD1	pCO2	logChl	P
HBD-02	Sal	MLPAR1	T	dT_dt
HBD-02	dN_dt	MLPAR1	Sal	Nstar
HBD-02	Nstar	logChl	dN_dt	dP_dt
HBD-02	T	Chl	dT_dt	dP_dt
HBD-02	Chl	MLD1	pCO2	PAR
HBD-03	MLD2	P	dT_dt	Nstar
HBD-03	T	P	dT_dt	MLD2
HBD-03	logMLD1	T	logSi	Nstar
HBD-03	logSi	MLD1	dN_dt	logP
HBD-03	logMLD1	dN_dt	Si	logP
HBD-04	dT_dt	MLD1	P	dN_dt
HBD-04	logP	Si	T	MLD1
HBD-04	Si	logP	T	Nstar
HBD-04	dP_dt	dN_dt	Nstar	dT_dt
HBD-04	dP_dt	P	Sistar	logSi
HBD-05	dT_dt	P	Nstar	Sistar
HBD-05	Nstar	P	T	Chl
HBD-05	Sistar	logChl	dMLD1_dt	T
HBD-05	dMLD1_dt	logChl	pCO2	MLD2
HBD-05	dT_dt	pCO2	MLD2	Chl

Taxon	Variable 1	Variable 2	Variable 3	Variable 4
HBD-06	dT_dt	dP_dt	Chl	Nstar
HBD-06	logMLD1	Chl	T	dT_dt
HBD-06	logChl	dP_dt	logMLD1	T
HBD-06	Nstar	MLD1	logChl	dN_dt
HBD-06	logMLD2	dN_dt	P	dMLD1_dt
HBD-07	PAR	dT_dt	logMLD2	dN_dt
HBD-07	PAR	MLD1	dN_dt	T
HBD-07	T	MLD1	dT_dt	N
HBD-07	MLD2	N	Chl	dP_dt
HBD-07	MLD2	Chl	P	dP_dt
HBD-09	P	dT_dt	MLD1	Sistar
HBD-09	MLD1	dT_dt	Sistar	Chl
HBD-09	Wind.CCMP	Chl	Nstar	T
HBD-09	T	logMLD1	P	Nstar
HBD-09	logMLD1	logChl	pCO2	PAR
Richelia	pCO2	logN	logChl	Sistar
Richelia	dN_dt	pCO2	T	logN
Richelia	P	logChl	T	dMLD1_dt
Richelia	dN_dt	dMLD1_dt	N	Sistar
Richelia	logP	MLPAR2	Chl	dP_dt
Richelia intracellularis	logN	dN_dt	pCO2	logChl
Richelia intracellularis	T	N	dN_dt	logChl
Richelia intracellularis	pCO2	P	Chl	Sistar
Richelia intracellularis	P	T	Chl	Sistar
Richelia intracellularis	logN	Si	dP_dt	Sal
Scytonematopsis_pilosa	Sal	logChl	MLPAR1	N
Scytonematopsis_pilosa	MLPAR1	Chl	dN_dt	logP
Scytonematopsis_pilosa	logP	Chl	dN_dt	T
Scytonematopsis_pilosa	Sal	logChl	N	T
Scytonematopsis_pilosa	pCO2	P	dT_dt	MLPAR2
Trichodesmium	logP	pCO2	Nstar	Sal
Trichodesmium	pCO2	N	Si	logChl
Trichodesmium	logChl	logN	T	Sal
Trichodesmium	logP	Nstar	T	Si
Trichodesmium	N	dP_dt	dMLD1_dt	Chl
Trichodesmium erythraeum	T	dMLD1_dt	logMLD1	N
Trichodesmium erythraeum	dMLD1_dt	pCO2	dN_dt	N
Trichodesmium erythraeum	T	Sistar	logP	logMLD1
Trichodesmium erythraeum	P	pCO2	Sistar	dN_dt
Trichodesmium erythraeum	logP	Nstar	MLD1	dP_dt
Trichodesmium thiebautii	T	MLPAR1	P	Sistar
Trichodesmium thiebautii	Sistar	logP	T	PAR
Trichodesmium thiebautii	PAR	logP	MLD2	Sal
Trichodesmium thiebautii	MLPAR2	Sal	P	Si
Trichodesmium thiebautii	MLD2	Si	pCO2	N
UCYN.A	dN_dt	N	pCO2	T
UCYN.A	logN	Sistar	pCO2	dN_dt
UCYN.A	T	P	MLPAR2	logChl
UCYN.A	logP	Sistar	logChl	MLPAR2
UCYN.A	logP	dMLD1_dt	Chl	MLD1
UCYN.A1	MLPAR2	Sistar	T	logChl
UCYN.A1	pCO2	Chl	MLPAR2	Sistar
UCYN.A1	logChl	dN_dt	pCO2	T
UCYN.A1	dN_dt	Wind.CCMP	logN	Chl
UCYN.A1	N	Wind.CCMP	dMLD1_dt	PAR
UCYN.A2	logMLD2	logN	dMLD1_dt	T
UCYN.A2	Sistar	dN_dt	N	pCO2
UCYN.A2	T	Sistar	N	dN_dt
UCYN.A2	pCO2	dMLD1_dt	MLD2	logN
UCYN.A2	logMLD2	dT_dt	Sal	PAR
UCYN.B	N	pCO2	logChl	T
UCYN.B	MLPAR2	pCO2	dN_dt	T
UCYN.B	dN_dt	logN	logChl	MLPAR2
UCYN.B	logN	Chl	MLD1	Sistar
UCYN.B	N	Chl	MLD1	Sistar
UCYN.C	T	MLD1	dN_dt	N
UCYN.C	dN_dt	pCO2	T	MLD1
UCYN.C	P	logMLD1	pCO2	PAR
UCYN.C	logMLD1	P	PAR	dP_dt
UCYN.C	logP	Wind.CCMP	dP_dt	dT_dt

SI References

1. Y.-W. Luo, *et al.*, Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates. *Earth Syst. Sci. Data* **4**, 47–73 (2012).

2. W. Tang, N. Cassar, Data-driven modeling of the distribution of diazotrophs in the global ocean. *Geophys. Res. Lett.* **46**, 12258–12269 (2019).

3. M. R. Gradoville, *et al.*, Latitudinal constraints on the abundance and activity of the cyanobacterium UCYN-A and other marine diazotrophs in the North Pacific. *Limnol. Oceanogr.* **65**, 1858–1875 (2020).

4. A. M. S. Detoni, A. Subramaniam, S. T. Haley, S. T. Dyhrman, P. H. R. Calil, Cyanobacterial diazotroph distributions in the western South Atlantic. *Front. Mar. Sci.* **9** (2022).

5. C. Martínez-Pérez, *et al.*, The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat Microbiol* **1**, 16163 (2016).

6. D. Righetti, M. Vogt, N. E. Zimmermann, M. D. Guiry, N. Gruber, PhytoBase: A global synthesis of open-ocean phytoplankton occurrences. *Earth Syst. Sci. Data* **12**, 907–933 (2020).

7. J. J. Pierella Karlusich, *et al.*, Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. *Nat. Commun.* **12**, 4160 (2021).

8. L. Paoli, *et al.*, Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).

9. Flanders Marine Institute (VLIZ), Belgium, Global Oceans and Seas, version 1 (2021) https:/doi.org/10.14284/542.

10. T. O. Delmont, *et al.*, Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* **3**, 804–813 (2018).

11. S. Sunagawa, *et al.*, Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

12. G. Salazar, *et al.*, Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068–1083.e21 (2019).

13. S. J. Biller, *et al.*, Marine microbial metagenomes sampled across space and time. *Sci Data* **5**, 180176 (2018).

14. S. G. Acinas, *et al.*, Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun Biol* **4**, 604 (2021).

15. T. Klemetsen, *et al.*, The MAR databases: development and implementation of databases

specific for marine metagenomics. *Nucleic Acids Res.* **46**, D692–D699 (2018).

16. M. G. Pachiadaki, *et al.*, Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* **179**, 1623–1635.e11 (2019).

17. D. H. Parks, *et al.*, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

18. D. Righetti, M. Vogt, N. Gruber, A. Psomas, N. E. Zimmermann, Global pattern of phytoplankton diversity driven by temperature and environmental variability. *Sci Adv* **5**, eaau6253 (2019).

19. P. Brun, *et al.*, Ecological niches of open ocean phytoplankton taxa. *Limnol. Oceanogr.* **60**, 1020–1038 (2015).

20. P. W. Boyd, R. Strzepek, F. Fu, D. A. Hutchins, Environmental control of open-ocean phytoplankton groups: Now and in the future. *Limnol. Oceanogr.* **55**, 1353–1376 (2010).

21. C. Amante, B. W. Eakins, ETOPO1 arc-minute global relief model: procedures, data sources and analysis (2009).

22. M. M. Zweng, *et al.*, World ocean atlas 2018, volume 2: Salinity (2019).

23. K. A. Whittaker, T. A. Rynearson, Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 2651–2656 (2017).

24. S. Lehtinen, T. Tamminen, R. Ptacnik, T. Andersen, Phytoplankton species richness, evenness, and production in relation to nutrient availability and imbalance. *Limnol. Oceanogr.* **62**, 1393–1408 (2017).

25. J. Ladau, *et al.*, Global marine bacterial diversity peaks at high latitudes in winter. *ISME J.* **7**, 1669–1677 (2013).

26. F. Benedetti, *et al.*, Major restructuring of marine plankton assemblages under global warming. *Nat. Commun.* **12**, 5226 (2021).

27. S. J. Phillips, *et al.*, Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* **19**, 181–197 (2009).

28. M. Barbet-Massin, F. Jiguet, C. H. Albert, W. Thuiller, Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.* **3**, 327–338 (2012).

29. L. Buisson, W. Thuiller, N. Casajus, S. Lek, G. Grenouillet, Uncertainty in ensemble forecasting of species distribution. *Glob. Chang. Biol.* **16**, 1145–1157 (2010).

30. P. Brun, *et al.*, Model complexity affects species distribution projections under climate change. *J. Biogeogr.* **47**, 130–142 (2020).

31. C. F. Dormann, *et al.*, Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27–46 (2013).

32. E. A. Freeman, G. Moisen, PresenceAbsence: An R Package for Presence Absence Analysis. *J. Stat. Softw.* **23**, 1–31 (2008).

33. A. Baselga, C. D. L. Orme, betapart : an R package for the study of beta diversity. *Methods Ecol. Evol.* **3**, 808–812 (2012).

34. W.-L. Wang, J. K. Moore, A. C. Martiny, F. W. Primeau, Convergent estimates of marine nitrogen fixation. *Nature* **566**, 205–211 (2019).

35. Z. Shao, *et al.*, Version 2 of the global oceanic diazotroph database. *Earth Syst. Sci. Data Discuss.* (2023) https:/doi.org/10.5194/essd-2023-13.